

Vladimir Spokoiny

Nonparametric estimation: parametric view

February 7, 2017

Springer

Berlin Heidelberg New York
Hong Kong London
Milan Paris Tokyo

Contents

Part I Model selection for linear methods

1 Quasi maximum likelihood estimation in linear models	17
1.1 Linear Modeling	17
1.1.1 Estimation under homogeneous noise assumption	19
1.1.2 Linear basis transformation	20
1.1.3 Orthogonal and orthonormal design	22
1.1.4 Spectral representation	23
1.2 Properties of the response estimate \tilde{f}	24
1.2.1 Decomposition into a deterministic and a stochastic component ...	24
1.2.2 Properties of the operator Π	25
1.2.3 Quadratic loss and risk of the response estimation	26
1.2.4 Misspecified “colored noise”	27
1.3 Properties of the MLE $\tilde{\theta}$	28
1.3.1 Properties of the stochastic component	28
1.3.2 Properties of the deterministic component	29
1.3.3 Risk of estimation. R-efficiency	31
1.3.4 The case of a misspecified noise	33
1.4 Linear models and quadratic log-likelihood	34
1.4.1 Inference based on the maximum likelihood	37
1.4.2 A misspecified LPA	40
1.4.3 A misspecified noise structure	40
2 Linear regression with random design	43
2.1 Random design linear regression	43
2.2 Design matrix and design distribution	43
2.2.1 Design with independent measurements	44
2.2.2 Aggregated random design	45

4 Contents

2.3	Fisher and Wilks expansions for the MLE under random design	47
2.4	A deviation bound for ξ	50
2.5	Misspecified linear modeling assumption	51
2.6	Application to instrumental regression	54
3	Linear smoothers	57
3.1	Regularization and ridge regression	58
3.2	Penalized likelihood. Bias and variance	58
3.3	Inference for the penalized MLE	61
3.4	Projection and shrinkage estimates	62
3.5	Smoothness constraints and roughness penalty approach	64
3.6	Shrinkage in a linear inverse problem	65
3.7	Spectral cut-off and spectral penalization. Diagonal estimates	65
3.8	Roughness penalty and random design	67
4	Sieve model selection in linear models	73
4.1	Projection estimation. Loss and risk	73
4.1.1	A linear model	73
4.1.2	Linear decomposition of the estimator $\tilde{\boldsymbol{\theta}}$ and quadratic risk	74
4.1.3	The case of Inhomogeneous errors	74
4.1.4	Prediction error $\tilde{\mathbf{f}} - \mathbf{f}^*$	75
4.1.5	Quadratic loss. Bias-variance decomposition	75
4.1.6	Projection estimation and the model choice problem	76
4.2	Unbiased risk estimation	78
4.2.1	AIC and pairwise comparison	80
4.2.2	Pairwise analysis	82
4.2.3	Uniform bounds and the zone of insensitivity	84
4.2.4	A bound on the excess	85
4.3	The approach based on multiple testing. “Smallest accepted” rule	87
4.3.1	A LR test	88
4.3.2	Multiplicity correction	89
4.3.3	Definition of the oracle and propagation property	90
4.3.4	A bound on the loss	91
4.3.5	Role of β	93
5	Ordered model selection for linear smoothers	95
5.1	Model and problem. Known noise variance	98
5.1.1	Smallest accepted (SmA) method in ordered model selection	101

5.1.2	Oracle choice	102
5.1.3	Tail function, multiplicity correction, critical values z_{m,m°	103
5.1.4	SmA choice and the oracle inequality	104
5.1.5	Analysis of the payment for adaptation \bar{z}_{m^*}	106
5.1.6	Power loss function	106
5.1.7	Application to projection estimation	109
5.1.8	Linear functional estimation	111
5.2	Bootstrap tuning	112
5.2.1	Presmoothing and wild bootstrap	112
5.2.2	Bootstrap validation. Range of applicability	116
5.3	Simulations	118
5.4	Proofs	121
5.4.1	Proof of Theorem 5.1.1	121
5.4.2	Proof of Proposition 5.1.2	122
5.4.3	Proof of Theorem 5.1.3	123
5.4.4	Proof of Proposition 5.1.1	124
5.4.5	Proof of Theorem 5.2.1	125
5.4.6	Proof of Theorem 5.2.2	126
5.4.7	Proof of Theorem 5.2.3	127
5.5	Linear non-Gaussian case and GAR	128
6	Unordered case. Anisotropic sets and subset selection	129
6.1	Subset selection procedure	129
6.1.1	SmA procedure and multilevel synchronization	130
6.1.2	Prediction loss	133
6.1.3	Estimation loss	134
6.1.4	Linear functional estimation	134
6.1.5	Subset selection problem	135
6.2	Anisotropic models	137
7	SmA and parameter tuning in high dimensional regression	141
7.1	SmA subset selection in high dimensional regression	142
8	Penalized model selection	147
8.1	Complexity penalization	147
8.1.1	Penalty tuning using propagation condition	149
8.1.2	Oracle inequality for $\hat{\chi}$ -choice	151
8.1.3	Numerical results	151

8.1.4	Bootstrap based tuning of penalty	151
8.2	Sparse penalty	152
8.2.1	Basic inequality	154
8.2.2	Dual problem and Danzig selector	155
8.2.3	Data-driven choice of λ	156

Part II General parametric theory

9	Fisher and Wilks expansion	161
9.1	Main results	162
9.2	Non-Gaussian case: conditions	163
9.3	Properties of the MLE $\tilde{\theta}$	166
9.4	Some auxiliary results and proofs	168
9.4.1	Local linear approximation of the gradient of the log-likelihood	168
9.4.2	Local quadratic approximation of the log-likelihood	170
9.4.3	Proof of Theorem 9.3.1	171
9.4.4	Proof of Theorem 9.3.2	172
9.4.5	Proof of Theorem 9.3.3	172
10	Bernstein – von Mises Theorem.....	175
10.1	Parametric BvM Theorem.....	176
10.2	The use of posterior mean and variance for credible sets.....	177
10.3	Extension to a flat Gaussian prior	179
10.4	Proof of Theorem 10.1.1	180
10.4.1	Local Gaussian approximation of the posterior. Upper bound.....	181
10.4.2	Tail posterior probability and contraction	183
10.4.3	Local Gaussian approximation of the posterior. Lower bound.....	185
10.4.4	Moments of the posterior	186
10.5	Proof of Theorem 10.3.1	187
11	Roughness penalty for dimension reduction	189
11.1	Fisher and Wilks Theorems under quadratic penalization	192
11.2	Effective dimension	194
11.3	Conditions	195
11.4	Concentration and a large deviation bound	198
11.5	Wilks and Fisher expansions	200
11.6	Quadratic risk bound and modeling bias	201
11.7	Proofs of the Fisher and Wilks expansions	203

11.8 Nonparametric BvM Theorem: Gaussian case	206
11.8.1 Finite dimensional projections and maxispaces	208
11.8.2 Concentration sets for the posterior	210
11.8.3 Frequentist coverage for Bayesian credible sets	211
11.8.4 Non-Gaussian errors	213
11.9 Nonparametric BvM Theorem: non-Gaussian case	214
11.9.1 A linear stochastic term	215
11.9.2 General likelihood	221
12 Semiparametric estimation	225
12.1 Fisher and Wilks results for a parameter subvector	226
12.1.1 Fisher expansion and semiparametric concentration	227
12.1.2 Semiparametric Wilks expansion	229
12.2 Likelihood ratio test statistic for a composite hypothesis	230
12.3 Semiparametric BvM approximation	232
12.4 Sieve semiparametric inference	237
12.5 Estimation of a nonlinear functional	237
12.6 Bias in semiparametric sieve approximation	239
12.6.1 Basis transformation for the nuisance	240
12.6.2 Smoothness conditions	241
13 Parametric i.i.d. models	247
13.1 Quasi MLE in an i.i.d. model	247
13.2 Conditions in the i.i.d. case	248
13.3 Results in the non-penalized i.i.d. case	250
13.4 Roughness penalization for an i.i.d. sample	251
13.5 BvM Theorem for the i.i.d. data	253
14 Generalized linear models	255
14.1 Linear models	255
14.2 Generalized linear models (GLM)	257
14.2.1 A general deviation bound for the MLE $\tilde{\theta}$	258
14.2.2 Fisher and Wilks expansions for $\tilde{\theta}$	259
14.2.3 Sufficient conditions on design and errors	260
14.3 Nonparametric sieve GLM estimation	264
14.4 Estimation for a penalized GLM	266
14.5 BvM Theorem for a GLM	268
14.5.1 A non-informative prior	268

14.5.2 Nonparametric BvM with a Gaussian prior	271
14.6 GLM with random design	272
14.6.1 Local concentration of $\tilde{\theta}$	274
14.6.2 Fisher and Wilks expansions	275
14.6.3 Sufficient conditions for the case of random design	276
14.6.4 Nonparametric BvM for a Gaussian prior	281
15 Estimation of a log-density	283
15.1 Log-density estimation. Conditions	283
15.2 Sieve nonparametric density estimation	290
15.3 Sieve likelihood ratio test	296
15.4 Error of estimation for the log-density function	297
15.4.1 Kullback-Leibler divergence	298
15.4.2 Hellinger loss	299
15.5 Penalized smooth density estimation	301
15.5.1 Loss in penalized density estimation	305
15.6 Parametric structural log-density modeling	306
16 Sieve parametric approach in nonparametric regression	313
16.1 Parametric and nonparametric regression	313
16.2 Conditions	314
16.2.1 Checking the local conditions (ED_0) and (ED_2)	316
16.2.2 Checking the local condition (L_0)	317
16.3 Large deviation result and Fisher expansion	322
16.4 Prediction loss and bias-variance decomposition	323
16.5 Sieve nonparametric estimation	324
16.6 Penalized regression	325
16.7 Semiparametric problem	326
16.8 Random design regression	330
16.8.1 Checking the condition (L_0)	332
16.8.2 Checking the conditions (ED_0) and (ED_2)	333
17 Structural regression	339
17.1 Single-index case	339
17.2 Error-in-variable nonparametric regression	345
17.3 Instrumental regression	349
18 Median and quantile regression	355

19 Generalized regression	357
--	------------

Part III Structural regression

20 Sieve Model Selection	361
20.1 Sieve SmA procedure	361
20.2 Resampling methods for parameter tuning in generalized regression	363
20.2.1 Generalized regression	363
20.2.2 Multiplier bootstrap	364
20.2.3 Numerical issues	366
20.3 Sieve Generalized Linear regression	367
20.3.1 Sieve MLE	368
20.3.2 Bootstrap counterpart	369
20.3.3 Bootstrap for the SmA procedure	370

Part IV Mathematical tools

A Some results for Gaussian law	373
A.1 Deviation bounds for a Gaussian vector	373
A.2 Gaussian integrals	374
B Deviation bounds for quadratic forms	379
B.1 Gaussian quadratic forms	379
B.2 Deviation bounds for non-Gaussian quadratic forms	382
B.2.1 Deviation bounds for the norm of a standardized non-Gaussian vector	383
B.2.2 A deviation bound for a general non-Gaussian quadratic form	387
B.3 Deviation probability for a normalized martingale	390
C Sums of random matrices	393
C.1 Matrix Bernstein inequality	393
C.2 Presmoothing and bias effects	400
C.3 Empirical covariance matrix	403
D Gaussian comparison via KL-divergence and Pinsker's inequality	405
D.1 Pinsker's inequality	405
D.2 Gaussian comparison	406

E	Random multiplicity correction	409
E.1	Gaussian measures with random covariance	409
E.2	Max-case	411
F	High-dimensional inference for a Gaussian law	413
F.1	Stein identity, Slepian bridge, and Gaussian comparison	413
F.2	Comparing of the maximum of Gaussians	416
F.3	Anti-concentration for Gaussian maxima	417
G	Gaussian approximation of a vector sum	419
G.1	A univariate case with Lindeberg telescopic sums	419
G.2	Berry-Esseen Theorem for a univariate sum	421
G.2.1	Characteristic functions for a univariate sum	422
G.2.2	Characteristic function of a sum. Simmerization	424
G.2.3	Methods based on the Fourier-Stieltjes transform	427
G.2.4	Berry-Esseen Theorem	430
G.2.5	Fourier transform for the norm of a vector	430
G.2.6	Fourier transform for the squared norm of a vector	432
G.3	GAR for the Euclidean norm of a vector sum	434
G.4	GAR for the sup-norm of a vector sum	437
G.5	GAR for the sup-norm of a vector sum. Improved	440
G.6	GAR for weighted sums	444
G.7	A uniform bound for the maximum of the norm of weighted vector sums	445
H	Deviation bounds for random processes	449
H.1	Chaining and covering numbers	449
H.2	Entropy and Dudley's integral	455
H.3	A local bound with generic chaining	455
H.4	Generic chaining with partitioning	458
H.5	A large deviation bound	458
H.6	Finite-dimensional smooth case	459
H.6.1	Covering and entropy for Euclidean distance	460
H.6.2	Generic chaining	462
H.7	Entropy of an ellipsoid	463
H.8	Roughness constraints for dimension reduction	468
H.9	Bound for a bivariate process	469
H.10	A bound for the norm of a vector random process	471
H.11	A bound for a family of quadratic forms	472

H.12 A bound for a smooth quadratic field	473
References	475

Part I

Model selection for linear methods

. . . the world will remain an eternal,
may be, comprehensible,
but still endless.

J. Brodsky

Quasi maximum likelihood estimation in linear models

1.1 Linear Modeling

A linear model assumes that the observations Y_i follow the equation:

$$Y_i = \Psi_i^\top \boldsymbol{\theta}^* + \varepsilon_i \quad (1.1)$$

for $i = 1, \dots, n$, where $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_p^*)^\top \in \mathbb{R}^p$ is an unknown parameter vector, Ψ_i are given vectors in \mathbb{R}^p and the ε_i 's are individual errors with zero mean. A typical example is given by linear regression when the vectors Ψ_i are the values of a set of functions (e.g. polynomial, trigonometric) series at the design points X_i .

A linear Gaussian model assumes in addition that the vector of errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$ is normally distributed with zero mean and a covariance matrix Σ :

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma).$$

In this chapter we suppose that Σ is given in advance. We will distinguish between three cases:

1. the errors ε_i are i.i.d. $\mathcal{N}(0, \sigma^2)$, or equivalently, the matrix Σ is equal to $\sigma^2 I_n$ with I_n being the unit matrix in \mathbb{R}^n .
2. the errors are independent but not homogeneous, that is, $E\varepsilon_i^2 = \sigma_i^2$. Then the matrix Σ is diagonal: $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$.
3. the errors ε_i are dependent with a covariance matrix Σ .

In practical applications one mostly starts with the white Gaussian noise assumption and more general cases 2 and 3 are only considered if there are clear indications of the noise inhomogeneity or correlation. The second situation is typical e.g. for the eigenvector decomposition in an inverse problem. The last case is the most general and includes the first two.

Denote by $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ (resp. $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$) the vector of observations (resp. of errors) in \mathbb{R}^n and by Ψ the $p \times n$ matrix with columns Ψ_i . Let also Ψ^\top denote its transpose. Then the model equation can be rewritten as:

$$\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma).$$

An equivalent formulation is that $\Sigma^{-1/2}(\mathbf{Y} - \Psi^\top \boldsymbol{\theta})$ is a standard normal vector in \mathbb{R}^n . The log-density of the distribution of the vector $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ w.r.t. the Lebesgue measure in \mathbb{R}^n is therefore of the form

$$\begin{aligned} L(\boldsymbol{\theta}) &= -\frac{n}{2} \log(2\pi) - \frac{\log(\det \Sigma)}{2} - \frac{1}{2} \|\Sigma^{-1/2}(\mathbf{Y} - \Psi^\top \boldsymbol{\theta})\|^2 \\ &= -\frac{n}{2} \log(2\pi) - \frac{\log(\det \Sigma)}{2} - \frac{1}{2} (\mathbf{Y} - \Psi^\top \boldsymbol{\theta})^\top \Sigma^{-1} (\mathbf{Y} - \Psi^\top \boldsymbol{\theta}). \end{aligned}$$

In case 1 this expression can be rewritten as

$$L(\boldsymbol{\theta}) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \Psi_i^\top \boldsymbol{\theta})^2.$$

In case 2 the expression is similar:

$$L(\boldsymbol{\theta}) = -\sum_{i=1}^n \left\{ \frac{1}{2} \log(2\pi\sigma_i^2) + \frac{(Y_i - \Psi_i^\top \boldsymbol{\theta})^2}{2\sigma_i^2} \right\}.$$

The *maximum likelihood estimate* (MLE) $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^*$ is defined by maximizing the log-likelihood $L(\boldsymbol{\theta})$:

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmax}} L(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} (\mathbf{Y} - \Psi^\top \boldsymbol{\theta})^\top \Sigma^{-1} (\mathbf{Y} - \Psi^\top \boldsymbol{\theta}). \quad (1.2)$$

We omit the other terms in the expression of $L(\boldsymbol{\theta})$ because they do not depend on $\boldsymbol{\theta}$. This estimate is the *least squares estimate* (LSE) because it minimizes the sum of squared distances between the observations Y_i and the linear responses $\Psi_i^\top \boldsymbol{\theta}$. Note that (1.2) is a quadratic optimization problem which has a closed form solution. Differentiating the right hand-side of (1.2) w.r.t. $\boldsymbol{\theta}$ yields the *normal equation*

$$\Psi \Sigma^{-1} \Psi^\top \tilde{\boldsymbol{\theta}} = \Psi \Sigma^{-1} \mathbf{Y}.$$

If the $p \times p$ -matrix $\Psi \Sigma^{-1} \Psi^\top$ is non-degenerate then the normal equation has the unique solution

$$\tilde{\boldsymbol{\theta}} = (\Psi \Sigma^{-1} \Psi^\top)^{-1} \Psi \Sigma^{-1} \mathbf{Y} = \mathcal{S} \mathbf{Y}, \quad (1.3)$$

where

$$\mathcal{S} = (\Psi \Sigma^{-1} \Psi^\top)^{-1} \Psi \Sigma^{-1}$$

is a $p \times n$ matrix. We denote by $\tilde{\theta}_j$ the entries of the vector $\tilde{\boldsymbol{\theta}}$, $j = 1, \dots, p$.

If the matrix $\Psi \Sigma^{-1} \Psi^\top$ is degenerate, then the normal equation has infinitely many solutions. However, one can still apply the formula (1.3) where $(\Psi \Sigma^{-1} \Psi^\top)^{-1}$ is a pseudo-inverse of the matrix $\Psi \Sigma^{-1} \Psi^\top$.

The ML-approach leads to the *parameter estimate* $\tilde{\boldsymbol{\theta}}$. Note that due to the model (1.1), the product $\tilde{\mathbf{f}} = \Psi^\top \tilde{\boldsymbol{\theta}}$ is an estimate of the mean $\mathbf{f}^* \stackrel{\text{def}}{=} \mathbb{E}\mathbf{Y}$ of the vector of observations \mathbf{Y} :

$$\tilde{\mathbf{f}} = \Psi^\top \tilde{\boldsymbol{\theta}} = \Psi^\top (\Psi \Sigma^{-1} \Psi^\top)^{-1} \Psi \Sigma^{-1} \mathbf{Y} = \Pi \mathbf{Y},$$

where

$$\Pi = \Psi^\top (\Psi \Sigma^{-1} \Psi^\top)^{-1} \Psi \Sigma^{-1}$$

is an $n \times n$ matrix (linear operator) in \mathbb{R}^n . The vector $\tilde{\mathbf{f}}$ is called a *prediction* or *response* regression estimate.

Below we study the properties of the estimates $\tilde{\boldsymbol{\theta}}$ and $\tilde{\mathbf{f}}$. In this study we try to address both types of possible model misspecification: due to a wrong assumption about the error distribution and due to a possibly wrong linear parametric structure. Namely we consider the model

$$Y_i = f_i + \varepsilon_i, \quad \varepsilon \sim \mathcal{N}(0, \Sigma_0). \quad (1.4)$$

The response values f_i are usually treated as the value of the regression function $f(\cdot)$ at the design points X_i . The parametric model (1.1) can be viewed as an approximation of (1.4) while Σ is an approximation of the true covariance matrix Σ_0 . If \mathbf{f}^* is indeed equal to $\Psi^\top \boldsymbol{\theta}^*$ and $\Sigma = \Sigma_0$, then $\tilde{\boldsymbol{\theta}}$ and $\tilde{\mathbf{f}}$ are MLEs, otherwise quasi MLEs. In our study we mostly restrict ourselves to the case 1 assumption about the noise ε : $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. The general case can be reduced to this one by a simple data transformation, namely, by multiplying the equation (1.4) $\mathbf{Y} = \mathbf{f}^* + \varepsilon$ with the matrix $\Sigma^{-1/2}$, see Section 1.4.1 for more detail.

1.1.1 Estimation under homogeneous noise assumption

If a homogeneous noise is assumed, that is, if $\Sigma = \sigma^2 I_n$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, then the formulae for the MLEs $\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{f}}$ slightly simplify. In particular, the variance σ^2 cancels and the resulting estimate is the *ordinary least squares* (oLSE):

$$\tilde{\boldsymbol{\theta}} = (\Psi \Psi^\top)^{-1} \Psi \mathbf{Y} = \mathcal{S} \mathbf{Y}$$

with $\mathcal{S} = (\Psi\Psi^\top)^{-1}\Psi$. Also

$$\tilde{\mathbf{f}} = \Psi^\top(\Psi\Psi^\top)^{-1}\Psi\mathbf{Y} = \Pi\mathbf{Y}$$

with $\Pi = \Psi^\top(\Psi\Psi^\top)^{-1}\Psi$.

Exercise 1.1.1. Derive the formulae for $\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{f}}$ directly from the log-likelihood $L(\boldsymbol{\theta})$ for homogeneous noise.

If the assumption $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$ about the errors is not precisely fulfilled, then the oLSE can be viewed as a quasi MLE.

1.1.2 Linear basis transformation

Denote by $\psi_1^\top, \dots, \psi_p^\top$ the rows of the matrix Ψ . Then the ψ_i 's are vectors in \mathbb{R}^n and we call them *the basis vectors*. In the linear regression case the ψ_i 's are obtained as the values of the basis functions at the design points. Our linear parametric assumption simply means that the underlying vector \mathbf{f}^* can be represented as a linear combination of the vectors ψ_1, \dots, ψ_p :

$$\mathbf{f}^* = \theta_1^*\psi_1 + \dots + \theta_p^*\psi_p.$$

In other words, \mathbf{f}^* belongs to the linear subspace in \mathbb{R}^n spanned by the vectors ψ_1, \dots, ψ_p . It is clear that this assumption still holds if we select another basis in this subspace.

Let U be any linear orthogonal transformation in \mathbb{R}^p with $UU^\top = I_p$. Then the linear relation $\mathbf{f}^* = \Psi^\top\boldsymbol{\theta}^*$ can be rewritten as

$$\mathbf{f}^* = \Psi^\top UU^\top\boldsymbol{\theta}^* = \check{\Psi}^\top\mathbf{u}^*$$

with $\check{\Psi} = U^\top\Psi$ and $\mathbf{u}^* = U^\top\boldsymbol{\theta}^*$. Here the columns of $\check{\Psi}$ mean the new basis vectors $\check{\psi}_m$ in the same subspace while \mathbf{u}^* is the vector of coefficients describing the decomposition of the vector \mathbf{f}^* w.r.t. this new basis:

$$\mathbf{f}^* = u_1^*\check{\psi}_1 + \dots + u_p^*\check{\psi}_p.$$

The natural question is how the expression for the MLEs $\tilde{\boldsymbol{\theta}}$ and $\tilde{\mathbf{f}}$ change with the change of the basis. The answer is straightforward. For notational simplicity, we only consider the case with $\Sigma = \sigma^2 I_n$. The model can be rewritten as

$$\mathbf{Y} = \check{\Psi}^\top\mathbf{u}^* + \boldsymbol{\varepsilon}$$

yielding the solutions

$$\tilde{\mathbf{u}} = (\breve{\Psi} \breve{\Psi}^\top)^{-1} \breve{\Psi} \mathbf{Y} = \check{\mathcal{S}} \mathbf{Y}, \quad \tilde{\mathbf{f}} = \breve{\Psi}^\top (\breve{\Psi} \breve{\Psi}^\top)^{-1} \breve{\Psi} \mathbf{Y} = \check{\Pi} \mathbf{Y},$$

where $\breve{\Psi} = U^\top \Psi$ implies

$$\begin{aligned}\check{\mathcal{S}} &= (\breve{\Psi} \breve{\Psi}^\top)^{-1} \breve{\Psi} = U^\top \mathcal{S}, \\ \check{\Pi} &= \breve{\Psi}^\top (\breve{\Psi} \breve{\Psi}^\top)^{-1} \breve{\Psi} = \Pi.\end{aligned}$$

This yields

$$\tilde{\mathbf{u}} = U^\top \tilde{\boldsymbol{\theta}}$$

and moreover, the estimate $\tilde{\mathbf{f}}$ is not changed for any linear transformation of the basis. The first statement can be expected in view of $\boldsymbol{\theta}^* = U \mathbf{u}^*$, while the second one will be explained in the next section: Π is the linear projector on the subspace spanned by the basis vectors and this projector is invariant w.r.t. basis transformations.

Exercise 1.1.2. Consider univariate polynomial regression of degree $p - 1$. This means that f is a polynomial function of degree $p - 1$ observed at the points X_i with errors ε_i that are assumed to be i.i.d. normal. The function f can be represented as

$$f(x) = \theta_1^* + \theta_2^* x + \dots + \theta_p^* x^{p-1}$$

using the basis functions $\psi_j(x) = x^{j-1}$ for $j = 0, \dots, p - 1$. At the same time, for any point x_0 , this function can also be written as

$$f(x) = u_1^* + u_2^*(x - x_0) + \dots + u_p^*(x - x_0)^{p-1}$$

using the basis functions $\breve{\psi}_j = (x - x_0)^{j-1}$.

- Write the matrices Ψ and $\Psi \Psi^\top$ and similarly $\breve{\Psi}$ and $\breve{\Psi} \breve{\Psi}^\top$.
- Describe the linear transformation A such that $\mathbf{u} = A \boldsymbol{\theta}$ for $p = 1$.
- Describe the transformation A such that $\mathbf{u} = A \boldsymbol{\theta}$ for $p > 1$.

Hint: use the formula

$$u_j^* = \frac{1}{(j-1)!} f^{(j-1)}(x_0), \quad j = 1, \dots, p$$

to identify the coefficient u_j^* via $\theta_1^*, \dots, \theta_p^*$.

1.1.3 Orthogonal and orthonormal design

Orthogonality of the design matrix Ψ means that the basis vectors ψ_1, \dots, ψ_p are orthonormal in the sense

$$\psi_j^\top \psi_{j'} = \sum_{i=1}^n \psi_{m,i} \psi_{m',i} = \begin{cases} 0 & \text{if } j \neq j', \\ \lambda_j & \text{if } j = j', \end{cases}$$

for some positive values $\lambda_1, \dots, \lambda_p$. Equivalently one can write

$$\Psi \Psi^\top = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_p).$$

This feature of the design is very useful and it essentially simplifies the computation and analysis of the properties of $\tilde{\theta}$. Indeed, $\Psi \Psi^\top = \Lambda$ implies

$$\tilde{\theta} = \Lambda^{-1} \Psi Y, \quad \tilde{f} = \Psi^\top \tilde{\theta} = \Psi^\top \Lambda^{-1} \Psi Y$$

with $\Lambda^{-1} = \text{diag}(\lambda_1^{-1}, \dots, \lambda_p^{-1})$. In particular, the first relation means

$$\tilde{\theta}_j = \lambda_j^{-1} \sum_{i=1}^n Y_i \psi_{j,i},$$

that is, $\tilde{\theta}_j$ is the scalar product of the data and the basis vector ψ_j for $j = 1, \dots, p$.

The estimate of the response f reads as

$$\tilde{f} = \tilde{\theta}_1 \psi_1 + \dots + \tilde{\theta}_p \psi_p.$$

Theorem 1.1.1. Consider the model $Y = \Psi^\top \theta + \varepsilon$ with homogeneous errors ε : $E \varepsilon \varepsilon^\top = \sigma^2 I_n$. If the design Ψ is orthogonal, that is, if $\Psi \Psi^\top = \Lambda$ for a diagonal matrix Λ , then the estimated coefficients $\tilde{\theta}_j$ are uncorrelated: $\text{Var}(\tilde{\theta}) = \sigma^2 \Lambda^{-1}$. Moreover, if $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, then $\tilde{\theta} \sim \mathcal{N}(\theta^*, \sigma^2 \Lambda^{-1})$.

An important message of this result is that the orthogonal design allows for splitting the original multivariate problem into a collection of independent univariate problems: each coefficient θ_j^* is estimated by $\tilde{\theta}_j$ independently on the remaining coefficients.

The calculus can be further simplified in the case of an orthogonal design with $\Psi \Psi^\top = I_p$. Then one speaks about an *orthonormal design*. This also implies that every basis function (vector) ψ_j is standardized: $\|\psi_j\|^2 = \sum_{i=1}^n \psi_{j,i}^2 = 1$. In the case of an orthonormal design, the estimate $\tilde{\theta}$ is particularly simple: $\tilde{\theta} = \Psi Y$. Correspondingly, the target of estimation θ^* satisfies $\theta^* = \Psi f^*$. In other words, the target is the collection (θ_j^*) of the Fourier coefficients of the underlying function (vector) f^* w.r.t. the basis Ψ while the estimate $\tilde{\theta}$ is the collection of empirical Fourier coefficients $\tilde{\theta}_j$:

$$\theta_j^* = \sum_{i=1}^n f_i \psi_{j,i}, \quad \tilde{\theta}_j = \sum_{i=1}^n Y_i \psi_{j,i}$$

An important feature of the orthonormal design is that it preserves the noise homogeneity:

$$\text{Var}(\tilde{\boldsymbol{\theta}}) = \sigma^2 I_p.$$

1.1.4 Spectral representation

Consider a linear model

$$\mathbf{Y} = \Psi^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon} \quad (1.5)$$

with homogeneous errors $\boldsymbol{\varepsilon}$: $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$. The rows of the matrix Ψ can be viewed as basis vectors in \mathbb{R}^n and the product $\Psi^\top \boldsymbol{\theta}$ is a linear combinations of these vectors with the coefficients $(\theta_1, \dots, \theta_p)$. Effectively linear least squares estimation does a kind of projection of the data onto the subspace generated by the basis functions. This projection is of course invariant w.r.t. a basis transformation within this linear subspace. This fact can be used to reduce the model to the case of an orthogonal design considered in the previous section. Namely, one can always find a linear orthogonal transformation $U : \mathbb{R}^p \rightarrow \mathbb{R}^p$ ensuring the orthogonality of the transformed basis. This means that the rows of the matrix $\check{\Psi} = U\Psi$ are orthogonal and the matrix $\check{\Psi}\check{\Psi}^\top$ is diagonal:

$$\check{\Psi}\check{\Psi}^\top = U\Psi\Psi^\top U^\top = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_p).$$

The original model reads after this transformation in the form

$$\mathbf{Y} = \check{\Psi}^\top \mathbf{u} + \boldsymbol{\varepsilon}, \quad \check{\Psi}\check{\Psi}^\top = \Lambda,$$

where $\mathbf{u} = U\boldsymbol{\theta} \in \mathbb{R}^p$. Within this model, the transformed parameter \mathbf{u} can be estimated using the empirical Fourier coefficients $Z_j = \check{\Psi}_j^\top \mathbf{Y}$, where $\check{\Psi}_j$ is the j th row of $\check{\Psi}$, $j = 1, \dots, p$. The original parameter vector $\boldsymbol{\theta}$ can be recovered via the equation $\boldsymbol{\theta} = U^\top \mathbf{u}$. This set of equations can be written in the form

$$\mathbf{Z} = \Lambda \mathbf{u} + \Lambda^{1/2} \boldsymbol{\xi} \quad (1.6)$$

where $\mathbf{Z} = \check{\Psi}\mathbf{Y} = U\Psi\mathbf{Y}$ is a vector in \mathbb{R}^p and $\boldsymbol{\xi} = \Lambda^{-1/2} \check{\Psi} \boldsymbol{\varepsilon} = \Lambda^{-1/2} U \Psi \boldsymbol{\varepsilon} \in \mathbb{R}^p$. The equation (1.6) is called the *spectral representation* of the linear model (1.5). The reason is that the basic transformation U can be built by a singular value decomposition of Ψ . This representation is widely used in context of linear inverse problems; see Section 3.6.

Theorem 1.1.2. Consider the model (1.5) with homogeneous errors ε , that is, $E\varepsilon\varepsilon^\top = \sigma^2 I_n$. Then there exists an orthogonal transform $U : \mathbb{R}^p \rightarrow \mathbb{R}^p$ leading to the spectral representation (1.6) with homogeneous uncorrelated errors ξ : $E\xi\xi^\top = \sigma^2 I_p$. If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, then the vector ξ is normal as well: $\xi \sim \mathcal{N}(0, \sigma^2 I_p)$.

Exercise 1.1.3. Prove the result of Theorem 1.1.2.

Hint: select any U ensuring $U^\top \Psi \Psi^\top U = \Lambda$. Then

$$E\xi\xi^\top = \Lambda^{-1/2} U \Psi E\varepsilon\varepsilon^\top \Psi^\top U^\top \Lambda^{-1/2} = \sigma^2 \Lambda^{-1/2} U^\top \Psi \Psi^\top U \Lambda^{-1/2} = \sigma^2 I_p.$$

A special case of the spectral representation corresponds to the orthonormal design with $\Psi \Psi^\top = I_p$. In this situation, the spectral model reads as $Z = u + \xi$, that is, we simply observe the target u corrupted with a homogeneous noise ξ . Such an equation is often called the *sequence space model* and it is intensively used in the literature for the theoretical study; cf. Section 3 below.

1.2 Properties of the response estimate \tilde{f}

This section discusses some properties of the estimate $\tilde{f} = \Psi^\top \tilde{\theta} = \Pi Y$ of the response vector f^* . It is worth noting that the first and essential part of the analysis does not rely on the underlying model distribution, only on our parametric assumptions that $f = \Psi^\top \theta^*$ and $\text{Cov}(\varepsilon) = \Sigma = \sigma^2 I_n$. The real model only appears when studying the risk of estimation. We will comment on the cases of misspecified f and Σ .

When $\Sigma = \sigma^2 I_n$, the operator Π in the representation $\tilde{f} = \Pi Y$ of the estimate \tilde{f} reads as

$$\Pi = \Psi^\top (\Psi \Psi^\top)^{-1} \Psi. \quad (1.7)$$

First we make use of the linear structure of the model (1.1) and of the estimate \tilde{f} to derive a number of its simple but important properties.

1.2.1 Decomposition into a deterministic and a stochastic component

The model equation $Y = f^* + \varepsilon$ yields

$$\tilde{f} = \Pi Y = \Pi(f^* + \varepsilon) = \Pi f^* + \Pi \varepsilon. \quad (1.8)$$

The first element of this sum, Πf^* is purely deterministic, but it depends on the unknown response vector f^* . Moreover, it will be shown in the next lemma that $\Pi f^* = f^*$ if the parametric assumption holds and the vector f^* indeed can be represented as $\Psi^\top \theta^*$.

The second element is stochastic as a linear transformation of the stochastic vector $\boldsymbol{\varepsilon}$ but is independent of the model response \mathbf{f}^* . The properties of the estimate $\tilde{\mathbf{f}}$ heavily rely on the properties of the linear operator Π from (1.7) which we collect in the next section.

1.2.2 Properties of the operator Π

Let ψ_1, \dots, ψ_p be the columns of the matrix Ψ^\top . These are the vectors in \mathbb{R}^n also called *the basis vectors*.

Lemma 1.2.1. *Let the matrix $\Psi\Psi^\top$ be non-degenerate. Then the operator Π fulfills the following conditions:*

- (i) Π is symmetric (self-adjoint), that is, $\Pi^\top = \Pi$.
- (ii) Π is a projector in \mathbb{R}^n , i.e. $\Pi^\top\Pi = \Pi^2 = \Pi$ and $\Pi(\mathbf{1}_n - \Pi) = 0$, where $\mathbf{1}_n$ means the unity operator in \mathbb{R}^n .
- (iii) For an arbitrary vector v from \mathbb{R}^n , it holds $\|v\|^2 = \|\Pi v\|^2 + \|v - \Pi v\|^2$.
- (iv) The trace of Π is equal to the dimension of its image, $\text{tr } \Pi = p$.
- (v) Π projects the linear space \mathbb{R}^n on the linear subspace $L_p = \langle \psi_1, \dots, \psi_p \rangle$, which is spanned by the basis vectors ψ_1, \dots, ψ_p , that is,

$$\|\mathbf{f}^* - \Pi\mathbf{f}^*\| = \inf_{\mathbf{g} \in L_p} \|\mathbf{f}^* - \mathbf{g}\|.$$

- (vi) The matrix Π can be represented in the form

$$\Pi = U^\top \Lambda_p U$$

where U is an orthonormal matrix and Λ_p is a diagonal matrix with the first p diagonal elements equal to 1 and the others equal to zero:

$$\Lambda_p = \text{diag}\{\underbrace{1, \dots, 1}_p, \underbrace{0, \dots, 0}_{n-p}\}.$$

Proof. It holds

$$\{\Psi^\top (\Psi\Psi^\top)^{-1}\Psi\}^\top = \Psi^\top (\Psi\Psi^\top)^{-1}\Psi$$

and

$$\Pi^2 = \Psi^\top (\Psi\Psi^\top)^{-1}\Psi\Psi^\top (\Psi\Psi^\top)^{-1}\Psi = \Psi^\top (\Psi\Psi^\top)^{-1}\Psi = \Pi,$$

which proves the first two statements of the lemma. The third one follows directly from the first two. Next,

$$\operatorname{tr} \Pi = \operatorname{tr} \Psi^\top (\Psi \Psi^\top)^{-1} \Psi = \operatorname{tr} \Psi \Psi^\top (\Psi \Psi^\top)^{-1} = \operatorname{tr} I_p = p.$$

The second property means that Π is a projector in \mathbb{R}^n and the fourth one means that the dimension of its image space is equal to p . The basis vectors ψ_1, \dots, ψ_p are the rows of the matrix Ψ . It is clear that

$$\Pi \Psi^\top = \Psi^\top (\Psi \Psi^\top)^{-1} \Psi \Psi^\top = \Psi^\top.$$

Therefore, the vectors ψ_j are invariants of the operator Π and in particular, all these vectors belong to the image space of this operator. If now \mathbf{g} is a vector in L_p , then it can be represented as $\mathbf{g} = c_1 \psi_1 + \dots + c_p \psi_p$ and therefore, $\Pi \mathbf{g} = \mathbf{g}$ and $\Pi L_p = L_p$. Finally, the non-singularity of the matrix $\Psi \Psi^\top$ means that the vectors ψ_1, \dots, ψ_p forming the rows of Ψ are linearly independent. Therefore, the space L_p spanned by the vectors ψ_1, \dots, ψ_p is of dimension p , and hence it coincides with the image space of the operation Π .

The last property is the usual diagonal decomposition of a projector.

Exercise 1.2.1. Consider the case of an orthogonal design with $\Psi \Psi^\top = I_p$. Specify the projector Π of Lemma 1.2.1 for this situation, particularly its decomposition from (vi).

1.2.3 Quadratic loss and risk of the response estimation

In this section we study the quadratic risk of estimating the response \mathbf{f}^* . The reason for studying the quadratic risk of estimating the response \mathbf{f}^* will be made clear when we discuss the properties of the fitted likelihood in the next section.

The loss $\varphi(\tilde{\mathbf{f}}, \mathbf{f}^*)$ of the estimate $\tilde{\mathbf{f}}$ can be naturally defined as the squared norm of the difference $\tilde{\mathbf{f}} - \mathbf{f}^*$:

$$\varphi(\tilde{\mathbf{f}}, \mathbf{f}^*) = \|\tilde{\mathbf{f}} - \mathbf{f}^*\|^2 = \sum_{i=1}^n |f_i - \tilde{f}_i|^2.$$

Correspondingly, the quadratic risk of the estimate $\tilde{\mathbf{f}}$ is the mean of this loss

$$\mathcal{R}(\tilde{\mathbf{f}}) = \mathbb{E} \varphi(\tilde{\mathbf{f}}, \mathbf{f}^*) = \mathbb{E} [(\tilde{\mathbf{f}} - \mathbf{f}^*)^\top (\tilde{\mathbf{f}} - \mathbf{f}^*)]. \quad (1.9)$$

The next result describes the loss and risk decomposition for two cases: when the parametric assumption $\mathbf{f}^* = \Psi^\top \boldsymbol{\theta}^*$ is correct and in the general case.

Theorem 1.2.1. Suppose that the errors ε_i from (1.1) are independent with $\mathbb{E} \varepsilon_i = 0$ and $\mathbb{E} \varepsilon_i^2 = \sigma^2$, i.e. $\Sigma = \sigma^2 I_n$. Then the loss $\varphi(\tilde{\mathbf{f}}, \mathbf{f}^*) = \|\Pi \mathbf{Y} - \mathbf{f}^*\|^2$ and the risk $\mathcal{R}(\tilde{\mathbf{f}})$ of the LSE $\tilde{\mathbf{f}}$ fulfill

$$\wp(\tilde{\mathbf{f}}, \mathbf{f}^*) = \|\mathbf{f}^* - \Pi \mathbf{f}^*\|^2 + \|\Pi \varepsilon\|^2,$$

$$\mathcal{R}(\tilde{\mathbf{f}}) = \|\mathbf{f}^* - \Pi \mathbf{f}^*\|^2 + p\sigma^2.$$

Moreover, if $\mathbf{f}^* = \Psi^\top \boldsymbol{\theta}^*$, then

$$\wp(\tilde{\mathbf{f}}, \mathbf{f}^*) = \|\Pi \varepsilon\|^2,$$

$$\mathcal{R}(\tilde{\mathbf{f}}) = p\sigma^2.$$

Proof. We apply (1.9) and the decomposition (1.8) of the estimate $\tilde{\mathbf{f}}$. It follows

$$\begin{aligned} \wp(\tilde{\mathbf{f}}, \mathbf{f}^*) &= \|\tilde{\mathbf{f}} - \mathbf{f}^*\|^2 = \|\mathbf{f}^* - \Pi \mathbf{f}^* - \Pi \varepsilon\|^2 \\ &= \|\mathbf{f}^* - \Pi \mathbf{f}^*\|^2 + 2(\mathbf{f}^* - \Pi \mathbf{f}^*)^\top \Pi \varepsilon + \|\Pi \varepsilon\|^2. \end{aligned}$$

This implies the decomposition for the loss of $\tilde{\mathbf{f}}$ by Lemma 1.2.1, (ii). Next we compute the mean of $\|\Pi \varepsilon\|^2$ applying again Lemma 1.2.1. Indeed

$$\begin{aligned} \mathbb{E}\|\Pi \varepsilon\|^2 &= \mathbb{E}(\Pi \varepsilon)^\top \Pi \varepsilon = \mathbb{E} \text{tr}\{\Pi \varepsilon (\Pi \varepsilon)^\top\} = \mathbb{E} \text{tr}(\Pi \varepsilon \varepsilon^\top \Pi^\top) \\ &= \text{tr}\{\Pi \mathbb{E}(\varepsilon \varepsilon^\top) \Pi\} = \sigma^2 \text{tr}(\Pi^2) = p\sigma^2. \end{aligned}$$

Now consider the case when $\mathbf{f}^* = \Psi^\top \boldsymbol{\theta}^*$. By Lemma 1.2.1 $\mathbf{f}^* = \Pi \mathbf{f}^*$ and the last two statements of the theorem clearly follow.

1.2.4 Misspecified “colored noise”

Here we briefly comment on the case when ε is not a white noise. So, our assumption about the errors ε_i is that they are uncorrelated and homogeneous, that is, $\Sigma = \sigma^2 I_n$ while the true covariance matrix is given by Σ_0 . Many properties of the estimate $\tilde{\mathbf{f}} = \Pi \mathbf{Y}$ which are simply based on the linearity of the model (1.1) and of the estimate $\tilde{\mathbf{f}}$ itself continue to apply. In particular, the loss $\wp(\tilde{\mathbf{f}}, \mathbf{f}^*) = \|\tilde{\mathbf{f}} - \mathbf{f}^*\|^2$ can again be decomposed as

$$\|\tilde{\mathbf{f}} - \mathbf{f}^*\|^2 = \|\mathbf{f}^* - \Pi \mathbf{f}^*\|^2 + \|\Pi \varepsilon\|^2.$$

Theorem 1.2.2. Suppose that $\mathbb{E} \varepsilon = 0$ and $\text{Var}(\varepsilon) = \Sigma_0$. Then the loss $\wp(\tilde{\mathbf{f}}, \mathbf{f})$ and the risk $\mathcal{R}(\tilde{\mathbf{f}})$ of the LSE $\tilde{\mathbf{f}}$ fulfill

$$\wp(\tilde{\mathbf{f}}, \mathbf{f}^*) = \|\mathbf{f}^* - \Pi \mathbf{f}^*\|^2 + \|\Pi \varepsilon\|^2,$$

$$\mathcal{R}(\tilde{\mathbf{f}}) = \|\mathbf{f}^* - \Pi \mathbf{f}^*\|^2 + \text{tr}(\Pi \Sigma_0 \Pi).$$

Moreover, if $\mathbf{f}^* = \Psi^\top \boldsymbol{\theta}^*$, then

$$\wp(\tilde{\mathbf{f}}, \mathbf{f}^*) = \|\Pi\boldsymbol{\varepsilon}\|^2,$$

$$\mathcal{R}(\tilde{\mathbf{f}}) = \text{tr}(\Pi\Sigma_0\Pi).$$

Proof. The decomposition of the loss from Theorem 1.2.1 only relies on the geometric properties of the projector Π and does not use the covariance structure of the noise. Hence, it only remains to check the expectation of $\|\Pi\boldsymbol{\varepsilon}\|^2$. Observe that

$$\mathbb{E}\|\Pi\boldsymbol{\varepsilon}\|^2 = \mathbb{E}\text{tr}[\Pi\boldsymbol{\varepsilon}(\Pi\boldsymbol{\varepsilon})^\top] = \text{tr}[\Pi\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)\Pi] = \text{tr}(\Pi\Sigma_0\Pi)$$

as required.

1.3 Properties of the MLE $\tilde{\boldsymbol{\theta}}$

In this section we focus on the properties of the quasi MLE $\tilde{\boldsymbol{\theta}}$ built for the idealized linear Gaussian model $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. As in the previous section, we do not assume the parametric structure of the underlying model and consider a more general model $\mathbf{Y} = \mathbf{f}^* + \boldsymbol{\varepsilon}$ with an unknown vector \mathbf{f}^* and errors $\boldsymbol{\varepsilon}$ with zero mean and covariance matrix Σ_0 . Due to (1.3), it holds $\tilde{\boldsymbol{\theta}} = \mathcal{S}\mathbf{Y}$ with $\mathcal{S} = (\Psi\Psi^\top)^{-1}\Psi$. An important feature of this estimate is its linear dependence on the data. The linear model equation $\mathbf{Y} = \mathbf{f}^* + \boldsymbol{\varepsilon}$ and linear structure of the estimate $\tilde{\boldsymbol{\theta}} = \mathcal{S}\mathbf{Y}$ allow us for decomposing the vector $\tilde{\boldsymbol{\theta}}$ into a deterministic and stochastic terms:

$$\tilde{\boldsymbol{\theta}} = \mathcal{S}\mathbf{Y} = \mathcal{S}(\mathbf{f}^* + \boldsymbol{\varepsilon}) = \mathcal{S}\mathbf{f}^* + \mathcal{S}\boldsymbol{\varepsilon}. \quad (1.10)$$

The first term $\mathcal{S}\mathbf{f}^*$ is deterministic but depends on the unknown vector \mathbf{f}^* while the second term $\mathcal{S}\boldsymbol{\varepsilon}$ is stochastic but it does not involve the model response \mathbf{f}^* . Below we study the properties of each component separately.

1.3.1 Properties of the stochastic component

The next result describes the distributional properties of the stochastic component $\boldsymbol{\delta} = \mathcal{S}\boldsymbol{\varepsilon}$ for $\mathcal{S} = (\Psi\Psi^\top)^{-1}\Psi$ and thus, of the estimate $\tilde{\boldsymbol{\theta}}$.

Theorem 1.3.1. *Assume $\mathbf{Y} = \mathbf{f}^* + \boldsymbol{\varepsilon}$ with $\mathbb{E}\boldsymbol{\varepsilon} = 0$ and $\text{Var}(\boldsymbol{\varepsilon}) = \Sigma_0$. The stochastic component $\boldsymbol{\delta} = \mathcal{S}\boldsymbol{\varepsilon}$ in (1.10) fulfills*

$$\mathbb{E}\boldsymbol{\delta} = 0, \quad W^2 \stackrel{\text{def}}{=} \text{Var}(\boldsymbol{\delta}) = \mathcal{S}\Sigma_0\mathcal{S}^\top, \quad \mathbb{E}\|\boldsymbol{\delta}\|^2 = \text{tr}W^2 = \text{tr}(\mathcal{S}\Sigma_0\mathcal{S}^\top).$$

Moreover, if $\Sigma = \Sigma_0 = \sigma^2 I_n$, then

$$W^2 = \sigma^2(\Psi\Psi^\top)^{-1}, \quad \mathbb{E}\|\boldsymbol{\delta}\|^2 = \text{tr}(W^2) = \sigma^2 \text{tr}[(\Psi\Psi^\top)^{-1}]. \quad (1.11)$$

Similarly for the estimate $\tilde{\boldsymbol{\theta}}$ it holds

$$\mathbb{E}\tilde{\boldsymbol{\theta}} = \mathcal{S}\mathbf{f}^*, \quad \text{Var}(\tilde{\boldsymbol{\theta}}) = W^2.$$

If the errors ε are Gaussian, then the both $\boldsymbol{\delta}$ and $\tilde{\boldsymbol{\theta}}$ are Gaussian as well:

$$\boldsymbol{\delta} \sim \mathcal{N}(0, W^2) \quad \tilde{\boldsymbol{\theta}} \sim \mathcal{N}(\mathcal{S}\mathbf{f}^*, W^2).$$

Proof. For the variance W^2 of $\boldsymbol{\delta}$ holds

$$\text{Var}(\boldsymbol{\delta}) = \mathbb{E}\boldsymbol{\delta}\boldsymbol{\delta}^\top = \mathbb{E}\mathcal{S}\varepsilon\varepsilon^\top\mathcal{S}^\top = \mathcal{S}\Sigma_0\mathcal{S}^\top.$$

Next we use that $\mathbb{E}\|\boldsymbol{\delta}\|^2 = \mathbb{E}\boldsymbol{\delta}^\top\boldsymbol{\delta} = \mathbb{E}\text{tr}(\boldsymbol{\delta}\boldsymbol{\delta}^\top) = \text{tr} W^2$. If $\Sigma = \Sigma_0 = \sigma^2 I_n$, then (1.11) follows by simple algebra.

If ε is a Gaussian vector, then $\boldsymbol{\delta}$ as its linear transformation is Gaussian as well. The properties of $\tilde{\boldsymbol{\theta}}$ follow directly from the decomposition (1.10).

With $\Sigma_0 \neq \sigma^2 I_n$, the variance W^2 can be represented as

$$W^2 = (\Psi\Psi^\top)^{-1}\Psi\Sigma_0\Psi^\top(\Psi\Psi^\top)^{-1}.$$

Exercise 1.3.1. Let $\boldsymbol{\delta}$ be the stochastic component of $\tilde{\boldsymbol{\theta}}$ built for the misspecified linear model $\mathbf{Y} = \Psi^\top\boldsymbol{\theta}^* + \varepsilon$ with $\text{Var}(\varepsilon) = \Sigma$. Let also the true noise variance is Σ_0 . Then $\text{Var}(\tilde{\boldsymbol{\theta}}) = W^2$ with

$$W^2 = (\Psi\Sigma^{-1}\Psi^\top)^{-1}\Psi\Sigma^{-1}\Sigma_0\Sigma^{-1}\Psi^\top(\Psi\Sigma^{-1}\Psi^\top)^{-1}.$$

The main finding in the presented study is that the stochastic part $\boldsymbol{\delta} = \mathcal{S}\varepsilon$ of the estimate $\tilde{\boldsymbol{\theta}}$ is completely independent of the structure of the vector \mathbf{f}^* . In other words, the behavior of the stochastic component $\boldsymbol{\delta}$ does not change even if the linear parametric assumption is misspecified.

1.3.2 Properties of the deterministic component

Now we study the deterministic term starting with the parametric situation $\mathbf{f}^* = \Psi^\top\boldsymbol{\theta}^*$. Here we only specify the results for the case 1 with $\Sigma = \sigma^2 I_n$.

Theorem 1.3.2. Let $\mathbf{f}^* = \Psi^\top\boldsymbol{\theta}^*$. Then $\tilde{\boldsymbol{\theta}} = \mathcal{S}\mathbf{Y}$ with $\mathcal{S} = (\Psi\Psi^\top)^{-1}\Psi$ is unbiased, that is, $\mathbb{E}\tilde{\boldsymbol{\theta}} = \mathcal{S}\mathbf{f}^* = \boldsymbol{\theta}^*$.

Proof. For the proof, just observe that $\mathcal{S}\mathbf{f}^* = (\Psi\Psi^\top)^{-1}\Psi\Psi^\top\boldsymbol{\theta}^* = \boldsymbol{\theta}^*$.

Now we briefly discuss what happens when the linear parametric assumption is not fulfilled, that is, \mathbf{f}^* cannot be represented as $\Psi^\top \boldsymbol{\theta}^*$. In this case it is not yet clear what $\tilde{\boldsymbol{\theta}}$ really estimates. The answer is given in the context of the general theory of minimum contrast estimation. Namely, define $\boldsymbol{\theta}^*$ as the point which maximizes the expectation of the (quasi) log-likelihood $L(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E} L(\boldsymbol{\theta}). \quad (1.12)$$

Theorem 1.3.3. *The solution $\boldsymbol{\theta}^*$ of the optimization problem (1.12) is given by*

$$\boldsymbol{\theta}^* = \mathcal{S} \mathbf{f}^* = (\Psi \Psi^\top)^{-1} \Psi \mathbf{f}^*.$$

Moreover,

$$\Psi^\top \boldsymbol{\theta}^* = \Pi \mathbf{f}^* = \Psi^\top (\Psi \Psi^\top)^{-1} \Psi \mathbf{f}^*.$$

In particular, if $\mathbf{f}^* = \Psi^\top \boldsymbol{\theta}^*$, then $\boldsymbol{\theta}^*$ follows (1.12).

Proof. The use of the model equation $\mathbf{Y} = \mathbf{f}^* + \boldsymbol{\varepsilon}$ and of the properties of the stochastic component $\boldsymbol{\delta}$ yield by simple algebra

$$\begin{aligned} \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E} L(\boldsymbol{\theta}) &= \operatorname{argmin}_{\boldsymbol{\theta}} \mathbb{E} (\mathbf{f}^* - \Psi^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon})^\top (\mathbf{f}^* - \Psi^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon}) \\ &= \operatorname{argmin}_{\boldsymbol{\theta}} \{ (\mathbf{f}^* - \Psi^\top \boldsymbol{\theta})^\top (\mathbf{f}^* - \Psi^\top \boldsymbol{\theta}) + \mathbb{E} (\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}) \} \\ &= \operatorname{argmin}_{\boldsymbol{\theta}} \{ (\mathbf{f}^* - \Psi^\top \boldsymbol{\theta})^\top (\mathbf{f}^* - \Psi^\top \boldsymbol{\theta}) \}. \end{aligned}$$

Differentiating w.r.t. $\boldsymbol{\theta}$ leads to the equation

$$\Psi(\mathbf{f}^* - \Psi^\top \boldsymbol{\theta}) = 0$$

and the solution $\boldsymbol{\theta}^* = (\Psi \Psi^\top)^{-1} \Psi \mathbf{f}^*$ which is exactly the expected value of $\tilde{\boldsymbol{\theta}}$ by Theorem 1.3.1.

Exercise 1.3.2. State the result of Theorems 1.3.2 and 1.3.3 for the MLE $\tilde{\boldsymbol{\theta}}$ built in the model $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with $\operatorname{Var}(\boldsymbol{\varepsilon}) = \Sigma$.

Hint: check that the statements continue to apply with $\mathcal{S} = (\Psi \Sigma^{-1} \Psi^\top)^{-1} \Psi \Sigma^{-1}$.

The last results and the decomposition (1.10) explain the behavior of the estimate $\tilde{\boldsymbol{\theta}}$ in a very general situation. The considered model is $\mathbf{Y} = \mathbf{f}^* + \boldsymbol{\varepsilon}$. We assume a linear parametric structure and independent homogeneous noise. The estimation procedure means in fact a kind of projection of the data \mathbf{Y} on a p -dimensional linear subspace in \mathbb{R}^n spanned by the given basis vectors ψ_1, \dots, ψ_p . This projection, as a linear operator,

can be decomposed into a projection of the deterministic vector f^* and a projection of the random noise ε . If the linear parametric assumption $f^* \in \langle \psi_1, \dots, \psi_p \rangle$ is correct, that is, $f^* = \theta_1^* \psi_1 + \dots + \theta_p^* \psi_p$, then this projection keeps f^* unchanged and only the random noise is reduced via this projection. If f^* cannot be exactly expanded using the basis ψ_1, \dots, ψ_p , then the procedure recovers the projection of f^* onto this subspace. The latter projection can be written as $\Psi^\top \theta^*$ and the vector θ^* can be viewed as the target of estimation.

1.3.3 Risk of estimation. R-efficiency

This section briefly discusses how the obtained properties of the estimate $\tilde{\theta}$ can be used to evaluate the risk of estimation. A particularly important question is the optimality of the MLE $\tilde{\theta}$. The main result of the section claims that $\tilde{\theta}$ is R-efficient if the model is correctly specified and is not if there is a misspecification.

We start with the case of a correct parametric specification $\mathbf{Y} = \Psi^\top \theta^* + \varepsilon$, that is, the linear parametric assumption $f^* = \Psi^\top \theta^*$ is exactly fulfilled and the noise ε is homogeneous: $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Later we extend the result to the case when the LPA $f^* = \Psi^\top \theta^*$ is not fulfilled and to the case when the noise is not homogeneous but still correctly specified. Finally we discuss the case when the noise structure is misspecified.

Under LPA $\mathbf{Y} = \Psi^\top \theta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, the estimate $\tilde{\theta}$ is also normal with mean θ^* and the variance $W^2 = \sigma^2 \mathcal{S} \mathcal{S}^\top = \sigma^2 (\Psi \Psi^\top)^{-1}$. Define a $p \times p$ symmetric matrix D by the equation

$$D^2 = \frac{1}{\sigma^2} \sum_{i=1}^n \Psi_i \Psi_i^\top = \frac{1}{\sigma^2} \Psi \Psi^\top.$$

Clearly $W^2 = D^{-2}$.

Now we show that $\tilde{\theta}$ is R-efficient. Actually this fact can be derived from the Cramér-Rao Theorem because the Gaussian model is a special case of an exponential family. However, we check this statement directly by computing the Cramér-Rao efficiency bound. Recall that the Fisher information matrix $\mathbb{F}(\theta)$ for the log-likelihood $L(\theta)$ is defined as the variance of $\nabla L(\theta)$ under \mathbb{P}_θ .

Theorem 1.3.4 (Gauss-Markov). *Let $\mathbf{Y} = \Psi^\top \theta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Then $\tilde{\theta}$ is R-efficient estimate of θ^* : $\mathbb{E}\tilde{\theta} = \theta^*$,*

$$\mathbb{E}[(\tilde{\theta} - \theta^*)(\tilde{\theta} - \theta^*)^\top] = \text{Var}(\tilde{\theta}) = D^{-2},$$

and for any unbiased linear estimate $\hat{\theta}$ satisfying $\mathbb{E}\hat{\theta} = \theta^*$, it holds

$$\text{Var}(\hat{\theta}) \geq \text{Var}(\tilde{\theta}) = D^{-2}.$$

Proof. Theorems 1.3.1 and 1.3.2 imply that $\tilde{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}^*, W^2)$ with $W^2 = \sigma^2(\Psi\Psi^\top)^{-1} = D^{-2}$. Next we show that for any $\boldsymbol{\theta}$

$$\text{Var}[\nabla L(\boldsymbol{\theta})] = D^2,$$

that is, the Fisher information does not depend on the model function \mathbf{f}^* . The log-likelihood $L(\boldsymbol{\theta})$ for the model $\mathbf{Y} \sim \mathcal{N}(\Psi^\top \boldsymbol{\theta}^*, \sigma^2 I_n)$ reads as

$$L(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}(\mathbf{Y} - \Psi^\top \boldsymbol{\theta})^\top (\mathbf{Y} - \Psi^\top \boldsymbol{\theta}) - \frac{n}{2} \log(2\pi\sigma^2).$$

This yields for its gradient $\nabla L(\boldsymbol{\theta})$:

$$\nabla L(\boldsymbol{\theta}) = \sigma^{-2}\Psi(\mathbf{Y} - \Psi^\top \boldsymbol{\theta})$$

and in view of $\text{Var}(\mathbf{Y}) = \Sigma = \sigma^2 I_n$, it holds

$$\text{Var}[\nabla L(\boldsymbol{\theta})] = \sigma^{-4}\Psi \text{Var}(\mathbf{Y})\Psi^\top = \sigma^{-2}\Psi\Psi^\top$$

as required.

The R-efficiency $\tilde{\boldsymbol{\theta}}$ follows from the Cramér-Rao efficiency bound because $\{\text{Var}(\tilde{\boldsymbol{\theta}})\}^{-1} = \text{Var}\{\nabla L(\boldsymbol{\theta})\}$. However, we present an independent proof of this fact. Actually we prove a sharper result that the variance of a linear unbiased estimate $\hat{\boldsymbol{\theta}}$ coincides with the variance of $\tilde{\boldsymbol{\theta}}$ only if $\hat{\boldsymbol{\theta}}$ coincides almost surely with $\tilde{\boldsymbol{\theta}}$, otherwise it is larger. The idea of the proof is quite simple. Consider the difference $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}$ and show that the condition $\mathbb{E}\hat{\boldsymbol{\theta}} = \mathbb{E}\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$ implies orthogonality $\mathbb{E}\{\tilde{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})^\top\} = 0$. This, in turns, implies $\text{Var}(\hat{\boldsymbol{\theta}}) = \text{Var}(\tilde{\boldsymbol{\theta}}) + \text{Var}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) \geq \text{Var}(\tilde{\boldsymbol{\theta}})$. So, it remains to check the orthogonality of $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}$. Let $\hat{\boldsymbol{\theta}} = A\mathbf{Y}$ for a $p \times n$ matrix A and $\mathbb{E}_{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}} \equiv \boldsymbol{\theta}$ and all $\boldsymbol{\theta}$. These two equalities and $\mathbb{E}\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^*$ imply that $A\Psi^\top \boldsymbol{\theta}^* \equiv \boldsymbol{\theta}^*$, i.e. $A\Psi^\top$ is the identity $p \times p$ matrix. The same is true for $\tilde{\boldsymbol{\theta}} = S\mathbf{Y}$ yielding $S\Psi^\top = I_p$. Next, in view of $\mathbb{E}\hat{\boldsymbol{\theta}} = \mathbb{E}\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$

$$\mathbb{E}\{(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})\tilde{\boldsymbol{\theta}}^\top\} = \mathbb{E}(A - S)\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top S^\top = \sigma^2(A - S)\Psi^\top(\Psi\Psi^\top)^{-1} = 0,$$

and the assertion follows.

Exercise 1.3.3. Check the details of the proof of the theorem. Show that the statement $\text{Var}(\hat{\boldsymbol{\theta}}) \geq \text{Var}(\tilde{\boldsymbol{\theta}})$ only uses that $\hat{\boldsymbol{\theta}}$ is unbiased and that $\mathbb{E}\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^*$ and $\text{Var}(\mathbf{Y}) = \sigma^2 I_n$.

Exercise 1.3.4. Compute $\nabla^2 L(\boldsymbol{\theta})$. Check that it is non-random, does not depend on $\boldsymbol{\theta}$, and fulfills for every $\boldsymbol{\theta}$ the identity

$$\nabla^2 L(\boldsymbol{\theta}) \equiv -\text{Var}[\nabla L(\boldsymbol{\theta})] = -D^2.$$

A colored noise

The majority of the presented results continue to apply in the case of heterogeneous and even dependent noise with $\text{Var}(\boldsymbol{\varepsilon}) = \Sigma_0$. The key facts behind this extension are the decomposition (1.10) and the properties of the stochastic component $\boldsymbol{\delta}$ from Section 1.3.1: $\boldsymbol{\delta} \sim \mathcal{N}(0, W^2)$. In the case of a colored noise, the definition of W and D is changed for

$$D^2 \stackrel{\text{def}}{=} W^{-2} = \Psi \Sigma_0^{-1} \Psi^\top.$$

Exercise 1.3.5. State and prove the analog of Theorem 1.3.4 for the colored noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma_0)$.

A misspecified LPA

An interesting feature of our results so far is that they equally apply for the correct linear specification $\mathbf{f}^* = \Psi^\top \boldsymbol{\theta}^*$ and for the case when the identity $\mathbf{f}^* = \Psi^\top \boldsymbol{\theta}$ is not precisely fulfilled whatever $\boldsymbol{\theta}$ is taken. In this situation the target of analysis is the vector $\boldsymbol{\theta}^*$ describing the best linear approximation of \mathbf{f}^* by $\Psi^\top \boldsymbol{\theta}$. We already know from the results of Section 1.3.1 and 1.3.2 that the estimate $\tilde{\boldsymbol{\theta}}$ is also normal with mean $\boldsymbol{\theta}^* = \mathcal{S}\mathbf{f}^* = (\Psi\Psi^\top)^{-1}\Psi\mathbf{f}^*$ and the variance $W^2 = \sigma^2 \mathcal{S}\mathcal{S}^\top = \sigma^2 (\Psi\Psi^\top)^{-1}$.

Theorem 1.3.5. Assume $\mathbf{Y} = \mathbf{f}^* + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Let $\boldsymbol{\theta}^* = \mathcal{S}\mathbf{f}^*$. Then $\tilde{\boldsymbol{\theta}}$ is R-efficient estimate of $\boldsymbol{\theta}^*$: $\mathbb{E}\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$,

$$\mathbb{E}[(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top] = \text{Var}(\tilde{\boldsymbol{\theta}}) = D^{-2},$$

and for any unbiased linear estimate $\hat{\boldsymbol{\theta}}$ satisfying $\mathbb{E}_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}} \equiv \boldsymbol{\theta}$, it holds

$$\text{Var}(\hat{\boldsymbol{\theta}}) \geq \text{Var}(\tilde{\boldsymbol{\theta}}) = D^{-2}.$$

Proof. The proofs only utilize that $\tilde{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}^*, W^2)$ with $W^2 = D^{-2}$. The only small remark concerns the equality $\text{Var}[\nabla L(\boldsymbol{\theta})] = D^2$ from Theorem 1.3.4.

Exercise 1.3.6. Check the identity $\text{Var}[\nabla L(\boldsymbol{\theta})] = D^2$ from Theorem 1.3.4 for $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma_0)$.

1.3.4 The case of a misspecified noise

Here we again consider the linear parametric assumption $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$. However, contrary to the previous section, we admit that the noise $\boldsymbol{\varepsilon}$ is not homogeneous normal:

$\varepsilon \sim \mathcal{N}(0, \Sigma_0)$ while our estimation procedure is the quasi MLE based on the assumption of noise homogeneity $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. We already know that the estimate $\tilde{\boldsymbol{\theta}}$ is unbiased with mean $\boldsymbol{\theta}^*$ and variance $W^2 = \mathcal{S}\Sigma_0\mathcal{S}^\top$, where $\mathcal{S} = (\Psi\Psi^\top)^{-1}\Psi$. This gives

$$W^2 = (\Psi\Psi^\top)^{-1}\Psi\Sigma_0\Psi^\top(\Psi\Psi^\top)^{-1}.$$

The question is whether the estimate $\tilde{\boldsymbol{\theta}}$ based on the misspecified distributional assumption is efficient. The Cramér-Rao result delivers the lower bound for the quadratic risk in form of $\text{Var}(\tilde{\boldsymbol{\theta}}) \geq [\text{Var}(\nabla L(\boldsymbol{\theta}))]^{-1}$. We already know that the use of the correctly specified covariance matrix of the errors leads to an R-efficient estimate $\tilde{\boldsymbol{\theta}}$. The next result show that the use of a misspecified matrix Σ results in an estimate which is unbiased but not R-efficient, that is, the best estimation risk is achieved if we apply the correct model assumptions.

Theorem 1.3.6. *Let $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \Sigma_0)$. Then*

$$\text{Var}[\nabla L(\boldsymbol{\theta})] = \Psi\Sigma_0^{-1}\Psi^\top.$$

The estimate $\tilde{\boldsymbol{\theta}} = (\Psi\Psi^\top)^{-1}\Psi\mathbf{Y}$ is unbiased, that is, $E\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^$, but it is not R-efficient unless $\Sigma_0 = \Sigma$.*

Proof. Let $\tilde{\boldsymbol{\theta}}_0$ be the MLE for the correct model specification with the noise $\varepsilon \sim \mathcal{N}(0, \Sigma_0)$. As $\tilde{\boldsymbol{\theta}}$ is unbiased, the difference $\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_0$ is orthogonal to $\tilde{\boldsymbol{\theta}}_0$ and it holds for the variance of $\tilde{\boldsymbol{\theta}}$

$$\text{Var}(\tilde{\boldsymbol{\theta}}) = \text{Var}(\tilde{\boldsymbol{\theta}}_0) + \text{Var}(\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_0);$$

cf. with the proof of Gauss-Markov-Theorem 1.3.4.

Exercise 1.3.7. Compare directly the variances of $\tilde{\boldsymbol{\theta}}$ and of $\tilde{\boldsymbol{\theta}}_0$.

1.4 Linear models and quadratic log-likelihood

Linear Gaussian modeling leads to a specific log-likelihood structure; see Section 1. Namely, the log-likelihood function $L(\boldsymbol{\theta})$ is quadratic in $\boldsymbol{\theta}$, the coefficients of the quadratic terms are deterministic and the cross term is linear both in $\boldsymbol{\theta}$ and in the observations Y_i . Here we show that this geometric structure of the log-likelihood characterizes linear models. We say that $L(\boldsymbol{\theta})$ is *quadratic* if it is a quadratic function of $\boldsymbol{\theta}$ and there is a deterministic symmetric matrix D^2 such that for any $\boldsymbol{\theta}^\circ, \boldsymbol{\theta}$

$$L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^\circ) = (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top \nabla L(\boldsymbol{\theta}^\circ) - (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top D^2(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)/2. \quad (1.13)$$

Here $\nabla L(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{dL(\boldsymbol{\theta})}{d\boldsymbol{\theta}}$. As usual we define

$$\begin{aligned}\tilde{\boldsymbol{\theta}} &\stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}), \\ \boldsymbol{\theta}^* &= \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E}L(\boldsymbol{\theta}).\end{aligned}$$

The next result describes some properties of the estimate $\tilde{\boldsymbol{\theta}}$ which are entirely based on the geometric (quadratic) structure of the function $L(\boldsymbol{\theta})$. All the results are stated by using the matrix D^2 and the vector $\zeta = \nabla L(\boldsymbol{\theta}^*)$.

Theorem 1.4.1. *Let $L(\boldsymbol{\theta})$ be quadratic for a matrix $D^2 > 0$. Then for any $\boldsymbol{\theta}^\circ$*

$$\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ = D^{-2} \nabla L(\boldsymbol{\theta}^\circ). \quad (1.14)$$

In particular, with $\boldsymbol{\theta}^\circ = 0$, it holds

$$\tilde{\boldsymbol{\theta}} = D^{-2} \nabla L(0).$$

Taking $\boldsymbol{\theta}^\circ = \boldsymbol{\theta}^*$ yields

$$\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = D^{-2} \zeta \quad (1.15)$$

with $\zeta \stackrel{\text{def}}{=} \nabla L(\boldsymbol{\theta}^*)$. Moreover, $\mathbb{E}\zeta = 0$, and it holds with $V^2 = \operatorname{Var}(\zeta) = \mathbb{E}\zeta\zeta^\top$

$$\mathbb{E}\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$$

$$\operatorname{Var}(\tilde{\boldsymbol{\theta}}) = D^{-2} V^2 D^{-2}.$$

Further, for any $\boldsymbol{\theta}$,

$$L(\tilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}) = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top D^2 (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) / 2 = \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\|^2 / 2. \quad (1.16)$$

Finally, it holds for the excess $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \stackrel{\text{def}}{=} L(\tilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$

$$2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top D^2 (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \zeta^\top D^{-2} \zeta = \|\xi\|^2 \quad (1.17)$$

with $\xi = D^{-1} \zeta$.

Proof. The extremal point equation $\nabla L(\boldsymbol{\theta}) = 0$ for the quadratic function $L(\boldsymbol{\theta})$ from (1.13) yields (1.14). The equation (1.13) with $\boldsymbol{\theta}^\circ = \boldsymbol{\theta}^*$ implies for any $\boldsymbol{\theta}$

$$\nabla L(\boldsymbol{\theta}) = \nabla L(\boldsymbol{\theta}^\circ) - D^2(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) = \zeta - D^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*). \quad (1.18)$$

Therefore, it holds for the expectation $\mathbb{E}L(\boldsymbol{\theta})$

$$\nabla \mathbb{E} L(\boldsymbol{\theta}) = \mathbb{E} \zeta - D^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*),$$

and the equation $\nabla \mathbb{E} L(\boldsymbol{\theta}^*) = 0$ implies $\mathbb{E} \zeta = 0$.

To show (1.16), apply again the property (1.13) with $\boldsymbol{\theta}^\circ = \tilde{\boldsymbol{\theta}}$:

$$\begin{aligned} L(\boldsymbol{\theta}) - L(\tilde{\boldsymbol{\theta}}) &= (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^\top \nabla L(\tilde{\boldsymbol{\theta}}) - (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^\top D^2(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})/2 \\ &= -(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top D^2(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})/2. \end{aligned}$$

Here we used that $\nabla L(\tilde{\boldsymbol{\theta}}) = 0$ because $\tilde{\boldsymbol{\theta}}$ is an extreme point of $L(\boldsymbol{\theta})$. The last result (1.17) is a special case with $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ in view of (1.15).

This theorem delivers an important message: the main properties of the MLE $\tilde{\boldsymbol{\theta}}$ can be explained via the geometric (quadratic) structure of the log-likelihood. An interesting question to clarify is whether a quadratic log-likelihood structure is specific for linear Gaussian model. The answer is positive: there is one-to-one correspondence between linear Gaussian models and quadratic log-likelihood functions. Indeed, the identity (1.18) with $\boldsymbol{\theta}^\circ = \boldsymbol{\theta}^*$ can be rewritten as

$$\nabla L(\boldsymbol{\theta}) + D^2\boldsymbol{\theta} \equiv \zeta + D^2\boldsymbol{\theta}^*.$$

If we fix any $\boldsymbol{\theta}$ and define $\mathbf{Y} = \nabla L(\boldsymbol{\theta}) + D^2\boldsymbol{\theta}$, this yields

$$\mathbf{Y} = D^2\boldsymbol{\theta}^* + \zeta.$$

Similarly, $\mathbf{Y} \stackrel{\text{def}}{=} D^{-1}\{\nabla L(\boldsymbol{\theta}) + D^2\boldsymbol{\theta}\}$ yields the equation

$$\mathbf{Y} = D\boldsymbol{\theta}^* + \xi, \tag{1.19}$$

where $\xi = D^{-1}\zeta$. We can summarize as follows.

Theorem 1.4.2. *Let $L(\boldsymbol{\theta})$ be quadratic with a non-degenerated matrix D^2 . Then $\mathbf{Y} \stackrel{\text{def}}{=} D^{-1}\{\nabla L(\boldsymbol{\theta}) + D^2\boldsymbol{\theta}\}$ does not depend on $\boldsymbol{\theta}$ and $L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$ is the quasi log-likelihood ratio for the linear Gaussian model (1.19) with ξ standard normal. It is the true log-likelihood if and only if $\zeta \sim \mathcal{N}(0, D^2)$.*

Proof. The model (1.19) with $\xi \sim \mathcal{N}(0, I_p)$ leads to the log-likelihood ratio

$$(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top D(\mathbf{Y} - D\boldsymbol{\theta}^*) - \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2 = (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \zeta - \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2$$

in view of the definition of \mathbf{Y} . The definition (1.13) implies

$$L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) = (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla L(\boldsymbol{\theta}^*) - \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2.$$

As these two expressions coincide, it follows that $L(\boldsymbol{\theta})$ is the true log-likelihood if and only if $\xi = D^{-1}\zeta$ is standard normal.

1.4.1 Inference based on the maximum likelihood

All the results presented above for linear models were based on the explicit representation of the (quasi) MLE $\tilde{\boldsymbol{\theta}}$. Here we present the approach based on the analysis of the maximum likelihood. This approach does not require to fix any analytic expression for the point of maximum of the (quasi) likelihood process $L(\boldsymbol{\theta})$. Instead we work directly with the maximum of this process. We establish exponential inequalities for the *excess* or the *maximum likelihood* $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$. We also show how these results can be used to study the accuracy of the MLE $\tilde{\boldsymbol{\theta}}$, in particular, for building confidence sets.

One more benefit of the ML-based approach is that it equally applies to a homogeneous and to a heterogeneous noise provided that the noise structure is not misspecified. The celebrated chi-squared result about the maximum likelihood $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ claims that the distribution of $2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ is chi-squared with p degrees of freedom χ_p^2 and it does not depend on the noise covariance; see Section 1.4.1.

Now we specify the setup. The starting point of the ML-approach is the linear Gaussian model assumption $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$. The corresponding log-likelihood ratio $L(\boldsymbol{\theta})$ can be written as

$$L(\boldsymbol{\theta}) = -\frac{1}{2}(\mathbf{Y} - \Psi^\top \boldsymbol{\theta})^\top \Sigma^{-1}(\mathbf{Y} - \Psi^\top \boldsymbol{\theta}) + R, \quad (1.20)$$

where the remainder term R does not depend on $\boldsymbol{\theta}$. Now one can see that $L(\boldsymbol{\theta})$ is a quadratic function of $\boldsymbol{\theta}$. Moreover, $\nabla^2 L(\boldsymbol{\theta}) = -\Psi \Sigma^{-1} \Psi^\top$, so that $L(\boldsymbol{\theta})$ is quadratic with $D^2 = \Psi \Sigma^{-1} \Psi^\top$. This enables us to apply the general results of Section 1.4 which are only based on the geometric (quadratic) structure of the log-likelihood $L(\boldsymbol{\theta})$: the true data distribution can be arbitrary.

Theorem 1.4.3. Consider $L(\boldsymbol{\theta})$ from (1.20). For any $\boldsymbol{\theta}$, it holds with $D^2 = \Psi \Sigma^{-1} \Psi^\top$

$$L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top D^2 (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) / 2. \quad (1.21)$$

In particular, if $\Sigma = \sigma^2 I_n$ then the fitted log-likelihood is proportional to the quadratic loss $\|\tilde{\mathbf{f}} - \mathbf{f}_\theta\|^2$ for $\tilde{\mathbf{f}} = \Psi^\top \tilde{\boldsymbol{\theta}}$ and $\mathbf{f}_\theta = \Psi^\top \boldsymbol{\theta}$:

$$L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{2\sigma^2} \|\Psi^\top (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\|^2 = \frac{1}{2\sigma^2} \|\tilde{\mathbf{f}} - \mathbf{f}_\theta\|^2.$$

If $\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E} L(\boldsymbol{\theta}) = D^{-2} \Psi \Sigma^{-1} \mathbf{f}^*$ for $\mathbf{f}^* = \mathbb{E} \mathbf{Y}$, then

$$2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \boldsymbol{\zeta}^\top D^{-2} \boldsymbol{\zeta} = \|\boldsymbol{\xi}\|^2 \quad (1.22)$$

with $\boldsymbol{\zeta} = \nabla L(\boldsymbol{\theta}^*)$ and $\boldsymbol{\xi} \stackrel{\text{def}}{=} D^{-1} \boldsymbol{\zeta}$.

Proof. The results (1.21) and (1.22) follow from Theorem 1.4.1; see (1.16) and (1.17).

If the model assumptions are not misspecified one can establish the remarkable χ^2 result.

Theorem 1.4.4. Let $L(\boldsymbol{\theta})$ from (1.20) be the log-likelihood for the model $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$. Then $\boldsymbol{\xi} = D^{-1}\boldsymbol{\zeta} \sim \mathcal{N}(0, I_p)$ and $2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \sim \chi_p^2$ is chi-squared with p degrees of freedom.

Proof. By direct calculus

$$\boldsymbol{\zeta} = \nabla L(\boldsymbol{\theta}^*) = \Psi \Sigma^{-1} (\mathbf{Y} - \Psi^\top \boldsymbol{\theta}^*) = \Psi \Sigma^{-1} \boldsymbol{\varepsilon}.$$

So, $\boldsymbol{\zeta}$ is a linear transformation of a Gaussian vector \mathbf{Y} and thus it is Gaussian as well.

By Theorem 1.4.1, $\mathbb{E}\boldsymbol{\zeta} = 0$. Moreover, $\text{Var}(\boldsymbol{\varepsilon}) = \Sigma$ implies

$$\text{Var}(\boldsymbol{\zeta}) = \mathbb{E}\boldsymbol{\zeta}\boldsymbol{\zeta}^\top = \mathbb{E}\Psi \Sigma^{-1} \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top \Sigma^{-1} \Psi^\top = \Psi \Sigma^{-1} \Psi^\top = D^2$$

yielding that $\boldsymbol{\xi} = D^{-1}\boldsymbol{\zeta}$ is standard normal.

The last result $2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \sim \chi_p^2$ is sometimes called the “chi-squared phenomenon”: the distribution of the maximum likelihood only depends on the number of parameters to be estimated and is independent of the design Ψ , of the noise covariance matrix Σ , etc. This particularly explains the use of word “phenomenon” in the name of the result.

Exercise 1.4.1. Check that the linear transformation $\check{\mathbf{Y}} = \Sigma^{-1/2}\mathbf{Y}$ of the data does not change the value of the log-likelihood ratio $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ and hence, of the maximum likelihood $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$.

Hint: use the representation

$$\begin{aligned} L(\boldsymbol{\theta}) &= \frac{1}{2}(\mathbf{Y} - \Psi^\top \boldsymbol{\theta})^\top \Sigma^{-1} (\mathbf{Y} - \Psi^\top \boldsymbol{\theta}) + R \\ &= \frac{1}{2}(\check{\mathbf{Y}} - \check{\Psi}^\top \boldsymbol{\theta})^\top (\check{\mathbf{Y}} - \check{\Psi}^\top \boldsymbol{\theta}) + R \end{aligned}$$

and check that the transformed data $\check{\mathbf{Y}}$ is described by the model $\check{\mathbf{Y}} = \check{\Psi}^\top \boldsymbol{\theta}^* + \check{\boldsymbol{\varepsilon}}$ with $\check{\Psi} = \Psi \Sigma^{-1/2}$ and $\check{\boldsymbol{\varepsilon}} = \Sigma^{-1/2} \boldsymbol{\varepsilon} \sim \mathcal{N}(0, I_n)$ yielding the same log-likelihood ratio as in the original model.

Exercise 1.4.2. Assume homogeneous noise in (1.20) with $\Sigma = \sigma^2 I_n$. Then it holds

$$2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \sigma^{-2} \|\Pi \boldsymbol{\varepsilon}\|^2$$

where $\Pi = \Psi^\top (\Psi \Psi^\top)^{-1} \Psi$ is the projector in \mathbb{R}^n on the subspace spanned by the vectors $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_p$.

Hint: use that $\zeta = \sigma^{-2}\Psi\varepsilon$, $D^2 = \sigma^{-2}\Psi\Psi^\top$, and

$$\sigma^{-2}\|\Pi\varepsilon\|^2 = \sigma^{-2}\varepsilon^\top\Pi^\top\Pi\varepsilon = \sigma^{-2}\varepsilon^\top\Pi\varepsilon = \zeta^\top D^{-2}\zeta.$$

We write the result of Theorem 1.4.3 in the form $2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \sim \chi_p^2$, where χ_p^2 stands for the chi-squared distribution with p degrees of freedom. This result can be used to build likelihood-based confidence ellipsoids for the parameter $\boldsymbol{\theta}^*$. Given $\mathfrak{z} > 0$, define

$$\mathcal{E}(\mathfrak{z}) = \{\boldsymbol{\theta} : L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \leq \mathfrak{z}\} = \left\{\boldsymbol{\theta} : \sup_{\boldsymbol{\theta}'} L(\boldsymbol{\theta}') - L(\boldsymbol{\theta}) \leq \mathfrak{z}\right\}. \quad (1.23)$$

Theorem 1.4.5. Assume $\mathbf{Y} = \Psi^\top\boldsymbol{\theta}^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \Sigma)$ and consider the MLE $\tilde{\boldsymbol{\theta}}$. Define \mathfrak{z}_α by $P(\chi_p^2 > 2\mathfrak{z}_\alpha) = \alpha$. Then $\mathcal{E}(\mathfrak{z}_\alpha)$ from (1.23) is an α -confidence set for $\boldsymbol{\theta}^*$.

Exercise 1.4.3. Let $D^2 = \Psi\Sigma^{-1}\Psi^\top$. Check that the likelihood-based CS $\mathcal{E}(\mathfrak{z}_\alpha)$ and estimate-based CS $E(z_\alpha) = \{\boldsymbol{\theta} : \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\| \leq z_\alpha\}$, $z_\alpha^2 = 2\mathfrak{z}_\alpha$, coincide in the case of the linear modeling:

$$\mathcal{E}(\mathfrak{z}_\alpha) = \{\boldsymbol{\theta} : \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\|^2 \leq 2\mathfrak{z}_\alpha\}.$$

Another corollary of the chi-squared result is a concentration bound for the maximum likelihood. A similar result was stated for the univariate exponential family model: the value $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ is stochastically bounded with exponential moments, and the bound does not depend on the particular family, parameter value, sample size, etc. Now we can extend this result to the case of a linear Gaussian model. Indeed, Theorem 1.4.3 states that the distribution of $2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ is chi-squared and only depends on the number of parameters to be estimated. The latter distribution concentrates on the ball of radius of order $p^{1/2}$ and the deviation probability is exponentially small.

Theorem 1.4.6. Assume $\mathbf{Y} = \Psi^\top\boldsymbol{\theta}^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \Sigma)$. Then for every $x > 0$

$$\begin{aligned} \mathbb{P}(2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) > p + 2\sqrt{xp} + 2x) \\ &= \mathbb{P}(\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 > p + 2\sqrt{xp} + 2x) \leq \exp(-x). \end{aligned} \quad (1.24)$$

Proof. Define $\xi \stackrel{\text{def}}{=} D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$. By Theorem 1.3.4 ξ is standard normal vector in \mathbb{R}^p and by Theorem 1.4.3 $2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \|\xi\|^2$. Now the statement (1.24) follows from the general deviation bound for the Gaussian quadratic forms; see Corollary B.1.2.

The main message of this result can be explained as follows: the deviation probability that the estimate $\tilde{\boldsymbol{\theta}}$ does not belong to the elliptic set $E(z) = \{\boldsymbol{\theta} : \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\| \leq z\}$

starts to vanish when z^2 exceeds the dimensionality p of the parameter space. Similarly, the coverage probability that the true parameter θ^* is not covered by the confidence set $\mathcal{E}(\mathfrak{z})$ starts to vanish when $2\mathfrak{z}$ exceeds p .

Corollary 1.4.1. *Assume $\mathbf{Y} = \Psi^\top \theta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \Sigma)$. Then for every $x > 0$, it holds with $2\mathfrak{z} = p + 2\sqrt{x p} + 2x$*

$$P(\mathcal{E}(\mathfrak{z}) \not\ni \theta^*) \leq \exp(-x).$$

Exercise 1.4.4. Compute \mathfrak{z} ensuring the covering of 95% in the dimension $p = 1, 2, 10, 20$.

1.4.2 A misspecified LPA

Now we discuss the behavior of the fitted log-likelihood for the misspecified linear parametric assumption $I\mathbb{E}\mathbf{Y} = \Psi^\top \theta^*$. Let the response function f^* not be linearly expandable as $f^* = \Psi^\top \theta^*$. Following to Theorem 1.3.3, define $\theta^* = \mathcal{S}f^*$ with $\mathcal{S} = (\Psi\Sigma^{-1}\Psi^\top)^{-1}\Psi\Sigma^{-1}$. This point provides the best approximation of the nonlinear response f^* by a linear parametric fit $\Psi^\top \theta$.

Theorem 1.4.7. *Assume $\mathbf{Y} = f^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \Sigma)$. Let $\theta^* = \mathcal{S}f^*$. Then $\tilde{\theta}$ is an R-efficient estimate of θ^* and*

$$2L(\tilde{\theta}, \theta^*) = \zeta^\top D^{-2} \zeta = \|\xi\|^2 \sim \chi_p^2,$$

where $D^2 = \Psi\Sigma^{-1}\Psi^\top$, $\zeta = \nabla L(\theta^*) = \Psi\Sigma^{-1}\varepsilon$, $\xi = D^{-1}\zeta$ is standard normal vector in \mathbb{R}^p and χ_p^2 is a chi-squared random variable with p degrees of freedom. In particular, $\mathcal{E}(\mathfrak{z}_\alpha)$ is an α -CS for the vector θ^* and the bound of Corollary 1.4.1 applies.

Exercise 1.4.5. Prove the result of Theorem 1.4.7.

1.4.3 A misspecified noise structure

This section addresses the question about the features of the maximum likelihood in the case when the likelihood is built under a wrong assumption about the noise structure. As one can expect, the chi-squared result is not valid anymore in this situation and the distribution of the maximum likelihood depends on the true noise covariance. However, the nice geometric structure of the maximum likelihood manifested by Theorems 1.4.3 and 1.4.5 does not rely on the true data distribution and it is only based on our structural assumptions on the considered model. This helps to get rigorous results about the behaviors of the maximum likelihood and particularly about its concentration properties.

Theorem 1.4.8. Let $\tilde{\boldsymbol{\theta}}$ be built for the model $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$, while the true noise covariance is $\Sigma_0 : \mathbb{E}\boldsymbol{\varepsilon} = 0$ and $\text{Var}(\boldsymbol{\varepsilon}) = \Sigma_0$. Then

$$\begin{aligned}\mathbb{E}\tilde{\boldsymbol{\theta}} &= \boldsymbol{\theta}^*, \\ \text{Var}(\tilde{\boldsymbol{\theta}}) &= D^{-2}W^2D^{-2},\end{aligned}$$

where

$$\begin{aligned}D^2 &= \Psi\Sigma^{-1}\Psi^\top, \\ W^2 &= \Psi\Sigma^{-1}\Sigma_0\Sigma^{-1}\Psi^\top.\end{aligned}$$

Further,

$$2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 = \|\boldsymbol{\xi}\|^2, \quad (1.25)$$

where $\boldsymbol{\xi}$ is a random vector in \mathbb{R}^p with $\mathbb{E}\boldsymbol{\xi} = 0$ and

$$\text{Var}(\boldsymbol{\xi}) = B \stackrel{\text{def}}{=} D^{-1}W^2D^{-1}.$$

Moreover, if $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma_0)$, then $\tilde{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}^*, D^{-2}W^2D^{-2})$ and $\boldsymbol{\xi} \sim \mathcal{N}(0, B)$.

Proof. The moments of $\tilde{\boldsymbol{\theta}}$ have been computed in Theorem 1.4.1 while the equality $2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 = \|\boldsymbol{\xi}\|^2$ is given in Theorem 1.4.3. Next, $\boldsymbol{\zeta} = \nabla L(\boldsymbol{\theta}^*) = \Psi\Sigma^{-1}\boldsymbol{\varepsilon}$ and

$$W^2 \stackrel{\text{def}}{=} \text{Var}(\boldsymbol{\zeta}) = \Psi\Sigma^{-1}\text{Var}(\boldsymbol{\varepsilon})\Sigma^{-1}\Psi^\top = \Psi\Sigma^{-1}\Sigma_0\Sigma^{-1}\Psi^\top.$$

This implies that

$$\text{Var}(\boldsymbol{\xi}) = \mathbb{E}\boldsymbol{\xi}\boldsymbol{\xi}^\top = D^{-1}\text{Var}(\boldsymbol{\zeta})D^{-1} = D^{-1}W^2D^{-1}.$$

It remains to note that if $\boldsymbol{\varepsilon}$ is a Gaussian vector, then $\boldsymbol{\zeta} = \Psi\Sigma^{-1}\boldsymbol{\varepsilon}$, $\boldsymbol{\xi} = D^{-1}\boldsymbol{\zeta}$, and $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = D^{-2}\boldsymbol{\zeta}$ are Gaussian as well.

Exercise 1.4.6. Check that $\Sigma_0 = \Sigma$ leads back to the χ^2 -result.

One can see that the chi-squared result is not valid any more if the noise structure is misspecified. An interesting question is whether the CS $\mathcal{E}(\mathfrak{z})$ can be applied in the case of a misspecified noise under some proper adjustment of the value \mathfrak{z} . Surprisingly, the answer is not entirely negative. The reason is that the vector $\boldsymbol{\xi}$ from (1.25) is zero mean and its norm has a similar behavior as in the case of the correct noise specification: the probability $\mathbb{P}(\|\boldsymbol{\xi}\| > z)$ starts to degenerate when z^2 exceeds $\mathbb{E}\|\boldsymbol{\xi}\|^2$. A general bound from Theorem B.1.1 in Section A.1 implies the following bound for the coverage probability.

Corollary 1.4.2. *Under the conditions of Theorem 1.4.8, for every $\mathbf{x} > 0$, it holds with $p = \text{tr}(B)$, $v^2 = \text{tr}(B^2)$, and $\lambda = \|B\|_\infty$*

$$\mathbb{P}(2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) > p + 2vx^{1/2} + 2\lambda x) \leq \exp(-x).$$

Exercise 1.4.7. Show that an overestimation of the noise in the sense $\Sigma \geq \Sigma_0$ preserves the coverage probability for the CS $\mathcal{E}(\mathfrak{z}_\alpha)$, that is, if $2\mathfrak{z}_\alpha$ is the $1 - \alpha$ quantile of χ_p^2 , then $\mathbb{P}(\mathcal{E}(\mathfrak{z}_\alpha) \not\ni \boldsymbol{\theta}^*) \leq \alpha$.

Linear regression with random design

2.1 Random design linear regression

Consider the linear regression equation

$$\mathbf{Y} = \boldsymbol{\Psi}^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}, \quad (2.1)$$

where $\boldsymbol{\theta}^* \in \mathbb{R}^p$ is the target parameter, \mathbf{Y} is the n -vector of responses, $\boldsymbol{\varepsilon}$ is the n -vector of errors, and $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1, \dots, \boldsymbol{\Psi}_n)$ is a $p \times n$ design matrix with columns $\boldsymbol{\Psi}_i \in \mathbb{R}^p$. The assumption of homogeneous Gaussian errors $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_p)$ yields the corresponding Gaussian log-likelihood

$$L(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2} \|\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}\|^2 + R$$

and the MLE

$$\tilde{\boldsymbol{\theta}} = (\boldsymbol{\Psi} \boldsymbol{\Psi}^\top)^{-1} \boldsymbol{\Psi} \mathbf{Y}.$$

Below we study its properties under the assumptions of independent errors ε_i and random independent design vectors $\boldsymbol{\Psi}_i$. Let the error vector $\boldsymbol{\varepsilon}$ satisfy

$$\mathbb{E}\boldsymbol{\varepsilon} = 0, \quad \text{Var}(\boldsymbol{\varepsilon}) = \Sigma = \text{diag}\{\sigma_1^2, \dots, \sigma_n^2\}. \quad (2.2)$$

For the case of a random design $\boldsymbol{\Psi}$, analysis of the MLE $\tilde{\boldsymbol{\theta}}$ becomes more involved because of inversion of the random matrix $\boldsymbol{\Psi} \boldsymbol{\Psi}^\top$. The results below present some conditions under which this random matrix can be replaced by its expectation.

2.2 Design matrix and design distribution

This section shows that in typical situations, the empirical design matrix is close to its population counterparts. We discuss separately two cases: design with independent measurements and an aggregated design.

2.2.1 Design with independent measurements

Let the feature vectors Ψ_1, \dots, Ψ_n in (2.1) be independent. We assume that the design is non degenerate, so that the matrix $\mathbf{M}^2 \stackrel{\text{def}}{=} \mathbb{E}(\Psi\Psi^\top)$ is positive. Consider

$$\mathbf{M}^{-1}\{\Psi\Psi^\top - \mathbb{E}(\Psi\Psi^\top)\}\mathbf{M}^{-1} = A_1 + \dots + A_n,$$

where

$$A_i = \mathbf{M}^{-1}\{\Psi_i\Psi_i^\top - \mathbb{E}(\Psi_i\Psi_i^\top)\}\mathbf{M}^{-1} \quad (2.3)$$

is a symmetric $p \times p$ random matrix with $\mathbb{E}A_i = 0$. Also define the variance parameter

$$S_n^2 \stackrel{\text{def}}{=} \|\mathbb{E}(A_1^2 + \dots + A_n^2)\|_{\text{op}}. \quad (2.4)$$

We also assume that all design vectors Ψ_i are uniformly bounded with probability one. This implies a uniform bound

$$\|A_i\|_{\text{op}} \leq u_n \quad a.s. \quad (2.5)$$

for a small constant u_n . In the case of an i.i.d. design, define

$$\begin{aligned} M_1^2 &\stackrel{\text{def}}{=} \mathbb{E}(\Psi_1\Psi_1^\top), \\ \sigma_1^2 &\stackrel{\text{def}}{=} \mathbb{E}(M_1^{-1}\Psi_1\Psi_1^\top M_1^{-1} - I_p)^2. \end{aligned}$$

Also suppose that with probability one

$$\|M_1^{-1}\Psi_1\Psi_1^\top M_1^{-1} - I_p\|_{\text{op}} \leq u^*.$$

Then it holds

$$\begin{aligned} \mathbf{M}^2 &= n M_1^2, \\ S_n^2 &= n^{-1} \sigma_1^2, \\ u_n &\leq n^{-1} u^* \end{aligned} \quad (2.6)$$

The matrix Bernstein inequality; see Theorem C.1.2, yields:

Theorem 2.2.1. Suppose that Ψ_i are independent and A_i from (2.3) fulfill (2.5). Then with $\mathbf{M}^2 = \mathbb{E}(\Psi\Psi^\top)$ and S_n^2 defined by (2.4), it holds for all $z > 0$

$$\mathbb{P}\left(\|\mathbf{M}^{-1}\Psi\Psi^\top\mathbf{M}^{-1} - I_p\|_{\text{op}} > z\right) \leq 2p \exp\left\{-\frac{z^2}{2S_n^2 + 2u_n z/3}\right\}.$$

If Ψ_i are i.i.d. and (2.6) holds then

$$\mathbb{P}\left(n^{1/2}\|\mathbf{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\mathbf{M}^{-1} - \mathbf{I}_p\|_{\text{op}} > z\right) \leq 2p \exp\left\{-\frac{z^2}{2\sigma_1^2 + 2n^{-1/2}u^*z/3}\right\}.$$

Proof. (please check)

For any fixed \mathbf{x} and $\delta > 0$, one can fix any n satisfying

$$n \geq (2\sigma_1^2\delta^{-2} + 2u^*\delta^{-1}/3)\{\mathbf{x} + \log(2p)\} \quad (2.7)$$

to ensure

$$\mathbb{P}\left(\|\mathbf{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\mathbf{M}^{-1} - \mathbf{I}_p\|_{\text{op}} > \delta\right) \leq e^{-\mathbf{x}}. \quad (2.8)$$

If n and \mathbf{x} are fixed, then one can check (2.8) for

$$\delta = \delta(p, \mathbf{x}) = \sqrt{\frac{2\sigma_1^2}{n}(\mathbf{x} + \log(2p))} + \frac{2u^*}{3n}(\mathbf{x} + \log(2p)). \quad (2.9)$$

Corollary 2.2.1. Suppose Ψ_i are i.i.d. and (2.6) holds. If n fulfills (2.7) for some fixed δ and \mathbf{x} , then (2.8) holds true. Similarly, if n and \mathbf{x} are fixed, then δ from (2.9) ensures (2.8).

Proof. (please check).

The result (2.8) guarantees that on a set $\Omega(\mathbf{x})$ with $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - e^{-\mathbf{x}}$

$$\|\mathbf{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\mathbf{M}^{-1} - \mathbf{I}_p\|_{\text{op}} \leq \delta. \quad (2.10)$$

This also implies for any $\boldsymbol{\gamma} \in \mathbb{R}^p$

$$(1 - \delta)\boldsymbol{\gamma}^\top\mathbf{M}^2\boldsymbol{\gamma} \leq \boldsymbol{\gamma}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\boldsymbol{\gamma} \leq (1 + \delta)\boldsymbol{\gamma}^\top\mathbf{M}^2\boldsymbol{\gamma}.$$

(please check).

2.2.2 Aggregated random design

Here we discuss another random design setup for the regression model (2.1). Namely, assume that the design matrix $\boldsymbol{\Psi}$ can be represented as a sum of independent $p \times q$ matrices Ψ_1, \dots, Ψ_n :

$$\boldsymbol{\Psi} = \Psi_1 + \dots + \Psi_n. \quad (2.11)$$

It is natural to expect that this model is close to the usual regression model in which the random matrix $\boldsymbol{\Psi}$ is replaced by its expectation $\mathbb{E}\boldsymbol{\Psi}$. Consider the product $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top$ which

has to be close to the corresponding product $\mathbf{M}^2 \stackrel{\text{def}}{=} \mathbb{E}\Psi \mathbb{E}(\Psi^\top)$. We aim at bounding the normalized difference $\Psi - \mathbb{E}\Psi$ in the operator norm. Define for $i = 1, \dots, n$

$$V_i^2 \stackrel{\text{def}}{=} \mathbb{E}(\Psi_i \Psi_i^\top) - \mathbb{E}\Psi_i \mathbb{E}\Psi_i^\top,$$

and

$$S_n^2 \stackrel{\text{def}}{=} \|\mathbf{M}^{-1}(V_1^2 + \dots + V_n^2)\mathbf{M}^{-1}\|_{\text{op}}.$$

Typically S_n^2 is inversely proportional to n . We again assume that the norms of all design vectors Ψ_i are uniformly bounded with probability one. This implies a uniform bound

$$\|\mathbf{M}^{-1}(\Psi_i - \mathbb{E}\Psi_i)\|_{\text{op}} \leq u_n \quad a.s.$$

for a small constant u_n . Now consider

$$\mathbf{M}^{-1}(\Psi - \mathbb{E}\Psi) = \sum_{i=1}^n \mathbf{M}^{-1}(\Psi_i - \mathbb{E}\Psi_i).$$

The matrix Bernstein inequality; see Theorem C.1.3, yields for any $z \geq 0$

$$\mathbb{P}(\|\mathbf{M}^{-1}(\Psi - \mathbb{E}\Psi)\|_{\text{op}} \geq z) \leq (p+q) \exp\left\{-\frac{z^2}{2S_n^2 + 2u_n z/3}\right\} \quad (2.12)$$

In the case with i.i.d. Ψ_i , define

$$\begin{aligned} M_1^2 &\stackrel{\text{def}}{=} \mathbb{E}\Psi_1 \mathbb{E}\Psi_1^\top, \\ \sigma_1^2 &\stackrel{\text{def}}{=} M_1^{-1} \mathbb{E}(\Psi_1 \Psi_1^\top) M_1^{-1} - I_p, \end{aligned}$$

and suppose that

$$\|M_1^{-1}(\Psi_1 - \mathbb{E}\Psi_1)\|_{\text{op}} \leq u_1.$$

Then it holds

$$\begin{aligned} \mathbf{M}^2 &= n^2 M_1^2, \\ u_n &\leq n^{-1} u_1^2, \end{aligned}$$

and

$$\mathbb{P}\left(n^{1/2} \|\mathbf{M}^{-1}(\Psi - \mathbb{E}\Psi)\|_{\text{op}} \geq z\right) \leq (p+q) \exp\left\{-\frac{z^2}{2\sigma_1^2 + 2u_1 z/(3n^{1/2})}\right\}.$$

The result (2.12) implies that Ψ is close to $E\Psi$. The MLE $\tilde{\theta}$ also involves the product $\Psi\Psi^\top$ which has to be close to the corresponding product $M^2 = E\Psi E(\Psi^\top)$. Below we assume that M is sufficiently large and bound the difference $M^{-1}\Psi\Psi^\top M^{-1} - I_p$.

Theorem 2.2.2. *Let $\|M^{-1}(\Psi - E\Psi)\|_{\text{op}} \leq \delta$ for some $\delta > 0$. Then*

$$\|M^{-1}\Psi\Psi^\top M^{-1} - I_p\|_{\text{op}} \leq \delta^2 + 2\delta.$$

Proof. One can bound

$$\begin{aligned} & M^{-1}\Psi\Psi^\top M^{-1} - I_p \\ &= M^{-1}(\Psi - E\Psi)(\Psi - E\Psi)^\top M^{-1} + 2M^{-1}(\Psi - E\Psi)E(\Psi^\top)M^{-1}. \end{aligned}$$

For any unit vector $\gamma \in \mathbb{R}^p$, the definition of M implies

$$\|E(\Psi^\top)M^{-1}\gamma\|^2 = \gamma^\top M^{-1}E(\Psi)E(\Psi^\top)M^{-1}\gamma = \|\gamma\|^2 = 1.$$

Therefore,

$$\|M^{-1}(\Psi - E\Psi)E(\Psi^\top)M^{-1}\gamma\| \leq \|M^{-1}(\Psi - E\Psi)\|_{\text{op}}$$

thus

$$\|M^{-1}\Psi\Psi^\top M^{-1} - I_p\|_{\text{op}} \leq \|M^{-1}(\Psi - E\Psi)\|_{\text{op}}^2 + 2\|M^{-1}(\Psi - E\Psi)\|_{\text{op}},$$

and the result follows.

Theorem 2.3.2 applies in this situation without any change.

2.3 Fisher and Wilks expansions for the MLE under random design

Now we apply this result to the MLE in the regression model (2.1) under possible misspecification of the error variance. The corresponding log-likelihood ratio can be written in the form

$$L(\theta, \theta^*) = \frac{1}{\sigma^2}(\mathbf{Y} - \Psi^\top \theta^*)^\top \Psi^\top (\theta - \theta^*) - \frac{1}{2\sigma^2}(\theta - \theta^*)^\top \Psi \Psi^\top (\theta - \theta^*).$$

Introduce also an approximating quadratic log-likelihood defined by

$$\mathbb{L}(\theta, \theta^*) = \frac{1}{\sigma^2}(\mathbf{Y} - \Psi^\top \theta^*)^\top \Psi^\top (\theta - \theta^*) - \frac{1}{2\sigma^2}(\theta - \theta^*)^\top M^2 (\theta - \theta^*) \quad (2.13)$$

with $M^2 = E\Psi \Psi^\top$. Note that two expressions $L(\theta, \theta^*)$, $\mathbb{L}(\theta, \theta^*)$ differ only in the quadratic term. Moreover, we already know that $\Psi \Psi^\top$ is close to its expectation $M^2 = E\Psi \Psi^\top$; see (2.10).

Theorem 2.3.1. Let (2.10) hold on a set $\Omega(\mathbf{x})$ with $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - e^{-\mathbf{x}}$. Then with

$$D^2 = \sigma^{-2} \mathbf{M}^2 = \sigma^{-2} \mathbb{E}(\Psi \Psi^\top), \quad (2.14)$$

it holds on $\Omega(\mathbf{x})$ for $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}$

$$\begin{aligned} |L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)| &\leq \delta \mathbf{r}^2 / 2, \\ \|D^{-1}\{\nabla L(\boldsymbol{\theta}) - \nabla \mathbb{L}(\boldsymbol{\theta})\}\| &\leq \delta \mathbf{r}. \end{aligned}$$

Proof. The difference between $L(\boldsymbol{\theta})$ and $\mathbb{L}(\boldsymbol{\theta})$ can be written as

$$\begin{aligned} \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= \frac{1}{2\sigma^2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top (\Psi \Psi^\top - \mathbf{M}^2) (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\ &= \frac{1}{2} \{D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}^\top (\mathbf{M}^{-1} \Psi \Psi^\top \mathbf{M}^{-1} - I_p) \{D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}. \end{aligned} \quad (2.15)$$

Therefore, it holds for $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}$ on $\Omega(\mathbf{x})$ by (2.10)

$$|L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)| \leq \delta \mathbf{r}^2 / 2.$$

Differentiating the identity (2.15) in $\boldsymbol{\theta}$ yields for the gradients $\nabla L(\boldsymbol{\theta})$ and $\nabla \mathbb{L}(\boldsymbol{\theta})$

$$D^{-1}\{\nabla L(\boldsymbol{\theta}) - \nabla \mathbb{L}(\boldsymbol{\theta})\} = (\mathbf{M}^{-1} \Psi \Psi^\top \mathbf{M}^{-1} - I_p) \{D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}$$

and thus, for any $\boldsymbol{\theta}$ with $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}$, (2.10) implies on $\Omega(\mathbf{x})$

$$\|D^{-1}\{\nabla L(\boldsymbol{\theta}) - \nabla \mathbb{L}(\boldsymbol{\theta})\}\| \leq \delta \mathbf{r}.$$

One can represent the approximating process $\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ from (2.13) as

$$\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \boldsymbol{\xi}^\top D(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top D^2 (\boldsymbol{\theta} - \boldsymbol{\theta}^*),$$

where the random p vector $\boldsymbol{\xi}$ is defined as

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} \sigma^{-2} D^{-1} \Psi (\mathbf{Y} - \Psi^\top \boldsymbol{\theta}^*) = \sigma^{-2} D^{-1} \Psi \boldsymbol{\varepsilon} = \sigma^{-1} \mathbf{M}^{-1} \Psi \boldsymbol{\varepsilon}. \quad (2.16)$$

For studying the properties of the MLE $\tilde{\boldsymbol{\theta}}$ and of the maximum log-likelihood $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\tilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta})$, one can use the expansions from Theorem 2.3.1. However, we present a direct proof based on quadraticity of $L(\boldsymbol{\theta})$.

Theorem 2.3.2. Consider the model (2.1) and suppose

$$\|\mathbf{M}^{-1} \Psi \Psi^\top \mathbf{M}^{-1} - I_p\|_{\text{op}} \leq \delta \quad (2.17)$$

for $\mathbf{M}^2 = \mathbb{E}(\Psi\Psi^\top)$ and some $\delta < 1/2$ on a dominating set $\Omega(\mathbf{x})$. Then the MLE $\tilde{\boldsymbol{\theta}}$ fulfills on $\Omega(\mathbf{x})$ for D^2 from (2.14) and ξ from (2.16)

$$\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \xi\| \leq \frac{\delta}{1-\delta} \|\xi\|, \quad (2.18)$$

$$|2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\xi\|^2| \leq \frac{\delta}{1-\delta} \|\xi\|^2, \quad (2.19)$$

$$\left| \sqrt{2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)} - \|\xi\| \right| \leq \frac{\delta}{1-\delta} \|\xi\|. \quad (2.20)$$

Proof. The bound (2.17) also implies

$$\|\mathbf{M}(\Psi\Psi^\top)^{-1}\mathbf{M} - I_p\|_{\text{op}} \leq \frac{\delta}{1-\delta}. \quad (2.21)$$

By using quadraticity of $L(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$, one obtains (cf. Theorem 1.4.1 in Section 1.4)

$$\begin{aligned} \tilde{\boldsymbol{\theta}} &= (\Psi\Psi^\top)^{-1}\Psi\mathbf{Y}, \\ L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= \frac{1}{2\sigma^2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \Psi\Psi^\top(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \frac{1}{2\sigma^2} \|\Psi^\top(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\|^2. \end{aligned}$$

Further, the model equation (2.1) implies with $\xi = \sigma^{-1}\mathbf{M}^{-1}\Psi\varepsilon$

$$D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = D(\Psi\Psi^\top)^{-1}\Psi\varepsilon = \mathbf{M}(\Psi\Psi^\top)^{-1}\mathbf{M}\xi = A_\Psi\xi$$

with $A_\Psi \stackrel{\text{def}}{=} \mathbf{M}(\Psi\Psi^\top)^{-1}\mathbf{M}$ and thus

$$\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \xi\| = \|(A_\Psi - I_p)\xi\|$$

so that (2.18) follows from (2.21). Similarly

$$\begin{aligned} 2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) &= \sigma^{-2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \Psi\Psi^\top(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ &= \sigma^{-2}((\Psi\Psi^\top)^{-1}\Psi\varepsilon)^\top \Psi\Psi^\top((\Psi\Psi^\top)^{-1}\Psi\varepsilon) \\ &= \xi^\top A_\Psi \xi \end{aligned}$$

yielding (2.19) on $\Omega(\mathbf{x})$. Finally, as $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \geq 0$, it holds

$$\left| \sqrt{2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)} - \|\xi\| \right| \leq \frac{|2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\xi\|^2|}{\|\xi\|}$$

yielding (2.20) by (2.19).

2.4 A deviation bound for ξ

This section presents a bound on the norm of a vector $\xi = \sigma^{-1} \mathbf{M}^{-1} \Psi \varepsilon$ for a random design Ψ and independent of Ψ Gaussian errors ε . The results easily extend to the case of non-Gaussian errors ε under exponential moment conditions.

One can easily compute the moments of ξ conditioned on the design: by (2.2)

$$\mathbb{E}(\xi | \Psi) = 0, \quad \text{Var}(\varepsilon | \Psi) = \Sigma, \quad \text{and}$$

$$\begin{aligned} \mathcal{A}_\xi &\stackrel{\text{def}}{=} \text{Var}(\xi | \Psi) = \sigma^{-2} \mathbf{M}^{-1} (\Psi \Sigma \Psi^\top) \mathbf{M}^{-1} \\ &= \sigma^{-2} \mathbf{M}^{-1} \left(\sum_{i=1}^n \sigma_i^2 \Psi_i \Psi_i^\top \right) \mathbf{M}^{-1}. \end{aligned} \quad (2.22)$$

Define

$$B_\xi \stackrel{\text{def}}{=} \text{Var}(\xi) = \mathbb{E} \mathcal{A}_\xi = \sigma^{-2} \mathbf{M}^{-1} \mathbb{E} \left(\sum_{i=1}^n \sigma_i^2 \Psi_i \Psi_i^\top \right) \mathbf{M}^{-1}. \quad (2.23)$$

Similarly to Theorem 2.2.1, the conditional variance $\text{Var}(\xi | \Psi)$ in (2.22) is close to its expectation B_ξ . This allows to state the following deviation bound for $\|\xi\|$.

Theorem 2.4.1. *Let the design Ψ and the noise variance Σ be such that*

$$\|B_\xi^{-1/2} \mathcal{A}_\xi B_\xi^{-1/2} - I_p\|_{\text{op}} \leq \delta_1 \quad (2.24)$$

with $\delta_1 = \delta_1(\mathbf{x})$ on a set $\Omega_1(\mathbf{x})$ with $\mathbb{P}(\Omega_1(\mathbf{x})) \geq 1 - e^{-\mathbf{x}}$. Let also the errors ε_i be conditionally on Ψ Gaussian zero mean with $\sigma_i^2 = \text{Var}(\varepsilon_i | \Psi_i)$. Then it holds for the vector $\xi = \sigma^{-1} \mathbf{M}^{-1} \Psi \varepsilon$

$$\mathbb{P}\left\{ \|\xi\| \geq (1 + \delta_1) z(B_\xi, \mathbf{x}) \right\} \leq 2e^{-\mathbf{x}}, \quad (2.25)$$

see (B.2) for the definition of $z(B, \mathbf{x})$.

Proof. Let (2.24) hold on a set $\Omega_1(\mathbf{x})$ with $\mathbb{P}(\Omega_1(\mathbf{x})) \geq 1 - e^{-\mathbf{x}}$ for some value $\delta_1 = \delta_1(\mathbf{x})$. It helps to bound the moment generating function of ξ . If ε is Gaussian conditioned on Ψ , then, it holds for any $\lambda > 0$ and any vector $\gamma \in \mathbb{R}^p$

$$\log \mathbb{E}\left\{ \exp(\lambda \gamma^\top \xi) | \Psi \right\} = \frac{\lambda^2 \|\mathcal{A}_\xi^{1/2} \gamma\|^2}{2}.$$

This implies by (2.24) on $\Omega_1(\mathbf{x})$

$$\begin{aligned} \log \mathbb{E}\left\{ \exp\left(\lambda \frac{\gamma^\top \xi}{\|B_\xi^{1/2} \gamma\|}\right) \mathbb{I}(\Omega_1(\mathbf{x})) \right\} &\leq \log \mathbb{E} \exp\left\{ \frac{\lambda^2 \|\mathcal{A}_\xi^{1/2} \gamma\|^2}{2 \|B_\xi^{1/2} \gamma\|^2} \mathbb{I}(\Omega_1(\mathbf{x})) \right\} \\ &\leq \frac{(1 + \delta_1)^2 \lambda^2}{2}. \end{aligned}$$

This implies the result by the deviation bound from Theorem B.1.1 for Gaussian quadratic form.

One can combine the expansions from previous section with the bound (2.25). In particular, putting all together yields on the set $\Omega_2(\mathbf{x}) = \Omega(\mathbf{x}) \cup \Omega_1(\mathbf{x})$ with $\mathbb{I}P(\Omega_2(\mathbf{x}) \geq 1 - 3e^{-\mathbf{x}})$

$$\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi}\| \leq \frac{\delta}{1-\delta} \|\boldsymbol{\xi}\| \leq \frac{\delta(1+\delta_1)}{1-\delta} z(B_{\boldsymbol{\xi}}, \mathbf{x})$$

(please check).

2.5 Misspecified linear modeling assumption

The derivations of previous sections explicitly used the linear modeling assumption (2.1). This section discusses what changes if this assumption is not fulfilled. Namely, we only assume that the response variable \mathbf{Y} and the feature vector $\boldsymbol{\Psi}$ are correlated. In this situation, the errors $\boldsymbol{\varepsilon}$ can be defined via conditional expectation $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbb{E}(\mathbf{Y} | \boldsymbol{\Psi})$, and the parameter vector $\boldsymbol{\theta}^*$ can be naturally associated with the canonical correlation coefficients:

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \{\mathbb{E}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)\}^{-1} \mathbb{E}(\boldsymbol{\Psi}\mathbf{Y}). \quad (2.26)$$

It is instructive to check that this definition becomes identity if (2.1) holds. In a slightly different way, one can define $\boldsymbol{\theta}^*$ by projection:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{arginf}} \mathbb{E}\|\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}\|^2.$$

The most important corollary of this definition is that $\boldsymbol{\Psi}$ and $\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}^*$ are orthogonal:

$$\mathbb{E}\{\boldsymbol{\Psi}(\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}^*)\} = \mathbb{E}\{\boldsymbol{\Psi}(\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}^*)\} = 0.$$

Further, the Fisher information matrix $D^2 = -\nabla^2 \mathbb{E}L(\boldsymbol{\theta}^*)$ is the same as in the case of correct specification:

$$D^2 = -\nabla^2 \mathbb{E}L(\boldsymbol{\theta}^*) = \sigma^{-2} \mathbb{E}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)$$

The corresponding score vector

$$\boldsymbol{\xi} = D^{-1} \nabla L(\boldsymbol{\theta}^*) = D^{-1} \{\nabla L(\boldsymbol{\theta}^*) - \nabla \mathbb{E}L(\boldsymbol{\theta}^*)\}$$

can be written as

$$\boldsymbol{\xi} = D^{-1}\sigma^{-2}\{\boldsymbol{\Psi}\mathbf{Y} - \boldsymbol{\Psi}\boldsymbol{\Psi}^\top\boldsymbol{\theta}^* - \mathbb{E}(\boldsymbol{\Psi}\mathbf{Y}) + \mathbb{E}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)\boldsymbol{\theta}^*\}.$$

Define a bias function

$$b(\boldsymbol{\Psi}) \stackrel{\text{def}}{=} \mathbb{E}(\mathbf{Y} | \boldsymbol{\Psi}) - \boldsymbol{\Psi}^\top\boldsymbol{\theta}^* \quad (2.27)$$

which measures the departure from the linear parametric assumption $\mathbb{E}(\mathbf{Y} | \boldsymbol{\Psi}) = \boldsymbol{\Psi}^\top\boldsymbol{\theta}^*$. By definition, the vector of observations \mathbf{Y} can be decomposed as

$$\mathbf{Y} = \boldsymbol{\Psi}^\top\boldsymbol{\theta}^* + b(\boldsymbol{\Psi}) + \boldsymbol{\varepsilon} \quad (2.28)$$

for $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbb{E}(\mathbf{Y} | \boldsymbol{\Psi})$; cf. with the model equation (2.1) under correct model specification. Moreover, the definition of $\boldsymbol{\varepsilon}$ implies $\mathbb{E}(\boldsymbol{\Psi}\boldsymbol{\varepsilon}) = 0$, and the definitions (2.26) and (2.27) imply $\mathbb{E}[\boldsymbol{\Psi}b(\boldsymbol{\Psi})] = 0$. For the vector $\boldsymbol{\xi}$, we obtain the decomposition

$$\boldsymbol{\xi} = \sigma^{-2}D^{-1}\{\boldsymbol{\Psi}b(\boldsymbol{\Psi}) + \boldsymbol{\Psi}\boldsymbol{\varepsilon}\} = \boldsymbol{\xi}_0 + \boldsymbol{\xi}_1, \quad (2.29)$$

where with $\mathbf{M}^2 = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top = \sigma^2D^2$

$$\begin{aligned} \boldsymbol{\xi}_0 &\stackrel{\text{def}}{=} \sigma^{-1}\mathbf{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\varepsilon} = \sigma^{-1}\mathbf{M}^{-1}\boldsymbol{\Psi}\{\mathbf{Y} - \mathbb{E}(\mathbf{Y} | \boldsymbol{\Psi})\}, \\ \boldsymbol{\xi}_1 &\stackrel{\text{def}}{=} \sigma^{-1}\mathbf{M}^{-1}\boldsymbol{\Psi}b(\boldsymbol{\Psi}) = \sigma^{-1}\mathbf{M}^{-1}\boldsymbol{\Psi}\{\mathbb{E}(\mathbf{Y} | \boldsymbol{\Psi}) - \boldsymbol{\Psi}^\top\boldsymbol{\theta}^*\}. \end{aligned} \quad (2.30)$$

The term $\boldsymbol{\xi}_0$ in the decomposition $\boldsymbol{\xi} = \boldsymbol{\xi}_0 + \boldsymbol{\xi}_1$ is identical to the case of correct model specification and it relies to the random noise in the observations \mathbf{Y} . The term $\boldsymbol{\xi}_1$ appears only if the model assumption $\mathbb{E}(\mathbf{Y} | \boldsymbol{\Psi}) = \boldsymbol{\Psi}^\top\boldsymbol{\theta}^*$ is not correct. It only relies to the random design distribution. By construction $\mathbb{E}(\boldsymbol{\xi}_0 | \boldsymbol{\Psi}) = 0$, in particular, $\boldsymbol{\xi}_0$ is zero mean. The same is true for $\boldsymbol{\xi}_1$.

For the LSE $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}\mathbf{Y}$, it holds from (2.28) in the case of a non-degenerated design matrix $\mathbf{M} = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top$

$$\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = (\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}\mathbf{Y} - \boldsymbol{\theta}^* = (\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}\boldsymbol{\varepsilon} + (\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}b(\boldsymbol{\Psi}).$$

so that in view of $D^2 = \sigma^{-2}\mathbf{M}^2$, it holds with $A_\Psi = \mathbf{M}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)^{-1}\mathbf{M}$

$$\begin{aligned} D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) &= \mathbf{M}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)^{-1}\mathbf{M}\boldsymbol{\xi}_0 + \mathbf{M}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)^{-1}\mathbf{M}\boldsymbol{\xi}_1 \\ &= A_\Psi(\boldsymbol{\xi}_0 + \boldsymbol{\xi}_1) = A_\Psi\boldsymbol{\xi}. \end{aligned} \quad (2.31)$$

Comparing with the case of a correct linear assumption reveals that model misspecification yields as additional error term $A_\Psi\boldsymbol{\xi}_1$ related to the bias $b(\boldsymbol{\Psi})$ from (2.27). The origin of this term is that the target $\boldsymbol{\theta}^* = \mathbb{E}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)^{-1}\mathbb{E}(\boldsymbol{\Psi}\mathbf{Y})$ is defined under the design

measure while the construction of the estimate $\tilde{\boldsymbol{\theta}}$ is based on its empirical counterpart. In particular, as $I\!\!E(\boldsymbol{\varepsilon} \mid \boldsymbol{\Psi}) = 0$, it holds

$$\begin{aligned} I\!\!E\left(\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 \mid \boldsymbol{\Psi}\right) &= I\!\!E(\|A_{\boldsymbol{\Psi}} \boldsymbol{\xi}_0\|^2 \mid \boldsymbol{\Psi}) + \|A_{\boldsymbol{\Psi}} \boldsymbol{\xi}_1\|^2 \\ &= A_{\boldsymbol{\Psi}} \text{Var}(\boldsymbol{\xi}_0 \mid \boldsymbol{\Psi}) A_{\boldsymbol{\Psi}} + \sigma^{-2} \|A_{\boldsymbol{\Psi}} \boldsymbol{M}^{-1} \boldsymbol{\Psi} b(\boldsymbol{\Psi})\|^2. \end{aligned}$$

This formula can be viewed as analog of the bias-variance decomposition for linear estimation with random design.

One can check that all the statements of Theorems 2.3.1 and 2.3.2 apply with such defined $\boldsymbol{\theta}^*$ and $\boldsymbol{\xi}$ even under model misspecification.

Theorem 2.5.1. *Consider the model (2.1) and suppose*

$$\|\boldsymbol{M}^{-1} \boldsymbol{\Psi} \boldsymbol{\Psi}^\top \boldsymbol{M}^{-1} - I_p\|_{\text{op}} \leq \delta \quad (2.32)$$

for $\boldsymbol{M}^2 = I\!\!E(\boldsymbol{\Psi} \boldsymbol{\Psi}^\top)$ and some $\delta < 1/2$ on a dominating set $\Omega(\mathbf{x})$. Then the MLE $\tilde{\boldsymbol{\theta}}$ fulfills on $\Omega(\mathbf{x})$ for $\boldsymbol{\xi}$ from (2.29)

$$\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi}\| \leq \frac{\delta}{1-\delta} \|\boldsymbol{\xi}\|, \quad (2.33)$$

$$|2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\boldsymbol{\xi}\|^2| \leq \frac{\delta}{1-\delta} \|\boldsymbol{\xi}\|^2. \quad (2.34)$$

Proof. The bound (2.32) implies $\|I_p - A_{\boldsymbol{\Psi}}\|_{\text{op}} \leq \delta/(1-\delta)$, and (2.33) follows from the decomposition (2.31). Further, for any $\boldsymbol{\theta}$, quadraticity of $L(\boldsymbol{\theta})$ implies

$$2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \sigma^{-2} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \boldsymbol{\Psi} \boldsymbol{\Psi}^\top (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

so that $D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = A_{\boldsymbol{\Psi}} \boldsymbol{\xi}$ yields by definition

$$2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \{D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\}^\top \boldsymbol{M}^{-1} \boldsymbol{\Psi} \boldsymbol{\Psi}^\top \boldsymbol{M}^{-1} \{D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\} = \boldsymbol{\xi}^\top A_{\boldsymbol{\Psi}} \boldsymbol{\xi}.$$

Now (2.34) follows from (2.31).

The result of this theorem should be combined with a bound on the random vector $\boldsymbol{\xi}$. Obviously

$$\|\boldsymbol{\xi}\| \leq \|\boldsymbol{\xi}_0\| + \|\boldsymbol{\xi}_1\| = \|\sigma^{-1} \boldsymbol{M}^{-1} \boldsymbol{\Psi} \boldsymbol{\varepsilon}\| + \|\sigma^{-1} \boldsymbol{M}^{-1} \boldsymbol{\Psi} b(\boldsymbol{\Psi})\|.$$

The term $\boldsymbol{\xi}_0$ from (2.30) can be bounded by Theorem 2.4.1 provided that the conditional variance $\mathcal{A}_{\boldsymbol{\xi}} = \text{Var}(\boldsymbol{\xi} \mid \boldsymbol{\Psi})$ of $\boldsymbol{\xi}$ is close to its unconditional counterpart $B_{\boldsymbol{\xi}} = I\!\!E \mathcal{A}_{\boldsymbol{\xi}}$ and the errors $\boldsymbol{\varepsilon}_i = Y_i - I\!\!E(Y_i \mid \boldsymbol{\Psi}_i)$ are Gaussian or subexponential. The additional term $\boldsymbol{\xi}_1$ depends on the design distribution only and can be bounded by general results on non-Gaussian quadratic forms.

Theorem 2.5.2. Let, given \mathbf{x} , condition (2.24) hold with $\delta_1 = \delta_1(\mathbf{x})$ on a set $\Omega_1(\mathbf{x})$ with $I\!\!P(\Omega_1(\mathbf{x})) \geq 1 - e^{-\mathbf{x}}$. Let also $\varepsilon_i | \Psi_i \sim \mathcal{N}(0, \sigma_i^2)$ and the matrix B_ξ is given by (2.23). Let, for the vector $b(\Psi)$, the matrix B_b be defined by

$$B_b \stackrel{\text{def}}{=} \text{Var}\{\mathbf{M}^{-1}\Psi b(\Psi)\} = \mathbf{M}^{-1}E\left\{\sum_{i=1}^n b_i^2 \Psi_i \Psi_i^\top\right\} \mathbf{M}^{-1}.$$

Then on a random set of probability $1 - 3e^{-\mathbf{x}}$

$$\|\xi\| \leq (1 + \delta_1) z(B_\xi, \mathbf{x}) + \sigma^{-1} z(B_b, \mathbf{x}).$$

In particular, if $\|b(\Psi)\|_\infty \leq b_\infty$, then

$$\|\xi\| \leq (1 + \delta_1) z(B_\xi, \mathbf{x}) + \sigma^{-1} b_\infty z(p, \mathbf{x}).$$

Proof. The bound for ξ_0 is already proved in Theorem 2.4.1. It remains to show that

$$I\!\!P(\|\mathbf{M}^{-1}\Psi b(\Psi)\| \geq z(B_b, \mathbf{x})) \leq 2e^{-\mathbf{x}}. \quad (2.35)$$

We already know that $E\{\mathbf{M}^{-1}\Psi b(\Psi)\} = 0$. For applying the general bounds from Section B.1 on the norm of a Gaussian random vector, we only need to evaluate the characteristics of its covariance matrix

$$\text{Var}\{\mathbf{M}^{-1}\Psi b(\Psi)\} = \mathbf{M}^{-1}E\{\Psi b(\Psi) b(\Psi)^\top \Psi^\top\} \mathbf{M}^{-1}.$$

Now the result (2.35) follows by Theorem B.1.1.

Under the constraint $\|b(\Psi)\|_\infty \leq b_\infty$

$$\text{Var}\{\mathbf{M}^{-1}\Psi b(\Psi)\} \leq b_\infty^2 \mathbf{M}^{-1}E(\Psi \Psi^\top) \mathbf{M}^{-1} = b_\infty^2 I_p,$$

and by Theorem B.2.1

$$\|\mathbf{M}^{-1}\Psi b(\Psi)\| \leq b_\infty z(p, \mathbf{x}) \leq b_\infty (\sqrt{p} + \sqrt{2\mathbf{x}}).$$

2.6 Application to instrumental regression

Observed: a sample from (Y, X, W) . Model

$$Y = f(X) + U, \quad E[U | W] = 0.$$

where Y , an explained variable, X , an explanatory variable, W , an instrument. The target is the regression function $f(\cdot)$.

Let $\psi_1(x), \dots, \psi_j(x), \dots$ be a functional basis. Consider a finite approximation

$$f(x) = \theta_1\psi_1(x) + \dots + \theta_p\psi_p(x)$$

or in vector form

$$f(x) = \boldsymbol{\psi}(x)^\top \boldsymbol{\theta}$$

with $\boldsymbol{\psi}(x) = (\psi_1(x), \dots, \psi_p(x))^\top \in \mathbb{R}^p$ and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top \in \mathbb{R}^p$. This leads to an approximating model

$$\mathbf{Y} = \boldsymbol{\psi}(X)^\top \boldsymbol{\theta}^* + U, \quad \mathbb{E}[U | W] = 0.$$

The constraint $\mathbb{E}[U | W] = 0$ means that for any function $\phi(W)$

$$\mathbb{E}[Y\phi(W)] = \mathbb{E}[\phi(W)\boldsymbol{\psi}(X)^\top]\boldsymbol{\theta}^*.$$

We apply a *discretization* or *finite dimensional approximation*: for a finite collection of functions $\boldsymbol{\phi}(w) = (\phi_1(w), \dots, \phi_q(w))^\top$, it holds

$$\mathbb{E}[Y\phi(W)] = \mathbb{E}[\boldsymbol{\psi}(X)\boldsymbol{\phi}(W)^\top]^\top \boldsymbol{\theta}^* = \mathbf{T}^\top \boldsymbol{\theta}^*$$

with

$$\mathbf{T} = \mathbb{E}[\boldsymbol{\psi}(X)\boldsymbol{\phi}(W)^\top] \in \mathbb{R}^{p \times q}.$$

Define

$$\begin{aligned} \mathbf{Z} &= \mathbb{E}_n[\mathbf{Y}\boldsymbol{\phi}(W)] = n^{-1} \sum_i Y_i \boldsymbol{\phi}(W_i) \in \mathbb{R}^q, \\ \mathbf{T}_n &= \mathbb{E}_n[\boldsymbol{\psi}(X)\boldsymbol{\phi}(W)^\top] = n^{-1} \sum_i \boldsymbol{\psi}(X_i)\boldsymbol{\phi}(W_i)^\top \in \mathbb{R}^{q \times p}, \\ \boldsymbol{\varepsilon} &= \mathbb{E}_n[\boldsymbol{\phi}(W)\mathbf{U}] = n^{-1} \sum_i \boldsymbol{\phi}(W_i)\mathbf{U}_i \in \mathbb{R}^q. \end{aligned}$$

The original problems reduces to

$$\mathbf{Z} = \mathbf{T}^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is the error q -vector, $\mathbf{T} = \mathbb{E}[\boldsymbol{\psi}(X)\boldsymbol{\phi}(W)^\top]$ is an unknown $p \times q$ matrix and only its empirical counterpart \mathbf{T}_n is available. In such cases one speaks of *an inverse problem with error in operator*. The main problem for the analysis in this model is that \mathbf{T}_n is random and correlated with \mathbf{Z} and $\boldsymbol{\varepsilon}$. The goal is to build an estimator $\tilde{\boldsymbol{\theta}}$ of the vector $\boldsymbol{\theta}^*$ leading to the estimator $\tilde{f}(x) = \boldsymbol{\psi}(x)^\top \tilde{\boldsymbol{\theta}}$ of the response.

The natural plug-in approach suggests to replace the unknown operator \mathbf{T} by its empirical counterpart \mathbf{T}_n leading to the approximating linear model

$$\mathbf{Z} = \mathbf{T}_n^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}.$$

with the random design $\mathbf{T}_n = n^{-1} \sum_i \psi(X_i) \phi(W_i)^\top$ so that the setup (2.11) applies. The corresponding least square estimator of $\boldsymbol{\theta}^*$ reads as

$$\tilde{\boldsymbol{\theta}} = (\mathbf{T}_n \mathbf{T}_n^\top)^{-1} \mathbf{T}_n \mathbf{Z}.$$

The results of Theorem 2.2.2 justify that the random matrix $\mathbf{T}_n \mathbf{T}_n^\top$ is very close to the product $\mathbb{E}(\mathbf{T}_n) \mathbb{E}(\mathbf{T}_n^\top) = \mathbf{T} \mathbf{T}^\top$ and the theoretical study of the properties of the estimator $\tilde{\boldsymbol{\theta}}$ can be done with $\mathbf{T} \mathbf{T}^\top$ in place of $\mathbf{T}_n \mathbf{T}_n^\top$ in (17.8). Similarly one can justify that the product $\mathbf{T}_n \mathbf{Z}$ behaves nearly as $\mathbf{T} \mathbf{Z}$.

Below we assume for simplicity that all triples (Y_i, X_i, W_i) are i.i.d. so that $T_i = \psi(X_i) \phi(W_i)^\top$ are also i.i.d. Define $\mathbf{M}^2 = \mathbf{T} \mathbf{T}^\top$ and

$$\sigma_1^2 = \|\mathbf{M}^{-1} \mathbb{E}(T_i T_i^\top) \mathbf{M}^{-1} - I_p\|_{\text{op}}$$

and suppose that it holds almost surely

$$\|\mathbf{M}^{-1} (T_i - \mathbf{T})\|_{\text{op}} \leq u.$$

Now Theorem 2.2.2 implies for any $z > 0$

$$\mathbb{P}\left(\sqrt{n} \|\mathbf{M}^{-1} (\mathbf{T}_n - \mathbf{T})\|_{\text{op}} > z\right) \leq 2(p+q) \exp\left\{-\frac{z^2}{2\sigma_1^2 + 2uz/(3n^{1/2})}\right\}.$$

Moreover, if $\|\mathbf{M}^{-1} (\mathbf{T}_n - \mathbf{T})\|_{\text{op}} \leq \delta$, then

$$\|\mathbf{M}^{-1} \mathbf{T}_n \mathbf{T}_n^\top \mathbf{M}^{-1} - I_p\|_{\text{op}} \leq \delta^2 + 2\delta.$$

Linear smoothers

Here we discuss the important situation when the number of predictors ψ_j and hence the number of parameters p in the linear model $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ is not small relative to the sample size. Then the least square or the maximum likelihood approach meets serious problems. The first one relates to the numerical issues. The definition of the LSE $\tilde{\boldsymbol{\theta}}$ involves the inversion of the $p \times p$ matrix $\Psi\Psi^\top$ and such an inversion becomes a delicate task for p large. The other problem concerns the inference for the estimated parameter $\boldsymbol{\theta}^*$. The risk bound and the width of the confidence set are proportional to the parameter dimension p and thus, with large p , the inference statements become almost uninformative. In particular, if p is of order the sample size n , even consistency is not achievable. One faces a really critical situation. We already know that the MLE is the efficient estimate in the class of all unbiased estimates. At the same time it is highly inefficient in overparametrized models. The only way out of this situation is to sacrifice the unbiasedness property in favor of reducing the model complexity: some procedures can be more efficient than MLE even if they are biased. This section discusses one way of resolving these problems by regularization or shrinkage. To be more specific, for the rest of the section we consider the following setup. The observed vector \mathbf{Y} follows the model

$$\mathbf{Y} = \mathbf{f}^* + \boldsymbol{\varepsilon} \quad (3.1)$$

with a homogeneous error vector $\boldsymbol{\varepsilon}$: $I\!\!E\boldsymbol{\varepsilon} = 0$, $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$. Noise misspecification is not considered in this section.

Furthermore, we assume a basis or a collection of basis vectors ψ_1, \dots, ψ_p is given with p large. This allows for approximating the response vector $\mathbf{f} = I\!\!E\mathbf{Y}$ in the form $\mathbf{f} = \Psi^\top \boldsymbol{\theta}^*$, or, equivalently,

$$\mathbf{f} = \theta_1^* \psi_1 + \dots + \theta_p^* \psi_p.$$

In many cases we will assume that the basis is already orthogonalized: $\Psi\Psi^\top = I_p$. The model (3.1) can be rewritten as

$$\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}, \quad \text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n.$$

The MLE or ordinary LSE of the parameter vector $\boldsymbol{\theta}^*$ for this model reads as

$$\tilde{\boldsymbol{\theta}} = (\Psi\Psi^\top)^{-1}\Psi\mathbf{Y}, \quad \tilde{\mathbf{f}} = \Psi^\top \tilde{\boldsymbol{\theta}} = \Psi^\top (\Psi\Psi^\top)^{-1}\Psi\mathbf{Y}.$$

If the matrix $\Psi\Psi^\top$ is degenerate or badly posed, computing the MLE $\tilde{\boldsymbol{\theta}}$ is a hard task. Below we discuss how this problem can be treated.

3.1 Regularization and ridge regression

Let R be a positive symmetric $p \times p$ matrix. Then the sum $\Psi\Psi^\top + R$ is positive symmetric as well and can be inverted whatever the matrix Ψ is. This suggests to replace $(\Psi\Psi^\top)^{-1}$ by $(\Psi\Psi^\top + R)^{-1}$ leading to the regularized least squares estimate $\tilde{\boldsymbol{\theta}}_R$ of the parameter vector $\boldsymbol{\theta}$ and the corresponding response estimate $\tilde{\mathbf{f}}_R$:

$$\tilde{\boldsymbol{\theta}}_R \stackrel{\text{def}}{=} (\Psi\Psi^\top + R)^{-1}\Psi\mathbf{Y}, \quad \tilde{\mathbf{f}}_R \stackrel{\text{def}}{=} \Psi^\top (\Psi\Psi^\top + R)^{-1}\Psi\mathbf{Y}. \quad (3.2)$$

Such a method is also called *ridge regression*. An example of choosing R is the multiple of the unit matrix: $R = \alpha I_p$ where $\alpha > 0$ and I_p stands for the unit matrix. This method is also called *Tikhonov regularization* and it results in the parameter estimate $\tilde{\boldsymbol{\theta}}_\alpha$ and the response estimate $\tilde{\mathbf{f}}_\alpha$:

$$\tilde{\boldsymbol{\theta}}_\alpha \stackrel{\text{def}}{=} (\Psi\Psi^\top + \alpha I_p)^{-1}\Psi\mathbf{Y}, \quad \tilde{\mathbf{f}}_\alpha \stackrel{\text{def}}{=} \Psi^\top (\Psi\Psi^\top + \alpha I_p)^{-1}\Psi\mathbf{Y}. \quad (3.3)$$

A proper choice of the matrix R for the ridge regression method (3.2) or the parameter α for the Tikhonov regularization (3.3) is an important issue. Below we discuss several approaches which lead to the estimate (3.2) with a specific choice of the matrix R . The properties of the estimates $\tilde{\boldsymbol{\theta}}_R$ and $\tilde{\mathbf{f}}_R$ will be studied in context of penalized likelihood estimation in the next section.

3.2 Penalized likelihood. Bias and variance

The estimate (3.2) can be obtained in a natural way within the (quasi) ML approach using the penalized least squares. The classical unpenalized method is based on minimizing the sum of residuals squared:

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} L(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{arginf}} \|\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}\|^2$$

with $L(\boldsymbol{\theta}) = \sigma^{-2} \|\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}\|^2 / 2$. (Here we omit the terms which do not depend on $\boldsymbol{\theta}$.) Now we introduce an additional penalty on the objective function which penalizes for the complexity of the candidate vector $\boldsymbol{\theta}$ which is expressed by the value $\|G\boldsymbol{\theta}\|^2 / 2$ for a given symmetric matrix G . This choice of complexity measure implicitly assumes that the vector $\boldsymbol{\theta} \equiv 0$ has the smallest complexity equal to zero and this complexity increases with the norm of $G\boldsymbol{\theta}$. Define the *penalized log-likelihood*

$$\begin{aligned} L_G(\boldsymbol{\theta}) &\stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2 / 2 \\ &= -(2\sigma^2)^{-1} \|\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}\|^2 - \|G\boldsymbol{\theta}\|^2 / 2 - (n/2) \log(2\pi\sigma^2). \end{aligned} \quad (3.4)$$

The penalized MLE reads as

$$\tilde{\boldsymbol{\theta}}_G = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} L_G(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \{(2\sigma^2)^{-1} \|\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}\|^2 + \|G\boldsymbol{\theta}\|^2 / 2\}.$$

A straightforward calculus leads to the expression (3.2) for $\tilde{\boldsymbol{\theta}}_G$ with $R = \sigma^2 G^2$:

$$\tilde{\boldsymbol{\theta}}_G \stackrel{\text{def}}{=} (\boldsymbol{\Psi} \boldsymbol{\Psi}^\top + \sigma^2 G^2)^{-1} \boldsymbol{\Psi} \mathbf{Y}. \quad (3.5)$$

We see that $\tilde{\boldsymbol{\theta}}_G$ is again a linear estimate: $\tilde{\boldsymbol{\theta}}_G = \mathcal{S}_G \mathbf{Y}$ with $\mathcal{S}_G = (\boldsymbol{\Psi} \boldsymbol{\Psi}^\top + \sigma^2 G^2)^{-1} \boldsymbol{\Psi}$. The results of Section 1.3 explains that $\tilde{\boldsymbol{\theta}}_G$ in fact estimates the value $\boldsymbol{\theta}_G$ defined by

$$\begin{aligned} \boldsymbol{\theta}_G &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \mathbb{E} L_G(\boldsymbol{\theta}) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{arginf}} \mathbb{E} \{\|\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}\|^2 + \sigma^2 \|G\boldsymbol{\theta}\|^2\} \\ &= (\boldsymbol{\Psi} \boldsymbol{\Psi}^\top + \sigma^2 G^2)^{-1} \boldsymbol{\Psi} \mathbf{f}^* = \mathcal{S}_G \mathbf{f}^*. \end{aligned} \quad (3.6)$$

In particular, if $\mathbf{f}^* = \boldsymbol{\Psi}^\top \boldsymbol{\theta}^*$, then

$$\boldsymbol{\theta}_G = (\boldsymbol{\Psi} \boldsymbol{\Psi}^\top + \sigma^2 G^2)^{-1} \boldsymbol{\Psi} \boldsymbol{\Psi}^\top \boldsymbol{\theta}^* \quad (3.7)$$

and $\boldsymbol{\theta}_G \neq \boldsymbol{\theta}^*$ unless $G = 0$. In other words, the penalized MLE $\tilde{\boldsymbol{\theta}}_G$ is biased.

Exercise 3.2.1. Check that $\mathbb{E} \tilde{\boldsymbol{\theta}}_\alpha = \boldsymbol{\theta}_\alpha$ for $\boldsymbol{\theta}_\alpha = (\boldsymbol{\Psi} \boldsymbol{\Psi}^\top + \alpha I_p)^{-1} \boldsymbol{\Psi} \boldsymbol{\Psi}^\top \boldsymbol{\theta}^*$, the bias $\|\boldsymbol{\theta}_\alpha - \boldsymbol{\theta}^*\|$ grows with the regularization parameter α .

The penalized MLE $\tilde{\boldsymbol{\theta}}_G$ leads to the response estimate $\tilde{\mathbf{f}}_G = \boldsymbol{\Psi}^\top \tilde{\boldsymbol{\theta}}_G$.

Exercise 3.2.2. Check that the penalized ML approach leads to the response estimate

$$\tilde{\mathbf{f}}_G = \boldsymbol{\Psi}^\top \tilde{\boldsymbol{\theta}}_G = \boldsymbol{\Psi}^\top (\boldsymbol{\Psi} \boldsymbol{\Psi}^\top + \sigma^2 G^2)^{-1} \boldsymbol{\Psi} \mathbf{Y} = \Pi_G \mathbf{Y}$$

with $\Pi_G = \Psi^\top (\Psi\Psi^\top + \sigma^2 G^2)^{-1} \Psi$. Show that Π_G is a sub-projector in the sense that $\|\Pi_G \mathbf{u}\| \leq \|\mathbf{u}\|$ for any $\mathbf{u} \in \mathbb{R}^n$.

Exercise 3.2.3. Let Ψ be orthonormal: $\Psi\Psi^\top = I_p$. Then the penalized MLE $\tilde{\boldsymbol{\theta}}_G$ can be represented as

$$\tilde{\boldsymbol{\theta}}_G = (I_p + \sigma^2 G^2)^{-1} \mathbf{Z},$$

where $\mathbf{Z} = \Psi \mathbf{Y}$ is the vector of empirical Fourier coefficients. Specify the result for the case of a diagonal matrix $G = \text{diag}(g_1, \dots, g_p)$ and describe the corresponding response estimate $\tilde{\mathbf{f}}_G$.

The previous results indicate that introducing the penalization leads to some bias of estimation. One can ask about a benefit of using a penalized procedure. The next result shows that penalization decreases the variance of estimation and thus, makes the procedure more stable.

Theorem 3.2.1. Let $\tilde{\boldsymbol{\theta}}_G$ be a penalized MLE from (3.5). Then $E\tilde{\boldsymbol{\theta}}_G = \boldsymbol{\theta}_G$, see (3.7), and under noise homogeneity $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$, it holds

$$\begin{aligned} \text{Var}(\tilde{\boldsymbol{\theta}}_G) &= (\sigma^{-2} \Psi \Psi^\top + G^2)^{-1} \sigma^{-2} \Psi \Psi^\top (\sigma^{-2} \Psi \Psi^\top + G^2)^{-1} \\ &= D_G^{-2} D^2 D_G^{-2} \end{aligned}$$

with $D_G^2 = \sigma^{-2} \Psi \Psi^\top + G^2$. In particular, $\text{Var}(\tilde{\boldsymbol{\theta}}_G) \leq D_G^{-2}$. If $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$, then $\tilde{\boldsymbol{\theta}}_G$ is also normal: $\tilde{\boldsymbol{\theta}}_G \sim \mathcal{N}(\boldsymbol{\theta}_G, D_G^{-2} D^2 D_G^{-2})$.

Moreover, the bias $\|\boldsymbol{\theta}_G - \boldsymbol{\theta}^*\|$ monotonously increases in G^2 while the variance monotonously decreases with the penalization G .

Proof. The first two moments of $\tilde{\boldsymbol{\theta}}_G$ are computed from $\tilde{\boldsymbol{\theta}}_G = S_G \mathbf{Y}$. Monotonicity of the bias and variance of $\tilde{\boldsymbol{\theta}}_G$ is proved below in Exercise 3.2.6.

Exercise 3.2.4. Let Ψ be orthonormal: $\Psi\Psi^\top = I_p$. Describe $\text{Var}(\tilde{\boldsymbol{\theta}}_G)$. Show that the variance decreases with the penalization G in the sense that $G_1 \geq G$ for two commutative matrices G and G_1 implies $\text{Var}(\tilde{\boldsymbol{\theta}}_{G_1}) \leq \text{Var}(\tilde{\boldsymbol{\theta}}_G)$.

Exercise 3.2.5. Let $\Psi\Psi^\top = I_p$ and let $G = \text{diag}(g_1, \dots, g_p)$ be a diagonal matrix. Compute the squared bias $\|\boldsymbol{\theta}_G - \boldsymbol{\theta}^*\|^2$ and show that it monotonously increases in each g_j for $j = 1, \dots, p$.

Exercise 3.2.6. Let G be a symmetric matrix and $\tilde{\boldsymbol{\theta}}_G$ the corresponding penalized MLE. Show that the variance $\text{Var}(\tilde{\boldsymbol{\theta}}_G)$ decreases while the bias $\|\boldsymbol{\theta}_G - \boldsymbol{\theta}^*\|$ increases in G^2 .

Hint: with $D^2 = \sigma^{-2}\Psi\Psi^\top$, show that for any vector $\mathbf{w} \in \mathbb{R}^p$ and $\mathbf{u} = D^{-1}\mathbf{w}$, it holds

$$\mathbf{w}^\top \text{Var}(\tilde{\boldsymbol{\theta}}_G)\mathbf{w} = \mathbf{u}^\top (I_p + D^{-1}G^2D^{-1})^{-2}\mathbf{u}$$

and this value decreases with G^2 because $I_p + D^{-1}G^2D^{-1}$ increases. Show in a similar way that

$$\|\boldsymbol{\theta}_G - \boldsymbol{\theta}^*\|^2 = \|(D^2 + G^2)^{-1}G^2\boldsymbol{\theta}^*\|^2 = \boldsymbol{\theta}^{*\top} \Gamma^{-1} \boldsymbol{\theta}^*$$

with $\Gamma = (I_p + G^{-2}D^2)(I_p + D^2G^{-2})$. Show that the matrix Γ monotonously increases and thus Γ^{-1} monotonously decreases as a function of the symmetric matrix $B = G^{-2}$.

Putting together the results about the bias and the variance of $\tilde{\boldsymbol{\theta}}_G$ yields the statement about the quadratic risk.

Theorem 3.2.2. *Assume the model $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$. Then the estimate $\tilde{\boldsymbol{\theta}}_G$ fulfills*

$$\mathbb{E}\|\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*\|^2 = \|\boldsymbol{\theta}_G - \boldsymbol{\theta}^*\|^2 + \text{tr}(D_G^{-2}D^2D_G^{-2}).$$

This result is called the *bias-variance decomposition*. The choice of a proper regularization is usually based on this decomposition: one selects a regularization from a given class to provide the minimal possible risk. This approach is referred to as *bias-variance trade-off*.

3.3 Inference for the penalized MLE

Here we discuss some properties of the penalized MLE $\tilde{\boldsymbol{\theta}}_G$. In particular, we focus on the construction of confidence and concentration sets based on the penalized log-likelihood. We know that the regularized estimate $\tilde{\boldsymbol{\theta}}_G$ is the empirical counterpart of the value $\boldsymbol{\theta}_G$ which solves the regularized deterministic problem (3.6). We also know that the key results are expressed via the value of the supremum $\sup_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}) - L_G(\boldsymbol{\theta}_G)$. The next result extends Theorem 1.4.3 to the penalized likelihood.

Theorem 3.3.1. *Let $L_G(\boldsymbol{\theta})$ be the penalized log-likelihood from (3.4). Then*

$$2L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G) = (\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G)^\top D_G^2(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G) \tag{3.8}$$

$$= \sigma^{-2} \boldsymbol{\varepsilon}^\top \Pi_G \boldsymbol{\varepsilon} \tag{3.9}$$

with $\Pi_G = \Psi^\top (\Psi\Psi^\top + \sigma^2 G^2)^{-1} \Psi$.

In general the matrix Π_G is not a projector and hence, $\sigma^{-2}\boldsymbol{\varepsilon}^\top \Pi_G \boldsymbol{\varepsilon}$ is not χ^2 -distributed, the chi-squared result does not apply.

Exercise 3.3.1. Prove (3.8).

Hint: apply the Taylor expansion to $L_G(\boldsymbol{\theta})$ at $\tilde{\boldsymbol{\theta}}_G$. Use that $\nabla L_G(\tilde{\boldsymbol{\theta}}_G) = 0$ and $-\nabla^2 L_G(\boldsymbol{\theta}) \equiv \sigma^{-2}\Psi\Psi^\top + G^2$.

Exercise 3.3.2. Prove (3.9).

Hint: show that $\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G = \mathcal{S}_G \boldsymbol{\varepsilon}$ with $\mathcal{S}_G = (\Psi\Psi^\top + \sigma^2 G^2)^{-1}\Psi$.

The straightforward corollaries of Theorem 3.3.1 are the concentration and confidence probabilities. Define the confidence set $\mathcal{E}_G(\mathfrak{z})$ for $\boldsymbol{\theta}_G$ as

$$\mathcal{E}_G(\mathfrak{z}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} : L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}) \leq \mathfrak{z}\}.$$

The definition implies the following result for the coverage probability:

$$\mathbb{P}(\mathcal{E}_G(\mathfrak{z}) \not\ni \boldsymbol{\theta}_G) \leq \mathbb{P}(L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G) > \mathfrak{z}).$$

Now the representation (3.9) for $L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G)$ reduces the problem to a deviation bound for a quadratic form. We apply the general result of Theorem B.1.1 in Section B.

Theorem 3.3.2. Let $L_G(\boldsymbol{\theta})$ be the penalized log-likelihood from (3.4) and let $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Then it holds with $p_G = \text{tr}(\Pi_G)$ and $v_G^2 = \text{tr}(\Pi_G^2)$ that

$$\mathbb{P}(2L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G) > p_G + 2v_G x^{1/2} + 2x) \leq \exp(-x).$$

Similarly one can state the concentration result. With $D_G^2 = \sigma^{-2}\Psi\Psi^\top + G^2$

$$2L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G) = \|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G)\|^2$$

and the result of Theorem 3.3.2 can be restated as the concentration bound:

$$\mathbb{P}(\|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G)\|^2 > p_G + 2v_G x^{1/2} + 2x) \leq \exp(-x).$$

In other words, $\tilde{\boldsymbol{\theta}}_G$ concentrates on the set $\mathcal{A}(\mathfrak{z}, \boldsymbol{\theta}_G) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_G\|^2 \leq 2\mathfrak{z}\}$ for $2\mathfrak{z} > p_G$.

3.4 Projection and shrinkage estimates

Consider a linear model $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ in which the matrix Ψ is orthonormal in the sense $\Psi\Psi^\top = I_p$. Then the multiplication with Ψ maps this model in the sequence space model $\mathbf{Z} = \boldsymbol{\theta}^* + \boldsymbol{\xi}$, where $\mathbf{Z} = \Psi\mathbf{Y} = (z_1, \dots, z_p)^\top$ is the vector of empirical Fourier

coefficients $z_j = \boldsymbol{\psi}_j^\top \mathbf{Y}$. The noise $\boldsymbol{\xi} = \Psi \boldsymbol{\varepsilon}$ borrows the feature of the original noise $\boldsymbol{\varepsilon}$: if $\boldsymbol{\varepsilon}$ is zero mean and homogeneous, the same applies to $\boldsymbol{\xi}$. The number of coefficients p can be large or even infinite. To get a sensible estimate, one has to apply some regularization method. The simplest one is called *projection*: one just considers the first m empirical coefficients z_1, \dots, z_m and drop the others. The corresponding parameter estimate $\tilde{\boldsymbol{\theta}}_m$ reads as

$$\tilde{\theta}_{m,j} = \begin{cases} z_j & \text{if } j \leq m, \\ 0 & \text{otherwise.} \end{cases}$$

The response vector $\mathbf{f}^* = I\mathbb{E}\mathbf{Y}$ is estimated by $\Psi^\top \tilde{\boldsymbol{\theta}}_m$ leading to the representation

$$\tilde{\mathbf{f}}_m = z_1 \boldsymbol{\psi}_1 + \dots + z_m \boldsymbol{\psi}_m$$

with $z_j = \boldsymbol{\psi}_j^\top \mathbf{Y}$. A disadvantage of the projection method is that it either keeps each empirical coefficient z_m or completely discards it. An extension of the projection method is called *shrinkage*: one multiplies every empirical coefficient z_j with a factor $\alpha_j \in (0, 1)$. This leads to the *shrinkage* estimate $\tilde{\boldsymbol{\theta}}_\alpha$ with

$$\tilde{\theta}_{\alpha,j} = \alpha_j z_j.$$

Here $\boldsymbol{\alpha}$ stands for the vector of coefficients α_j for $j = 1, \dots, p$. A projection method is a special case of this shrinkage with α_j equal to one or zero. Another popular choice of the coefficients α_j is given by

$$\alpha_j = (1 - j/m)^\beta \mathbf{1}(j \leq m) \tag{3.10}$$

for some $\beta > 0$ and $m \leq p$. This choice ensures that the coefficients α_j smoothly approach zero as j approach the value m , and α_j vanish for $j > m$. In this case, the vector $\boldsymbol{\alpha}$ is completely specified by two parameters m and β . The projection method corresponds to $\beta = 0$. The design orthogonality $\Psi \Psi^\top = I_p$ yields again that the estimation risk $I\mathbb{E} \|\tilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}^*\|^2$ coincides with the prediction risk $I\mathbb{E} \|\tilde{\mathbf{f}}_\alpha - \mathbf{f}^*\|^2$.

Exercise 3.4.1. Let $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_p$. The risk $\mathcal{R}(\tilde{\mathbf{f}}_\alpha)$ of the shrinkage estimate $\tilde{\mathbf{f}}_\alpha$ fulfills

$$\mathcal{R}(\tilde{\mathbf{f}}_\alpha) \stackrel{\text{def}}{=} I\mathbb{E} \|\tilde{\mathbf{f}}_\alpha - \mathbf{f}^*\|^2 = \sum_{j=1}^p \theta_j^{*2} (1 - \alpha_j)^2 + \sum_{j=1}^p \alpha_j^2 \sigma^2.$$

Specify the cases of $\boldsymbol{\alpha} = \boldsymbol{\alpha}(m, \beta)$ from (3.10). Evaluate the variance term $\sum_j \alpha_j^2 \sigma^2$.

Hint: approximate the sum over j by the integral $\int (1 - x/m)_+^{2\beta} dx$.

The oracle choice is again defined by risk minimization:

$$\boldsymbol{\alpha}^* \stackrel{\text{def}}{=} \operatorname{argmin}_{\boldsymbol{\alpha}} \mathcal{R}(\tilde{\mathbf{f}}_{\boldsymbol{\alpha}}),$$

where minimization is taken over the class of all considered coefficient vectors $\boldsymbol{\alpha}$.

One way of obtaining a shrinkage estimate in the sequence space model $\mathbf{Z} = \boldsymbol{\theta}^* + \boldsymbol{\xi}$ is by using a roughness penalization. Let G be a symmetric matrix. Consider the regularized estimate $\tilde{\boldsymbol{\theta}}_G$ from (3.2). The next result claims that if G is a diagonal matrix, then $\tilde{\boldsymbol{\theta}}_G$ is a shrinkage estimate. Moreover, a general penalized MLE can be represented as shrinkage by an orthogonal basis transformation.

Theorem 3.4.1. *Let G be a diagonal matrix, $G = \operatorname{diag}(g_1, \dots, g_p)$. The penalized MLE $\tilde{\boldsymbol{\theta}}_G$ in the sequence space model $\mathbf{Z} = \boldsymbol{\theta}^* + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 I_p)$ coincides with the shrinkage estimate $\tilde{\boldsymbol{\theta}}_{\boldsymbol{\alpha}}$ for $\alpha_j = (1 + \sigma^2 g_j^2)^{-1} \leq 1$. Moreover, a penalized MLE $\tilde{\boldsymbol{\theta}}_G$ for a general matrix G can be reduced to a shrinkage estimate by a basis transformation in the sequence space model.*

Proof. The first statement for a diagonal matrix G follows from the representation $\tilde{\boldsymbol{\theta}}_G = (I_p + \sigma^2 G^2)^{-1} \mathbf{Z}$. Next, let U be an orthogonal transform leading to the diagonal representation $G^2 = U^\top D^2 U$ with $D^2 = \operatorname{diag}(g_1, \dots, g_p)$. Then

$$U \tilde{\boldsymbol{\theta}}_G = (I_p + \sigma^2 D^2)^{-1} U \mathbf{Z}$$

that is, $U \tilde{\boldsymbol{\theta}}_G$ is a shrinkage estimate in the transformed model $U \mathbf{Z} = U \boldsymbol{\theta}^* + U \boldsymbol{\xi}$.

In other words, roughness penalization results in some kind of shrinkage. Interestingly, the inverse statement holds as well.

Exercise 3.4.2. Let $\tilde{\boldsymbol{\theta}}_{\boldsymbol{\alpha}}$ is a shrinkage estimate for a vector $\boldsymbol{\alpha} = (\alpha_j)$. Then there is a diagonal penalty matrix G such that $\tilde{\boldsymbol{\theta}}_{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\theta}}_G$.

Hint: define the j th diagonal entry g_j by the equation $\alpha_j = (1 + \sigma^2 g_j^2)^{-1}$.

3.5 Smoothness constraints and roughness penalty approach

Another way of reducing the complexity of the estimation procedure is based on smoothness constraints. The notion of smoothness originates from regression estimation. A non-linear regression function f is expanded using a Fourier or some other functional basis and $\boldsymbol{\theta}^*$ is the corresponding vector of coefficients. Smoothness properties of the regression function imply certain rate of decay of the corresponding Fourier coefficients: the larger frequency is, the fewer amount of information about the regression function is

contained in the related coefficient. This leads to the natural idea to replace the original optimization problem over the whole parameter space with the constrained optimization over a subset of “smooth” parameter vectors. Here we consider one popular example of Sobolev smoothness constraints which effectively means that the s th derivative of the function \mathbf{f}^* has a bounded L_2 -norm. A general Sobolev ball can be defined using a diagonal matrix G :

$$\mathcal{B}_G(R) \stackrel{\text{def}}{=} \|G\boldsymbol{\theta}\| \leq R.$$

Now we consider a constrained ML problem:

$$\tilde{\boldsymbol{\theta}}_{G,R} = \underset{\boldsymbol{\theta} \in \mathcal{B}_G(R)}{\operatorname{argmax}} L(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta} \in \Theta: \|G\boldsymbol{\theta}\| \leq R}{\operatorname{argmin}} \|\mathbf{Y} - \Psi^\top \boldsymbol{\theta}\|^2. \quad (3.11)$$

The Lagrange multiplier method leads to an unconstrained problem

$$\tilde{\boldsymbol{\theta}}_{G,\lambda} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ \|\mathbf{Y} - \Psi^\top \boldsymbol{\theta}\|^2 + \lambda \|G\boldsymbol{\theta}\|^2 \right\}.$$

A proper choice of λ ensures that the solution $\tilde{\boldsymbol{\theta}}_{G,\lambda}$ belongs to $\mathcal{B}_G(R)$ and solves also the problem (3.11). So, the approach based on a Sobolev smoothness assumption, leads back to regularization and shrinkage.

3.6 Shrinkage in a linear inverse problem

This section extends the previous approaches to the situation with indirect observations. More precisely, we focus on the model

$$\mathbf{Y} = A\mathbf{f}^* + \boldsymbol{\varepsilon},$$

where A is a given linear operator (matrix) and \mathbf{f}^* is the target of analysis. With the obvious change of notation this problem can be put back in the general linear setup $\mathbf{Y} = \Psi^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon}$. The special focus is due to the facts that the target can be high dimensional or even functional and that the product $A^\top A$ is usually badly posed and its inversion is a hard task. Below we consider separately the cases when the spectral representation for this problem is available and the general case.

3.7 Spectral cut-off and spectral penalization. Diagonal estimates

Suppose that the eigenvectors of the matrix $A^\top A$ are available. This allows for reducing the model to the spectral representation by an orthogonal change of the coordinate

system: $\mathbf{Z} = \Lambda \mathbf{u} + \Lambda^{1/2} \boldsymbol{\xi}$ with a diagonal matrix $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_p\}$ and a homogeneous noise $\text{Var}(\boldsymbol{\xi}) = \sigma^2 I_p$; see Section 1.1.4. Below we assume without loss of generality that the eigenvalues λ_j are ordered and decrease with j . This spectral representation means that one observes empirical Fourier coefficients z_m described by the equation $z_j = \lambda_j u_j + \lambda_j^{1/2} \xi_j$ for $j = 1, \dots, p$. The LSE or qMLE estimate of the spectral parameter \mathbf{u} is given by

$$\tilde{\mathbf{u}} = \Lambda^{-1} \mathbf{Z} = (\lambda_1^{-1} z_1, \dots, \lambda_p^{-1} z_p)^\top.$$

Exercise 3.7.1. Consider the spectral representation $\mathbf{Z} = \Lambda \mathbf{u} + \Lambda^{1/2} \boldsymbol{\xi}$. The LSE $\tilde{\mathbf{u}}$ reads as $\tilde{\mathbf{u}} = \Lambda^{-1} \mathbf{Z}$.

If the dimension p of the model is high or, specifically, if the spectral values λ_j rapidly go to zero, it might be useful to only track few coefficients u_1, \dots, u_m and to set all the remaining ones to zero. The corresponding estimate $\tilde{\mathbf{u}}_m = (\tilde{u}_{m,1}, \dots, \tilde{u}_{m,p})^\top$ reads as

$$\tilde{u}_{m,j} \stackrel{\text{def}}{=} \begin{cases} \lambda_j^{-1} z_j & \text{if } j \leq m, \\ 0 & \text{otherwise.} \end{cases}$$

It is usually referred to as a *spectral cut-off* estimate.

Exercise 3.7.2. Consider the linear model $\mathbf{Y} = A \mathbf{f}^* + \boldsymbol{\varepsilon}$. Let U be an orthogonal transform in \mathbb{R}^p providing $U A^\top A U^\top = \Lambda$ with a diagonal matrix Λ leading to the spectral representation for $\mathbf{Z} = U A \mathbf{Y}$. Write the corresponding spectral cut-off estimate $\tilde{\mathbf{f}}_m$ for the original vector \mathbf{f}^* . Show that computing this estimate only requires to know the first m eigenvalues and eigenvectors of the matrix $A^\top A$.

Similarly to the direct case, a spectral cut-off can be extended to *spectral shrinkage*: one multiplies every empirical coefficient z_j with a factor $\alpha_j \in (0, 1)$. This leads to the *spectral shrinkage* estimate $\tilde{\mathbf{u}}_\alpha$ with $\tilde{u}_{\alpha,j} = \alpha_j \lambda_j^{-1} z_j$. Here α stands for the vector of coefficients α_j for $j = 1, \dots, p$. A spectral cut-off method is a special case of this shrinkage with α_j equal to one or zero.

Exercise 3.7.3. Specify the spectral shrinkage $\tilde{\mathbf{u}}_\alpha$ with a given vector α for the situation of Exercise 3.7.2.

The spectral cut-off method can be described as follows. Let ψ_1, ψ_2, \dots be the intrinsic orthonormal basis of the problem composed of the standardized eigenvectors of

$A^\top A$ and leading to the spectral representation $\mathbf{Z} = \Lambda \mathbf{u} + \Lambda^{1/2} \boldsymbol{\xi}$ with the target vector \mathbf{u} . In terms of the original target \mathbf{f}^* , one is looking for a solution or an estimate in the form $\mathbf{f} = \sum_j u_j \boldsymbol{\psi}_j$. The design orthogonality allows to estimate every coefficient u_j independently of the others using the empirical Fourier coefficient $\boldsymbol{\psi}_j^\top \mathbf{Y}$. Namely, $\tilde{u}_j = \lambda_j^{-1} \boldsymbol{\psi}_j^\top \mathbf{Y} = \lambda_j^{-1} z_j$. The LSE procedure tries to recover \mathbf{f} as the full sum $\tilde{\mathbf{f}} = \sum_j \tilde{u}_j \boldsymbol{\psi}_j$. The projection method suggests to cut this sum at the index m : $\tilde{\mathbf{f}}_m = \sum_{j \leq m} \tilde{u}_j \boldsymbol{\psi}_j$, while the shrinkage procedure is based on downweighting the empirical coefficients \tilde{u}_j : $\tilde{\mathbf{f}}_\alpha = \sum_j \alpha_j \tilde{u}_j \boldsymbol{\psi}_j$.

Next we study the risk of the shrinkage method. Orthonormality of the basis $\boldsymbol{\psi}_j$ allows to represent the loss as $\|\tilde{\mathbf{u}}_\alpha - \mathbf{u}^*\|^2 = \|\tilde{\mathbf{f}}_\alpha - \mathbf{f}^*\|^2$. Under the noise homogeneity one obtains the following result.

Theorem 3.7.1. *Let $\mathbf{Z} = \Lambda \mathbf{u}^* + \Lambda^{1/2} \boldsymbol{\xi}$ with $\text{Var}(\boldsymbol{\xi}) = \sigma^2 I_p$. It holds for the shrinkage estimate $\tilde{\mathbf{u}}_\alpha$*

$$\mathcal{R}(\tilde{\mathbf{u}}_\alpha) \stackrel{\text{def}}{=} \mathbb{E} \|\tilde{\mathbf{u}}_\alpha - \mathbf{u}^*\|^2 = \sum_{j=1}^p |\alpha_j - 1|^2 u_j^{*2} + \sum_{j=1}^p \alpha_j^2 \sigma^2 \lambda_j^{-1}.$$

Proof. The empirical Fourier coefficients z_j are uncorrelated and $\mathbb{E} z_j = \lambda_j u_j^*$, $\text{Var} z_j = \sigma^2 \lambda_j$. This implies

$$\mathbb{E} \|\tilde{\mathbf{u}}_\alpha - \mathbf{u}^*\|^2 = \sum_{j=1}^p \mathbb{E} |\alpha_j \lambda_j^{-1} z_j - u_j^*|^2 = \sum_{j=1}^p \{ |\alpha_j - 1|^2 u_j^{*2} + \alpha_j^2 \sigma^2 \lambda_j^{-1} \}$$

as required.

Risk minimization leads to the oracle choice of the vector α or

$$\alpha^* = \underset{\alpha}{\operatorname{argmin}} \mathcal{R}(\tilde{\mathbf{u}}_\alpha)$$

where the minimum is taken over the set of all admissible vectors α .

Similar analysis can be done for the spectral cut-off method.

Exercise 3.7.4. The risk of the spectral cut-off estimate $\tilde{\mathbf{u}}_m$ fulfills

$$\mathcal{R}(\tilde{\mathbf{u}}_m) = \sum_{j=1}^m \lambda_j^{-1} \sigma^2 + \sum_{j=m+1}^p u_j^{*2}.$$

Specify the choice of the oracle cut-off index m^* .

3.8 Roughness penalty and random design

This section discusses how penalization works for random design regression. We consider a penalized log-likelihood

$$L_G(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \frac{1}{2} \|G\boldsymbol{\theta}\|^2 = -\frac{1}{2\sigma^2} \|\mathbf{Y} - \boldsymbol{\Psi}\boldsymbol{\theta}\|^2 - \frac{1}{2} \|G\boldsymbol{\theta}\|^2;$$

(here we ignore the terms which do not depend on $\boldsymbol{\theta}$). The penalty matrix G^2 can depend on the design $\boldsymbol{\Psi}$ and therefore, can be random as well. As usual, define

$$\tilde{\boldsymbol{\theta}}_G \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}),$$

$$\boldsymbol{\theta}_G^* \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E} L_G(\boldsymbol{\theta}).$$

Quadraticity of L_G implies

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_G &= (\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2)^{-1} \boldsymbol{\Psi}\mathbf{Y}, \\ \boldsymbol{\theta}_G^* &= \left\{ \mathbb{E}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2) \right\}^{-1} \mathbb{E}(\boldsymbol{\Psi}\mathbf{Y}) = D_G^{-2} \mathbb{E}(\boldsymbol{\Psi}\mathbf{Y}) \end{aligned}$$

with

$$D_G^2 \stackrel{\text{def}}{=} \sigma^{-2} \mathbb{E}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2).$$

The question of study is whether $\tilde{\boldsymbol{\theta}}_G$ is a good estimator of $\boldsymbol{\theta}_G^*$. Define also \mathbf{M}_G by

$$\mathbf{M}_G^2 \stackrel{\text{def}}{=} \mathbb{E}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2) = \sigma^2 D_G^2.$$

The key step in the analysis is the same as in the non-penalized case: to show that the empirical matrix $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2$ is close to its expectation. We already know that the empirical design matrix $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top$ is close to its expectation. Now we also need to check the concentration properties of the penalty matrix G^2 . This is of course trivial if G^2 is deterministic. However, in some cases, e.g. in the roughness penalty approach, the penalty may depend on the design and therefore, it can be random. Instead of specifying the form of dependence, we just assume in the next result that G^2 is close to $\mathbb{E}G^2$.

Lemma 3.8.1. *Let a set $\Omega(\mathbf{x})$ be such that $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - e^{-x}$, and it holds on $\Omega(\mathbf{x})$ for some $\delta = \delta(\mathbf{x})$*

$$\begin{aligned} \|\mathbf{M}^{-1}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)\mathbf{M}^{-1} - I_p\|_{\text{op}} &\leq \delta, \\ \|(\mathbb{E}G^2)^{-1/2} G^2 (\mathbb{E}G^2)^{-1/2} - I_p\|_{\text{op}} &\leq \delta. \end{aligned} \tag{3.12}$$

Then

$$\|\mathbf{M}_G^{-1}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2)\mathbf{M}_G^{-1} - I_p\|_{\text{op}} \leq \delta.$$

Proof. Conditions (3.12) imply

$$-\delta \mathbf{M}^2 \leq \Psi \Psi^\top - \mathbf{M}^2 \leq \delta \mathbf{M}^2,$$

$$-\delta \mathbb{E} G^2 \leq G^2 - \mathbb{E} G^2 \leq \delta \mathbb{E} G^2$$

and thus

$$-\delta(\mathbf{M}^2 + \sigma^2 \mathbb{E} G^2) \leq \Psi \Psi^\top + \sigma^2 G^2 - (\mathbf{M}^2 + \sigma^2 \mathbb{E} G^2) \leq \delta(\mathbf{M}^2 + \sigma^2 \mathbb{E} G^2)$$

as required.

The standardized score ξ_G can be written as

$$\begin{aligned} \xi_G &\stackrel{\text{def}}{=} D_G^{-1} \nabla L_G(\theta_G^*) \\ &= \sigma^{-2} D_G^{-1} \Psi (\mathbf{Y} - \Psi^\top \theta_G^*) - D_G^{-1} G^2 \theta_G^* \\ &= \sigma^{-1} \mathbf{M}_G^{-1} \Psi \{ \mathbf{Y} - \mathbb{E}(\mathbf{Y} | \Psi) \} + \sigma^{-1} \mathbf{M}_G^{-1} \{ \Psi \mathbb{E}(\mathbf{Y} | \Psi) - (\Psi \Psi^\top + \sigma^2 G^2) \theta_G^* \} \\ &= \sigma^{-1} \mathbf{M}_G^{-1} \Psi \varepsilon + \delta_G(\Psi), \end{aligned} \quad (3.13)$$

where

$$\begin{aligned} \varepsilon &\stackrel{\text{def}}{=} \mathbf{Y} - \mathbb{E}(\mathbf{Y} | \Psi) \\ \delta_G(\Psi) &\stackrel{\text{def}}{=} \sigma^{-1} \mathbf{M}_G^{-1} \{ \Psi \mathbb{E}(\mathbf{Y} | \Psi) - (\Psi \Psi^\top + \sigma^2 G^2) \theta_G^* \}. \end{aligned} \quad (3.14)$$

The vector $\delta_G(\Psi) \in \mathbb{R}^p$ can be viewed as design-dependent estimation error, induced by random design. Remind that $\mathbb{E}(\Psi \Psi^\top + \sigma^2 G^2) \theta_G^* = \mathbb{E}(\Psi \mathbf{Y})$ so that $\delta_G(\Psi)$ vanishes for a deterministic design. For a random design, one can only claim that

$$\mathbb{E} \delta_G(\Psi) = \sigma^{-1} \mathbf{M}_G^{-1} \{ \mathbb{E}(\Psi \mathbf{Y}) - \mathbb{E}(\Psi \Psi^\top + \sigma^2 G^2) \theta_G^* \} = 0.$$

Now we are prepared to state the result on Fisher/Wilks expansions for the penalized MLE $\tilde{\theta}_G$ under random design.

Theorem 3.8.1. *Consider the model (2.1) and suppose*

$$\| \mathbf{M}_G^{-1} (\Psi \Psi^\top + \sigma^2 G^2) \mathbf{M}_G^{-1} - I_p \|_{\text{op}} \leq \delta \quad (3.15)$$

for some $\delta < 1/2$ on a dominating set $\Omega(\mathbf{x})$. Then the penalized MLE $\tilde{\theta}_G$ fulfills on $\Omega(\mathbf{x})$ for ξ_G from (3.13)

$$\| D_G(\tilde{\theta}_G - \theta_G^*) - \xi_G \| \leq \frac{\delta}{1 - \delta} \|\xi_G\|, \quad (3.16)$$

$$|2L_G(\tilde{\theta}_G, \theta_G^*) - \|\xi_G\|^2| \leq \frac{\delta}{1 - \delta} \|\xi_G\|^2. \quad (3.17)$$

Proof. The bound (3.15) also implies

$$\|\mathbf{M}_G(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2)^{-1}\mathbf{M}_G - I_p\|_{\text{op}} \leq \frac{\delta}{1-\delta}. \quad (3.18)$$

By using quadraticity of $L_G(\boldsymbol{\theta})$, one obtains (see Theorem 1.4.1 in Section 1.4)

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_G &= (\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2)^{-1}\boldsymbol{\Psi}\mathbf{Y}, \\ L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*) &= \frac{1}{2\sigma^2}(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*)^\top(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2)(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*). \end{aligned}$$

Further, the model equation (2.1) and the decomposition $\mathbf{Y} = \boldsymbol{\varepsilon} + \mathbb{E}(\mathbf{Y} | \boldsymbol{\Psi})$ imply for $\delta_G(\boldsymbol{\Psi})$ from (3.14) and $\boldsymbol{\xi}_G$ from (3.13)

$$\begin{aligned} D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*) &= D_G(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2)^{-1}\left\{\boldsymbol{\Psi}\boldsymbol{\varepsilon} + \boldsymbol{\Psi}\mathbb{E}(\mathbf{Y} | \boldsymbol{\Psi}) - (\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2)\boldsymbol{\theta}_G^*\right\} \\ &= \mathbf{M}_G(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2)^{-1}\mathbf{M}_G\boldsymbol{\xi}_G = A_G\boldsymbol{\xi}_G \end{aligned}$$

with $A_G \stackrel{\text{def}}{=} \mathbf{M}_G(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2)^{-1}\mathbf{M}_G$. Now we obtain

$$\|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*) - \boldsymbol{\xi}_G\| = \|(A_G - I_p)\boldsymbol{\xi}_G\|$$

so that (3.16) follows from (3.18). Similarly

$$\begin{aligned} 2L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*) &= \sigma^{-2}(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*)^\top(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2)(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*) \\ &= \sigma^{-2}\{D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*)\}^\top D_G^{-1}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2)D_G^{-1}\{D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*)\} \\ &= \boldsymbol{\xi}_G^\top A_G \boldsymbol{\xi}_G \end{aligned}$$

yielding (3.17) on $\Omega(\mathbf{x})$.

Similarly to Sections 2.4 and 2.5, one can establish a deviation bound for the norm $\|\boldsymbol{\xi}_G\|$ entering in the error bound. Definition (3.13) can be rewritten as

$$\begin{aligned} \boldsymbol{\xi}_G &= \boldsymbol{\xi}_{G,0} + \delta_G(\boldsymbol{\Psi}), \\ \boldsymbol{\xi}_{G,0} &\stackrel{\text{def}}{=} \sigma^{-1}\mathbf{M}_G^{-1}\boldsymbol{\Psi}\boldsymbol{\varepsilon}. \end{aligned}$$

The variance of $\boldsymbol{\xi}_{G,0}$ satisfies

$$\begin{aligned} \text{Var}(\boldsymbol{\xi}_{G,0} | \boldsymbol{\Psi}) &= \text{Var}(\sigma^{-1}\mathbf{M}_G^{-1}\boldsymbol{\Psi}\boldsymbol{\varepsilon} | \boldsymbol{\Psi}) = \sigma^{-2}\mathbf{M}_G^{-1}(\boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^\top)\mathbf{M}_G^{-1}, \\ \text{Var}(\boldsymbol{\xi}_{G,0}) &= B_{G,0} \stackrel{\text{def}}{=} \sigma^{-2}\mathbf{M}_G^{-1}\mathbb{E}(\boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^\top)\mathbf{M}_G^{-1}. \end{aligned}$$

The norm $\|\boldsymbol{\xi}_{G,0}\|$ of $\boldsymbol{\xi}_{G,0}$ can be bounded in two steps similarly to the non-penalized case. First we consider a set $\Omega_1(\mathbf{x})$ of dominating measure $1 - 2e^{-\mathbf{x}}$ on which the conditional

expectation $\text{Var}(\boldsymbol{\xi}_{G,0} \mid \boldsymbol{\Psi})$ is close to the unconditional one up the value $\delta_1 = \delta_1(\mathbf{x})$. Then we can apply the deviation bound of Theorem B.2.2 conditionally on $\boldsymbol{\Psi}$ on the set $\Omega_1(\mathbf{x})$ yielding the bound for $\|\boldsymbol{\xi}_{G,0}\|$ similar to (2.25) with $B_{G,0}$ in place of B :

$$\|\boldsymbol{\xi}_{G,0}\| \leq \sqrt{\text{tr}(B_{G,0})} + \sqrt{2\mathbf{x}}$$

with a high probability.

Theorem 3.8.2. Suppose (3.15) on a set $\Omega(\mathbf{x})$. Let the error vector $\boldsymbol{\varepsilon}$ satisfy the exponential moment conditions ... Then

$$\mathbb{P}\left\{\|\boldsymbol{\xi}_{G,0}\| \geq (1 + \delta_1) z(B_{G,0}, \mathbf{x})\right\} \leq 2e^{-\mathbf{x}}.$$

Now we show how the norm $\|\delta_G(\boldsymbol{\Psi})\|$ of the error vector $\delta_G(\boldsymbol{\Psi})$ can be bounded. Introduce the random vector $\mathbf{b}_G \in \mathbb{R}^n$

$$\mathbf{b}_G \stackrel{\text{def}}{=} \mathbb{E}(\mathbf{Y} \mid \boldsymbol{\Psi}) - \boldsymbol{\Psi}^\top \boldsymbol{\theta}_G^*.$$

One can interpret this vector as prediction error in penalized estimation. The next result assumes this error to be small in the sup-norm. Equivalently one can say that each entry $b_i = \mathbb{E}(Y_i \mid \boldsymbol{\Psi}_i) - \boldsymbol{\Psi}_i^\top \boldsymbol{\theta}_G^*$ does not exceed some small value b_∞ .

Theorem 3.8.3. Suppose that $\|\mathbb{E}(\mathbf{Y} \mid \boldsymbol{\Psi}) - \boldsymbol{\Psi}^\top \boldsymbol{\theta}_G^*\|_\infty \leq b_\infty$. Then

$$\mathbb{P}\left\{\|\delta_G(\boldsymbol{\Psi})\| \geq \sigma b_\infty z(B_G, \mathbf{x})\right\} \leq 2e^{-\mathbf{x}}$$

with $B_G \stackrel{\text{def}}{=} \mathbf{M}_G^{-1} \mathbb{E}(\boldsymbol{\Psi} \boldsymbol{\Psi}^\top) \mathbf{M}_G^{-1}$.

Proof. Denote $f_i = \mathbb{E}(Y_i \mid \boldsymbol{\Psi}_i)$ and $b_i = f_i - \boldsymbol{\Psi}_i^\top \boldsymbol{\theta}_G^*$. We use the representation

$$\begin{aligned} \delta_G(\boldsymbol{\Psi}) &= D_G^{-1} \{ \boldsymbol{\Psi} \mathbb{E}(\mathbf{Y} \mid \boldsymbol{\Psi}) - (\boldsymbol{\Psi} \boldsymbol{\Psi}^\top + \sigma^2 G^2) \boldsymbol{\theta}_G^* \} \\ &= D_G^{-1} \sum_i \boldsymbol{\Psi}_i (f_i - \boldsymbol{\Psi}_i^\top \boldsymbol{\theta}_G^*) - \sigma^2 D_G^{-1} G^2 \boldsymbol{\theta}_G^* \\ &= D_G^{-1} \sum_i \boldsymbol{\Psi}_i b_i - \sigma^2 D_G^{-1} G^2 \boldsymbol{\theta}_G^* = D_G^{-1} \sum_i \{\boldsymbol{\Psi}_i b_i - \mathbb{E}(\boldsymbol{\Psi}_i b_i)\}. \end{aligned}$$

The last identity here holds because $\mathbb{E}\delta_G(\boldsymbol{\Psi}) = 0$. Now one can see that Theorem B.2.2 applies to the norm of $\delta_G(\boldsymbol{\Psi})$. The condition $\|\mathbf{b}_G\|_\infty \leq b_\infty$ implies

$$\text{Var}\{\delta_G(\boldsymbol{\Psi})\} = D_G^{-1} \sum_i \text{Var}(\boldsymbol{\Psi}_i b_i) D_G^{-1} \leq \sigma^2 b_\infty^2 \mathbf{M}_G^{-1} \sum_i \mathbb{E}(\boldsymbol{\Psi}_i \boldsymbol{\Psi}_i^\top) \mathbf{M}_G^{-1} = \sigma^2 b_\infty^2 B_G.$$

Now the result follows from Theorem B.2.2 which only relies on the covariance matrix of $\delta_G(\boldsymbol{\Psi})$.

Sieve model selection in linear models

Here we consider the problem of sieve model selection in linear regression model. A high dimensional linear model is approximated by its projection, the main issue is a proper choice of the cut-off parameter.

4.1 Projection estimation. Loss and risk

This section presents the main definitions and properties of the projection estimate.

4.1.1 A linear model

The linear parametric assumption can be stated in the following form: the observed vector \mathbf{Y} is supposed to follow the linear regression model with a homogeneous Gaussian error vector $\boldsymbol{\varepsilon}$:

$$\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n).$$

Here Ψ is the $n \times p$ design matrix formed by a collection of values of basis functions ψ_1, \dots, ψ_p at n design points, and p is the parameter dimension assumed to be large or even infinity. The parametric assumption is only an idealization, for the true data distribution \mathbb{P} of the vector $\mathbf{Y} \in \mathbb{R}^n$ we assume that the observations Y_i are independent and the errors $\varepsilon_i = Y_i - \mathbb{E}Y_i$ have some moments. We also assume that the response vector $\mathbf{f} = \mathbb{E}\mathbf{Y}$ can be well approximated by $\Psi^\top \boldsymbol{\theta}$ for a proper choice of $\boldsymbol{\theta}$, or, equivalently,

$$\mathbf{f} \approx \theta_1^* \psi_1 + \dots + \theta_p^* \psi_p.$$

The MLE or oLSE of the parameter vector $\boldsymbol{\theta}^*$ for this model reads as

$$\tilde{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} \|\mathbf{Y} - \Psi^\top \boldsymbol{\theta}\|^2 = (\Psi \Psi^\top)^{-1} \Psi \mathbf{Y} = \mathcal{S} \mathbf{Y}$$

with $\mathcal{S} = (\Psi \Psi^\top)^{-1} \Psi$.

4.1.2 Linear decomposition of the estimator $\tilde{\theta}$ and quadratic risk

The estimate $\tilde{\theta} = \mathcal{S}\mathbf{Y} = (\Psi\Psi^\top)^{-1}\Psi\mathbf{Y}$ is linear in \mathbf{Y} . This implies by the model equation $\mathbf{Y} = \mathbf{f}^* + \varepsilon$ the decomposition

$$\tilde{\theta} = \mathcal{S}\mathbf{Y} = \mathcal{S}\mathbf{f}^* + \mathcal{S}\varepsilon.$$

We already know that $\tilde{\theta}$ is unbiased, that is,

$$\mathbb{E}\tilde{\theta} = \theta^*,$$

where the target θ^* can be defined as the vector of coefficients of the best linear fit:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \|f - \Psi^\top \theta\|^2 = (\Psi\Psi^\top)^{-1}\Psi f.$$

If the error homogeneity assumption is correct, that is, if $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, then the variance of $\tilde{\theta}$ follows

$$\operatorname{Var}(\tilde{\theta}) = \mathbb{E}\{(\tilde{\theta} - \theta^*)(\tilde{\theta} - \theta^*)^\top\} = (\Psi\Psi^\top)^{-1}\Psi\mathbb{E}(\varepsilon\varepsilon^\top)\Psi^\top(\Psi\Psi^\top)^{-1} = \sigma^2(\Psi\Psi^\top)^{-1}.$$

Moreover, for any $q \times p$ matrix W , it holds in the same way

$$\mathbb{E}[W(\tilde{\theta} - \theta^*)]^2 = \operatorname{tr}[W(\tilde{\theta} - \theta^*)(\tilde{\theta} - \theta^*)^\top W^\top] = \sigma^2 \operatorname{tr}[W(\Psi\Psi^\top)^{-1}W^\top].$$

4.1.3 The case of Inhomogeneous errors

In the general case of $\operatorname{Var}(\varepsilon) = \Sigma$,

$$\begin{aligned} \operatorname{Var}(\tilde{\theta}) &= \mathbb{E}\mathbb{E}\{(\tilde{\theta} - \theta^*)(\tilde{\theta} - \theta^*)^\top\} = (\Psi\Psi^\top)^{-1}\Psi\mathbb{E}(\varepsilon\varepsilon^\top)\Psi^\top(\Psi\Psi^\top)^{-1} \\ &= (\Psi\Psi^\top)^{-1}\Psi\Sigma\Psi^\top(\Psi\Psi^\top)^{-1}. \end{aligned}$$

Exercise 4.1.1. Consider the regression model

$$Y_i = \theta_1^*\psi_1(X_i) + \dots + \theta_p^*\psi_p(X_i) + \varepsilon_i \tag{4.1}$$

with independent heterogeneous errors $\operatorname{Var}(\varepsilon_i) = \sigma_i^2$. Consider the MLE $\tilde{\theta}$ and the LSE $\tilde{\theta}_{\text{LSE}} = (\Psi\Psi^\top)^{-1}\Psi\mathbf{Y}$ and corresponding to homogeneous errors.

- Compute $\tilde{\theta}$
- Show that $\mathbb{E}\tilde{\theta} = \mathbb{E}\tilde{\theta}_{\text{LSE}} = \theta^*$.
- Compute the variance $\operatorname{Var}(\tilde{\theta})$ and the variance $\operatorname{Var}(\tilde{\theta}_{\text{LSE}})$;
- show that $\operatorname{Var}(\tilde{\theta}_{\text{LSE}}) \geq \operatorname{Var}(\tilde{\theta})$;

- check that $\text{Var}(\tilde{\boldsymbol{\theta}}_{\text{LSE}}) = \text{Var}(\tilde{\boldsymbol{\theta}})$ iff all the σ_i are equal to each other.

To illustrate the performance of the MLE $\tilde{\boldsymbol{\theta}}$, consider the case of the orthonormal design $\Psi\Psi^\top = I_p$ and homogeneous errors $\boldsymbol{\varepsilon}$. Then

$$\tilde{\boldsymbol{\theta}} = \Psi\mathbf{Y}, \quad \text{Var}(\tilde{\boldsymbol{\theta}}) = \sigma^2 I_p.$$

and for any symmetric positive matrix W , it holds

$$\mathbb{E}\{W(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\}^2 = \sigma^2 \text{tr}(WW^\top)$$

Now we consider the prediction $\tilde{\mathbf{f}} = \Psi^\top \tilde{\boldsymbol{\theta}}$.

4.1.4 Prediction error $\tilde{\mathbf{f}} - \mathbf{f}^*$

It follows

$$\tilde{\mathbf{f}} = \Psi^\top \tilde{\boldsymbol{\theta}} = \Psi^\top (\Psi\Psi^\top)^{-1} \Psi\mathbf{Y} = \Pi\mathbf{Y}.$$

Here Π is the projector in the space \mathbb{R}^n on the linear subspace spanned by the basis vectors ψ_1, \dots, ψ_p . The model equation $\mathbf{Y} = \mathbf{f}^* + \boldsymbol{\varepsilon}$ implies the decomposition

$$\tilde{\mathbf{f}} = \Pi\mathbf{f}^* + \Pi\boldsymbol{\varepsilon}.$$

It implies

$$\mathbb{E}\tilde{\mathbf{f}} = \Pi\mathbf{f}^*.$$

Moreover, under the noise homogeneity $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$, it holds

$$\text{Var}(\tilde{\mathbf{f}}) = \mathbb{E}(\Pi\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\Pi) = \sigma^2\Pi.$$

4.1.5 Quadratic loss. Bias-variance decomposition

For the quadratic loss function $\varrho(\tilde{\mathbf{f}}, \mathbf{f}^*) = \|\tilde{\mathbf{f}} - \mathbf{f}^*\|^2$, it follows

$$\begin{aligned} \varrho(\tilde{\mathbf{f}}, \mathbf{f}^*) &= \|\tilde{\mathbf{f}} - \mathbf{f}^*\|^2 = \|\Pi\mathbf{f}^* - \mathbf{f}^* + \Pi\boldsymbol{\varepsilon}\|^2 \\ &= \|(I_p - \Pi)\mathbf{f}^*\|^2 + \|\Pi\boldsymbol{\varepsilon}\|^2 + 2(\Pi\boldsymbol{\varepsilon})^\top(I_p - \Pi)\mathbf{f}^* \\ &= \|(I_p - \Pi)\mathbf{f}^*\|^2 + \|\Pi\boldsymbol{\varepsilon}\|^2. \end{aligned} \tag{4.2}$$

Here we have used that Π is a projector in \mathbb{R}^n which implies $\Pi^\top(I_p - \Pi) = 0$. Therefore, the quadratic risk $\mathcal{R}(\tilde{\mathbf{f}}, \mathbf{f}^*) = \mathbb{E}\varrho(\tilde{\mathbf{f}}, \mathbf{f}^*)$ satisfies under homogeneous errors $\text{Var}(\varepsilon) = \sigma^2 I_p$

$$\mathcal{R}(\tilde{\mathbf{f}}, \mathbf{f}^*) = \|(I_p - \Pi)\mathbf{f}^*\|^2 + \|\Pi\varepsilon\|^2 = \|(I_p - \Pi)\mathbf{f}^*\|^2 + \sigma^2 p$$

as

$$\mathbb{E}\|\Pi\varepsilon\|^2 = \mathbb{E}(\Pi\varepsilon)^\top(\Pi\varepsilon) = \mathbb{E}\text{tr}(\Pi\varepsilon)(\Pi\varepsilon)^\top = \text{tr}(\Pi\mathbb{E}\varepsilon\varepsilon^\top\Pi^\top) = \sigma^2 \text{tr}(\Pi\Pi^\top) = \sigma^2 p.$$

The term $\|(I_p - \Pi)\mathbf{f}^*\|^2$ is usually called *the squared bias* and it describes the accuracy of approximation of \mathbf{f}^* by its projection $\Pi\mathbf{f}^*$ on the space generated by the basis functions ψ_1, \dots, ψ_p . The term $\sigma^2 p$ called *the variance* measures the statistical error related to estimation of p unknown coefficients in the decomposition of $\Pi\mathbf{f}^* = \theta_1^*\psi_1 + \dots + \theta_p^*\psi_p$.

Exercise 4.1.2. Consider the nonparametric model

$$Y_i = f(X_i) + \varepsilon_i, \quad \text{Var}(\varepsilon_i) = \sigma_i^2$$

and the parametric approximation (4.1). Consider the qMLE (LSE) $\tilde{\mathbf{f}}_{\text{LSE}} = \Pi\mathbf{Y}$ for $\Pi = \Psi^\top(\Psi\Psi^\top)^{-1}\Psi$.

- Derive the bias-variance decomposition for the quadratic losses $\|\tilde{\mathbf{f}}_{\text{LSE}} - \mathbf{f}^*\|^2$ and of the risk $\mathbb{E}\|\tilde{\mathbf{f}}_{\text{LSE}} - \mathbf{f}^*\|^2$.
- Compute the variance term of $\tilde{\mathbf{f}}_{\text{LSE}}$ and of $\tilde{\mathbf{f}} = \Psi^\top\tilde{\boldsymbol{\theta}}$ for the MLE $\tilde{\boldsymbol{\theta}}$.

4.1.6 Projection estimation and the model choice problem

In this section we consider the linear model $\mathbf{Y} = \Psi^\top\boldsymbol{\theta} + \varepsilon$ with a p -dimensional parameter p which is large or even infinity. The full dimensional estimation of the parameter $\boldsymbol{\theta}$ can be highly inefficient. Here we consider the simplest method of complexity reduction called *projection*. The idea is to use just a submodel corresponding to the reduced subset of parameters.

We associate the rows of the design matrix Ψ with basis vectors in \mathbb{R}^n . By Ψ_m we denote a submatrix of Ψ composed of the first m rows ψ_1, \dots, ψ_m . It corresponds to the reduced regression model

$$\mathbf{Y} = \Psi_m^\top\boldsymbol{\theta}_m + \varepsilon$$

with the parameter $\boldsymbol{\theta}_m$ from \mathbb{R}^m . The corresponding estimate $\tilde{\boldsymbol{\theta}}_m$ and the predictor $\tilde{\mathbf{f}}_m$ read as

$$\begin{aligned}\tilde{\boldsymbol{\theta}}_m &= (\Psi_m \Psi_m^\top)^{-1} \Psi_m \mathbf{Y}, \\ \tilde{\mathbf{f}}_m &= \Psi_m (\Psi_m \Psi_m^\top)^{-1} \Psi_m \mathbf{Y} = \Pi_m \mathbf{Y}\end{aligned}$$

where Π_m is a projector in \mathbb{R}^n on the subspace spanned by the basis vectors ψ_1, \dots, ψ_m .

In the case of an orthonormal design, one just considers the first m empirical coefficients z_1, \dots, z_m and drop the others. The corresponding parameter estimate $\tilde{\boldsymbol{\theta}}_m$ reads as

$$\tilde{\theta}_{m,j} = \begin{cases} z_j & \text{if } j \leq m, \\ 0 & \text{otherwise} \end{cases}$$

with $z_j = \psi_j^\top \mathbf{Y}$. The response vector $\mathbf{f}^* = E\mathbf{Y}$ is estimated by $\Psi^\top \tilde{\boldsymbol{\theta}}_m$ leading to the representation

$$\tilde{\mathbf{f}}_m = z_1 \psi_1 + \dots + z_m \psi_m.$$

In other words, $\tilde{\mathbf{f}}_m$ is just a projection of the observed vector \mathbf{Y} onto the subspace L_m spanned by the first m basis vectors ψ_1, \dots, ψ_m : $L_m = \langle \psi_1, \dots, \psi_m \rangle$. This explains the name of the method. Clearly one can study the properties of $\tilde{\boldsymbol{\theta}}_m$ or $\tilde{\mathbf{f}}_m$ using the methods of previous sections. However, one more question for this approach is still open: a proper choice of m . The standard way of accessing this issue is based on the analysis of the quadratic risk.

Consider first the prediction risk defined as $\mathcal{R}(\tilde{\mathbf{f}}_m, \mathbf{f}^*) = E\|\tilde{\mathbf{f}}_m - \mathbf{f}^*\|^2$. Below we focus on the case of a homogeneous noise with $\text{Var}(\varepsilon) = \sigma^2 I_p$. An extension to the colored noise is possible. Recall that $\tilde{\mathbf{f}}_m$ effectively estimates the vector $\mathbf{f}_m = \Pi_m \mathbf{f}^*$, where Π_m is the projector on L_m ; see Section 1.2.3. Moreover, the quadratic risk $\mathcal{R}(\tilde{\mathbf{f}}_m, \mathbf{f}^*)$ can be decomposed as

$$\mathcal{R}(\tilde{\mathbf{f}}_m, \mathbf{f}^*) = \|\mathbf{f}^* - \Pi_m \mathbf{f}^*\|^2 + \sigma^2 m = \sigma^2 m + \sum_{j=m+1}^p \theta_j^{*2}. \quad (4.3)$$

Obviously the squared bias $\|\mathbf{f}^* - \Pi_m \mathbf{f}^*\|^2$ decreases with m while the variance $\sigma^2 m$ linearly grows with m . Risk minimization leads to the so called *bias-variance trade-off*: one selects m which minimizes the risk $\mathcal{R}(\tilde{\mathbf{f}}_m, \mathbf{f}^*)$ over all possible m :

$$m^* \stackrel{\text{def}}{=} \operatorname{argmin}_m \mathcal{R}(\tilde{\mathbf{f}}_m, \mathbf{f}^*) = \operatorname{argmin}_m \{\|\mathbf{f}^* - \Pi_m \mathbf{f}^*\|^2 + \sigma^2 m\}. \quad (4.4)$$

Unfortunately this choice requires some information about the bias $\|\mathbf{f}^* - \Pi_m \mathbf{f}^*\|$ which depends on the unknown vector \mathbf{f}^* . As this information is not available in typical situation, the value m^* is also called an *oracle* choice. A data-driven choice of m is one of the central issue in the nonparametric statistics.

The situation is not changed if we consider the estimation risk $E\|\tilde{\theta}_m - \theta^*\|^2$. Indeed, the basis orthogonality $\Psi\Psi^\top = I_p$ implies for $\mathbf{f}^* = \Psi^\top\theta^*$

$$\|\tilde{\mathbf{f}}_m - \mathbf{f}^*\|^2 = \|\Psi^\top\tilde{\theta}_m - \Psi^\top\theta^*\|^2 = \|\tilde{\theta}_m - \theta^*\|^2$$

and minimization of the estimation risk coincides with minimization of the prediction risk.

The problem of selecting the model m^* can be stated in different ways depending on what is the target and objective of the method. Usually the problem is formulated as the problem of *adaptive estimation* and the one aims at constructing an estimate $\hat{\theta}$ whose risk is close to the risk of the oracle $\tilde{\theta}_{m^*}$. The problem of *model selection* mainly focuses choosing a proper model \hat{m} on the base of available data. The latter problem is appealing if one is concerned with inference, prediction, or some other model-based question. To understand the difference between two possible setups, consider the ideal situation when the risk is completely flat: $\mathcal{R}_m \equiv C$. Then any model choice yields the same risk and one can free to take any model. In terms of building a confidence statement, for prediction or testing, the model choice matters a lot and a smaller model (in term of complexity) will be much more useful. In some sense, two mentioned objectives are contradictory: a flat risk is very good for estimation and enables us to apply a simple rule-of-thumb for choosing the parameter m . However, identification of a good model is very hard for models with a flat risk function. At the same time, the case of a profiled risk makes the choice of the model crucial but it can be identified much easier. Below we try to address both issues: estimation of the parameter θ^* and of the oracle model m^* .

4.2 Unbiased risk estimation

The “oracle” choice m^* cannot be implemented because the bias term $\|\mathbf{f}^* - \Pi_m \mathbf{f}^*\|^2$ depends on the target object \mathbf{f}^* . Now we want to develop a data-driven rule which attempts to reproduce (mimic) the oracle. The first naive idea is to look at the empirical risk (data fit) $\|\mathbf{Y} - \tilde{\mathbf{f}}_m\|^2$ in which we replace the function \mathbf{f} by the data \mathbf{Y} . Unfortunately, this rule leads to the trivial solution

$$\hat{m} = \operatorname{argmin}_m \|\mathbf{Y} - \tilde{\mathbf{f}}_m\|^2 = p.$$

Indeed, the value $\|\mathbf{Y} - \tilde{\mathbf{f}}_m\|^2$ monotonously decreases with m as follows from the next lemma.

Lemma 4.2.1. *Consider the projection estimator $\tilde{\mathbf{f}}_m = \Pi_m \mathbf{Y}$. For two different values $m' > m$, the following statements hold:*

- $\Pi_{m',m} \stackrel{\text{def}}{=} \Pi_{m'} - \Pi_m$ is a projector in \mathbb{R}^n .
- If Ψ is orthogonal then $\Pi_{m',m}$ projects onto subspace generated by $\psi_{m+1}, \dots, \psi_{m'}$.
- The next identity is fulfilled:

$$\|\mathbf{Y} - \Pi_m \mathbf{Y}\|^2 - \|\mathbf{Y} - \Pi_{m'} \mathbf{Y}\|^2 = \|\Pi_{m',m} \mathbf{Y}\|^2 = \|\Pi_{m',m} \mathbf{f} + \Pi_{m',m} \boldsymbol{\varepsilon}\|^2 \geq 0.$$

Proof. It obviously holds

$$\begin{aligned} \|\mathbf{Y} - \Pi_m \mathbf{Y}\|^2 - \|\mathbf{Y} - \Pi_{m'} \mathbf{Y}\|^2 &= \mathbf{Y}^\top (I - \Pi_m) \mathbf{Y} - \mathbf{Y}^\top (I - \Pi_{m'}) \mathbf{Y} \\ &= \mathbf{Y}^\top (\Pi_{m'} - \Pi_m) \mathbf{Y} = \|\Pi_{m',m} \mathbf{Y}\|^2. \end{aligned}$$

Therefore, empirical risk minimization always tries to select the largest possible model which provides the best data fit. In the extreme case of $m = n$, we obtain the perfect fit $\tilde{\mathbf{f}}_m = \mathbf{Y}$, that is, the estimate coincides with the data. This is formally correct but the corresponding squared risk is equal to $\sigma^2 p$ which can be a very large number. So, the empirical risk minimization does not do the required job, it does not mimic the “oracle” risk minimizer. Now we try to look more attentively at the empirical risk to understand the origin of the problem. First compute its expectation.

Lemma 4.2.2. *It holds under homogeneous errors $\boldsymbol{\varepsilon}$:*

- For each m

$$\mathbb{E} \|\mathbf{Y} - \Pi_m \mathbf{Y}\|^2 = \|\mathbf{f}^* - \Pi_m \mathbf{f}^*\|^2 + \sigma^2(n - m). \quad (4.5)$$

- For any $m' > m$

$$\mathbb{E} \|\mathbf{Y} - \Pi_m \mathbf{Y}\|^2 - \mathbb{E} \|\mathbf{Y} - \Pi_{m'} \mathbf{Y}\|^2 = \|\Pi_{m',m} \mathbf{f}^*\|^2 + \sigma^2(m' - m)$$

Proof. Obvious.

The first term in the statement (4.5) is exactly the squared bias which is a good news: the empirical risk contains the same term which we need in the squared risk evaluation. Unfortunately, the second term $\sigma^2(n - m)$ behaves differently than the similar variance term $\sigma^2 m$. Another good news is that both variance terms are known to us. Therefore, one can easily make a correction of the empirical risk which delivers an unbiased risk estimate: just add $\sigma^2(2m - n)$. Define

$$\tilde{\mathcal{R}}_m = \|\mathbf{Y} - \Pi_m \mathbf{Y}\|^2 + 2\sigma^2 m.$$

Then it holds

$$\mathbb{E}\tilde{\mathcal{R}}_m = \mathcal{R}(\tilde{\mathbf{f}}_m, \mathbf{f}^*) + \sigma^2 n.$$

In words, the expectation of $\tilde{\mathcal{R}}_m$ is equal to the risk $\mathcal{R}(\tilde{\mathbf{f}}_m, \mathbf{f}^*)$ up to the fixed term $\sigma^2 n$ which does not affect the model choice. This suggests to define

$$\hat{m} \stackrel{\text{def}}{=} \operatorname{argmin}_m \tilde{\mathcal{R}}_m = \operatorname{argmin}_m (\|\mathbf{Y} - \Pi_m \mathbf{Y}\|^2 + 2\sigma^2 m). \quad (4.6)$$

This rule is known as Akaike information criteria (AIC) and it is very popular in practical applications. It suggests to balance the data fit measured by $\|\mathbf{Y} - \Pi_m \mathbf{Y}\|^2$ and the model complexity $2\sigma^2 m$. One can say that this rule selects a model with a possibly small complexity $\sigma^2 m$ still providing a reasonable data fit $\|\mathbf{Y} - \Pi_m \mathbf{Y}\|^2$.

Exercise 4.2.1. Consider the projection estimator $\tilde{\mathbf{f}}_m = \Pi_m \mathbf{Y}$ for the model (4.1) with $\Pi_m = \Psi_m^\top (\Psi_m \Psi_m^\top)^{-1} \Psi_m$. For two different values $m' > m$:

- Check that $\Pi_{m',m} \stackrel{\text{def}}{=} \Pi_{m'} - \Pi_m$ is a projector in \mathbb{R}^n . Describe its image in the orthogonal case when $\Psi \Psi^\top$ is a diagonal matrix.
- Check the identities

$$\begin{aligned} \|\mathbf{Y} - \Pi_m \mathbf{Y}\|^2 - \|\mathbf{Y} - \Pi_{m'} \mathbf{Y}\|^2 &= \|\tilde{\mathbf{f}}_{m'} - \tilde{\mathbf{f}}_m\|^2 = \|\Pi_{m',m} \mathbf{f} + \Pi_{m',m} \boldsymbol{\varepsilon}\|^2, \\ \|\tilde{\mathbf{f}}_{m'} - \mathbf{f}\|^2 - \|\tilde{\mathbf{f}}_m - \mathbf{f}\|^2 &= -\|\Pi_{m',m} \mathbf{f}\|^2 + \|\Pi_{m',m} \boldsymbol{\varepsilon}\|^2. \end{aligned}$$

- compute $\mathbb{E}\|\tilde{\mathbf{f}}_{m'} - \tilde{\mathbf{f}}_m\|^2$ and $\mathbb{E}[\|\tilde{\mathbf{f}}_{m'} - \mathbf{f}\|^2 - \|\tilde{\mathbf{f}}_m - \mathbf{f}\|^2]$.

4.2.1 AIC and pairwise comparison

Here we try to understand whether the AIC rule does a good job in model selection. In particular, whether it mimics the oracle. Our study will be based on pairwise comparison. More precisely, we check two situations: when the data-driven choice \hat{m} is larger than the oracle and the inverse case. The most important problem is to bound the probability and the risk associated with the event $\{\hat{m} > m^*\}$.

The definition of m^* (4.4) implies for $m > m^*$ with $\mathcal{R}_m \stackrel{\text{def}}{=} \mathcal{R}(\tilde{\mathbf{f}}_m, \mathbf{f}^*)$

$$\begin{aligned} \mathcal{R}_m - \mathcal{R}_{m^*} &= \|\mathbf{f}^* - \Pi_m \mathbf{f}^*\|^2 - \|\mathbf{f}^* - \Pi_{m^*} \mathbf{f}^*\|^2 + \sigma^2(m - m^*) \\ &= -\|b_{m,m^*}\|^2 + \sigma^2(m - m^*) \geq 0, \end{aligned} \quad (4.7)$$

where $b_{m,m^*} \stackrel{\text{def}}{=} \Pi_{m,m^*} \mathbf{f}^*$.

Exercise 4.2.2. Consider the model $\mathbf{Y} = \mathbf{f}^* + \boldsymbol{\varepsilon}$ with homogeneous errors $\sigma_i \equiv \sigma$. Let m^* be the oracle choice from (4.4).

- check (4.7);
- check that for $m < m^*$, it holds

$$\|b_{m^*,m}\|^2 = \|\Pi_{m^*,m} \mathbf{f}\|^2 \geq \sigma^2(m^* - m). \quad (4.8)$$

- check that for $m > m^*$, it holds

$$\|b_{m,m^*}\|^2 = \|\Pi_{m,m^*} \mathbf{f}\|^2 \leq \sigma^2(m - m^*) \quad (4.9)$$

In words, due to (4.8), it is reasonable to increase the model complexity towards m^* , the gain in the quality of approximation is larger than the additional complexity. However, (4.9) shows that if we increase the complexity of the model over the oracle m^* , then our additional loss due to increased complexity exceeds the gain due to bias reduction.

The next question is whether the data-driven choice \hat{m} reproduces this situation. The selected model \hat{m} is a winner in a pairwise competition with all other models, in particular, in competition with the “oracle” choice m^* . This means that $\tilde{\mathcal{R}}_{\hat{m}} \leq \tilde{\mathcal{R}}_{m^*}$. If the value $\tilde{\mathcal{R}}_m$ is close to its expectation \mathcal{R}_m and if \mathcal{R}_m is significantly larger than the oracle risk \mathcal{R}_{m^*} then the probability of the event $\tilde{\mathcal{R}}_m \leq \tilde{\mathcal{R}}_{m^*}$ is very small. So, one can expect that the selected model \hat{m} is mainly located on the set where the risk \mathcal{R}_m does not deviate much from \mathcal{R}_{m^*} . The next result quantifies this relation. We use the decomposition

$$\begin{aligned} \tilde{\mathcal{R}}_m - \tilde{\mathcal{R}}_{m^*} &= \|\mathbf{Y} - \Pi_m \mathbf{Y}\|^2 + 2\sigma^2 m - \|\mathbf{Y} - \Pi_{m^*} \mathbf{Y}\|^2 - 2\sigma^2 m^* \\ &= -\|\Pi_{m,m^*} \mathbf{Y}\|^2 + 2\sigma^2(m - m^*) \\ &= -\|\Pi_{m,m^*} \boldsymbol{\varepsilon} + b_{m,m^*}\|^2 + 2\sigma^2(m - m^*) \\ &= \mathcal{R}_m - \mathcal{R}_{m^*} - \{\|\Pi_{m,m^*} \boldsymbol{\varepsilon}\|^2 - \sigma^2(m - m^*)\} - 2b_{m,m^*}^\top \Pi_{m,m^*} \boldsymbol{\varepsilon}. \end{aligned} \quad (4.10)$$

The first stochastic term $\|\Pi_{m,m^*} \boldsymbol{\varepsilon}\|^2 - \sigma^2(m - m^*)$ of this difference is a centered quadratic form of the errors $\boldsymbol{\varepsilon}$ and the unknown regression function f does not show up there. The second one $2b_{m,m^*}^\top \Pi_{m,m^*} \boldsymbol{\varepsilon}$ involves the bias b_{m,m^*} but it is linear in $\boldsymbol{\varepsilon}$. Both terms can be easily bounded for the Gaussian errors $\boldsymbol{\varepsilon}$.

Lemma 4.2.3. *Let $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Then it holds for $b_{m,m^*} = \Pi_{m,m^*} \mathbf{f}^*$*

$$\mathbb{P}\left(\sigma^{-2} \|\Pi_{m,m^*} \boldsymbol{\varepsilon}\|^2 > \mathfrak{z}^+(m - m^*, \mathbf{x})\right) \leq \frac{1}{2} e^{-x},$$

$$\mathbb{P}\left(\sigma^{-2} \|\Pi_{m,m^*} \boldsymbol{\varepsilon}\|^2 < \mathfrak{z}^-(m - m^*, \mathbf{x})\right) \leq \frac{1}{2} e^{-x},$$

$$\mathbb{P}\left(\sigma^{-2} |b_{m,m^*}^\top \Pi_{m,m^*} \boldsymbol{\varepsilon}| > \sigma^{-1} \|b_{m,m^*}\| z_1(\mathbf{x})\right) \leq e^{-x}.$$

where $\mathfrak{z}^+(k, \mathbf{x})$ is the upper $1 - 0.5e^{-\mathbf{x}}$ quantile of χ_k^2 , $\mathfrak{z}^-(k, \mathbf{x})$ is its lower $0.5e^{-\mathbf{x}}$ quantile, $z_1(\mathbf{x})$ is the quantile of ξ for a standard normal r.v. $\xi \sim \mathcal{N}(0, 1)$: $\mathbb{P}(|\xi| > z_1(\mathbf{x})) \leq e^{-\mathbf{x}}$.

It holds for any $k \geq 1$ and $\mathbf{x} > 0$ with $\mathbf{x}_1 = \mathbf{x} + \log(2)$

$$\begin{aligned}\mathfrak{z}^+(k, \mathbf{x}) &\leq k + 2\sqrt{k \mathbf{x}_1} + 2\mathbf{x}_1, \\ \mathfrak{z}^-(k, \mathbf{x}) &\geq k - 2\sqrt{k \mathbf{x}_1}.\end{aligned}\tag{4.11}$$

The proof only uses that $\sigma^{-1}\boldsymbol{\varepsilon}$ is a standard Gaussian vector in \mathbb{R}^n and thus, $\sigma^{-1}\Pi_{m,m^*}\boldsymbol{\varepsilon}$ is standard normal in \mathbb{R}^{m-m^*} , while $(\sigma\|b_{m,m^*}\|)^{-1}b_{m,m^*}^\top\Pi_{m,m^*}\boldsymbol{\varepsilon}$ is a standard normal r.v. if the bias b_{m,m^*} does not vanish.

The presented bounds show that for moderate values of \mathbf{x}

$$z^\pm(k, \mathbf{x}) \stackrel{\text{def}}{=} \mathfrak{z}^+(k, \mathbf{x}) - \mathfrak{z}^-(k, \mathbf{x}) \leq C\sqrt{k \mathbf{x}}.$$

for a fixed constant C . Therefore, for large k , the interquartile range $z^\pm(k, \mathbf{x}) = \mathfrak{z}^+(k, \mathbf{x}) - \mathfrak{z}^-(k, \mathbf{x})$ is small relative to k which is the expectation of $\sigma^{-2}\mathbb{E}\|\Pi_k\boldsymbol{\varepsilon}\|^2$. This effect is called *concentration* and it explains why the AIC rule works: the difference between empirical risk and its population counterpart is small relatively to the risk itself.

4.2.2 Pairwise analysis

Now we make a more precise analysis of the term $\tilde{\mathcal{R}}_m - \tilde{\mathcal{R}}_{m^*}$ in (4.10). It is based on the following general property of the Gaussian distribution.

Lemma 4.2.4. Let $\boldsymbol{\xi}$ be standard Gaussian vector in \mathbb{R}^k and $\boldsymbol{\delta}$ be a deterministic vector in \mathbb{R}^k with $\|\boldsymbol{\delta}\|^2 = \Delta$. Then

- the distribution of $\|\boldsymbol{\xi} + \boldsymbol{\delta}\|^2$ only depends on k and Δ ;
- let, for a given \mathbf{x} , the quantiles $\mathfrak{z}^+(k, \Delta; \mathbf{x})$ and $\mathfrak{z}^-(k, \Delta; \mathbf{x})$ be defined as

$$\begin{aligned}\mathbb{P}(\|\boldsymbol{\xi} + \boldsymbol{\delta}\|^2 \geq \mathfrak{z}^+(k, \Delta; \mathbf{x})) &= e^{-\mathbf{x}}, \\ \mathbb{P}(\|\boldsymbol{\xi} + \boldsymbol{\delta}\|^2 \leq \mathfrak{z}^-(k, \Delta; \mathbf{x})) &= e^{-\mathbf{x}}.\end{aligned}\tag{4.12}$$

Then

$$\begin{aligned}\mathfrak{z}^+(k, \Delta; \mathbf{x}) &\leq \Delta + \mathfrak{z}^+(k, \mathbf{x}) + 2\Delta^{1/2}z_1(\mathbf{x}), \\ \mathfrak{z}^-(k, \Delta; \mathbf{x}) &\geq \Delta + \mathfrak{z}^-(k, \mathbf{x}) - 2\Delta^{1/2}z_1(\mathbf{x}),\end{aligned}\tag{4.13}$$

Proof. Use the decomposition

$$\|\boldsymbol{\xi} + \boldsymbol{\delta}\|^2 = \Delta + \|\boldsymbol{\xi}\|^2 + 2\boldsymbol{\xi}^\top\boldsymbol{\delta}$$

and Lemma 4.2.3.

We apply this result to $\pm\sigma^{-2}\|\Pi_{m,m^*}\mathbf{Y}\|^2$ entering in the difference $\tilde{\mathcal{R}}_m - \tilde{\mathcal{R}}_{m^*}$. The bound $\tilde{\mathcal{R}}_m - \tilde{\mathcal{R}}_{m^*}$ can be rewritten for $m > m^*$ as $\sigma^{-2}\|\Pi_{m,m^*}\mathbf{Y}\|^2 > 2(m - m^*)$ which is directly related to the upper quantile of non-central chi-squared. Define the value of non-centrality parameter Δ to have $\mathfrak{z}^\pm(k, \Delta^\pm; \mathbf{x})$ exactly equal to $2k$:

$$\mathfrak{z}^+(k, \Delta^+(k, \mathbf{x}); \mathbf{x}) = 2k, \quad \mathfrak{z}^-(k, \Delta^-(k, \mathbf{x}); \mathbf{x}) = 2k. \quad (4.14)$$

This definition can be rewritten as follows:

$$\mathbb{P}(\|\boldsymbol{\xi} + \boldsymbol{\delta}^+\|^2 > 2k) \leq e^{-\mathbf{x}}, \quad \text{if } \|\boldsymbol{\delta}^+\|^2 \leq \Delta^+(k, \mathbf{x}), \quad (4.15)$$

$$\mathbb{P}(\|\boldsymbol{\xi} + \boldsymbol{\delta}^-\|^2 < 2k) \leq e^{-\mathbf{x}}, \quad \text{if } \|\boldsymbol{\delta}^-\|^2 \geq \Delta^-(k, \mathbf{x}). \quad (4.16)$$

Exercise 4.2.3. For the quantities $\Delta^+(k, \mathbf{x}), \Delta^-(k, \mathbf{x})$ from (4.14) and $\mathbf{x} \geq 1$

- show that $\Delta^+(k, \mathbf{x}) < k$, $\Delta^-(k, \mathbf{x}) > k$;
- check that Lemma 4.2.4 implies

$$\Delta^+(k, \mathbf{x}) \geq \mathfrak{z}^-(k, \mathbf{x}) - 2k^{1/2}z_1(\mathbf{x}), \quad (4.17)$$

$$\Delta^-(k, \mathbf{x}) \leq \mathfrak{z}^+(k, \mathbf{x}) + 2k^{1/2}z_1(\mathbf{x}), \quad (4.18)$$

We conclude with the following statement.

Proposition 4.2.1. *Let the errors $\boldsymbol{\varepsilon}$ be normal and homogeneous: $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Then the inequalities*

$$\sigma^{-2}\|b_{m,m^*}\|^2 \leq \Delta^+(m - m^*, \mathbf{x}), \quad m > m^* \quad (4.19)$$

$$\sigma^{-2}\|b_{m^*,m}\|^2 \geq \Delta^-(m^* - m, \mathbf{x}), \quad m < m^*$$

ensures

$$\mathbb{P}(\tilde{\mathcal{R}}_m \leq \tilde{\mathcal{R}}_{m^*}) \leq e^{-\mathbf{x}}.$$

In particular, if the bias term $\|b_m\|$ is uniformly bounded for all $m \geq m^*$ by a fixed constant $C(\mathcal{F})$, then $\|b_{m,m^*}\| \leq C(\mathcal{F})$ and the inequality (4.19) is fulfilled if

$$m - m^* > (2\sigma^{-1}C(\mathcal{F}) + C\mathbf{x})^2 \quad (4.20)$$

for $C \geq 3$.

Proof. Consider the case $m > m^*$. We apply the decomposition

$$\sigma^{-2}(\tilde{\mathcal{R}}_m - \tilde{\mathcal{R}}_{m^*}) = -\|\sigma^{-1}\Pi_{m,m^*}\varepsilon + \sigma^{-1}b_{m,m^*}\|^2 + 2(m - m^*).$$

Further, $\xi \stackrel{\text{def}}{=} \sigma^{-1}\Pi_{m,m^*}\varepsilon$ is standard normal in \mathbb{R}^k for $k = m - m^*$. The condition (4.15) with $\delta^+ = \sigma^{-1}b_{m,m^*}$ implies

$$\mathbb{P}(\tilde{\mathcal{R}}_m \leq \tilde{\mathcal{R}}_{m^*}) = \mathbb{P}(\|\xi + \delta^+\|^2 > 2k) \leq e^{-x}.$$

The case $m < m^*$ can be done in a similar way using (4.16) in place of (4.15).

The inequality $\|b_m\| \leq C(\mathcal{F})$ implies $\|b_{m,m^*}\| \leq C(\mathcal{F})$ for all $m > m^*$; see (4.21).

Now it remains to check by (4.17) and (4.11) that (4.20) implies (4.19).

To be done: complete the proof

Exercise 4.2.4. Show that the value $\|b_m\| = \|\mathbf{f}^* - \Pi_m \mathbf{f}^*\|$ monotonously decreases with m . Moreover, for any $m' > m$, the relative bias $b_{m',m} = \Pi_{m',m} \mathbf{f}^*$ satisfies

$$\|b_{m',m}\| \leq \|b_m\|, \quad m' > m. \quad (4.21)$$

Check whether also holds

$$\|b_{m',m}\| \leq \|b_{m'}\|, \quad m' > m.$$

4.2.3 Uniform bounds and the zone of insensitivity

This section introduces the *set of insensitivity* $\mathcal{M}^\circ(x)$ which describes the quality of model selection. Namely, we aim at describing the set $\mathcal{M}^\circ(x)$ which contains all possible values of \hat{m} with a high probability. The ideal situation would be $\mathcal{M}^\circ(x) = \{m^*\}$, but it is rare the case. Usually $\mathcal{M}^\circ(x)$ is a larger set containing m^* . Below we specify this set in terms of the difference $\mathcal{R}_m - \mathcal{R}_{m^*}$ and its decomposition (4.7).

The study of the previous section quantifies the pairwise relation $\tilde{\mathcal{R}}_m - \tilde{\mathcal{R}}_{m^*} \leq 0$: under $\sigma^{-2}\|b_{m,m^*}\|^2 \leq \Delta^+(m - m^*, x)$; see (4.19), it holds

$$\mathbb{P}(\tilde{\mathcal{R}}_m \leq \tilde{\mathcal{R}}_{m^*}) \leq e^{-x}. \quad (4.22)$$

Now we need its uniform version over the complement of $\mathcal{M}^\circ(x)$:

$$\mathbb{P}\left(\max_{m \notin \mathcal{M}^\circ(x)} \{\tilde{\mathcal{R}}_m - \tilde{\mathcal{R}}_{m^*}\} \geq 0\right) \leq e^{-x}.$$

One can use a uniform adjustment in each bound (4.22) by increasing the value x to another slightly larger level x_s . A simple way is based on the so called Bonferroni correction: $x_s \equiv x + \log(|\mathcal{M}|)$.

Proposition 4.2.2. Let $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. If $\mathcal{M}^\circ(\mathbf{x})$ is the set of indices m such that

$$\mathcal{M}^\circ(\mathbf{x}) \stackrel{\text{def}}{=} \begin{cases} \sigma^{-2} \|b_{m,m^*}\|^2 \geq \Delta^+(m - m^*, \mathbf{x}_s), & m > m^*, \\ \sigma^{-2} \|b_{m^*,m}\|^2 \leq \Delta^-(m^* - m, \mathbf{x}_s), & m < m^*, \end{cases} \quad (4.23)$$

for $\mathbf{x}_s = \mathbf{x} + \log(|\mathcal{M}|)$, then

$$\mathbb{P}(\hat{m} \notin \mathcal{M}^\circ(\mathbf{x})) \leq e^{-\mathbf{x}}. \quad (4.24)$$

Proof. By definition of $\mathcal{M}^\circ(\mathbf{x})$ and Proposition 4.2.1

$$\mathbb{P}(\hat{m} \notin \mathcal{M}^\circ(\mathbf{x})) \leq \sum_{m \notin \mathcal{M}^\circ(\mathbf{x})} \mathbb{P}(\hat{m} = m) \leq \sum_{m \notin \mathcal{M}^\circ(\mathbf{x})} \mathbb{P}(\tilde{\mathcal{R}}_m \leq \tilde{\mathcal{R}}_{m^*}) \leq \sum_{m \in \mathcal{M}} e^{-\mathbf{x}_s} \leq e^{-\mathbf{x}}.$$

One can conclude that if the m lies beyond the *insensitivity zone* $\mathcal{M}^\circ(\mathbf{x})$ around m^* , on which the difference $\mathcal{R}_m - \mathcal{R}_{m^*}$ is not sufficiently large, then the event $\hat{m} = m$ is very unlikely. The result (4.24) can be stated in the form that there exists a random set $\Omega(\mathbf{x})$ such that $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - e^{-\mathbf{x}}$, and on this set, it holds

$$\tilde{\mathcal{R}}_m > \tilde{\mathcal{R}}_{m^*}, \quad m \notin \mathcal{M}^\circ(\mathbf{x})$$

and hence $\hat{m} \in \mathcal{M}^\circ(\mathbf{x})$ on $\Omega(\mathbf{x})$.

4.2.4 A bound on the excess

Introduce another random set $\Omega_0(\mathbf{x})$ such that

$$\begin{aligned} \|\sigma^{-1} \Pi_{m,m^*} \varepsilon\|^2 &\leq \mathfrak{z}^+(m - m^*, \mathbf{x}_s), & m > m^*, \\ \|\sigma^{-1} \Pi_{m^*,m} \varepsilon\|^2 &\geq \mathfrak{z}^-(m^* - m, \mathbf{x}_s), & m < m^*. \end{aligned} \quad (4.25)$$

Lemma 4.2.3 implies that $\mathbb{P}(\Omega_0(\mathbf{x})) \geq 1 - e^{-\mathbf{x}}$.

The next question is what happens if $\hat{m} \in \mathcal{M}^\circ(\mathbf{x})$ and how big this set is. We will try to bound the loss difference $\varrho(\tilde{\mathbf{f}}_m, \mathbf{f}^*) - \varrho(\tilde{\mathbf{f}}_{m^*}, \mathbf{f}^*)$. It follows from (4.2) that for $m > m^*$

$$\sigma^{-2} \{ \varrho(\tilde{\mathbf{f}}_m, \mathbf{f}^*) - \varrho(\tilde{\mathbf{f}}_{m^*}, \mathbf{f}^*) \} = -\|\sigma^{-1} b_{m,m^*}\|^2 + \|\sigma^{-1} \Pi_{m,m^*} \varepsilon\|^2.$$

This implies for $m \in \mathcal{M}_+(\mathbf{x})$ by (4.23) and (4.17) on $\Omega_0(\mathbf{x})$

$$\begin{aligned} \sigma^{-2} \{ \varrho(\tilde{\mathbf{f}}_m, \mathbf{f}^*) - \varrho(\tilde{\mathbf{f}}_{m^*}, \mathbf{f}^*) \} &\leq -\Delta^+(m - m^*, \mathbf{x}_s) + \mathfrak{z}^+(m - m^*, \mathbf{x}_s) \\ &\leq \mathfrak{z}^+(m - m^*, \mathbf{x}_s) - \mathfrak{z}^-(m - m^*, \mathbf{x}_s) + 2z_1(\mathbf{x}_s)\sqrt{m - m^*} \\ &= z^\pm(m - m^*, \mathbf{x}_s) + 2z_1(\mathbf{x}_s)\sqrt{m - m^*}; \end{aligned} \quad (4.26)$$

here $z^\pm(k, \mathbf{x}) = \mathfrak{z}^+(k, \mathbf{x}) - \mathfrak{z}^-(k, \mathbf{x})$.

Now we consider the similar difference for the parameter m from the insensitivity zone $\mathcal{M}_-(\mathbf{x})$ with $m < m^*$. It holds

$$\sigma^{-2}\{\varrho(\tilde{\mathbf{f}}_m, \mathbf{f}^*) - \varrho(\tilde{\mathbf{f}}_{m^*}, \mathbf{f}^*)\} = \|\sigma^{-1}b_{m^*, m}\|^2 - \|\sigma^{-1}\Pi_{m^*, m}\boldsymbol{\varepsilon}\|^2.$$

This implies for $m \in \mathcal{M}^\circ(\mathbf{x})$ by (4.23) and (4.18) on $\Omega_0(\mathbf{x})$

$$\begin{aligned} & \sigma^{-2}\{\varrho(\tilde{\mathbf{f}}_m, \mathbf{f}^*) - \varrho(\tilde{\mathbf{f}}_{m^*}, \mathbf{f}^*)\} \\ & \leq \|\sigma^{-1}b_{m^*, m}\|^2 - \|\sigma^{-1}\Pi_{m^*, m}\boldsymbol{\varepsilon}\|^2 \\ & \leq \Delta^-(m^* - m, \mathbf{x}_s) - \mathfrak{z}^-(m^* - m, \mathbf{x}_s) \\ & \leq \mathfrak{z}^+(m^* - m, \mathbf{x}_s) - \mathfrak{z}^-(m^* - m, \mathbf{x}_s) + 2z_1(\mathbf{x}_s)\sqrt{m^* - m} \\ & = z^\pm(m^* - m, \mathbf{x}_s) + 2z_1(\mathbf{x}_s)\sqrt{m^* - m}. \end{aligned} \tag{4.27}$$

Now we can summarize. Define the radius $R = R(\mathcal{M}^\circ(\mathbf{x}))$ of the set $\mathcal{M}^\circ(\mathbf{x})$:

$$R \stackrel{\text{def}}{=} \max_{m \in \mathcal{M}^\circ(\mathbf{x})} |m - m^*|.$$

Theorem 4.2.1. *Let \hat{m} be defined by (4.6) and m^* be the oracle choice from (4.4). Suppose that $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Let $\mathbf{x}_s = \mathbf{x} + \log(|\mathcal{M}|)$. For the insensitivity zone $\mathcal{M}^\circ(\mathbf{x})$ from (4.23), it holds $\hat{m} \in \mathcal{M}^\circ(\mathbf{x})$ on a random set $\Omega(\mathbf{x})$ with $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - e^{-x}$. Moreover, on a random set $\Omega_0(\mathbf{x})$ with $\mathbb{P}(\Omega_0(\mathbf{x})) \geq 1 - 2e^{-x}$*

$$\sigma^{-2}\{\varrho(\tilde{\mathbf{f}}_{\hat{m}}, \mathbf{f}^*) - \varrho(\tilde{\mathbf{f}}_{m^*}, \mathbf{f}^*)\} \leq z^\pm(R, \mathbf{x}_s) + 2z_1(\mathbf{x}_s)\sqrt{R}.$$

Proof. The result follows from (4.26) and (4.27) by monotonicity of the function $z^\pm(k, \mathbf{x})$ in k and \mathbf{x} .

In view of $z^\pm(R, \mathbf{x}_s) \asymp \sqrt{R \mathbf{x}_s}$, we conclude that the data-driven choice of the parameter m leads to additional loss of order $\sigma^2 \sqrt{R \mathbf{x}_s}$. One can say that the model selection based on unbiased risk estimation works well if the size of the zone of insensitivity $R = R(\mathcal{M}^\circ(\mathbf{x}))$ is not too large compared with the loss and risk of the oracle $\tilde{\mathbf{f}}_{m^*}$.

Note that the loss $\varrho(\tilde{\mathbf{f}}_{m^*}, \mathbf{f}^*)$ fulfills

$$\varrho(\tilde{\mathbf{f}}_{m^*}, \mathbf{f}^*) = \|b_{m^*}\|^2 + \|\Pi_{m^*}\boldsymbol{\varepsilon}\|^2 \geq \|\Pi_{m^*}\boldsymbol{\varepsilon}\|^2.$$

By Lemma 4.2.3, it is of order m^* .

Exercise 4.2.5. Let $\Omega_0(\mathbf{x})$ be a random set on which (4.25) holds. Show that $\mathbb{P}(\Omega_0(\mathbf{x})) \geq 1 - e^{-x}$ and for every m , the loss $\varrho(\tilde{\mathbf{f}}_m, \mathbf{f}^*)$ satisfies on $\Omega_0(\mathbf{x})$

$$\sigma^{-2} \varrho(\tilde{\mathbf{f}}_m, \mathbf{f}^*) \geq \mathfrak{z}^-(m, \mathbf{x}).$$

For the case when the value $R = \max_{m \in \mathcal{M}^\circ(\mathbf{x})} |m - m^*|$ is small relative to m^* , the loss of the estimate $\hat{\mathbf{f}} = \tilde{\mathbf{f}}_{\hat{m}}$ corresponding to the data-driven selector \hat{m} is not significantly larger than the loss of the oracle estimate $\tilde{\mathbf{f}}_{m^*}$. Unfortunately, the set $\mathcal{M}^\circ(\mathbf{x})$ can be very large if the risk \mathcal{R}_m is a flat function of m . The extreme case is given by the so called “noise reproducing” model. This model is described by the equations $|\theta_j^*|^2 \equiv \sigma^2$; see (4.3). One can easily check that

$$\begin{aligned} \|b_{m,m^*}\|^2 &\equiv \sigma^2(m - m^*), & m > m^*, \\ \|b_{m^*,m}\|^2 &\equiv \sigma^2(m^* - m), & m < m^*. \end{aligned}$$

This implies that the risk function \mathcal{R}_m is constant in m , and therefore, the set $\mathcal{M}^\circ(\mathbf{x})$ coincides with the whole set \mathcal{M} .

Exercise 4.2.6. Build an example in which the radius R is twice as large as the oracle risk \mathcal{R}_{m^*} .

4.3 The approach based on multiple testing. “Smallest accepted” rule

This section discusses the alternative approach to model selection based on the idea of multiple testing. Let m^* be a good choice in the sense of “bias-variance trade-off”. Now we aim to develop a procedure that would treat m^* as a good choice with a high probability. We interpret the model selection procedure as pairwise comparison: the “oracle” model wins in terms of the risk against all other models:

$$\mathcal{R}_{m^*} \leq \mathcal{R}_m, \quad m \neq m^*.$$

The selector \hat{m} suggests to apply the model which wins in term of the unbiased risk estimate $\tilde{\mathcal{R}}_m$:

$$\tilde{\mathcal{R}}_{\hat{m}} \leq \tilde{\mathcal{R}}_m, \quad m \neq \hat{m}.$$

Now we reconsider this approach in terms of hypothesis testing.

4.3.1 A LR test

Our null hypothesis will be that a model-candidate m° is “good” in the sense that it delivers a kind of bias-variance trade-off. This means that there is no reason for considering a larger model: the bias improvement will not be compensated by increase of model complexity. Now we interpret this check as a test of the model-candidate m° against any larger model $m > m^\circ$. The null hypothesis H_{m° means that $\theta_j^* \equiv 0$ for all $j > m^\circ$. The alternative is that there are significant coefficients θ_j^* for $m^\circ < j \leq m$.

Define

$$\Theta_m \stackrel{\text{def}}{=} \{\boldsymbol{\theta} = (\theta_1, \dots, \theta_m, 0, \dots, 0)^\top \in I\!\!R^p\}.$$

For the log-likelihood function $L(\boldsymbol{\theta}) = -(2\sigma^2)^{-1}\|\mathbf{Y} - \Psi^\top \boldsymbol{\theta}\|^2$, the likelihood ratio test statistic reads as

$$\begin{aligned} T_{m,m^\circ} &\stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta} \in \Theta_m} L(\boldsymbol{\theta}) - \sup_{\boldsymbol{\theta} \in \Theta_{m^\circ}} L(\boldsymbol{\theta}) = \frac{1}{2\sigma^2} (\|\mathbf{Y} - \Psi^\top \tilde{\boldsymbol{\theta}}_{m^\circ}\|^2 - \|\mathbf{Y} - \Psi^\top \tilde{\boldsymbol{\theta}}_m\|^2) \\ &= \frac{1}{2\sigma^2} (\|\mathbf{Y} - \Pi_{m^\circ} \mathbf{Y}\|^2 - \|\mathbf{Y} - \Pi_m \mathbf{Y}\|^2) = \frac{1}{2\sigma^2} \|\Pi_{m,m^\circ} \mathbf{Y}\|^2. \end{aligned}$$

We aim to calibrate this test in a way that the null hypothesis of no significant bias b_{m,m° between m° and m is not rejected if the bias is indeed insignificant. Significance can be measured by the energy $\sigma^{-2}\|b_{m,m^\circ}\|^2$ of this bias. Namely, we say that the bias b_{m,m° is not significant if

$$\sigma^{-2}\|b_{m,m^\circ}\|^2 \leq \beta(m - m^\circ) \quad (4.28)$$

for a fixed value β . Remind that the definition of the oracle m^* yields the inequality $\|b_{m,m^*}\|^2 \leq \sigma^2(m - m^*)$ corresponding to $\beta = 1$. We apply Lemma 4.2.4 to choose a critical value for the test statistic T_{m,m° . Define

$$z_\beta(k, \mathbf{x}) \stackrel{\text{def}}{=} \mathfrak{z}^+(k, \beta k; \mathbf{x}),$$

where $\mathfrak{z}^+(k, \Delta; \mathbf{x})$ is the quantile of a non-central chi-squared from (4.12). Lemma 4.2.4 also implies

$$z_\beta(k, \mathbf{x}) \leq \beta k + \mathfrak{z}^+(k, \mathbf{x}) + 2\sqrt{\beta k} z_1(\mathbf{x}). \quad (4.29)$$

These definitions yield the following statement.

Proposition 4.3.1. *Let $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Then $\|b_{m,m^\circ}\|^2 \leq \beta(m - m^\circ)$ implies*

$$I\!\!P(2T_{m,m^\circ} > z_\beta(m - m^\circ, \mathbf{x})) \leq e^{-\mathbf{x}}.$$

Exercise 4.3.1. Prove Proposition 4.3.1.

4.3.2 Multiplicity correction

The hypothesis H_{m° is not rejected if each test T_{m,m° for $m > m^\circ$ does not reject the null. This requires a multiple testing procedure and a correction for multiplicity. A simple way is the Bonferroni correction: one increases each \mathbf{x} by the same value $q_{m^\circ} = \log(|\mathcal{M}(m^\circ)|) = \log(p - m^\circ)$. Here $\mathcal{M}(m^\circ) = \{m \in \mathcal{M}: m > m^\circ\}$. Then

$$\begin{aligned} \mathbb{P} \left(\bigcup_{m \in \mathcal{M}(m^\circ)} \{2T_{m,m^\circ} > z_\beta(m - m^\circ, \mathbf{x} + q_{m^\circ})\} \right) \\ \leq \sum_{m \in \mathcal{M}(m^\circ)} e^{-\mathbf{x}-q_{m^\circ}} = |\mathcal{M}(m^\circ)| \exp\{-\mathbf{x} - \log(|\mathcal{M}(m^\circ)|)\} = e^{-\mathbf{x}}. \end{aligned}$$

However, the Bonferroni correction is known to be rather conservative especially if the test statistics T_{m,m° are correlated for different m . This is exactly the case under consideration. Another more careful way to choose the correction q_{m° is based on calibration for one special model.

First we consider the special case $\beta = 0$. Then the null hypothesis means that $\theta_j^* \equiv 0$ for all $j > m^\circ$. In this situation the relative bias $b_{m,m^\circ} = \Pi_{m,m^\circ} \mathbf{f}^*$ is equal to zero, and does not depend on the first m° coefficients θ_j^* for $j \leq m^\circ$. This allows to define q_{m° by the condition

$$\mathbb{P}_0 \left(\bigcup_{m \in \mathcal{M}(m^\circ)} \{2T_{m,m^\circ} > \mathfrak{z}^+(m - m^\circ, \mathbf{x} + q_{m^\circ})\} \right) = e^{-\mathbf{x}}, \quad (4.30)$$

where \mathbb{P}_0 is the measure corresponding to zero signal $\theta_j^* \equiv 0$, and $\mathfrak{z}^+(k, \mathbf{x})$ is the upper quantile of the χ_k^2 from Lemma 4.2.3. The meaning of (4.30) is that each particular test based on the test statistic $2T_{m,m^\circ}$ is performed at a higher level $e^{-\mathbf{x}-q_{m^\circ}} = A^{-1}e^{-\mathbf{x}}$ with $A = e^{q_{m^\circ}}$. The value q_{m° is the smallest number providing the familywise error probability $e^{-\mathbf{x}}$. The Bonferroni correction uses $A = \#\{\text{set of hypotheses}\}$ but this choice is conservative especially if the test statistics are strongly dependent.

In the case of β positive, consider another special “frontier” model $\mathbf{f}^* = \Psi^\top \boldsymbol{\theta}^*$ with the vector of parameters $\boldsymbol{\theta}^*$ with the equalities $\sigma^{-2} \|b_{m,m^\circ}\|^2 = \beta(m - m^\circ)$ in place of inequalities in (4.28).

Exercise 4.3.2. For the case of an orthonormal design Ψ , find a vector $\boldsymbol{\theta}_\beta^*$ for which $\sigma^{-2} \|b_{m,m^\circ}\|^2 \equiv \beta(m - m^\circ)$.

Now we calibrate the critical values z_β for this special extreme model. Namely, we fix the smallest value $q_{m^\circ} = q_{m^\circ}(\beta)$ for which holds

$$\mathbb{P}_{\boldsymbol{\theta}_\beta^*} \left(\bigcup_{m \in \mathcal{M}_+(m^\circ)} \{2T_{m,m^\circ} > z_\beta(m - m^\circ, \mathbf{x} + q_{m^\circ})\} \right) = e^{-\mathbf{x}}. \quad (4.31)$$

Suppose that such correction $q_{m^\circ} = q_{m^\circ}(\beta)$ is defined for each $m^\circ \in \mathcal{M}$.

Exercise 4.3.3. Consider the frontier model of Exercise 4.3.2. Let $q_{m^\circ}(\beta)$ be the corresponding value for the condition (4.31). Check that for each m° ,

$$q_{m^\circ}(\beta) \geq q_{m^\circ+1}(\beta)$$

yielding

$$\mathbf{z}_\beta(k, \mathbf{x} + q_{m^\circ}) \geq \mathbf{z}_\beta(k, \mathbf{x} + q_{m^\circ-1}).$$

Denote

$$\mathbf{x}_{m^\circ} = \mathbf{x} + q_{m^\circ}.$$

The suggested procedure selects the smallest m° which is accepted by the multiple test H_{m° against H_m for all $m > m^\circ$. This acceptance rule for the null hypothesis H_{m° reads as follows:

$$2T_{m,m^\circ} \leq \mathbf{z}_\beta(m - m^\circ, \mathbf{x}_{m^\circ}), \quad \forall m > m^\circ.$$

Now the “smallest accepted” procedure can be stated in the following form:

$$\begin{aligned} \hat{m} &\stackrel{\text{def}}{=} \min \{m^\circ \in \mathcal{M}: m^\circ \text{ is accepted}\} \\ &= \min \left\{ m^\circ \in \mathcal{M}: \max_{m \in \mathcal{M}(m^\circ)} \{2T_{m,m^\circ} - \mathbf{z}_\beta(m - m^\circ, \mathbf{x}_{m^\circ})\} \leq 0 \right\}. \end{aligned} \quad (4.32)$$

4.3.3 Definition of the oracle and propagation property

Further we aim at establishing the oracle inequality for this method. It is desirable to show that the data driven selector \hat{m} behaves as good as the oracle one. As the procedure does no rely on the quadratic risk we slightly extend the oracle definition. Let m^* be the oracle value which is now defined as the smallest value m^* satisfying the conditions $\sigma^{-2} \|b_{m,m^*}\|^2 \leq \beta(m - m^*)$ for $m > m^*$:

$$m^* \stackrel{\text{def}}{=} \operatorname{argmin} \{m: \sigma^{-2} \|b_{m,m^*}\|^2 \leq \beta(m - m^*), m > m^*\}. \quad (4.33)$$

This definition follows the structure of the null hypothesis considered in the procedure. For $\beta = 1$ it is consistent with the oracle definition based on the risk minimization. Now we aim at establishing the oracle property: the data-driven selector \hat{m} behaves essentially as good as the oracle choice m^* . The study is done in two steps. The first step is to check that the model m^* will be accepted with a high probability. This would mean that the selected model \hat{m} satisfies $\hat{m} \leq m^*$. The second step is in applying the test $T_{m^*, \hat{m}}$ for this situation.

Theorem 4.3.1. Let m^* be defined by (4.33). Let also the selector \hat{m} be calibrated by (4.31). Then

$$\mathbb{P}(m^* \text{ is not accepted}) \leq e^{-x}. \quad (4.34)$$

Proof. We use the following fact.

Lemma 4.3.1. Let $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Let m° and x be fixed.

- The values $\theta_1^*, \dots, \theta_{m^\circ}^*$ do not enter in the distribution of T_{m,m° .
- Consider the set $F_\beta(m^\circ)$ of all vectors f^* with $\sigma^{-2} \|b_{m,m^\circ}\|^2 \leq \beta(m - m^\circ)$. Within this set, the probability

$$\mathbb{P}\left(\max_{m > m^\circ} \{2T_{m,m^\circ} - z_\beta(m - m^\circ, x)\} > 0\right)$$

is maximized for the case $\sigma^{-2} \|b_{m,m^\circ}\|^2 \equiv \beta(m - m^\circ)$.

To be done: Proof of Lemma 4.3.1

Now we check (4.34). The definition of m^* ensures that $f^* \in F_\beta(m^*)$, that is, (4.33) is fulfilled. The last statement of Lemma 4.3.1 ensures that if the value q_{m^*} is fixed for the extreme model with $\sigma^{-2} \|b_{m,m^*}\|^2 \equiv \beta(m - m^*)$, then

$$\mathbb{P}(m^* \text{ is not accepted}) = \mathbb{P}\left(\max_{m > m^*} \{2T_{m,m^*} - z_\beta(m - m^*, x_{m^*})\} > 0\right) \leq e^{-x}$$

because the similar inequality (with $m^\circ = m^*$) is fulfilled for the extreme model.

The “propagation” property (4.34) is very important and it is usually not fulfilled for classical procedures like unbiased risk estimation (SURE). The advantage of the proposed approach is that this property is in fact intrinsic for the method and is postulated by the calibration step.

4.3.4 A bound on the loss

It remains to clarify the situation if the selected model is smaller than m^* . This probability is not small but we can control the difference of the losses similarly to the SURE procedure. First we check that the new procedure is also able to identify a significant bias $b_{m^*,m}$ measured by $\sigma^{-2} \|b_{m^*,m}\|^2$.

Proposition 4.3.2. Define $x_s \stackrel{\text{def}}{=} x + \log(m^*)$ and let

$$\mathcal{M}^\circ(x) = \left\{m \leq m^* : \mathfrak{z}^-(m^* - m, \Delta_m; x_s) \leq z_\beta(m^* - m, x_s)\right\}$$

with $\Delta_m \stackrel{\text{def}}{=} \sigma^{-2} \|b_{m^*,m}\|^2$. Then

$$\mathbb{P}(\hat{m} \notin \mathcal{M}^\circ(\mathbf{x})) \leq 2e^{-x}.$$

Moreover, $m \in \mathcal{M}^\circ(\mathbf{x})$ is impossible if the bias $\Delta_m \stackrel{\text{def}}{=} \sigma^{-2}\|b_{m^*,m}\|^2$ fulfills with $k = m^* - m$ and with $z^\pm(k, \mathbf{x}_s) = \mathfrak{z}^+(k, \mathbf{x}_s) - \mathfrak{z}^-(k, \mathbf{x}_s)$

$$\{\Delta_m^{1/2} - z_1(\mathbf{x}_s)\}^2 \geq z^\pm(k, \mathbf{x}_s) + \{\sqrt{\beta k} + z_1(\mathbf{x}_s)\}^2 \quad (4.35)$$

or a stronger condition

$$\Delta_m^{1/2} \geq \sqrt{z^\pm(k, \mathbf{x}_s)} + \sqrt{\beta k} + 2z_1(\mathbf{x}_s). \quad (4.36)$$

Proof. We already proved that $\mathbb{P}(\hat{m} > m^*) \leq e^{-x}$. It remains to study the case $\hat{m} < m^*$. Lemma 4.2.4 yields for $m < m^*$

$$2T_{m^*,m} = \sigma^{-2}\|\Pi_{m^*,m}\mathbf{Y}\|^2 \geq \mathfrak{z}^-(m^* - m, \Delta_m; \mathbf{x})$$

with the probability at least $1 - e^{-x}$. This implies a uniform bound

$$2T_{m^*,m} = \sigma^{-2}\|\Pi_{m^*,m}\mathbf{Y}\|^2 \geq \mathfrak{z}^-(m^* - m, \Delta_m; \mathbf{x}_s), \quad m < m^*$$

on a random set $\Omega_0(\mathbf{x})$ of probability at least $1 - e^{-x}$. For $m \notin \mathcal{M}^\circ(\mathbf{x})$, this implies

$$2T_{m^*,m} > z_\beta(m^* - m, \mathbf{x}_m),$$

which makes the event “ m is accepted” impossible on $\Omega_0(\mathbf{x})$ for $m \notin \mathcal{M}^\circ(\mathbf{x})$.

The bound $\mathfrak{z}^-(m^* - m, \Delta_m; \mathbf{x}_s) > z_\beta(m^* - m, \mathbf{x}_m)$ is implicit in Δ_m . To make it explicit, we use the lower bound (4.13) of Lemma 4.2.4 and (4.29). In view of $\mathbf{x}_s \geq \mathbf{x}_m$ for all $m < m^*$, with $k = m^* - m$, the following inequality is sufficient for checking that $m \notin \mathcal{M}^\circ(\mathbf{x})$:

$$\Delta_m + \mathfrak{z}^-(k, \mathbf{x}_s) - 2\Delta_m^{1/2}z_1(\mathbf{x}_s) \geq \beta k + \mathfrak{z}^+(k, \mathbf{x}_s) + 2\sqrt{\beta k}z_1(\mathbf{x}_s), \quad (4.37)$$

which yields (4.35). The latter can be slightly simplified by using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$: the bound (4.37) holds if

$$\Delta_m^{1/2} - z_1(\mathbf{x}_s) \geq \sqrt{z^\pm(m^* - m, \mathbf{x}_s)} + \sqrt{\beta(m^* - m)} + z_1(\mathbf{x}_s)$$

which coincides with (4.36).

We conclude that the “smallest accepted” rule leads to the choice of \hat{m} in the insensitivity zone $\mathcal{M}^\circ(\mathbf{x})$ with a high probability. It remains to bound the loss difference $\varrho(\tilde{\mathbf{f}}_m, \mathbf{f}^*) - \varrho(\tilde{\mathbf{f}}_{m^*}, \mathbf{f}^*)$ for $m \in \mathcal{M}^\circ(\mathbf{x})$. We use that for $m < m^*$

$$\sigma^{-2}\{\varrho(\tilde{\mathbf{f}}_m, \mathbf{f}^*) - \varrho(\tilde{\mathbf{f}}_{m^*}, \mathbf{f}^*)\} = \|\sigma^{-1}b_{m^*, m}\|^2 - \|\sigma^{-1}\Pi_{m^*, m}\varepsilon\|^2.$$

By Lemma 4.2.3, the stochastic form $\|\sigma^{-1}\Pi_{m^*, m}\varepsilon\|^2$ can be bounded from below by $\mathfrak{z}^-(m^* - m, \mathbf{x})$ with probability $1 - e^{-\mathbf{x}}$. This implies a uniform probability bound $\|\sigma^{-1}\Pi_{m^*, m}\varepsilon\|^2 \geq \mathfrak{z}^-(m^* - m, \mathbf{x}_s)$ with $\mathbf{x}_s = \mathbf{x} + \log(m^*)$ for $m < m^*$. For $m \in \mathcal{M}^\circ(\mathbf{x})$, this implies

$$\sigma^{-2}\{\varrho(\tilde{\mathbf{f}}_m, \mathbf{f}^*) - \varrho(\tilde{\mathbf{f}}_{m^*}, \mathbf{f}^*)\} \leq \Delta_m - \mathfrak{z}^-(m^* - m, \mathbf{x}_s)$$

on a random set of probability at least $1 - 3e^{-\mathbf{x}}$, where Δ_m follows the bound (4.35) or (4.36).

We state the following result:

Theorem 4.3.2. *Let $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. For the selector \hat{m} from (4.32) and the oracle m^* from (4.33), it holds*

$$\mathbb{P}(\hat{m} < m^*) \leq e^{-\mathbf{x}},$$

and for the insensitivity zone $\mathcal{M}^\circ(\mathbf{x})$, it follows on a random set $\Omega(\mathbf{x})$ with $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - 3e^{-\mathbf{x}}$

$$\begin{aligned} \sigma^{-2}\{\varrho(\tilde{\mathbf{f}}_m, \mathbf{f}^*) - \varrho(\tilde{\mathbf{f}}_{m^*}, \mathbf{f}^*)\} &\leq \max_{m \in \mathcal{M}^\circ(\mathbf{x})}\{\Delta_m - \mathfrak{z}^-(m^* - m, \mathbf{x}_s)\} \\ &\leq \max_{m \in \mathcal{M}^\circ(\mathbf{x})}\left\{\left(\sqrt{z^\pm(m^* - m, \mathbf{x}_s)} + \sqrt{\beta(m^* - m)} + 2z_1(\mathbf{x}_s)\right)^2 - \mathfrak{z}^-(m^* - m, \mathbf{x}_s)\right\}. \end{aligned}$$

4.3.5 Role of β

The SURE method corresponds to $\beta = 1$. This choice is natural as long the squared risk is concerned. However, the procedure is meaningful for every $\beta > 0$. It even applies for $\beta = 0$. This is a very special case leading to a testing problem instead of estimation.

In general, consider the situation when we apply the SURE procedure with the noise level σ which is misspecified and the true noise level is $\sigma_0^2 \leq \sigma^2$. Then the true risk \mathcal{R}_m is

$$\mathcal{R}_m = \sigma_0^2 m + \|b_m\|^2 = \beta^{-1} \sigma^2 m + \|b_m\|^2$$

for $\beta = \sigma^2/\sigma_0^2$. The use of the wrong noise level is equivalent to applying the procedure with such β .

The choice of a small β results in a strong condition on the bias, so the oracle choice will be shifted towards larger (more complex) models. With $\beta > 1$, one allows for more space for the bias, the procedure becomes more robust and tends to oversmooth the model, that is, to select \hat{m} which is smaller than the risk minimizer m^* .

Ordered model selection for linear smoothers

Model selection is one of the key topics in mathematical statistics. A choice between models of differing complexity can often be viewed as a trade-off between overfitting the data by choosing a model which has too many degrees of freedom and smoothing out the underlying structure in the data by choosing a model which has too few degrees of freedom. This trade-off which shows up in most methods as the classical bias-variance trade-off is at the heart of every model selection method (as for example in unbiased risk estimation, [Kneip \(1994\)](#) or in penalized model selection, [Barron et al. \(1999a\)](#), [Massart \(2007\)](#)). This is also the case in Lepski's method, [Lepski \(1990\)](#), [Lepski \(1991\)](#), [Lepski \(1992\)](#), [Lepski and Spokoiny \(1997\)](#), [Lepski et al. \(1997\)](#), [Birgé \(2001\)](#) and risk hull minimization, [Cavalier and Golubev \(2006\)](#). Many of these methods allow their strongest theoretical results only for highly idealized situations (for example sequence space models), are very specific to the type of problem under consideration (for instance, signal or functional estimation), require to know the noise behavior (like homogeneity) and the exact noise level. Moreover, they typically involve an unwieldy number of calibration constants whose choice is crucial to the applicability of the method and is not addressed by the theoretical considerations. For instance, any Lepski-type method requires to fix a numerical constant in the definition of the threshold, the theoretical results only apply if this constant is sufficiently large while the numerical results benefit from the choice of a rather small constant. [Spokoiny and Vial \(2009\)](#) offered a propagation approach to calibration of Lepski's method in the case of the estimation of a one-dimensional quantity of interest. However, the proposal still requires the exact knowledge of the noise level and only applies to linear functional estimation. A similar approach has been applied to local constant density estimation with sup-norm risk in [Gach et al. \(2013\)](#) and to local quantile estimation in [Spokoiny et al. \(2013\)](#).

In the case of unknown but homogeneous noise, generalized cross validation can be used instead of unbiased risk estimation method. For the penalized model selection, recently a number of proposals appeared to apply one or another resampling method. [Arlot](#)

(2009) suggested the use of resampling methods for the choice of an optimal penalization, following the framework of penalized model selection, Barron et al. (1999a), Birgé and Massart (2007a). The validity of a bootstrapping procedure for Lepski's method has also been studied in Chernozhukov et al. (2014) with new innovative technical tools with applications to honest adaptive confidence bands.

An alternative approach to adaptive estimation is based on aggregation of different estimates; see Goldenshluger (2009) and Dalalyan and Salmon (2012) for an overview of the existing results. However, the proposed aggregation procedures either require two independent copies of the data or involves a data splitting for estimating the noise variance. Each of these requirements is very restrictive for practical applications.

Another point to mention is that the majority of the obtained results on adaptive estimation focus on the quality of estimating the unknown response, that is, the loss is measured by the difference between the true response and its estimate. At the same time, inference questions like confidence estimation would require to know some additional information about the right model parameter. Only few results address the issue of estimating the true (oracle) model. Moreover, there are some negative results showing that a construction of adaptive honest confidence sets is impossible without special conditions like self-similarity; see, e.g. Gine and Nickl (2010).

This chapter aims at developing a unified approach to the problem of ordered model selection with the focus on the quality of model selection rather than on accuracy of adaptive estimation under realistic assumptions on the model. Our setup covers linear regression and linear inverse problems, and equally applies to estimation of the whole parameter vectors, a subvector or linear mapping, as well as estimation of a linear functional. The proposed procedure and the theoretical study are also unified and do not distinguish between models and problems. In the case of a linear inverse problem, it is applicable to mild and severely ill-posed problems without prior knowledge of the type and degree of ill-posedness; cf. Tsybakov (2000), Cavalier et al. (2002). Another important issue is that the procedure does not use any prior information about the variance structure of the noise under assumption of minimal Hölder smoothness $1/4$ on the underlying function. The method automatically adjusts the parameters to the underlying possibly heterogeneous noise: the resampling technique allows to achieve the same quality of estimation as if the noise structure were precisely known. Also we allow for a model misspecification: the linear structure of the response can be violated, in this case the procedure adaptively recovers the best linear projection.

Consider a linear model $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ in \mathbb{R}^n for an unknown parameter vector $\boldsymbol{\theta}^* \in \mathbb{R}^p$ and a given $p \times n$ design matrix Ψ . Suppose that a family of linear smoothers $\tilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \mathbf{Y}$ is given, where \mathcal{S}_m is for each $m \in \mathcal{M}$ a given $p \times n$ matrix. We also

assume that this family is ordered by the complexity of the method. The task is to develop a data-based model selector \hat{m} which performs nearly as good as the optimal choice, which depends on the model and is not available. The proposed procedure called the “smallest accepted” (SmA) rule can be viewed as a calibrated Lepski-type method. The idea how the parameters of the method can be tuned, originates from [Spokoiny and Vial \(2009\)](#) and is related to a multiple testing problem. The whole procedure is based on family of pairwise tests, each model is tested against all larger ones. Finally the smallest accepted model is selected. The critical values for this multiple testing procedure are fixed using the so-called *propagation condition*. Theorem 5.1.1 presents finite sample results on the behavior of the proposed selector \hat{m} and the corresponding estimator $\hat{\theta} = \tilde{\theta}_{\hat{m}}$. In particular, it describes a concentration set \mathcal{M}° for the selected index \hat{m} and states an oracle bound for the resulting estimator $\hat{\theta} = \tilde{\theta}_{\hat{m}}$. Usual rate results can be easily derived from these statements. Further results address the important issue called “the payment for adaptation” which can be defined as the gap between oracle and adaptive bounds. Theorem 5.1.2 gives a general description of this quantity. Then we specify the results to important special cases like projection estimation and estimation of a linear functional. It appears, that in some cases the obtained results yield sharp asymptotic bounds. In some other cases they lead to the usual log-price for adaptation. However, all these results require a known noise distribution. Section 5.2 explains how the proposed procedure can be tuned in the case of unknown noise using a bootstrap procedure. We establish explicit error bounds on the accuracy of the bootstrap approximation and show that the procedure with bootstrap tuning does essentially the same job as the ideal procedure designed for the known noise. The study is quite involved because the procedure uses the same data twice for parameter tuning and for model selection.

The chapter is structured as follows. The next section presents the procedure and the results for an idealistic situation when the noise distribution is precisely known. We introduce the SmA selector \hat{m} and explain how it can be calibrated. Then we describe the set of possible \hat{m} -values and establish probabilistic oracle bounds. Section 5.1.6 explains how the method and the results can be extended to the case of a polynomial loss function. The results are also specified to the particular problems of projection and linear functional estimation. Section 5.2 extends the method and the study to the realistic case with unknown heteroscedastic noise by using a resampling technique. The proofs and a detailed study of the bootstrap procedure in the linear Gaussian setup are given in the appendix. We also collect there some useful technical statements for Gaussian quadratic forms and sums of random matrices.

5.1 Model and problem. Known noise variance

This section presents the model selector for the idealistic case when the noise distribution is precisely known. In the next section we explain how the unknown noise structure can be recovered from the data using a resampling technique. First we specify our setup. We consider the following linear Gaussian model:

$$Y_i = \Psi_i^\top \boldsymbol{\theta}^* + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}, \quad i = 1, \dots, n, \quad (5.1)$$

with given design Ψ_1, \dots, Ψ_n in \mathbb{R}^p . We also write this equation in vector form $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon} \in \mathbb{R}^n$, where Ψ is $p \times n$ design matrix and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Below we assume a deterministic design, otherwise one can understand the results conditioned on the design realization.

In what follows, we allow the model (5.1) to be completely misspecified. We mainly assume that the observations Y_i are independent and define the response vector $\mathbf{f}^* = \mathbb{E}\mathbf{Y}$ with entries f_i . Such a model can be written as

$$Y_i = f_i + \varepsilon_i. \quad (5.2)$$

Our study allows that the linear parametric assumption $\mathbf{f}^* = \Psi^\top \boldsymbol{\theta}^*$ is violated, and the underlying noise $\boldsymbol{\varepsilon} = (\varepsilon_i)$ can be heterogeneous and non-Gaussian. However, in this section we assume the noise distribution to be known. The main oracle results of Theorem 5.1.1 below do not require any further conditions on the noise. Some upper bounds on the quantities $\bar{\mathbf{z}}_{m^*}$ entering in the oracle bounds are established under i.i.d. Gaussian noise, but can be easily extended to non-Gaussian heterogeneous noise under moment conditions. For the linear model (5.2), define $\boldsymbol{\theta}^* \in \mathbb{R}^p$ as the vector providing the best linear fit:

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \mathbb{E}\|\mathbf{Y} - \Psi^\top \boldsymbol{\theta}\|^2 = (\Psi\Psi^\top)^{-1}\Psi\mathbf{f}^*.$$

As usual, a pseudo-inversion is assumed if the matrix $\Psi\Psi^\top$ is degenerated.

Below we assume a family $\{\tilde{\boldsymbol{\theta}}_m\}$ of linear estimators $\tilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \mathbf{Y}$ of $\boldsymbol{\theta}^*$ to be given. Typical examples include projection estimation on an m -dimensional subspace or regularized estimation with a regularization parameter α_m , penalized estimators with a quadratic penalty function, etc. To include specific problems like subvector/functional estimation, we also introduce a weighting $q \times p$ -matrix W for some fixed $q \geq 1$ and define quadratic loss and risk with this weighting matrix W :

$$\varrho_m \stackrel{\text{def}}{=} \|W(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2, \quad \mathcal{R}_m \stackrel{\text{def}}{=} \mathbb{E}\|W(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2.$$

Of course, the loss and the risk depend on the choice of W . We do not indicate this dependence explicitly but it is important to keep in mind the role of W in the definition of ϱ_m . Typical examples of W are as follows.

Estimation of the whole vector $\boldsymbol{\theta}^$*

Let W be the identity matrix $W = I_p$ with $q = p$. This means that the estimation loss is measured by the usual squared Euclidean norm $\|\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\|^2$.

Prediction

Let W be the square root of the total Fisher information matrix $\mathbb{F} = \sigma^{-2}\Psi\Psi^\top$, that is, $W^2 = \mathbb{F}$. Such a type of loss is usually referred to as *prediction loss* because it measures the fit and the prediction ability of the true model by the model with the parameter $\boldsymbol{\theta}$.

Semiparametric estimation

Let the target of estimation not be the whole vector $\boldsymbol{\theta}^*$ but some subvector $\boldsymbol{\theta}_0^*$ of dimension q . The estimate $\Pi\tilde{\boldsymbol{\theta}}_m$ is called the *profile maximum likelihood estimate*. The matrix W can be defined as the projector Π_0 on the $\boldsymbol{\theta}_0^*$ subspace. The corresponding loss is equal to the squared Euclidean norm in this subspace:

$$\varrho_m = \|\Pi_0(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2.$$

Alternatively, one can select W^2 as the efficient Fisher information matrix defined by the relation

$$W^2 \stackrel{\text{def}}{=} \mathbb{I}_0 = (\Pi_0 \mathbb{F}^{-1} \Pi_0^\top)^{-1}.$$

Linear functional estimation

The choice of the weighting matrix W can be adjusted to address the problem of estimating some functionals of the whole parameter $\boldsymbol{\theta}^*$. For instance, in the regression problem $\mathbb{E}Y_i = f(X_i)$ with the Fourier expansion the target function f can be represented as

$$f(x) = \sum_{j \geq 0} \theta_j^* \psi_j(x) = \sum_{j \geq 0} \{\theta_{2j}^* \cos(2\pi jx) + \theta_{2j+1}^* \sin(2\pi jx)\}.$$

The value of this function at zero coincides with the functional $f(0) = \sum_j \theta_{2j}^*$. The first derivative of this function leads to the functional $f'(0) = 2\pi \sum_{j \geq 0} j \theta_{2j+1}^*$.

In all cases, the most important feature of the estimators $\tilde{\boldsymbol{\theta}}_m$ is *linearity*. It greatly simplifies the study of their properties including the prominent bias-variance decomposition of the risk of $\tilde{\boldsymbol{\theta}}_m$. Namely, for the model (5.2) with $\mathbb{E}\boldsymbol{\varepsilon} = 0$, it holds

$$\begin{aligned}
E\tilde{\theta}_m &= \theta_m^* = \mathcal{S}_m f^*, \\
\mathcal{R}_m &= \|W(\theta_m^* - \theta^*)\|^2 + \text{tr}\{W\mathcal{S}_m \text{Var}(\varepsilon) \mathcal{S}_m^\top W^\top\} \\
&= \|W(\mathcal{S}_m - \mathcal{S})f^*\|^2 + \text{tr}\{W\mathcal{S}_m \text{Var}(\varepsilon) \mathcal{S}_m^\top W^\top\}.
\end{aligned} \tag{5.3}$$

The optimal choice of the parameter m can be defined by risk minimization:

$$m^* \stackrel{\text{def}}{=} \underset{m \in \mathcal{M}}{\operatorname{argmin}} \mathcal{R}_m.$$

The *model selection* problem can be described as the choice of m by data which *mimics the oracle*, that is, we aim at constructing a selector \hat{m} leading to the adaptive estimate $\hat{\theta} = \tilde{\theta}_{\hat{m}}$ with properties similar to the oracle estimate $\tilde{\theta}_{m^*}$.

Below we discuss the *ordered case*. The parameter $m \in \mathcal{M}$ is treated as complexity of the method $\tilde{\theta}_m$. In some cases the set \mathcal{M} of possible m choices can be countable and/or continuous and even unbounded. For simplicity of presentation, we assume that \mathcal{M} is a finite set of positive numbers, $|\mathcal{M}|$ stands for its cardinality. Typical examples are given by the number of terms in the Fourier expansion, or by the bandwidth in the kernel smoothing. In general, complexity can be naturally expressed via the variance of the stochastic term of the estimator $\tilde{\theta}_m$: the larger m , the larger is the variance $\text{Var}(W\tilde{\theta}_m)$. In the case of projection estimation with m -dimensional projectors \mathcal{S}_m , this variance is linear in m , $\text{Var}(\tilde{\theta}_m) = \sigma^2 m$. In general, dependence of the variance term on m may be more complicated but the monotonicity constraint (5.4) has to be preserved.

Further, it is implicitly assumed that the bias term $\|W(\theta^* - \theta_m^*)\|^2$ becomes small when m increases. The smallest index $m = m_0$ corresponds to the simplest (zero) model with probably a large bias, while m large ensures a good approximation quality $\theta_m^* \approx \theta^*$ and a small bias at cost of a big complexity measured by the variance term. In the case of projection estimation, the bias term in (5.3) describes the accuracy of approximating the response f^* by an m -dimensional linear subspace and this approximation improves as m grows. However, in general, in contrast to the case of projection estimation, one cannot require that the bias term $\|W(\theta^* - \theta_m^*)\|^2$ monotonously decreases with m . One example is given by an estimation-at-a-point problem.

Example 5.1.1. Suppose that a signal θ^* is observed with noise: $Y_i = \theta_j^* + \varepsilon_j$. Consider the set of projection estimates $\tilde{\theta}_m$ on the first m coordinates and the target is $\phi^* \stackrel{\text{def}}{=} W\theta = \sum_j \theta_j$. If θ^* is composed of alternating blocks of 1's and -1's with equal length, then the bias $|\phi^* - \phi_m^*|$ for $\phi_m^* = \sum_{j \leq m} \theta_j^*$ is not monotonous in m .

5.1.1 Smallest accepted (SmA) method in ordered model selection

First we recall our setup. Due to the linear structure of the estimators $\tilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \mathbf{Y}$ and of the loss function W , one can consider $\tilde{\boldsymbol{\phi}}_m = \mathcal{K}_m \mathbf{Y}$ with $\mathcal{K}_m = W\mathcal{S}_m: \mathbb{R}^n \rightarrow \mathbb{R}^q$, $m \in \mathcal{M}$, as a family of linear estimators of the q -dimensional target of estimation $\boldsymbol{\phi}^* = W\boldsymbol{\theta}^* = W\mathcal{S}\mathbf{f}^* = \mathcal{K}\mathbf{f}^*$ for $\mathcal{K} = W\mathcal{S}$.

Now we discuss a general approach to model selection problems based on multiple testing. Suppose that the given family $\{\tilde{\boldsymbol{\phi}}_m, m \in \mathcal{M}\}$ of estimators is naturally ordered by their complexity (variance). Due to (5.3), this condition can be written as

$$\mathcal{K}_m \text{Var}(\boldsymbol{\varepsilon}) \mathcal{K}_m^\top \leq \mathcal{K}_{m'} \text{Var}(\boldsymbol{\varepsilon}) \mathcal{K}_{m'}^\top, \quad m' > m. \quad (5.4)$$

One would like to pick up a smallest possible index $m \in \mathcal{M}$ which still provides a reasonable fit. The latter means that the bias component

$$\|b_m\|^2 = \|\boldsymbol{\phi}_m^* - \boldsymbol{\phi}^*\|^2 = \|(\mathcal{K}_m - \mathcal{K})\mathbf{f}^*\|^2$$

in the risk decomposition (5.3) is not significantly larger than the variance

$$\text{tr}\{\text{Var}(\tilde{\boldsymbol{\phi}}_m)\} = \text{tr}\{\mathcal{K}_m \text{Var}(\boldsymbol{\varepsilon}) \mathcal{K}_m^\top\}.$$

If $m^\circ \in \mathcal{M}$ is such a “good” choice, then our ordering assumption yields that a further increase of the index m over m° only increases the complexity (variance) of the method without real gain in the quality of approximation. This latter fact can be interpreted in term of pairwise comparison: whatever $m \in \mathcal{M}$ with $m > m^\circ$ we take, there is no significant bias reduction in using a larger model m instead of m° . This leads to a multiple testing procedure: for each pair $m > m^\circ$ from \mathcal{M} , we consider a hypothesis of no significant bias between the models m° and m , and let τ_{m,m° be the corresponding test. The model m° is accepted if $\tau_{m,m^\circ} = 0$ for all $m > m^\circ$. Finally, the selected model is the “smallest accepted”:

$$\hat{m} \stackrel{\text{def}}{=} \underset{m^\circ \in \mathcal{M}}{\text{argmin}} \{m^\circ: \tau_{m,m^\circ} = 0, \forall m > m^\circ\}.$$

Usually the test τ_{m,m° can be written in the form

$$\tau_{m,m^\circ} = \mathbb{I}\{\mathbb{T}_{m,m^\circ} > z_{m,m^\circ}\}$$

for some *test statistics* \mathbb{T}_{m,m° and for *critical values* z_{m,m° . The information-based criteria like AIC or BIC use the likelihood ratio test statistics $\mathbb{T}_{m,m^\circ} = \sigma^{-2} \|\Psi^\top (\tilde{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_{m^\circ})\|^2$. A great advantage of such tests is that the test statistic \mathbb{T}_{m,m° is pivotal (χ^2 with $m-m^\circ$ degrees of freedom) under the correct null hypothesis, this makes it simple to

compute the corresponding critical values. Below we apply another choice corresponding to Lepski-type procedure and based on the norm of differences $\tilde{\phi}_m - \tilde{\phi}_{m^\circ}$:

$$\mathbb{T}_{m,m^\circ} = \|\tilde{\phi}_m - \tilde{\phi}_{m^\circ}\| = \|\mathcal{K}_{m,m^\circ} \mathbf{Y}\|, \quad \mathcal{K}_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{K}_m - \mathcal{K}_{m^\circ}.$$

The main issue for such a method is a proper choice of the critical values \mathbf{z}_{m,m° . One can say that the procedure is specified by a way of selecting these critical values. Below we offer a novel way of carrying out this choice in a general situation by using a so-called *propagation condition*: if a model m° is “good” it has to be accepted with a high probability. This rule can be seen as an analog of the family-wise level condition in a multiple testing problem. Rejecting a “good” model is the family-wise error of first kind, and this error has to be controlled.

5.1.2 Oracle choice

To specify precisely the meaning of a good model, we use below for each pair $m > m^\circ$ from \mathcal{M} the decomposition

$$\mathbb{T}_{m,m^\circ} = \|\tilde{\phi}_m - \tilde{\phi}_{m^\circ}\| = \|\mathcal{K}_{m,m^\circ} \mathbf{Y}\| = \|\mathcal{K}_{m,m^\circ}(\mathbf{f}^* + \boldsymbol{\varepsilon})\| = \|b_{m,m^\circ} + \boldsymbol{\xi}_{m,m^\circ}\|, \quad (5.5)$$

where with $\mathcal{K}_{m,m^\circ} = \mathcal{K}_m - \mathcal{K}_{m^\circ}$

$$b_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{K}_{m,m^\circ} \mathbf{f}^*, \quad \boldsymbol{\xi}_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{K}_{m,m^\circ} \boldsymbol{\varepsilon}.$$

We also define

$$\mathbf{b}_m \stackrel{\text{def}}{=} \mathcal{K}_m \mathbf{f}^*, \quad \boldsymbol{\xi}_m \stackrel{\text{def}}{=} \mathcal{K}_m \boldsymbol{\varepsilon}.$$

It obviously holds $E\boldsymbol{\xi}_{m,m^\circ} = 0$. Introduce the $q \times q$ -matrix \mathbb{V}_{m,m° as the variance of $\tilde{\phi}_m - \tilde{\phi}_{m^\circ}$:

$$\mathbb{V}_{m,m^\circ} \stackrel{\text{def}}{=} \text{Var}(\tilde{\phi}_m - \tilde{\phi}_{m^\circ}) = \text{Var}(\mathcal{K}_{m,m^\circ} \mathbf{Y}) = \mathcal{K}_{m,m^\circ} \text{Var}(\boldsymbol{\varepsilon}) \mathcal{K}_{m,m^\circ}^\top.$$

If the noise $\boldsymbol{\varepsilon}$ is homogeneous with $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$, it holds

$$\mathbb{V}_{m,m^\circ} = \sigma^2 \mathcal{K}_{m,m^\circ} \mathcal{K}_{m,m^\circ}^\top.$$

Further,

$$\begin{aligned} E\mathbb{T}_{m,m^\circ}^2 &= \|b_{m,m^\circ}\|^2 + E\|\boldsymbol{\xi}_{m,m^\circ}\|^2 = \|b_{m,m^\circ}\|^2 + p_{m,m^\circ}, \\ p_{m,m^\circ} &\stackrel{\text{def}}{=} \text{tr}(\mathbb{V}_{m,m^\circ}) = E\|\boldsymbol{\xi}_{m,m^\circ}\|^2. \end{aligned} \quad (5.6)$$

The bias term $b_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{K}_{m,m^\circ} \mathbf{f}^*$ is significant if its squared norm is competitive with the variance term $\mathbf{p}_{m,m^\circ} = \text{tr}(\mathbb{V}_{m,m^\circ})$. We say that m° is a “good” choice if there is no significant bias b_{m,m° for any $m > m^\circ$. This condition can be quantified in the following “bias-variance trade-off”:

$$\|b_{m,m^\circ}\|^2 \leq \beta^2 \mathbf{p}_{m,m^\circ}, \quad m > m^\circ \quad (5.7)$$

for a given parameter β which controls the bias component in the risk due to decomposition (5.6). Now define the *oracle* m^* as the minimal m° with the property (5.7):

$$m^* \stackrel{\text{def}}{=} \min \left\{ m^\circ : \max_{m > m^\circ} \{ \|b_{m,m^\circ}\|^2 - \beta^2 \mathbf{p}_{m,m^\circ} \} \leq 0 \right\}. \quad (5.8)$$

5.1.3 Tail function, multiplicity correction, critical values z_{m,m°

Now we explain a possible choice of critical values z_{m,m° in the situation when the noise distribution is known. A particular example is the case of Gaussian errors $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Then the distribution of the stochastic component $\boldsymbol{\xi}_{m,m^\circ}$ is known as well. In the Gaussian case, it is $\mathcal{N}(0, \mathbb{V}_{m,m^\circ})$ with the covariance matrix \mathbb{V}_{m,m° . Introduce for each pair $m > m^\circ$ from \mathcal{M} a *tail function* $z_{m,m^\circ}(t)$ of the argument t such that

$$\mathbb{P}(\|\boldsymbol{\xi}_{m,m^\circ}\| > z_{m,m^\circ}(t)) = e^{-t}. \quad (5.9)$$

Here we assume that the distribution of $\|\boldsymbol{\xi}_{m,m^\circ}\|$ is continuous and the value $z_{m,m^\circ}(t)$ is well defined. Otherwise one has to define $z_{m,m^\circ}(t)$ as the smallest value for which the error probability is smaller than e^{-t} .

For checking the propagation condition, we need a uniform in $m > m^\circ$ version of the probability bound (5.9). Let

$$\mathcal{M}^+(m^\circ) \stackrel{\text{def}}{=} \{m \in \mathcal{M} : m > m^\circ\}.$$

Given \mathbf{x} , by $q_{m^\circ} = q_{m^\circ}(\mathbf{x})$ denote the corresponding multiplicity correction:

$$\mathbb{P} \left(\bigcup_{m \in \mathcal{M}^+(m^\circ)} \{ \|\boldsymbol{\xi}_{m,m^\circ}\| \geq z_{m,m^\circ}(\mathbf{x} + q_{m^\circ}) \} \right) = e^{-\mathbf{x}}. \quad (5.10)$$

A simple way of computing the multiplicity correction q_{m° is based on the Bonferroni bound: $q_{m^\circ} = \log(\#\mathcal{M}^+(m^\circ))$. However, it is well known that the Bonferroni bound is very conservative and leads to a large correction q_{m° , especially if the random vectors $\boldsymbol{\xi}_{m,m^\circ}$ are strongly correlated. This is exactly the case under consideration. Note that the joint distribution of the $\boldsymbol{\xi}_{m,m^\circ}$'s is precisely known. This allows to define the correction $q_{m^\circ} = q_{m^\circ}(\mathbf{x})$ just by condition (5.10). Finally we define the critical values z_{m,m° by one more correction for the bias:

$$z_{m,m^\circ} \stackrel{\text{def}}{=} z_{m,m^\circ}(x + q_{m^\circ}) + \beta \sqrt{p_{m,m^\circ}} \quad (5.11)$$

for $p_{m,m^\circ} = \text{tr}(\mathbb{V}_{m,m^\circ})$. This definition still involves two numerical tuning constants x and β . The first value x controls the nominal rejection probability under the null, a usual choice $x = 3$ does a good job in most of cases. The value β controls the amount of admissible bias in the definition of a good choice; cf. (5.7) and (5.8). This value is mainly for theoretical study, in practice one can always take $\beta = 0$.

5.1.4 SmA choice and the oracle inequality

Define the selector \hat{m} by the “smallest accepted” (SmA) rule. Namely, with z_{m,m° from (5.11), the acceptance rule reads as follows:

$$\{m^\circ \text{ is accepted}\} \Leftrightarrow \left\{ \max_{m \in \mathcal{M}_+(m^\circ)} \{T_{m,m^\circ} - z_{m,m^\circ}\} \leq 0 \right\}.$$

The SmA rule is

$$\begin{aligned} \hat{m} &\stackrel{\text{def}}{=} \text{“smallest accepted”} \\ &= \min \left\{ m^\circ : \max_{m \in \mathcal{M}_+(m^\circ)} \{T_{m,m^\circ} - z_{m,m^\circ}\} \leq 0 \right\}. \end{aligned} \quad (5.12)$$

Our study mainly focuses on the behavior of the selector \hat{m} . The performance of the resulting estimator $\hat{\phi} = \tilde{\phi}_{\hat{m}}$ is a kind of corollary from statements about the selected model \hat{m} . The ideal solution would be $\hat{m} \equiv m^*$, then the adaptive estimator $\hat{\phi}$ coincides with the oracle estimate $\tilde{\phi}_{m^*}$.

The bound (5.9) automatically ensures the desired *propagation property*: any good model m° in the sense (5.7) will be accepted with probability at least $1 - e^{-x}$. In some sense, this property is built-in by the construction of the procedure. By definition, the oracle m^* is also a “good” choice, this yields

$$P(m^* \text{ is rejected}) \leq e^{-x}. \quad (5.13)$$

Therefore, the selector \hat{m} typically takes its value in $\mathcal{M}_-(m^*)$, where

$$\mathcal{M}_-(m^*) = \{m \in \mathcal{M} : m < m^*\}$$

is the set of all models in \mathcal{M} smaller than m^* . It remains to check the performance of the method in this region. The next step is to specify a subset \mathcal{M}° of $\mathcal{M}_-(m^*)$ of highly probable \hat{m} -values. We will refer to this subset as the *zone of insensitivity*. The definition of m^* implies that there is a significant bias for each $m \in \mathcal{M}_-(m^*)$. If this bias is really large, then, again, the probability of selecting m can be bounded from above by a small

value. Therefore, the zone of insensitivity is composed of m -values for which the bias is significant but not very large.

Now we present a formal description which specifies a subset \mathcal{M}_- of $\mathcal{M}_-(m^*)$ for which the bias $\|b_{m^*,m}\|$ is sufficiently large and hence, the probability of the event $\{\hat{m} \in \mathcal{M}_-\}$ is negligible.

Theorem 5.1.1. *Let $z_{m,m^\circ}(\cdot)$ be the tail function from (5.9) for each pair $m > m^\circ \in \mathcal{M}$. Given \mathbf{x} and β , let \mathbf{z}_{m,m° be given by (5.10) and (5.11). Then the propagation property (5.13) is fulfilled for the SmA selector \hat{m} . Moreover, for any subset $\mathcal{M}_- \subseteq \mathcal{M}_-(m^*)$ s.t.*

$$\|b_{m^*,m}\| > \mathbf{z}_{m^*,m} + z_{m^*,m}(\mathbf{x}_s), \quad m \in \mathcal{M}_-, \quad (5.14)$$

for $\mathbf{x}_s \stackrel{\text{def}}{=} \mathbf{x} + \log(|\mathcal{M}_-|)$ with $|\mathcal{M}_-|$ being the cardinality of \mathcal{M}_- , it holds

$$\mathbb{P}(\hat{m} \in \mathcal{M}_-) \leq e^{-x}.$$

The SmA estimator $\hat{\phi} = \tilde{\phi}_{\hat{m}}$ satisfies the following bound:

$$\mathbb{P}\left(\|\hat{\phi} - \tilde{\phi}_{m^*}\| > \bar{z}_{m^*}\right) \leq 2e^{-x}, \quad (5.15)$$

where \bar{z}_{m^*} is defined with $\mathcal{M}^\circ \stackrel{\text{def}}{=} \mathcal{M}_-(m^*) \setminus \mathcal{M}_-$ as

$$\bar{z}_{m^*} \stackrel{\text{def}}{=} \max_{m \in \mathcal{M}^\circ} \mathbf{z}_{m^*,m}. \quad (5.16)$$

This implies the probabilistic oracle bound: with probability at least $1 - 2e^{-x}$

$$\|\hat{\phi} - \phi^*\| \leq \|\tilde{\phi}_{m^*} - \phi^*\| + \bar{z}_{m^*}. \quad (5.17)$$

Remark 5.1.1. Note that the choice $\mathbf{x}_s = \mathbf{x} + \log(|\mathcal{M}_-|)$ relies on crude Bonferroni arguments and the definition of \mathcal{M}_- can be refined by choosing \mathbf{x}_s more carefully. However, this value only enters in the theoretical bound and is not used in the procedure, a fine tuning for this value is not required. Obviously $\mathbf{x}_s \leq \mathbf{x} + \log(|\mathcal{M}_-(m^*)|)$.

Remark 5.1.2. The result (5.17) is called the *oracle bound* because it compares the loss of the data-driven selector \hat{m} and of the optimal choice m^* . The value \bar{z}_{m^*} in (5.16) can be viewed as a “payment for adaptation”. An interesting feature of the presented result is that not only the oracle quality but also the payment for adaptation depend upon the unknown response \mathbf{f}^* and the corresponding oracle choice m^* . In the worst case of a model with a flat risk profile \mathcal{R}_m , the set \mathcal{M}° can coincide with the whole range $\mathcal{M}_-(m^*)$. Even in this case the bounds (5.15) and (5.17) are meaningful. However, the payment for adaptation \bar{z}_{m^*} in this case can be larger than the oracle risk. In the contrary, if the risk function \mathcal{R}_m grows rapidly as m decreases below m^* , then the set \mathcal{M}° is small and the value \bar{z}_{m^*} is much smaller than the oracle risk \mathcal{R}_{m^*} .

5.1.5 Analysis of the payment for adaptation \bar{z}_{m^*}

Here we present an upper bound on \bar{z}_{m^*} for a special case of Gaussian independent errors ε_i . The benefit of considering the Gaussian case is that each vector $\xi_{m',m}$ is Gaussian as well, which simplifies the analysis of the tail function $z_{m',m}(\cdot)$. However, the results can be extended to non-Gaussian errors ε_i under exponential moment conditions. Below m_0 denotes the smallest model in \mathcal{M} . Writing $\mathbb{V}_m \stackrel{\text{def}}{=} \sigma^2 \mathcal{K}_m \mathcal{K}_m^\top$, we define

$$\begin{aligned} p_m &\stackrel{\text{def}}{=} \text{tr}(\mathbb{V}_m) \\ \lambda_m &\stackrel{\text{def}}{=} \|\mathbb{V}_m\|_{\text{op}}. \end{aligned}$$

Theorem 5.1.2. *Assume the conditions of Theorem 5.1.1. Let also $p_{m,m^\circ} = \text{tr}(\mathbb{V}_{m,m^\circ})$ and $\lambda_{m,m^\circ} = \|\mathbb{V}_{m,m^\circ}\|_{\text{op}}$ with $\mathbb{V}_{m,m^\circ} = \text{Var}(\xi_{m,m^\circ})$ satisfy $p_{m^*,m} \leq p_{m^*,m_0} \leq p_{m^*}$ and $\lambda_{m^*,m} \leq \lambda_{m^*,m_0} \leq \lambda_{m^*}$ for all $m_0 \leq m < m^*$. If the errors ε_i are normal zero mean then the critical values z_{m,m° given by (5.11) satisfy*

$$z_{m,m^\circ} \leq (1 + \beta) \sqrt{p_{m,m^\circ}} + \sqrt{2\lambda_{m,m^\circ} \{x + \log(|\mathcal{M}|)\}},$$

while the payment for adaptation \bar{z}_{m^*} follows the bound

$$\begin{aligned} \bar{z}_{m^*} &\leq (1 + \beta) \sqrt{p_{m^*,m_0}} + \sqrt{2\lambda_{m^*,m_0} \{x + \log(|\mathcal{M}_-(m^*)|\)} \\ &\leq (1 + \beta) \sqrt{p_{m^*}} + \sqrt{2\lambda_{m^*} \{x + \log(|\mathcal{M}|)\}}. \end{aligned}$$

Some special cases of this result for projection and linear functional estimation will be discussed in Sections 5.1.7 and 5.1.8 below.

5.1.6 Power loss function

The probabilistic oracle bound of Theorem 5.1.1 provides some statement about typical behavior of the adaptive SmA estimate $\hat{\phi} = \tilde{\phi}_{\hat{m}}$. Unfortunately, this bound does not yield a risk bound for quadratic or polynomial losses: even if big losses occur with a small probability, the related risk can still be large. It happens that the SmA procedure can be easily tuned to secure an oracle risk bound.

For simplicity of notation, we only consider the quadratic risk

$$\mathcal{R}(\hat{\phi}) \stackrel{\text{def}}{=} \mathbb{E} \|\hat{\phi} - \phi^*\|^2.$$

We aim at comparing the risk of the SmA procedure with the risk \mathcal{R}_{m^*} of the oracle estimate $\tilde{\phi}_{m^*}$. Recall the representation

$$\mathcal{R}_m \stackrel{\text{def}}{=} \mathbb{E} \|\tilde{\phi}_m - \phi^*\|^2 = \mathbb{E} \|\xi_m\|^2 + \|b_m\|^2 = p_m + \|b_m\|^2$$

with $p_m = \text{tr}(\mathbb{V}_m)$ and $\mathbb{V}_m = \text{Var}(\boldsymbol{\xi}_m)$. For our analysis, we have to slightly modify the definition of the oracle (5.8). Namely, to ensure an oracle risk bound, we require that not only the model m^* is “good” but also all the larger models $m > m^*$ are “good” as well:

$$m^* \stackrel{\text{def}}{=} \min \left\{ m^\circ : \max_{m' \in \mathcal{M}_+(m^\circ) : m' > m} \{ \|b_{m',m}\|^2 - \beta^2 p_{m',m} \} \leq 0 \right\}. \quad (5.18)$$

Below we also suppose that the bias component $\|b_m\|^2$ fulfills

$$\|b_m\| \leq \|b_{m^*}\|, \quad m > m^*. \quad (5.19)$$

Otherwise, one can define $\|b_{m^*}\| \stackrel{\text{def}}{=} \max_{m \in \mathcal{M}_+(m^*)} \|b_m\|$.

The choice of the critical values $z_{m',m}$ for the SmA procedure has to be slightly changed to ensure a risk bound for quadratic loss. For this, we need a bit more detailed analysis of the SmA procedure in the propagation zone $m > m^*$. In this zone the variance dominates the bias, therefore, the SmA procedure can be tuned in the situation when there is no signal and hence no bias at all:

$$\mathbb{T}_{m',m} = \|\tilde{\boldsymbol{\phi}}_{m'} - \tilde{\boldsymbol{\phi}}_m\| = \|\boldsymbol{\xi}_{m',m}\|.$$

The analysis is based on a simple but important observation that if $\hat{m} = m > m^*$, then the good model $m^\circ = m_{[-1]}$ is rejected, where $m_{[-1]}$ denotes the next smaller model with respect to m . The latter means that at least one check based on $\mathbb{T}_{m',m_{(-1)}}$ fails. The same can be expressed as follows: the maximum of the r.v.’s $\mathbb{T}_{m',m_{(-1)}} \mathbb{I}(\mathbb{T}_{m',m_{(-1)}} > z_{m',m_{(-1)}})$ is positive. For a formal description, introduce for each m and \mathbf{x} a random event

$$A_m(\mathbf{x}) \stackrel{\text{def}}{=} \mathbb{I} \left(\max_{m' \in \mathcal{M}_+(m)} \{ \|\boldsymbol{\xi}_{m',m}\| - z_{m',m}(\mathbf{x}) \} > 0 \right)$$

on which at least one of the test statistics $\mathbb{T}_{m',m} = \|\boldsymbol{\xi}_{m',m}\|$ exceeds the critical value $z_{m',m}(\mathbf{x})$. The case of probabilistic loss focuses on the probability of this event, the value \mathbf{x} is selected to make it small enough. Now, under the polynomial loss function, we need a bound for the moment of the corresponding loss. Namely, for each m , consider the expectation of $p_m^{-1} \|\boldsymbol{\xi}_m\|^2$ on the random set $A_{m_{(-1)}}(\mathbf{x})$:

$$\mathcal{R}_m^+(\mathbf{x}) \stackrel{\text{def}}{=} \mathbb{E} \left[(p_m^{-1} \|\boldsymbol{\xi}_m\|^2 \vee 1) \mathbb{I} \left(\max_{m' \in \mathcal{M}_+(m_{(-1)})} \{ \|\boldsymbol{\xi}_{m',m_{(-1)}}\| - z_{m',m_{(-1)}}(\mathbf{x}) \} > 0 \right) \right].$$

Similarly one can consider any other power loss function by replacing $(p_m^{-1/2} \|\boldsymbol{\xi}_m\|)^2$ with $(p_m^{-1/2} \|\boldsymbol{\xi}_m\|)^q$. In particular, $q = 0$ yields the probability loss considered before.

Now we define the value $\mathbf{x}_{m_{(-1)}}$ in such a way that the related deviation risk $\mathcal{R}_m^+(\mathbf{x})$ can be controlled from above. Let α_m be a given decreasing sequence. Its choice will be discussed below. We fix for each m the value $\mathbf{x}_{m_{(-1)}}$ such that

$$\mathcal{R}_m^+(\mathbf{x}_{m(-1)}) = \alpha_m. \quad (5.20)$$

It implies

$$\begin{aligned} \mathbb{E}\left[\|\boldsymbol{\xi}_m\|^2 \mathbb{I}(A_{m(-1)}(\mathbf{x}))\right] &\leq \alpha_m p_m, \\ \mathbb{P}(A_{m(-1)}(\mathbf{x})) &\leq \alpha_m. \end{aligned} \quad (5.21)$$

Now define the critical values \mathbf{z}_{m,m° of the SmA procedure as

$$\mathbf{z}_{m,m^\circ} = z_{m,m^\circ}(\mathbf{x}_{m^\circ}) + \beta p_{m,m^\circ}^{1/2}. \quad (5.22)$$

The resulting procedure reads exactly as in the case of probabilistic loss:

$$\hat{m} = \min \left\{ m^\circ : \max_{m \in \mathcal{M}_+(m^\circ)} \{ \mathbb{T}_{m,m^\circ} - \mathbf{z}_{m,m^\circ} \} \leq 0 \right\}. \quad (5.23)$$

It is worth mentioning that the procedure is the same, and even the critical values \mathbf{z}_{m,m° are given by the same formula, as in the case of probabilistic loss. The only difference is in the propagation condition (5.20) which is a bit stronger than a similar condition for indicator loss. This implies that the values \mathbf{x}_{m° and \mathbf{z}_{m,m° are a bit larger in the case of a power loss function.

Theorem 5.1.3. *Let the SmA procedure (5.23) be applied with the critical values \mathbf{z}_{m,m° from (5.22), where the values \mathbf{x}_m are defined by (5.20) with the coefficients α_m satisfying*

$$\sum_{m \in \mathcal{M}_+(m^*)} \alpha_m p_m \leq \bar{\alpha}_{m^*} p_{m^*} \quad (5.24)$$

for some $\bar{\alpha}_{m^*}$. If the errors ε_i are normal zero mean, then

$$\mathbb{E}\|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\|^2 \leq 2\bar{\alpha}_{m^*} \mathcal{R}_{m^*} + (\mathcal{R}_{m^*}^{1/2} + \bar{\mathbf{z}}_{m^*})^2, \quad (5.25)$$

where

$$\bar{\mathbf{z}}_{m^*} \stackrel{\text{def}}{=} \max_{m \in \mathcal{M}_-(m^*)} \mathbf{z}_{m^*,m}.$$

Similarly to the probabilistic loss function, the result can be refined by considering the zone of insensitivity in the region $m < m^*$.

Now we briefly discuss the choice of constants α_m entering into (5.24). Suppose that the p_m 's satisfy

$$\sum_{m \in \mathcal{M}_+(m^*)} (p_{m^*}/p_m)^a \leq C \quad (5.26)$$

for some $a > 0$ and a fixed constant C . Then one can take

$$\alpha_m = (\mathbf{p}_m / \mathbf{p}_{m_0})^{-1-a}.$$

Below we focus on a situation when the effective dimension \mathbf{p}_m grows exponentially with m . Note that this situation is typical in model selection and often one can reduce the general case to this one by a proper discretization. Then (5.26) is fulfilled for any $a > 0$ with $C = C(a)$.

The further step is an upper bound on the values \mathbf{x}_m , $z_{m,m^*}(\mathbf{x}_m)$, and \mathbf{z}_{m,m^*} , as well as on the payment for adaptation \bar{z}_{m^*} . These bounds require some exponential moment conditions on the errors ε_i . To reduce the computational burden, we again focus on the case of Gaussian errors.

Proposition 5.1.1. *Suppose (5.26) for $a > 0$. If the errors ε_i are normal zero mean, then the choice*

$$\alpha_m = \sqrt{3}(\mathbf{p}_m / \mathbf{p}_{m_0})^{-1-a}, \quad \mathbf{x}_{m_{(-1)}} = 2(1+a) \log(\mathbf{p}_m / \mathbf{p}_{m_0}), \quad (5.27)$$

ensures conditions (5.24), (5.20), and therefore, the oracle bound (5.25) with $\bar{\alpha}_{m^*} = \sqrt{3}C(\mathbf{p}_{m_0} / \mathbf{p}_{m^*})^{1+a}$. Furthermore,

$$\bar{z}_{m^*} \leq \beta \sqrt{\mathbf{p}_{m^*}} + \sqrt{2\lambda_{m^*} \{2(1+a) \log(\mathbf{p}_{m^*} / \mathbf{p}_{m_0}) + \log(|\mathcal{M}|)\}}. \quad (5.28)$$

5.1.7 Application to projection estimation

An important feature of the obtained oracle statements is their universality: they equally apply to various setups and problems and provide some quantitative explicit error bounds even for finite samples. Below we briefly comment on two popular cases of projection estimation and estimation of a linear functional. In some sense, these are two extreme cases of relation between \mathbf{p}_{m^*} and λ_{m^*} .

This section discusses the case of projection estimation in the linear model $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with homogeneous errors $\varepsilon_i : \text{Var}(\varepsilon_i) = \sigma^2$. All the conclusions can be easily extended to heterogeneous errors whose variances are contained in some fixed interval. We also focus on probabilistic loss, the case of polynomial loss can be considered in the same way.

Let us assume an ordering on the features of Ψ_m and let for each $m \in \mathbb{N}$ denote Ψ_m as the submatrix Ψ corresponding to first m features, i. e. the projector onto the first m features. We use m to denote the model and the number of features. The related estimator $\tilde{\boldsymbol{\theta}}_m$ is the standard LSE with $\mathcal{S}_m = (\Psi_m \Psi_m^\top)^{-1} \Psi_m$ and the prediction problem with $W = \Psi^\top$ yields $\mathcal{K}_m \mathbf{Y} = \Pi_m \mathbf{Y}$ where $\Pi_m = \Psi_m^\top (\Psi_m \Psi_m^\top)^{-1} \Psi_m$ is the projector

onto the corresponding feature subspace. For homogeneous errors ε_i with $\text{Var}(\varepsilon_i) = \sigma^2$, the variance $\mathbb{V}_m = \text{Var}(\Pi_m \mathbf{Y})$ satisfies

$$\mathbf{p}_m = \text{tr}\{\text{Var}(\Pi_m \mathbf{Y})\} = \sigma^2 \text{tr}(\Pi_m) = \sigma^2 m.$$

Moreover, for each pair $m > m^\circ$, it holds

$$\Psi^\top (\tilde{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_{m^\circ}) = (\Pi_m - \Pi_{m^\circ}) \mathbf{Y} = \Pi_{m,m^\circ} \mathbf{Y},$$

where Π_{m,m° projects on the subspace of features entering in m but not in m° .

Corollary 5.1.1. *Consider the problem of projection estimation with homogeneous Gaussian errors ε_i and probabilistic loss. Then $\mathbf{p}_{m,m^\circ} = \sigma^2(m - m^\circ)$, $\lambda_{m,m^\circ} = \sigma^2$, and*

$$\begin{aligned} z_{m,m^\circ} &\leq \sigma(1 + \beta)\sqrt{m - m^\circ} + \sigma\sqrt{2x + 2\log(|\mathcal{M}|)}, \\ \bar{z}_{m^*} &\leq \sigma(1 + \beta)\sqrt{m^*} + \sigma\sqrt{2x + 2\log(|\mathcal{M}|)}. \end{aligned}$$

The first term in the expression for \bar{z}_{m^*} is of order $\sqrt{m^*}$ and it is a leading one provided that the effective dimension m^* is essentially larger than $\log(|\mathcal{M}|)$. Usually the cardinality of the set \mathcal{M} is only logarithmic in the sample size n ; cf. Lepski (1991); Lepski et al. (1997). Then $\log(|\mathcal{M}|) \approx \log \log n$ and $\bar{z}_{m^*} \approx \sigma\sqrt{m^*}$ for $m^* \gg \log \log n$. For the oracle risk \mathcal{R}_{m^*} , it holds $\mathcal{R}_{m^*} = \mathbf{p}_{m^*} + \|b_{m^*}\|^2 \geq \sigma^2 m^*$. Therefore, the payment for adaptation \bar{z}_{m^*} is of the same order as the square root of the oracle risk, and the result of Proposition 5.1.2 has a surprising corollary: rate adaptive estimation is possible if the oracle dimension m^* is significantly larger than $\log \log n$.

Remark 5.1.3. The payment for adaptation can be drastically reduced in the situations with a narrow zone of insensitivity. If the bias grows rapidly when m decreases from m^* to m_0 , more precisely, if $\|b_{m^*,m}\|^2 \geq C\sigma^2(m^* - m + 2x + 2\log(|\mathcal{M}|))$ for some fixed constant C and all $m \leq m^\circ$ with $m^\circ < m^*$, then

$$\bar{z}_{m^*} \leq \sigma(1 + \beta)\sqrt{m^* - m^\circ} + \sigma\sqrt{2x + 2\log(|\mathcal{M}|)}.$$

So, if $(m^* - m^\circ)/m^*$ is small, the payment for adaptation is smaller in order than the oracle risk, and the procedure is sharp adaptive. In particular, one can easily see that the self-similarity condition of Gine and Nickl (2010) ensures a rapid growth of the bias when the index m becomes smaller than m^* . This in turn yields a narrow zone of insensitivity and hence, a sharp adaptive estimation.

Remark 5.1.4. It is worth mentioning the relation of the proposed procedure to the popular Akaike (AIC) criterion. AIC defines \hat{m} by minimizing

$$\widehat{m} = \operatorname{argmin}_m \{ \| \mathbf{Y} - \Pi_m \mathbf{Y} \|^2 + 2\sigma^2 m \}.$$

One can easily see that this rule is equivalent to the SmA rule (5.12) with $z_{m,m^\circ}^2 = 2\sigma^2(m - m^\circ)$. However, this choice does not guarantee the propagation condition (5.13).

5.1.8 Linear functional estimation

In this section, we discuss the problem of linear functional estimation. As previously, we assume a family of estimators $\tilde{\phi}_m = \mathcal{K}_m \mathbf{Y}$, $m \in \mathcal{M}$, to be given, where the rank of each \mathcal{K}_m is equal to 1. The ordering condition means that these estimators are ordered by their variance:

$$v_m^2 \stackrel{\text{def}}{=} \operatorname{Var}(\mathcal{K}_m \mathbf{Y}) = \mathcal{K}_m \operatorname{Var}(\boldsymbol{\varepsilon}) \mathcal{K}_m^\top \quad (5.29)$$

grows with m . Further, each stochastic component $\xi_{m,m^\circ} = \mathcal{K}_{m,m^\circ} \boldsymbol{\varepsilon}$ is one-dimensional, and it holds

$$\lambda_{m,m^\circ} = p_{m,m^\circ} = v_{m,m^\circ}^2 = \mathcal{K}_{m,m^\circ} \operatorname{Var}(\boldsymbol{\varepsilon}) \mathcal{K}_{m,m^\circ}^\top.$$

Note that in the case of Gaussian errors, ξ_{m,m° is also Gaussian: $\xi_{m,m^\circ} \sim \mathcal{N}(0, v_{m,m^\circ}^2)$. The tail function $z_{m,m^\circ}(x)$ of ξ_{m,m° can be upper-bounded by $v_{m,m^\circ} \sqrt{2x}$. In the case of probabilistic loss, a Bonferroni correction and a bias adjustment lead to the upper bound for the critical values z_{m,m° :

$$z_{m,m^\circ} \leq v_{m,m^\circ} \left(\beta + \sqrt{2x + 2 \log(|\mathcal{M}|)} \right), \quad (5.30)$$

where $|\mathcal{M}|$ is the number of elements in \mathcal{M} . This implies

$$\bar{z}_{m^*} \leq v_{m^*} \left(\beta + \sqrt{2x + 2 \log(|\mathcal{M}|)} \right). \quad (5.31)$$

Theorem 5.1.4. *Let the errors $\boldsymbol{\varepsilon}_i$ be Gaussian zero mean. Consider a problem of linear functional estimation of $\phi^* = \mathcal{K} \mathbf{f}^*$ by a given family $\tilde{\phi}_m = \mathcal{K}_m \mathbf{Y}$ with $\operatorname{rank}(\mathcal{K}_m) = \operatorname{rank}(\mathcal{K}) = 1$, $m \in \mathcal{M}$. Then the critical values z_{m,m° from (5.11) fulfill (5.30) and the oracle inequality (5.17) holds with the payment for adaptation \bar{z}_{m^*} obeying (5.31).*

Remark 5.1.5. One can conclude that for the problem of functional estimation with probabilistic loss, the squared payment for adaptation $\bar{z}_{m^*}^2$ is by factor $\log(|\mathcal{M}|)$ larger than the oracle variance $v_{m^*}^2$. If $|\mathcal{M}|$ itself is logarithmic in the sample size n , we end up with the extra $(\log \log n)$ -factor in the accuracy of adaptive estimation.

In the case of *polynomial loss*, similar arguments yield due to (5.22) and (5.27)

$$\begin{aligned} z_{m,m^\circ} &\leq v_{m,m^\circ} (\beta + \sqrt{2x_{m^\circ} + 2\log(|\mathcal{M}|)}) \\ &\leq v_{m,m^\circ} (\beta + \sqrt{2(1+a)\log(p_{m^\circ}/p_{m_0}) + 2\log(|\mathcal{M}|)}) \end{aligned}$$

Spokoiny and Vial (2009) showed that the bound $z_{m,m^\circ}^2 \geq Cv_{m,m^\circ}^2(m^\circ - m_0)$ is necessary to ensure a propagation condition for geometrically growing variance $p_m = v_m^2$. The bound (5.30) yields

$$\bar{z}_{m^*} \leq v_{m^*} \left(\beta + \sqrt{2(1+a)\log(v_{m^*}^2/v_{m_0}^2) + 2\log(|\mathcal{M}|)} \right). \quad (5.32)$$

Theorem 5.1.5. Suppose that the errors ε_i are Gaussian zero mean. Let the family of functional estimators $\mathcal{K}_m \mathbf{Y}$ be such that the variances $p_m = v_m^2$ from (5.29) fulfill the condition (5.26) with $a > 0$. Then the critical values z_{m,m° from (5.22) for the SmA procedure fulfill (5.30). For the resulting selector \hat{m} , the oracle inequality (5.25) holds and the payment for adaptation \bar{z}_{m^*} follows (5.32).

Remark 5.1.6. It appears that polynomial loss yields a larger price for adaptation: $\bar{z}_{m^*}^2 \asymp v_{m^*}^2 \log(v_{m^*}^2/v_{m_0}^2)$. This conclusion is consistent with the results by Lepski (1992) and Cai and Low (2003, 2005) which show that the log-price for adaptation cannot be avoided if a polynomial loss function is considered. Our result seems to be even more informative because it delivers a non-asymptotic error bound which adapts to the underlying unknown model.

5.2 Bootstrap tuning

This section explains how the proposed SmA procedure can be applied if no information about the noise $\varepsilon = \mathbf{Y} - I\mathbf{E}\mathbf{Y}$ is available.

5.2.1 Presmoothing and wild bootstrap

Let the observed data \mathbf{Y} follow the model $\mathbf{Y} = \mathbf{f}^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \Sigma)$, where $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ is an unknown diagonal covariance matrix. We assume that the response vector \mathbf{f}^* can be well approximated by a linear expansion for a given basis Ψ in the form $\mathbf{f}^* \approx \Psi^\top \boldsymbol{\theta}^*$. The vector $\boldsymbol{\theta}^*$ can be naturally treated as target of estimation. Assume we are given the ordered family of the estimators $(\tilde{\boldsymbol{\theta}}_m)$ of $\boldsymbol{\theta}^*$:

$$\tilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \mathbf{Y} = (\Psi_m \Psi_m^\top)^{-1} \Psi_m \mathbf{Y}, \quad m \in \mathcal{M}.$$

For each pair $m > m^\circ$ from \mathcal{M} , we consider the test statistic \mathbb{T}_{m,m° and its decomposition from (5.5): with $\mathcal{K}_{m,m^\circ} = W(\mathcal{S}_m - \mathcal{S}_{m^\circ})$

$$\mathbb{T}_{m,m^\circ} = \|\mathcal{K}_{m,m^\circ} \mathbf{Y}\| = \|\mathcal{K}_{m,m^\circ}(\mathbf{f}^* + \boldsymbol{\varepsilon})\| = \|b_{m,m^\circ} + \boldsymbol{\xi}_{m,m^\circ}\|,$$

Calibration of the SmA model selection procedure requires to know the joint distribution of all corresponding stochastic terms $\|\boldsymbol{\xi}_{m,m^\circ}\|$ for $m > m^\circ$ which is uniquely determined by the noise covariance matrix Σ . In the case when this matrix is unknown, we are going to use a bootstrapping procedure to approximate this distribution.

The proposed procedure relates to the concept of the *wild* bootstrap, [Wu \(1986\)](#), [Beran \(1986\)](#). In the framework of a regression problem, it suggests to model the unknown heteroscedastic noise using randomly weighted residuals from pilot estimation. We apply normal weights. For other possible bootstrap weights see for example [Mammen \(1993\)](#).

Suppose we are given a pilot estimator (presmoothing) $\tilde{\mathbf{f}}$ of the response vector $\mathbf{f}^* \in \mathbb{R}^n$. Define the residuals:

$$\check{\mathbf{Y}} \stackrel{\text{def}}{=} \mathbf{Y} - \tilde{\mathbf{f}}.$$

This pilot is supposed to undersmooth, that is, the bias is negligible and the variance of $\check{\mathbf{Y}}$ is close to Σ . This pre-smoothing requires some minimal smoothness of the regression function, and this condition seems to be unavoidable if no information about the noise is given: otherwise one cannot distinguish between signal and noise. Below we suppose that $\tilde{\mathbf{f}}$ is a linear predictor, $\tilde{\mathbf{f}} = \Pi \mathbf{Y}$, where Π is a sub-projector in the space \mathbb{R}^n . For example, one can take $\Pi = \Psi_{m^\dagger}^\top (\Psi_{m^\dagger} \Psi_{m^\dagger}^\top)^{-1} \Psi_{m^\dagger}$ where m^\dagger is a large model, e.g. the largest model M in our collection.

The wild bootstrap proposes to resample from the heteroscedastic Gaussian noise $\mathbb{P}^b = \mathcal{N}(0, \check{\Sigma})$ with

$$\check{\Sigma} = \text{diag}(\check{\mathbf{Y}} \cdot \check{\mathbf{Y}}),$$

where $\check{\mathbf{Y}} \cdot \check{\mathbf{Y}}$ denotes the coordinate-wise product of the vector $\check{\mathbf{Y}}$ with itself and $\text{diag}(\check{\mathbf{Y}} \cdot \check{\mathbf{Y}})$ denotes the diagonal matrix with entries from $\check{\mathbf{Y}} \cdot \check{\mathbf{Y}}$. These entries depend on \mathbf{Y} and thus are random. Therefore, the bootstrap distribution \mathbb{P}^b is a random measure on \mathbb{R}^n and the aim of our study is to show that this random measure mimics well the underlying data distribution for typical realizations of \mathbf{Y} . Clearly $\text{diag}(\check{\mathbf{Y}} \cdot \check{\mathbf{Y}})$ is a very bad estimator of the covariance matrix Σ . However, below we show that under realistic conditions on the pilot $\tilde{\mathbf{f}}$ and on the model, it does a good job and allows to obtain essentially the same results as in the case of known Σ .

Let \mathbf{w}^b denote the n -vector of bootstrap weights $\mathbf{w}^b \sim \mathcal{N}(0, I_n)$. Clearly the product $\boldsymbol{\varepsilon}^b = \text{diag}(\check{\mathbf{Y}}) \mathbf{w}^b$ is conditionally on \mathbf{Y} normal,

$$\boldsymbol{\varepsilon}^b = \text{diag}(\check{\mathbf{Y}}) \mathbf{w}^b \mid \mathbf{Y} \sim \mathcal{N}(0, \check{\Sigma}).$$

Bootstrap analog of $\xi_{m,m^\circ} = \mathcal{K}_{m,m^\circ}\varepsilon$ reads $\xi_{m,m^\circ}^b = \mathcal{K}_{m,m^\circ}\varepsilon^b = \mathcal{K}_{m,m^\circ}\text{diag}(\check{\mathbf{Y}})\mathbf{w}^b$ and

$$\|\xi_{m,m^\circ}^b\| \stackrel{\text{def}}{=} \|\mathcal{K}_{m,m^\circ}\text{diag}(\check{\mathbf{Y}})\mathbf{w}^b\|. \quad (5.33)$$

The idea is to calibrate the SmA procedure under the bootstrap measure \mathbb{I}^b using $\|\xi_{m,m^\circ}^b\|$ in place of $\|\xi_{m,m^\circ}\|$. The bootstrap quantiles $z_{m,m^\circ}^b(t)$ are given by analog of (5.9):

$$\mathbb{I}^b\left(\|\xi_{m,m^\circ}^b\| > z_{m,m^\circ}^b(t)\right) = e^{-t}. \quad (5.34)$$

The multiplicity correction $q_{m^\circ}^b = q_{m^\circ}^b(\mathbf{x})$ is specified by the condition

$$\mathbb{I}^b\left(\bigcup_{m \in \mathcal{M}^+(m^\circ)} \{\|\xi_{m,m^\circ}^b\| \geq z_{m,m^\circ}^b(\mathbf{x} + q_{m^\circ}^b)\}\right) = e^{-\mathbf{x}}. \quad (5.35)$$

Finally, the bootstrap critical values are fixed by the analog of (5.11):

$$\mathbf{z}_{m,m^\circ}^b \stackrel{\text{def}}{=} z_{m,m^\circ}^b(\mathbf{x} + q_{m^\circ}^b) + \beta \sqrt{\mathbf{p}_{m,m^\circ}^b}$$

for $\mathbf{p}_{m,m^\circ}^b = \mathbb{E}^b\|\xi_{m,m^\circ}^b\|^2$ given by

$$\mathbf{p}_{m^\circ,m}^b \stackrel{\text{def}}{=} \text{tr}(\mathcal{K}_{m^\circ,m}^\top \text{diag}(\check{\mathbf{Y}} \cdot \check{\mathbf{Y}}) \mathcal{K}_{m^\circ,m}).$$

Recall that all these quantities are data-driven and depend upon the original data. Now we apply the SmA procedure with the critical values \mathbf{z}_{m,m°^b defined in such a way. Our main result claims that this choice still ensures the propagation condition (5.10) and therefore, all the obtained results including the oracle bounds, apply for this choice as well; see Theorem 5.2.2. Moreover, we evaluate the distance between the unknown underlying data distribution \mathbb{P} and the bootstrap distribution \mathbb{I}^b . The latter is random, however, we show that with high probability, it is close to its deterministic counterpart \mathbb{I}^b . To make the results transparent and concise we assume a heterogeneous Gaussian noise ε . All the statements can be extended to a non-Gaussian noise under some exponential moment conditions at the cost of many technical details.

Let \mathbb{Q} denote the joint distribution of all stochastic vectors ξ_{m,m° entering in the decomposition of the test statistics \mathbb{T}_{m,m° for $m > m^\circ$. Let also \mathbb{Q}^b be the similar distribution of the bootstrapized stochastic vectors ξ_{m,m°^b entering in the test statistics \mathbb{T}_{m,m°^b . The next result allows to upper bound the total variation distance between \mathbb{Q} and \mathbb{Q}^b in terms of the following quantities:

Design Regularity is measured by the value δ_Ψ

$$\delta_\Psi \stackrel{\text{def}}{=} \max_{i=1,\dots,n} \|S^{-1/2}\Psi_i\|\sigma_i, \quad \text{where} \quad S \stackrel{\text{def}}{=} \sum_{i=1}^n \Psi_i \Psi_i^\top \sigma_i^2; \quad (5.36)$$

Obviously

$$\sum_{i=1}^n \|S^{-1/2}\Psi_i\|^2 \sigma_i^2 = \text{tr} \left(\sum_{i=1}^n S^{-2} \Psi_i \Psi_i^\top \sigma_i^2 \right) = \text{tr } I_p = p,$$

and therefore in typical situations the value δ_Ψ is of order $\sqrt{p/n}$.

Presmoothing bias for a projector Π is described by the vector

$$\mathbf{B} = \Sigma^{-1/2}(\mathbf{f}^* - \Pi \mathbf{f}^*). \quad (5.37)$$

We will use the sup-norm $\|\mathbf{B}\|_\infty = \max_i |b_i|$ and the squared ℓ_2 -norm $\|\mathbf{B}\|^2 = \sum_i b_i^2$ to measure the bias after presmoothing.

Stochastic noise after presmoothing is described via the covariance matrix $\text{Var}(\check{\boldsymbol{\varepsilon}})$ of the smoothed noise $\check{\boldsymbol{\varepsilon}} = \Sigma^{-1/2}(\boldsymbol{\varepsilon} - \Pi \boldsymbol{\varepsilon})$. Namely, this matrix is assumed to be sufficiently close to the unit matrix I_n , in particular, its diagonal elements should be close to one. This is measured by the operator norm of $\text{Var}(\check{\boldsymbol{\varepsilon}}) - I_n$ and by deviations of the individual variances $\mathbb{E}\check{\varepsilon}_i^2$ from one:

$$\begin{aligned} \delta_1 &\stackrel{\text{def}}{=} \|\text{Var}(\check{\boldsymbol{\varepsilon}}) - I_n\|_{\text{op}}, \\ \delta_\varepsilon &\stackrel{\text{def}}{=} \max_i |\mathbb{E}\check{\varepsilon}_i^2 - 1|. \end{aligned} \quad (5.38)$$

In particular, in the case of homogeneous errors $\Sigma = \sigma^2 I_n$ and the smoothing operator Π as a p -dimensional projector, it holds

$$\begin{aligned} \text{Var}(\check{\boldsymbol{\varepsilon}}) &= (I_n - \Pi)^2 = I_n - \Pi \leq I_n, \\ \delta_1 &= \|\text{Var}(\check{\boldsymbol{\varepsilon}}) - I_n\|_{\text{op}} = \|\Pi\|_{\text{op}} = 1, \\ \delta_\varepsilon &= \max_i |\mathbb{E}\check{\varepsilon}_i^2 - 1| = \max_i |\Pi_{ii}|. \end{aligned}$$

One can check that $\Pi_{ii} \asymp \sqrt{p/n}$ for typical smoothing operators like local average or kernel smoothing. Similar bounds with an additional constant can be established for general regular noise $\boldsymbol{\varepsilon}$ and a general smoothing operator Π .

Regularity of the smoothing operator Π is required in Theorem 5.2.2. This condition will be expressed via the norm of the rows Υ_i^\top of the matrix $\Upsilon \stackrel{\text{def}}{=} \Sigma^{-1/2} \Pi \Sigma^{1/2}$ fulfill

$$\|\Upsilon_i^\top\| \leq \delta_\Pi, \quad i = 1, \dots, n. \quad (5.39)$$

This condition is in fact very close to the design regularity condition (5.36). To see this, consider the case of a homogeneous noise with $\Sigma = \sigma^2 I_n$ and $\Pi = \Psi^\top (\Psi \Psi^\top)^{-1} \Psi$. Then $\Upsilon = \Pi$ and (5.36) implies

$$\|\Upsilon_i^\top\| = \|\Psi^\top (\Psi\Psi^\top)^{-1}\Psi_i\| = \|(\Psi\Psi^\top)^{-1/2}\Psi_i\| \leq \delta_\Psi.$$

In general one can expect that (5.39) is fulfilled with some other constant which however, is of the same magnitude as δ_Ψ . For simplicity, we use the same symbol.

5.2.2 Bootstrap validation. Range of applicability

This section states the main results justifying the proposed bootstrap procedure. They claim that the joint distribution \mathbb{Q}^b of the bootstrap stochastic components ξ_{m,m^o}^b for $m > m^o$ nicely reproduces the underlying distribution \mathbb{Q} of the ξ_{m,m^o} 's, and hence, all the probabilistic results obtained in Section 5.1 for known noise continue to apply after bootstrap parameter tuning. In the next result, we give a bound on the total variation distance $\|\mathbb{Q} - \mathbb{Q}^b\|_{TV}$ between \mathbb{Q} and \mathbb{Q}^b .

Theorem 5.2.1. *Let $\mathbf{Y} = \mathbf{f}^* + \boldsymbol{\varepsilon}$ be a Gaussian vector in \mathbb{R}^n with independent components, $\mathbf{Y} \sim \mathcal{N}(\mathbf{f}^*, \Sigma)$ for $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$. Let also Ψ be a $p \times n$ feature matrix such that the $p \times p$ -matrix $S = \Psi \Sigma \Psi^\top$ is non-degenerated. For a given presmothing operator $\Pi: \mathbb{R}^n \rightarrow \mathbb{R}^n$, assume that δ_1 from (5.38) satisfies $\delta_1 \leq 1$. Let $\mathbb{Q} = \mathcal{L}(\xi_{m,m^o}, m, m^o \in \mathcal{M})$ and let \mathbb{Q}^b be the joint conditional distribution of the bootstrap stochastic terms ξ_{m,m^o}^b for $m, m^o \in \mathcal{M}$ given the data \mathbf{Y} . Then it holds on a random set $\Omega_2(\mathbf{x})$ with $\mathbb{P}(\Omega_2(\mathbf{x})) \geq 1 - 3e^{-\mathbf{x}}$:*

$$\begin{aligned} \|\mathbb{Q} - \mathbb{Q}^b\|_{TV} &\leq \frac{1}{2} \Delta_2(\mathbf{x}), \\ \Delta_2(\mathbf{x}) &\stackrel{\text{def}}{=} 2\sqrt{\delta_\Psi^2 p \mathbf{x}_n} + \sqrt{\delta_\varepsilon^2 p} + \sqrt{\|\mathbf{B}\|_\infty^4 p} + 4\delta_\Psi^2 \|\mathbf{B}\| (1 + \sqrt{\mathbf{x}}). \end{aligned} \quad (5.40)$$

where $\mathbf{x}_n = \mathbf{x} + \log(n)$, the bias \mathbf{B} is given by (5.37) and δ_1 , δ_ε by (5.38).

The result (5.40) gives us a way to control differences $\mathbb{Q}(A) - \mathbb{Q}^b(A)$ for fixed sets A . To justify the propagation property for the bootstrap-based set of critical values $z_{m,m^o}^b(\mathbf{x} + q_{m^o}^b)$, given according to (5.33), (5.34), and (5.35) with $\check{\mathbf{Y}} = \mathbf{Y} - \Pi \mathbf{Y}$, we also need to take into account the \mathbf{Y} -dependence of $z_{m,m^o}^b(\mathbf{x} + q_{m^o}^b)$. This is done by the following theorem.

Theorem 5.2.2. *Assume the conditions of Theorem 5.2.1, and let the rows Υ_i^\top of the matrix $\Upsilon \stackrel{\text{def}}{=} \Sigma^{-1/2} \Pi \Sigma^{1/2}$ satisfy (5.39). Then for each $m^o \in \mathcal{M}$*

$$\mathbb{P}\left(\max_{m > m^o} \left\{ \|\xi_{m,m^o}\| - z_{m,m^o}^b(\mathbf{x} + q_{m^o}^b) \right\} \geq 0\right) \leq 6e^{-\mathbf{x}} + \sqrt{p} \Delta_0(\mathbf{x}), \quad (5.41)$$

where with $\mathbf{x}_n = \mathbf{x} + \log(n)$ and $\mathbf{x}_p = \mathbf{x} + \log(2p)$

$$\Delta_0(\mathbf{x}) \stackrel{\text{def}}{=} \|\mathbf{B}\|_\infty^2 + \delta_\Psi^2 \|\mathbf{B}\| \sqrt{2\mathbf{x}} + 2\delta_\Pi \mathbf{x}_n + \delta_\Pi^2 \mathbf{x}_n + 2\delta_\Psi \sqrt{\mathbf{x}_p} + 2\delta_\Psi^2 \mathbf{x}_p.$$

The SmA procedure also involves the values p_{m,m° , which are unknown and depend on the noise ε . The next result shows the bootstrap counterparts p_{m,m°^b can be well used in place of p_{m,m° .

Theorem 5.2.3. *Assume the conditions of Theorem 5.2.1. Then it holds on a set $\Omega_1(\mathbf{x})$ with $I\!\!P(\Omega_1(\mathbf{x})) \geq 1 - 3e^{-x}$ for all pairs $m < m^\circ \in \mathcal{M}$*

$$\left| \frac{p_{m,m^\circ}^b}{p_{m,m^\circ}} - 1 \right| \leq \Delta_p,$$

$$\Delta_p \stackrel{\text{def}}{=} \|B\|_\infty^2 + 4x_M^{1/2} \delta_n^2 \|B\| + 4x_M^{1/2} \delta_n + 4x_M \delta_n^2 + \delta_\varepsilon,$$

where $p_{m,m^\circ}^b = I\!\!E^b \|\xi_{m,m^\circ}^b\|^2$, $p_{m,m^\circ} = I\!\!E \|\xi_{m,m^\circ}\|^2$, and $x_M = x + 2 \log(|\mathcal{M}|)$.

The above results immediately imply all the oracle bounds for probabilistic loss of Section 5.1 with the obvious correction of the error terms.

Now we discuss the sense of the required conditions for bootstrap validity. Our results are only meaningful and the bootstrap approximation is accurate if the values $\Delta_2(\mathbf{x})$ and $\sqrt{p} \Delta_0(\mathbf{x})$ are small. One easily gets

$$\Delta_2(\mathbf{x}) \asymp \sqrt{p} \Delta_0(\mathbf{x}) \leq C p^{1/2} (\|B\|_\infty^2 + \delta_\Psi^2 \|B\| + \delta_\varepsilon + \delta_\varepsilon),$$

where C is a generic notation for absolute constants and log-terms like x_n, x_p etc. So, keeping the errors of bootstrap approximation small requires that the values $\delta_\Psi^2 p$, $\delta_\varepsilon^2 p$, $\|B\|_\infty^4 p$, and $\delta_\Psi^2 \|B\|$ are sufficiently small. Now we spell this condition in the typical situation with $\delta_\Psi \asymp \sqrt{p/n}$ and $\delta_\varepsilon \asymp \sqrt{p/n}$. Then we need that $p^2 \log(n)/n$ is small. Further, the bias component does not destroy the bootstrap validity result if the values $\|B\|_\infty^4 p$ and $p n^{-1} \|B\| \leq p n^{-1/2} \|B\|_\infty$ are small. If f^* is Hölder-smooth with the parameter s :

$$\|B\|_\infty \leq C p^{-s} \tag{5.42}$$

then the bootstrap procedure is justified for $s > 1/4$ if $p = p_n \rightarrow \infty$ but $p_n^2/n \rightarrow 0$ as $n \rightarrow \infty$. We state one asymptotic result of this sort.

Corollary 5.2.1. *Assume the conditions of Theorem 5.2.2 and let $p = p_n$ fulfill $p_n^2 \log(n)/n \rightarrow 0$ as $n \rightarrow \infty$, and (5.42) hold for $s > 1/4$. Then the results of Theorem 5.2.1 and 5.2.2 apply with a small value $\Delta_n = (\sqrt{p_n} \Delta_0(x_n)) \vee \Delta_2(x_n) \rightarrow 0$ as $n \rightarrow 0$.*

5.3 Simulations

This section illustrates the performance of the proposed procedure by means of simulated examples. We consider a regression problem for an unknown univariate function on $[0, 1]$ with unknown inhomogeneous noise. The aim is to compare the bootstrap-calibrated procedure with the SmA procedure for the known noise and with the oracle estimator. We also check the sensitivity of the method to the choice of the presmoothing parameter m^\dagger .

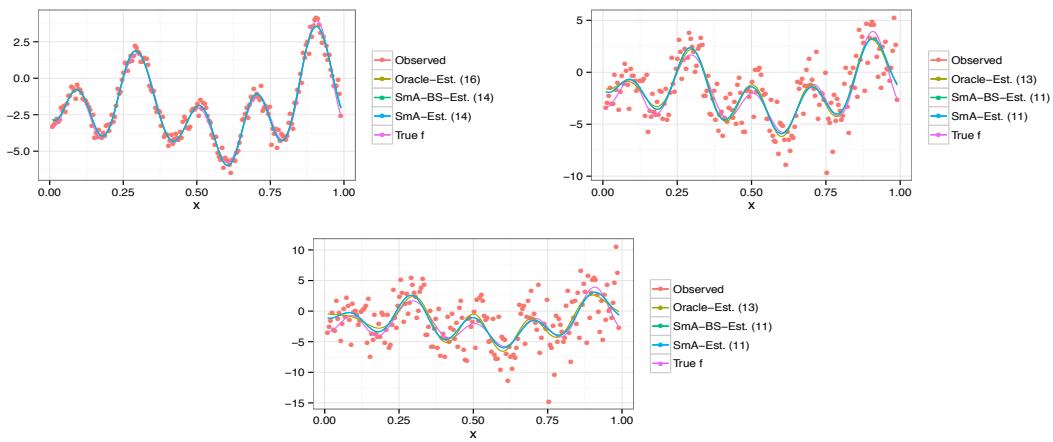


Fig. 5.1. True functions and observed values plotted with oracle estimator, the known-variance SmA-Estimator (SmA-Est.) and the Bootstrap-SmA-Estimator (SmA-BS-Est.) for 3 different functions with different noise structure going from low noise to high noise. The numbers in parentheses indicate the chosen model dimension.

We use a uniform design on $[0, 1]$ and the Fourier basis $\{\psi_j(x)\}_{j=1}^\infty$ for approximation of the regression function f which is modelled in the form

$$f(x) = c_1\psi_1(x) + \dots + c_p\psi_p(x),$$

where the $(c_j)_{1 \leq j \leq p}$ are chosen randomly: with γ_j i.i.d. standard normal

$$c_j = \begin{cases} \gamma_j, & 1 \leq j \leq 10, \\ \gamma_j/(j-10)^2, & 11 \leq j \leq 200. \end{cases}$$

The noise intensity grows from low to high as x increases to one. We use $n_{\text{sim-bs}} = n_{\text{sim-theo}} = n_{\text{sim-calib}} = 1000$ samples for computing the bootstrap marginal quantiles and the theoretical quantiles and for checking the calibration condition. The maximal model dimension is $M = 37$ and we also choose $m^\dagger = 20$. The calibration is run with $x = 2$ and $\beta = 1$.

We start by considering examples for $W = \Psi_n^\top$, i.e. the estimation of the whole function vector with prediction loss. One can see in Figure 5.1 three examples with different intensity of the noise term comparing the Bootstrap-method to the oracle estimator and the known-variance SmA-Method. Figure 5.2 illustrates the dependence of the choice of the estimated dimension on our calibration dimension m^\dagger and the sample size n . We see that in the specific example we are considering, the sensitivity of the chosen dimension \tilde{m} on m^\dagger decreases very fast. In the case $n = 200$, we have no variation in the choice of \tilde{m} with respect to m^\dagger . The oracles are respectively $m^* = 12$ for $n = 100, 200$ and $m^* = 10$ for $n = 50$. We also want to compare the true quantiles and their bootstrap

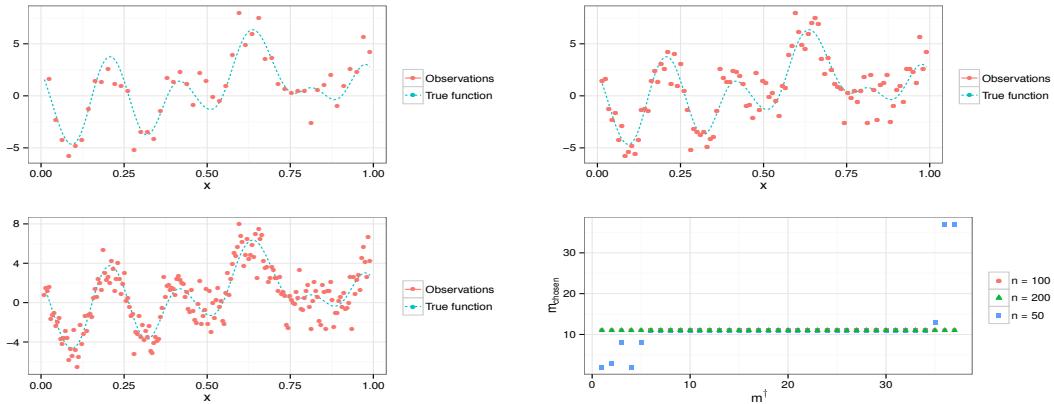


Fig. 5.2. The first three plots show an exemplary function with $n = 50, 100, 200$ observations. The right plot shows the \hat{m} chosen by the Bootstrap-SmA-Method as a function of the calibration dimension m^\dagger and the number of observations.

substitute. Figure 5.3 plots the ratios of quantiles for all possible comparisons (m_1, m_2) for the same function as before. Here we see that there is, as one would expect, still significant variation in the quantile ratios for small differences $|m_1 - m_2|$. Nonetheless the method works very well as seen in Fig. 5.2, but the variability in the ratios implies the possibility to stabilize the procedure even more by introducing some smoothing scheme for the quantiles.

Figure 5.4 again demonstrates the dependence of the ratios on m^\dagger . It is remarkable that the ratio is varying very slowly above $m^* = 12$. We also give the results on the simulation of $n_{\text{hist}} = 100$ repeated applications of the method to the same true underlying function observed with different realizations of the errors in Figure 5.5.

The case of the estimation of the first derivative is similar. Figure 5.6 shows the numerical results for estimation of the derivative in the same model as above. One can see that the bootstrap-version of the SmA-procedure is again competitive with the procedure

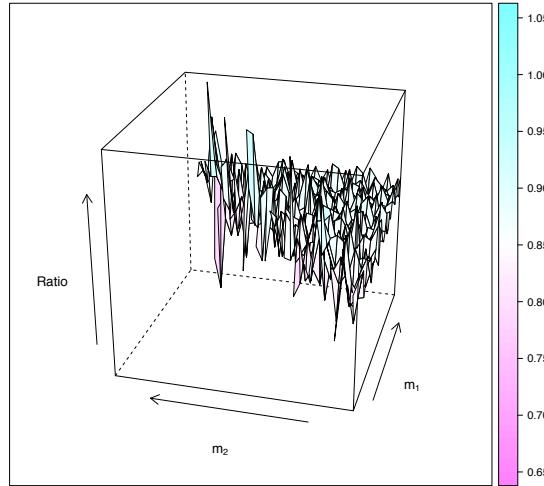


Fig. 5.3. Ratio of quantiles $|z_{m_1, m_2}^b / z_{m_1, m_2}|^2$ for $m^\dagger = 20$ and $n = 200$ with the data and true function as in Fig. 5.2.

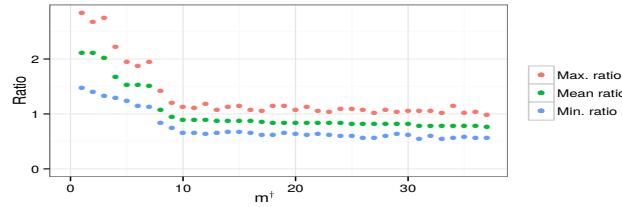


Fig. 5.4. Maximal, minimal and mean ratio of the bootstrap and theoretical tail functions at $x = 2$, $|z_{m_1, m_2}^b / z_{m_1, m_2}|^2$ as a function of m^\dagger .

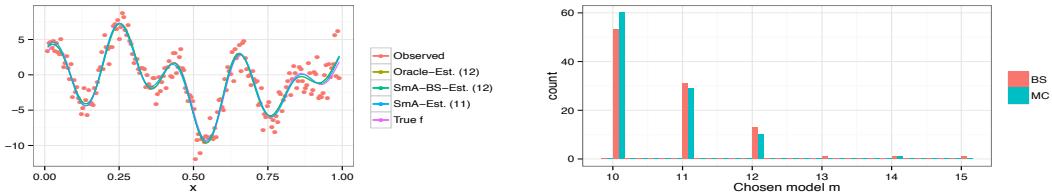


Fig. 5.5. In the left plot, the true function and observed values are plotted for one realization together with the oracle estimator, the known-variance SmA-Estimator (SmA-Est.) and the Bootstrap-SmA-Estimator (SmA-BS-Est.). The numbers in parentheses indicate the chosen model dimension. In the right plot, histograms for the selected model are given for the bootstrap (BS) and the known-variance method (MC) for repeated observations of the same underlying function with a simulation size $n_{\text{hist}} = 100$.

based on a known noise structure and the method does a good job of mimicking the oracle.

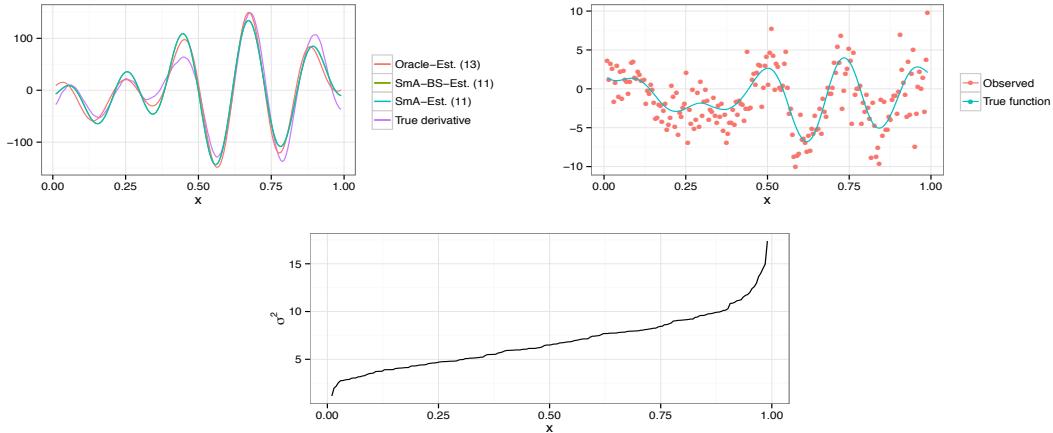


Fig. 5.6. The upper left plot shows the true derivative, the oracle estimator, the known-variance SmA-Estimator (SmA-Est.) and the Bootstrap-SmA-Estimator (SmA-BS-Est.). The upper right plot shows the true function and the observations and in the lower plot one can find the standard deviation of the errors.

One can conclude that the proposed procedure is really universal and demonstrates a very good performance in various settings.

5.4 Proofs

This section collects the proofs of announced results.

5.4.1 Proof of Theorem 5.1.1

The propagation property (5.13) claims that the oracle model m^* will be accepted with high probability. This yields that the selected model is not larger than m^* , that is, $\hat{m} \leq m^*$ with a probability at least $1 - e^{-x}$. Below we consider only this event. Let $m \in \mathcal{M}_-(m^*)$. Acceptance of m requires in particular that $\mathbb{T}_{m^*,m} \leq z_{m^*,m}$. The representation $\mathbb{T}_{m^*,m} = \|b_{m^*,m} + \xi_{m^*,m}\|$ implies

$$\mathbb{P}(\mathbb{T}_{m^*,m} < z_{m^*,m}) \leq \mathbb{P}(\|\xi_{m^*,m}\| > \|b_{m^*,m}\| - z_{m^*,m}).$$

Under (5.14) this yields

$$\begin{aligned} \mathbb{P}(m \text{ is accepted}) &\leq \mathbb{P}(\|b_{m^*,m} + \xi_{m^*,m}\| \leq z_{m^*,m}) \\ &\leq \mathbb{P}(\|\xi_{m^*,m}\| \geq z_{m^*,m}(x_s)) \leq e^{-x_s}. \end{aligned} \quad (5.43)$$

If the lower bound on the bias is fulfilled for all $m \in \mathcal{M}_-$, then (5.43) helps to bound the probability of the event $\{\hat{m} \in \mathcal{M}_-\}$:

$$\mathbb{P}(\hat{m} \in \mathcal{M}_-) \leq \sum_{m \in \mathcal{M}_-} \mathbb{P}(\|b_{m^*,m} + \xi_{m^*,m}\| < z_{m^*,m}) \leq \sum_{m \in \mathcal{M}_-} e^{-x_s} \leq e^{-x}.$$

Therefore, the probability that the SmA-selector picks up a value $m > m^*$ or $m \in \mathcal{M}_-$ is very small:

$$\mathbb{P}(\hat{m} \in \mathcal{M}_+(m^*) \cup \mathcal{M}_-) \leq 2e^{-x}.$$

It remains to study the case when $\hat{m} = m \in \mathcal{M}^\circ = \mathcal{M}_-(m^*) \setminus \mathcal{M}_-$. We can use that \hat{m} is accepted, which implies by definition

$$\mathbb{T}_{m^*,m} = \|\tilde{\phi}_m - \tilde{\phi}_{m^*}\| \leq z_{m^*,m}.$$

This yields (5.15). The bound (5.17) now follows by the triangle inequality.

5.4.2 Proof of Proposition 5.1.2

Below we use the deviation bound (B.3) for a Gaussian quadratic form from Theorem B.1.1. Note that similar results are available for non-Gaussian quadratic forms under exponential moment conditions; see e.g. Spokoiny (2012). The result (B.3) combined with the Bonferroni correction $q_{m^\circ} = \log(|\mathcal{M}_+(m^\circ)|) \leq \log(|\mathcal{M}|)$ yields the following upper bound for the critical values z_{m,m° :

$$\begin{aligned} z_{m,m^\circ} &\leq z_{m,m^\circ}(x + q_{m^\circ}) + \beta p_{m,m^\circ}^{1/2} \\ &\leq (1 + \beta)\sqrt{p_{m,m^\circ}} + \sqrt{2\lambda_{m,m^\circ} \{x + \log(|\mathcal{M}_+(m^\circ)|)\}} \\ &\leq (1 + \beta)\sqrt{p_{m,m^\circ}} + \sqrt{2\lambda_{m,m^\circ} \{x + \log(|\mathcal{M}|)\}}. \end{aligned} \tag{5.44}$$

For the payment for adaptation \bar{z}_{m^*} , the result (5.44) and the monotonicity condition $p_{m^*,m} \leq p_{m^*,m_0} \leq p_{m^*}$ and $\lambda_{m^*,m} \leq \lambda_{m^*,m_0} \leq \lambda_{m^*}$ imply the following upper bound:

$$\begin{aligned} \bar{z}_{m^*} &\leq (1 + \beta)\sqrt{p_{m^*,m_0}} + \sqrt{2\lambda_{m^*,m_0} \{x + \log(|\mathcal{M}_-(m^*)|\})} \\ &\leq (1 + \beta)\sqrt{p_{m^*}} + \sqrt{2\lambda_{m^*} \{x + \log(|\mathcal{M}|)\}} \end{aligned}$$

which yields the claim.

5.4.3 Proof of Theorem 5.1.3

The result will be proved in two steps. First we bound the risk on the set $\hat{m} > m^*$:

$$\mathbb{E}\{\|\hat{\phi} - \phi^*\|^2 \mathbb{I}(\hat{m} > m^*)\} \leq 2\bar{\alpha}_{m^*} \mathcal{R}_{m^*}. \quad (5.45)$$

Then we consider the region $\hat{m} < m^*$ and prove an oracle inequality

$$\|\hat{\phi} - \tilde{\phi}_{m^*}\| \mathbb{I}(\hat{m} < m^*) \leq \bar{z}_{m^*} \quad (5.46)$$

and the oracle bound (5.25). We start by proving (5.45). Let us fix $m \in \mathcal{M}_+(m^*)$ and $m' \geq m$. The definition (5.18) of the oracle m^* and the formula (5.22) for the critical value $z_{m',m_{(-1)}}$ implies for the test statistic $\mathbb{T}_{m',m_{(-1)}} = \|\xi_{m',m_{(-1)}} + b_{m',m_{(-1)}}\|$

$$\{\mathbb{T}_{m',m_{(-1)}} > z_{m',m_{(-1)}}\} \subseteq \{\|\xi_{m',m_{(-1)}}\| > z_{m',m_{(-1)}}(\mathbf{x}_{m_{(-1)}})\}.$$

Now we can bound the risk of $\hat{\phi}$ on the set $\hat{m} > m^*$. We use that for $\hat{m} = m > m^*$ in view of (5.19)

$$\begin{aligned} \|\hat{\phi} - \phi^*\|^2 &= \|\tilde{\phi}_m - \phi^*\|^2 = \|\xi_m + b_m\|^2 \\ &\leq 2\|\xi_m\|^2 + 2\|b_m\|^2 \leq 2\|\xi_m\|^2 + 2\|b_{m^*}\|^2 \end{aligned}$$

and it holds by (5.21) and monotonicity $p_m > p_{m^*}$

$$\begin{aligned} &\mathbb{E}\{\|\hat{\phi} - \phi^*\|^2 \mathbb{I}(\hat{m} > m^*)\} \\ &\leq 2 \sum_{m \in \mathcal{M}_+(m^*)} \mathbb{E}\{(\|\xi_m\|^2 + \|b_{m^*}\|^2) \mathbb{I}(\hat{m} = m)\} \\ &\leq 2 \sum_{m \in \mathcal{M}_+(m^*)} \mathbb{E}\{(\|\xi_m\|^2 + \|b_{m^*}\|^2) \mathbb{I}(m_{(-1)} \text{ is rejected})\} \\ &= 2 \sum_{m \in \mathcal{M}_+(m^*)} \mathbb{E}\left[(\|\xi_m\|^2 + \|b_{m^*}\|^2) \mathbb{I}\left(\max_{m' \in \mathcal{M}_+(m)} \{\|\xi_{m',m_{(-1)}}\| - z_{m',m_{(-1)}}(\mathbf{x}_m)\} > 0\right)\right] \\ &\leq 2 \sum_{m \in \mathcal{M}_+(m^*)} \alpha_m (p_m + \|b_{m^*}\|^2) \leq 2\bar{\alpha}_{m^*} (p_{m^*} + \|b_{m^*}\|^2) = 2\bar{\alpha}_{m^*} \mathcal{R}_{m^*}. \end{aligned}$$

Here we have used that (5.24) and $p_m \geq p_{m^*}$ imply $\sum_{m \in \mathcal{M}_+(m^*)} \alpha_m \leq \bar{\alpha}_{m^*}$. This completes the proof of (5.45).

In the situation when $\hat{m} = m < m^*$, we can use the stability property: as m is accepted, it holds

$$\|\tilde{\phi}_m - \tilde{\phi}_{m^*}\| \mathbb{I}(\hat{m} = m) \leq z_{m^*,m},$$

which implies (5.46) by definition of \bar{z}_{m^*} . This yields

$$\begin{aligned} I\!\!E \|\hat{\phi} - \phi^*\|^2 &\leq 2\bar{\alpha}_{m^*} \mathcal{R}_{m^*} + I\!\!E \{ \|\hat{\phi} - \phi^*\|^2 \mathbb{I}(\hat{m} < m^*) \} \\ &\leq 2\bar{\alpha}_{m^*} \mathcal{R}_{m^*} + I\!\!E (\|\tilde{\phi}_{m^*} - \phi^*\| + \bar{z}_{m^*})^2 \\ &\leq 2\bar{\alpha}_{m^*} \mathcal{R}_{m^*} + (\mathcal{R}_{m^*}^{1/2} + \bar{z}_{m^*})^2 \end{aligned}$$

as required.

5.4.4 Proof of Proposition 5.1.1

Observe first that the choice $\alpha_m = (p_m/p_{m_0})^{-1-a}$ yields

$$\sum_{m \in \mathcal{M}_+(m^*)} \alpha_m p_m \leq p_{m_0}^{1+a} \sum_{m \in \mathcal{M}_+(m^*)} p_m^{-a} \leq C p_{m^*}^{-a} p_{m_0}^{1+a} = C \bar{\alpha}_{m^*} p_{m^*}$$

with $\bar{\alpha}_{m^*} = C(p_{m_0}/p_{m^*})^{1+a}$.

For any random vector ξ with $\text{Var}(\xi) = B$ and $p = \text{tr}(B)$ and any random event A , it holds

$$I\!\!E \left[p^{-1} \|\xi\|^2 \mathbb{I}(A) \right] \leq \{1 + p^{-2} \text{Var}(\|\xi\|^2)\}^{1/2} I\!\!P^{1/2}(A). \quad (5.47)$$

Indeed, the Cauchy-Schwartz inequality implies

$$\begin{aligned} I\!\!E \left\{ p^{-1} \|\xi\|^2 \mathbb{I}(A) \right\} &\leq I\!\!E^{1/2} \{ p^{-1} \|\xi\|^2 \}^2 I\!\!P^{1/2}(A) \\ &= \{1 + p^{-2} \text{Var}(\|\xi\|^2)\}^{1/2} I\!\!P^{1/2}(A). \end{aligned}$$

Moreover, in the Gaussian case $\xi \sim \mathcal{N}(0, B)$ with $\|B\|_{\text{op}} \leq 1$, it holds $\text{Var}(\|\xi\|^2) \leq 2p$. If p is large then $\text{Var}(\|\xi\|^2)/p^2$ is small. In general $\text{Var}(\|\xi\|^2)/p^2 \leq 2$.

Result (5.47) and the choice $\alpha_m = \sqrt{3} p_m^{-1-a}$ allow to specify an upper bound on x_m . Namely, the choice $x_m = C \log(p_m)$ ensures the propagation condition (5.20). To see this, fix m and $m' \geq m$. Let

$$A'_m(x) \stackrel{\text{def}}{=} \mathbb{I} \left(\max_{m' \in \mathcal{M}_+(m)} \{ \|\xi_{m',m}\| - \sqrt{p_{m',m}} - \sqrt{2\lambda_{m',m} \{x + \log(|\mathcal{M}|)\}} \} > 0 \right)$$

The arguments after Lemma B.1.1 with $x_{m(-1)} = 2(1+a) \log(p_m)$ and (5.47) imply

$$I\!\!E \left[p_m^{-1} \|\xi_m\|^2 \mathbb{I}\{A'_{m(-1)}(x_{m(-1)})\} \right] \leq \sqrt{3} e^{-(1+a) \log(p_m)} = \sqrt{3} p_m^{-1-a}$$

and by (5.22)

$$z_{m,m^\circ} \leq \sqrt{p_{m,m^\circ}} + \sqrt{2\lambda_{m,m^\circ} \{(1+a) \log(p_{m^\circ+1}) + \log(|\mathcal{M}|)\}}.$$

This implies the upper bound (5.28) on the payment for adaptation \bar{z}_{m^*} .

5.4.5 Proof of Theorem 5.2.1

Any statement on the use of bootstrap-tuned parameters faces the same fundamental problem: the bootstrap distribution is random and depends on the underlying sample. When we use such values for the original procedure, we have to account for this dependence. The statement of Theorem 5.2.1 is even more involved due to the presmoothing step and multiplicity correction (5.35). The proof will be split into a couple of steps. First we evaluate the effect of the presmoothing bias and variance and reduce the study to an artificial situation where one uses the errors ε_i for resampling in place of the residuals \check{Y}_i . Then we compare \mathbb{Q} and \mathbb{Q}^b using the Pinsker inequality.

Below we write Ψ in place of Ψ_M , where M is the largest model in the collection. This does not conflict with our general setup, it is implicitly assumed that the largest model coincides with the original one. By p we denote the corresponding parameter dimension, that is, Ψ is a $p \times n$ matrix. Further, the feature matrix Ψ_m can be written as the product $\Psi_m = \Pi_m \Psi$, where Π_m is the projector on the subspace of the feature space spanned by the features from the model m : $\Pi_m = \Psi_m^\top (\Psi_m \Psi_m^\top)^{-1} \Psi_m$. This allows to represent each estimator $\tilde{\phi}_m$ in the form

$$\begin{aligned}\tilde{\phi}_m &= W \tilde{\theta}_m = W \mathcal{S}_m \mathbf{Y} = W (\Psi_m \Psi_m^\top)^{-1} \Psi_m \mathbf{Y} = \mathcal{T}_m \Psi \mathbf{Y} \\ \mathcal{T}_m &\stackrel{\text{def}}{=} W (\Psi_m \Psi_m^\top)^{-1} \Pi_m.\end{aligned}$$

This implies the following representation of the stochastic components ξ_{m,m° :

$$\xi_{m,m^\circ} = \mathcal{T}_{m,m^\circ} \Psi \varepsilon = \mathcal{T}_{m,m^\circ} \nabla, \quad \mathcal{T}_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{T}_m - \mathcal{T}_{m^\circ},$$

where $\nabla = \Psi \varepsilon$. One can say that each stochastic vector ξ_{m,m° is a linear function of the vector ∇ . A similar representation holds true in the bootstrap world:

$$\xi_{m,m^\circ}^b = \mathcal{T}_{m,m^\circ} \Psi \operatorname{diag}(\check{\mathbf{Y}}) \mathbf{w}^b = \mathcal{T}_{m,m^\circ} \nabla^b, \quad \nabla^b \stackrel{\text{def}}{=} \Psi \operatorname{diag}(\check{\mathbf{Y}}) \mathbf{w}^b.$$

Here the original errors ε are replaced by their bootstrap surrogates $\varepsilon^b = \operatorname{diag}(\check{\mathbf{Y}}) \mathbf{w}^b$. Therefore, it suffices to compare the distribution of $\nabla = \Psi \varepsilon$ with the conditional distribution of $\nabla^b = \Psi \operatorname{diag}(\check{\mathbf{Y}}) \mathbf{w}^b$ given \mathbf{Y} . Then the results will be automatically extended to any deterministic mapping of these two vectors.

Normality of the errors $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ implies that $\nabla = \Psi \varepsilon$ is also normal zero mean:

$$\nabla \sim \mathcal{N}(0, S), \quad S \stackrel{\text{def}}{=} \Psi \Sigma \Psi^\top, \quad \Sigma = \operatorname{Var}(\varepsilon) = \operatorname{diag}(\sigma_1^2, \dots, \sigma_n^2).$$

Similarly we can use standard normality of the bootstrap weights w_i^b . Given the data \mathbf{Y} , the vector ∇^b is conditionally normal zero mean with the conditional variance

$$S^\flat \stackrel{\text{def}}{=} \text{Var}^\flat(\nabla^\flat) = \Psi \text{diag}(\check{Y}_1^2, \dots, \check{Y}_n^2) \Psi^\top = \Psi \text{diag}(\check{\mathbf{Y}} \cdot \check{\mathbf{Y}}) \Psi^\top.$$

Therefore, the problem is reduced to comparing two p -dimensional Gaussian distributions with different covariance matrices. Equivalently, we have to bound the value $\Delta = \sqrt{\text{tr}(\mathcal{B}^2)}$ for a random $p \times p$ matrix \mathcal{B} given by

$$\mathcal{B} \stackrel{\text{def}}{=} S^{-1/2}(S^\flat - S)S^{-1/2}.$$

Define a $p \times n$ matrix $\mathcal{U} = S^{-1/2}\Psi\Sigma^{1/2}$ so that $\mathcal{U}\mathcal{U}^\top = I_p$. We will use the decomposition

$$\Sigma^{-1/2}\check{\mathbf{Y}} = \Sigma^{-1/2}(\mathbf{Y} - \Pi\mathbf{Y}) = \Sigma^{-1/2}(\boldsymbol{\varepsilon} - \Pi\boldsymbol{\varepsilon}) + \Sigma^{-1/2}(\mathbf{f}^* - \Pi\mathbf{f}^*) = \boldsymbol{\eta} + \mathbf{B}$$

with

$$\boldsymbol{\eta} \stackrel{\text{def}}{=} \Sigma^{-1/2}(\boldsymbol{\varepsilon} - \Pi\boldsymbol{\varepsilon}), \quad \mathbf{B} \stackrel{\text{def}}{=} \Sigma^{-1/2}(\mathbf{f}^* - \Pi\mathbf{f}^*). \quad (5.48)$$

With the matrix \mathcal{B} can now be represented as

$$\begin{aligned} \mathcal{B} &= \mathcal{U} \text{diag}\{(\boldsymbol{\eta} + \mathbf{B}) \cdot (\boldsymbol{\eta} + \mathbf{B}) - I_n\} \mathcal{U}^\top && (5.49) \\ &= \mathcal{U} \text{diag}\{(\boldsymbol{\eta} + \mathbf{B}) \cdot (\boldsymbol{\eta} + \mathbf{B}) - \boldsymbol{\eta} \cdot \boldsymbol{\eta}\} \mathcal{U}^\top &\stackrel{\text{def}}{=} \mathcal{B}_1 \\ &\quad + \mathcal{U} \text{diag}\{\boldsymbol{\eta} \cdot \boldsymbol{\eta} - \mathbb{E}(\boldsymbol{\eta} \cdot \boldsymbol{\eta})\} \mathcal{U}^\top &\stackrel{\text{def}}{=} \mathcal{B}_2 \\ &\quad + \mathcal{U} \text{diag}\{\mathbb{E}(\boldsymbol{\eta} \cdot \boldsymbol{\eta}) - I_n\} \mathcal{U}^\top &\stackrel{\text{def}}{=} \mathcal{B}_3 \end{aligned}$$

The first term \mathcal{B}_1 in this decomposition expresses the impact of the bias \mathbf{B} remaining after presmoothing, the last two terms \mathcal{B}_2 and \mathcal{B}_3 measure the change of the noise covariance due to presmoothing. The triangle inequality in the Frobenius norm $\|\mathcal{B}\|_{\text{Fr}} \stackrel{\text{def}}{=} \sqrt{\text{tr}(\mathcal{B}^2)}$ and bounds from Propositions ??, ??, and ?? with $\mathcal{U}\mathcal{U}^\top = I_p$ and $p = p = \text{tr}(\mathcal{U}\mathcal{U}^\top) = p$ imply on a random set $\Omega_2(\mathbf{x}) = \Omega_{12}(\mathbf{x}) \cup \Omega_{22}(\mathbf{x})$ with $\mathbb{P}(\Omega_2(\mathbf{x})) \geq 1 - 2e^{-x}$

$$\begin{aligned} \|\mathcal{B}\|_{\text{Fr}} &\leq \|\mathcal{B}_1\|_{\text{Fr}} + \|\mathcal{B}_2\|_{\text{Fr}} + \|\mathcal{B}_3\|_{\text{Fr}} \\ &\leq \Delta_1(\mathbf{x}) + \Delta_2(\mathbf{x}) + \Delta_3(\mathbf{x}) \\ &= 2\sqrt{\delta_\Psi^2 p(x + \log(n))} + \sqrt{\delta_\varepsilon^2 p} + \sqrt{\|\mathbf{B}\|_\infty^4 p} + 4\delta_\Psi^2 \|\mathbf{B}\| (1 + \sqrt{x}). \end{aligned}$$

This proves (5.40) in view of Pinsker's Lemma D.1.1 with $\mathbf{b} = \mathbf{b}^\flat = 0$.

5.4.6 Proof of Theorem 5.2.2

The result of Theorem 5.2.1 justifies the bootstrap-phenomenon, namely it explains why the known bootstrap distribution can be used as a proxy for the unknown error distribution. However, it cannot be applied directly to (5.41) because the quantities $z_{m,m^\circ}^\flat(\mathbf{x})$

and q_m^b are random and depend on the original data. This especially concerns the multiplicity correction q_m^b which is based on the joint distribution of the vectors ξ_{m,m°^b from (5.33) and is defined in (5.35). The latter distribution is a random measure in the bootstrap world which is normal conditioned on the original sample. To cope with the problem of this cross-dependence, we apply the statement of Theorem E.1.1 in the Appendix. The underlying idea is to use geometric arguments to sandwich the random probability in (5.35) in two deterministic probabilities. Then the error of bootstrap approximation can again be bounded by using the Pinsker inequality. The statement of Theorem 5.2.2 can be derived from Theorem E.1.1 if an operator norm bound $\|\mathcal{B}\|_{\text{op}}$ is available. Note that Theorem 5.2.1 only requires a bound for the Frobenius norm. By Proposition C.2.1, it holds with $\delta_n = \delta_\Psi$, $\mathbf{x}_n = \mathbf{x} + \log(n)$, and $\mathbf{x}_p = \mathbf{x} + 2 \log(p)$

$$\begin{aligned} \|\mathcal{B}\|_{\text{op}} &\leq \Delta_{\text{op}}(\mathbf{x}), \\ \Delta_{\text{op}}(\mathbf{x}) &\stackrel{\text{def}}{=} \|\mathbf{B}\|_\infty^2 + \delta_\Psi^2 \|\mathbf{B}\| \sqrt{2\mathbf{x}} + 2\delta_\Psi \mathbf{x}_p^{1/2} + 2\delta_\Psi^2 \mathbf{x}_p + 2\delta_\Pi \mathbf{x}_n + \delta_\Pi^2 \mathbf{x}_n. \end{aligned}$$

The result of the theorem follows now by Theorem E.1.1.

5.4.7 Proof of Theorem 5.2.3

For a fixed pair $m > m^\circ$ from \mathcal{M} , consider $\mathbf{p}_{m,m^\circ}^b = \mathbb{E}^b \|\xi_{m,m^\circ}^b\|^2$ and $\mathbf{p}_{m,m^\circ} = \mathbb{E} \|\xi_{m,m^\circ}\|^2$. As $\text{diag}(\check{\mathbf{Y}})$ and Σ are diagonal matrices, the definitions (5.33) and (5.48) imply

$$\begin{aligned} \xi_{m,m^\circ}^b &= \mathcal{K}_{m,m^\circ} \text{diag}(\check{\mathbf{Y}}) \mathbf{w}^b = \mathcal{K}_{m,m^\circ} \Sigma^{1/2} \Sigma^{-1/2} \text{diag}(\check{\mathbf{Y}}) \mathbf{w}^b \\ &= \mathcal{U}_{m,m^\circ} \text{diag}(\boldsymbol{\eta} + \mathbf{B}) \mathbf{w}^b, \end{aligned}$$

where $\mathcal{U}_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{K}_{m,m^\circ} \Sigma^{1/2}$. It holds for \mathbf{p}_{m,m°^b

$$\mathbf{p}_{m,m^\circ}^b = \mathbb{E}^b \|\xi_{m,m^\circ}^b\|^2 = \text{tr}(\mathcal{U}_{m,m^\circ} \text{diag}\{(\boldsymbol{\eta} + \mathbf{B}) \cdot (\boldsymbol{\eta} + \mathbf{B})\} \mathcal{U}_{m,m^\circ}^\top)$$

while $\xi_{m,m^\circ} = \mathcal{K}_{m,m^\circ} \Sigma^{1/2} \Sigma^{-1/2} \boldsymbol{\varepsilon}$ and

$$\mathbf{p}_{m,m^\circ} = \mathbb{E} \|\xi_{m,m^\circ}\|^2 = \text{tr}(\mathcal{U}_{m,m^\circ} \mathcal{U}_{m,m^\circ}^\top).$$

As we are interested in the ratio $\mathbf{p}_{m,m^\circ}^b / \mathbf{p}_{m,m^\circ}$, one can assume without loss of generality that $\|\mathcal{U}_{m,m^\circ} \mathcal{U}_{m,m^\circ}^\top\|_{\text{op}} = 1$ and $\mathbf{p}_{m,m^\circ} \geq 1$. Now we again apply the decomposition (5.49). The bounds (??) of Proposition ??, (??) of Proposition ??, and (??) of Proposition ?? imply on a set $\Omega_{m,m^\circ}(\mathbf{x})$ with $\mathbb{P}(\Omega_{m,m^\circ}(\mathbf{x})) \geq 1 - 3e^{-\mathbf{x}}$

$$\left| \frac{\mathbf{p}_{m,m^\circ}^b}{\mathbf{p}_{m,m^\circ}} - 1 \right| \leq \|\mathbf{B}\|_\infty^2 + 4\mathbf{x}^{1/2} \delta_n^2 \|\mathbf{B}\| + 4\mathbf{x}^{1/2} \delta_n + 4\mathbf{x} \delta_n^2 + \delta_\varepsilon.$$

The choice of $\mathbf{x} = \mathbf{x}_{\mathcal{M}} = \mathbf{x} + 2 \log(|\mathcal{M}|)$ ensures a uniform bound for all pairs $m > m^{\circ}$ from \mathcal{M} .

5.5 Linear non-Gaussian case and GAR

This section briefly comment why the bootstrap procedure can be validated even if the true error distribution is not Gaussian. This means that we again consider the linear Gaussian likelihood and the corresponding qMLE $\tilde{\boldsymbol{\theta}}$ is given by $\tilde{\boldsymbol{\theta}} = D^{-2}\Psi\mathbf{Y}$, the errors $\boldsymbol{\varepsilon}$ are independent but no more Gaussian. The discussion of the previous section shows that the most challenging step of analysis is to check that two vectors $\nabla = \Psi\boldsymbol{\varepsilon}$ and $\nabla^b = \Psi\mathcal{E}^b\mathbf{Y}$ have a similar distribution under the corresponding measures. In the Gaussian case, both vectors are normal zero mean and it suffices to compare their covariance matrices. In the non-Gaussian case the situation is more involved. A nice feature of Gaussian bootstrap multipliers is that the distribution of $\nabla^b = \Psi\mathcal{E}^b\mathbf{Y}$ given \mathbf{Y} is again Gaussian, and this fact does not rely on the true data distribution. It is entirely due to the construction of the bootstrap multipliers: ∇^b is normal because it is a linear combination of standard normal weights $e_i^b = w_i^b - 1$. The real score $\nabla = \Psi\boldsymbol{\varepsilon}$ is again a linear combination of errors ε_i , however these errors can be non-normal. If fact, in typical applications, there is no reason to assume that the errors are exactly normal. However, $\Psi\boldsymbol{\varepsilon}$ can be viewed as a linear combination of the errors ε_i . In combination with the condition that the value δ_{Ψ} from (5.36) is small, the central limit theorem applies and the zero mean standardized vector $V^{-1}\nabla$ is nearly standard normal under some further regularity and moment conditions. This allows to extend the result on bootstrap validity to the non-Gaussian case in some asymptotic sense. In the univariate case with $p = 1$ one can use the famous Berry-Esseen theorem, which can be also extended to the multivariate case in various special setups.

Unordered case. Anisotropic sets and subset selection

The SmA method of the previous section is quite general and can be extended to many statistical models and problem. However, it essentially requires the ordered structure of the set of considered models/methods. This section discusses how the SmA procedure can be extended to some other setups without ordered structure. To distinguish ordered and unordered cases, we denote by $\mathcal{A} = \{\varkappa\}$ the set of all considered models. The basic idea is to assume a kind of partial ordering which enables to define an acceptance rule:

\varkappa° is accepted if it is *not rejected against any larger model*.

This rule allows to fix a set of accepted models. Further we need some global measure of complexity which can be used for final selection:

$\hat{\varkappa}$ is the *simplest accepted* model.

Below we illustrate how this method works in two important examples: *anisotropic classes* and *subset selection* problems.

6.1 Subset selection procedure

Consider a linear model $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \varepsilon$. The dimension p of the vector $\boldsymbol{\theta}$ can be very large and we implicitly assumes a kind of *sparse* structure:

most of $\boldsymbol{\theta}^*$ -entries are nearly zero and can be dropped, there is a relatively small subvector of $\boldsymbol{\theta}^*$ containing the important features.

The aim is to find this subset and to estimate the whole vector $\boldsymbol{\theta}^*$. As in the ordered case, one can separate between prediction $W = \Psi^\top$ and estimation $W = I_p$ loss. Of course, these two problems coincide in the sequence space model with $n = p$ and $\Psi = I_p$.

6.1.1 SmA procedure and multilevel synchronization

Let κ mean a subset of the entire index set $\{1, 2, \dots, p\}$. We use the obvious notation $\kappa \vee \kappa'$ for the union, $\kappa \wedge \kappa^*$ for the overlap of two subsets κ and κ' , $\kappa' - \kappa$ for the complement of κ within κ' . Further we consider the usual partial ordering: $\kappa' > \kappa$ means that $\kappa \subseteq \kappa'$. The subset κ is good if there is no significant bias in its complement κ^c . The approach is to design a procedure which accepts any such good model with a high probability. The proposed SmA rule will be again to select the simplest (smallest in complexity) accepted model.

The acceptance rule is based on pairwise comparison with a family of tests T_{κ, κ^o} for $\kappa > \kappa^o$. The model-candidate κ^o is accepted if no test among T_{κ, κ^o} rejects the hypothesis of “no bias”. Given the loss matrix W , the test statistic T_{κ, κ^o} reads as in the ordered case:

$$T_{\kappa, \kappa^o} = \|W(\tilde{\theta}_\kappa - \tilde{\theta}_{\kappa^o})\|. \quad (6.1)$$

The acceptance rule can be written as

$$\kappa^o \text{ is accepted iff } T_{\kappa, \kappa^o} \leq z_{\kappa, \kappa^o} \quad \forall \kappa > \kappa^o. \quad (6.2)$$

Now we discuss how the critical values z_{κ, κ^o} can be fixed by *synchronization (multiplicity correction)* of the individual *tail functions*. We use the decomposition of the test statistic T_{κ, κ^o} from (6.1):

$$T_{\kappa, \kappa^o} = \|\xi_{\kappa, \kappa^o} + b_{\kappa, \kappa^o}\|.$$

The increase of complexity between κ^o and κ can be measured by via the variance of ξ_{κ, κ^o} . Namely, define p_{κ, κ^o} as expectation of $\|\xi_{\kappa, \kappa^o}\|^2$:

$$p_{\kappa, \kappa^o} \stackrel{\text{def}}{=} I\!\!E \|\xi_{\kappa, \kappa^o}\|^2 = \text{tr}\{\text{Var}(\xi_{\kappa, \kappa^o})\}.$$

This value will be used in the bias-variance relation: the bias b_{κ, κ^o} is insignificant if $\|b_{\kappa, \kappa^o}\|^2$ is smaller than the variance p_{κ, κ^o} . More precisely, define a *good choice* κ^o as previously by “no significant bias” condition:

$$\|b_{\kappa, \kappa^o}\| \leq \beta p_{\kappa, \kappa^o}^{1/2} \quad \forall \kappa > \kappa^o. \quad (6.3)$$

We aim at designing a procedure which accepts any such good model with a high probability. Suppose we are given for each pair $\kappa > \kappa^o$ a tail function $z_{\kappa, \kappa^o}(x)$ of the noise component $\|\xi_{\kappa, \kappa^o}\|$ providing

$$I\!\!P(\|\xi_{\kappa, \kappa^o}\| > z_{\kappa, \kappa^o}(x)) \leq e^{-x}.$$

This tail function can be used for testing the hypothesis of no significant bias component in the test statistic $\mathbb{T}_{\varkappa, \varkappa^\circ}$. The model-candidate \varkappa° is accepted by the SmA method if all such tests for $\varkappa > \varkappa^\circ$ do. To keep the overall test level, we have to synchronize all performed $\mathbb{T}_{\varkappa, \varkappa^\circ}$ -based tests by correcting for multiple check. The simplest way of multiplicity correction is done by a uniform increase of the level \mathbf{x} to control the overall rejection probability: define $q_{\varkappa^\circ}(\mathbf{x})$ by the condition

$$\mathbb{P}\left(\bigcup_{\varkappa \in \mathcal{M}(\varkappa^\circ)} \left\{ \|\boldsymbol{\xi}_{\varkappa, \varkappa^\circ}\| > z_{\varkappa, \varkappa^\circ}(\mathbf{x} + q_{\varkappa^\circ}(\mathbf{x})) \right\}\right) \leq e^{-\mathbf{x}}.$$

Denote $\mathbf{x}_{\varkappa^\circ} = \mathbf{x} + q_{\varkappa^\circ}(\mathbf{x})$ and apply the acceptance rule (6.2) with $\mathbf{z}_{\varkappa, \varkappa^\circ}$ equal to such defined $z_{\varkappa, \varkappa^\circ}(\mathbf{x}_{\varkappa^\circ})$ after a small bias correction:

$$\mathbf{z}_{\varkappa, \varkappa^\circ}(\beta) \stackrel{\text{def}}{=} z_{\varkappa, \varkappa^\circ}(\mathbf{x}_{\varkappa^\circ}) + \beta p_{\varkappa, \varkappa^\circ}^{1/2}.$$

One can use a more sophisticated *multilevel synchronization* procedure which accounts for the complexity of the alternative model \varkappa . Let $|\varkappa^\circ|$ mean the cardinality (complexity) of \varkappa° . For a given growing sequence $0 < \tau_1 < \tau_2 < \dots < \tau_K$, define

$$\mathcal{M}_m(\varkappa^\circ) \stackrel{\text{def}}{=} \{\varkappa > \varkappa^\circ : |\varkappa| + \tau_{m-1} < |\varkappa| \leq |\varkappa^\circ| + \tau_m\}$$

If $\tau_m = m$ for all m then

$$\mathcal{M}_m(\varkappa^\circ) \stackrel{\text{def}}{=} \{\varkappa > \varkappa^\circ : |\varkappa| = |\varkappa^\circ| + m\}.$$

The corrections $q_{1, \varkappa^\circ}, q_{2, \varkappa^\circ}, \dots, q_{m, \varkappa^\circ}$ can be defined step by step: first we fix the correction $q_{1, \varkappa^\circ} = q_{1, \varkappa^\circ}(\mathbf{x})$ for all $\varkappa \in \mathcal{M}_1(\varkappa^\circ)$

$$\mathbb{P}(\mathcal{A}_1) \leq \frac{1}{2}e^{-\mathbf{x}}$$

with

$$\mathcal{A}_1 \stackrel{\text{def}}{=} \bigcup_{\varkappa \in \mathcal{M}_1(\varkappa^\circ)} \left\{ \|\boldsymbol{\xi}_{\varkappa, \varkappa^\circ}\| > z_{\varkappa, \varkappa^\circ}(\mathbf{x} + q_{1, \varkappa^\circ}) \right\}.$$

With such defined q_{1, \varkappa° , define $q_{2, \varkappa^\circ} = q_{2, \varkappa^\circ}(\mathbf{x})$ such that

$$\begin{aligned} \mathbb{P}(\mathcal{A}_1 \cup \mathcal{A}_2) &\leq \frac{3}{4}e^{-\mathbf{x}}, \\ \mathcal{A}_2 &\stackrel{\text{def}}{=} \bigcup_{\varkappa \in \mathcal{M}_2(\varkappa^\circ)} \left\{ \|\boldsymbol{\xi}_{\varkappa, \varkappa^\circ}\| > z_{\varkappa, \varkappa^\circ}(\mathbf{x} + q_{1, \varkappa^\circ} + q_{2, \varkappa^\circ}) \right\}. \end{aligned}$$

We continue this way and define q_{m, \varkappa° by induction: if $q_{1, \varkappa^\circ}, \dots, q_{m-1, \varkappa^\circ}$ are fixed then the correction for the set $\mathcal{M}_m(\varkappa^\circ)$ is selected as the sum $\mathbf{x} + q_{1, \varkappa^\circ} + \dots + q_{m, \varkappa^\circ}$ to ensure

$$\mathbb{P}(\mathcal{A}_1 \cup \mathcal{A}_2 \cup \dots \cup \mathcal{A}_m) \leq (1 - 2^{-|m|})e^{-x} \quad (6.4)$$

with

$$\mathcal{A}_m \stackrel{\text{def}}{=} \bigcup_{\varkappa \in \mathcal{M}_m(\varkappa^\circ)} \left\{ \|\xi_{\varkappa, \varkappa^\circ}\| > z_{\varkappa, \varkappa^\circ}(x + q_{1, \varkappa^\circ} + q_{2, \varkappa^\circ} + \dots + q_{m, \varkappa^\circ}) \right\}.$$

For each $\varkappa > \varkappa^\circ$, there exists a unique $m = m(\varkappa)$ corresponding to the smallest set $\mathcal{M}_m(\varkappa^\circ)$ containing \varkappa . Finally, we define

$$z_{\varkappa, \varkappa^\circ} = z_{\varkappa, \varkappa^\circ}(x + q_{1, \varkappa^\circ} + q_{2, \varkappa^\circ} + \dots + q_{m, \varkappa^\circ}) + \beta p_{\varkappa, \varkappa^\circ}^{1/2}, \quad m = m(\varkappa). \quad (6.5)$$

Our selection rule chooses the smallest accepted model. It can be written as

$$\hat{\varkappa} = \operatorname{argmin}_{\varkappa^\circ \in \mathcal{M}} \{ |\varkappa^\circ| : \mathbb{T}_{\varkappa, \varkappa^\circ} \leq z_{\varkappa, \varkappa^\circ}, \quad \varkappa \in \mathcal{M}(\varkappa^\circ) \}. \quad (6.6)$$

If there are many such \varkappa° , one can select arbitrarily among them. The construction ensures that a good model in the sense (6.3) will be accepted with a high probability.

Theorem 6.1.1. *Let \varkappa° be a good model in the sense (6.3). Then it holds for the SmA procedure with the critical values $z_{\varkappa, \varkappa^\circ}$ from (6.4) and (6.5)*

$$\mathbb{P}(\varkappa^\circ \text{ is rejected}) \leq e^{-x}.$$

Now we define the oracle choice \varkappa^* as the simplest (in complexity $|\varkappa^*|$) model under the constraint (6.3):

$$\varkappa^* \stackrel{\text{def}}{=} \operatorname{argmin}_{\varkappa^\circ \in \mathcal{M}} \{ |\varkappa^\circ| : \|b_{\varkappa, \varkappa^\circ}\| \leq \beta p_{\varkappa, \varkappa^\circ}^{1/2}, \quad \varkappa \in \mathcal{M}(\varkappa^\circ) \}. \quad (6.7)$$

Again, this relation does not uniquely define the \varkappa^* value, if there are many \varkappa^* with this property, any of them can be taken. The oracle bound compares the risk of the oracle estimate $\mathcal{R}_{\varkappa^*}$ with the risk of the adaptive estimate $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}_{\hat{\varkappa}}$ for the SmA rule $\hat{\varkappa}$.

Theorem 6.1.2. *It holds on a random set of probability at least $1 - e^{-x}$*

$$\|W(\tilde{\boldsymbol{\theta}}_{\varkappa^*} - \tilde{\boldsymbol{\theta}}_{\hat{\varkappa}})\| \leq \bar{z}_{\varkappa^*}$$

with $k^* = |\varkappa^*|$ and \bar{z}_{\varkappa^*} defined by

$$\bar{z}_{\varkappa^*} \stackrel{\text{def}}{=} \max_{|\varkappa| \leq k^*} (z_{\varkappa \vee \varkappa^*, \varkappa^*} + z_{\varkappa \vee \varkappa^*, \varkappa}). \quad (6.8)$$

Proof. The construction and the propagation property ensures that \varkappa^* is accepted with a high probability $1 - e^{-x}$. Below we focus on this case. Then the adaptive choice $\hat{\varkappa}$ from (6.6) has to fulfill

$$|\widehat{\varkappa}| \leq |\varkappa^*|.$$

Due to partial ordering, the models \varkappa^* and $\widehat{\varkappa}$ are not directly comparable. We use the model $\check{\varkappa} = \varkappa^* \vee \widehat{\varkappa}$ which contains both and is the smallest one with this property. As \varkappa^* and $\widehat{\varkappa}$ are both accepted, it holds

$$\|W(\widetilde{\boldsymbol{\theta}}_{\check{\varkappa}} - \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}})\| \leq z_{\check{\varkappa}, \widehat{\varkappa}}, \quad \|W(\widetilde{\boldsymbol{\theta}}_{\check{\varkappa}} - \widetilde{\boldsymbol{\theta}}_{\varkappa^*})\| \leq z_{\check{\varkappa}, \varkappa^*}.$$

We conclude that

$$\|W(\widetilde{\boldsymbol{\theta}}_{\varkappa^*} - \widetilde{\boldsymbol{\theta}}_{\check{\varkappa}})\| \leq z_{\check{\varkappa}, \widehat{\varkappa}} + z_{\check{\varkappa}, \varkappa^*}.$$

and the assertion follows.

6.1.2 Prediction loss

The case of prediction loss ($W = \Psi^\top$) in combination with projection estimates $\widetilde{\boldsymbol{\theta}}_\varkappa$ allows to reduce the study to the sequence space model with $\Psi = I_p$ provided that $p \leq n$. This dramatically simplifies the situation. Below we denote by Π_\varkappa the projector in \mathbb{R}^n onto the subspace corresponding to \varkappa . For a couple $\varkappa^o < \varkappa$, we also consider the projector $\Pi_{\varkappa, \varkappa^o} = \Pi_\varkappa - \Pi_{\varkappa^o}$ which projects to the orthogonal complement of Π_{\varkappa^o} within the \varkappa -related subspace. Below we suppose a homogeneous noise ε with

$$\text{Var}(\varepsilon) = \sigma^2 I_n.$$

We also use that

$$p_{\varkappa, \varkappa^o} = \mathbb{E}\|\boldsymbol{\xi}_{\varkappa, \varkappa^o}\|^2 = \sigma^2 |\varkappa - \varkappa^o|.$$

Theorem 6.1.3. *For the regression model $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \varepsilon$ with homogeneous Gaussian errors and for $W = \Psi^\top$, the tail functions $z_{\varkappa, \varkappa^o}(\mathbf{x})$ only depends on $|\varkappa - \varkappa^o|$. Moreover, for \mathbf{x} fixed, the multiplicity corrections $q_{m, \varkappa^o} = q_{m, \varkappa^o}(\mathbf{x})$ only depends on $p - |\varkappa^o|$ and on τ_m .*

Proof. We use that $\Psi^\top \widetilde{\boldsymbol{\theta}}_\varkappa = \Pi_\varkappa \mathbf{Y}$ and

$$\sigma^{-2} \mathbb{T}_{\varkappa, \varkappa^o}^2 = \sigma^{-2} \|\Psi^\top (\widetilde{\boldsymbol{\theta}}_\varkappa - \widetilde{\boldsymbol{\theta}}_{\varkappa^o})\|^2 = \sigma^{-2} \|\Pi_{\varkappa, \varkappa^o} \mathbf{Y}\|^2 \sim \chi^2_{|\varkappa - \varkappa^o|}.$$

To be done: complete the proof

The definition (6.7) of the oracle choice \varkappa^* can be restated as

$$\varkappa^* \stackrel{\text{def}}{=} \underset{\varkappa^o \in \mathcal{M}}{\operatorname{argmin}} \{ |\varkappa^o| : \|b_{\varkappa, \varkappa^o}\| \leq \beta |\varkappa - \varkappa^o|^{1/2}, \quad \varkappa \in \mathcal{M}(\varkappa^o)\}.$$

Theorem 6.1.4. *The result of Theorem 6.1.2 holds with $\bar{z}_{\varkappa^*} = \bar{z}_{k^*}$ only depending on the cardinality $k^* = |\varkappa^*|$. Any \varkappa^* with this cardinality can be used in definition (6.8).*

To be done: An upper bound on q_{m,\varkappa^*} using Bonferroni and Stirling formulas

To be done: An upper bound on \bar{z}_{k^*}

To be done: Algorithmic implementation

6.1.3 Estimation loss

The analysis for the problem of estimation loss $W = I_p$ is similar, but some nice features of the prediction loss do not apply here. We use

$$\begin{aligned}\tilde{\boldsymbol{\theta}}_{\varkappa} &= \mathcal{S}_{\varkappa} \mathbf{Y}, \\ \mathcal{S}_{\varkappa} &= (\Psi_{\varkappa} \Psi_{\varkappa}^{\top})^{-1} \Psi_{\varkappa}.\end{aligned}$$

Therefore,

$$\mathbb{T}_{\varkappa,\varkappa^*} = \|(\mathcal{S}_{\varkappa} - \mathcal{S}_{\varkappa^*}) \mathbf{Y}\| = \|\boldsymbol{\xi}_{\varkappa,\varkappa^*} + b_{\varkappa,\varkappa^*}\|.$$

If the bias component vanishes, the distribution of this test statistic is completely described by the matrices Ψ_{\varkappa} and Ψ_{\varkappa^*} but in a more complicated way than for the prediction loss. Noise homogeneity allows us to define

$$p_{\varkappa,\varkappa^*} = \mathbb{E} \|\boldsymbol{\xi}_{\varkappa,\varkappa^*}\|^2 = \sigma^2 \operatorname{tr}\{(\mathcal{S}_{\varkappa} - \mathcal{S}_{\varkappa^*})(\mathcal{S}_{\varkappa} - \mathcal{S}_{\varkappa^*})^{\top}\}.$$

The general results of Theorems 6.1.1 and 6.1.2 apply here for such defined values $p_{\varkappa,\varkappa^*}$. Unfortunately, the nice simplification of the formulas as in the prediction case is only possible if the design is orthonormal. Then the estimation and prediction problems coincide.

6.1.4 Linear functional estimation

Now we briefly discuss the case when W is a matrix of rank one which corresponds to estimation of a linear functional. An advantage of this situation is that each difference $W(\tilde{\boldsymbol{\theta}}_{\varkappa'} - \tilde{\boldsymbol{\theta}}_{\varkappa})$ is univariate normal. This helps a lot in evaluating the distribution of each test statistic $\mathbb{T}_{\varkappa',\varkappa}$.

As in the ordered case, each estimate $\tilde{\phi}_{\varkappa} = W \tilde{\boldsymbol{\theta}}_{\varkappa}$ fulfills

$$\tilde{\phi}_{\varkappa} - \phi^* = W(\tilde{\boldsymbol{\theta}}_{\varkappa} - \boldsymbol{\theta}^*) = W \mathcal{S}_{\varkappa} \boldsymbol{\varepsilon} + W(\mathcal{S}_{\varkappa} \mathbf{f}^* - \boldsymbol{\theta}^*) = \xi_{\varkappa} + b_{\varkappa},$$

where $\xi_{\varkappa} = W \mathcal{S}_{\varkappa} \boldsymbol{\varepsilon}$ is a zero mean random variable and $b_{\varkappa} = W(\mathcal{S}_{\varkappa} \mathbf{f}^* - \boldsymbol{\theta}^*)$ is the deterministic bias. The squared risk of $\tilde{\phi}_{\varkappa}$ is given by the usual bias-variance decomposition:

$$\mathcal{R}_\kappa = \mathbb{E}(\tilde{\phi}_\kappa - \phi^*)^2 = \mathbb{E}(\xi_\kappa + b_\kappa)^2 = b_\kappa^2 + \text{Var}(\xi_\kappa) = b_\kappa^2 + s_\kappa^2$$

with

$$s_\kappa^2 = \sigma^2 W \mathcal{S}_\kappa \mathcal{S}_\kappa^\top W^\top.$$

Monotonicity condition $\mathcal{S}_\kappa \mathcal{S}_\kappa^\top \geq \mathcal{S}_{\kappa^\circ} \mathcal{S}_{\kappa^\circ}^\top$ for $\kappa > \kappa^\circ$ yields monotonicity $s_\kappa \geq s_{\kappa^\circ}$ for the functional estimate. For each pair $\kappa^\circ < \kappa$

$$\tilde{\phi}_\kappa - \tilde{\phi}_{\kappa^\circ} = W(\tilde{\theta}_\kappa - \tilde{\theta}_{\kappa^\circ}) = W(\mathcal{S}_\kappa - \mathcal{S}_{\kappa^\circ}) \mathbf{Y} = W \mathcal{S}_{\kappa, \kappa^\circ} \mathbf{Y}.$$

The variance of this difference reads as

$$s_{\kappa, \kappa^\circ}^2 = \text{Var}(\xi_{\kappa, \kappa^\circ}) = \sigma^2 W \mathcal{S}_{\kappa, \kappa^\circ} \mathcal{S}_{\kappa, \kappa^\circ}^\top W^\top.$$

The scaled test statistic $\mathbb{T}_{\kappa, \kappa^\circ}$ is given for $\kappa > \kappa^\circ$ by

$$\mathbb{T}_{\kappa, \kappa^\circ} = s_{\kappa, \kappa^\circ}^{-1} |W \mathcal{S}_{\kappa, \kappa^\circ} \mathbf{Y}|.$$

One can use that the stochastic component $\xi_{\kappa, \kappa^\circ}$ of $s_{\kappa, \kappa^\circ}^{-1} W \mathcal{S}_{\kappa, \kappa^\circ} \mathbf{Y}$ is standard normal. Thus, the multiplicity corrections are computed from the same bound (6.4) with $z_1(\cdot)$ in place of $z_{\kappa, \kappa^\circ}(\cdot)$. Moreover, $p_{\kappa, \kappa^\circ} \equiv 1$, and the oracle definition reads as

$$\kappa^* \stackrel{\text{def}}{=} \underset{\kappa^\circ \in \mathcal{M}}{\operatorname{argmin}} \{|\kappa^\circ| : \|b_{\kappa, \kappa^\circ}\| \leq \beta, \quad \kappa \in \mathcal{M}(\kappa^\circ)\}.$$

The general results of Theorems 6.1.1 and 6.1.2 apply here without any change. However, the involved values can be made more precise.

To be done: An upper bound on q_{m, κ°

To be done: An upper bound on \bar{z}_{k^*}

To be done: Algorithmic implementation

6.1.5 Subset selection problem

The presented oracle bound of Theorem 6.1.2 claims that the risk of the adaptive estimate $\hat{\theta}$ is linked to the risk of the oracle $\tilde{\theta}_{\kappa^*}$. However, it tells nothing about the selected set $\hat{\kappa}$. Now we discuss this issue of choosing the active set of important features represented by non-zero entries of θ^* .

Note that this problem has to be put in a right way: one can suppose that θ^* is sparse and only significant entries are non-zero. Alternatively, one tries to find an approximating model with another vector θ^* having a sparse representation and delivering nearly the same approximation and prediction quality. We follow our oracle result and define κ^*

by (6.7). This will be our target. We aim at finding some sufficient conditions ensuring that $\widehat{\varkappa} \approx \varkappa^*$. We already know that the set \varkappa^* will be accepted with a high probability. This particularly implies that $|\widehat{\varkappa}| \leq |\varkappa^*|$. So, the question under study is the probability of a situation when $\widehat{\varkappa}$ selects some other features instead of those in \varkappa^* .

For simplicity of notation, we consider the sequence space model with $\Psi = I_p$ and also assume $\sigma^2 = 1$. Our first result describes which candidate set \varkappa will be rejected with a high probability. The definition of the oracle \varkappa^* implies that the component of $\boldsymbol{\theta}^*$ lying outside of \varkappa^* is not massive in the sense that $\|\Pi_{\varkappa \setminus \varkappa^*} \boldsymbol{\theta}^*\|$ does not exceed $\beta |\varkappa - \varkappa^*|^{1/2}$ for any $\varkappa > \varkappa^*$.

Let now \varkappa be any subset not equal to \varkappa^* with $|\varkappa| \leq |\varkappa^*|$. This implies that some features from the active set \varkappa^* do not enter in \varkappa . We therefore, measure the related loss of information by the norm of $\Pi_{\varkappa^* \setminus \varkappa} \boldsymbol{\theta}^* = \Pi_{\varkappa^*, \varkappa^* \wedge \varkappa} \boldsymbol{\theta}^*$, which is the projection of the true signal onto components which are in \varkappa^* but not in \varkappa . Our result says that if this loss of information is large, such a candidate will be killed with a high probability. This result can be viewed as an extension of the zone-of-insensitivity result from the ordered case. Below we use the short notation $z_{\varkappa, \varkappa^\circ}^+$ for the tail functions of $\|\boldsymbol{\xi}_{\varkappa, \varkappa^\circ}\|$ with the proper multiplicity correction; cf (6.5):

$$z_{\varkappa, \varkappa^\circ}^+ = z_{\varkappa, \varkappa^\circ}(\mathbf{x} + q_{1, \varkappa^\circ} + q_{2, \varkappa^\circ} + \dots + q_{m, \varkappa^\circ}), \quad \varkappa \in \mathcal{M}_m(\varkappa^\circ).$$

Theorem 6.1.5. *Let \varkappa be such that*

$$\|b_{\varkappa \vee \varkappa^*, \varkappa}\| \geq \|\Pi_{\varkappa^* \setminus \varkappa} \boldsymbol{\theta}^*\| > 2z_{\varkappa \vee \varkappa^*, \varkappa}^+ + \beta |\varkappa^* \setminus \varkappa|^{1/2}, \quad (6.9)$$

and let $\mathcal{M}^\circ(\varkappa^)$ be the collection of all such \varkappa . Then*

$$\mathbb{P}(\text{any of } \varkappa \in \mathcal{M}^\circ(\varkappa^*) \text{ is accepted}) \leq e^{-x}.$$

Proof. We just apply the acceptance rule for \varkappa requiring

$$\|\tilde{\boldsymbol{\theta}}_{\varkappa \vee \varkappa^*} - \tilde{\boldsymbol{\theta}}_\varkappa\| \leq z_{\varkappa \vee \varkappa^*, \varkappa} = z_{\varkappa \vee \varkappa^*, \varkappa}^+ + \beta |\varkappa^* \setminus \varkappa|^{1/2}. \quad (6.10)$$

The usual decomposition of $\tilde{\boldsymbol{\theta}}_\varkappa$ implies on a dominating set $\Omega(\mathbf{x})$ by (6.4)

$$\begin{aligned} \|\tilde{\boldsymbol{\theta}}_{\varkappa \vee \varkappa^*} - \tilde{\boldsymbol{\theta}}_\varkappa\| &\geq \|b_{\varkappa \vee \varkappa^*, \varkappa}\| - \|\boldsymbol{\xi}_{\varkappa \vee \varkappa^*, \varkappa}\| \\ &\geq \|b_{\varkappa \vee \varkappa^*, \varkappa}\| - z_{\varkappa \vee \varkappa^*, \varkappa}^+. \end{aligned} \quad (6.11)$$

It remains to check that the inequalities (6.10) and (6.11) are incompatible under (6.9).

This result can be directly applied to check for a subset \varkappa_0^* of \varkappa^* whether it will be completely missed by $\widehat{\varkappa}$.

Corollary 6.1.1. *Let \varkappa_0^* fulfill*

$$\|\Pi_{\varkappa_0^*} \boldsymbol{\theta}^*\| \geq \bar{z}_{\varkappa^*} \stackrel{\text{def}}{=} \max_{|\varkappa| \leq |\varkappa^*|} \{2z_{\varkappa \vee \varkappa^*, \varkappa} + \beta |\varkappa^* \setminus \varkappa|^{1/2}\}.$$

Then

$$IP(\varkappa_0^* \cap \widehat{\varkappa} = \emptyset) \leq e^{-x}.$$

In particular, any coefficient θ_j^ of $\boldsymbol{\theta}^*$ with $|\theta_j^*| > \bar{z}_{\varkappa^*}$ will be included in $\widehat{\varkappa}$ with a probability at least $1 - e^{-x}$.*

6.2 Anisotropic models

This section studies the so called anisotropic models when one has a number tuning parameters to be selected, and each of them is ordered. In other words, \varkappa is a vector with two or more components, and we consider the set of the estimates $\tilde{\boldsymbol{\theta}}_\varkappa$. When only one component of \varkappa is varying and the other are fixed, the monotonicity assumption is assumed to be fulfilled. However, this only yields a componentwise partial ordering of the set \mathcal{M} of all considered models.

To simplify the presentation, we consider below the two dimensional case and a product structure. An extension to the general case is straightforward.

Let $\varkappa = (\varkappa_1, \varkappa_2)$ with $\varkappa_j \in \mathcal{M}_j$ for $j = 1, 2$ and $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$. We write $\varkappa = (\varkappa_1, \varkappa_2) \geq \varkappa^\circ = (\varkappa_1^\circ, \varkappa_2^\circ)$ if $\varkappa_1 \geq \varkappa_1^\circ$ and $\varkappa_2 \geq \varkappa_2^\circ$. Assume we are given a collection of linear smoothers $\tilde{\boldsymbol{\theta}}_\varkappa$ with a partial ordering: if $\varkappa > \varkappa^\circ$ then

$$\text{Var}(W\tilde{\boldsymbol{\theta}}_\varkappa) > \text{Var}(W\tilde{\boldsymbol{\theta}}_{\varkappa^\circ}).$$

This particularly implies

$$p_\varkappa \stackrel{\text{def}}{=} \text{tr}\{\text{Var}(W\tilde{\boldsymbol{\theta}}_\varkappa)\} > p_{\varkappa^\circ} \stackrel{\text{def}}{=} \text{tr}\{\text{Var}(W\tilde{\boldsymbol{\theta}}_{\varkappa^\circ})\}.$$

We aim at applying the SmA method to this special situation. The setup and notation of the previous section are kept. We focus on a linear regression model $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \varepsilon$ with homogeneous Gaussian error $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and consider a collection of pairwise test statistics

$$\mathbb{T}_{\varkappa, \varkappa^\circ} = \|W(\tilde{\boldsymbol{\theta}}_\varkappa - \tilde{\boldsymbol{\theta}}_{\varkappa^\circ})\| = \|\xi_{\varkappa, \varkappa^\circ} + b_{\varkappa, \varkappa^\circ}\|.$$

As previously, define

$$p_{\varkappa, \varkappa^\circ} = \text{tr}\{\text{Var}(\xi_{\varkappa, \varkappa^\circ})\}.$$

The acceptance rule can be written as

$$\varkappa^\circ \text{ is accepted iff } T_{\varkappa, \varkappa^\circ} \leq z_{\varkappa, \varkappa^\circ} \quad \forall \varkappa \in \mathcal{M}(\varkappa^\circ), \quad (6.12)$$

where $\mathcal{M}(\varkappa^\circ) = \{\varkappa \in \mathcal{M}: \varkappa > \varkappa^\circ\}$. In words, \varkappa° is accepted if it is competitive with any larger model \varkappa . Now we discuss how the critical values $z_{\varkappa, \varkappa^\circ}$ can be fixed by the multiplicity correction of the individual tail functions. Here we only discuss a uniform correction, however, it can be easily done in a multilevel form. Define $q_{\varkappa^\circ}(\mathbf{x})$ by the condition

$$IP\left(\bigcup_{\varkappa \in \mathcal{M}(\varkappa^\circ)} \left\{ \| \boldsymbol{\xi}_{\varkappa, \varkappa^\circ} \| > z_{\varkappa, \varkappa^\circ}(\mathbf{x} + q_{\varkappa^\circ}(\mathbf{x})) \right\}\right) \leq e^{-x}. \quad (6.13)$$

Denote $\mathbf{x}_{\varkappa^\circ} = \mathbf{x} + q_{\varkappa^\circ}(\mathbf{x})$ and apply the acceptance rule (6.2) with $z_{\varkappa, \varkappa^\circ}$ equal to such defined $z_{\varkappa, \varkappa^\circ}(\mathbf{x}_{\varkappa^\circ})$ after a small bias correction:

$$z_{\varkappa, \varkappa^\circ} \stackrel{\text{def}}{=} z_{\varkappa, \varkappa^\circ}(\mathbf{x}_{\varkappa^\circ}) + \beta p_{\varkappa, \varkappa^\circ}^{1/2}. \quad (6.14)$$

A good choice \varkappa° can be defined as previously by “no significant bias” condition:

$$\| b_{\varkappa, \varkappa^\circ} \| \leq \beta p_{\varkappa, \varkappa^\circ}^{1/2} \quad \varkappa \in \mathcal{M}(\varkappa^\circ). \quad (6.15)$$

The construction ensures that a good model will be accepted with a high probability.

Theorem 6.2.1. *Let \varkappa° be a good model in the sense (6.15). Then it holds for the acceptance rule (6.12) with the critical values $z_{\varkappa, \varkappa^\circ}$ from (6.13) and (6.14)*

$$IP(\varkappa^\circ \text{ is rejected}) \leq e^{-x}.$$

So, the construction allows to figure out a set of good models, each of them will be kept by the procedure with a high probability. It remains to introduce a natural ordering on the set of such good models. This can be done by the value p_{\varkappa° which is proportional to $|\varkappa^\circ|$ for the sequence space model. Define the oracle choice \varkappa^* as the smallest (in complexity p_{\varkappa°) model under the constraint (6.15):

$$\varkappa^* \stackrel{\text{def}}{=} \operatorname{argmin}_{\varkappa^\circ \in \mathcal{M}} \{ p_{\varkappa^\circ} : \| b_{\varkappa, \varkappa^\circ} \| \leq \beta p_{\varkappa, \varkappa^\circ}^{1/2}, \quad \varkappa \in \mathcal{M}(\varkappa^\circ) \}.$$

Our selection rule chooses the smallest accepted model. It can be written as

$$\hat{\varkappa} = \operatorname{argmin}_{\varkappa^\circ \in \mathcal{M}} \{ p_{\varkappa^\circ} : T_{\varkappa, \varkappa^\circ} \leq z_{\varkappa, \varkappa^\circ}, \quad \varkappa \in \mathcal{M}(\varkappa^\circ) \}. \quad (6.16)$$

The oracle bound compares the risk of the oracle estimate $\mathcal{R}_{\varkappa^*}$ with the risk of the adaptive estimate $\hat{\theta} = \tilde{\theta}_{\hat{\varkappa}}$ for the SmA rule $\hat{\varkappa}$. Below for two given models \varkappa and \varkappa° , we denote by $\varkappa \vee \varkappa^\circ$ the smallest model which is larger than each:

$$\varkappa \vee \varkappa^\circ \stackrel{\text{def}}{=} (\varkappa_1 \vee \varkappa_1^\circ, \varkappa_2 \vee \varkappa_2^\circ).$$

Theorem 6.2.2. *It holds on a random set of probability at least $1 - e^{-x}$*

$$\|W(\tilde{\boldsymbol{\theta}}_{\varkappa^*} - \tilde{\boldsymbol{\theta}}_{\hat{\varkappa}})\| \leq \bar{z}_{\varkappa^*}$$

where \bar{z}_{\varkappa^*} is defined by

$$\bar{z}_{\varkappa^*} \stackrel{\text{def}}{=} \max_{p_{\varkappa} \leq p_{\varkappa^*}} (z_{\varkappa \vee \varkappa^*, \varkappa^*} + z_{\varkappa \vee \varkappa^*, \varkappa}).$$

Proof. The construction and the propagation property ensures that \varkappa^* is accepted with a high probability $1 - e^{-x}$. Below we focus on this case. Then the adaptive choice $\hat{\varkappa}$ from (6.16) has to fulfill

$$p_{\hat{\varkappa}} \leq p_{\varkappa^*}.$$

Consider $\check{\varkappa} = \varkappa^* \vee \hat{\varkappa}$ which contains both $\hat{\varkappa}$ and \varkappa^* . As \varkappa^* and $\hat{\varkappa}$ are both accepted, it holds

$$\|W(\tilde{\boldsymbol{\theta}}_{\check{\varkappa}} - \tilde{\boldsymbol{\theta}}_{\hat{\varkappa}})\| \leq z_{\check{\varkappa}, \hat{\varkappa}}, \quad \|W(\tilde{\boldsymbol{\theta}}_{\check{\varkappa}} - \tilde{\boldsymbol{\theta}}_{\varkappa^*})\| \leq z_{\check{\varkappa}, \varkappa^*}.$$

Therefore,

$$\|W(\tilde{\boldsymbol{\theta}}_{\varkappa^*} - \tilde{\boldsymbol{\theta}}_{\hat{\varkappa}})\| \leq z_{\check{\varkappa}, \hat{\varkappa}} + z_{\check{\varkappa}, \varkappa^*},$$

and the assertion follows.

The value \bar{z}_{\varkappa^*} is the ‘‘payment for adaptation’’, and it can be quite large relative to the oracle standard deviation $p_{\varkappa^*}^{1/2}$. The worst case is given by the anisotropic situation with the oracle of the form $\varkappa^* = (\varkappa_1^*, \varkappa_{2,\max})$. In this case the set of competitive models includes $\varkappa = (\varkappa_{1,\max}, \varkappa_2)$, the maximum of \varkappa^* and \varkappa is the largest possible model

$$\varkappa^* \vee \varkappa = (\varkappa_{1,\max}, \varkappa_{2,\max})$$

and the corresponding critical value $z_{\varkappa \vee \varkappa^*, \varkappa^*}$ can be very large.

To be done: In the isotropic case with $\varkappa_1^* \asymp \varkappa_2^*$, this problem disappears.

To be done: One can refine the result by considering the zone of insensitivity $\mathcal{M}^\circ(\varkappa^*)$.

SmA and parameter tuning in high dimensional regression

Extending the methods and results on model selection to situation with a large or even huge parameter dimension is one of the main challenge of modern statistics. An important requirement to any such method is an automatic parameter tuning. If a proposed procedure involves some tuning parameters without explaining their automatic choice, then one problem is just replaced by another. This chapter focuses a special problem of subset selection for linear models with unknown noise structure. We aim to bring together the SmA procedure of Section 6.1 and the resampling idea of Chapter 20. In this chapter we restrict ourselves to a linear model: the observation vector $\mathbf{Y} \in \mathbb{R}^n$ is described by the equation

$$\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \varepsilon$$

for a given dictionary Ψ . We allow that the dictionary is overcomplete, that is, the dimension p of the vector $\boldsymbol{\theta}^*$ can be much larger than the number of observations n . Some structural assumptions are necessary to make the problem of recovering $\boldsymbol{\theta}^*$ meaningful. As in Chapter 20 we assume that $\boldsymbol{\theta}^*$ is sparse or can be approximated by a sparse vector. For a subset κ , by $\tilde{\boldsymbol{\theta}}_\kappa$ we denote the corresponding LSE $\tilde{\boldsymbol{\theta}}_\kappa$:

$$\tilde{\boldsymbol{\theta}}_\kappa = (\Psi_\kappa \Psi_\kappa^\top)^{-1} \Psi_\kappa \mathbf{Y} = D_\kappa^{-2} \Pi_\kappa \Psi \mathbf{Y},$$

where $\Psi_\kappa = \Pi_\kappa \Psi$ with Π_κ being a projector on the κ -subspace of the $\boldsymbol{\theta}$ -space and $\nabla = \Psi \varepsilon$ is the score vector in \mathbb{R}^p . By D_κ^{-2} we denote the pseudo inverse of the matrix $D_\kappa^2 = \Psi_\kappa \Psi_\kappa^\top$. Given a weighing loss matrix W , the SmA procedure can be applied as soon as the family of tail functions $z_{\kappa, \kappa^\circ}(\mathbf{x})$ for the norm of the stochastic component

$$\xi_{\kappa, \kappa^\circ} = W(D_\kappa^{-2} - D_{\kappa^\circ}^{-2}) \nabla = S_{\kappa, \kappa^\circ} \nabla$$

is fixed. Prior information about the noise ε makes this possible. Here we discuss how this information can be recovered from the data using a resampling method.

7.1 SmA subset selection in high dimensional regression

This section discusses the parameter choice in the problem of subset selection for a high dimensional linear regression model with unknown noise structure. The SmA procedure of Section 6.1 will be extended by the bootstrap step for choosing the critical values z_{κ, κ^0} . We follow the approach of previous chapters and consider for a given sample \mathbf{Y} a family of Gaussian multipliers $\mathbf{w}^b = (w_i^b)$. We also need a pilot estimate $\tilde{\boldsymbol{\theta}}$ which removes most of systematic part from the data \mathbf{Y} . Further we proceed with the residuals $\check{\boldsymbol{\varepsilon}} = \mathbf{Y} - \Psi^\top \tilde{\boldsymbol{\theta}}$ for the bootstrap step. For each pair $\kappa > \kappa^0$, the bootstrap counterpart ξ_{κ, κ^0}^b of ξ_{κ, κ^0} looks as

$$\xi_{\kappa, \kappa^0}^b = W(D_\kappa^{-2} \Pi_\kappa - D_{\kappa^0}^{-2} \Pi_{\kappa^0}) \Psi \mathcal{E}^b \check{\boldsymbol{\varepsilon}},$$

where \mathcal{E}^b is the diagonal matrix of the centered bootstrap multipliers; cf. (??). The proposed approach uses the bootstrap tail function $z_{\kappa, \kappa^0}(\mathbf{x})$ as a proxy for the true one:

$$\mathbb{P}^b(\|\xi_{\kappa, \kappa^0}^b\| > z_{\kappa, \kappa^0}(\mathbf{x})) \leq e^{-\mathbf{x}}.$$

After each tail function $z_{\kappa, \kappa^0}(\mathbf{x})$ is built, one can apply either uniform or multilevel synchronization of such tail functions for fixing the set of critical values z_{κ, κ^0} . The main argument in validating this bootstrap procedure in the ordered case was that all real-world stochastic terms ξ_{κ, κ^0} and bootstrap-world stochastic components ξ_{κ, κ^0}^b are deterministic linear functions of the corresponding scores: $\check{\nabla}^b = \Psi \mathcal{E}^b \check{\boldsymbol{\varepsilon}}$ in the bootstrap world and $\nabla = \Psi \boldsymbol{\varepsilon}$ in the underlying model. This is still the case.

If the errors $\boldsymbol{\varepsilon}$ are Gaussian then the score vector $\nabla = \Psi \boldsymbol{\varepsilon}$ is Gaussian as well. The bootstrap score $\check{\nabla}^b$ is Gaussian under the bootstrap measure \mathbb{P}^b by construction. So, validation of the bootstrap procedure can now be restated as comparison of two Gaussian distributions in a rather special sense. Below we admit a possibly inhomogeneous noise. The only necessary assumptions are independence of the errors ε_i and a kind of the Lindeberg condition on the rows of the matrix Ψ . If $\psi_j = (\psi_{i,j})$ denotes the j th row of Ψ , then each component $\nabla_j = \psi_j \boldsymbol{\varepsilon}$ of the score $\nabla = \Psi \boldsymbol{\varepsilon}$ reads

$$\nabla_j = \sum_{i=1}^n \psi_{i,j} \varepsilon_i$$

Under the assumption that the errors ε_i are Gaussian, the same holds for ∇_j :

$$\nabla_j \sim \mathcal{N}(0, v_j^2), \quad v_j^2 = \sum_{i=1}^n \psi_{i,j}^2 \sigma_i^2.$$

The corresponding bootstrap-score component can be represented as

$$\nabla_j^b = \sum_{i=1}^n \psi_{i,j} \check{\varepsilon}_i w_i^b,$$

where $\check{\varepsilon}_i$ are the components of $\check{\varepsilon}$. With Gaussian multipliers, it is normal as well conditioned on the data \mathbf{Y} :

$$\nabla_j^b \sim \mathcal{N}(0, v_j^b)^2, \quad v_j^b = \sum_{i=1}^n \psi_{i,j}^2 \check{\varepsilon}_i^2.$$

Even if we ignore the systematic component in the residuals $\check{\varepsilon}_i$ and use ε_i in place of $\check{\varepsilon}_i$, we can only hope that two covariances v_j^2 and v_j^b are close to each other with high probability under Lindeberg type conditions on ε_i . The same applies to cross-covariance of the different component of ∇ and similar components of $\check{\nabla}^b$. Therefore, the problem can be reduced to comparing of two high dimensional Gaussian measures with similar covariance structure. Unfortunately the dimension p of the vectors ∇ and $\check{\nabla}^b$ can be very large. The tools of previous chapters based on the Pinsker inequality hardly apply here, because the dimension p enters in the error bound. In regular situation the error term is of order $\sqrt{p/n}$, and this value is not small if p is larger than n . One needs another technique which applies in a very high dimensional space.

The SmA procedure involves a multiple comparison of many test statistics each of them is based on the norm of a Gaussian vector. Suppose we are given a family of vectors $\{\xi_\varkappa, \varkappa \in \mathcal{M}_m\}$ each of dimension m . For each ξ_\varkappa suppose that the corresponding critical value z_\varkappa is fixed in a way that

$$\mathbb{P}\left(\bigcup_{\varkappa \in \mathcal{M}_m} \{\|\xi_\varkappa\| > z_\varkappa\}\right) \leq e^{-x}.$$

We now want to control a similar deviation probability for another family of vectors $\{\xi_\varkappa^b\}$ whose covariance structure is close to that of $\{\xi_\varkappa\}$. To justify the bootstrap procedure in the case of high parameter dimension, we need to reduce this problem to the problem of comparison of maxima for two large Gaussian vectors.

Let \mathcal{S}_m denote a unit sphere in \mathbb{R}^m . We use that for any vector ξ in \mathbb{R}^m , it holds

$$\|\xi\| = \sup_{\gamma \in \mathcal{S}_m} \gamma^\top \xi.$$

Further we have to replace the maximum over the whole sphere by the maximum over a finite subset. Given $\delta > 0$, consider a finite δ -net $\mathcal{S}_m(\delta)$ in \mathcal{S}_m . It is obvious that

$$(1 - \delta)\|\xi\| \leq \max_{\mathcal{S}_m(\delta)} \gamma^\top \xi \leq \|\xi\|. \quad (7.1)$$

Putting together for all \varkappa implies

$$\begin{aligned} \mathbb{P}\left(\bigcup_{\varkappa \in \mathcal{M}_m} \bigcup_{\gamma \in \mathcal{S}_m} \left\{ \frac{1}{z_\varkappa} \gamma^\top \xi_\varkappa > 1 \right\} \right) &\leq \mathbb{P}\left(\bigcup_{\varkappa \in \mathcal{M}_m} \{\|\xi_\varkappa\| > z_\varkappa\}\right) \\ &\leq \mathbb{P}\left(\bigcup_{\varkappa \in \mathcal{M}_m} \bigcup_{\gamma \in \mathcal{S}_m(\delta)} \left\{ \frac{1}{z_\varkappa} \gamma^\top \xi_\varkappa > 1 - \delta \right\} \right) \end{aligned}$$

Now introduce the vector \mathbf{X} of dimension $\mathbb{M}_m = \mathbb{M}_m(\delta) = |\mathcal{S}_m(\delta)| \times |\mathcal{M}_m|$ with the entries $z_\varkappa^{-1} \gamma^\top \xi_\varkappa$ for $\gamma \in \mathcal{S}_m(\delta)$ and $\varkappa \in \mathcal{M}_m$.

Below we consider a special case when $\xi_\varkappa = \Pi_\varkappa \xi$ for some linear mapping $\Pi_\varkappa: \mathbb{R}^p \rightarrow \mathbb{R}^m$ for $\varkappa \in \mathcal{M}_m$. Suppose also that another Gaussian zero mean vector ξ^\flat is given and $\xi_\varkappa^\flat = \Pi_\varkappa \xi^\flat$. Build a vector \mathbf{X}^\flat out of the ξ_\varkappa^\flat 's in the same way as \mathbf{X} was constructed out of the ξ_\varkappa 's. As a next step we evaluate the distance between two covariance operators for \mathbf{X} and \mathbf{X}^\flat . Let $\Sigma = \text{Var}(\xi)$. Obviously, for each two pairs (\varkappa, γ) and (\varkappa_1, γ_1) ,

$$\begin{aligned} \mathbb{E}[\gamma^\top \xi_\varkappa \gamma_1^\top \xi_{\varkappa_1}] &= \mathbb{E}[\gamma^\top \xi_\varkappa \xi_{\varkappa_1}^\top \gamma_1] = \mathbb{E}[\gamma^\top \Pi_\varkappa \xi \xi^\top \Pi_{\varkappa_1}^\top \gamma_1] \\ &= \gamma^\top \Pi_\varkappa \Sigma \Pi_{\varkappa_1}^\top \gamma_1 \leq \|\Pi_\varkappa \Sigma \Pi_{\varkappa_1}^\top\|. \end{aligned} \quad (7.2)$$

A similar formula holds for the covariance operator of ξ^\flat . Below we denote

$$\Delta_m \stackrel{\text{def}}{=} \max_{\varkappa, \varkappa_1 \in \mathcal{M}_m} \|\Pi_\varkappa (\Sigma - \Sigma^\flat) \Pi_{\varkappa_1}^\top\|. \quad (7.3)$$

To be done: Exp-Bernstein inequality implies $\Delta_m \leq Cn^{-1/2} \log(p)$

Then (7.2) and (7.3) imply for $z_\varkappa \geq 1$

$$\|\Sigma_{\mathbf{X}} - \Sigma_{\mathbf{X}^\flat}\|_\infty \leq \Delta_m.$$

This and the result (F.10) of Theorem F.2.1 imply

$$\begin{aligned} \mathbb{P}\left(\max_{\varkappa \in \mathcal{M}_m} \max_{\gamma \in \mathcal{S}_m(\delta)} \frac{1}{z_\varkappa} \gamma^\top \xi_\varkappa \geq 1\right) &\leq \mathbb{P}\left(\max_{\varkappa \in \mathcal{M}_m} \max_{\gamma \in \mathcal{S}_m(\delta)} \frac{1}{z_\varkappa} \gamma^\top \xi_\varkappa^\flat \geq 1 - 2\Delta\right) + 2\Delta^{-2} \{\log(\mathbb{M}_m) + 1\} \Delta_m. \end{aligned}$$

Together with the norm approximation (7.1) this yields

$$\begin{aligned} \mathbb{P}\left(\max_{\varkappa \in \mathcal{M}_m} \frac{1}{z_\varkappa} \|\xi_\varkappa\| \geq 1\right) &\leq \mathbb{P}\left(\max_{\varkappa \in \mathcal{M}_m} \frac{1}{z_\varkappa} \|\xi_\varkappa^\flat\| \geq (1 - \delta)(1 - 2\Delta)\right) + 2\Delta^{-2} \{\log(\mathbb{M}_m) + 1\} \Delta_m. \end{aligned} \quad (7.4)$$

It remains to account for δ . The cardinality $|\mathcal{S}_m(\delta)|$ can be roughly upper bounded by $(1 + 2\delta^{-1})^m$, see Lemma 5.2 in www-personal.umich.edu/~romanv/papers/non-asymptotic-rmt-plain.pdf, yielding

$$\mathbb{M}_m < |\mathcal{M}_m| (1 + 2\delta^{-1})^m. \quad (7.5)$$

The above calculus with $\delta = \Delta$ imply the following result.

Theorem 7.1.1. *Let ξ and ξ^b be two zero mean Gaussian vectors in \mathbb{R}^p , and let $\{\Pi_\kappa, \kappa \in \mathcal{M}_m\}$ be a collection of linear mappings $\mathbb{R}^{\mathcal{M}} \rightarrow \mathbb{R}^m$. For the quantity Δ_m from (7.3) and for any set of critical values $z_\kappa \geq 1$, it holds with any $\Delta < 1/3$*

$$\begin{aligned} & \mathbb{P}\left(\max_{\kappa \in \mathcal{M}_m} \frac{1}{z_\kappa} \|\xi_\kappa\| \geq 1\right) \\ & \leq \mathbb{P}\left(\max_{\kappa \in \mathcal{M}_m} \frac{1}{z_\kappa} \|\xi_\kappa^b\| \geq 1 - 3\Delta\right) + 2\Delta^{-2} \left\{ \log(|\mathcal{M}_m|) + m \log(1 + 2/\Delta) \right\} \Delta_m. \end{aligned}$$

Proof. Apply (7.4) and (7.5) and use that $(1 - \Delta)(1 - 2\Delta) \geq 1 - 3\Delta$.

If \mathcal{M}_m is the set of all subsets of the full index set $\{1, \dots, p\}$, then by the Stirling formula for $m!$

$$\log |\mathcal{M}_m| \leq \log \binom{p}{m} \leq \log(p^m/m!) \leq m \log(ep/m).$$

In particular, with $\Delta < 1/3$ we obtain

$$\begin{aligned} & \mathbb{P}\left(\max_{\kappa \in \mathcal{M}_m} \frac{1}{z_\kappa} \|\xi_\kappa\| \geq 1\right) \\ & \leq \mathbb{P}^b\left(\max_{\kappa \in \mathcal{M}_m} \frac{1}{z_\kappa} \|\xi_\kappa^b\| \geq 1 - 3\Delta\right) + 2\Delta^{-2} m \log\left(\frac{2ep}{m\Delta}\right) \Delta_m. \end{aligned}$$

To be done: put altogether and find the bound on n , m , and p

If the dimension p of the vector θ is large, the approach of Section ?? based on the Pinsker inequality faces a crucial problem: the error term is proportional to $p^{1/2}$ and can be very large. One needs a different technique which allows comparing two Gaussian measures in a high dimensional space in terms of the corresponding covariance operators.

Penalized model selection

This chapter discusses a class of procedures which can be represented as penalized minimization of the empirical risk. We consider the linear model $\mathbf{Y} = \boldsymbol{\Psi}^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ in which the parameter dimension p can be very large. The empirical risk is just the squared norm of the difference $\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}$. The penalized procedure tries to minimize this empirical risk penalized by the complexity of the vector $\boldsymbol{\theta}$ used for prediction:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \{ \| \mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta} \|^2 + \operatorname{pen}(\boldsymbol{\theta}) \} \quad (8.1)$$

for a penalty function $\operatorname{pen}(\boldsymbol{\theta})$.

The roughness penalty has been already discussed in Chapter 3. The resulting estimate is again linear and the general approach of linear model selection continues to apply. Here we consider two special choices of $\operatorname{pen}(\boldsymbol{\theta})$ which are essentially non-linear. The complexity penalty $\operatorname{pen}(\boldsymbol{\theta}) = C\|\boldsymbol{\theta}\|_0$ just counts the number of non-zero $\boldsymbol{\theta}$ -coefficients. The sparse penalty $\operatorname{pen}(\boldsymbol{\theta}) = C\|\boldsymbol{\theta}\|_q^q$ use the q -norm of $\boldsymbol{\theta}$ for some $q < 2$. The most popular sparse penalty corresponds to $q = 1$.

8.1 Complexity penalization

This section considers the important special case of penalization by complexity. The famous Akaike criteria is a special case of such penalization. We offer another viewpoint based on the SmA idea. As previously in Section 4.1, for a subset κ of the index set $\{1, \dots, p\}$, the estimate $\tilde{\boldsymbol{\theta}}_\kappa$ is the corresponding projection MLE:

$$\tilde{\boldsymbol{\theta}}_\kappa = (\boldsymbol{\Psi}_\kappa \boldsymbol{\Psi}_\kappa^\top)^{-1} \boldsymbol{\Psi}_\kappa \mathbf{Y}.$$

Theorem 8.1.1. *Let $\operatorname{pen}(\boldsymbol{\theta}) = C\|\boldsymbol{\theta}\|_0$. Then the solution $\hat{\boldsymbol{\theta}}$ of the problem (8.1) satisfies*

$$\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}_{\hat{\kappa}},$$

where

$$\hat{\kappa} = \operatorname{argmin}_{\kappa} \{ \|\tilde{\varepsilon}_{\kappa}\|^2 + C|\kappa| \}, \quad (8.2)$$

with

$$\tilde{\varepsilon}_{\kappa} = \mathbf{Y} - \boldsymbol{\Psi}_{\kappa}^{\top} \tilde{\boldsymbol{\theta}}_{\kappa} = (I_n - \Pi_{\kappa}) \mathbf{Y}$$

and $|\kappa|$ means the cardinality of κ or, equivalently the number of coefficients in the support set κ .

The unbiased risk estimation procedure corresponds to the special choice of constant $C = 2\sigma^2$. This results in selecting a proper subset κ which provides a reasonable fit under complexity constraint:

$$\hat{\kappa} = \operatorname{argmin}_{\kappa} \{ \|\tilde{\varepsilon}_{\kappa}\|^2 + 2\sigma^2|\kappa| \}.$$

This procedure faces two essential problems when the dimension p becomes large. One of them is algorithmic: the procedure requires to compute the MLE $\tilde{\boldsymbol{\theta}}_{\kappa}$ and the related empirical risk for any subset κ ; such a problem is in general NP-hard and can be solved in very special cases, e.g. if the design matrix $\boldsymbol{\Psi}$ is orthogonal. Then the procedure can be reduced to thresholding of individual Fourier coefficients $\tilde{\theta}_j = \psi_j \mathbf{Y}$, where ψ_j denotes the j th row of $\boldsymbol{\Psi}$.

Theorem 8.1.2. Let $n \geq p$ and the matrix $\boldsymbol{\Psi}$ be orthonormal, that is, $\boldsymbol{\Psi}\boldsymbol{\Psi}^T = I_p$. Then the active set $\hat{\kappa}$ from (8.2) is given by hard thresholding

$$\hat{\kappa} = \{j : |\tilde{\theta}_j| \geq \lambda\}$$

for a proper $\lambda > 0$. The corresponding hard thresholding estimate $\hat{\boldsymbol{\theta}} = (\hat{\theta}_j)$ reads as

$$\hat{\theta}_j = \begin{cases} \psi_j \mathbf{Y}, & |\psi_j \mathbf{Y}| > \lambda, \\ 0, & \text{otherwise.} \end{cases}$$

The other problem is statistical. First of all, the procedure assumes a homogeneous noise and requires that the noise variance σ^2 is given. Second, the penalization in the form $2\sigma^2|\kappa|$ is too mild and does not ensure a proper model selection for p large. The reason is that there are very many models to select between, this number is exponential in p , and therefore, the price for this choice has to be much larger than in the ordered case.

Below we discuss several ways of solving the statistical problem. First we consider the penalized model selection procedure (8.1) or equivalently (8.2) with a data-driven constant C . Then we extend it to a more sophisticated non-linear choice of the penalty function $\text{pen}(\boldsymbol{\theta})$. Finally we discuss the saddle-point bivariate model selection.

8.1.1 Penalty tuning using propagation condition

Complexity penalization requires to fix a constant C in (8.2), and this choice is crucial for the performance of the procedure. Below we discuss the approach based on the *propagation* idea: if the model is “good” in the sense that it provides a reasonable data fit, it should be competitive against larger models. In other words, if one already achieved a proper prediction ability, there is no reason to increase the complexity of the estimate. A typical situation is as follows: we have a model-candidate κ° which is not too complex, that is, $|\kappa^\circ|$ is small relative to the sample size n and the total dimension p . One says in such cases that κ° is *sparse*. Further, this choice κ° is *good* if the coefficients θ_j^* for $j \notin \kappa^\circ$ are nearly zero. We aim at designing a data-driven procedure such that the criterium (8.2) keeps κ° alive in competition with all larger models. This precisely means that

$$\|\tilde{\varepsilon}_\kappa\|^2 + C|\kappa| \geq \|\tilde{\varepsilon}_{\kappa^\circ}\|^2 + C|\kappa^\circ|, \quad \kappa > \kappa^\circ.$$

This inequality has to be verified for all $\kappa > \kappa^\circ$ with a high probability. Rearranging yields in view of $\tilde{\varepsilon}_\kappa = (I_n - \Pi_\kappa)\mathbf{Y}$

$$\|\tilde{\varepsilon}_{\kappa^\circ}\|^2 - \|\tilde{\varepsilon}_\kappa\|^2 = \|\Pi_{\kappa, \kappa^\circ}\mathbf{Y}\|^2 = \|\boldsymbol{\Psi}^\top(\tilde{\boldsymbol{\theta}}_\kappa - \tilde{\boldsymbol{\theta}}_{\kappa^\circ})\|^2 \leq C(|\kappa| - |\kappa^\circ|). \quad (8.3)$$

If we knew the noise distribution then we can fix the constant C in the “pure noise” situation. Indeed, the underlying structural assumption means that there is no significant signal θ_j^* for $j \notin \kappa^\circ$, and hence, one can simply ignore such signal and consider $\theta_j^* \equiv 0$ in the complement of κ° . This leads to the propagation condition

$$\mathbb{P}\left(\bigcup_{\kappa > \kappa^\circ} \{\|\Pi_{\kappa, \kappa^\circ}\varepsilon\|^2 > C_0(|\kappa| - |\kappa^\circ|)\}\right) \leq e^{-x} \quad (8.4)$$

for a constant C_0 . In the case of a homogeneous noise, it is obvious that this condition becomes stronger if the set κ° is taken smaller. The hardest case corresponds to the empty set κ° , yielding the constraint

$$\mathbb{P}\left(\bigcup_{\kappa} \{\|\Pi_{\kappa}\varepsilon\|^2 > C_0|\kappa|\}\right) \leq e^{-x}. \quad (8.5)$$

If the dimension of κ only slightly higher than the dimension of κ° , the inequality $\|\Pi_{\kappa, \kappa^\circ}\varepsilon\|^2 > C_0(|\kappa| - |\kappa^\circ|)$ would require a very large constant C_0 . At the same time, this constant rapidly stabilizes if $|\kappa| - |\kappa^\circ|$ exceeds some prescribed value. This suggests to extend the condition (8.4) (resp. (8.5)): given $\tau \geq 1$

$$\mathbb{P}\left(\bigcup_{\kappa \in \mathcal{M}_\tau(\kappa^\circ)} \{\|\Pi_{\kappa, \kappa^\circ}\varepsilon\|^2 > C_0(|\kappa| - |\kappa^\circ|)\}\right) \leq e^{-x}. \quad (8.6)$$

Here $\mathcal{M}_\tau(\varkappa^\circ) = \{\varkappa > \varkappa^\circ : |\varkappa| \geq |\varkappa^\circ| + \tau\}$ is the set of all models whose complexity exceed $|\varkappa^\circ|$ by τ or more, it is the complement of the set $\mathcal{M}_\tau(\varkappa^\circ)$; cf. (6.4).

Now suppose that such a constant $C_0 = C_0(\tau)$ is fixed. Then the procedure can be applied with

$$C \stackrel{\text{def}}{=} C_0(\tau) + \beta,$$

where β appears in the definition of a “good” choice: \varkappa° is good if

$$\|\Pi_{\varkappa, \varkappa^\circ} f^*\|^2 \leq \beta(|\varkappa| - |\varkappa^\circ|). \quad (8.7)$$

We now bound the difference between $\|\Pi_{\varkappa, \varkappa^\circ} \varepsilon\|^2$ and its expectation using Bernstein-type inequality for quadratic forms; see Corollary B.1.2 in Section ??: for homogeneous Gaussian noise ε and $m \stackrel{\text{def}}{=} |\varkappa| - |\varkappa^\circ|$ with $\alpha > 0$

$$\mathbb{P}(\sigma^{-1}\|\Pi_{\varkappa, \varkappa^\circ} \varepsilon\| > z(m, x)) \leq e^{-x},$$

where

$$z^2(m, x) \leq m + 2\sqrt{m x} + 2x.$$

Given C_0 , define x_m by

$$z^2(m, x_m) = C_0 m.$$

Then, with $N_m = \#\mathcal{M}_m(\varkappa^\circ)$ for $\mathcal{M}_m(\varkappa^\circ) \stackrel{\text{def}}{=} \{\varkappa > \varkappa^\circ : |\varkappa - \varkappa^\circ| = m\}$

$$\begin{aligned} \mathbb{P}\left(\bigcup_{\varkappa \in \mathcal{M}_\tau(\varkappa^\circ)} \{\sigma^{-2}\|\Pi_{\varkappa, \varkappa^\circ} \varepsilon\|^2 > C_0 m\}\right) \\ \leq \sum_{m=\tau}^p \sum_{\varkappa \in \mathcal{M}_m(\varkappa^\circ)} \mathbb{P}(\sigma^{-2}\|\Pi_{\varkappa, \varkappa^\circ} \varepsilon\|^2 > z^2(m, x_m)) \leq \sum_{m=\tau}^p N_m e^{-x_m}. \end{aligned}$$

The Stirling formula yields $N_m \leq (p/m)^m$ and the sum can be bounded for $\alpha > \log(p/\tau)$.

Theorem 8.1.3. *Consider a linear model with a homogeneous Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. If $\hat{\varkappa}$ is defined by (8.2) with $C = C_0 + \beta$, then for any good model \varkappa° , it holds*

$$\mathbb{P}(\varkappa^\circ \text{ is rejected}) \leq \sum_{m=\tau}^p N_m e^{-x_m}.$$

8.1.2 Oracle inequality for $\widehat{\kappa}$ -choice

Let $\kappa_0 \supseteq \text{supp}(\boldsymbol{\theta}^*)$ and let $\widehat{\kappa}$ be the selected model. First we bound the probability that $\widehat{\kappa}$ is not contained in κ_0 . Consider $\overline{\kappa} = \widehat{\kappa} \setminus \kappa_0$. The definition of $\widehat{\kappa}$ ensures by (8.3) that

$$\|\boldsymbol{\Psi}^\top (\widetilde{\boldsymbol{\theta}}_{\widehat{\kappa}} - \widetilde{\boldsymbol{\theta}}_{\overline{\kappa}})\|^2 \geq C(|\widehat{\kappa}| - |\overline{\kappa}|).$$

But

$$\|\boldsymbol{\Psi}^\top (\widetilde{\boldsymbol{\theta}}_{\widehat{\kappa}} - \widetilde{\boldsymbol{\theta}}_{\overline{\kappa}})\|^2 = \|I_{\widehat{\kappa} \cup \kappa_0, \kappa_0} \mathbf{Y}\|^2 = \|I_{\widehat{\kappa} \cup \kappa_0, \kappa_0} \boldsymbol{\varepsilon}\|^2$$

and by (8.4), this event can only happen with a very small probability. Otherwise $\widehat{\kappa} \subseteq \kappa_0$ and we obtain by the acceptance rule

$$\|\boldsymbol{\Psi}^\top (\widetilde{\boldsymbol{\theta}}_{\widehat{\kappa}} - \widetilde{\boldsymbol{\theta}}_{\kappa_0})\|^2 \leq C(|\kappa_0| - |\widehat{\kappa}|) \leq C|\kappa_0|. \quad (8.8)$$

Theorem 8.1.4. *Consider a linear model with a homogeneous Gaussian noise $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n)$. Let κ_0 be a good model due to (8.7). If C is selected to ensure the propagation condition (8.6) then the probability of the event $\widehat{\kappa} \leq \kappa_0$ is at least $1 - e^{-x}$ and on this event, it holds (8.8).*

8.1.3 Numerical results

This section presents numerical results for one special case when the procedure can be easily implemented. Namely, we consider a sequence space model with just one observation per parameter:

$$Y_i = \theta_i^* + \varepsilon_i.$$

8.1.4 Bootstrap based tuning of penalty

Now we discuss the situation when the noise distribution is unknown. Then the relation (8.5) cannot be used for fixing the constant C_0 . Below we discuss the bootstrap based choice of this constant. As in Section ?? we fix some model-candidate κ° and consider the family of estimates

$$\widetilde{\boldsymbol{\theta}}_\kappa = (\boldsymbol{\Psi}_\kappa \boldsymbol{\Psi}_\kappa^\top)^{-1} \boldsymbol{\Psi}_\kappa \mathbf{Y}$$

for $\kappa > \kappa^\circ$. We also suppose a pilot estimate $\widetilde{\boldsymbol{\theta}}$ to be given which provides a reasonable data fit but is probably too volatile. Define the residuals $\check{\boldsymbol{\varepsilon}} = \mathbf{Y} - \boldsymbol{\Psi}^\top \widetilde{\boldsymbol{\theta}}$. The procedure

follows the same path as in the ordered case. For each \varkappa we compute and store the corresponding MLE $\tilde{\boldsymbol{\theta}}_\varkappa$ and a collection of the bootstrap-based stochastic vectors ζ_\varkappa^b :

$$\zeta_\varkappa^b = (\boldsymbol{\Psi}_\varkappa \boldsymbol{\Psi}_\varkappa^\top)^{-1} \boldsymbol{\Psi}_\varkappa \mathcal{E}^b \check{\boldsymbol{\varepsilon}}.$$

Further we can fix the value C_0 using the bootstrap differences

$$\xi_{\varkappa, \varkappa^o}^b = W(\zeta_\varkappa^b - \zeta_{\varkappa^o}^b)$$

for the weighting loss matrix W . The bootstrap critical values can be computed from these differences by the bootstrap analog of the propagation condition (8.4)

$$\mathbb{P}^b \left(\bigcup_{\varkappa \in \mathcal{M}_\tau(\varkappa^o)} \{ \|\xi_{\varkappa, \varkappa^o}^b\|^2 > C_0(|\varkappa| - |\varkappa^o|) \} \right) \leq e^{-x}.$$

Here \mathbb{P}^b is to be understood as the empirical bootstrap measure.

To be done: Bootstrap validity

8.2 Sparse penalty

This section discusses the use of a sparse penalty based on the ℓ_1 -norm of the vector $\boldsymbol{\theta}$. Below for a vector $\boldsymbol{\theta} \in \mathbb{R}^p$ we denote

$$\|\boldsymbol{\theta}\|_1 = \sum_{j=1}^p |\theta_j|, \quad \|\boldsymbol{\theta}\|^2 = \sum_{j=1}^p \theta_j^2, \quad \|\boldsymbol{\theta}\|_\infty = \max_{j \leq p} |\theta_j|.$$

We consider the LASSO type procedure which is based on minimization of the empirical risk $\|\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}\|^2$ penalized by $\lambda \|\boldsymbol{\theta}\|_1$:

$$\hat{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{J}_\lambda(\boldsymbol{\theta}) = \operatorname{argmin}_{\boldsymbol{\theta}} \{ \|\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|_1 \}. \quad (8.9)$$

Note that the formulation of the problem implicitly assumes that all components θ_j of the vector $\boldsymbol{\theta}$ have the same impact. This can be translated into the scaling condition on the matrix $\boldsymbol{\Psi}$. Usually this matrix and thus, the coefficients θ_j are rescaled in a way that the diagonal elements of the matrix $\boldsymbol{\Psi} \boldsymbol{\Psi}^\top$ are equal to one:

$$\sum_{i=1}^n \psi_{i,j}^2 = 1.$$

The problem (8.9) has a closed form solution only in very special situation. One of them is when the matrix $\boldsymbol{\Psi}$ is orthogonal.

Theorem 8.2.1. Let $n \geq p$ and $\Psi\Psi^\top = I_p$. Then $\hat{\theta}$ is obtained by soft-thresholding of $\tilde{\theta} = \Psi Y$:

$$\hat{\theta}_j = \begin{cases} (\tilde{\theta}_j - \lambda)_+ & \tilde{\theta}_j \geq 0, \\ -(|\tilde{\theta}_j| - \lambda)_+ & \tilde{\theta}_j < 0 \end{cases}$$

However, compared to the complexity penalization, the problem (8.9) can be solved numerically because the objective function is convex.

Now we briefly discuss some properties of the solution. The underlying structural assumption is that the true vector θ^* is sparse, that is, most of its entries vanish. By κ^* we denote the corresponding oracle support. We first consider the case when Ψ_{κ^*} is orthogonal to the rest of Ψ . Our first result shows that a proper choice of the parameter λ ensures a sparse solution: the non-zero coefficients of $\hat{\theta}$ are all located within κ^* .

Theorem 8.2.2. Let θ^* be supported on κ_0 , κ_0^c be the complement of κ_0 , and

$$\Psi_{\kappa_0}\Psi_{\kappa_0^c}^\top = 0. \quad (8.10)$$

If the coefficient λ fulfills

$$2\|\Psi\varepsilon\|_\infty \leq \lambda,$$

then

$$\hat{\kappa} \subseteq \kappa_0. \quad (8.11)$$

Proof. It suffices to check for each candidate θ that the criteria in the optimization problem (8.9) only improves if we kill all its entries which do not enter in the set κ_0 . Let κ be the support of θ . Define $\theta_{\kappa_0} = \Pi_{\kappa_0}\theta$ as the restriction of the parameter vector θ to κ_0 , and similarly $\theta_{\kappa_0^c} = \Pi_{\kappa_0^c}\theta$ is the projection on the complement set κ_0^c . Obviously $\theta_{\kappa_0^c} = \theta - \theta_{\kappa_0}$. Then the model equation $Y = \Psi^\top\theta^* + \varepsilon$ implies

$$\begin{aligned} \|Y - \Psi^\top\theta\|^2 - \|Y - \Psi^\top\theta_{\kappa_0}\|^2 &= \|\varepsilon - \Psi^\top(\theta_{\kappa_0} + \theta_{\kappa_0^c} - \theta^*)\|^2 - \|\varepsilon - \Psi^\top(\theta_{\kappa_0} - \theta^*)\|^2 \\ &= \|\Psi^\top\theta_{\kappa_0^c}\|^2 - 2\{\varepsilon - \Psi^\top(\theta_{\kappa_0} - \theta^*)\}^\top\Psi^\top\theta_{\kappa_0^c}. \end{aligned}$$

Exercise 8.2.1. Check that

$$\{\Psi^\top(\theta_0 - \theta^*)\}^\top\Psi^\top\theta_{\kappa_0^c} = 0.$$

Hint: use that $\theta_{\kappa_0} - \theta^*$ is supported on κ_0 , and $\theta_{\kappa_0^c} = \theta - \theta_{\kappa_0}$ on κ_0^c . Then the result follows from (8.10).

Now $\|\boldsymbol{\theta}\|_1 - \|\boldsymbol{\theta}_{\varkappa_0}\|_1 = \|\boldsymbol{\theta}_{\varkappa_0^c}\|_1$ and

$$\begin{aligned} & \|\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}\|^2 - \|\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}_{\varkappa_0}\|^2 + \lambda(\|\boldsymbol{\theta}\|_1 - \|\boldsymbol{\theta}_{\varkappa_0}\|_1) \\ &= \|\boldsymbol{\Psi}^\top \boldsymbol{\theta}_{\varkappa_0^c}\|^2 + \lambda\|\boldsymbol{\theta}_{\varkappa_0^c}\|_1 - 2(\boldsymbol{\Psi}\boldsymbol{\varepsilon})^\top \boldsymbol{\theta}_{\varkappa_0^c}. \end{aligned}$$

It remains to check that the condition $2\|\boldsymbol{\Psi}\boldsymbol{\varepsilon}\|_\infty \leq \lambda$ ensures that

$$\lambda\|\boldsymbol{\theta}_{\varkappa_0^c}\|_1 - 2(\boldsymbol{\Psi}\boldsymbol{\varepsilon})^\top \boldsymbol{\theta}_{\varkappa_0^c} \geq 0.$$

Therefore, reduction of $\boldsymbol{\theta}$ to $\boldsymbol{\theta}_{\varkappa_0}$ only improves the objective function, and the result follows.

8.2.1 Basic inequality

If we drop the orthogonality condition (8.10), the wonderful oracle result (8.11) does not hold any more. However, one can establish an oracle bound on the quadratic risk in terms of the sparsity value $\|\boldsymbol{\theta}^*\|_1$. We again check the condition that $\boldsymbol{\theta}$ is better than $\boldsymbol{\theta}^*$ w.r.t. the criterion $\mathcal{J}_\lambda(\boldsymbol{\theta})$ from (8.9). It holds due to the model equation $\mathbf{Y} = \boldsymbol{\Psi}^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$

$$\begin{aligned} \|\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}\|^2 - \|\mathbf{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}^*\|^2 &= \|\boldsymbol{\varepsilon} - \boldsymbol{\Psi}^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 - \|\boldsymbol{\varepsilon}\|^2 \\ &= -2(\boldsymbol{\Psi}\boldsymbol{\varepsilon})^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \|\boldsymbol{\Psi}^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2. \end{aligned}$$

The event $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ is only possible if $\mathcal{J}_\lambda(\boldsymbol{\theta}) \leq \mathcal{J}_\lambda(\boldsymbol{\theta}^*)$. This yields

$$\|\boldsymbol{\Psi}^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + \lambda\|\boldsymbol{\theta}\|_1 \leq 2(\boldsymbol{\Psi}\boldsymbol{\varepsilon})^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \lambda\|\boldsymbol{\theta}^*\|_1 \quad (8.12)$$

Moreover, if the score $\nabla = \boldsymbol{\Psi}\boldsymbol{\varepsilon}$ fulfills $\|\nabla\|_\infty \leq C_0$, then

$$2(\boldsymbol{\Psi}\boldsymbol{\varepsilon})^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \leq 2C_0\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 \quad (8.13)$$

By the triangle inequality $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 \geq \|\boldsymbol{\theta}\|_1 - \|\boldsymbol{\theta}^*\|_1$, and (8.12) and (8.13) imply

$$\|\boldsymbol{\Psi}^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + (\lambda - 2C_0)\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 \leq 2\lambda\|\boldsymbol{\theta}^*\|_1$$

If $\lambda > 2C_0$, this inequality provides a number of informative messages. The prediction loss $\|\boldsymbol{\Psi}^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2$ and the estimation loss $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1$ can be bounded as

$$\begin{aligned} \|\boldsymbol{\Psi}^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 &\leq 2\lambda\|\boldsymbol{\theta}^*\|_1, \\ \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 &\leq \frac{2\lambda}{\lambda - 2C_0}\|\boldsymbol{\theta}^*\|_1. \end{aligned}$$

The last inequality can be made more precise if the true value is supported on \varkappa_0 . Consider for any $\boldsymbol{\theta}$ the decomposition $\boldsymbol{\theta} = \boldsymbol{\theta}_{\varkappa_0} + \boldsymbol{\theta}_{\varkappa_0^c}$, where $\boldsymbol{\theta}_{\varkappa_0}$ is supported on

\varkappa_0 and $\boldsymbol{\theta}_{\varkappa_0^c}$ on its complement \varkappa_0^c . Obviously $\|\boldsymbol{\theta}\|_1 = \|\boldsymbol{\theta}_{\varkappa_0}\|_1 + \|\boldsymbol{\theta}_{\varkappa_0^c}\|_1$. Denote also $\mathbf{u} = \boldsymbol{\theta}_{\varkappa_0} - \boldsymbol{\theta}^*$. By construction, \mathbf{u} is supported on \varkappa_0 . It holds

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 = \|\mathbf{u}\| + \|\boldsymbol{\theta}_{\varkappa_0^c}\|_1.$$

Now (8.12) and (8.13) imply

$$\|\boldsymbol{\Psi}^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + \lambda(\|\boldsymbol{\theta}_{\varkappa_0}\|_1 + \|\boldsymbol{\theta}_{\varkappa_0^c}\|_1) \leq 2C_0(\|\mathbf{u}\| + \|\boldsymbol{\theta}_{\varkappa_0^c}\|_1) + \lambda\|\boldsymbol{\theta}^*\|_1$$

and therefore, the component $\boldsymbol{\theta}_{\varkappa_0^c}$ of $\boldsymbol{\theta}$ fulfills

$$\|\boldsymbol{\Psi}^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + (\lambda - 2C_0)\|\boldsymbol{\theta}_{\varkappa_0^c}\|_1 \leq 2C_0\|\mathbf{u}\|_1 + \lambda\|\boldsymbol{\theta}^*\|_1 - \lambda\|\boldsymbol{\theta}_{\varkappa_0}\|_1 \leq (\lambda + 2C_0)\|\mathbf{u}\|_1.$$

Theorem 8.2.3. *Let $\mathbf{Y} = \boldsymbol{\Psi}^\top\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ and $\nabla = \boldsymbol{\Psi}\boldsymbol{\varepsilon}$ fulfill $\|\boldsymbol{\Psi}\boldsymbol{\varepsilon}\|_\infty \leq C_0$. If $\lambda > 2C_0$, then the estimate $\hat{\boldsymbol{\theta}}$ satisfies*

$$\|\boldsymbol{\Psi}^\top(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 + (\lambda - 2C_0)\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq 2\lambda\|\boldsymbol{\theta}^*\|_1$$

In addition, if $\boldsymbol{\theta}^*$ is supported on \varkappa_0 , then the projections $\hat{\boldsymbol{\theta}}_{\varkappa_0}$ and $\hat{\boldsymbol{\theta}}_{\varkappa_0^c}$ of $\hat{\boldsymbol{\theta}}$ to \varkappa_0 and \varkappa_0^c can be related as

$$\|\boldsymbol{\Psi}^\top(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 + (\lambda - 2C_0)\|\hat{\boldsymbol{\theta}}_{\varkappa_0^c}\|_1 \leq (\lambda + 2C_0)\|\hat{\boldsymbol{\theta}}_{\varkappa_0} - \boldsymbol{\theta}^*\|_1.$$

To be done: Compatibility condition

To be done: Restricted isometry and oracle bound

8.2.2 Dual problem and Danzig selector

The LASSO optimization can be viewed as minimizing the fit $\|\mathbf{Y} - \boldsymbol{\Psi}^\top\boldsymbol{\theta}\|^2$ under the constraint on the ℓ_1 -norm of $\boldsymbol{\theta}$. Then the objective (8.9) is obtained by Lagrange multiplier method. One can also consider the dual problem: minimizing ℓ_1 -norm of $\boldsymbol{\theta}$ under the fit constraints. The dual problem is known as Danzig selector and reads as

$$\hat{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname{arginf}_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_1 \quad \text{subject to} \quad 2\|\boldsymbol{\Psi}(\mathbf{Y} - \boldsymbol{\Psi}^\top\boldsymbol{\theta})\|_\infty \leq \lambda. \quad (8.14)$$

The true value $\boldsymbol{\theta}^*$ is a natural candidate. Then $\|\boldsymbol{\Psi}(\mathbf{Y} - \boldsymbol{\Psi}^\top\boldsymbol{\theta}^*)\|_\infty = \|\boldsymbol{\Psi}\boldsymbol{\varepsilon}\|_\infty$. If λ is selected properly to ensure $2\|\boldsymbol{\Psi}\boldsymbol{\varepsilon}\|_\infty \leq \lambda$, then the constraints meet, and the objective functional is equal to $\|\boldsymbol{\theta}^*\|_1$. Therefore, the solution $\hat{\boldsymbol{\theta}}$ cannot be less sparse than $\boldsymbol{\theta}^*$ in the sense $\|\hat{\boldsymbol{\theta}}\|_1 \leq \|\boldsymbol{\theta}^*\|_1$.

Theorem 8.2.4. *Let $\hat{\boldsymbol{\theta}}$ be defined by (8.14). If $2\|\boldsymbol{\Psi}\boldsymbol{\varepsilon}\|_\infty \leq \lambda$ then $\|\hat{\boldsymbol{\theta}}\|_1 \leq \|\boldsymbol{\theta}^*\|_1$.*

8.2.3 Data-driven choice of λ

The necessary requirement to the choice of λ is that the score vector $\nabla = \Psi \varepsilon$

$$2\|\Psi \varepsilon\|_\infty \leq \lambda.$$

This condition can be assessed by replacing the true noise by its bootstrap counterpart $\nabla^b = \Psi \varepsilon^b$. With a pilot $\tilde{\theta}$, define $\varepsilon^b = \mathcal{W}^b(\mathbf{Y} - \Psi^\top \tilde{\theta})$ and fix λ^b by the condition

$$\mathbb{P}^b(2\|\nabla^b\|_\infty > \lambda^b) \leq e^{-x}.$$

The original procedure has to be applied with $\lambda = \lambda^b + \beta$.

Part II

General parametric theory

... And, then, there were only
illusion and the road.

J. Brodsky

Fisher and Wilks expansion

This chapter presents two prominent results of classical parametric statistics, namely the Fisher and Wilks Theorems, in a non-classical framework. The main features to be addressed here are a finite sample setup with large parameter dimension and a possible model misspecification.

First we specify our set-up. Let \mathbf{Y} denote the observed data and \mathcal{P} mean their distribution. A general parametric assumption (PA) means that \mathcal{P} belongs to p -dimensional family $(\mathcal{P}_\theta, \theta \in \Theta \subseteq \mathbb{R}^p)$ dominated by a measure μ_0 . This family yields the log-likelihood function $L(\theta) = L(\mathbf{Y}, \theta) \stackrel{\text{def}}{=} \log \frac{d\mathcal{P}_\theta}{d\mu_0}(\mathbf{Y})$. The PA can be misspecified, so, in general, $L(\theta)$ is a *quasi log-likelihood*. The classical likelihood principle suggests to estimate θ by maximizing the function $L(\theta)$:

$$\tilde{\theta} \stackrel{\text{def}}{=} \operatorname{argmax}_{\theta \in \Theta} L(\theta). \quad (9.1)$$

If $\mathcal{P} \notin (\mathcal{P}_\theta)$, then the (quasi) MLE estimate $\tilde{\theta}$ from (9.1) is still meaningful and it appears to be an estimate of the value θ^* defined by maximizing the expected value of $L(\theta)$:

$$\theta^* \stackrel{\text{def}}{=} \operatorname{argmax}_{\theta \in \Theta} \mathbb{E} L(\theta).$$

Here θ^* is the true value in the parametric situation and can be viewed as the parameter of the best parametric fit in the general case. The study is non-asymptotic, that is, we proceed with only one sample \mathbf{Y} . One can easily extend it to an asymptotic setup in which the data, its distribution, the parameter space and the parametric family depend on the asymptotic parameter like the sample size. One example is given below in Section 13.3 for the case of an i.i.d. sample.

The Fisher expansion of the qMLE $\tilde{\theta}$ is given as follows:

$$D(\tilde{\theta} - \theta^*) \approx \xi \stackrel{\text{def}}{=} D^{-1} \nabla L(\theta^*),$$

where $\nabla L(\boldsymbol{\theta}) = \frac{dL}{d\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $D^2 \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}^*)$ is the analog of the total Fisher information matrix. In classical situations, the standardized score $\boldsymbol{\xi}$ is asymptotically standard normal yielding asymptotic root-n normality and efficiency of the MLE $\tilde{\boldsymbol{\theta}}$. Theorem 9.3.2 carefully describes how the error of this expansion depends on the parameter dimension p and the regularity of the model. The Wilks expansion means

$$L(\tilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) \approx \|\boldsymbol{\xi}\|^2/2.$$

Again, if the vector $\boldsymbol{\xi}$ is asymptotically standard normal, the expansion yields the classical χ_p^2 asymptotic distribution for the excess $L(\tilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$.

The whole study is nonasymptotic and all “small” terms are carefully described. This helps to understand how the parameter dimension is involved and particularly to address the question of a *critical dimension*; see Section 13.3 which specifies the result to the i.i.d. case with n observations and links the obtained results to the classical literature.

9.1 Main results

This section presents our main results which include the Fisher and Wilks expansions for a non-classical and non-asymptotic framework. First we present the frequentist results: concentration and large deviation properties of the maximum likelihood estimator $\tilde{\boldsymbol{\theta}}$, the Fisher expansion for the difference $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}$ and the Wilks expansion for the excess $L(\tilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta})$. The results are stated in a concise way, all the terms are given explicitly. Surprisingly, the leading terms in all bounds are sharp, in particular, the classical results on asymptotic efficiency can be easily derived from the obtained expansions.

Introduce the notation $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$ for the (quasi) log-likelihood ratio. The main step in the approach is the following *uniform local bracketing result*:

$$\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \Delta \leq L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \leq \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \Delta, \quad \boldsymbol{\theta} \in \Theta_0. \quad (9.2)$$

Here $\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is a quadratic in $\boldsymbol{\theta} - \boldsymbol{\theta}^*$ expression, Δ is a small error only depending on Θ_0 which is a local vicinity of the central point $\boldsymbol{\theta}^*$. This result can be viewed as an extension of the famous Le Cam *local asymptotic normality* (LAN) condition. The LAN condition postulates an approximation of the log-likelihood $L(\boldsymbol{\theta})$ by a nearly Gaussian process; see e.g. Ibragimov and Khas'minskij (1981) or Kleijn and van der Vaart (2012) for an extension of this condition (stochastic LAN). The bracketing bound (9.2) requires only some general conditions listed in Section 9.2. A model misspecification case is included. Similarly to the LAN theory, the bracketing result has a number of remarkable corollaries like the Wilks and Fisher Theorems; see Theorems 9.3.2 and 9.3.3.

For making a precise statement, we have to specify the ingredients of the bracketing device. The most important one is a symmetric positive $p \times p$ -matrix D^2 . In typical situations, it can be defined as the negative Hessian of the expected log-likelihood: $D^2 = \mathbb{F}(\boldsymbol{\theta}^*) = -\nabla^2 \mathbb{E}L(\boldsymbol{\theta}^*)$. Also one has to specify a radius \mathbf{r}_0 entering in the definition of the local vicinity $\Theta_0(\mathbf{r}_0)$ of the central point $\boldsymbol{\theta}^*$: $\Theta_0(\mathbf{r}_0) = \{\boldsymbol{\theta}: \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}_0\}$. The bracketing result (9.2) can be stated for $\Theta_0 = \Theta_0(\mathbf{r}_0)$ with

$$\begin{aligned}\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &\stackrel{\text{def}}{=} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla L(\boldsymbol{\theta}^*) - \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2 \\ &= \boldsymbol{\xi}^\top D(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2\end{aligned}$$

and

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} D^{-1} \nabla L(\boldsymbol{\theta}^*). \quad (9.3)$$

The construction is essentially changed relative to Spokoiny (2012) (in fact, it is simplified) by using only one matrix D^2 while Spokoiny (2012) used three matrices D_ϵ^2 , $D_{\epsilon\epsilon}^2$, and V^2 . The bracketing bound (9.2) becomes useful if the error Δ is relatively small and can be neglected.

9.2 Non-Gaussian case: conditions

This section collects the conditions which are systematically used in the text. The conditions are quite general and seem to be non-restrictive; see the discussion at the end of the section. Moreover, these conditions can be viewed just as definitions of some quantities which measure some features of the model and will enter in the final error bounds. We mainly require some regularity and smoothness of the log-likelihood process $L(\boldsymbol{\theta})$. Below we assume that the log-likelihood function $L(\boldsymbol{\theta})$ is twice differentiable and denote $\nabla L(\boldsymbol{\theta}) = \frac{dL(\boldsymbol{\theta})}{d\boldsymbol{\theta}}$ and $\nabla^2 L(\boldsymbol{\theta}) = \frac{d^2L(\boldsymbol{\theta})}{d\boldsymbol{\theta}^2}$ for the Hessian of $L(\boldsymbol{\theta})$. We also suppose that one can exchange expectation and differentiation and use $\nabla \mathbb{E}L(\boldsymbol{\theta}) = \mathbb{E}\{\nabla L(\boldsymbol{\theta})\}$, $\nabla^2 \mathbb{E}L(\boldsymbol{\theta}) = \mathbb{E}\nabla^2 L(\boldsymbol{\theta})$.

With $D^2 = -\nabla^2 \mathbb{E}L(\boldsymbol{\theta}^*)$, define the local elliptic sets $\Theta_0(\mathbf{r})$ as

$$\Theta_0(\mathbf{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} \in \Theta: \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}\}.$$

We distinguish between local and global conditions. Local ones are stated on $\Theta_0(\mathbf{r}_0)$, while the global one corresponds to $\mathbf{r} \geq \mathbf{r}_0$, where the value \mathbf{r}_0 will be specified later.

The first condition requires that the expected log-likelihood $\mathbb{E}L(\boldsymbol{\theta})$ is twice continuously differentiable.

(L₀) For each $\mathbf{r} \leq \mathbf{r}_0$, there is a constant $\delta(\mathbf{r}) \leq 1/2$ such that it holds for any $\boldsymbol{\theta} \in \Theta_0(\mathbf{r})$ and $\mathbb{F}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E}L(\boldsymbol{\theta})$:

$$\|D^{-1}\mathbb{F}(\boldsymbol{\theta})D^{-1} - I_p\|_{\text{op}} \leq \delta(\mathbf{r}).$$

In fact, this condition just defines the modulus of continuity of the matrix function $\mathbb{F}(\boldsymbol{\theta})$ in $\Theta_0(\mathbf{r})$. Under **(L₀)**, it follows from the second order Taylor expansion at $\boldsymbol{\theta}^*$:

$$|-2\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2| \leq \delta(\mathbf{r})\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2, \quad \boldsymbol{\theta} \in \Theta_0(\mathbf{r}).$$

For $\mathbf{r} > \mathbf{r}_0$, we need a global identification property which ensures that the deterministic component $\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ of the log-likelihood is competitive with the variation of the stochastic component.

(L) For some constant $C_1 \geq 0$, it holds for $\mathbf{r} > \mathbf{r}_0$

$$\mathbb{E}L(\boldsymbol{\theta}^*) - \mathbb{E}L(\boldsymbol{\theta}) \geq \{1 - \delta(\mathbf{r}_0)\}(\mathbf{r}_0 \mathbf{r} - \frac{1}{2}\mathbf{r}_0^2) + C_1 \mathbf{r}^2, \quad \mathbf{r} = \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|.$$

Remark 9.2.1. Conditions **(L₀)** and **(L)** can be effectively checked if the function $f(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\mathbb{E}L(\boldsymbol{\theta})$ is smooth and convex in $\boldsymbol{\theta}$. Equivalently, the matrix $\nabla^2 f(\boldsymbol{\theta})$ is positively semidefinite everywhere. Continuity of the second derivative $\nabla^2 f(\boldsymbol{\theta})$ in $\Theta_0(\mathbf{r}_0)$ implies **(L₀)**. Convexity of f implies for any $\boldsymbol{\theta}^\circ = \boldsymbol{\theta}^* + \rho(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ with $\rho = \mathbf{r}_0/\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq 1$

$$f(\boldsymbol{\theta}) \geq f(\boldsymbol{\theta}^\circ) + (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top \nabla f(\boldsymbol{\theta}^\circ) \geq f(\boldsymbol{\theta}^\circ) + \{D(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\}^\top D^{-1} \nabla f(\boldsymbol{\theta}^\circ).$$

By construction, $\|D(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*)\| = \mathbf{r}_0$, and it follows by condition **(L₀)** that $f(\boldsymbol{\theta}^\circ) \geq f(\boldsymbol{\theta}^*) + (1 - \delta)\mathbf{r}_0^2/2$ and $\nabla^2 f(\boldsymbol{\theta}^\circ) \geq (1 - \delta)D^2$ with $\delta = \delta(\mathbf{r}_0)$. As $\|D(\boldsymbol{\theta}^\circ - \boldsymbol{\theta})\| = \mathbf{r} - \mathbf{r}_0$, we conclude that

$$f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^*) \geq (1 - \delta) \left\{ \frac{\mathbf{r}_0^2}{2} + \mathbf{r}_0(\mathbf{r} - \mathbf{r}_0) \right\}$$

and **(L)** follows for $C_1 = 0$.

Remark 9.2.2. The case $C_1 > 0$ and $1 - \delta - C_1 > 0$ can be similarly checked under a strong convexity of $f(\boldsymbol{\theta})$:

$$\nabla^2 f(\boldsymbol{\theta}) \geq C_1 D^2.$$

This guarantees that the difference $f(\boldsymbol{\theta}) = -\mathbb{E}L(\boldsymbol{\theta}) - C_1\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2$ is a convex function and the above arguments apply with obvious corrections: $f(\boldsymbol{\theta}^\circ) \geq f(\boldsymbol{\theta}^*) + (1 - \delta - C_1)\mathbf{r}_0^2/2$ and $\nabla^2 f(\boldsymbol{\theta}^\circ) \geq (1 - \delta - C_1)D^2$. Then the same arguments ensure condition **(L)** with δ replaced by $\delta + C_1$.

In the case of linear or generalized linear models, one can use $C_1 = 0$. In regular situations, it suffices that (\mathcal{L}) holds with C_1 of order $n^{-1/2}$.

Now we consider the stochastic component of the process $L(\boldsymbol{\theta})$:

$$\zeta(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - \mathbb{E}L(\boldsymbol{\theta}).$$

We assume that it is twice differentiable and denote by $\nabla\zeta(\boldsymbol{\theta})$ its gradient and by $\nabla^2\zeta(\boldsymbol{\theta})$ its Hessian matrix.

(ED₀) *There exist a positive symmetric matrix V^2 , and constants $g > 0$, $\nu_0 \geq 1$ such that $\text{Var}\{\nabla\zeta(\boldsymbol{\theta}^*)\} \leq V^2$ and*

$$\sup_{\gamma \in \mathbb{R}^p} \log \mathbb{E} \exp \left\{ \lambda \frac{\gamma^\top \nabla\zeta(\boldsymbol{\theta}^*)}{\|V\gamma\|} \right\} \leq \frac{\nu_0^2 \lambda^2}{2}, \quad |\lambda| \leq g.$$

(ED₂) *There exist a value $\omega > 0$ and for each $r > 0$, a constant $g(r) > 0$ such that it holds for any $\boldsymbol{\theta} \in \Theta_0(r)$:*

$$\sup_{\gamma_1, \gamma_2 \in \mathbb{R}^p} \log \mathbb{E} \exp \left\{ \frac{\lambda}{\omega} \frac{\gamma_1^\top \nabla^2\zeta(\boldsymbol{\theta}) \gamma_2}{\|D\gamma_1\| \cdot \|D\gamma_2\|} \right\} \leq \frac{\nu_0^2 \lambda^2}{2}, \quad |\lambda| \leq g(r).$$

Condition **(ED₀)** and **(ED₂)** basically require that the stochastic component $\zeta(\boldsymbol{\theta})$ and its first and second derivative have finite exponential moments. Then one can use the fact that existence of the exponential moment $\mathbb{E}e^{\lambda_0 \xi}$ for a centered random variable ξ and some fixed λ_0 implies that the moment generating function $f_\xi(\lambda) \stackrel{\text{def}}{=} \log \mathbb{E}e^{\lambda \xi}$ is analytic in $\lambda \in (0, \lambda_0)$ and can be well majorated by a quadratic function in a smaller interval $(0, \lambda_1]$ for $\lambda_1 < \lambda_0$. In words, sub-exponential tail condition implies local sub-Gaussian behavior for a restricted range $\lambda \leq g(r)$. Below we only need that the constant $g(r)$ is larger than Cp for a fixed constant C and all r .

The *identifiability condition* relates the matrices V^2 and D^2 .

(I) There is a constant $a > 0$ such that

$$a^2 D^2 \geq V^2.$$

Remark 9.2.3. The conditions involve some constants. We distinguish between important constants and technical ones. The impact of the important constants is shown in our results, the list includes $\delta(r)$, ω , and a . The constant a can be viewed as the largest eigenvalue of $B = D^{-1}V^2D^{-1}$ and it enters in the definition of the upper quantile function $z(B, x)$ for $\|\xi\|$; see Proposition B.2.1 below. The other constants like ν_0 or $g(r)$ are technical. The constant ν_0 is introduced for convenience only, it can be omitted by rescaling the matrix V . In the asymptotic setup it can usually be selected very close to one.

Remark 9.2.4. We briefly comment how restrictive the imposed conditions are. Spokoiny (2012), Section 5.1, considered in details the i.i.d. case and presented some mild sufficient conditions on the parametric family which imply the above general conditions. Condition **(ED₀)** requires some exponential moments of the observations (errors). Usually one only assumes some finite moments of the errors; cf. Ibragimov and Khas'minskij (1981), Chapter 2. Our condition is a bit more restrictive but it allows to obtain some finite sample bounds. Condition **(L₀)** only requires some regularity of the considered parametric family and is not restrictive. Conditions **(ED₂)** with $g(r) \equiv g > 0$ and **(L)** with $C_1 > 0$ are easy to verify if the parameter set Θ is compact and the sample size n exceeds Cp for a fixed constant C . It suffices to check a usual identifiability condition that the value $IEL(\theta, \theta^*)$ does not vanish for $\theta \neq \theta^*$.

The regression and generalized regression models are included as well; cf. Ghosal (1999, 2000) or Kim (2006). Spokoiny (2012), Section 5.2, argued that **(ED₂)** is automatically fulfilled for a generalized linear model, while **(ED₀)** requires that regression errors have to fulfill some exponential moments conditions. If this condition is too restrictive and a more stable (robust) estimation procedure is desirable, one can apply the LAD-type contrast leading to median regression. Spokoiny (2012), Section 5.3, showed for the case of linear median regression that all the required conditions are fulfilled automatically if the sample size n exceeds Cp for a fixed constant C . Spokoiny et al. (2013) applied this approach for local polynomial quantile regression. Zaitsev et al. (2013) applied the approach to the problem of regression with Gaussian process where the unknown parameters enter in the likelihood in a rather complicated way.

9.3 Properties of the MLE $\tilde{\theta}$

This section collects the main results about the MLE $\tilde{\theta}$. We begin by a large deviation bound which ensures a small probability of the event $\tilde{\theta} \notin \Theta_0(r_0)$. Then we present the Fisher and Wilks expansions. The formulation involves two growing functions of the argument x : $z(B, x)$ and $\mathfrak{z}_{\mathbb{H}}(x)$. The functions are given analytically and only depend on the parameters of the model. The function $z(B, x)$ with $B = D^{-1}V^2D^{-1}$ describes the quantiles of the norm of the vector $\|\xi\|$ from (9.3). One can use an upper bound $z(B, x) \leq \sqrt{\text{tr } B} + \sqrt{2\lambda_{\max}(B)x}$, the exact definition is given in (B.22). Further, the function $\mathfrak{z}_{\mathbb{H}}(x)$ is related to the entropy of the parameter space and it is given by (H.19). In typical situations $\mathfrak{z}_{\mathbb{H}}(x) \approx 2\sqrt{6p} + \sqrt{8x}$, and one can use the upper bound $C\sqrt{p+x}$ for both functions $z(B, x)$ and $\mathfrak{z}_{\mathbb{H}}(x)$. The first result explains the choice of r_0 ensuring with a high probability that $\tilde{\theta} \in \Theta_0(r_0)$.

Theorem 9.3.1. Suppose **(ED₀)** and **(ED₂)**, **(L₀)**, **(L)**, and **(I)**. Let also

$$r_0 \{1 - \delta(r_0)\} \geq 2z(B, x) \quad (9.4)$$

and let the constant C_1 in (\mathcal{L}) satisfy

$$C_1 \geq \varrho(r, x), \quad r > r_0,$$

where

$$\varrho(r, x) \stackrel{\text{def}}{=} \sqrt{8} \nu_0 \mathfrak{z}_{\mathbb{H}}(x + \log(2r/r_0)) \omega \quad (9.5)$$

and the function $\mathfrak{z}_{\mathbb{H}}(x)$ is given in Theorem H.6.1. Then

$$IP(\tilde{\boldsymbol{\theta}} \notin \Theta_0(r_0)) \leq 3e^{-x}.$$

Remark 9.3.1. The radius r_0 has to fulfill (9.4). In typical situations $1 - \delta(r_0) \approx 1$, and a simple rule $r_0 \geq (2 + \delta)z(B, x)$ for some $\delta > 0$ works in most of cases.

Now we state the result about the Fisher expansion for the qMLE $\tilde{\boldsymbol{\theta}}$.

Theorem 9.3.2. Suppose the conditions of Theorem 9.3.1. On a random set $\Omega(x)$ of a dominating probability at least $1 - 4e^{-x}$, it holds

$$\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi}\| \leq \diamond(r_0, x), \quad (9.6)$$

where the value $\diamond(r_0, x)$ is defined by

$$\diamond(r_0, x) \stackrel{\text{def}}{=} \{\delta(r_0) + \sqrt{8} \nu_0 \mathfrak{z}_{\mathbb{H}}(x) \omega\} r_0. \quad (9.7)$$

Our version of the Wilks result can be stated in the following form.

Theorem 9.3.3. Suppose the conditions of Theorem 9.3.1. On the same random set $\Omega(x)$ as in Theorem 9.3.2 with $IP(\Omega(x)) \geq 1 - 4e^{-x}$, it holds for $\diamond(r_0, x)$ from (9.7)

$$\begin{aligned} |2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2| &\leq 2r_0 \diamond(r_0, x) + \diamond^2(r_0, x), \\ |2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\boldsymbol{\xi}\|^2| &\leq 2r_0 \diamond(r_0, x) + \diamond^2(r_0, x). \end{aligned} \quad (9.8)$$

Furthermore,

$$\begin{aligned} \left| \sqrt{2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)} - \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \right| &\leq 2\diamond(r_0, x), \\ \left| \sqrt{2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)} - \|\boldsymbol{\xi}\| \right| &\leq 3\diamond(r_0, x), \end{aligned} \quad (9.9)$$

and for any $\boldsymbol{\theta}^\circ \in \Theta_0(r_0)$, it holds on $\Omega(x)$

$$\begin{aligned} \left| \sqrt{2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{\circ})} - \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\circ})\| \right| &\leq 4\Diamond(\mathbf{r}_0, \mathbf{x}), \\ \left| \sqrt{2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{\circ})} - \|\boldsymbol{\xi} + D(\boldsymbol{\theta}^* - \boldsymbol{\theta}^{\circ})\| \right| &\leq 5\Diamond(\mathbf{r}_0, \mathbf{x}). \end{aligned}$$

Remark 9.3.2. The classical Fisher and Wilks results describe asymptotic behavior of the MLE $\tilde{\boldsymbol{\theta}}$ and of the excess $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$. The whole derivations are based on expansions similar to (9.6) and (9.8) and on the limiting behavior of the vector $\boldsymbol{\xi}$ and its squared norm. Under standard assumptions in the regression or i.i.d. setup the vector $\boldsymbol{\xi}$ is standard normal and $\|\boldsymbol{\xi}\|^2$ is asymptotically χ^2 with p degrees of freedom. The asymptotic distribution of the MLE $\tilde{\boldsymbol{\theta}}$ or of the excess $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ can be used for building the confidence sets or for test critical values. However, the use of asymptotic arguments is limited and faces serious problems in practical applications.

This especially concerns the likelihood ratio statistic $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$. It is well recognized that the accuracy of χ^2 -approximation of the tails of $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ is very poor and a reasonable quality requires a huge sample size. If the parameter dimension grows with n this problem becomes even more crucial. The qualitative tail behavior of $\|\boldsymbol{\xi}\|^2$ is described in Proposition B.2.1 but the upper bounds given there appear to be too conservative for practical use.

Remark 9.3.3. Another issue is a possible model misspecification. The expansions (9.8) and (9.9) apply, even if $L(\boldsymbol{\theta})$ is a quasi-log-likelihood function. However, the covariance matrix $V^2 = \text{Var}\{\nabla L(\boldsymbol{\theta}^*)\}$ of the score does not necessarily coincide with the information matrix D^2 . Then the covariance matrix of the vector $\boldsymbol{\xi}$ follows the famous “sandwich” formula $\text{Var}(\boldsymbol{\xi}) = D^{-1}V^2D^{-1}$, and the distribution of the squared norm $\|\boldsymbol{\xi}\|^2$ depends on the unknown covariance matrix V^2 .

9.4 Some auxiliary results and proofs

This section collects some auxiliary results about the behavior of the posterior measures which might be of independent interest.

9.4.1 Local linear approximation of the gradient of the log-likelihood

The principle step of the proof is a bound on the local linear approximation of the gradient $\nabla L(\boldsymbol{\theta})$. Below we study separately its stochastic and deterministic components coming from the decomposition $L(\boldsymbol{\theta}) = \mathbb{E}L(\boldsymbol{\theta}) + \zeta(\boldsymbol{\theta})$. With $D^2 = \mathbb{F}(\boldsymbol{\theta}^*) = -\nabla^2\mathbb{E}L(\boldsymbol{\theta}^*)$, this leads to the decomposition

$$\begin{aligned}\chi(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &\stackrel{\text{def}}{=} D^{-1}\{\nabla L(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta}^*) + D^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\} \\ &= D^{-1}\{\nabla \zeta(\boldsymbol{\theta}) - \nabla \zeta(\boldsymbol{\theta}^*) + \nabla \mathbb{E}L(\boldsymbol{\theta}) - \nabla \mathbb{E}L(\boldsymbol{\theta}^*) + D^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}.\end{aligned}\quad (9.10)$$

First we check the deterministic part. For any $\boldsymbol{\theta}$ with $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}$ and any unit vector \mathbf{u} , it holds

$$\begin{aligned}\mathbf{u}^\top \mathbb{E}\chi(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &\stackrel{\text{def}}{=} \mathbf{u}^\top D^{-1}\{\nabla \mathbb{E}L(\boldsymbol{\theta}) - \nabla \mathbb{E}L(\boldsymbol{\theta}^*) + D^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\} \\ &= \mathbf{u}^\top \{I_p - D^{-1}\mathbb{F}(\boldsymbol{\theta}^*)D^{-1}\} D(\boldsymbol{\theta} - \boldsymbol{\theta}^*),\end{aligned}$$

where $\boldsymbol{\theta}^\circ = \boldsymbol{\theta}^\circ(\mathbf{u})$ is a point on the line connecting $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$. This implies by **(L₀)**

$$\|\mathbb{E}\chi(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\| \leq \|I_p - D^{-1}\mathbb{F}(\boldsymbol{\theta}^*)D^{-1}\|_{\text{op}} \mathbf{r} \leq \delta(\mathbf{r})\mathbf{r}. \quad (9.11)$$

Now we study the stochastic part. Consider the vector process

$$\mathcal{U}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \stackrel{\text{def}}{=} D^{-1}\{\nabla \zeta(\boldsymbol{\theta}) - \nabla \zeta(\boldsymbol{\theta}^*)\}.$$

It is convenient to change the variable by $\mathbf{v} = D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ and consider the vector process $\mathcal{Y}(\mathbf{v}) = \mathcal{U}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$. It obviously holds $\nabla \mathcal{Y}(\mathbf{v}) = D^{-1}\nabla^2 \zeta(\boldsymbol{\theta})D^{-1}$. Moreover, for any unit vectors $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^p$, condition **(ED₂)** implies

$$\log \mathbb{E} \exp \left\{ \frac{\lambda}{\omega} \mathbf{u}_1^\top \nabla \mathcal{Y}(\mathbf{v}) \mathbf{u}_2 \right\} = \log \mathbb{E} \exp \left\{ \frac{\lambda}{\omega} \mathbf{u}_1^\top D^{-1} \nabla^2 \zeta(\boldsymbol{\theta}) D^{-1} \mathbf{u}_2 \right\} \leq \frac{\nu_0^2 \lambda^2}{2}.$$

Define $\Upsilon_\circ(\mathbf{r}) \stackrel{\text{def}}{=} \{\mathbf{v} : \|\mathbf{v}\| \leq \mathbf{r}\}$. Then

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r})} \|\mathcal{U}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\| = \sup_{\mathbf{v} \in \Upsilon_\circ(\mathbf{r})} \|\mathcal{Y}(\mathbf{v})\|. \quad (9.12)$$

Theorem H.10.1 yields with $\varrho(\mathbf{x}) \stackrel{\text{def}}{=} \sqrt{8\nu_0} \mathfrak{z}_{\mathbb{H}}(\mathbf{x}) \omega$

$$\sup_{\mathbf{v} \in \Upsilon_\circ(\mathbf{r})} \|\mathcal{Y}(\mathbf{v})\| \leq \mathbf{r} \varrho(\mathbf{x}) = \sqrt{8\nu_0} \mathfrak{z}_{\mathbb{H}}(\mathbf{x}) \omega \mathbf{r}$$

on a set of a dominating probability at least $1 - e^{-\mathbf{x}}$, where the function $\mathfrak{z}_{\mathbb{H}}(\mathbf{x})$ is given by the following rules:

$$\mathfrak{z}_{\mathbb{H}}(\mathbf{x}) = \begin{cases} 2\sqrt{6p + 2\mathbf{x}}, & \text{if } 6p + 2\mathbf{x} \leq g^2, \\ 2g^{-1}\mathbf{x} + 6g^{-1}p + g, & \text{if } 6p + 2\mathbf{x} > g^2; \end{cases}$$

see Theorem H.6.1 in the Appendix.

Putting together the bounds (9.11) and (9.12) imply the following result.

Proposition 9.4.1. Suppose that the matrix $\mathbb{F}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} L(\boldsymbol{\theta})$ fulfills the condition (\mathcal{L}_0) and let also (\mathbf{ED}_2) be fulfilled on $\Theta_0(\mathbf{r})$ for any fixed \mathbf{r} . Then

$$\mathbb{P} \left\{ \sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r})} \| D^{-1} \{ \nabla L(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta}^*) \} + D(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \| \geq \diamond(r, x) \right\} \leq e^{-x},$$

where

$$\diamond(r, x) \stackrel{\text{def}}{=} \{ \delta(r) + \varrho(x) \} r = \{ \delta(r) + \sqrt{8} \nu_0 \mathfrak{z}_{\mathbb{H}}(x) \omega \} r. \quad (9.13)$$

The result of Proposition 9.4.1 can be extended to the differences $\mathcal{U}(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) = D^{-1} \{ \nabla \zeta(\boldsymbol{\theta}) - \nabla \zeta(\boldsymbol{\theta}^\circ) \}$: on a set of probability at least $1 - e^{-x}$, it holds for any $\boldsymbol{\theta}, \boldsymbol{\theta}^\circ \in \Theta_0(\mathbf{r})$ and $\chi(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) = D^{-1} \{ \nabla L(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta}^\circ) \} + D(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)$

$$\begin{aligned} \mathbb{E}[\chi(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)] &\leq \delta(r) \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\| \leq 2r \delta(r), \\ \|\mathcal{U}(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)\| &\leq \|\mathcal{U}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\| + \|\mathcal{U}(\boldsymbol{\theta}^\circ, \boldsymbol{\theta}^*)\| \leq 2r \varrho(x), \\ \|\chi(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)\| &\leq 2 \diamond(r, x). \end{aligned} \quad (9.14)$$

9.4.2 Local quadratic approximation of the log-likelihood

As the next step, we derive a uniform deviation bound on the error of a quadratic approximation $\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) = (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top \nabla L(\boldsymbol{\theta}^\circ) - \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\|^2/2$ of $L(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)$:

$$\begin{aligned} \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) &\stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^\circ) - (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top \nabla L(\boldsymbol{\theta}^\circ) + \frac{1}{2} \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\|^2 \\ &= L(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) - \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) \end{aligned}$$

in all $\boldsymbol{\theta}, \boldsymbol{\theta}^\circ \in \Theta_0$, where Θ_0 is some vicinity of a fixed point $\boldsymbol{\theta}^*$. With $\boldsymbol{\theta}^\circ$ fixed, the gradient $\nabla \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) \stackrel{\text{def}}{=} \frac{d}{d\boldsymbol{\theta}} \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)$ fulfills

$$\nabla \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) = \nabla L(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta}^\circ) + D^2(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) = D \chi(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ);$$

cf. (9.10). This implies

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) = (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top \nabla \alpha(\boldsymbol{\theta}', \boldsymbol{\theta}^\circ),$$

where $\boldsymbol{\theta}'$ is a point on the line connecting $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\circ$. Further,

$$|\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)| = |(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top D D^{-1} \nabla \alpha(\boldsymbol{\theta}', \boldsymbol{\theta}^\circ)| \leq \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\| \sup_{\boldsymbol{\theta}' \in \Theta_0(\mathbf{r})} |\chi(\boldsymbol{\theta}', \boldsymbol{\theta}^\circ)|.$$

and one can apply (9.14). This yields the following result.

Proposition 9.4.2. Suppose **(L₀)** and **(ED₂)**. For each \mathbf{r} , it holds on a random set $\Omega(\mathbf{r}, \mathbf{x})$ of a probability at least $1 - e^{-\mathbf{x}}$ with any $\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \Theta_0(\mathbf{r})$

$$\begin{aligned}\frac{|\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*)|}{\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|} &\leq \diamond(\mathbf{r}, \mathbf{x}), \quad |\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*)| \leq \mathbf{r} \diamond(\mathbf{r}, \mathbf{x}), \\ \frac{|\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta})|}{\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|} &\leq 2\diamond(\mathbf{r}, \mathbf{x}), \quad |\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta})| \leq 2\mathbf{r} \diamond(\mathbf{r}, \mathbf{x}), \\ \frac{|\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*)|}{\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|} &\leq 2\diamond(\mathbf{r}, \mathbf{x}), \quad |\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*)| \leq 4\mathbf{r} \diamond(\mathbf{r}, \mathbf{x}),\end{aligned}$$

where $\diamond(\mathbf{r}, \mathbf{x})$ is from (9.13).

9.4.3 Proof of Theorem 9.3.1

By definition $\sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \geq 0$. So, it suffices to check that $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) < 0$ for all $\boldsymbol{\theta} \in \Theta \setminus \Theta_0(\mathbf{r}_0)$. The proof is based on the following bound: for each \mathbf{r}

$$I\!P\left(\sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r})} |\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla \zeta(\boldsymbol{\theta}^*)| \geq \sqrt{8} \nu_0 \mathfrak{z}_{\mathbb{H}}(\mathbf{x}) \omega \mathbf{r}^2\right) \leq e^{-\mathbf{x}}.$$

This bound is a special case of the result of Proposition 9.4.2 applied to the stochastic component of $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ and can be formally obtained from this result with $\delta(\mathbf{r}) = 0$. It implies by Theorem H.5.1 with $\rho = 1/2$ on a set $\Omega(\mathbf{x})$ of probability at least $1 - e^{-\mathbf{x}}$ that for all $\mathbf{r} \geq \mathbf{r}_0$ and all $\boldsymbol{\theta}$ with $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}$

$$|\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla \zeta(\boldsymbol{\theta}^*)| \leq \varrho(\mathbf{r}, \mathbf{x}) \mathbf{r}^2,$$

where

$$\varrho(\mathbf{r}, \mathbf{x}) = \varrho(\mathbf{x} + \log(2\mathbf{r}/\mathbf{r}_0)) = \sqrt{8} \nu_0 \mathfrak{z}_{\mathbb{H}}(\mathbf{x} + \log(2\mathbf{r}/\mathbf{r}_0)) \omega.$$

The use of $\nabla I\!E L(\boldsymbol{\theta}^*) = 0$ yields

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r})} |L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - I\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla L(\boldsymbol{\theta}^*)| \leq \varrho(\mathbf{r}, \mathbf{x}) \mathbf{r}^2.$$

By Theorem B.2.2, the vector $\boldsymbol{\xi} = D^{-1} \nabla \zeta(\boldsymbol{\theta}^*)$ fulfills $I\!P(\|\boldsymbol{\xi}\| \geq z(B, \mathbf{x})) \leq 2e^{-\mathbf{x}}$. We ignore here the negligible term of order $e^{-\mathbf{x}_c}$. The condition $\|\boldsymbol{\xi}\| \leq z(B, \mathbf{x})$ implies for $\mathbf{r} \geq \mathbf{r}_0$

$$\begin{aligned}\sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r})} |(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla L(\boldsymbol{\theta}^*)| \\ \leq \sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r})} \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \times \|D^{-1} \nabla L(\boldsymbol{\theta}^*)\| = \mathbf{r} \|\boldsymbol{\xi}\| \leq z(B, \mathbf{x}) \mathbf{r}.\end{aligned}$$

Condition (\mathcal{L}) implies for each $\boldsymbol{\theta}$ with $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = r > r_0$ and $\delta = \delta(r_0)$ that

$$-\mathbb{E} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) > (1 - \delta) \left(r_0 r - \frac{1}{2} r_0^2 \right) + C_1 r^2 \geq z(B, x) r + \varrho(r, x) r^2$$

provided that $r_0 \geq 2(1 - \delta)^{-1} z(B, x)$ and $C_1 \geq \varrho(r, x)$. This ensures that $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) < 0$ for all $\boldsymbol{\theta} \notin \Theta_0(r_0)$ with a dominating probability.

9.4.4 Proof of Theorem 9.3.2

Let r_0 be selected to ensure that $\mathbb{P}\{\tilde{\boldsymbol{\theta}} \notin \Theta_0(r_0)\} \leq e^{-x}$. Furthermore, the definition of $\tilde{\boldsymbol{\theta}}$ yields $\nabla L(\tilde{\boldsymbol{\theta}}) = 0$ and

$$\chi(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = -D^{-1} \nabla L(\boldsymbol{\theta}^*) + D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*).$$

Now Proposition 9.4.1 implies on a set of a dominating probability

$$\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \xi\| \leq \diamond(r_0, x) \quad (9.15)$$

and the assertion follows.

9.4.5 Proof of Theorem 9.3.3

We apply the result of Proposition 9.4.2 on a random set of dominating probability $1 - 2e^{-x}$ on which $\tilde{\boldsymbol{\theta}} \in \Theta_0(r_0)$ and the inequalities from that proposition are fulfilled. For the special case with $\boldsymbol{\theta}^\circ = \tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ we obtain in view of $\nabla L(\tilde{\boldsymbol{\theta}}) = 0$ that

$$\left| L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 / 2 \right| \leq |\alpha(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\theta}})| \leq 2r_0 \diamond(r_0, x). \quad (9.16)$$

Furthermore, with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^\circ = \boldsymbol{\theta}^*$

$$\left| L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \xi^\top D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 / 2 \right| = |\alpha(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)| \leq r_0 \diamond(r_0, x)$$

which implies

$$\left| L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\xi\|^2 / 2 + \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \xi\|^2 / 2 \right| \leq r_0 \diamond(r_0, x).$$

Now it follows by (9.15) that

$$\left| L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\xi\|^2 / 2 \right| \leq r_0 \diamond(r_0, x) + \diamond^2(r_0, x) / 2.$$

For the squared root of the excess, (9.16) implies

$$\begin{aligned}
\left| \{2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)\}^{1/2} - \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \right| &\leq \frac{|2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2|}{\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|} \\
&\leq \frac{2|\alpha(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)|}{\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|} \leq \sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)} \frac{2|\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*)|}{\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|} \leq 2 \diamond (\mathbf{r}_0, \mathbf{x}). \tag{9.17}
\end{aligned}$$

Similarly, for any $\boldsymbol{\theta}^\circ \in \Theta_0(\mathbf{r})$, it holds

$$\left| \{2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^\circ)\}^{1/2} - \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ)\| \right| \leq \sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)} \frac{2|\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)|}{\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\|} \leq 4 \diamond (\mathbf{r}_0, \mathbf{x}).$$

The Fisher expansion (9.15) allows to replace in (9.17) the norm of the standardized error $D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ with the norm of the normalized score $\boldsymbol{\xi}$. This completes the proof of Theorem 9.3.3.

Bernstein – von Mises Theorem

This chapter discusses the properties of the posterior distribution for a non-informative and a regular prior and the log-likelihood function $L(\boldsymbol{\theta})$. The Bernstein – von Mises result claims that the posterior is nearly normal with the mean $\tilde{\boldsymbol{\theta}}$ and the variance D^{-2} , where D^2 is the information matrix from condition (\mathcal{L}_0) . We refer to Kleijn and van der Vaart (2012) for a detailed historical overview around the BvM result. There is a number of papers in this direction recently appeared. We mention Ghosal et al. (2000); Ghosal and van der Vaart (2007) for a general theory in the i.i.d. case; Ghosal (1999), Ghosal (2000) for high dimensional linear models; Boucheron and Gassiat (2009), Kim (2006) for some special non-Gaussian models; Shen (2002), Bickel and Kleijn (2012), Rivoirard and Rousseau (2012), Castillo (2012), Castillo and Rousseau (2013) for a semiparametric version of the BvM result for different models; Kleijn and van der Vaart (2006), Bunke and Milhaud (1998), for the misspecified parametric case, Castillo and Rousseau (2013), among many others. A general framework for the BvM result is given in Kleijn and van der Vaart (2012) in terms of the so called stochastic LAN conditions. This condition extends the classical LAN condition and basically means a kind of quadratic expansion of the log-likelihood in a root-n vicinity of the central point. The approach applies even if the parametric assumption is misspecified, however, it requires a fixed parametric model and large samples. Extensions to nonparametric models with infinite or growing parameter dimension p exist for some special situations, see e.g. Freedman (1999) and Ghosal (1999, 2000) for linear models or Bontemps (2011) for Gaussian regression, or Castillo and Nickl (2013) for the functional Gaussian case. Though the main arguments behind BvM results are similar in all studies, the way of bounding the error terms in the BvM results are essentially different. The approach of this paper allows to get explicit upper bounds on the error of Gaussian approximation for the posterior law which apply for finite samples and admit model misspecification. In a special case of an i.i.d. sample, one can precisely control how the error terms depend on the sample size n and the dimension p and judge about the applicability of the approach when p grows with n . We also show

that the posterior mean is a very good approximation of the MLE, while the posterior variance estimates the inverse of the Fisher information matrix. Section 10.2 discusses a possible construction of credible sets based on the posterior mean and variance. Our results are stated for the non-informative prior. In the Bayesian nonparametric literature the contraction rate is heavily influenced by the prior. However, in the considered setup, the prior structure does not significantly affect the results and the main statements continue to hold for any a regular prior; see Section 10.3.

10.1 Parametric BvM Theorem

This section considers the classical situation in Bayesian calculus with a finite dimensional parametric likelihood and a non-informative prior. An extension to a regular prior is straightforward and we comment on it later in ??. We will focus on a special case of using a Gaussian prior and keep in mind an extension to the nonparametric situation.

Let $\boldsymbol{\vartheta}$ mean a random element on the parameter set Θ , by $\Pi(\boldsymbol{\theta})$ we denote a prior density. In this section we assume that $\boldsymbol{\vartheta}$ is uniformly distributed on Θ with $\Pi(\boldsymbol{\theta}) \equiv 1$. The posterior distribution of $\boldsymbol{\vartheta}$ given \mathbf{Y} is described by the product density $\exp\{L(\boldsymbol{\theta})\}$ normalized by the marginal density $p(\mathbf{Y}) = \int_{\Theta} \exp\{L(\boldsymbol{\theta})\} \Pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$. Introduce the posterior moments

$$\bar{\boldsymbol{\vartheta}} \stackrel{\text{def}}{=} \mathbb{E}(\boldsymbol{\vartheta} | \mathbf{Y}), \quad \mathfrak{S}^2 \stackrel{\text{def}}{=} \text{Cov}(\boldsymbol{\vartheta} | \mathbf{Y}) \stackrel{\text{def}}{=} \mathbb{E}\{(\boldsymbol{\vartheta} - \bar{\boldsymbol{\vartheta}})(\boldsymbol{\vartheta} - \bar{\boldsymbol{\vartheta}})^{\top} | \mathbf{Y}\}. \quad (10.1)$$

An important feature of the posterior distribution is that it can be numerically assessed. In particular, one can evaluate its moments from (10.1). If we know in addition that the posterior is nearly normal, then the posterior is completely specified. This information can be effectively used for building Bayesian credible sets with an elliptic shape; see the next section. Before stating the results, introduce some more notations. Define

$$\check{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + D^{-2} \nabla L(\boldsymbol{\theta}^*) = \boldsymbol{\theta}^* + D^{-1} \boldsymbol{\xi}.$$

The result of Theorem 9.3.2 implies the expansion of the MLE $\tilde{\boldsymbol{\theta}}$ in the form $\|D(\tilde{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}})\| \leq \diamond(r_0, \mathbf{x})$. This section presents a version of the BvM result in the considered nonasymptotic setup which claims that $\bar{\boldsymbol{\vartheta}}$ is close to $\check{\boldsymbol{\theta}}$ and thus to $\tilde{\boldsymbol{\theta}}$, \mathfrak{S}^2 is nearly equal to D^{-2} , and $D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}})$ is nearly standard normal conditionally on \mathbf{Y} . In the next result and below $\boldsymbol{\gamma}$ denotes a standard normal random vector in \mathbb{R}^p .

Theorem 10.1.1. *Suppose **(ED₀)** and **(ED₂)**, **(L₀)**, and **(I)**. Let also for any $\boldsymbol{\theta} \in \Theta$ with $r = \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \geq r_0$*

$$\mathbb{E}L(\boldsymbol{\theta}^*) - \mathbb{E}L(\boldsymbol{\theta}) \geq C_0 r_0 r - \frac{C_0}{2} r_0^2 + r^2 \varrho(r, \mathbf{x}), \quad (10.2)$$

where C_0 fulfills $1/2 < C_0 \leq 1$ and $\varrho(\mathbf{r}, \mathbf{x})$ from (9.5). Let also \mathbf{r}_0 satisfy

$$C_0 \mathbf{r}_0 \geq 2\{(p + \mathbf{x})^{1/2} + (p + \mathbf{x})^{-1/2} + \|\boldsymbol{\xi}\|\}. \quad (10.3)$$

Then it holds on a random set $\Omega(\mathbf{x})$ of probability at least $1 - 5e^{-\mathbf{x}}$

$$\mathbb{P}(\boldsymbol{\vartheta} \notin \Theta_0(\mathbf{r}_0) \mid \mathbf{Y}) \leq e^{-\mathbf{x}}.$$

Also on $\Omega(\mathbf{x})$, it holds with $\Delta_o(\mathbf{x}) = \mathbf{r}_0 \diamond (\mathbf{r}_0, \mathbf{x})$ (see (9.7))

$$\begin{aligned} \|D(\bar{\boldsymbol{\vartheta}} - \check{\boldsymbol{\theta}})\|^2 &\leq 4\Delta_o(\mathbf{x}) + 4e^{-\mathbf{x}}, \\ \|I_p - D\mathfrak{S}^2 D\|_{\text{op}} &\leq 4\Delta_o(\mathbf{x}) + 4e^{-\mathbf{x}}. \end{aligned} \quad (10.4)$$

Moreover, for any $\boldsymbol{\lambda} \in \mathbb{R}^p$ with $\|\boldsymbol{\lambda}\|^2 \leq p$, it holds on $\Omega(\mathbf{x})$

$$\left| \log \mathbb{E} \left[\exp \{ \boldsymbol{\lambda}^\top D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}}) \} \mid \mathbf{Y} \right] - \|\boldsymbol{\lambda}\|^2/2 \right| \leq 2\Delta_o(\mathbf{x}) + e^{-\mathbf{x}},$$

and for any measurable set $A \subset \mathbb{R}^p$ on $\Omega(\mathbf{x})$

$$\begin{aligned} \mathbb{P}(D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}}) \in A \mid \mathbf{Y}) &\geq \exp \{ -2\Delta_o(\mathbf{x}) - 3e^{-\mathbf{x}} \} \mathbb{P}(\boldsymbol{\gamma} \in A) - e^{-\mathbf{x}}, \\ \mathbb{P}(D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}}) \in A \mid \mathbf{Y}) &\leq \exp \{ 2\Delta_o(\mathbf{x}) + 2e^{-\mathbf{x}} \} \mathbb{P}(\boldsymbol{\gamma} \in A) + e^{-\mathbf{x}}. \end{aligned}$$

One can see that all statements of Theorem 10.1.1 require “ $\Delta_o(\mathbf{x}) = \mathbf{r}_0 \diamond (\mathbf{r}_0, \mathbf{x})$ is small”. Later we show that the results continue to hold if $\check{\boldsymbol{\theta}}$ is replaced by any efficient estimate $\hat{\boldsymbol{\theta}}$, e.g. by the MLE $\tilde{\boldsymbol{\theta}}$, satisfying the condition “ $\|D(\hat{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}})\|$ is small” with a dominating probability.

Remark 10.1.1. The BvM result is stated under essentially the same list of conditions as the frequentist results of Theorem 9.3.1 through 9.3.3. Similarly to the previous results, the normal approximation of the posterior is entirely based on the smoothness properties of the likelihood function and does not involve any asymptotic arguments like weak convergence or convergence in probability, or the Central Limit Theorem.

Remark 10.1.2. The bound (10.2) can be easily checked if $\mathbb{E}L(\boldsymbol{\theta})$ is a convex function of $\boldsymbol{\theta}$.

10.2 The use of posterior mean and variance for credible sets

This section discusses a possibility of building some Bayesian credible sets in the elliptic form motivated by the Gaussian approximation of the posterior. The BvM result ensures that the posterior can be well approximated by the normal law with the mean $\check{\boldsymbol{\theta}}$ and the covariance D^{-2} . This means that the posterior probability of the set

$$\mathcal{C}^\circ(A) = \{\boldsymbol{\theta}: D(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}) \in A\}$$

is close to $\mathbb{P}(\gamma \in A)$ up to an error term of order $\Delta_o(x)$. Unfortunately, the quantities $\check{\boldsymbol{\theta}}$ and D^2 are unknown and cannot be used for building the elliptic credible sets. A natural question is whether one can replace these values by some empirical counterparts without any substantial change of the posterior mass. An answer is given by the following result.

Theorem 10.2.1. *Let a vector $\hat{\boldsymbol{\theta}}$ and a symmetric matrix \hat{D}^2 fulfill*

$$\begin{aligned} \text{tr}(D^{-1}\hat{D}^2D^{-1} - I_p)^2 &\leq \Delta^2, & \|D^{-1}\hat{D}^2D^{-1} - I_p\|_{\text{op}} &\leq \epsilon \leq 1/2, \\ \|D(\hat{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}})\| &\leq \rho. \end{aligned}$$

Then it holds on a set $\Omega(x)$ of probability $1 - 5e^{-x}$

$$\begin{aligned} \mathbb{P}(\hat{D}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\theta}}) \in A \mid \mathbf{Y}) &\geq \exp(-2\Delta_o(x) - 3e^{-x}) \{ \mathbb{P}(\gamma \in A) - \tau \} - e^{-x}, \\ \mathbb{P}(\hat{D}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\theta}}) \in A \mid \mathbf{Y}) &\leq \exp(2\Delta_o(x) + 2e^{-x}) \{ \mathbb{P}(\gamma \in A) + \tau \} + e^{-x}, \end{aligned}$$

where

$$\tau \stackrel{\text{def}}{=} \frac{1}{2} \sqrt{(1 + \epsilon)\rho^2 + \Delta^2}.$$

Proof. With $U = \hat{D}D^{-1}$, $\boldsymbol{\eta} = D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}})$, and $\boldsymbol{\beta} = D(\hat{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}})$

$$\mathbb{P}(\hat{D}(\boldsymbol{\vartheta} - \hat{\boldsymbol{\theta}}) \in A \mid \mathbf{Y}) = \mathbb{P}(U(\boldsymbol{\eta} - \boldsymbol{\beta}) \in A \mid \mathbf{Y}) = \mathbb{P}(U(\gamma - \boldsymbol{\beta}) \in A \mid \mathbf{Y}).$$

Now the result follows from Theorem 10.1.1 and Lemma D.1.1 below.

Note that the spectral norm bound $\|D^{-1}\hat{D}^2D^{-1} - I_p\|_{\text{op}} \leq \epsilon$ obviously implies $\Delta^2 \leq p\epsilon^2$. However, in some cases this upper bound can be too rough.

We conclude that the use of the estimates $\hat{\boldsymbol{\theta}}$ and \hat{D}^2 in place of $\check{\boldsymbol{\theta}}$ and D^2 does not significantly affect the posterior mass of any set A provided that the quantities $\|D(\hat{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}})\|$ and $\text{tr}(D^{-1}\hat{D}^2D^{-1} - I_p)^2$ are small. Theorem 10.1.1 justifies the use of the posterior mean $\bar{\boldsymbol{\vartheta}}$ in place of $\check{\boldsymbol{\theta}}$. The next important question is whether the posterior covariance \mathfrak{S}^2 is a reasonable estimate of D^{-2} . Unfortunately, (10.4) only implies

$$\text{tr}(D^{-1}\mathfrak{S}^2D^{-1} - I_p)^2 \leq C_p \Delta_o^2(x).$$

This yields that the use of credible sets in the form

$$\mathcal{C}(A) = \{\boldsymbol{\theta}: \mathfrak{S}^{-1}(\boldsymbol{\theta} - \bar{\boldsymbol{\vartheta}}) \in A\}$$

is only justified if $p \Delta_o^2(\mathbf{x})$ is small. If the dimension p is fixed we only need $\Delta_o(\mathbf{x})$ small. If p is large, the use of posterior covariance requires a stronger condition “ $p \Delta_o^2(\mathbf{x})$ small”. In the regular i.i.d. case, $\Delta_o(\mathbf{x}) \asymp \sqrt{(p + \mathbf{x})^3/n}$, and $p \Delta_o^2(\mathbf{x}) \asymp (p + \mathbf{x})^4/n$.

Alternatively, one can use a plug-in estimator of the matrix D^2 . Namely, suppose that the matrix $D^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} L(\boldsymbol{\theta})$ at the point $\boldsymbol{\theta}$ is available. This is always the case when the model assumption $\mathcal{IP} \in (\mathcal{IP}_{\boldsymbol{\theta}})$ is correct. Then one can define $\widehat{D}^2 = D^2(\widehat{\boldsymbol{\theta}})$, where $\widehat{\boldsymbol{\theta}}$ is a pilot estimator of $\boldsymbol{\theta}^*$. Due to Theorem 10.1.1, the posterior mean $\bar{\boldsymbol{\theta}}$ is a natural candidate for $\widehat{\boldsymbol{\theta}}$ leading to the credible sets of the form

$$\bar{\mathcal{C}}(A) \stackrel{\text{def}}{=} \{\boldsymbol{\theta}: D(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \in A\}, \quad A \subset \mathbb{R}^p. \quad (10.5)$$

The only condition to check is that $\text{tr}(D^{-1}D^2(\boldsymbol{\theta})D^{-1} - I_p)^2$ is small for all $\boldsymbol{\theta}$ from the set Θ_0 on which the estimator $\bar{\boldsymbol{\theta}}$ concentrates with a dominating probability. The condition **(L₀)** implies $\|D^{-1}D^2(\boldsymbol{\theta})D^{-1} - I_p\| \leq \delta(\mathbf{r}_0)$ for $\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)$ that implies

$$\text{tr}(D^{-1}D^2(\boldsymbol{\theta})D^{-1} - I_p)^2 \leq p \delta^2(\mathbf{r}_0), \quad \boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0).$$

The condition (??) forces $\mathbf{r}_0^2 > p$, so $p \delta^2(\mathbf{r}_0) < \mathbf{r}_0^2 \delta^2(\mathbf{r}_0) \leq \Delta_o^2(\mathbf{x})$. One can see that the plug-in approach is well justified provided that the matrix function $D^2(\boldsymbol{\theta})$ is available, in particular, under the true model specification.

Corollary 10.2.1. *Let the conditions of Theorem 10.1.1 be fulfilled, and in addition, the matrix function $D^2(\boldsymbol{\theta})$ be known. Then Theorem 10.2.1 applies to the credible sets of the form (10.5) with $a = 1 + \delta(\mathbf{r}_0)$ and $\varepsilon^2 = p \delta^2(\mathbf{r}_0)$.*

10.3 Extension to a flat Gaussian prior

The previous results for a non-informative prior can be extended to the case of a flat prior $\Pi(d\boldsymbol{\theta})$. To be more specific we restrict ourselves to the case of a Gaussian prior. This is a prototypic situation because any smooth prior can be locally approximated by a Gaussian one. Without loss of generality the prior mean will be set to zero: $\Pi = \mathcal{N}(0, G^{-2})$ with the density $\Pi(\boldsymbol{\theta}) \propto \exp\{-\|G\boldsymbol{\theta}\|^2/2\}$ for some positive symmetric matrix G^2 . A non-informative prior can be viewed as a limiting case of a Gaussian prior as $G \rightarrow 0$. We are interested in quantifying this relation. How small should G be to ensure a reasonable Gaussian approximation of the posterior? To explain the result, we first consider the Gaussian case when $\mathcal{IP}_{\boldsymbol{\theta}} = \mathcal{N}(\boldsymbol{\theta}, D^{-2})$ and $\boldsymbol{\theta}^*$ is the true point. It is well known that in this situation the non-informative prior leads to the Gaussian posterior $\mathcal{N}(\widetilde{\boldsymbol{\theta}}, D^{-2})$, while the Gaussian prior $\Pi = \mathcal{N}(0, G^{-2})$ yields the Gaussian posterior $\mathcal{N}(\widetilde{\boldsymbol{\theta}}_G, D_G^{-2})$ with $D_G^2 = D^2 + G^2$ and $\widetilde{\boldsymbol{\theta}}_G = D_G^{-2}D^2\widetilde{\boldsymbol{\theta}}$. Therefore, the Gaussian prior Π does not

significantly affect the posterior if two Gaussian measures $\mathcal{N}(\tilde{\boldsymbol{\theta}}, D^{-2})$ and $\mathcal{N}(\tilde{\boldsymbol{\theta}}_G, D_G^{-2})$ are nearly equivalent. The corresponding condition is represented in Lemma D.1.1. It requires the values $\Delta^2 = \text{tr}(D^{-1}D_G^2D^{-1} - I_p)^2 = \text{tr}(D^{-1}G^2D^{-1})^2$ and $\|D_G(\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_G)\| = \|D_G^{-1}G^2\tilde{\boldsymbol{\theta}}\|$ to be small.

Theorem 10.3.1. Suppose the conditions of Theorem 10.1.1. Let also $\Pi = \mathcal{N}(0, G^{-2})$ be a Gaussian prior measure on \mathbb{R}^p such that with $\boldsymbol{\xi} = D^{-1}\nabla L(\boldsymbol{\theta}^*)$ and $\check{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + D^{-1}\boldsymbol{\xi}$

$$\begin{aligned} \|D^{-1}G^2D^{-1}\|_{\text{op}} &\leq \varepsilon \leq 1/2, & \text{tr}(D^{-1}G^2D^{-1})^2 &\leq \Delta^2, \\ \|D_G^{-1}G^2\boldsymbol{\theta}^*\| &\leq \rho, & \|D_G^{-1}G^2D^{-1}\boldsymbol{\xi}\| &\leq \beta_1 \end{aligned} \quad (10.6)$$

for some constants $\varepsilon, \Delta, \rho, \beta_1$. Then it holds on a set $\Omega(\mathbf{x})$ of probability $1 - 5e^{-\mathbf{x}}$

$$\begin{aligned} \mathbb{P}(D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}}) \in A \mid \mathbf{Y}) &\geq \exp(-2\Delta_o(\mathbf{x}) - 3e^{-\mathbf{x}}) \{ \mathbb{P}(\boldsymbol{\gamma} \in A) - \tau \} - e^{-\mathbf{x}}, \\ \mathbb{P}(D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}}) \in A \mid \mathbf{Y}) &\leq \exp(2\Delta_o(\mathbf{x}) + 2e^{-\mathbf{x}}) \{ \mathbb{P}(\boldsymbol{\gamma} \in A) + \tau \} + e^{-\mathbf{x}}, \end{aligned}$$

where

$$\tau \stackrel{\text{def}}{=} \frac{1}{2} \sqrt{(\rho + \beta_1)^2 + \Delta^2}.$$

Remark 10.3.1. Note that under our conditions, for $\boldsymbol{\xi} = D^{-1}\nabla L(\boldsymbol{\theta}^*)$, it holds by Theorem B.2.1 and (I) on a set $\Omega(\mathbf{x})$ of a dominating probability in view of $D \leq D_G$

$$\begin{aligned} \|D_G^{-1}G^2D^{-1}\boldsymbol{\xi}\|^2 &\leq C(\mathbf{x}) \mathbb{E} \|D_G^{-1}G^2D^{-1}\boldsymbol{\xi}\|^2 = C(\mathbf{x}) \text{tr}(D_G^{-1}G^2D^{-2}V^2D^{-2}G^2D_G^{-1}) \\ &\leq C(\mathbf{x}) \alpha^2 \text{tr}(D^{-1}G^2D^{-1})^2 \leq C(\mathbf{x}) \alpha^2 \Delta^2 \end{aligned}$$

with $C(\mathbf{x}) = (1 + \sqrt{2\mathbf{x}/p})^2$.

Similar results in an asymptotic form can be found in the literature for some special cases. A Gaussian sequence space model is studied e.g. in Bontemps (2011). Johnstone (2010) considered a very particular case of a Gaussian model with $D^2 = \sigma_n^{-2}I_p$ and a Gaussian prior $\mathcal{N}(0, \tau_n^2 I_p)$. Our condition (10.6) for validity of the BvM result translate for $p = p_n$ into $p_n(\sigma_n/\tau_n)^4 \rightarrow 0$ which coincides with the condition of Johnstone (2010) in terms of σ_n and τ_n .

10.4 Proof of Theorem 10.1.1

The whole proof is split into few important steps. Everywhere $\boldsymbol{\gamma} \sim \mathcal{N}(0, I_p)$ means a standard normal vector in \mathbb{R}^p . For a random variable η , denote

$$\mathbb{E}^\circ \eta \stackrel{\text{def}}{=} \mathbb{E}[\eta \mid \mathbf{Y}].$$

Below we apply for each \mathbf{r} the bracketing bound of Proposition 9.4.2: on a random set $\Omega(\mathbf{r}, \mathbf{x})$ of probability at least $1 - e^{-\mathbf{x}}$

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r})} |L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)| \leq \mathbf{r} \diamond (\mathbf{r}, \mathbf{x}) \quad \text{on } \Omega(\mathbf{r}, \mathbf{x}). \quad (10.7)$$

The bound from Theorem B.2.2 implies

$$\|\boldsymbol{\xi}\| \leq z(B, \mathbf{x}) \quad \text{on } \Omega(B, \mathbf{x}),$$

on a set $\Omega(B, \mathbf{x})$ of a probability at least $1 - 2e^{-\mathbf{x}}$. Obviously, the probability of the overlap $\Omega(\mathbf{r}, \mathbf{x}) \cap \Omega(B, \mathbf{x})$ is at least $1 - 3e^{-\mathbf{x}}$. Finally, we assume the radius \mathbf{r}_0 to be fixed which has to ensure the concentration of the posterior on the local set $\Theta_0(\mathbf{r}_0)$ similarly to concentration of the MLE $\tilde{\boldsymbol{\theta}}$ shown in Theorem 9.3.1.

10.4.1 Local Gaussian approximation of the posterior. Upper bound

As the first step, we study the properties of $D(\boldsymbol{\vartheta} - \tilde{\boldsymbol{\theta}})$, where

$$\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + D^{-1}\boldsymbol{\xi} = \boldsymbol{\theta}^* + D^{-2}\nabla L(\boldsymbol{\theta}^*)$$

and $\boldsymbol{\xi} = D^{-1}\nabla L(\boldsymbol{\theta}^*)$. For any nonnegative function f , it holds by (10.7)

$$\int_{\Theta_0(\mathbf{r}_0)} e^{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)} f(D(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})) d\boldsymbol{\theta} \leq e^{\Delta_0(\mathbf{x})} \int_{\Theta_0(\mathbf{r}_0)} e^{\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)} f(D(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})) d\boldsymbol{\theta}.$$

Similarly,

$$\int_{\Theta_0(\mathbf{r}_0)} e^{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)} f(D(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})) d\boldsymbol{\theta} \geq e^{-\Delta_0(\mathbf{x})} \int_{\Theta_0(\mathbf{r}_0)} e^{\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)} f(D(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})) d\boldsymbol{\theta}.$$

The main benefit of these bounds is that $\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is quadratic in $\boldsymbol{\theta}$. This enables to explicitly evaluate the posterior and to show that the posterior measure is nearly Gaussian. In what follows $\boldsymbol{\gamma}$ is a standard normal vector in \mathbb{R}^p independent of \mathbf{Y} . By Corollary B.1.2

$$\mathbb{P}(\|\boldsymbol{\gamma}\| \geq z(p, \mathbf{x})) \leq e^{-\mathbf{x}},$$

$$z^2(p, \mathbf{x}) = p + 2\sqrt{p\mathbf{x}} + 2\mathbf{x}.$$

Proposition 10.4.1. Suppose (10.7) for $\mathbf{r} = \mathbf{r}_0$ with

$$\mathbf{r}_0 \geq z(B, \mathbf{x}) + z(p, \mathbf{x}).$$

Then for any nonnegative function $f(\cdot)$ on \mathbb{R}^p , it holds on $\Omega(\mathbf{r}_0, \mathbf{x})$

$$\mathbb{E}^\circ[f(D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}})) \mathbb{I}\{\boldsymbol{\vartheta} \in \Theta_0(\mathbf{r}_0)\}] \leq \exp\{\Delta_\circ^+(\mathbf{x})\} \mathbb{E}f(\boldsymbol{\gamma}), \quad (10.8)$$

where

$$\begin{aligned} \Delta_\circ^+(\mathbf{x}) &= 2\Delta_\circ(\mathbf{x}) + \nu(\mathbf{r}_0) \leq 2\Delta_\circ(\mathbf{x}) + 2e^{-\mathbf{x}}, \\ \nu(\mathbf{r}_0) &\stackrel{\text{def}}{=} -\log \mathbb{P}^\circ(\|\boldsymbol{\gamma} + \boldsymbol{\xi}\| \leq \mathbf{r}_0) \leq 2e^{-\mathbf{x}}. \end{aligned} \quad (10.9)$$

Proof. We use that $\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \boldsymbol{\xi}^\top D(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2$ is proportional to the density of a Gaussian distribution. More precisely, define

$$m(\boldsymbol{\xi}) \stackrel{\text{def}}{=} -\|\boldsymbol{\xi}\|^2/2 + \log(\det D) - p \log(\sqrt{2\pi}).$$

Then

$$m(\boldsymbol{\xi}) + \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = -\|D(\boldsymbol{\theta} - \check{\boldsymbol{\theta}})\|^2/2 + \log(\det D) - p \log(\sqrt{2\pi}) \quad (10.10)$$

is (conditionally on \mathbf{Y}) the log-density of the normal law with the mean $\check{\boldsymbol{\theta}} = D^{-1}\boldsymbol{\xi} + \boldsymbol{\theta}^*$ and the covariance matrix D^{-2} . Change of variables $\mathbf{u} = D(\boldsymbol{\theta} - \check{\boldsymbol{\theta}})$ implies by (10.10) for any nonnegative function f that

$$\begin{aligned} &\int_{\Theta_0(\mathbf{r}_0)} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + m(\boldsymbol{\xi})\} f(D(\boldsymbol{\theta} - \check{\boldsymbol{\theta}})) d\boldsymbol{\theta} \\ &\leq e^{\Delta_\circ(\mathbf{x})} \int \exp\{\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + m(\boldsymbol{\xi})\} f(D(\boldsymbol{\theta} - \check{\boldsymbol{\theta}})) d\boldsymbol{\theta} \\ &= e^{\Delta_\circ(\mathbf{x})} \int \phi(\mathbf{u}) f(\mathbf{u}) d\mathbf{u} = e^{\Delta_\circ(\mathbf{x})} \mathbb{E}f(\boldsymbol{\gamma}). \end{aligned} \quad (10.11)$$

Similarly, for any nonnegative function f , it follows by change of variables $\mathbf{u} = D(\boldsymbol{\theta} - \check{\boldsymbol{\theta}})$ and $D(\boldsymbol{\theta} - \boldsymbol{\theta}^*) = \mathbf{u} + \boldsymbol{\xi}$ that

$$\begin{aligned} &\int \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} f(D(\boldsymbol{\theta} - \check{\boldsymbol{\theta}})) \mathbb{I}\{\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}_0\} d\boldsymbol{\theta} \\ &\geq \exp\{-\Delta_\circ(\mathbf{x}) - m(\boldsymbol{\xi})\} \int \phi(\mathbf{u}) f(\mathbf{u}) \mathbb{I}\{\|\mathbf{u} + \boldsymbol{\xi}\| \leq \mathbf{r}_0\} d\mathbf{u}. \end{aligned} \quad (10.12)$$

A special case of (10.12) with $f(\mathbf{u}) \equiv 1$ implies by definition of $\nu(\mathbf{r}_0)$:

$$\int_{\Theta_0(\mathbf{r}_0)} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} d\boldsymbol{\theta} \geq \exp\{-\Delta_\circ(\mathbf{x}) - m(\boldsymbol{\xi}) - \nu(\mathbf{r}_0)\}. \quad (10.13)$$

Further, (10.11) and (10.13) imply on $\Omega(\mathbf{r}_0, \mathbf{x})$

$$\frac{\int_{\Theta_0(\mathbf{r}_0)} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} f(D(\boldsymbol{\theta} - \check{\boldsymbol{\theta}})) d\boldsymbol{\theta}}{\int \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} d\boldsymbol{\theta}} \leq \exp\{2\Delta_\circ(\mathbf{x}) + \nu(\mathbf{r}_0)\} \mathbb{E}f(\boldsymbol{\gamma})$$

and (10.8) follows. As $\|\xi\| \leq z(B, \mathbf{x})$ on $\Omega(B, \mathbf{x})$ and $\mathbf{r}_0 \geq z(B, \mathbf{x}) + z(p, \mathbf{x})$, this and Theorem B.1.1 imply for $\boldsymbol{\gamma} \sim \mathcal{N}(0, I_p)$

$$\nu(\mathbf{r}_0) = -\log \mathbb{P}^\circ(\|\boldsymbol{\gamma} + \xi\| \leq \mathbf{r}_0) \leq -\log \mathbb{P}(\|\boldsymbol{\gamma}\| \leq z(p, \mathbf{x})) \leq 2e^{-x},$$

and the last assertion follows.

The condition “ $\Delta_o(x)$ is small” allows us to ignore the exp-factor in (10.8) and this result yields an upper bound $\mathbb{E}f(\boldsymbol{\gamma})$ for the posterior expectation of $f(D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}}))$ conditioned on \mathbf{Y} and on $\boldsymbol{\vartheta} \in \Theta_0(\mathbf{r}_0)$.

The next result considers some special cases of (10.8) with $f(\mathbf{u}) = \exp(\boldsymbol{\lambda}^\top \mathbf{u})$ and $f(\mathbf{u}) = \mathbb{I}(\mathbf{u} \in A)$ for a measurable subset $A \subset \mathbb{R}^p$.

Corollary 10.4.1. *Suppose (10.7) for $\mathbf{r} = \mathbf{r}_0$. For any $\boldsymbol{\lambda} \in \mathbb{R}^p$, it holds on $\Omega(\mathbf{r}_0, \mathbf{x})$*

$$\log \mathbb{E}^\circ[\exp\{\boldsymbol{\lambda}^\top D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}})\} \mathbb{I}\{\boldsymbol{\vartheta} \in \Theta_0(\mathbf{r}_0)\}] \leq \|\boldsymbol{\lambda}\|^2/2 + \Delta_o^+(\mathbf{x}).$$

For any measurable set A , it holds on $\Omega(\mathbf{r}_0, \mathbf{x})$

$$\mathbb{P}^\circ(D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}}) \in A) \stackrel{\text{def}}{=} \mathbb{P}(D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}}) \in A \mid \mathbf{Y}) \leq \exp\{\Delta_o^+(\mathbf{x})\} \mathbb{P}(\boldsymbol{\gamma} \in A).$$

In the next result we describe the local concentration properties of the posterior. Namely, the centered and scaled posterior vector $D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}})$ concentrates on a coronary set $\{\mathbf{u}: z_-(p, \mathbf{x}) \leq \|\mathbf{u}\| \leq z(p, \mathbf{x})\}$ for $z_-^2(p, \mathbf{x}) = p - 2\sqrt{p\mathbf{x}}$ with \mathbb{P}° -probability of order $1 - 2e^{-x}$.

Corollary 10.4.2. *For $x \geq 0$, it holds*

$$\begin{aligned} \mathbb{P}^\circ\{\|D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}})\|^2 \geq p + 2\sqrt{p\mathbf{x}} + 2x\} &\leq \exp\{-x + \Delta_o^+(\mathbf{x})\}, \\ \mathbb{P}^\circ(\|D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}})\|^2 \leq p - 2\sqrt{p\mathbf{x}}) &\leq \exp\{-x + \Delta_o^+(\mathbf{x})\}. \end{aligned}$$

Proof. This result is a combination of the bound from Corollary 10.4.1 and the bounds for the standard normal distributions from Corollary B.1.2.

10.4.2 Tail posterior probability and contraction

The next important step in our analysis is to check that $\boldsymbol{\vartheta}$ concentrates in a small vicinity $\Theta_0 = \Theta_0(\mathbf{r}_0)$ of the point $\boldsymbol{\theta}^*$ with a properly selected \mathbf{r}_0 . This will be described by using the random quantity

$$\rho(\mathbf{r}_0) \stackrel{\text{def}}{=} \frac{\int_{\Theta \setminus \Theta_0} \exp\{L(\boldsymbol{\theta})\} d\boldsymbol{\theta}}{\int_{\Theta_0} \exp\{L(\boldsymbol{\theta})\} d\boldsymbol{\theta}} = \frac{\int_{\Theta \setminus \Theta_0} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} d\boldsymbol{\theta}}{\int_{\Theta_0} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} d\boldsymbol{\theta}}.$$

Obviously $I\!\!P\{\boldsymbol{\vartheta} \notin \Theta_0(\mathbf{r}_0) \mid \mathbf{Y}\} \leq \rho(\mathbf{r}_0)$. Therefore, small values of $\rho(\mathbf{r}_0)$ indicate a small posterior probability of the set $\Theta \setminus \Theta_0$.

Proposition 10.4.2. *Suppose the conditions **(ED₀)**, **(ED₂)**, and **(10.2)**. Then with $\Delta_{\circ}^+(\mathbf{x})$ from **(10.9)**, it holds on a set $\Omega_1(\mathbf{x})$ of probability at least $1 - 4e^{-\mathbf{x}}$*

$$\rho(\mathbf{r}_0) \leq \exp\left\{-\frac{p + \mathbf{x}}{2} + \Delta_{\circ}^+(\mathbf{x})\right\}. \quad (10.14)$$

The proof only uses condition **(10.2)** and the fact that there exists a random set $\Omega(\mathbf{x})$ of probability at least $1 - e^{-\mathbf{x}}$ such that

$$|\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \boldsymbol{\xi}^\top D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)| \leq \mathbf{r}^2 \varrho(\mathbf{r}, \mathbf{x}) \quad \text{on } \Omega(\mathbf{x}) \quad (10.15)$$

for $\mathbf{r} = \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|$ and $\varrho(\mathbf{r}, \mathbf{x})$ from **(9.5)**; cf. the proof of Theorem **9.3.1**.

For the denominator of $\rho(\mathbf{r}_0)$ we apply the lower bound **(10.7)**: on $\Omega(\mathbf{x})$

$$\begin{aligned} \int_{\Theta_0(\mathbf{r}_0)} e^{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)} d\boldsymbol{\theta} &\geq e^{-\Delta_{\circ}(\mathbf{x})} \int_{\Theta_0(\mathbf{r}_0)} e^{\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)} d\boldsymbol{\theta} \\ &= e^{-\Delta_{\circ}(\mathbf{x})} \int_{\Theta_0(\mathbf{r}_0)} \exp\left\{\boldsymbol{\xi}^\top D(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2\right\} d\boldsymbol{\theta} \\ &= \frac{-e^{\Delta_{\circ}(\mathbf{x})}}{\det(D)} \int_{\|\mathbf{u}\| \leq \mathbf{r}_0} \exp\{\boldsymbol{\xi}^\top \mathbf{u} - \|\mathbf{u}\|^2/2\} d\mathbf{u}. \end{aligned} \quad (10.16)$$

It remains to bound from above the integral over the complement of the local set $\Theta_0(\mathbf{r}_0)$. The conditions **(10.2)** and **(10.15)** imply on $\Omega(\mathbf{x})$ for any $\boldsymbol{\theta}$ with $\mathbf{r} = \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| > \mathbf{r}_0$

$$L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \leq -\mathbf{r}_0 \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| + \boldsymbol{\xi}^\top D(\boldsymbol{\theta} - \boldsymbol{\theta}^*).$$

Therefore,

$$\begin{aligned} \int_{\Theta \setminus \Theta_0} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} d\boldsymbol{\theta} &\leq \int_{\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| > \mathbf{r}_0} \exp\{-\mathbf{r}_0 \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| + \boldsymbol{\xi}^\top D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\} d\boldsymbol{\theta} \\ &= \frac{1}{\det(D)} \int_{\|\mathbf{u}\| > \mathbf{r}_0} \exp\{-\mathbf{r}_0 \|\mathbf{u}\| + \boldsymbol{\xi}^\top \mathbf{u}\} d\mathbf{u} \end{aligned}$$

This and **(10.16)** together imply on $\Omega(\mathbf{x})$ by Theorem **A.2.3**

$$\rho(\mathbf{r}_0) \leq e^{\Delta_{\circ}(\mathbf{x})} \frac{\int_{\|\mathbf{u}\| > \mathbf{r}_0} e^{-\mathbf{r}_0 \|\mathbf{u}\| + \boldsymbol{\xi}^\top \mathbf{u}} d\mathbf{u}}{\int_{\|\mathbf{u}\| \leq \mathbf{r}_0} e^{-\|\mathbf{u}\|^2/2 + \boldsymbol{\xi}^\top \mathbf{u}} d\mathbf{u}} \leq \exp\left\{-\frac{p + \mathbf{x}}{2} + \Delta_{\circ}(\mathbf{x})\right\}.$$

Proposition 10.4.3. *Assume the conditions of Proposition **10.4.2**. It holds on a set $\Omega_2(\mathbf{x})$ of probability at least $1 - 4e^{-\mathbf{x}}$ for any unit vector $\mathbf{a} \in I\!\!R^p$*

$$\rho_2(\mathbf{r}_0) \stackrel{\text{def}}{=} \frac{\int_{\Theta \setminus \Theta_0} |\mathbf{a}^\top D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)|^2 e^{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)} d\boldsymbol{\theta}}{\int_{\Theta_0} e^{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)} d\boldsymbol{\theta}} \leq 2 \exp\left\{-\frac{p + \mathbf{x}}{2} + \Delta_{\circ}^+(\mathbf{x})\right\}. \quad (10.17)$$

Proof. The arguments are similar to the proof of Proposition 10.4.2 with the use of (A.4) in place of (A.3).

10.4.3 Local Gaussian approximation of the posterior. Lower bound

Now we present a local lower bound for the posterior probability. The reason for separating the upper and lower bounds is that the lower bound also requires a tail probability estimation; see (10.14) and (10.17).

Proposition 10.4.4. *Suppose (10.7) for $\mathbf{r} = \mathbf{r}_0$ and (10.14). Then for any nonnegative function $f(\cdot)$ on \mathbb{R}^p , it holds on $\Omega(\mathbf{x})$*

$$\begin{aligned} & \mathbb{E}^\circ\{f(D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}})) \mathbb{I}\{\boldsymbol{\vartheta} \in \Theta_0(\mathbf{r}_0)\}\} \\ & \geq \exp\{-\Delta_o^-(\mathbf{x})\} \mathbb{E}^\circ\left\{f(\boldsymbol{\gamma}) \mathbb{I}\{\|\boldsymbol{\gamma} + \boldsymbol{\xi}\| \leq \mathbf{r}_0\}\right\}, \end{aligned} \quad (10.18)$$

where $\Delta_o^-(\mathbf{x}) = \Delta_o^+(\mathbf{x}) + \rho(\mathbf{r}_0)$.

Proof. On the set $\Omega(\mathbf{x})$, it holds by (10.11) with $f(\cdot) = 1$:

$$\begin{aligned} \int \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} d\boldsymbol{\theta} & \leq \int_{\Theta_0} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} d\boldsymbol{\theta} + \int_{\Theta \setminus \Theta_0} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} d\boldsymbol{\theta} \\ & \leq \{1 + \rho(\mathbf{r}_0)\} \int_{\Theta_0} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} d\boldsymbol{\theta} \\ & \leq \{1 + \rho(\mathbf{r}_0)\} \exp\{\Delta_o(\mathbf{x}) - m(\boldsymbol{\xi}) + \nu(\mathbf{r}_0)\} \\ & \leq \exp\{\Delta_o(\mathbf{x}) - m(\boldsymbol{\xi}) + \nu(\mathbf{r}_0) + \rho(\mathbf{r}_0)\}. \end{aligned}$$

This and the bound (10.12) imply

$$\begin{aligned} & \frac{\int_{\Theta_0(\mathbf{r}_0)} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} f(D(\boldsymbol{\theta} - \check{\boldsymbol{\theta}})) d\boldsymbol{\theta}}{\int \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} d\boldsymbol{\theta}} \\ & \geq \frac{\exp\{-\Delta_o(\mathbf{x}) - m(\boldsymbol{\xi})\} \int \phi(\mathbf{u}) f(\mathbf{u}) \mathbb{I}\{\|\mathbf{u} + \boldsymbol{\xi}\| \leq \mathbf{r}_0\} d\mathbf{u}}{\exp\{\Delta_o(\mathbf{x}) - m(\boldsymbol{\xi}) + \nu(\mathbf{r}_0) + \rho(\mathbf{r}_0)\}} \\ & \geq \exp\{-\Delta_o^-(\mathbf{x})\} \mathbb{E}^\circ[f(\boldsymbol{\gamma}) \mathbb{I}\{\|\boldsymbol{\gamma} + \boldsymbol{\xi}\| \leq \mathbf{r}_0\}]. \end{aligned}$$

This yields (10.18).

Note that the bound $\|\boldsymbol{\xi}\| \leq z(B, \mathbf{x})$ implies for $\mathbf{r}_0 > z(B, \mathbf{x})$

$$\{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u} + \boldsymbol{\xi}\| \leq \mathbf{r}_0\} \supseteq \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}\| \leq \mathbf{r}_0 - z(B, \mathbf{x})\}.$$

As a corollary, we state the results for the distribution and moment generating functions of $D(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}})$. We assume that the value \mathbf{r}_0^2 be selected to ensure the tail probability bound (10.17).

Corollary 10.4.3. Suppose (10.15) and (10.14). Let r_0^2 ensure (10.3). On $\Omega(x) \cap \Omega(B, x)$, it holds for any $\lambda \in I\!\!R^p$ with $\|\lambda\|^2 \leq p$

$$\log I\!\!E^\circ [\exp\{\lambda^\top D(\vartheta - \check{\theta})\} \mathbb{I}\{\vartheta \in \Theta_0(r_0)\}] \geq \|\lambda\|^2/2 - \Delta_o^-(x) - 2e^{-x}. \quad (10.19)$$

Moreover, for any $A \subset I\!\!R^p$, it holds on $\Omega(x)$

$$I\!\!P^\circ(D(\vartheta - \check{\theta}) \in A) \geq \exp\{\Delta_o^-(x)\} I\!\!P(\gamma \in A) - e^{-x}.$$

Proof. The first result follows from Proposition 10.4.4. The only important additional step is an evaluation of the integral $I\!\!E\{\exp(\lambda^\top \gamma) \mathbb{I}(\|\gamma\| \leq r)\}$. The bound (A.2) yields (10.19) in view of $\log(1 - e^{-3x/2}) \geq -e^{-x}$ for $x \geq 1$. The second statement can be proved similarly to Corollary 10.4.1.

10.4.4 Moments of the posterior

Here we show that the first two moments of the posterior are pretty close to the moments of the standard normal law. The results are entirely based on our obtained statements from Corollary 10.4.1 and Proposition 10.4.4. Due to our previous results, it is convenient to decompose the r.v. $\eta = D(\vartheta - \check{\theta})$ in the form

$$\eta = \eta \mathbb{I}_{r_0} + \eta \mathbb{I}(\vartheta \notin \Theta_0(r_0)) = \eta^\circ + \eta^c.$$

The large deviation result yields that the posterior distribution of the part η^c is negligible provided a proper choice of r_0 . Below we show that η° is nearly standard normal which yields the BvM result. Define also the first two moments of η° :

$$\bar{\eta} \stackrel{\text{def}}{=} I\!\!E^\circ \eta^\circ, \quad S_o^2 \stackrel{\text{def}}{=} I\!\!E^\circ \{(\eta^\circ - \bar{\eta})(\eta^\circ - \bar{\eta})^\top\}.$$

Similarly to the proof of Corollary 10.4.1 and Proposition 10.4.4 one derives for any unit vector $u \in I\!\!R^p$

$$\exp \Delta_o^-(x) \leq I\!\!E^\circ |u^\top \eta^\circ|^2 \leq \exp \Delta_o^+(x); \quad (10.20)$$

see (10.8), (10.17), and (10.18). It suffices to show that (10.20) implies

$$\|\bar{\eta}\|^2 \leq 2\Delta_o^*(x), \quad \|S_o^2 - I_p\|_{\text{op}} \leq 2\Delta_o^*(x) \quad (10.21)$$

with $\Delta_o^*(x) = \max\{\Delta_o^+(x), \Delta_o^-(x)\} \leq 1/2$. Note now that

$$I\!\!E^\circ |u^\top \eta^\circ|^2 = u^\top S_o^2 u + |u^\top \bar{\eta}|^2.$$

Hence

$$\exp\{-\Delta_{\circ}^-(\mathbf{x})\} \leq \mathbf{u}^\top S_{\circ}^2 \mathbf{u} + |\mathbf{u}^\top \bar{\boldsymbol{\eta}}|^2 \leq \exp\{\Delta_{\circ}^+(\mathbf{x})\}. \quad (10.22)$$

In a similar way with $\mathbf{u} = \bar{\boldsymbol{\eta}}/\|\bar{\boldsymbol{\eta}}\|$ and $\boldsymbol{\gamma} \sim \mathcal{N}(0, I_p)$

$$\begin{aligned} \mathbf{u}^\top S_{\circ}^2 \mathbf{u} &= \mathbb{E}^{\circ} |\mathbf{u}^\top (\boldsymbol{\eta} - \bar{\boldsymbol{\eta}})|^2 \\ &\geq \exp\{-\Delta_{\circ}^-(\mathbf{x})\} \mathbb{E}^{\circ} |\mathbf{u}^\top (\boldsymbol{\gamma} - \bar{\boldsymbol{\eta}})|^2 = \exp\{-\Delta_{\circ}^-(\mathbf{x})\} (1 + \|\bar{\boldsymbol{\eta}}\|^2) \end{aligned}$$

yielding

$$\mathbf{u}^\top S_{\circ}^2 \mathbf{u} \geq (1 + \|\bar{\boldsymbol{\eta}}\|^2) \exp\{-\Delta_{\circ}^-(\mathbf{x})\}.$$

This inequality contradicts (10.22) if $\|\bar{\boldsymbol{\eta}}\|^2 > 2\Delta_{\circ}^*(\mathbf{x})$ for $\Delta_{\circ}^*(\mathbf{x}) \leq 1/2$, and (10.21) follows.

The bound for the first moment implies

$$\|D(\mathbb{E}^{\circ} \boldsymbol{\vartheta} - \check{\boldsymbol{\theta}})\|^2 \leq 2\Delta_{\circ}^*(\mathbf{x})$$

while the second bound yields $\|D\mathfrak{S}_{\circ}^2 D - I_p\|_{\text{op}} \leq 2\Delta_{\circ}^*(\mathbf{x})$. This completes the proof of (10.4).

10.5 Proof of Theorem 10.3.1

Define $L_G(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2/2$. The deterministic quadratic penalty $\|G\boldsymbol{\theta}\|^2/2$ does not change the stochastic component of $L(\boldsymbol{\theta})$ coincides with one of $L(\boldsymbol{\theta})$. Also it does not deteriorate the smoothness properties of the expected process $\mathbb{E}L_G(\boldsymbol{\theta})$. Consider first the deterministic part of $L_G(\boldsymbol{\theta})$ in the vicinity $\Theta_0(\mathbf{r}_0)$. It holds

$$\begin{aligned} 2\mathbb{E}L_G(\boldsymbol{\theta}) - 2\mathbb{E}L_G(\boldsymbol{\theta}^*) &= 2\mathbb{E}L(\boldsymbol{\theta}) - 2\mathbb{E}L(\boldsymbol{\theta}^*) - \|G\boldsymbol{\theta}\|^2 + \|G\boldsymbol{\theta}^*\|^2 \\ &\approx -\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 - \|G\boldsymbol{\theta}\|^2 + \|G\boldsymbol{\theta}^*\|^2. \end{aligned}$$

Define the point $\boldsymbol{\theta}_G^*$ by maximizing the latter expression w.r.t. $\boldsymbol{\theta}$:

$$\begin{aligned} D_G^2 &\stackrel{\text{def}}{=} -\nabla^2 \mathbb{E}L(\boldsymbol{\theta}^*) + G^2 = D^2 + G^2, \\ \boldsymbol{\theta}_G^* &\stackrel{\text{def}}{=} D_G^{-2} D^2 \boldsymbol{\theta}^*. \end{aligned}$$

Then it holds

$$2\mathbb{E}L_G(\boldsymbol{\theta}) - 2\mathbb{E}L_G(\boldsymbol{\theta}^*) \approx -\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\|^2,$$

where the error of approximation is the same as in the non-penalized case with $G = 0$. Note that

$$\begin{aligned}\|D_G(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*)\| &= \|D_G(I_p - D_G^{-2}D^2)\boldsymbol{\theta}^*\| = \|D_G^{-1}(D_G^2 - D^2)\boldsymbol{\theta}^*\| \\ &= \|D_G^{-1}G^2\boldsymbol{\theta}^*\| = \beta^*. \end{aligned}\quad (10.23)$$

So, a small β^* -value ensures that $\boldsymbol{\theta}_G^*$ belongs to a vicinity $\Theta_0(\mathbf{r}_0)$ of $\boldsymbol{\theta}^*$. Now one can easily see that all the conditions of Theorem 10.1.1 are fulfilled for the process $L_G(\boldsymbol{\theta})$ when $\boldsymbol{\theta}^*$ is replaced by $\boldsymbol{\theta}_G^*$ and D by D_G . The random vector $\boldsymbol{\xi}_G$ can be taken as $\boldsymbol{\xi}_G = D_G^{-1}\nabla L(\boldsymbol{\theta}^*)$ yielding

$$\check{\boldsymbol{\theta}}_G = \boldsymbol{\theta}_G^* + D_G^{-1}\boldsymbol{\xi}_G = \boldsymbol{\theta}_G^* + D_G^{-2}\nabla L(\boldsymbol{\theta}^*).$$

The result approximates the posterior $\boldsymbol{\vartheta} | \mathbf{Y}$ for the Gaussian prior Π by the normal law $\mathcal{N}(\check{\boldsymbol{\theta}}_G, D_G^{-2})$. Now the final result follows by Lemma D.1.1 if we can bound $D^{-1}D_G^2D^{-1} - I_p$ and $\|D_G(\check{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}_G)\|$. By definition

$$D^{-1}D_G^2D^{-1} - I_p = D^{-1}G^2D^{-1}.$$

Further, by (10.23)

$$\begin{aligned}\|D_G(\check{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}_G)\| &= \|D_G(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*) + D_G(D^{-2} - D_G^{-2})\nabla L(\boldsymbol{\theta}^*)\| \\ &\leq \|D_G(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*)\| + \|D_G(D^{-2} - D_G^{-2})D\boldsymbol{\xi}\| \\ &\leq \beta^* + \|D_G^{-1}(D_G^2 - D^2)D^{-1}\boldsymbol{\xi}\| = \beta^* + \|D_G^{-1}G^2D^{-1}\boldsymbol{\xi}\|. \end{aligned}$$

This yields the result by the Pinsker Lemma D.1.1.

Roughness penalty for dimension reduction

The Fisher and Wilks Theorems belong to the short list of most fascinating results in the statistical theory. In particular, the Wilks result in its simple form claims that the likelihood ratio test statistic is close in distribution to the χ_p^2 distribution as the sample size increases, where p means the parameter dimension. So, the limiting distribution of this test statistic only depends on the dimension of the parameter space whatever the parametric model is. This explains why this result is sometimes called the *Wilks phenomenon*. This paper aims at reconsidering the mentioned results from different viewpoints. One important issue is that the presented results are stated for *finite samples*. There are only few general finite-sample results in statistical inference; see [Boucheron and Massart \(2011\)](#) and references therein in context of i.i.d. modeling. The novel approach from [Spokoiny \(2012\)](#) offered a general framework for a *finite sample theory*, and the present paper makes a further step in this direction: the classical large sample results are extended to the finite sample case with *explicit and sharp* error bounds.

Another important point is a possible *model misspecification*. The classical parametric theory requires the parametric assumption to be exactly fulfilled. Any violation of the parametric specification may destroy the Fisher and Wilks results; cf. [Huber \(1967\)](#). This study admits from the very beginning that the parametric specification is probably wrong. This automatically extends the applicability of the proposed approach.

The further issue is the use of *penalization* for reducing the *model complexity*. If the parameter dimension is too large, the classical statistical results become almost intractable because the corresponding error is proportional to the dimension of parameter space. Sieve parametric approach is often used to replace the an infinite dimensional problem with a finite dimensional one; see e.g. [Shen and Wong \(1994\)](#), [Shen \(1997\)](#), [Van de Geer \(2000\)](#), [Birgé and Massart \(1998\)](#); [Barron et al. \(1999b\)](#), and references therein. Some asymptotic results for generalized regression models are available in [Fan et al. \(2001\)](#).

Another standard way of reducing the complexity of the model is by introducing some penalty in the likelihood function. In this paper we focus on quadratic-type penalization.

Roughness penalty approach provides a popular example; cf. [Green and Silverman \(1994\)](#). [Koenker et al. \(1994\)](#) explained how roughness penalty works in context of quantile regression. Tikhonov regularization and ridge regression are the other examples which are often used in linear inverse problems. It is well known that the use of a penalization in context of an inverse problem provides regularization and uncertainty reduction at the same time. Our results show that the use of penalization indeed leads to some improvement in the obtained error bounds. Namely, one can replace the original parameter dimension p by the so called *effective dimension* p_G which can be much smaller than p . Even the case of a functional parameter $\boldsymbol{\theta}$ with $p = \infty$ can be included. In this paper the penalty term is supposed to be given in advance. In general, a model selection procedure based on a proper choice of penalization is a high topic, one of the central in nonparametric statistics. We refer to [Shen \(1997\)](#), [Birgé and Massart \(1998\)](#), [van de Geer \(2002\)](#) for the general models and to [Birgé and Massart \(2001, 2007b\)](#) for Gaussian model selection where one can find an extensive overview of the vast literature on this problem.

The final issue is the *critical parameter dimension* which is measured by the effective dimension p_G . The problem of statistical inference for models with growing parameter dimension is quite involved. There are some specific issues even if a simple linear or exponential model is considered, the results from [Portnoy \(1984, 1985\)](#) requires “ p^2/n small” for asymptotic normality of the MLE. Depending on the considered problem and the model at hand, the conditions on the critical parameter dimension p may differ. For instance, [Portnoy \(1988\)](#) obtained the Fisher and Wilks results for a generalized linear model under $p^{3/2}/n \rightarrow 0$, [Mammen \(1996\)](#) established similar results for high-dimensional linear models. A general Wilks result can be stated under the condition that p^3/n is small; see e.g. [Belloni and Chernozhukov \(2009\)](#). Below we show that the conditions on the critical dimension in penalized ML estimation can be given in terms of the effective dimension p_G rather than the parameter dimension p . In particular, in the i.i.d. case, the Fisher expansion can be stated under “ p_G^2/n small” and “ p_G^3/n small” is sufficient for the Wilks result.

First we specify our set-up. Let \mathbf{Y} denote the observed data and \mathcal{IP} mean their distribution. A general parametric assumption (PA) means that \mathcal{IP} belongs to p -dimensional family ($\mathcal{IP}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$) dominated by a measure μ_0 . This family yields the log-likelihood function $L(\boldsymbol{\theta}) = L(\mathbf{Y}, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \log \frac{d\mathcal{IP}_{\boldsymbol{\theta}}}{d\mu_0}(\mathbf{Y})$. The PA can be misspecified, so, in general, $L(\boldsymbol{\theta})$ is a *quasi log-likelihood*. The classical likelihood principle suggests to estimate $\boldsymbol{\theta}$ by maximizing the function $L(\boldsymbol{\theta})$:

$$\tilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} L(\boldsymbol{\theta}). \quad (11.1)$$

If $\mathbb{P} \notin (\mathbb{P}_{\boldsymbol{\theta}})$, then the quasi MLE estimate $\tilde{\boldsymbol{\theta}}$ from (11.1) is still meaningful and it can be viewed as estimate of the value $\boldsymbol{\theta}^*$ defined by maximizing the expected value of $L(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \mathbb{E} L(\boldsymbol{\theta})$$

which is the true value in the parametric situation and can be viewed as the parameter of the best parametric fit in the general case. The classical *Fisher Theorem* claims the expansion for the MLE $\tilde{\boldsymbol{\theta}}$:

$$D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi} \xrightarrow{P} 0,$$

where $D^2 = -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}^*)$ and $\boldsymbol{\xi} \stackrel{\text{def}}{=} D^{-1} \nabla L(\boldsymbol{\theta}^*)$. Under the correct model specification, D^2 is the total Fisher information matrix and the vector $\boldsymbol{\xi}$ is centered and standardized. So, it is asymptotically standard normal under general CLT conditions.

It is well known that many important properties of the quasi MLE $\tilde{\boldsymbol{\theta}}$ like concentration or coverage probability can be described in terms of the *excess* or *quasi maximum likelihood* $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \stackrel{\text{def}}{=} L(\tilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) = \max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$, which is the difference between the maximum of the process $L(\boldsymbol{\theta})$ and its value at the “true” point $\boldsymbol{\theta}^*$. The *Wilks phenomenon* claims that the distribution of the twice excess $2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ can be approximated by $\|\boldsymbol{\xi}\|^2$ which is asymptotically χ_p^2 , where p is the dimension of the parameter space:

$$2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\boldsymbol{\xi}\|^2 \xrightarrow{P} 0, \quad \|\boldsymbol{\xi}\|^2 \xrightarrow{w} \chi_p^2.$$

This fact is very attractive and yields asymptotic confidence and concentration sets as well as the limiting critical values for the likelihood ratio tests. However, practical applications of all mentioned results are limited: they require true parametric distribution, large samples and a fixed parameter dimension.

Modern applications stimulate a further extension of the classical theory beyond the classical parametric assumptions. Spokoiny (2012) offers a general approach which appears to be very useful for such an extension. The whole approach is based on the following local bracketing result:

$$\mathbb{L}_{\epsilon}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \diamondsuit_{\epsilon} \leq L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) \leq \mathbb{L}_{\epsilon}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \diamondsuit_{\epsilon}, \quad \boldsymbol{\theta} \in \Theta_0. \quad (11.2)$$

Here $\mathbb{L}_{\epsilon}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ and $\mathbb{L}_{\epsilon}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ are quadratic in $\boldsymbol{\theta} - \boldsymbol{\theta}^*$ expressions and Θ_0 is a local vicinity of the central point $\boldsymbol{\theta}^*$. This result can be viewed as an extension of the famous Le Cam *local asymptotic normality* (LAN) condition. The LAN condition considers just one quadratic process for approximating the log-likelihood $L(\boldsymbol{\theta})$. The use of bracketing with

two different quadratic expressions allows one to keep control of the error terms $\diamond_{\epsilon}, \diamond_{\underline{\epsilon}}$ even for relatively large neighborhoods Θ_0 of $\boldsymbol{\theta}^*$ while the LAN approach is essentially restricted to a root-n vicinity of $\boldsymbol{\theta}^*$. It also allows to incorporate a large parameter dimension and a model misspecification. However, the approach from [Spokoiny \(2012\)](#) has natural limitations: the parameter dimension p cannot be too large. For instance, in the i.i.d. case, the error terms \diamond_{ϵ} and $\diamond_{\underline{\epsilon}}$ are of order $\sqrt{p^3/n}$ which destroys the Wilks result if $p > n^{1/3}$.

A standard way of overcoming this difficulty is to impose a kind of smoothness assumption on the unknown parameter value $\boldsymbol{\theta}^*$. Here we discuss one general way to deal with such smoothness assumptions using a quadratic penalization. Section 11.1 offers a new approach to studying the properties of the penalized MLE which is based on a linear approximation of the gradient of the log-likelihood process. Compared to the bracketing approach (11.2), it allows to establish a Fisher type expansion for the penalized MLE under weaker conditions on the critical dimension of the problem. Another important novelty of the approach is the systematic use of the *effective dimension* p_G in place of the original dimension p of the parameter space. Usually p_G is much smaller than p . It is even possible to treat the case of a functional parameter if the effective dimension of the parameter set remains finite. Our main results include the Fisher and Wilks expansions for the penalized MLE. In the important special case of an i.i.d. model, the error term in the Wilks expansion is of order p_G^3/n , while the Fisher expansion requires p_G^2/n small.

Also we discuss an implication of these results to the bias-variance decomposition of the squared risk of the penalized MLE. In all our results, the error terms only depend on the effective dimension p_G .

11.1 Fisher and Wilks Theorems under quadratic penalization

Let $\text{pen}(\boldsymbol{\theta})$ be a penalty function on Θ . A big value of $\text{pen}(\boldsymbol{\theta})$ corresponds to a large degree of roughness or a small amount of smoothness of $\boldsymbol{\theta}$. The underlying assumption on the model is that the true value $\boldsymbol{\theta}^*$ is smooth in the sense that $\text{pen}(\boldsymbol{\theta}^*)$ is relatively small. A penalized (quasi) MLE approach leads to maximizing the penalized log-likelihood:

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\text{argmax}} \{L(\boldsymbol{\theta}) - \text{pen}(\boldsymbol{\theta})\}.$$

Below we discuss an important special case of a quadratic penalty $\text{pen}(\boldsymbol{\theta}) = \|G\boldsymbol{\theta}\|^2/2$ for a given symmetric matrix G ; see e.g. [Green and Silverman \(1994\)](#) or [Koenker et al. \(1994\)](#) for particular examples. Denote

$$L_G(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2/2,$$

$$\tilde{\boldsymbol{\theta}}_G \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} L_G(\boldsymbol{\theta}).$$

The use of a penalty changes the target of estimation which is now defined as

$$\boldsymbol{\theta}_G^* \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} IEL_G(\boldsymbol{\theta}). \quad (11.3)$$

So, introducing a penalty leads to some estimation bias: the new target $\boldsymbol{\theta}_G^*$ may be different from $\boldsymbol{\theta}^*$. At the same time, similarly to linear modeling, the use of penalization reduces the variability of the estimate $\tilde{\boldsymbol{\theta}}_G$ and improves its concentration properties. An interesting question is the total impact and a possible gain of using the penalized procedure. A preliminary answer is that the penalty term $\|G\boldsymbol{\theta}^*\|^2$ at the true point should not be too large relative to the squared error of estimation for the penalized model. This rule is known under the name “bias-variance trade-off”.

Another important message of this study is that the use of penalization allows to reduce the parameter dimension to the *effective dimension* which characterizes the entropy of the penalized parameter space. The resulting confidence and concentration sets depend on the effective dimension rather than on the real parameter dimension and they can be much more narrow than in the non-penalized case.

The principle steps of the study are as follows. The *concentration* step claims that the penalized MLE $\tilde{\boldsymbol{\theta}}_G$ is concentrated in a local vicinity $\Theta_{0,G}(\mathbf{r}_G)$ of the point $\boldsymbol{\theta}_G^*$. It is based on the upper function method which bounds the penalized log-likelihood $L_G(\boldsymbol{\theta})$ from above by a deterministic function. Theorem 11.4.1 states that $\tilde{\boldsymbol{\theta}}_G$ belongs to the local set $\Theta_{0,G}(\mathbf{r}_G)$ with a dominating probability, and this local set can be much smaller than the similar set for the non-penalized results. As the next step, Spokoiny (2012) applied the *bracketing* approach to bound from above and from below the log-likelihood process $L(\boldsymbol{\theta})$ by two quadratic in $\boldsymbol{\theta} - \boldsymbol{\theta}^*$ expressions. Here the bracketing step is changed essentially by using a local linear approximation of the vector gradient process $\nabla L(\boldsymbol{\theta})$. This helps to get a sharper bound on the error of approximation and improve the quality of the Fisher expansion. Similarly to Spokoiny (2012), the obtained results are stated for finite samples and do not involve any asymptotic arguments. An advantage of the proposed approach is that it combines an accurate local approximation with rather rough large deviation arguments and allows one to obtain usual asymptotic statements including asymptotic normality of the penalized MLE; see Section ?? for the i.i.d. case.

11.2 Effective dimension

Let V^2 be the matrix shown in condition (E_0G) in Section 11.3. Typically $V^2 = \text{Var}\{\nabla L(\theta_G^*)\}$ and this matrix measures the local variability of the process $L_G(\cdot)$. Let also D_G^2 be a penalized information matrix defined as

$$D_G^2 = -\nabla^2 IEL_G(\theta_G^*) = D^2 + G^2$$

with $D^2 = -\nabla^2 IEL(\theta_G^*)$. One can redefine $D^2 = -\nabla^2 IEL(\theta^*)$ under condition (L_0G) below and the so called small modeling bias condition; see Section 11.6. The *effective dimension* p_G is defined as the trace of the matrix $B_G \stackrel{\text{def}}{=} D_G^{-1}V^2D_G^{-1}$:

$$p_G \stackrel{\text{def}}{=} \text{tr}(B_G). \quad (11.4)$$

Below we show that the use of penalization enables us to replace the original dimension p in our risk bounds with the effective dimension p_G which can be much smaller than p depending on relations between the matrices D^2 , V^2 , and G^2 .

In our results the value p_G will be used via another quantity $z(B_G, \mathbf{x})$ which also depends on a fixed constant \mathbf{x} and for moderate values of \mathbf{x} can be defined as

$$z(B_G, \mathbf{x}) = \sqrt{p_G} + \sqrt{2\mathbf{x}\lambda_G}, \quad (11.5)$$

where $\lambda_G \stackrel{\text{def}}{=} \lambda_{\max}(B_G)$ is the largest eigenvalue of B_G ; see (B.22) for a precise definition.

Now we present a couple of typical examples of using the quadratic penalty: blockwise penalization and estimation under a Sobolev smoothness constraint. For simplicity of presentation we assume that $V^2 = D^2 = nI_p$, while G^2 is diagonal with non-decreasing eigenvalues g_j^2 . Then $D_G^2 = D^2 + G^2 = \text{diag}\{n + g_1^2, \dots, n + g_p^2\}$. It holds that $B_G = \text{diag}\{(1 + n^{-1}g_1^2)^{-1}, \dots, (1 + n^{-1}g_p^2)^{-1}\}$, and we apply (11.4) for computing the effective dimension p_G .

Block penalization

Consider the case when G is of a simple two-block structure: $G = \text{diag}\{0, G_1\}$. Many blocks can be considered in the similar way. The first block of dimension p_0 corresponds to the unconstrained part of the parameter vector while the second block of dimension p_1 corresponds to the low energy component. An interesting question is the minimal penalization G_1 making the impact of the low energy part inessential. Assume for simplicity that $G_1 = gI_{p_1}$. Then

$$p_G = \text{tr } B_G = p_0 + p_1 / (1 + n^{-1}g^2).$$

One can see that the impact of the second block G_1 in the effective dimension is inessential if $g^2/n \gg p_1/p_0$.

Sobolev smoothness constraint

Consider the case with $D^2 = V^2 = nI_p$ and $G^2 = \text{diag}\{g_1^2, \dots, g_p^2\}$ with $g_j = Lj^\beta$ for $\beta > 1/2$. The value β is usually considered as the Sobolev smoothness parameter. It holds

$$p_G = \sum_{j=1}^p \frac{1}{1 + L^2 j^{2\beta}/n}.$$

Define also the index p_e as the largest j satisfying $L^2 j^{2\beta} \leq n$. It is straightforward to see that $\beta > 1/2$ yields $p_G \leq C(\beta, L)p_e$ for a constant $C(\beta, L)$ depending on β, L only.

Linear inverse problem

The next example corresponds to the case of a linear inverse problem. Assume for simplicity of notation the sequence space representation, the noise is inhomogeneous with increasing eigenvalues $V^2 = \text{diag}\{v_1^2, \dots, v_p^2\}$ and the information matrix D^2 is proportional to identity, that is, $D^2 = nI_p$. Then the effective dimension is given by the sum

$$p_G = \sum_{j=1}^p \frac{v_j^2}{n + g_j^2}.$$

To keep the effective dimension small, one has to compensate the increase of the eigenvalues v_j^2 by the penalization g_j^2 .

11.3 Conditions

This section presents the list of conditions which are similar to ones from the non-penalized case in [Spokoiny \(2012\)](#). However, the use of penalization leads to some change in each condition. Most important fact is that the use of penalization helps to state the large deviation (LD) result for much smaller local neighborhoods than in the non-penalized case. [Spokoiny \(2012\)](#) presented the LD result for local sets of the form $\Theta_0(r) = \{\boldsymbol{\theta} : \|V(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq r\}$ with a proper $r \asymp p^{1/2}$. Now we redefine this set by using D_G^2 in place of V^2 and $\boldsymbol{\theta}_G^*$ in place of $\boldsymbol{\theta}^*$:

$$\Theta_{0,G}(r) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} : \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\| \leq r\}.$$

Moreover, the radius r can be selected of order $p_G^{1/2}$, which can be very useful for large or infinite p . Our conditions mainly assume some regularity and smoothness of

the penalized log-likelihood process $L_G(\boldsymbol{\theta})$. The first condition states some smoothness properties of the expected log-likelihood $\mathbb{E}L_G(\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ in a vicinity $\Theta_{0,G}(\mathbf{r})$ of $\boldsymbol{\theta}_G^*$. More precisely, it effectively means that the expected log-likelihood $\mathbb{E}L(\boldsymbol{\theta})$ is twice continuously differentiable on the local set $\Theta_{0,G}(\mathbf{r})$.

Below each condition is given in penalized and non-penalized form for the sake of comparison. Already now it is worth saying that the use of penalization helps to relax most of conditions. Define

$$\mathbb{F}_G(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E}L_G(\boldsymbol{\theta}) = -\nabla^2 \mathbb{E}L(\boldsymbol{\theta}) + G^2.$$

Then $D_G^2 = \mathbb{F}_G(\boldsymbol{\theta}_G^*)$. The conditions involve a radius \mathbf{r}_G which separates the local zone and the zone of large deviations. This value will be made precise in Theorem 11.4.1.

First we consider the stochastic component of the log-likelihood process $L_G(\boldsymbol{\theta})$ which is the same as in the non-penalized case:

$$\zeta(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L_G(\boldsymbol{\theta}) - \mathbb{E}L_G(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathbb{E}L(\boldsymbol{\theta}).$$

We assume that it is twice differentiable and denote by $\nabla\zeta(\boldsymbol{\theta})$ its gradient and by $\nabla^2\zeta(\boldsymbol{\theta})$ its Hessian matrix. The next two conditions are to ensure that the random vector $\nabla\zeta(\boldsymbol{\theta}_G^*)$ and the random processes $\nabla^2\zeta(\boldsymbol{\theta})$ are stochastically bounded with exponential moments. The conditions involve a $p \times p$ -matrix V which normalizes the vector $\nabla\zeta(\boldsymbol{\theta}_G^*)$, and a similar matrix V_2 normalizing $\nabla^2\zeta(\boldsymbol{\theta})$.

(E₀G) *There exist a positively semi-definite symmetric matrix V^2 , and constants $g > 0$, $\nu_0 \geq 1$ such that $\text{Var}\{\nabla\zeta(\boldsymbol{\theta}_G^*)\} \leq V^2$ and*

$$\sup_{\gamma \in \mathbb{R}^p} \log \mathbb{E} \exp \left\{ \lambda \frac{\gamma^\top \nabla\zeta(\boldsymbol{\theta}_G^*)}{\|V\gamma\|} \right\} \leq \frac{\nu_0^2 \lambda^2}{2}, \quad |\lambda| \leq g.$$

(E₂G) *There exist a positively semi-definite symmetric matrix V_2^2 and a value $\omega > 0$ such that it holds for any $\boldsymbol{\theta} \in \Theta_{0,G}(\mathbf{r}_0)$:*

$$\sup_{\gamma_1, \gamma_2 \in \mathbb{R}^p} \log \mathbb{E} \exp \left\{ \frac{\lambda}{\omega} \frac{\gamma_1^\top \nabla^2\zeta(\boldsymbol{\theta}) \gamma_2}{\|V_2\gamma_1\| \cdot \|V_2\gamma_2\|} \right\} \leq \frac{\nu_0^2 \lambda^2}{2}, \quad |\lambda| \leq g.$$

Below we only need that the constant $g(\mathbf{r})$ is larger than Cp_G for a fixed constant C . This allows to reduce the condition to the case with a fixed g which does not depend on the distance \mathbf{r} .

Their non-penalized versions are almost identical: one has to replace $\boldsymbol{\theta}_G^*$ with $\boldsymbol{\theta}^*$ and $\Theta_{0,G}(\mathbf{r})$ with $\Theta_0(\mathbf{r})$.

$$\begin{aligned} (\mathbf{E}_0) \quad & \sup_{\gamma \in \mathbb{R}^p} \log \mathbb{E} \exp \left\{ \lambda \frac{\gamma^\top \nabla \zeta(\boldsymbol{\theta}^*)}{\|V\gamma\|} \right\} \leq \frac{\nu_0^2 \lambda^2}{2}, \quad |\lambda| \leq g. \\ (\mathbf{E}_2) \quad & \sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)} \sup_{\gamma_1, \gamma_2 \in \mathbb{R}^p} \log \mathbb{E} \exp \left\{ \frac{\lambda}{\omega} \frac{\gamma_1^\top \nabla^2 \zeta(\boldsymbol{\theta}) \gamma_2}{\|V_2 \gamma_1\| \cdot \|V_2 \gamma_2\|} \right\} \leq \frac{\nu_0^2 \lambda^2}{2}, \quad |\lambda| \leq g(\mathbf{r}). \end{aligned}$$

The conditions (\mathbf{E}_0) and $(\mathbf{E}_0\mathbf{G})$ are very similar while $(\mathbf{E}_2\mathbf{G})$ is restricted to the vicinity $\Theta_{0,G}(\mathbf{r})$ which can be much smaller than $\Theta_0(\mathbf{r})$.

The *identifiability condition* relates the matrices V^2 and V_2^2 and to D_G^2 .

$(\mathcal{I}\mathbf{G})$ There is a constant $a_G > 0$ such that

$$a_G^2 D_G^2 \geq V^2, \quad a_G^2 D_G^2 \geq V_2^2.$$

In the non-penalized case of Chapter 9, this condition reads as

$$(\mathcal{I}) \quad a^2 D^2 \geq V^2 \text{ with } D^2 = -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}^*).$$

Therefore, the use of regularization helps to improve the identifiability in the regularized problem relative to the non-penalized one as $D^2 \leq D_G^2$.

Finally, we consider the expected log-likelihood $\mathbb{E} L_G(\boldsymbol{\theta})$. The local condition requires that it is nearly quadratic in the vicinity $\Theta_{0,G}(\mathbf{r}_G)$ of $\boldsymbol{\theta}_G^*$ while the global condition assumes a linear growth in the complement of this vicinity. Here and below $\|A\|_{\text{op}}$ means the operator norm of a matrix A .

$(\mathcal{L}_0\mathbf{G})$ For each $\mathbf{r} \leq \mathbf{r}_G$, there is a constant $\delta_G(\mathbf{r}) \leq 1/2$ such that

$$\|D_G^{-1} \mathbb{F}_G(\boldsymbol{\theta}) D_G^{-1} - I_p\|_{\text{op}} \leq \delta_G(\mathbf{r}), \quad \boldsymbol{\theta} \in \Theta_{0,G}(\mathbf{r}). \quad (11.6)$$

Under condition $(\mathcal{L}_0\mathbf{G})$, it follows from the second order Taylor expansion at $\boldsymbol{\theta}_G^*$:

$$|-2\mathbb{E} L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*) - \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\|^2| \leq \delta_G(\mathbf{r}) \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\|^2, \quad \boldsymbol{\theta} \in \Theta_{0,G}(\mathbf{r}).$$

A non-penalized version of (11.6) claims a similar approximation of $\mathbb{F}(\boldsymbol{\theta}) = -\nabla^2 \mathbb{E} L(\boldsymbol{\theta})$ by $D^2 \stackrel{\text{def}}{=} \mathbb{F}(\boldsymbol{\theta}^*)$ in the vicinity $\Theta_0(\mathbf{r}_0)$ centered at $\boldsymbol{\theta}^*$ instead of $\boldsymbol{\theta}_G^*$:

$$(\mathcal{L}_0) \quad \|D^{-1} \mathbb{F}(\boldsymbol{\theta}) D^{-1} - I_p\|_{\text{op}} \leq \delta(\mathbf{r}_0), \quad \boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0) = \{\boldsymbol{\theta}: \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}_0\}.$$

As the quadratic penalty $\|G\boldsymbol{\theta}\|^2$ does not change the smoothness properties of the expected contrast $\mathbb{E} L_G(\boldsymbol{\theta})$, the conditions $(\mathcal{L}_0\mathbf{G})$ and (\mathcal{L}_0) are essentially equivalent provided that the points $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_G^*$ are not too far from each others.

The local condition $(\mathcal{L}_0\mathbf{G})$ describes the behavior of $\mathbb{E} L_G(\boldsymbol{\theta})$ within $\Theta_{0,G}(\mathbf{r}_G)$. In particular, $\mathbb{E} L_G(\boldsymbol{\theta}_G^*) - \mathbb{E} L_G(\boldsymbol{\theta}) \approx \mathbf{r}_G^2/2$ on the boundary of this local set. The global condition means that $\mathbb{E} L_G(\boldsymbol{\theta}_G^*) - \mathbb{E} L_G(\boldsymbol{\theta})$ can be lower bounded by a linear function on the complement of this set.

($\mathcal{L}G$) For each $\boldsymbol{\theta}$ with $\mathbf{r} = \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\| \geq \mathbf{r}_G$

$$\mathbb{E}L_G(\boldsymbol{\theta}_G^*) - \mathbb{E}L_G(\boldsymbol{\theta}) \geq \{1 - \delta(\mathbf{r}_G)\} \left(\mathbf{r}_G \mathbf{r} - \frac{\mathbf{r}_G^2}{2} \right) + C_1 \mathbf{r}^2, \quad (11.7)$$

for a small constant C_1 ; see Theorem 11.4.1 below for a precise bound.

A non-penalized version of this condition is obtained by letting $G^2 = 0$.

(\mathcal{L}) $\mathbb{E}L(\boldsymbol{\theta}^*) - \mathbb{E}L(\boldsymbol{\theta}) \geq \{1 - \delta(\mathbf{r}_0)\} (\mathbf{r}_0 \mathbf{r} - \mathbf{r}_0^2/2) + C_1 \mathbf{r}^2$ for $\mathbf{r} = \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|$.

Remark 11.3.1. Conditions (\mathcal{L}_0G) and ($\mathcal{L}G$) can be effectively checked if the function $f(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\mathbb{E}L_G(\boldsymbol{\theta})$ is smooth and convex in $\boldsymbol{\theta}$. Continuity of the second derivative $\nabla^2 f(\boldsymbol{\theta})$ in $\Theta_{0,G}(\mathbf{r}_G)$ implies (\mathcal{L}_0G). Convexity of f implies for any $\boldsymbol{\theta}^\circ = \boldsymbol{\theta}_G^* + \rho(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)$ with $\rho = \mathbf{r}_G/\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\| \leq 1$

$$f(\boldsymbol{\theta}) \geq f(\boldsymbol{\theta}^\circ) + (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top \nabla f(\boldsymbol{\theta}^\circ) \geq f(\boldsymbol{\theta}^\circ) + \{D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\}^\top D_G^{-1} \nabla f(\boldsymbol{\theta}^\circ).$$

Condition (\mathcal{L}_0G) implies in view of $\|D_G(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}_G^*)\| = \mathbf{r}_G$ that $f(\boldsymbol{\theta}^\circ) \geq (1 - \delta)\mathbf{r}_G^2/2$ and $\nabla^2 f(\boldsymbol{\theta}^\circ) \geq (1 - \delta)D_G^2$ for $\delta = \delta(\mathbf{r}_G)$. As $\|D_G(\boldsymbol{\theta}^\circ - \boldsymbol{\theta})\| = \mathbf{r} - \mathbf{r}_G$, we conclude that

$$f(\boldsymbol{\theta}) \geq (1 - \delta) \left\{ \frac{\mathbf{r}_G^2}{2} + \mathbf{r}_G(\mathbf{r} - \mathbf{r}_G) \right\}$$

and (11.7) follows for $C_1 = 0$. The case $C_1 > 0$ can be similarly checked under a strong convexity of $-\mathbb{E}L_G(\boldsymbol{\theta})$. In the case of linear or generalized linear models, one can use $C_1 = 0$, in regular situations, it suffices that ($\mathcal{L}G$) holds with C_1 of order $n^{-1/2}$.

11.4 Concentration and a large deviation bound

This section demonstrates that the use of the penalty term helps to strengthen the concentration properties of the penalized quasi maximum likelihood estimator (qMLE) $\tilde{\boldsymbol{\theta}}_G$. Namely, we show that $\tilde{\boldsymbol{\theta}}_G$ belongs with a dominating probability to a set $\Theta_{0,G}(\mathbf{r}_G)$ which can be much smaller than a similar set from the non-penalized case; see Remark 11.4.1.

All our results involve a value \mathbf{x} . We say that a generic random set $\Omega(\mathbf{x})$ is of a *dominating probability* if $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - Ce^{-\mathbf{x}}$ for a fixed constant C like 1 or 2. We also use two growing functions $z(B_G, \mathbf{x})$ and $\mathfrak{z}_{\mathbb{H}}(\mathbf{x})$ of the argument \mathbf{x} . The functions $z(B_G, \mathbf{x})$ already mentioned in (11.5) and it describes the quantiles of the norm of the normalized score vector ξ_G ; see (11.10) below. The formal definition of $z(B_G, \mathbf{x})$ is given in (B.22). The function $\mathfrak{z}_{\mathbb{H}}(\mathbf{x})$ is related to the penalized entropy of the parameter space and it is given by (H.3). In typical situations one can use the upper bounds $z^2(B_G, \mathbf{x}) \leq C(p_G + \mathbf{x})$ and $\mathfrak{z}_{\mathbb{H}}^2(\mathbf{x}) \leq C(p_G + \mathbf{x})$ for both functions.

Theorem 11.4.1. Let (E_0G) , (E_2G) , $(\mathcal{I}G)$, (\mathcal{L}_0G) , and $(\mathcal{L}G)$ hold with

$$\{1 - \delta(r_G)\}r_G \geq 2z(B_G, x), \quad (11.8)$$

where $z(B_G, x)$ is from (B.22), and let the constant C_1 in $(\mathcal{L}G)$ satisfy

$$C_1 \geq \sup_{r > r_G} \varrho_G(r, x), \quad \text{with } \varrho_G(r, x) \stackrel{\text{def}}{=} \sqrt{8} \nu_0 \alpha_G \mathfrak{z}_{\mathbb{H}}(x + \log(2r/r_G)) \omega$$

with the function $\mathfrak{z}_{\mathbb{H}}(x)$ given by (H.3). Then

$$\mathbb{P}(\tilde{\boldsymbol{\theta}}_G \notin \Theta_{0,G}(r_G)) \leq 3e^{-x}. \quad (11.9)$$

Remark 11.4.1. This result explains a proper r_G ensuring (11.9). In the non-penalized case of Spokoiny (2012), a similar condition reads as $r_0 \geq C(\sqrt{p} + \sqrt{2x})$, so the use of penalization helps to improve the concentration properties of the penalized MLE. We conclude that the use of penalization leads to weaker conditions and to a stronger concentration property. The only problem is that the corresponding estimate $\tilde{\boldsymbol{\theta}}_G$ concentrates around $\boldsymbol{\theta}_G^*$ instead of $\boldsymbol{\theta}^*$. This can yield a bias effect; see Section 11.6 below.

Proof. By definition $\sup_{\boldsymbol{\theta} \in \Theta_{0,G}(r_G)} L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*) \geq 0$. So, it suffices to check that $L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*) < 0$ for all $\boldsymbol{\theta} \in \Theta \setminus \Theta_{0,G}(r_G)$. The proof is based on the following bound: for each r

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta} \in \Theta_{0,G}(r)} |\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)^\top \nabla \zeta(\boldsymbol{\theta}_G^*)| \geq \sqrt{8} \nu_0 \alpha_G \mathfrak{z}_{\mathbb{H}}(x) \omega r^2\right) \leq e^{-x}.$$

This bound follows from Theorem H.11.1; see (11.20) for more details. It implies by Theorem H.5.1 with $\rho = 1/2$ on a set of dominating probability at least $1 - e^{-x}$ that for all $r \geq r_G$ and all $\boldsymbol{\theta}$ with $\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\| \leq r$

$$|\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)^\top \nabla \zeta(\boldsymbol{\theta}_G^*)| \leq \varrho_G(r, x) r^2,$$

where $\varrho_G(r, x) = \nu_0 \alpha_G \mathfrak{z}_{\mathbb{H}}(x + \log(2r/r_G)) \omega$. The use of $\nabla \mathbb{E} L_G(\boldsymbol{\theta}_G^*) = 0$ yields

$$\sup_{\boldsymbol{\theta} \in \Theta_{0,G}(r)} |L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*) - \mathbb{E} L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)^\top \nabla L_G(\boldsymbol{\theta}_G^*)| \leq \varrho_G(r, x) r^2.$$

Also the vector $\boldsymbol{\xi}_G = D_G^{-1} \nabla L_G(\boldsymbol{\theta}_G^*) = D_G^{-1} \nabla \zeta(\boldsymbol{\theta}_G^*)$ can be bounded with a dominating probability: by Theorem B.2.1 $\mathbb{P}(\|\boldsymbol{\xi}_G\| \geq z(B_G, x)) \leq 2e^{-x}$. We ignore here the negligible term Ce^{-x_c} . The condition $\|\boldsymbol{\xi}_G\| \leq z(B_G, x)$ implies for each $r \geq r_G$

$$\begin{aligned} \sup_{\boldsymbol{\theta} \in \Theta_{0,G}(r)} & |(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)^\top \nabla L_G(\boldsymbol{\theta}_G^*)| \\ & \leq \sup_{\boldsymbol{\theta} \in \Theta_{0,G}(r)} \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\| \times \|D_G^{-1} \nabla \zeta(\boldsymbol{\theta}_G^*)\| = r \|\boldsymbol{\xi}_G\| \leq z(B_G, x) r. \end{aligned}$$

Condition **(LG)** implies for each $\boldsymbol{\theta}$ with $\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\| = \mathbf{r} > \mathbf{r}_0$ that

$$-\mathbb{E}L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*) > (1 - \delta) \left(\mathbf{r}_0 \mathbf{r} - \frac{1}{2} \mathbf{r}_0^2 \right) + C_1 \mathbf{r}^2 \geq z(B_G, \mathbf{x}) \mathbf{r} + \varrho_G(\mathbf{r}, \mathbf{x}) \mathbf{r}^2$$

provided that $\mathbf{r}_0 \geq 2(1 - \delta)^{-1}z(B_G, \mathbf{x})$ and $C_1 \geq \varrho_G(\mathbf{r}, \mathbf{x})$. This ensures that $L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*) < 0$ for all $\boldsymbol{\theta} \notin \Theta_{0,G}(\mathbf{r}_G)$ with a dominating probability.

11.5 Wilks and Fisher expansions

This section collects the main results of the paper. Let $\boldsymbol{\theta}_G^*$ be the point of concentration from (11.3) and let $\zeta(\boldsymbol{\theta}) = L_G(\boldsymbol{\theta}) - \mathbb{E}L_G(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathbb{E}L(\boldsymbol{\theta})$. Define a random p -vector

$$\boldsymbol{\xi}_G \stackrel{\text{def}}{=} D_G^{-1} \nabla \zeta(\boldsymbol{\theta}_G^*) = D_G^{-1} \{ \nabla L(\boldsymbol{\theta}_G^*) - G^2 \boldsymbol{\theta}_G^* \}. \quad (11.10)$$

Theorem 11.5.1. *Suppose that \mathbf{r}_G is selected to ensure (11.8). Suppose also that the conditions **(E₀G)**, **(E₂G)**, **(IG)** hold. On a random set $\Omega(\mathbf{x})$ of a dominating probability at least $1 - 4e^{-\mathbf{x}}$, it holds*

$$\|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*) - \boldsymbol{\xi}_G\| \leq \diamondsuit_G(\mathbf{x}), \quad (11.11)$$

where $\diamondsuit_G(\mathbf{x})$ is given by

$$\diamondsuit_G(\mathbf{x}) \stackrel{\text{def}}{=} \{ \delta_G(\mathbf{r}_G) + \sqrt{8} \nu_0 \mathfrak{a}_G \mathfrak{z}_{\mathbb{H}}(\mathbf{x}) \omega \} \mathbf{r}_G \quad (11.12)$$

for $\mathfrak{z}_{\mathbb{H}}(\mathbf{x})$ given by (H.3).

The proof of this and the next result is based on a linear expansion of the gradient $\nabla L_G(\boldsymbol{\theta})$ and will be given in Section 11.7.

Now we present a result on the excess $L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*) = L_G(\tilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}_G^*)$. The classical Wilks result claims that the twice excess is nearly χ_p^2 . Our result describes the quality of its approximation by a quadratic form $\|\boldsymbol{\xi}_G\|^2$.

Theorem 11.5.2. *Suppose that **(L₀G)**, **(E₀G)**, and **(E₂G)** hold. Suppose also that \mathbf{r}_G is selected to ensure (11.8). On a random set $\Omega(\mathbf{x})$ of a dominating probability at least $1 - 5e^{-\mathbf{x}}$, it holds with $\diamondsuit_G(\mathbf{x})$ from (11.12)*

$$|2L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*) - \|\boldsymbol{\xi}_G\|^2| \leq 2\mathbf{r}_G \diamondsuit_G(\mathbf{x}) + \diamondsuit_G^2(\mathbf{x}), \quad (11.13)$$

$$\left| \sqrt{2L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*)} - \|\boldsymbol{\xi}_G\| \right| \leq 3 \diamondsuit_G(\mathbf{x}). \quad (11.14)$$

One can see that the Fisher expansion (11.11) and the square root Wilks expansion (11.14) require $\diamondsuit_G(\mathbf{x})$ small, while the standard Wilks expansion (11.13) is accurate if

$r_G \diamondsuit_G(x)$ is small. This makes some difference if the parameter dimension is large. Below we address this question for the important special case of an i.i.d. likelihood.

The classical Fisher and Wilks results include some statements about the limiting behavior of the vector ξ_G and of the quadratic form $\|\xi_G\|^2$. In the i.i.d. case, one can easily show that the vector ξ_G is asymptotically standard normal as $n \rightarrow \infty$; see Section 13.4 below. However, it is well known that the convergence of $\|\xi_G\|^2$ to the χ^2 -distribution is quite slow even in the case of a fixed dimension p . For finite sample inference, we recommend to combine the approximations (11.11) to (11.14) with any resampling technique which mimics the specific behavior of the quadratic form $\|\xi_G\|^2$; see Spokoiny and Zhilova (2015).

11.6 Quadratic risk bound and modeling bias

This section demonstrates the applicability of the obtained general results to bounding the quadratic risk of estimation. For the penalized MLE $\tilde{\theta}_G$ of the parameter θ , consider the quadratic loss of estimation $\|W(\tilde{\theta}_G - \theta^*)\|^2$ for a given non-negative symmetric matrix W . A special case includes the usual quadratic loss $\|\tilde{\theta}_G - \theta^*\|^2$. Here the point $\theta^* \in \Theta$ is a proxy for the true parameter value which describes the best parametric fit of the true measure \mathbb{P} by the family (\mathbb{P}_θ) :

$$\theta^* \stackrel{\text{def}}{=} \operatorname{argmax}_{\theta \in \Theta} \mathbb{E} L(\theta).$$

The use of penalization $\|G\theta\|^2/2$ introduces some estimation bias: the penalized MLE $\tilde{\theta}_G$ estimates θ_G^* from (11.3) rather than θ^* . The value $\|W(\theta^* - \tilde{\theta}_G^*)\|^2$ is called the modeling bias and it describes the modeling error caused by using the penalization. The variance term $\|W(\tilde{\theta}_G - \theta_G^*)\|^2$ describes the error *within the penalized model*, and it can be studied with the help of the Fisher expansion of Theorem 11.5.1: $\|D_G(\tilde{\theta}_G - \theta_G^*) - \xi_G\| \leq \diamondsuit_G(x)$ on a set $\Omega(x)$ of dominating probability for $\xi_G = D_G^{-1}\nabla\zeta(\theta_G^*)$. This yields the following result on $\Omega(x)$:

$$\|D_G(\tilde{\theta}_G - \theta^* - b_G) - \xi_G\| \leq \diamondsuit_G(x)$$

with the *bias* $b_G = \theta_G^* - \theta^*$. For any positive symmetric $p \times p$ matrix W satisfying $W^2 \leq D_G^2$, it implies the probability bound for the squared loss

$$\|W(\tilde{\theta}_G - \theta^*)\| = \|Wb_G + WD_G^{-1}\xi_G\| \pm \diamondsuit_G(x).$$

One can see that analysis of the quadratic risk of the penalized MLE $\tilde{\theta}_G$ can be reduced to the analysis of $\|Wb_G + WD_G^{-1}\xi_G\|^2$. Now we consider an implication of this bound

to the squared risk $\mathbb{E}\|W(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|^2$. The use of the identity $\mathbb{E}\nabla\zeta(\boldsymbol{\theta}_G^*) = 0$ and $\text{Var}(\nabla\zeta(\boldsymbol{\theta}_G^*)) \leq V^2$ yields

$$\begin{aligned}\mathbb{E}\|Wb_G + WD_G^{-1}\xi_G\|^2 &= \|Wb_G\|^2 + \mathbb{E}\|WD_G^{-2}\nabla\zeta(\boldsymbol{\theta}_G^*)\|^2 \\ &= \|Wb_G\|^2 + \text{tr}(WD_G^{-2}\text{Var}\{\nabla\zeta(\boldsymbol{\theta}_G^*)\}D_G^{-2}W) \\ &\leq \|Wb_G\|^2 + \text{tr}(WD_G^{-2}V^2D_G^{-2}W).\end{aligned}$$

Denote $\mathcal{X}_G \stackrel{\text{def}}{=} \text{tr}(WD_G^{-2}V^2D_G^{-2}W)$ and

$$\mathcal{R}_G \stackrel{\text{def}}{=} \|Wb_G\|^2 + \mathcal{X}_G = \|Wb_G\|^2 + \text{tr}(WD_G^{-2}V^2D_G^{-2}W). \quad (11.15)$$

Theorem 11.6.1. *Let (E_0G) , (E_2G) , (L_0G) , (I_0G) , and (L_2G) hold. If $W^2 \leq D_G^2$, then it holds with \mathcal{R}_G from (11.15)*

$$\mathbb{E}\|W(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|^2 \leq \{\mathcal{R}_G^{1/2} + \diamondsuit_G^*\}^2, \quad (11.16)$$

where

$$\diamondsuit_G^* = 4\left\{\delta_G(\mathbf{r}_G)\mathbf{r}_G + 2\nu_0\mathbf{a}_G\mathbf{r}_G(\mathbb{H}_1 + \mathbb{H}_2/g + 4)\omega\right\}.$$

Remark 11.6.1. If the error term \diamondsuit_G^* in (11.16) is relatively small, this result implies $\mathbb{E}\|W(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|^2 \approx \mathcal{R}_G = \|Wb_G\|^2 + \mathcal{X}_G$. This is the usual decomposition of the quadratic risk in term of the squared bias $\|W(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*)\|^2$ and the variance term \mathcal{X}_G . The condition “ $\|Wb_G\|^2/\mathcal{X}_G$ is small” yields $\mathcal{R}_G \approx \mathcal{X}_G$. This condition can be naturally called the *small modeling bias* (SMB) condition, often it is referred to as *undersmoothing*. The bias-variance trade-off corresponds to the situation with $\|Wb_G\|^2 \asymp \mathcal{X}_G$. *Oversmoothing* means that the bias terms $\|Wb_G\|^2$ dominates.

Remark 11.6.2. As already mentioned, the result (11.16) is informative if the remainder \diamondsuit_G^* is relatively small and can be ignored. For the special case $W^2 = D_G^2$, it holds $\mathcal{X}_G = p_G \asymp r_G^2$. In the i.i.d. situation (see Section 13.4 below)

$$r_G^{-1}\diamondsuit_G^* \leq C\sqrt{p_G/n}$$

which yields a sharp risk bound $\mathbb{E}\|W(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|^2 = \mathcal{R}_G(1 + o(1))$ under “ p_G/n small”.

Remark 11.6.3. The bias induced by penalization can be measured in terms of the value $\|G\boldsymbol{\theta}^*\|^2$. To be more precise, consider the case with $W^2 = D^2$, where $D^2 = -\nabla^2\mathbb{E}L(\boldsymbol{\theta}^*)$ is the non-penalized Fisher information matrix. The definition of $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_G^*$ implies

$$\mathbb{E}L(\boldsymbol{\theta}^*) - \|G\boldsymbol{\theta}^*\|^2/2 \leq \mathbb{E}L(\boldsymbol{\theta}_G^*) - \|G\boldsymbol{\theta}_G^*\|^2/2 \leq \mathbb{E}L(\boldsymbol{\theta}_G^*).$$

Condition $(\mathcal{L}_0 G)$ implies $\mathbb{E}L(\boldsymbol{\theta}^*) - \mathbb{E}L(\boldsymbol{\theta}_G^*) \approx \|D(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*)\|^2/2$ and

$$\|D(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*)\|^2 \leq \|G\boldsymbol{\theta}^*\|^2 - \|G\boldsymbol{\theta}_G^*\|^2 \leq \|G\boldsymbol{\theta}^*\|^2.$$

So, if the true point is “smooth” in there sense that $\|G\boldsymbol{\theta}^*\|^2$ is small, then the squared bias $\|D(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*)\|^2$ caused by penalization is small as well.

Proof. The Fisher expansion from Theorem 11.5.1 can be written as

$$\mathbb{P}\left(\|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*) - D_G b_G - \boldsymbol{\xi}_G\| \geq \diamond_G(\mathbf{x})\right) \leq 4e^{-x}.$$

The definition (11.12) of $\diamond_G(\mathbf{x})$ and (H.4) of Theorem H.1.1 imply

$$\mathbb{E}^{1/2}\|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*) - D_G b_G - \boldsymbol{\xi}_G\|^2 \leq 4\left\{\delta_G(\mathbf{r}_G)\mathbf{r}_G + 2\nu_0 \mathbf{a}_G \mathbf{r}_G (\mathbb{H}_1 + \mathbb{H}_2/\mathbf{g} + 4)\omega\right\}.$$

By the result follows by the triangle inequality

$$\mathbb{E}^{1/2}\|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|^2 \leq \mathbb{E}^{1/2}\|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*) - D_G b_G - \boldsymbol{\xi}_G\|^2 + \mathbb{E}^{1/2}\|D_G b_G + \boldsymbol{\xi}_G\|^2.$$

This yields the assertion of the theorem.

11.7 Proofs of the Fisher and Wilks expansions

This section presents the proofs of the main results and some additional statements which can be of independent interest. The principle step of the proof is a bound on the local linear approximation of the gradient $\nabla L_G(\boldsymbol{\theta})$. Below we study separately its stochastic and deterministic components coming from the decomposition $L(\boldsymbol{\theta}) = \mathbb{E}L(\boldsymbol{\theta}) + \zeta(\boldsymbol{\theta})$. With $D_G^2 = -\nabla^2 \mathbb{E}L_G(\boldsymbol{\theta}_G^*)$, this leads to the decomposition

$$\begin{aligned} \chi(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*) &\stackrel{\text{def}}{=} D_G^{-1}\{\nabla L_G(\boldsymbol{\theta}) - \nabla L_G(\boldsymbol{\theta}_G^*)\} + D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*) \\ &= D_G^{-1}\{\nabla \zeta(\boldsymbol{\theta}) - \nabla \zeta(\boldsymbol{\theta}_G^*)\} \\ &\quad + D_G^{-1}\{\nabla \mathbb{E}L_G(\boldsymbol{\theta}) - \nabla \mathbb{E}L_G(\boldsymbol{\theta}_G^*)\} + D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*). \end{aligned}$$

First we check the deterministic part. For any $\boldsymbol{\theta}$ with $\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\| \leq \mathbf{r}$ and any unit vector $\mathbf{u} \in \mathbb{R}^p$, it holds

$$\begin{aligned} \mathbf{u}^\top \mathbb{E}\chi(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*) &= \mathbf{u}^\top D_G^{-1}\{\nabla \mathbb{E}L_G(\boldsymbol{\theta}) - \nabla \mathbb{E}L_G(\boldsymbol{\theta}_G^*)\} + D_G^2(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*) \\ &= \mathbf{u}^\top \{I_p - D_G^{-1}\mathbb{F}_G(\boldsymbol{\theta}^\circ)D_G^{-1}\} D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*), \end{aligned}$$

where $\boldsymbol{\theta}^\circ = \boldsymbol{\theta}^\circ(\mathbf{u})$ is a point on the line connecting $\boldsymbol{\theta}_G^*$ and $\boldsymbol{\theta}$. This implies by $(\mathcal{L}_0 G)$

$$\|\mathbb{E}\chi(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*)\| \leq \|I_p - D_G^{-1}\mathbb{F}_G(\boldsymbol{\theta}^\circ)D_G^{-1}\|_{\text{op}} \mathbf{r} \leq \delta_G(\mathbf{r})\mathbf{r}. \quad (11.17)$$

Now we study the stochastic part. Consider the vector process

$$\mathcal{U}(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*) \stackrel{\text{def}}{=} D_G^{-1}\{\nabla\zeta(\boldsymbol{\theta}) - \nabla\zeta(\boldsymbol{\theta}_G^*)\}. \quad (11.18)$$

Further, define $\mathbf{v} = V_2(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)$ and introduce a vector process $\mathcal{Y}(\mathbf{v})$ with

$$\mathcal{Y}(\mathbf{v}) \stackrel{\text{def}}{=} V_2^{-1}[\nabla\zeta(\boldsymbol{\theta}) - \nabla\zeta(\boldsymbol{\theta}_G^*)].$$

It obviously holds $\nabla\mathcal{Y}(\mathbf{v}) = V_2^{-1}\nabla^2\zeta(\boldsymbol{\theta})V_2^{-1}$. Moreover, for any $\gamma_1, \gamma_2 \in \mathbb{R}^p$ with $\|\gamma_1\| = \|\gamma_2\| = 1$, condition **(E₂G)** implies for $|\lambda| \leq g(\mathbf{r})$

$$\log \mathbb{E} \exp \left\{ \frac{\lambda}{\omega} \gamma_1^\top \nabla\mathcal{Y}(\mathbf{v}) \gamma_2 \right\} = \log \mathbb{E} \exp \left\{ \frac{\lambda}{\omega} \gamma_1^\top V_2^{-1} \nabla^2\zeta(\boldsymbol{\theta}) V_2^{-1} \gamma_2 \right\} \leq \frac{\nu_0^2 \lambda^2}{2}.$$

Define $\Upsilon_\circ(\mathbf{r}) \stackrel{\text{def}}{=} \{\mathbf{v}: \|\mathbf{v}\| \leq \mathbf{r}, \|S\mathbf{v}\| \leq \mathbf{r}\}$ for $S^{-2} = \mathfrak{a}_G^{-2} D_G^{-1} V_2^2 D_G^{-1}$. Then

$$\sup_{\boldsymbol{\theta} \in \Theta_{0,G}(\mathbf{r})} \|\mathcal{U}(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*)\| \leq \sup_{\mathbf{v} \in \Upsilon_\circ(\mathbf{r})} \|A\mathcal{Y}(\mathbf{v})\| \quad (11.19)$$

for $A = \mathfrak{a}_G^{-1} D_G^{-1} V_2$. Theorem H.11.1 yields

$$\sup_{\mathbf{v} \in \Upsilon_\circ(\mathbf{r})} \|A\mathcal{Y}(\mathbf{v})\| \leq \sqrt{8} \nu_0 \mathfrak{z}_{\mathbb{H}}(\mathbf{x}) \mathfrak{a}_G \omega \mathbf{r} \quad (11.20)$$

on a set of a dominating probability at least $1 - e^{-x}$, where the function $\mathfrak{z}_{\mathbb{H}}(\mathbf{x})$ is given by (H.3). Putting together the bounds (11.17) and (11.19) imply the following result.

Theorem 11.7.1. *Suppose that the matrix $\mathbb{F}_G(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} L_G(\boldsymbol{\theta})$ fulfills the condition **(L₀G)** and let **(E₀G)** and **(E₂G)** be fulfilled on $\Theta_{0,G}(\mathbf{r})$ for any fixed $\mathbf{r} \leq \mathbf{r}^*$. Then*

$$\mathbb{P} \left\{ \sup_{\boldsymbol{\theta} \in \Theta_{0,G}(\mathbf{r})} \|D_G^{-1}\{\nabla L_G(\boldsymbol{\theta}) - \nabla L_G(\boldsymbol{\theta}_G^*)\} + D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\| \geq \diamondsuit_G(\mathbf{r}, \mathbf{x}) \right\} \leq e^{-x},$$

where

$$\diamondsuit_G(\mathbf{r}, \mathbf{x}) \stackrel{\text{def}}{=} \{\delta_G(\mathbf{r}) + \sqrt{8} \nu_0 \mathfrak{z}_{\mathbb{H}}(\mathbf{x}) \mathfrak{a}_G \omega\} \mathbf{r}. \quad (11.21)$$

The result of Theorem 11.7.1 can be extended to the increments of the process $\mathcal{U}(\boldsymbol{\theta})$: on a random set of probability at least $1 - e^{-x}$, it holds for any $\boldsymbol{\theta}, \boldsymbol{\theta}^\circ \in \Theta_{0,G}(\mathbf{r})$ and $\chi(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) = D_G^{-1}\{\nabla L_G(\boldsymbol{\theta}) - \nabla L_G(\boldsymbol{\theta}^\circ)\} + D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)$

$$\begin{aligned} \mathbb{E}[\chi(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)] &\leq \delta_G(\mathbf{r}) \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\| \leq 2\mathbf{r} \delta_G(\mathbf{r}), \\ \|\chi(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)\| &\leq 2\diamondsuit_G(\mathbf{r}, \mathbf{x}). \end{aligned} \quad (11.22)$$

Now we present the proof of Theorem 11.5.1 about the Fisher expansion for the qMLE $\tilde{\boldsymbol{\theta}}_G$ defined by maximization of $L_G(\boldsymbol{\theta})$. Let \mathbf{r}_G be selected to ensure that $\mathbb{P}\{\tilde{\boldsymbol{\theta}}_G \notin \Theta_{0,G}(\mathbf{r}_G)\} \leq e^{-\mathbf{x}}$. Furthermore, the definition of $\tilde{\boldsymbol{\theta}}_G$ yields $\nabla L_G(\tilde{\boldsymbol{\theta}}_G) = 0$ and

$$\chi(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*) = -D_G^{-1} \nabla L_G(\boldsymbol{\theta}_G^*) + D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*).$$

By Theorem 11.7.1, it holds on a set of a dominating probability

$$\|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*) - \boldsymbol{\xi}_G\| \leq \diamondsuit_G(\mathbf{x}) \quad (11.23)$$

as required.

As the next step, we apply the obtained results to evaluate the quality of the Wilks expansion $2L_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_G^*) \approx \|\boldsymbol{\xi}_G\|^2$. For this we derive a uniform deviation bound on the error of a quadratic approximation

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) \stackrel{\text{def}}{=} L_G(\boldsymbol{\theta}) - L_G(\boldsymbol{\theta}^\circ) - (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top \nabla L_G(\boldsymbol{\theta}^\circ) + \frac{1}{2} \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\|^2$$

in all $\boldsymbol{\theta}, \boldsymbol{\theta}^\circ \in \Theta_0$, where Θ_0 is some vicinity of a fixed point $\boldsymbol{\theta}_G^*$. With $\boldsymbol{\theta}^\circ$ fixed, the gradient $\nabla \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) \stackrel{\text{def}}{=} \frac{d}{d\boldsymbol{\theta}} \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)$ fulfills

$$\nabla \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) = \nabla L_G(\boldsymbol{\theta}) - \nabla L_G(\boldsymbol{\theta}^\circ) + D_G^2(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) = D_G \chi(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ);$$

cf. (11.18). This implies

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) = (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top \nabla \alpha(\boldsymbol{\theta}', \boldsymbol{\theta}^\circ),$$

where $\boldsymbol{\theta}'$ is a point on the line connecting $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\circ$. Further,

$$|\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)| = |(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top D_G D_G^{-1} \nabla \alpha(\boldsymbol{\theta}', \boldsymbol{\theta}^\circ)| \leq \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\| \sup_{\boldsymbol{\theta}' \in \Theta_{0,G}(\mathbf{r})} |\chi(\boldsymbol{\theta}', \boldsymbol{\theta}^\circ)|,$$

and one can apply (11.22). This yields the following result.

Theorem 11.7.2. Suppose $(\mathcal{L}_0 G)$, $(E_0 G)$, and $(E_2 G)$. For each \mathbf{r} , it holds on a random set $\Omega(\mathbf{x})$ of a dominating probability at least $1 - e^{-\mathbf{x}}$, it holds with any $\boldsymbol{\theta}, \boldsymbol{\theta}^\circ \in \Theta_{0,G}(\mathbf{r})$ and $\diamondsuit_G(\mathbf{r}, \mathbf{x})$ is from (11.21)

$$\begin{aligned} \frac{|\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*)|}{\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\|} &\leq \diamondsuit_G(\mathbf{r}, \mathbf{x}), \quad |\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*)| \leq \mathbf{r} \diamondsuit_G(\mathbf{r}, \mathbf{x}), \\ \frac{|\alpha(\boldsymbol{\theta}_G^*, \boldsymbol{\theta})|}{\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\|} &\leq 2\diamondsuit_G(\mathbf{r}, \mathbf{x}), \quad |\alpha(\boldsymbol{\theta}_G^*, \boldsymbol{\theta})| \leq 2\mathbf{r} \diamondsuit_G(\mathbf{r}, \mathbf{x}), \\ \frac{|\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)|}{\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\|} &\leq 2\diamondsuit_G(\mathbf{r}, \mathbf{x}), \quad |\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)| \leq 4\mathbf{r} \diamondsuit_G(\mathbf{r}, \mathbf{x}). \end{aligned}$$

The result of Theorem 11.7.2 for the special case with $\boldsymbol{\theta} = \boldsymbol{\theta}_G^*$ and $\boldsymbol{\theta}^\circ = \tilde{\boldsymbol{\theta}}_G$ yields in view of $\nabla L_G(\tilde{\boldsymbol{\theta}}_G) = 0$ for $\mathbf{r} = \mathbf{r}_G$ and $\diamondsuit_G(\mathbf{x}) = \diamondsuit_G(\mathbf{r}_G, \mathbf{x})$ under the condition $\tilde{\boldsymbol{\theta}}_G \in \Theta_{0,G}(\mathbf{r}_G)$

$$\left| L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*) - \|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*)\|^2/2 \right| = |\alpha(\boldsymbol{\theta}_G^*, \tilde{\boldsymbol{\theta}}_G)| \leq 2\mathbf{r}_G \diamondsuit_G(\mathbf{x}).$$

Furthermore, with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_G$ and $\boldsymbol{\theta}^\circ = \boldsymbol{\theta}_G^*$

$$\left| L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*) - \boldsymbol{\xi}_G^\top D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*) + \|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*)\|^2/2 \right| = |\alpha(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*)| \leq \mathbf{r}_G \diamondsuit_G(\mathbf{x})$$

which implies

$$\left| L(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*) - \|\boldsymbol{\xi}_G\|^2 + \|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*) - \boldsymbol{\xi}_G\|^2 \right| \leq 2\mathbf{r}_G \diamondsuit_G(\mathbf{x}).$$

Now it follows by (11.23) that

$$\left| L(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*) - \|\boldsymbol{\xi}_G\|^2/2 \right| \leq \mathbf{r}_G \diamondsuit_G(\mathbf{x}) + \diamondsuit_G^2(\mathbf{x})/2.$$

The error term can be improved if the squared root of the excess is considered. Indeed, if $\tilde{\boldsymbol{\theta}}_G \in \Theta_{0,G}(\mathbf{r}_G)$

$$\begin{aligned} \left| \{2L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*)\}^{1/2} - \|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*)\| \right| &\leq \frac{|2L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*) - \|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*)\|^2|}{\|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*)\|} \\ &\leq \frac{2|\alpha(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*)|}{\|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*)\|} \leq \sup_{\boldsymbol{\theta} \in \Theta_{0,G}(\mathbf{r}_G)} \frac{2|\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*)|}{\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\|} \leq 2\diamondsuit_G(\mathbf{x}). \end{aligned}$$

The Fisher expansion (11.23) allows to replace here the norm of the standardized error $D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*)$ with the norm of the normalized score $\boldsymbol{\xi}_G$. This completes the proof of Theorem 11.5.2.

11.8 Nonparametric BvM Theorem: Gaussian case

This section discusses the parametric BvM result for a Gaussian model and a Gaussian prior in the case of an infinite dimensional parameter $\boldsymbol{\theta}$. The case of any regular model will be considered in Section 11.9.

The main problem with an extension of the BvM result to the nonparametric situation is that a Gaussian measure in the infinite dimensional space is only defined in a weak sense while the BvM result is oriented towards the strong total variation distance. In this section, given a Gaussian prior $\mathcal{N}(0, G^{-2})$ with a compact operator G^{-2} , we aim at describing a finite-dimensional subspace of the parameter space for which the BvM result still applies. We follow the notation of the previous section. We suppose that the

parameter space Θ coincides with the space \mathbb{R}^∞ . The study easily extends to the case when $\Theta \subset \mathbb{R}^\infty$.

Consider a linear Gaussian model $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with a standard Gaussian error $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I_n)$ and with a vector $\boldsymbol{\theta}^* \in \mathbb{R}^\infty$ and a linear design operator $\Psi: \mathbb{R}^\infty \rightarrow \mathbb{R}^n$. Define the self-adjoint operator \mathbb{F} in \mathbb{R}^∞ and its square root D as

$$\mathbb{F} \stackrel{\text{def}}{=} \Psi\Psi^\top, \quad D^2 = \mathbb{F}.$$

Let $\mathcal{I}_\Psi = \text{Image}(\Psi)$ be the subspace in \mathbb{R}^∞ spanned by the columns of Ψ and let $\Pi_\Psi \stackrel{\text{def}}{=} \Psi^\top(\Psi\Psi^\top)^{-1}\Psi$ be the projector on \mathcal{I}_Ψ . Here and in the definition of the MLE $\tilde{\boldsymbol{\theta}} = (\Psi\Psi^\top)^{-1}\Psi\mathbf{Y}$ the inversion should be understood as pseudo-inversion within \mathcal{I}_Ψ . Then the posterior $\boldsymbol{\vartheta}_G$ for the Gaussian prior $\mathcal{N}(0, G^{-2})$ is also normal, centered at $\tilde{\boldsymbol{\theta}}_G = (\mathbb{F} + G^2)^{-1}\Psi\mathbf{Y}$, and

$$D(\boldsymbol{\vartheta}_G - \tilde{\boldsymbol{\theta}}_G) \mid \mathbf{Y} \sim \mathcal{N}(0, \mathbb{P}_G),$$

that is, $D(\boldsymbol{\vartheta}_G - \tilde{\boldsymbol{\theta}}_G)$ given \mathbf{Y} is normal zero mean vector in \mathbb{R}^∞ with the covariance operator $\mathbb{P}_G \stackrel{\text{def}}{=} \text{Var}(D\boldsymbol{\vartheta}_G \mid \mathbf{Y})$, where

$$\begin{aligned} \mathbb{P}_G &\stackrel{\text{def}}{=} DD_G^{-2}D = D(D^2 + G^2)^{-1}D, \\ D_G^2 &\stackrel{\text{def}}{=} D^2 + G^2. \end{aligned}$$

One can see that the mean and the variance of the posterior $\boldsymbol{\vartheta}_G$ are affected by the prior covariance G^{-2} . A formal extension of the BvM result to the infinite dimensional situation would mean that $D(\boldsymbol{\vartheta}_G - \tilde{\boldsymbol{\theta}})$ given \mathbf{Y} is nearly standard normal. Now we want to understand in which sense the BvM result can be validated. The first look yields a negative answer. Indeed, already the scaled difference of the posterior mean $\tilde{\boldsymbol{\theta}}_G = D_G^{-2}\Psi\mathbf{Y}$ and the MLE $\tilde{\boldsymbol{\theta}} = D^{-2}\Psi\mathbf{Y}$ as vectors in \mathbb{R}^∞ is significant:

$$\begin{aligned} D(\tilde{\boldsymbol{\theta}}_G - \tilde{\boldsymbol{\theta}}) &= D(D_G^{-2} - D^{-2})\Psi\mathbf{Y} \\ &= (\mathbb{P}_G - \Pi_\Psi)D^{-1}\Psi\mathbf{Y} \\ &= (\mathbb{P}_G - \Pi_\Psi)D^{-1}\Psi\mathbf{f}^* + (\mathbb{P}_G - \Pi_\Psi)\boldsymbol{\xi}, \end{aligned} \tag{11.24}$$

where $\boldsymbol{\xi} \stackrel{\text{def}}{=} D^{-1}\Psi\boldsymbol{\varepsilon}$. The deterministic term here can be bounded under smoothness conditions on the true response \mathbf{f}^* . However, if $\text{Cov}(\boldsymbol{\varepsilon}) = I_n$, then $\text{Cov}(\boldsymbol{\xi}) = \Pi_\Psi$, and for the stochastic part $(\mathbb{P}_G - \Pi_\Psi)\boldsymbol{\xi}$, it holds

$$\mathbb{E}\|(\mathbb{P}_G - \Pi_\Psi)\boldsymbol{\xi}\|^2 = \text{tr}\{(\mathbb{P}_G - \Pi_\Psi)^2\Pi_\Psi\} \rightarrow \infty, \quad \Pi_\Psi \rightarrow I,$$

because \mathbb{P}_G is a trace operator. Similar problems arise when we compare the posterior covariance operator \mathbb{P}_G with the identity. The operator \mathbb{P}_G can be treated as a smoothed projector on the effective subspace for which $\mathbb{F} \gg G^2$. At the same time, it shrinks towards zero in all the directions $\boldsymbol{\theta}$ for which $\|D\boldsymbol{\theta}\| \ll \|G\boldsymbol{\theta}\|$. By definition, the maximal eigenvalue $\lambda(\mathbb{P}_G)$ of \mathbb{P}_G fulfills $\lambda(\mathbb{P}_G) \leq 1$. Usually one observes a kind of smooth transition: the ordered eigenvalues of \mathbb{P}_G smoothly decreases from one to zero. So, there is no chance to achieve the full dimensional BvM result on the entire space \mathbb{R}^∞ . Now we discuss how the BvM statement can be adjusted to the infinite dimensional situation.

11.8.1 Finite dimensional projections and maxispaces

This section aims at describing the largest possible subspace \mathcal{I} such that after restricting on this subspace, the posterior is close to the standard normal distribution in total variation distance. Due to the Pinsker bound of Lemma D.1.1, the total variation distance between the distribution of $D(\boldsymbol{\vartheta}_G - \tilde{\boldsymbol{\theta}})$ and the standard normal law after restricting to \mathcal{I} can be measured via two quantities:

$$\Delta_{G,\mathcal{I}}^2 \stackrel{\text{def}}{=} \text{tr}\{\Pi_{\mathcal{I}}(\mathbb{P}_G - \Pi_{\mathbb{Y}})^2\Pi_{\mathcal{I}}\}, \quad (11.25)$$

$$b_{G,\mathcal{I}} \stackrel{\text{def}}{=} \|\Pi_{\mathcal{I}}D(\tilde{\boldsymbol{\theta}}_G - \tilde{\boldsymbol{\theta}})\|. \quad (11.26)$$

Although a BvM type statement does not hold on the entire space \mathbb{R}^∞ , the next result claims that it is nearly true after projecting on the subspace \mathcal{I} under the constraint “ $\Delta_{G,\mathcal{I}}$ and $b_{G,\mathcal{I}}$ are small”.

Theorem 11.8.1. *Consider the posterior $\boldsymbol{\vartheta}_G$ for a linear Gaussian model $\mathbf{Y} = \Psi^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon}$ with a Gaussian error $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$ for a positive covariance operator Σ and for a Gaussian prior $\mathcal{N}(0, G^{-2})$. Define $\Omega = D^{-1}G^2D^{-1}$*

$$D^2 \stackrel{\text{def}}{=} \Psi\Sigma^{-1}\Psi^\top,$$

$$D_G^2 \stackrel{\text{def}}{=} D^2 + G^2 = \Psi\Sigma^{-1}\Psi^\top + G^2,$$

$$\mathbb{P}_G \stackrel{\text{def}}{=} D D_G^{-2} D = D(D^2 + G^2)^{-1}D = (I + \Omega)^{-1}.$$

Let $\tilde{\boldsymbol{\theta}} = D^{-2}\Psi\Sigma^{-1}\mathbf{Y}$ and let \mathbb{Q}_G denote the distribution of $D(\boldsymbol{\vartheta}_G - \tilde{\boldsymbol{\theta}})$ given \mathbf{Y} which is a random measure on \mathbb{R}^∞ . For a subspace $\mathcal{I} \subset \mathbb{R}^\infty$, let the values $\Delta_{G,\mathcal{I}}$ and $b_{G,\mathcal{I}}$ are given by (11.25) and (11.26). Then it holds on a set $\Omega(\mathbf{x})$ with $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - e^{-\mathbf{x}}$

$$\begin{aligned} \|\mathbb{Q}_G - \mathcal{N}(0, I)\|_{\text{TV}, \mathcal{I}} &\stackrel{\text{def}}{=} \sup_{A \in \mathcal{B}(\mathcal{I})} |\mathbb{P}^\circ\{\Pi_{\mathcal{I}}D(\boldsymbol{\vartheta}_G - \tilde{\boldsymbol{\theta}}) \in A\} - \mathbb{P}(\gamma_{\mathcal{I}} \in A)| \\ &\leq \frac{1}{2} \sqrt{\Delta_{G,\mathcal{I}}^2 + b_{G,\mathcal{I}}^2} \end{aligned} \quad (11.27)$$

for a standard Gaussian vector $\gamma_{\mathcal{I}}$ on \mathcal{I} .

Below we consider a couple of examples met in the literature.

Example 11.8.1. Let $\mathbb{F} = \mathbb{F}_n = nI$ and $G^2 = \text{diag}\{g_1^2 \leq g_2^2 \leq \dots\}$. In the roughness penalty approach, $g_j^2 \approx Lj^{2\beta}$ for $\beta > 0$. Then $D_G^2 = \mathbb{F}_n + G^2$ and the operator $\mathbb{P}_G = DD_G^{-2}D$ is also diagonal with

$$\mathbb{P}_G = \text{diag}\left\{\frac{1}{1+g_1^2/n}, \frac{1}{1+g_2^2/n}, \dots\right\}.$$

Now for a fixed p , consider the subspace \mathcal{I}_p spanned by the first p basis elements. For $j \leq p$, it holds

$$1 - \frac{1}{1+g_j^2/n} = \frac{g_j^2/n}{1+g_j^2/n} \leq \frac{g_p^2/n}{1+g_p^2/n},$$

The value $\Delta_{G,\mathcal{I}}$ can be bounded as follows:

$$\Delta_{G,\mathcal{I}}^2 = \sum_{j=1}^p \left(1 - \frac{1}{1+g_j^2/n}\right)^2 = \sum_{j=1}^p \frac{g_j^4/n^2}{(1+g_j^2/n)^2} \leq n^{-2} \sum_{j=1}^p g_j^4 \leq \frac{L^2}{4\beta+1} n^{-2} p^{4\beta+1}.$$

Asymptotically, as $n \rightarrow \infty$ and the dimension $p = p_n$ grows with n , the condition “ $\Delta_{G,\mathcal{I}}$ small” reads as $n^{-2} p_n^{4\beta+1} \rightarrow 0$.

Example 11.8.2. Let as in previous example, $\mathbb{F} = \sigma_n^2 I$. Now we consider a slightly different situation when the prior covariance operator $G^2 = \tau_n^2 \Pi_{\mathcal{I}}$ coincides up to a constant factor with the projector Π_0 on the subspace \mathcal{I} of dimension $p = p_n$ spanned by the first basis eigenvectors. Then $D_G^2 = \sigma_n^2 I + \tau_n^2 I_{\mathcal{I}}$ and $\mathbb{P}_G = (I + (\tau_n/\sigma_n)^2 I_{\mathcal{I}})^{-1}$. Therefore,

$$\Delta_{G,\mathcal{I}}^2 = \sum_{j=1}^{p_n} \left(1 - \frac{1}{1+(\tau_n/\sigma_n)^2}\right)^2 \leq p_n (\tau_n/\sigma_n)^4.$$

This leads to the condition on the dimensionality p_n of \mathcal{I} : $p_n (\tau_n/\sigma_n)^4 \rightarrow 0$ as $n \rightarrow \infty$ which coincides with [Bontemps \(2011\)](#) and [Johnstone \(2010\)](#).

Now we consider the bias term $b_{G,\mathcal{I}}$. The decomposition (11.24) implies

$$\begin{aligned} \Pi_{\mathcal{I}} D(\tilde{\boldsymbol{\theta}}_G - \tilde{\boldsymbol{\theta}}) &= \Pi_{\mathcal{I}} D(D_G^{-2} - D^{-2}) \Psi \mathbf{Y} \\ &= \Pi_{\mathcal{I}} (\mathbb{P}_G - \Pi_{\Psi}) D^{-1} \Psi \mathbf{Y} \\ &= \Pi_{\mathcal{I}} (\mathbb{P}_G - \Pi_{\Psi}) D^{-1} \Psi \mathbf{f}^* + \Pi_{\mathcal{I}} (\mathbb{P}_G - \Pi_{\Psi}) \boldsymbol{\xi}, \end{aligned} \quad (11.28)$$

where $\boldsymbol{\xi} \stackrel{\text{def}}{=} D^{-1} \Psi \boldsymbol{\varepsilon}$. Under $\text{Cov}(\boldsymbol{\xi}) = \Pi_{\Psi}$, it holds for the stochastic component $(\mathbb{P}_G - \Pi_{\Psi}) \boldsymbol{\xi}$ after projecting on \mathcal{I} :

$$\mathbb{E}\|\Pi_{\mathcal{I}}(\mathbb{P}_G - \Pi_{\Psi})\xi\|^2 = \text{tr}\{\Pi_{\mathcal{I}}(\mathbb{P}_G - \Pi_{\Psi})^2\Pi_{\mathcal{I}}\} = \Delta_{G,\mathcal{I}}^2.$$

If ξ is standard Gaussian one can apply the result of Corollary B.1.2 to bound

$$\|(\mathbb{P}_G - \Pi_{\Psi})\xi\| \leq C\Delta_{G,\mathcal{I}}$$

on a set of dominating probability. In the non-Gaussian situation the value $\|(\mathbb{P}_G - \Pi_{\Psi})\xi\|$ can be bounded in a similar way using Theorem B.2.2.

Under noise misspecification, one can obtain a similar bound for a regular noise $\text{Cov}(\xi) \leq \alpha^2 I_p$ with an additional α factor. So, the stochastic term in (11.28) can be controlled if $\Delta_{G,\mathcal{I}}$ is small. The value $\|\Pi_{\mathcal{I}}(\mathbb{P}_G - \Pi_{\Psi})D^{-1}\Psi f^*\|$ depends on the smoothness properties of f^* . Obviously

$$\|\Pi_{\mathcal{I}}(\mathbb{P}_G - \Pi_{\Psi})D^{-1}\Psi f^*\| \leq \|(\mathbb{P}_G - \Pi_{\Psi})D^{-1}\Psi f^*\|$$

and it is sufficient to show that a smooth projection \mathbb{P}_G does not change significantly the vector $D^{-1}\Psi f^*$.

11.8.2 Concentration sets for the posterior

In a linear Gaussian model and a Gaussian prior, the posterior is also normal, centered at $\tilde{\theta}_G = (D^2 + G^2)^{-1}\Psi Y$, and

$$(\vartheta_G - \tilde{\theta}_G) | Y \sim \mathcal{N}(0, D_G^{-2}),$$

that is, $\vartheta_G - \tilde{\theta}_G$ given Y is normal zero mean vector in \mathbb{R}^∞ with the covariance operator $D_G^{-2} \stackrel{\text{def}}{=} \text{Var}(\vartheta_G | Y)$. In other words, $D_G(\vartheta_G - \tilde{\theta}_G)$ given Y is a standard Gaussian measure in the Hilbert space \mathbb{R}^∞ . Unfortunately, this measure does not concentrate on a ball of any fixed radius r . Equivalently, $\vartheta_G | Y$ does not concentrate on the ball $\{\theta : \|D_G(\vartheta_G - \tilde{\theta}_G)\| \leq r\}$. An important question is to describe the “smallest” subset in \mathbb{R}^∞ on which the posterior concentrates with a large probability. The Gaussian structure helps to easily answer this question. Indeed, for any p -matrix $Q \leq D_G$, consider the rescaled posterior $Q(\vartheta_G - \tilde{\theta}_G) | Y$. It is conditionally on Y normal zero mean with the covariance matrix $B_{Q|G} = QD_G^{-2}Q$. If $B_{Q|G}$ is a trace operator with $P_{Q|G} = \text{tr}(B_{Q|G}) < \infty$, then the results of Theorem B.1.1 imply that the posterior $Q(\vartheta_G - \tilde{\theta}_G) | Y = \mathcal{N}(0, B_{Q|G})$ concentrates on $\Theta_Q(r)$ for $r \geq \sqrt{pq} + \sqrt{2x}$. In particular, with $Q = D$, it holds $D(\vartheta_G - \tilde{\theta}_G) | Y \sim \mathcal{N}(0, B)$ for $B = D D_G^{-2} D$, and we only need B to be a trace operator.

Theorem 11.8.2. *Let a self-adjoint operator Q be such that $B_{Q|G} = QD_G^{-2}Q$ satisfies*

$$\mathbf{p}_{Q|G} \stackrel{\text{def}}{=} \text{tr}(B_{Q|G}) \leq \infty.$$

Then for any $\mathbf{x} \geq 0$, it holds almost surely

$$\mathbb{P}\left(\|Q(\boldsymbol{\vartheta}_G - \tilde{\boldsymbol{\theta}}_G)\| \geq z(\mathbf{x}, B_{Q|G}) \mid \mathbf{Y}\right) = e^{-\mathbf{x}}.$$

11.8.3 Frequentist coverage for Bayesian credible sets

The obtained results help to understand under which conditions one can use the Bayesian credible sets \mathcal{E} as frequentist confidence sets. This question relies to the probability of the coverage $\boldsymbol{\theta}^* \in \mathcal{E}$. One can choose one of two strategies. The sieve approach suggests to fix a finite dimensional subspace \mathcal{I} and consider all sets \mathcal{A} in this subspace. The corresponding credible sets read as

$$\mathcal{E}_{G,\mathcal{I}}(z) = \mathbb{I}\left(\|D_{G,\mathcal{I}}(\boldsymbol{\vartheta}_{G,\mathcal{I}} - \tilde{\boldsymbol{\theta}}_{G,\mathcal{I}})\| \leq z\right). \quad (11.29)$$

Here $\boldsymbol{\vartheta}_{G,\mathcal{I}}$ is the projection of $\boldsymbol{\vartheta}_G$ on \mathcal{I} and similarly for $\tilde{\boldsymbol{\theta}}_{G,\mathcal{I}}$, and $D_{G,\mathcal{I}}$ is defined by

$$D_{G,\mathcal{I}}^{-2} = \text{Var}(\boldsymbol{\vartheta}_{G,\mathcal{I}} \mid \mathbf{Y}) = \text{Var}(\Pi_{\mathcal{I}} \boldsymbol{\vartheta}_G \mid \mathbf{Y}) = \Pi_{\mathcal{I}} D_G^{-2} \Pi_{\mathcal{I}}^\top,$$

so that $D_{G,\mathcal{I}}(\boldsymbol{\vartheta}_{G,\mathcal{I}} - \tilde{\boldsymbol{\theta}}_{G,\mathcal{I}})$ is a standard Gaussian vector $\boldsymbol{\gamma}_{\mathcal{I}}$ in the Euclidean space \mathcal{I} of dimension $p = \dim(\mathcal{I})$. The choice z as a quantile of χ_p^2 ensures a prescribed credible probability:

$$\mathbb{P}(\mathcal{E}_{G,\mathcal{I}}(z) \mid \mathbf{Y}) = \mathbb{P}^o(\|\boldsymbol{\gamma}_{\mathcal{I}}\| \leq z).$$

The other strategy is to fix an operator Q such that $\mathbf{p}_{Q|G} = \text{tr}(D_G^{-1} Q^2 D_G^{-1}) < \infty$. Then $Q(\boldsymbol{\vartheta}_G - \tilde{\boldsymbol{\theta}}_G)$ is conditionally on \mathbf{Y} normal zero mean with the covariance $B_Q = D_G^{-1} Q^2 D_G^{-1}$ and one can build elliptic credible sets in the form

$$\mathcal{E}_Q(z) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} : \|Q(\boldsymbol{\vartheta}_G - \tilde{\boldsymbol{\theta}}_G)\| \leq z\} = \{\boldsymbol{\theta} : \|Q D_G^{-1} \boldsymbol{\gamma}\| \leq z\}, \quad (11.30)$$

where $\boldsymbol{\gamma}$ is a standard normal law. Usually $z = z_\alpha$ is to be selected to provide the prescribed credible probability:

$$\mathbb{P}^o(\|Q D_G^{-1} \boldsymbol{\gamma}\| \leq z_\alpha) = 1 - \alpha \quad (11.31)$$

for the given nominal level $1 - \alpha$. Note that (11.29) corresponds to the special choice $Q = D_{G,\mathcal{I}} \Pi_{\mathcal{I}}$.

Now we discuss whether this set can be used as the frequentist confidence set. A particular question to check is the coverage probability

$$\mathbb{I}P(\boldsymbol{\theta}^* \in \mathcal{E}(z_\alpha)).$$

Ideally it coincides or is close to the credible probability $1 - \alpha$. To answer this question, we use the decomposition of $\tilde{\boldsymbol{\theta}}_G$:

$$\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^* = \boldsymbol{\theta}_G^* - \boldsymbol{\theta}^* + D_G^{-2}\nabla = \boldsymbol{\theta}_G^* - \boldsymbol{\theta}^* + D_G^{-2}D\xi, \quad (11.32)$$

where $\boldsymbol{\theta}^* = D^{-2}\Psi\Sigma^{-1}\mathbf{f}^*$, $\boldsymbol{\theta}_G^* = D_G^{-2}\Psi\Sigma^{-1}\mathbf{f}^* = D_G^{-2}D^2\boldsymbol{\theta}^*$, $\nabla = \Psi\Sigma^{-1}\boldsymbol{\varepsilon}$, and $\xi = D^{-1}\nabla$. The event $\{\boldsymbol{\theta}^* \in \mathcal{E}(z)\}$ can be rewritten as

$$\{\|Q(\boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}}_G)\| \leq z\} = \{\|Q(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^* + D_G^{-2}D\xi)\| \leq z\}.$$

Therefore, we have to compare two probabilities

$$\mathbb{I}P^\circ(\|QD_G^{-1}\boldsymbol{\gamma}\| \leq z) \text{ vs } \mathbb{I}P(\|Q(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^* + D_G^{-2}D\xi)\| \leq z). \quad (11.33)$$

Under correct noise specification, ξ is standard normal and validity of Bayes credible sets $\mathcal{E}(z_\alpha)$ relies to the “small bias” condition “ $\|D^{-1}G^2\boldsymbol{\theta}^*\|$ is small”.

Theorem 11.8.3. *Let $\mathbf{Y} = \Psi^\top\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with a zero mean Gaussian noise $\boldsymbol{\varepsilon}$. For the credible set $\mathcal{E}(z_\alpha)$ from (11.30) and (11.31), it holds*

$$\mathbb{I}P(\boldsymbol{\theta}^* \in \mathcal{E}(z_\alpha)) \geq 1 - \alpha - \frac{1}{2}\|D^{-1}G^2\boldsymbol{\theta}^*\|. \quad (11.34)$$

Proof. The vector $QD_G^{-1}\boldsymbol{\gamma}$ is normal with $QD_G^{-1}\boldsymbol{\gamma} \sim \mathcal{N}(0, S)$ for $S = QD_G^{-2}Q$. Under the correct noise specification, the vector ξ is standard normal and the vector $Q(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^* + D_G^{-2}D\xi)$ is normal as well with the mean

$$\mathbf{b} = Q(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*) = Q(I - D_G^{-2}D^2)\boldsymbol{\theta}^* = QD_G^{-2}G^2\boldsymbol{\theta}^*$$

and the variance $S_1 = QD_G^{-2}D^2D_G^{-2}Q$. We proceed in two steps. Observe that $D_G^2 \geq D^2$ implies $S_1 \leq S$, therefore

$$\mathbb{I}P^\circ(\boldsymbol{\gamma}^\top S_1 \boldsymbol{\gamma} \leq z_\alpha) \geq \mathbb{I}P^\circ(\boldsymbol{\gamma}^\top S \boldsymbol{\gamma} \leq z_\alpha) = 1 - \alpha.$$

Next we compare the probability of the centered ball $\mathbb{I}P^\circ\{\boldsymbol{\gamma}^\top S_1 \boldsymbol{\gamma} \leq z_\alpha\}$ with a similar probability of a non-centered ball $\mathbb{I}P^\circ\{(\boldsymbol{\gamma} - \mathbf{b})^\top S_1(\boldsymbol{\gamma} - \mathbf{b}) \leq z_\alpha\}$. Simple algebra yields for Q positive and invertible

$$\mathbf{b}^\top S_1^{-1} \mathbf{b} = \boldsymbol{\theta}^{*\top} G^2 D^{-2} G^2 \boldsymbol{\theta}^* = \|D^{-1}G^2\boldsymbol{\theta}^*\|^2.$$

If Q is nonnegative then

$$\mathbf{b} S_1^{-1} \mathbf{b} \leq \|D^{-1} G^2 \boldsymbol{\theta}^*\|^2.$$

Now by Pinsker's inequality and Lemma D.1.1

$$\begin{aligned} & \left| \mathbb{P}^\circ \{ \boldsymbol{\gamma}^\top S_1 \boldsymbol{\gamma} \leq z_\alpha \} - \mathbb{P}^\circ \{ (\boldsymbol{\gamma} - \mathbf{b})^\top S_1 (\boldsymbol{\gamma} - \mathbf{b}) \leq z_\alpha \} \right| \\ & \leq \frac{1}{2} \sqrt{\mathbf{b} S_1^{-1} \mathbf{b}} = \frac{1}{2} \|D^{-1} G^2 \boldsymbol{\theta}^*\|. \end{aligned}$$

as required.

Remark 11.8.1. The bound (11.34) of Theorem 11.8.3 is one-sided: the coverage probability is up to the bias term at least as big as the credible one. An interesting question is whether this bound is sharp and the use of the credible set $\mathcal{E}(z_\alpha)$ ensures the nominal confidence level $1 - \alpha$. To answer this question, we have to account for different covariance structure S and S_1 . It is already addressed in the result (11.27) of Theorem 11.8.1 for the case when Q is a projector. If, however, the image of the operator Q is infinite dimensional, the term Δ_G in (11.27) explodes.

Now we discuss the case of a possible noise misspecification when the noise component $\boldsymbol{\varepsilon}$ is a Gaussian zero mean vector with the covariance matrix Σ_0 , while the likelihood $L(\boldsymbol{\theta})$ uses the other matrix Σ . This fact does not affect the Bayesian procedure which is entirely based on the likelihood structure and the prior distribution. However, the coverage probability of the credible set $\mathcal{E}(z_\alpha)$ strictly depends on the noise variance. We present one more one-sided result which extends the one-sided bound of Theorem 11.8.3.

Theorem 11.8.4. *Let $\mathbf{Y} = \boldsymbol{\Psi}^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma_0)$, while $L(\boldsymbol{\theta})$ uses another matrix Σ for $\Sigma_0 \leq \Sigma$. For the credible set $\mathcal{E}(z_\alpha)$ from (11.30) and (11.31), it holds*

$$\mathbb{P}(\boldsymbol{\theta}^* \in \mathcal{E}(z_\alpha)) \geq 1 - \alpha - \frac{1}{2} \|D^{-1} G^2 \boldsymbol{\theta}^*\|.$$

Proof. We use again the decomposition (11.32). The deterministic term $\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*$ remains the same as in the case of a correct model. The stochastic term $Q D_G^{-2} \nabla$ is normal zero mean with the covariance

$$S_1 = \text{Var}(Q D_G^{-2} \nabla) = \text{Var}(Q D_G^{-2} \boldsymbol{\Psi} \Sigma^{-1} \boldsymbol{\varepsilon}) = Q D_G^{-2} \boldsymbol{\Psi} \Sigma^{-1} \Sigma_0 \Sigma^{-1} \boldsymbol{\Psi}^\top D_G^{-2} Q.$$

If $\Sigma_0 \leq \Sigma$, then $S_1 \leq S = Q D_G^{-2} \boldsymbol{\Psi} \Sigma^{-1} \boldsymbol{\Psi}^\top D_G^{-2} Q$, and we can proceed exactly as in the proof of Theorem 11.8.3.

11.8.4 Non-Gaussian errors

Here we briefly comment on the case when the errors $\boldsymbol{\varepsilon}$ in the regression model $\mathbf{Y} = \boldsymbol{\Psi}^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon}$ are not Gaussian, but we still use the Gaussian likelihood. Note that the Bayesian

inference is being done conditionally on \mathbf{Y} , that is, the data distribution does not show up in Bayesian calculus. In the contrary, the frequentist covering probability depends heavily on the distribution of the errors $\boldsymbol{\varepsilon}$. More precisely, it relies on the distribution of the vector $\boldsymbol{\xi} = D^{-1}\nabla$; cf. (11.33). However, under regularity conditions, one can use the usual Gaussian approximation technique to show that the coverage probability $\mathbb{P}(\|\mathbf{b} + QD_G^{-2}D\boldsymbol{\xi}\| \leq z)$ is close to the Gaussian probability $\mathbb{P}(\|\mathbf{b} + QD_G^{-2}D\tilde{\boldsymbol{\xi}}\| \leq z)$, where $\tilde{\boldsymbol{\xi}}$ is a Gaussian zero mean vector with $\text{Var}(\boldsymbol{\xi}) = \text{Var}(\tilde{\boldsymbol{\xi}})$.

11.9 Nonparametric BvM Theorem: non-Gaussian case

This section discusses the BvM result for a general model with a high-dimensional or infinite dimensional parameter set with a Gaussian prior. It appears that the impact of the Gaussian prior is very similar to the impact of the roughness penalty in the parameter estimation. It allows to reduce the effective dimension of the problem and to establish a Gaussian approximation of the posterior with the error bound depending on the effective dimension only. However, the BvM result appears to be more involved than the Fisher or Wilks expansions by two reasons. The first problem is related to the concentration result. The posterior concentration requires to bound the integral of the likelihood process in the complement of the local vicinity and this is a hard task in the nonparametric setup. The second problem is due to fact that a standard Gaussian measure on \mathbb{R}^∞ is only defined in a weak sense. In particular, it does not concentrate on any ℓ_2 ball in \mathbb{R}^∞ . This makes it difficult to study the total variation distance between the posterior and the Gaussian law.

Consider a model given by a log-likelihood $L(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. Without loss of generality, a Gaussian prior $\Pi(\boldsymbol{\theta})$ will be assumed to be centered at zero. By G^{-2} we denote its covariance matrix, so that, $\Pi \sim \mathcal{N}(0, G^{-2})$. The main question studied below is to understand under which conditions on the prior covariance G^{-2} and the model, the BvM-type result holds and what is the error term in the BvM approximation. For the case when the log-likelihood function is not quadratic in $\boldsymbol{\theta}$, the study is more involved. The posterior is obtained by normalizing the product $L_G(\boldsymbol{\theta})$ given by

$$L_G(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2/2.$$

This expression arises in penalized maximum likelihood estimation, one can treat the prior term $\|G\boldsymbol{\theta}\|^2/2$ as roughness penalty. Therefore, we expect the same effect of using the Gaussian prior as in the penalized MLE case: it improves the concentration properties but can introduce some bias in estimation. We already know that the penalized MLE $\tilde{\boldsymbol{\theta}}_G$ concentrates in the vicinity $\Theta_{0,G}(\mathbf{r}_G)$ of the point $\boldsymbol{\theta}_G^* = \operatorname{argmin}_{\boldsymbol{\theta}} \mathbb{E}L_G(\boldsymbol{\theta})$. for a proper

choice of $\mathbf{r}_G \approx \sqrt{\mathbf{p}_G} + \sqrt{2\mathbf{x}}$. Now we aim at describing the local set of concentration for the posterior. Unfortunately, even in the Gaussian case, the posterior does not concentrate on $\Theta_{0,G}(\mathbf{r}_G)$. Instead, we show that the posterior well concentrates on a larger local set $U_Q(\mathbf{r}_0)$ with a proper choice of Q and \mathbf{r}_0 :

$$U_Q(\mathbf{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} : \|Q(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}\}, \quad (11.35)$$

where the matrix Q fulfills

$$D^2 \leq Q^2 \leq D_G^2. \quad (11.36)$$

Similarly to the Gaussian case, the posterior concentration requires that $QD_G^{-2}Q$ is a trace operator with $\mathbf{p}_{Q|G} \stackrel{\text{def}}{=} \text{tr}(QD_G^{-2}Q) < \infty$. The results below extend the concentration result under conditions (\mathcal{L}_0) and (\mathcal{L}) with $\mathbf{r} \approx 2(\mathbf{p}_{Q|G}^{1/2} + \sqrt{2\mathbf{x}})$. Moreover, we will show that the rescaled posterior $Q(\boldsymbol{\vartheta}_G - \tilde{\boldsymbol{\theta}}_G) \mid \mathbf{Y}$ is very close to a Gaussian distribution $\mathcal{N}(0, QD_G^{-2}Q)$ as in the Gaussian case.

11.9.1 A linear stochastic term

First we consider one special case of a *linear* (in $\boldsymbol{\theta}$) stochastic term $\zeta(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathbb{E}L(\boldsymbol{\theta})$. A typical example when such situations arise is given by generalized linear models; see Chapter 14. The study is much more transparent in this case. Linearity of the stochastic component $\zeta(\boldsymbol{\theta})$ implies for any $\boldsymbol{\theta}^\circ$ the representation

$$\zeta(\boldsymbol{\theta}) - \zeta(\boldsymbol{\theta}^\circ) = (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top \nabla \zeta(\boldsymbol{\theta}^*) = \boldsymbol{\xi}_G^\top D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) \quad (11.37)$$

with $\boldsymbol{\xi}_G = D_G^{-1} \nabla \zeta(\boldsymbol{\theta}^*)$ similarly to the Gaussian case. Below we assume that the random vector $\boldsymbol{\xi}_G$ fulfill $\|\boldsymbol{\xi}_G\| \leq z(B_G, \mathbf{x})$ for $B_G = D_G^{-1} V^2 D_G^{-1}$ on a random set $\Omega(\mathbf{x})$ with $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - e^{-\mathbf{x}}$.

First we establish an upper bound on the random quantity

$$\rho_Q(\mathbf{r}) \stackrel{\text{def}}{=} \frac{\int_{\boldsymbol{\theta} \setminus U_Q(\mathbf{r})} \exp\{L_G(\boldsymbol{\theta})\} d\boldsymbol{\theta}}{\int_{U_Q(\mathbf{r})} \exp\{L_G(\boldsymbol{\theta})\} d\boldsymbol{\theta}},$$

which describes the posterior probability of the complement of $U_Q(\mathbf{r})$. Obviously $\mathbb{P}\{\boldsymbol{\vartheta}_G \notin U_Q(\mathbf{r}) \mid \mathbf{Y}\} \leq \rho_Q(\mathbf{r})$. Therefore, small values of $\rho_Q(\mathbf{r})$ indicate a concentration of the posterior on the set $U_Q(\mathbf{r})$. The next result presents sufficient conditions for the contraction property for the local set $U_Q(\mathbf{r}_0)$ from (11.35).

Theorem 11.9.1. *Let the stochastic component $\zeta(\boldsymbol{\theta})$ of $L(\boldsymbol{\theta})$ satisfy (11.37). Suppose that for a fixed \mathbf{r}_0 , the local neighborhood $U_Q(\mathbf{r}_0)$ is given by (11.35) with a matrix Q satisfying (11.36). Further, let the expected log-likelihood $\mathbb{E}L(\boldsymbol{\theta})$ fulfill*

$$\left| \mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \frac{1}{2}\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \right| \leq \frac{1}{2}\mathbf{r}_0^2\delta(\mathbf{r}_0), \quad \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}_0,$$

and

$$-\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \geq \mathbf{C}_0\mathbf{r}_0\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| - \frac{1}{2}\mathbf{C}_0\mathbf{r}_0^2, \quad \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| > \mathbf{r}_0, \quad (11.38)$$

with $\mathbf{C}_0 = 1 - \delta(\mathbf{r}_0)$. If \mathbf{r}_0 is large enough to provide

$$\mathbf{C}_0\mathbf{r}_0 \geq 2\left(\sqrt{\mathbf{p}_{Q|G} + \mathbf{x}} + 1 + z(B_G, \mathbf{x}) + b_G\right) \quad (11.39)$$

with $\mathbf{p}_{Q|G} = \text{tr}(QD_G^{-2}Q)$, $B_G = D_G^{-1}V^2D_G^{-1}$, and

$$b_G = \|DD_G^{-2}G^2\boldsymbol{\theta}^*\| \leq \|D_G^{-1}G^2\boldsymbol{\theta}^*\| \leq \|G\boldsymbol{\theta}^*\|, \quad (11.40)$$

then the quantity $\rho_Q(\mathbf{r}_0)$ fulfills under condition $\|\boldsymbol{\xi}_G\| \leq z(B_G, \mathbf{x})$

$$\rho_Q(\mathbf{r}_0) \leq e^{4\Delta_o(\mathbf{x})} \left\{ e^{-\frac{1}{2}(\mathbf{p}_{Q|G} + \mathbf{x})} + \frac{e^{-\mathbf{x}}}{1 - e^{-\mathbf{x}}} \right\}, \quad (11.41)$$

where $\Delta_o(\mathbf{x}) = \mathbf{r}_0^2\delta(\mathbf{r}_0)/2$.

Proof. Define the point $\boldsymbol{\theta}_G^\dagger$ by

$$D_G^2\boldsymbol{\theta}_G^\dagger = D^2\boldsymbol{\theta}^*. \quad (11.42)$$

This point is obtained by minimizing the quadratic expression $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + \|G\boldsymbol{\theta}\|^2$ and it approximates the point $\boldsymbol{\theta}_G^*$ which maximizes $\mathbb{E}L_G(\boldsymbol{\theta})$. It is easy to check that

$$\|D(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^\dagger)\| = \|DD_G^{-2}G^2\boldsymbol{\theta}^*\| = b_G,$$

and condition (11.39) implies that $\boldsymbol{\theta}_G^\dagger \in \Theta_0(\mathbf{r}_0)$. Obviously

$$\rho_Q(\mathbf{r}) = \frac{\int_{\Theta \setminus U_Q(\mathbf{r})} \exp\{L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger)\} d\boldsymbol{\theta}}{\int_{U_Q(\mathbf{r})} \exp\{L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger)\} d\boldsymbol{\theta}},$$

where $L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) = L_G(\boldsymbol{\theta}) - L_G(\boldsymbol{\theta}_G^\dagger)$. The first step of the proof describes the behavior of the expected log-likelihood within and outside of $\Theta_0(\mathbf{r}_0)$.

Lemma 11.9.1. *With $\boldsymbol{\theta}_G^\dagger = D_G^{-2}D^2\boldsymbol{\theta}^*$, under condition (\mathcal{L}_0) , the expected penalized log-likelihood $\mathbb{E}L_G(\boldsymbol{\theta})$ fulfills for all $\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)$*

$$\left| \mathbb{E}L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) + \frac{1}{2}\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)\|^2 \right| \leq \mathbf{r}_0^2\delta(\mathbf{r}_0) = 2\Delta_o(\mathbf{x}), \quad (11.43)$$

and for $\boldsymbol{\theta} \in \Theta \setminus \Theta_0(\mathbf{r}_0)$

$$\begin{aligned}
-\mathbb{E}L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) &\geq -\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \frac{1}{2}\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + \frac{1}{2}\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)\|^2 - \Delta_o(\mathbf{x}) \\
&\geq C_0 r_0 \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| - \frac{1}{2}C_0 r_0^2 - \frac{1}{2}\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \\
&\quad + \frac{1}{2}\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)\|^2 - \Delta_o(\mathbf{x}). \tag{11.44}
\end{aligned}$$

Proof. The use of $D_G^2 \boldsymbol{\theta}_G^\dagger = D^2 \boldsymbol{\theta}^*$ yields

$$\|G\boldsymbol{\theta}\|^2 - \|G\boldsymbol{\theta}_G^\dagger\|^2 = -\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + \|D(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^\dagger)\|^2 + \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)\|^2. \tag{11.45}$$

This identity can be checked by observing that both sides are quadratic in $\boldsymbol{\theta}$ and their values at $\boldsymbol{\theta}_G^\dagger$ coincide as well as the gradient and the Hessian. For any $\boldsymbol{\theta} \in \Theta_0(r_0)$, this implies

$$\begin{aligned}
2\mathbb{E}L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) + \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)\|^2 \\
= 2\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + 2\mathbb{E}L(\boldsymbol{\theta}^*, \boldsymbol{\theta}_G^\dagger) - \|G\boldsymbol{\theta}\|^2 + \|G\boldsymbol{\theta}_G^\dagger\|^2 + \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)\|^2 \\
= 2\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + 2\mathbb{E}L(\boldsymbol{\theta}^*, \boldsymbol{\theta}_G^\dagger) + \|D(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^\dagger)\|^2,
\end{aligned}$$

and the inequality (11.43) follows by **(L₀)**.

Further, outside of $\Theta_0(r_0)$, it holds by definition and by (11.45)

$$\begin{aligned}
-\mathbb{E}L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) &= -\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) + \frac{1}{2}\|G\boldsymbol{\theta}\|^2 - \frac{1}{2}\|G\boldsymbol{\theta}_G^\dagger\|^2 \\
&= -\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbb{E}L(\boldsymbol{\theta}^*, \boldsymbol{\theta}_G^\dagger) + \frac{1}{2}\|D(\boldsymbol{\theta}_G^\dagger - \boldsymbol{\theta}^*)\|^2 \\
&\quad - \frac{1}{2}\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + \frac{1}{2}\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)\|^2. \tag{11.46}
\end{aligned}$$

It remains to note that condition **(L₀)** implies $|\mathbb{E}L(\boldsymbol{\theta}^*, \boldsymbol{\theta}_G^\dagger) - \frac{1}{2}\|D(\boldsymbol{\theta}_G^\dagger - \boldsymbol{\theta}^*)\|^2| \leq \Delta_o(r_0)$, and (11.44) follows by (11.38).

Now we apply the usual decomposition $L_G(\boldsymbol{\theta}) = \mathbb{E}L_G(\boldsymbol{\theta}) + \zeta(\boldsymbol{\theta})$. The stochastic term $\zeta(\boldsymbol{\theta})$ is linear and follows (11.37), for the deterministic one we use (11.43). This yield for $\boldsymbol{\theta} \in \Theta_0(r_0)$

$$\begin{aligned}
L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) &\geq -\frac{1}{2}\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)\|^2 + \boldsymbol{\xi}_G^\top D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger) - 2\Delta_o(\mathbf{x}) \\
&= -\frac{1}{2}\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger) - \boldsymbol{\xi}_G\|^2 + \frac{1}{2}\|\boldsymbol{\xi}_G\|^2 - 2\Delta_o(\mathbf{x}), \tag{11.47}
\end{aligned}$$

$$L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) \leq -\frac{1}{2}\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger) - \boldsymbol{\xi}_G\|^2 + \frac{1}{2}\|\boldsymbol{\xi}_G\|^2 + 2\Delta_o(\mathbf{x}). \tag{11.48}$$

For $\boldsymbol{\theta}$ outside of $\Theta_0(r_0)$, the use of (11.38) implies in a similar way

$$\begin{aligned}
L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) &\leq -C_0 r_0 \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| + \frac{1}{2} C_0 r_0^2 \\
&\quad + \frac{1}{2} \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 - \frac{1}{2} \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)\|^2 + \boldsymbol{\xi}_G^\top D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger) + \Delta_o(\mathbf{x}) \\
&= -C_0 r_0 \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| + \frac{1}{2} C_0 r_0^2 + \frac{1}{2} \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \\
&\quad - \frac{1}{2} \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger) - \boldsymbol{\xi}_G\|^2 + \frac{1}{2} \|\boldsymbol{\xi}_G\|^2 + \Delta_o(\mathbf{x}).
\end{aligned}$$

Moreover, the condition $Q \geq D$ implies $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \|Q(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|$. As the function $f(r) = -C_0 r_0 r - r^2/2$ increases in r for $r \geq r_0$ due to $f'(r) = -C_0 r_0 + r \geq 0$, it follows

$$-C_0 r_0 \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| + \frac{1}{2} \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \leq -C_0 r_0 \|Q(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| + \frac{1}{2} \|Q(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2$$

for $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \geq r_0$. Define new variable $\mathbf{u} = D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger) - \boldsymbol{\xi}_G$, the vector $\mathbf{b} = D_G(\boldsymbol{\theta}_G^\dagger - \boldsymbol{\theta}^*) - \boldsymbol{\xi}_G$, and the operator $T = QD_G^{-1}$. Then $Q(\boldsymbol{\theta} - \boldsymbol{\theta}^*) = T(\mathbf{u} - \mathbf{b})$ and

$$\begin{aligned}
L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) &\leq -C_0 r_0 \|T(\mathbf{u} - \mathbf{b})\| + \frac{C_0 r_0^2}{2} + \frac{1}{2} \|T(\mathbf{u} - \mathbf{b})\|^2 - \frac{\|\mathbf{u}\|^2}{2} \\
&\quad + \frac{1}{2} \|\boldsymbol{\xi}_G\|^2 + \Delta_o(\mathbf{x})
\end{aligned} \tag{11.49}$$

for all $\boldsymbol{\theta}$ with $\|T(\mathbf{u} - \mathbf{b})\| \geq r_0$. Now we decompose the tail posterior probability into two zones $\Theta \setminus \Theta_0(r_0)$ and $\Theta_0(r_0) \setminus U_Q(r_0)$ yielding

$$\begin{aligned}
\rho_Q(r_0) &= \frac{\int_{\Theta \setminus U_Q(r_0)} \exp\{L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger)\} d\boldsymbol{\theta}}{\int_{U_Q(r_0)} \exp\{L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger)\} d\boldsymbol{\theta}} \\
&\leq \frac{\int_{\Theta \setminus \Theta_0(r_0)} \exp\{L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger)\} d\boldsymbol{\theta}}{\int_{U_Q(r_0)} \exp\{L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger)\} d\boldsymbol{\theta}} + \frac{\int_{\Theta_0(r_0) \setminus U_Q(r_0)} \exp\{L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger)\} d\boldsymbol{\theta}}{\int_{U_Q(r_0)} \exp\{L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger)\} d\boldsymbol{\theta}} \\
&= I_{\Theta \setminus \Theta_0(r_0)} + I_{\Theta_0(r_0) \setminus U_Q(r_0)}.
\end{aligned}$$

The second zone is still within $\Theta_0(r_0)$ and one can use the quadratic approximation of $L_G(\boldsymbol{\theta})$. As $\|T\|_{\text{op}} = \|QD_G^{-2}Q\|_{\text{op}}^{1/2} \leq 1$, the bound $\|\boldsymbol{\xi}_G\| \leq z(B_G, \mathbf{x})$ implies $\|T\boldsymbol{\xi}_G\| \leq z(B_G, \mathbf{x})$. Therefore, on $\Omega(\mathbf{x})$

$$\begin{aligned}
\|T\mathbf{b}\| &\leq \|T\boldsymbol{\xi}_G\| + \|TD_G(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^\dagger)\| \leq z(B_G, \mathbf{x}) + \|D(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^\dagger)\| \\
&= z(B_G, \mathbf{x}) + \|DD_G^{-2}G^2\boldsymbol{\theta}^*\| = z(B_G, \mathbf{x}) + b_G.
\end{aligned}$$

By (11.47) and (11.48) and , it holds

$$\begin{aligned} I_{\Theta_0(\mathbf{r}_0) \setminus U_Q(\mathbf{r}_0)} &\leq e^{4\Delta_\circ(\mathbf{x})} \frac{\int_{\|T(\mathbf{u}-\mathbf{b})\| \geq \mathbf{r}_0} \exp\left(-\frac{\|\mathbf{u}\|^2}{2}\right) d\mathbf{u}}{\int_{\|T(\mathbf{u}-\mathbf{b})\| \leq \mathbf{r}_0} \exp\left(-\frac{\|\mathbf{u}\|^2}{2}\right) d\mathbf{u}} \\ &\leq e^{4\Delta_\circ(\mathbf{x})} \frac{\mathbb{P}^\circ(T\gamma\| \geq \mathbf{r}_0 - \|T\mathbf{b}\|)}{\mathbb{P}^\circ(T\gamma\| \leq \mathbf{r}_0 - \|T\mathbf{b}\|)}. \end{aligned}$$

The bound (11.39) on \mathbf{r}_0 yields

$$\mathbf{r}_0 - \|T\mathbf{b}\| \geq 2\sqrt{p_{Q|G} + \mathbf{x}} \geq \sqrt{p_{Q|G} + 2\mathbf{x} + 2\sqrt{\mathbf{x} p_{Q|G}}}$$

and by Corollary B.1.2

$$\mathbb{P}^\circ\left(\|T\gamma\| \geq 2\sqrt{p_{Q|G} + \mathbf{x}}\right) \leq \mathbb{P}^\circ\left(\|T\gamma\|^2 \geq p_{Q|G} + 2\mathbf{x} + 2\sqrt{\mathbf{x} p_{Q|G}}\right) \leq e^{-\mathbf{x}}.$$

Therefore,

$$I_{\Theta_0(\mathbf{r}_0) \setminus U_Q(\mathbf{r}_0)} \leq e^{4\Delta_\circ(\mathbf{x})} \frac{e^{-\mathbf{x}}}{1 - e^{-\mathbf{x}}}.$$

In the complement of $\Theta_0(\mathbf{r}_0)$, we use (11.47) and (11.49):

$$\begin{aligned} I_{\Theta \setminus \Theta_0(\mathbf{r}_0)} &\leq e^{3\Delta_\circ(\mathbf{x})} \frac{\int_{\|T(\mathbf{u}-\mathbf{b})\| > \mathbf{r}_0} \exp\left\{-C_0\mathbf{r}_0\|T(\mathbf{u}-\mathbf{b})\| + \frac{C_0\mathbf{r}_0^2}{2} + \frac{1}{2}\|T(\mathbf{u}-\mathbf{b})\|^2 - \frac{\|\mathbf{u}\|^2}{2}\right\} d\mathbf{u}}{\int_{\|T(\mathbf{u}-\mathbf{b})\| \leq \mathbf{r}_0} \exp\left(-\frac{\|\mathbf{u}\|^2}{2}\right) d\mathbf{u}}. \end{aligned}$$

By Theorem A.2.2

$$I_{\Theta \setminus \Theta_0(\mathbf{r}_0)} \leq \exp\left(3\Delta_\circ(\mathbf{x}) - \frac{p_{Q|G} + \mathbf{x}}{2}\right).$$

This implies the overall bound (11.41) on $\rho_Q(\mathbf{r}_0)$.

As the next step, we evaluate the quality of Gaussian approximation of the posterior. In view of the concentration result, it is natural to consider the centered and rescaled posterior $Q(\boldsymbol{\vartheta}_G - \check{\boldsymbol{\vartheta}}_G) | \mathbf{Y}$ with

$$\check{\boldsymbol{\vartheta}}_G = \boldsymbol{\vartheta}_G^\dagger + D_G^{-1} \nabla \zeta(\boldsymbol{\vartheta}^*) = \boldsymbol{\vartheta}_G^\dagger + D_G^{-1} \boldsymbol{\xi}_G, . \quad (11.50)$$

Similarly to the Gaussian case, one can expect that this distribution is nearly Gaussian zero mean with the covariance matrix $Q D_G^{-2} Q$.

Theorem 11.9.2. Let Q satisfy $D^2 \leq Q^2 \leq D_G^2$ and $p_{Q|G} = \text{tr}(Q D_G^{-2} Q) < \infty$. Let also the stochastic component $\zeta(\boldsymbol{\vartheta})$ of $L(\boldsymbol{\vartheta})$ be linear in $\boldsymbol{\vartheta}$ and satisfy (11.37), and let the random vector $\boldsymbol{\xi}_G$ fulfill $\|\boldsymbol{\xi}_G\| \leq z(B_G, \mathbf{x})$ on a random set $\Omega(\mathbf{x})$ with $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - e^{-\mathbf{x}}$. Suppose that it holds for a \mathbf{r}_0 fixed

$$\left| \mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \frac{1}{2}\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \right| \leq \frac{1}{2}\mathbf{r}_0^2 \delta(\mathbf{r}_0), \quad \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}_0,$$

and

$$-\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \geq \mathbf{C}_0 \mathbf{r}_0 \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| - \frac{1}{2}\mathbf{C}_0 \mathbf{r}_0^2, \quad \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| > \mathbf{r}_0,$$

with $\mathbf{C}_0 = 1 - \delta(\mathbf{r}_0)$. If \mathbf{r}_0 is large enough to ensure with b_G from (11.40)

$$\mathbf{C}_0 \mathbf{r}_0 \geq 2 \left\{ \sqrt{\mathbf{p}_{Q|G} + \mathbf{x}} + 1 + z(B_G, \mathbf{x}) + b_G \right\}, \quad (11.51)$$

then it holds on $\Omega(\mathbf{x})$ with $\check{\boldsymbol{\theta}}_G$ from (11.50) and $\Delta_o(\mathbf{x}) = \mathbf{r}_0^2 \delta(\mathbf{r}_0)/2$ for any measurable set $\mathcal{A} \subset \mathbb{R}^p$

$$\mathbb{P}(Q(\boldsymbol{\vartheta}_G - \check{\boldsymbol{\theta}}_G) \in \mathcal{A} \mid \mathbf{Y}) \geq e^{4\Delta_o(\mathbf{x})} \mathbb{P}^\circ(QD_G^{-1}\boldsymbol{\gamma} \in \mathcal{A}) - e^{-\mathbf{x}} - \rho_Q(\mathbf{x}), \quad (11.52)$$

$$\mathbb{P}(Q(\boldsymbol{\vartheta}_G - \check{\boldsymbol{\theta}}_G) \in \mathcal{A} \mid \mathbf{Y}) \leq \frac{e^{4\Delta_o(\mathbf{x})}}{1 - e^{-\mathbf{x}}} \mathbb{P}^\circ(QD_G^{-1}\boldsymbol{\gamma} \in \mathcal{A}) + \rho_Q(\mathbf{x}), \quad (11.53)$$

where $\rho_G(\mathbf{x})$ follows (11.41).

Proof. We proceed as in the case of a non-informative prior. Let the local neighborhood $U_Q(\mathbf{r}_0)$ be given by (11.35) and let \mathcal{A} be a measurable subset in $U_Q(\mathbf{r}_0)$. Then, with $\mathbf{u} = D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger) - \boldsymbol{\xi}_G$, $\mathbf{b} = D_G(\boldsymbol{\theta}_G^\dagger - \boldsymbol{\theta}^*) - \boldsymbol{\xi}_G$ and $T = QD_G^{-1}$, it holds

$$Q(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_G) = T\mathbf{u},$$

and (11.47) implies

$$\begin{aligned} & \mathbb{P}(Q(\boldsymbol{\vartheta}_G - \check{\boldsymbol{\theta}}_G) \in \mathcal{A} \mid \mathbf{Y}) \\ &= \frac{\int \mathbb{I}\{Q(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_G) \in \mathcal{A}\} \exp\{L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger)\} d\boldsymbol{\theta}}{\int \mathbb{I}\{\boldsymbol{\theta} \in U_Q(\mathbf{r}_0)\} \exp\{L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger)\} d\boldsymbol{\theta}} \\ &\geq e^{-4\Delta_o(\mathbf{x})} \frac{\int \mathbb{I}\{Q(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_G) \in \mathcal{A}\} \exp\{-\frac{1}{2}\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger) - \boldsymbol{\xi}_G\|^2\} d\boldsymbol{\theta}}{\int \mathbb{I}\{\boldsymbol{\theta} \in U_Q(\mathbf{r}_0)\} \exp\{-\frac{1}{2}\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger) - \boldsymbol{\xi}_G\|^2\} d\boldsymbol{\theta}} \\ &= e^{-4\Delta_o(\mathbf{x})} \frac{\int \mathbb{I}\{T\mathbf{u} \in \mathcal{A}\} e^{-\|\mathbf{u}\|^2/2} d\mathbf{u}}{\int \mathbb{I}\{\|T(\mathbf{u} - \mathbf{b})\| \leq \mathbf{r}_0\} e^{-\|\mathbf{u}\|^2/2} d\mathbf{u}} \\ &= e^{-4\Delta_o(\mathbf{x})} \frac{\mathbb{P}^\circ(QD_G^{-1}\boldsymbol{\gamma} \in \mathcal{A})}{\mathbb{P}^\circ(\|QD_G^{-1}(\boldsymbol{\gamma} - \mathbf{b})\| \leq \mathbf{r}_0)} \\ &\geq e^{-4\Delta_o(\mathbf{x})} \mathbb{P}^\circ(QD_G^{-1}\boldsymbol{\gamma} \in \mathcal{A}). \end{aligned}$$

For a general set \mathcal{A} , it follows on $\Omega(\mathbf{x})$

$$\begin{aligned}
& \mathbb{P}(Q(\boldsymbol{\vartheta}_G - \check{\boldsymbol{\theta}}_G) \in \mathcal{A} \mid \mathbf{Y}) \\
& \geq \mathbb{P}(Q(\boldsymbol{\vartheta}_G - \check{\boldsymbol{\theta}}_G) \in \mathcal{A} \cap U_Q(\mathbf{r}_0) \mid \mathbf{Y}) - \mathbb{P}(\Theta \setminus U_Q(\mathbf{r}_0) \mid \mathbf{Y}) \\
& \geq e^{-4\Delta_{\circ}(\mathbf{x})} \{ \mathbb{P}^{\circ}(QD_G^{-1}\boldsymbol{\gamma} \in \mathcal{A}) - \mathbb{P}^{\circ}(\|QD_G^{-1}(\boldsymbol{\gamma} - \mathbf{b})\| > \mathbf{r}_0) \} - \rho_Q(\mathbf{x}).
\end{aligned}$$

The bound (11.51) on \mathbf{r}_0 and the inequality $\|\boldsymbol{\xi}_G\| \leq z(B_G, \mathbf{x})$ ensure that

$$\mathbb{P}^{\circ}(\|QD_G^{-1}(\boldsymbol{\gamma} - \mathbf{b})\| > \mathbf{r}_0) \leq \mathbb{P}^{\circ}(\|QD_G^{-1}\boldsymbol{\gamma} - \mathbf{b}\| > \mathbf{r}_0 - b_G) \leq e^{-\mathbf{x}}$$

and (11.52) follows. To check the lower bound for the posterior probability $\mathbb{P}(Q(\boldsymbol{\vartheta}_G - \check{\boldsymbol{\theta}}_G) \in \mathcal{A} \mid \mathbf{Y})$, we proceed in a similar way. For $\mathcal{A} \subseteq U_Q(\mathbf{r}_0)$

$$\begin{aligned}
\mathbb{P}(Q(\boldsymbol{\vartheta}_G - \check{\boldsymbol{\theta}}_G) \in \mathcal{A} \mid \mathbf{Y}) & \leq e^{4\Delta_{\circ}(\mathbf{x})} \frac{\mathbb{P}^{\circ}(QD_G^{-1}\boldsymbol{\gamma} \in \mathcal{A})}{\mathbb{P}^{\circ}(\|QD_G^{-1}(\boldsymbol{\gamma} - \mathbf{b})\| \leq \mathbf{r}_0)} \\
& \leq e^{4\Delta_{\circ}(\mathbf{x})} \frac{\mathbb{P}^{\circ}(QD_G^{-1}\boldsymbol{\gamma} \in \mathcal{A})}{1 - e^{-\mathbf{x}}}
\end{aligned}$$

and for a general $\mathcal{A} \subset \mathbb{R}^p$

$$\mathbb{P}(Q(\boldsymbol{\vartheta}_G - \check{\boldsymbol{\theta}}_G) \in \mathcal{A} \mid \mathbf{Y}) \leq e^{4\Delta_{\circ}(\mathbf{x})} \frac{\mathbb{P}^{\circ}(QD_G^{-1}\boldsymbol{\gamma} \in \mathcal{A})}{1 - e^{-\mathbf{x}}} + \rho_Q(\mathbf{x}).$$

This yields (11.53).

11.9.2 General likelihood

Let the prior variance G^{-2} be fixed and $D_G^2 = D^2 + G^2$. It was already mentioned that in the case of a large or infinite parameter dimension p , the local sets $\Theta_{0,G}(\mathbf{r})$ are too small to ensure the posterior concentration. In the Gaussian case and, more generally, in the case of a linear stochastic term, the posterior well concentrates on a larger local set $U_Q(\mathbf{r})$ from (11.35) with Q satisfying (11.36). In the general case, to establish the BvM-type result, we have to ensure a reasonable approximation of the stochastic term $\zeta(\boldsymbol{\theta})$ by a linear in $\boldsymbol{\theta}$ expression. The error of linear approximation is proportional to the entropy of the local set $U_Q(\mathbf{r})$. This leads to an additional constraint on the choice of Q to ensure a small or moderate value for the corresponding entropy $\mathfrak{z}_{\mathbb{H}}(\mathbf{x})$ which is of order $\sqrt{\text{tr}(Q^{-1}V_2^2Q^{-1})}$ or, equivalently, $\sqrt{\text{tr}(Q^{-1}D^2Q^{-1})}$ in view of (T). In particular, we cannot take here $Q = D$ because the sets $\Theta_0(\mathbf{r})$ are too big in the sense that their entropy is proportional to the dimension p which can be large or equal to infinity. To state the result, consider intermediate local sets $U_Q(\mathbf{r})$ with the operator Q for which both values $\text{tr}(QD_G^{-2}Q)$ and $\text{tr}(Q^{-1}D^2Q^{-1})$ can be controlled at the same time. The result of Theorem 11.9.2 extends to the general case under this condition and the error term of the BvM approximation also depends on both of them.

Theorem 11.9.3. Let Q satisfy (11.36) and

$$\begin{aligned} p_{Q|G} &= \text{tr}(Q D_G^{-2} Q) < \infty, \\ p_{D|Q} &= \text{tr}(Q^{-1} D^2 Q^{-1}) < \infty. \end{aligned} \quad (11.54)$$

Let also the stochastic component $\zeta(\boldsymbol{\theta})$ of $L(\boldsymbol{\theta})$ satisfy **(ED₂)** and the random vector ξ_G fulfill $\|\xi_G\| \leq z(B_G, \mathbf{x})$ on a random set $\Omega(\mathbf{x})$ with $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - e^{-\mathbf{x}}$. Suppose that \mathbf{r}_0 be selected and the local neighborhood $U_Q(\mathbf{r}_0)$ is given by (11.35). Further, let the expected log-likelihood $\mathbb{E}L(\boldsymbol{\theta})$ satisfy for $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}_0$

$$\left| \mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \frac{1}{2}\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \right| \leq \frac{1}{2}\mathbf{r}_0^2 \delta(\mathbf{r}_0), \quad (11.55)$$

and for $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| > \mathbf{r}_0$, it holds

$$-\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \geq C_0 \mathbf{r}_0 \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| - \frac{C_0 \mathbf{r}_0^2}{2} + \frac{C_1}{2} \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \quad (11.56)$$

with $C_0 = 1 - \delta(\mathbf{r}_0)$ and $C_1 \geq \sup_{\mathbf{r} > \mathbf{r}_0} \varrho_Q(\mathbf{r}, \mathbf{x})$. If $C_0 + C_1 \leq 1$ and \mathbf{r}_0 is large enough to ensure

$$C_0 \mathbf{r}_0 \geq 2 \left\{ \sqrt{p_{Q|G} + \mathbf{x}} + 1 + z(B_G, \mathbf{x}) + b_G \right\} \quad (11.57)$$

with b_G from (11.40), then it holds on $\Omega(\mathbf{x})$

$$\rho_Q(\mathbf{r}_0) \stackrel{\text{def}}{=} \frac{\int_{\boldsymbol{\theta} \setminus U_Q(\mathbf{r})} \exp\{L_G(\boldsymbol{\theta})\} d\boldsymbol{\theta}}{\int_{U_Q(\mathbf{r})} \exp\{L_G(\boldsymbol{\theta})\} d\boldsymbol{\theta}} \leq e^{4\Delta_o(\mathbf{x})} \left\{ e^{-\frac{1}{2}(p_{Q|G} + \mathbf{x})} + \frac{e^{-\mathbf{x}}}{1 - e^{-\mathbf{x}}} \right\},$$

where

$$\Delta_o(\mathbf{x}) = \frac{1}{2} \{ \delta(\mathbf{r}_0) + \varrho_Q(\mathbf{r}_0, \mathbf{x}) \} \mathbf{r}_0^2.$$

Furthermore, with $\check{\boldsymbol{\theta}}_G = \boldsymbol{\theta}_G^\dagger + D_G^{-2} \nabla \zeta(\boldsymbol{\theta}^*)$, it holds for any measurable set $\mathcal{A} \subset \mathbb{R}^p$

$$\begin{aligned} \mathbb{P}\left(Q(\boldsymbol{\vartheta}_G - \check{\boldsymbol{\theta}}_G) \in \mathcal{A} \mid \mathbf{Y}\right) &\geq e^{-4\Delta_o(\mathbf{x})} \mathbb{P}^o(Q D_G^{-1} \boldsymbol{\gamma} \in \mathcal{A}) - e^{-\mathbf{x}} - \rho_Q(\mathbf{x}), \\ \mathbb{P}\left(Q(\boldsymbol{\vartheta}_G - \check{\boldsymbol{\theta}}_G) \in \mathcal{A} \mid \mathbf{Y}\right) &\leq \frac{e^{4\Delta_o(\mathbf{x})}}{1 - e^{-\mathbf{x}}} \mathbb{P}^o(Q D_G^{-1} \boldsymbol{\gamma} \in \mathcal{A}) + \rho_Q(\mathbf{x}). \end{aligned}$$

Remark 11.9.1. The results of Theorems 11.9.2 and 11.9.3 are almost identical. The only difference is in the error term $\Delta_o(\mathbf{x})$. However, it is worth stressing once again, this difference is not only due to the additional term $\varrho_Q(\mathbf{r}_0, \mathbf{x})$ which is used for describing the accuracy of the linear approximation of the stochastic component. An additional problem is in the condition (11.54) which enforces us to take Q essentially larger than D . This leads to an increase of the effective dimension $p_{Q|G} = \text{tr}(Q D_G^{-2} Q)$. This dimension

enters in the inequality (11.57) on the radius of concentration and the error terms grows linearly in r_0^2 and thus, in $p_{Q|G}$.

Proof. The proof is similar to one of Theorem 11.9.3 and we only comment on the issues specific to nonlinearity of $\zeta(\boldsymbol{\theta})$. The deviation bound of Theorem H.8.1 implies on $\Omega_Q(\mathbf{x})$

$$|\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) - (\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)^\top \nabla \zeta(\boldsymbol{\theta}^*)| \leq r_0^2 \varrho_Q(r_0, \mathbf{x}), \quad \boldsymbol{\theta} \in U_Q(r_0).$$

The decomposition $L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) = \mathbb{E} L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) + \zeta(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger)$ yields for $\boldsymbol{\theta} \in U_Q(r_0)$ in view of (11.55)

$$\begin{aligned} L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) &\geq -\frac{1}{2} \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)\|^2 + (\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)^\top \nabla \zeta(\boldsymbol{\theta}^*) - \frac{r_0^2}{2} \{\delta(r_0) + \varrho_Q(r_0, \mathbf{x})\} \\ &\geq -\frac{1}{2} \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)\|^2 + \boldsymbol{\xi}_G^\top D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger) - \Delta_o(\mathbf{x}), \\ L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) &\leq -\frac{1}{2} \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)\|^2 + \boldsymbol{\xi}_G^\top D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger) + \Delta_o(\mathbf{x}). \end{aligned}$$

Outside of $\Theta_0(r_0)$ one can use (11.45) and (11.56) to bound the expected value of the penalized log-likelihood similarly to (11.44) and (11.46) as

$$\begin{aligned} -2\mathbb{E} L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) &= -2\mathbb{E} L(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) + \|G\boldsymbol{\theta}\|^2 - \|G\boldsymbol{\theta}_G^\dagger\|^2 \\ &\geq 2C_0 r_0 \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| - (1 - C_1) \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \\ &\quad - C_0 r_0^2 + \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)\|^2 - r_0^2 \delta(r_0). \end{aligned}$$

For $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \geq r_0$, the condition $C_0 \leq 1 - C_1$ implies

$$C_0 r_0 \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| - \frac{1 - C_1}{2} \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \leq C_0 r_0 \|Q(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| - \frac{1 - C_1}{2} \|Q(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2$$

and

$$\begin{aligned} -\mathbb{E} L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) &\geq C_0 r_0 \|Q(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| - \frac{1 - C_1}{2} \|Q(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \\ &\quad - \frac{C_0 r_0^2}{2} + \frac{1}{2} \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)\|^2 - \frac{1}{2} r_0^2 \delta(r_0). \end{aligned}$$

The uniform deviation bound from Theorem H.5.1 implies on a random set $\Omega_Q(\mathbf{x})$ for all $r \geq r_0$ and all $\boldsymbol{\theta} \in \Theta$ with $r = \|Q(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|$

$$|\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) - (\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)^\top \nabla \zeta(\boldsymbol{\theta}^*)| \leq r^2 \varrho_Q(r, \mathbf{x}) \leq \frac{C_1}{2} r^2, \quad \boldsymbol{\theta} \in U_Q(r),$$

yielding on $\Omega_Q(\mathbf{x})$ for any $\boldsymbol{\theta} \in \Theta$ with $r = \|D(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)\| > r_0$

$$\begin{aligned}
L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^\dagger) &\leq -C_0 r_0 \|Q(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| + \frac{1}{2} \|Q(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + \frac{1}{2} \|Q(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \\
&\quad + \frac{C_0 r_0^2}{2} - \frac{1}{2} \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger)\|^2 + \boldsymbol{\xi}_G^\top D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^\dagger) + \frac{1}{2} r_0^2 \delta(r_0).
\end{aligned}$$

Now we can continue as in the case of linear stochastic component from Theorem 11.9.2.

12

Semiparametric estimation

This chapter specifies the general results to the case of semiparametric estimation. We consider statistical problems described by the log-likelihood of the form $L(\boldsymbol{\theta}, \boldsymbol{\phi})$, where $\boldsymbol{\theta}$ is a finite-dimensional target parameter from a parameter set $\Theta \subset \mathbb{R}^p$, while $\boldsymbol{\phi}$ is a possibly infinite dimensional nuisance parameter. The results mainly describe the inference on the target parameter $\boldsymbol{\theta}$ although one cannot avoid some pilot estimation of the nuisance $\boldsymbol{\phi}$. The main problem in the analysis is that the general approach of previous chapter involves an approximation error which polynomially grows with the total parameter dimension, therefore the obtained results are not directly applicable. This chapter discusses two possible approaches to deal with this problem. One is based on the finite-dimensional *sieve* approximation of the nuisance parameter: one approximates the infinite dimensional nuisance parameter $\boldsymbol{\phi}$ by a finite dimensional parameter $\boldsymbol{\eta}$, estimates the parameters in the obtained finite-dimensional model with parameter $\boldsymbol{v} = (\boldsymbol{\theta}, \boldsymbol{\eta})$, also evaluates the bias induced by the sieve approximation. The overall error in the sieve approach is the sum of the estimation error in the sieve model and the sieve bias.

The other approach is similar but the sieve truncation is replaced by penalization. The use of penalty allows to reduce the error of estimation: although the parameter dimension is infinite, the effective dimension becomes finite and it is in some sense equivalent to sieve dimension.

Section 12.6 presents some sufficient conditions which allow to evaluate the estimation bias induced by sieve approximation. Section 12.1 explains how the Wilks and BvM results for the sieve model can be refined if the inference is reduced to the target parameter. Finally, Section 12.4 puts together the obtained results to get the final semiparametric statements.

12.1 Fisher and Wilks results for a parameter subvector

This section consider the situation when the whole model is described via a finite dimensional parameter \boldsymbol{v} while the target of analysis is its subvector $\boldsymbol{\theta}$. Of course, the obtained general results in terms of \boldsymbol{v} can be directly applied to this situation by projecting onto the $\boldsymbol{\theta}$ -subspace. However, at some points the results can be refined. This particularly concerns the Wilks expansion: the error term there depends on the interplay between the total parameter dimension p^* and the dimension p of the target parameter $\boldsymbol{\theta}$. Let $L(\boldsymbol{v})$ be the log-likelihood for a structural parametric model with $\boldsymbol{v} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^{p^*}$. Let $\mathcal{D}^2 = -\nabla^2 L(\boldsymbol{v}^*)$ denote the full-dimensional information matrix. By $\Upsilon_o(\mathbf{s})$ we denote the elliptic local set in the full parameter space:

$$\Upsilon_o(\mathbf{s}) = \{\boldsymbol{v} : \|\mathcal{D}(\boldsymbol{v} - \boldsymbol{v}^*)\| \leq \mathbf{s}\}.$$

The key result of Proposition 9.4.2 claims a kind of local linear approximation of the gradient $\nabla L(\boldsymbol{v})$ in the vicinity $\Upsilon_o(\mathbf{s})$: on a random set $\Omega(\mathbf{x})$ with $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - e^{-x}$, it holds

$$\|\mathcal{D}^{-1}\{\nabla L(\boldsymbol{v}) - \nabla L(\boldsymbol{v}^*)\} - \mathcal{D}(\boldsymbol{v} - \boldsymbol{v}^*)\| \leq \diamond(\mathbf{s}, \mathbf{x}), \quad \boldsymbol{v} \in \Upsilon_o(\mathbf{s}), \quad (12.1)$$

where the error term $\diamond(\mathbf{s}, \mathbf{x})$ linearly grows with \mathbf{s}^2 . The choice $\mathbf{s} = \mathbf{s}_0 = C\sqrt{p^* + x}$ ensures that the full parameter MLE $\tilde{\boldsymbol{v}}$ concentrates in $\Upsilon_o(\mathbf{s}_0)$. The Fisher expansion is obtained from (12.1) under the condition $\tilde{\boldsymbol{v}} \in \Upsilon_o(\mathbf{s}_0)$ by plugging $\tilde{\boldsymbol{v}}$ in place of \boldsymbol{v} and using $\nabla L(\tilde{\boldsymbol{v}}) = 0$:

$$\|\mathcal{D}(\tilde{\boldsymbol{v}} - \boldsymbol{v}^*) - \boldsymbol{\xi}\| \leq \diamond(\mathbf{x}) = \diamond(\mathbf{s}_0, \mathbf{x}), \quad (12.2)$$

where $\boldsymbol{\xi} = \mathcal{D}^{-1}\nabla L(\boldsymbol{v}^*)$; cf. Theorem 9.3.2. This result and the probabilistic bound on the norm of $\boldsymbol{\xi}$

$$\mathbb{P}(\|\boldsymbol{\xi}\| \leq z(B, \mathbf{x})) \geq 1 - 2e^{-x}$$

with $B = \text{Var}(\boldsymbol{\xi})$ ensures that $\tilde{\boldsymbol{v}}$ lies in $\Upsilon_o(\mathbf{s}_0(\mathbf{x}))$ for

$$\mathbf{s}_0(\mathbf{x}) \stackrel{\text{def}}{=} z(B, \mathbf{x}) + \diamond(\mathbf{x})$$

on a random set $\Omega_o(\mathbf{x})$ with a probability at least $1 - 3e^{-x}$.

The Wilks expansion explains a similar error for the excess $L(\tilde{\boldsymbol{v}}) - L(\boldsymbol{v}^*)$:

$$|L(\tilde{\boldsymbol{v}}) - L(\boldsymbol{v}^*) - \|\boldsymbol{\xi}\|^2/2| \leq \mathbf{s}_0(\mathbf{x}) \diamond(\mathbf{x}) + \diamond^2(\mathbf{x})/2$$

on the same set $\Omega_0(\mathbf{x})$. Below we aim at spelling these results in the structural situation when the target dimension p is much smaller than the dimension q of the nuisance component: $p \ll q$.

Now we want to understand what these result yield for the estimator $\tilde{\boldsymbol{\theta}}$ of the target parameter $\boldsymbol{\theta}$. Consider the block representation of the matrix $\mathbb{F}(\mathbf{v}) = -\nabla^2 \mathbb{E}L(\mathbf{v})$:

$$\mathbb{F}(\mathbf{v}) = \begin{pmatrix} \mathbb{F}_{\boldsymbol{\theta}}(\mathbf{v}) & \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}(\mathbf{v}) \\ \mathbb{F}_{\boldsymbol{\eta}\boldsymbol{\theta}}(\mathbf{v}) & \mathbb{F}_{\boldsymbol{\eta}}(\mathbf{v}) \end{pmatrix}.$$

For the central point \mathbf{v}^* we write these block decomposition in the form

$$\mathcal{D}^2 = \mathbb{F}(\mathbf{v}^*) = \begin{pmatrix} D^2 & A \\ A^\top & H^2 \end{pmatrix}.$$

Also decompose the score vector ∇ as

$$\nabla = \begin{pmatrix} \nabla_{\boldsymbol{\theta}} \\ \nabla_{\boldsymbol{\eta}} \end{pmatrix}$$

and define *efficient Fisher matrix* $\mathbb{I}_{\boldsymbol{\theta}}$ and the influence vector $\check{\xi}_{\boldsymbol{\theta}}$ as

$$\begin{aligned} \mathbb{I}_{\boldsymbol{\theta}} &= D^2 - A^\top H^{-2} A, \\ \check{\xi}_{\boldsymbol{\theta}} &= \mathbb{I}_{\boldsymbol{\theta}}^{-1/2} (\nabla_{\boldsymbol{\theta}} - A H^{-2} \nabla_{\boldsymbol{\eta}}). \end{aligned} \tag{12.3}$$

12.1.1 Fisher expansion and semiparametric concentration

This section explains how the obtained Fisher expansion can be used for a significant refinement of the concentration bounds for the target parameter $\boldsymbol{\theta}$.

Theorem 12.1.1. *Let (12.2) hold on a random set $\Omega_0(\mathbf{x})$ with $\mathbb{P}(\Omega_0(\mathbf{x})) \geq 1 - e^{-\mathbf{x}}$. Let also the vector $\check{\xi}_{\boldsymbol{\theta}}$ from (12.3) fulfill on a set $\Omega_1(\mathbf{x})$ with $\mathbb{P}(\Omega_1(\mathbf{x})) \geq 1 - 2e^{-\mathbf{x}}$*

$$\|\check{\xi}_{\boldsymbol{\theta}}\| \leq z(B_{\boldsymbol{\theta}}, \mathbf{x}).$$

Then for $\mathbf{r}_0 = z(B_{\boldsymbol{\theta}}, \mathbf{x}) + \diamond(\mathbf{x})$ with $B_{\boldsymbol{\theta}} = \text{Var}(\check{\xi}_{\boldsymbol{\theta}})$, it holds on $\Omega_2(\mathbf{x})$

$$\begin{aligned} \|\mathbb{I}_{\boldsymbol{\theta}}^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \check{\xi}_{\boldsymbol{\theta}}\| &\leq \diamond(\mathbf{x}), \\ \|\mathbb{I}_{\boldsymbol{\theta}}^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| &\leq \mathbf{r}_0 = z(B_{\boldsymbol{\theta}}, \mathbf{x}) + \diamond(\mathbf{x}). \end{aligned}$$

Proof. First we consider the orthogonal case with $A = 0$ and $\mathbb{F} = \text{block}\{D^2, H^2\}$. Then

$$\boldsymbol{\xi} = \mathcal{D}^{-1} \nabla = \begin{pmatrix} D^{-1} & 0 \\ 0 & H^{-1} \end{pmatrix} \begin{pmatrix} \nabla_{\boldsymbol{\theta}} \\ \nabla_{\boldsymbol{\eta}} \end{pmatrix} = \begin{pmatrix} D^{-1} \nabla_{\boldsymbol{\theta}} \\ H^{-1} \nabla_{\boldsymbol{\eta}} \end{pmatrix} = \begin{pmatrix} \xi_{\boldsymbol{\theta}} \\ \xi_{\boldsymbol{\eta}} \end{pmatrix}$$

and $\check{\xi}_{\boldsymbol{\theta}} = \xi_{\boldsymbol{\theta}}$. Semiparametric Fisher expansion in the orthogonal case reads as

$$\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \xi_{\boldsymbol{\theta}}\|^2 + \|H(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) - \xi_{\boldsymbol{\eta}}\|^2 \leq \diamond(x) \quad (12.4)$$

with $\diamond(x) = \diamond(s_0, x)$ yielding

$$\begin{aligned} \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \xi_{\boldsymbol{\theta}}\| &\leq \diamond(x), \\ \|H(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) - \xi_{\boldsymbol{\eta}}\| &\leq \diamond(x). \end{aligned} \quad (12.5)$$

These bounds hold on a set $\Omega_0(x)$ with $\text{IP}(\Omega_0(x)) \geq 1 - e^{-x}$. Furthermore, it can be combined with the probabilistic bounds

$$\begin{aligned} \|\xi_{\boldsymbol{\theta}}\| &\leq z(B_{\boldsymbol{\theta}}, x), \\ \|\xi_{\boldsymbol{\eta}}\| &\leq z(B_{\boldsymbol{\eta}}, x), \end{aligned} \quad (12.6)$$

where $B_{\boldsymbol{\theta}} = \text{Var}(\xi_{\boldsymbol{\theta}})$, $B_{\boldsymbol{\eta}} = \text{Var}(\xi_{\boldsymbol{\eta}})$. Again, these bounds hold on a random set $\Omega_1(x)$ with $\text{IP}(\Omega_1(x)) \geq 1 - e^{-x}$. On the overlap $\Omega_2(x) = \Omega_0(x) \cap \Omega_1(x)$, we obtain the deviation bound

$$\begin{aligned} \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| &\leq z(B_{\boldsymbol{\theta}}, x) + \diamond(x), \\ \|H(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)\| &\leq z(B_{\boldsymbol{\eta}}, x) + \diamond(x). \end{aligned}$$

In particular, restricting to $\Omega_2(x)$, this implies that the target component $\tilde{\boldsymbol{\theta}}$ well concentrates on the set $\Theta_0(r_0) = \{\boldsymbol{\theta}: \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq r_0\}$ for $r_0 = z(B_{\boldsymbol{\theta}}, x) + \diamond(x)$:

$$\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \leq r_0 = z(B_{\boldsymbol{\theta}}, x) + \diamond(x). \quad (12.7)$$

Similarly for $\tilde{\boldsymbol{\eta}}$ with $h_0 = z(B_{\boldsymbol{\eta}}, x) + \diamond(x)$:

$$\|H(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)\| \leq h_0 = z(B_{\boldsymbol{\eta}}, x) + \diamond(x).$$

The result can be accomplished by a similar concentration for the partial nuisance estimator $\tilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*)$ obtained by maximization of $L(\boldsymbol{\theta}^*, \boldsymbol{\eta})$ w.r.t. $\boldsymbol{\eta}$:

$$\tilde{\boldsymbol{\eta}}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \underset{\boldsymbol{\eta}}{\operatorname{argmax}} L(\boldsymbol{\theta}, \boldsymbol{\eta})$$

It holds on the same random set $\Omega_2(x)$

$$\|H(\tilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*) - \boldsymbol{\eta}^*)\| \leq h_0. \quad (12.8)$$

As usual, the general situation can be reduced to the orthogonal one by a linear change of variable with the new nuisance parameter $\check{\boldsymbol{\eta}} \stackrel{\text{def}}{=} \boldsymbol{\eta} - H^{-2}A^\top\boldsymbol{\theta}$ and the efficient information matrix $\mathbb{I}_{\boldsymbol{\theta}}$ from (12.3). Then

$$\|\mathcal{D}(\boldsymbol{v} - \boldsymbol{v}^*)\|^2 = \|\mathbb{I}_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + \|H(\tilde{\boldsymbol{\eta}} - \tilde{\boldsymbol{\eta}}^*)\|^2,$$

leading back to the orthogonal case.

12.1.2 Semiparametric Wilks expansion

Now we consider the semiparametric excess $\check{L}(\tilde{\boldsymbol{\theta}}) - \check{L}(\boldsymbol{\theta}^*)$ for

$$\check{L}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sup_{\boldsymbol{\eta}} L(\boldsymbol{\theta}, \boldsymbol{\eta}). \quad (12.9)$$

Theorem 12.1.2. *Let (12.4) and (12.6) hold on a random set $\Omega_2(\mathbf{x})$ with $\mathbb{P}(\Omega_2(\mathbf{x})) \geq 1 - 2e^{-x}$ and r_0 be given by (12.7). Then the profile MLE $\tilde{\boldsymbol{\theta}}$ and partial maximum likelihood $\check{L}(\boldsymbol{\theta}) = \max_{\boldsymbol{\eta}} L(\boldsymbol{\theta}, \boldsymbol{\eta})$ satisfy on $\Omega_2(\mathbf{x})$*

$$\|\mathbb{I}_{\boldsymbol{\theta}}^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \check{\boldsymbol{\xi}}_{\boldsymbol{\theta}}\| \leq \diamond(x), \quad (12.10)$$

$$\begin{aligned} |\check{L}(\tilde{\boldsymbol{\theta}}) - \check{L}(\boldsymbol{\theta}^*) - \|\check{\boldsymbol{\xi}}_{\boldsymbol{\theta}}\|^2/2| &\leq r_0 \diamond(x) + \diamond^2(x)/2, \\ \left| \sqrt{2\check{L}(\tilde{\boldsymbol{\theta}}) - 2\check{L}(\boldsymbol{\theta}^*)} - \|\check{\boldsymbol{\xi}}_{\boldsymbol{\theta}}\| \right| &\leq 7\diamond(x). \end{aligned} \quad (12.11)$$

Moreover, for any point $\boldsymbol{\theta}^\circ \in \Theta_0(r_0)$, it holds with $\mathbf{b}_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \mathbb{I}_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*)$

$$\begin{aligned} \|\mathbb{I}_{\boldsymbol{\theta}}^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ) - \check{\boldsymbol{\xi}}_{\boldsymbol{\theta}} - \mathbf{b}_{\boldsymbol{\theta}}\| &\leq \diamond(x), \\ \left| \sqrt{2\check{L}(\tilde{\boldsymbol{\theta}}) - 2\check{L}(\boldsymbol{\theta}^\circ)} - \|\check{\boldsymbol{\xi}}_{\boldsymbol{\theta}} + \mathbf{b}_{\boldsymbol{\theta}}\| \right| &\leq 9\diamond(x). \end{aligned} \quad (12.12)$$

Proof. The concentration results (12.7) and (12.8) allow to reduce the consideration to the case $\tilde{\boldsymbol{\theta}} \in \Theta_0(r_0)$, $\tilde{\boldsymbol{\eta}} \in \mathcal{H}_0(h_0)$, and $\tilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*) \in \mathcal{H}_0(h_0)$.

Consider first the orthogonal case with $\mathcal{D}^2 = \text{block}\{D^2, H^2\}$ and $\mathbb{I}_{\boldsymbol{\theta}} = D^2$. For any $\boldsymbol{v} = (\boldsymbol{\theta}, \boldsymbol{\eta})$ with $\boldsymbol{\theta} \in \Theta_0(r_0)$ and $\boldsymbol{\eta} \in \mathcal{H}_0(h_0)$, it holds by Proposition 9.4.2

$$\begin{aligned} &|L(\boldsymbol{\theta}, \boldsymbol{\eta}) - L(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \{D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}^\top \check{\boldsymbol{\xi}}_{\boldsymbol{\theta}} + \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2| \\ &\leq \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \diamond(x) \leq r_0 \diamond(x). \end{aligned} \quad (12.13)$$

One can see from this bound that the difference $L(\boldsymbol{\theta}, \boldsymbol{\eta}) - L(\boldsymbol{\theta}^*, \boldsymbol{\eta})$ can be uniformly in $\boldsymbol{\eta} \in \mathcal{H}_0(h_0)$ approximated by the quadratic in $\boldsymbol{\theta} - \boldsymbol{\theta}^*$ expression $D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \check{\boldsymbol{\xi}}_{\boldsymbol{\theta}} - \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2$ and the error term is only of order $r_0 \diamond(x)$. Note that for the difference $L(\boldsymbol{\theta}, \boldsymbol{\eta}) - L(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$, the error of approximation is of order $\|\mathcal{D}(\boldsymbol{v} - \boldsymbol{v}^*)\| \diamond(x) \leq s_0 \diamond(x)$.

The definition of $\tilde{\boldsymbol{v}}$ and of $\tilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*)$ implies in view of $\tilde{\boldsymbol{\eta}} \in \mathcal{H}_0(h_0)$ and $\tilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*) \in \mathcal{H}_0(h_0)$

$$\check{L}(\tilde{\boldsymbol{\theta}}) - \check{L}(\boldsymbol{\theta}^*) = L(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}) - L(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*)) \leq L(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}) - L(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\eta}}),$$

$$\check{L}(\tilde{\boldsymbol{\theta}}) - \check{L}(\boldsymbol{\theta}^*) = L(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}) - L(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*)) \geq L(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*)) - L(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*)).$$

Application of (12.13) two times with $\tilde{\boldsymbol{\theta}}$ in place of $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\eta}}$ or $\tilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^*)$ in place of $\boldsymbol{\eta}$ yields for $\tilde{\mathbf{u}} = D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$

$$|\check{L}(\tilde{\boldsymbol{\theta}}) - \check{L}(\boldsymbol{\theta}^*) - \tilde{\mathbf{u}}^\top \xi_{\boldsymbol{\theta}} + \|\tilde{\mathbf{u}}\|^2/2| \leq 2\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \diamond(x) \leq 2r_0 \diamond(x). \quad (12.14)$$

Together with (12.5) this implies by $-2\tilde{\mathbf{u}}^\top \xi_{\boldsymbol{\theta}} + \|\tilde{\mathbf{u}}\|^2 = -\|\xi_{\boldsymbol{\theta}}\|^2 + \|\tilde{\mathbf{u}} - \xi_{\boldsymbol{\theta}}\|^2$

$$|\check{L}(\tilde{\boldsymbol{\theta}}) - \check{L}(\boldsymbol{\theta}^*) - \|\xi_{\boldsymbol{\theta}}\|^2/2| \leq 2r_0 \diamond(x) + \diamond^2(x)/2.$$

The square-root expansion can be obtained in a similar way. Indeed, (12.14) and (12.10) imply

$$\begin{aligned} & \left| \sqrt{2\check{L}(\tilde{\boldsymbol{\theta}}) - 2\check{L}(\boldsymbol{\theta}^*)} - \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \right| \\ & \leq \frac{|2\check{L}(\tilde{\boldsymbol{\theta}}) - 2\check{L}(\boldsymbol{\theta}^*) - \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2|}{\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|} \leq \frac{|2\tilde{\mathbf{u}}^\top \xi_{\boldsymbol{\theta}} - 2\|\tilde{\mathbf{u}}\|^2| + 4\|\tilde{\mathbf{u}}\| \diamond(x)}{\|\tilde{\mathbf{u}}\|} \\ & \leq 2\|\tilde{\mathbf{u}} - \xi_{\boldsymbol{\theta}}\| + 4\diamond(x) \leq 6\diamond(x). \end{aligned}$$

One more application of (12.10) yields (12.11). If $\boldsymbol{\theta}^*$ is replaced by any other point $\boldsymbol{\theta}^\circ$ from $\Theta_0(r_0)$, all the bounds continue to apply with an obvious correction in (12.13) that $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\| \leq 2r_0$.

Change of variable helps to reduce everything to the orthogonal situation. As previously, one obtains the expansion for $L(\boldsymbol{\theta}, \check{\boldsymbol{\eta}}) = L(\boldsymbol{\theta}, \boldsymbol{\eta} + H^{-2}A^\top \boldsymbol{\theta})$ in the form

$$|L(\boldsymbol{\theta}, \check{\boldsymbol{\eta}}) - L(\boldsymbol{\theta}^*, \check{\boldsymbol{\eta}}) - \mathbf{u}^\top \check{\xi}_{\boldsymbol{\theta}} + \|\mathbf{u}\|^2/2| \leq \|\mathbf{u}\| \diamond(x),$$

where $\mathbf{u} = \mathbb{I}_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ and $\check{\xi}_{\boldsymbol{\theta}}$ from (12.3). Now we can continue as in the orthogonal case.

12.2 Likelihood ratio test statistic for a composite hypothesis

This section provides an expansion for the likelihood ratio (LR) test statistic for a composite null hypothesis. Let the model be described by a composite parameter $(\boldsymbol{\theta}, \boldsymbol{\eta})$ and $L(\boldsymbol{\theta}, \boldsymbol{\eta})$ denote the corresponding log-likelihood function. We consider a testing problem when the null hypothesis only concerns the parameter $\boldsymbol{\theta}$, namely, $H_0: \boldsymbol{\theta} = \boldsymbol{\theta}_0$ for a given point $\boldsymbol{\theta}_0$. The related LR test statistics reads as

$$T \stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta}, \boldsymbol{\eta}} L(\boldsymbol{\theta}, \boldsymbol{\eta}) - \sup_{\boldsymbol{\eta}} L(\boldsymbol{\theta}_0, \boldsymbol{\eta}). \quad (12.15)$$

In the classical statistical theory, the prominent Wlks phenomenon claims that $2T$ is under the null hypothesis nearly χ_p^2 random variable with p degrees of freedom, where

p is dimension of $\boldsymbol{\theta}$. The result below extends the Wilks result to the finite sample setup. We show that $\sqrt{2T}$ can be well approximated by $\|\check{\xi}_{\boldsymbol{\theta}} + \mathbf{b}_{\boldsymbol{\theta}}\|$, where $\check{\xi}_{\boldsymbol{\theta}}$ is the a zero mean random vector and $\mathbf{b}_{\boldsymbol{\theta}}$ is a deterministic non-centrality vector which disappears under the null. The semiparametric Wilks expansion from Theorem 12.1.2 helps to explicitly describe the vectors $\check{\xi}_{\boldsymbol{\theta}}$ and $\mathbf{b}_{\boldsymbol{\theta}}$ and also bounds the accuracy of approximation.

First we state the result under the null, that is, when $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^*$. This result is, in fact, a reformulation of (12.11), because the LR test statistic T coincides with the semiparametric excess $\check{L}(\tilde{\boldsymbol{\theta}}) - \check{L}(\boldsymbol{\theta}^*)$; cf. the definitions (12.15) and (12.9).

Theorem 12.2.1. *Let (12.4) and (12.6) hold on a random set $\Omega_2(\mathbf{x})$ with $\mathbb{I}\mathcal{P}(\Omega_2(\mathbf{x})) \geq 1 - 2e^{-\mathbf{x}}$ and \mathbf{r}_0 be given by (12.7). If $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^*$, then the LR test statistic T from (12.15) fulfills on $\Omega_2(\mathbf{x})$*

$$\left| \sqrt{2T} - \|\check{\xi}_{\boldsymbol{\theta}}\| \right| \leq 7\Diamond(\mathbf{x}).$$

Now we extend the result to the case when $\boldsymbol{\theta}_0$ deviates from $\boldsymbol{\theta}^*$. Again, this is a reformulation of the result from Theorem 12.1.2; see (12.12).

Theorem 12.2.2. *Let (12.4) and (12.6) hold on a random set $\Omega_2(\mathbf{x})$ with $\mathbb{I}\mathcal{P}(\Omega_2(\mathbf{x})) \geq 1 - 2e^{-\mathbf{x}}$ and \mathbf{r}_0 be given by (12.7). If $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}^*$, then the LR test statistic T from (12.15) fulfills on $\Omega_2(\mathbf{x})$*

$$\left| \sqrt{2T} - \|\check{\xi}_{\boldsymbol{\theta}} + \mathbf{b}_{\boldsymbol{\theta}}\| \right| \leq 9\Diamond(\mathbf{x}),$$

where

$$\mathbf{b}_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \mathbb{I}_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*).$$

The obtained bounds can be used to describe the power of the LR test. The level result suggests to select the critical value \mathfrak{z}_α to ensure

$$\mathbb{I}\mathcal{P}(\|\check{\xi}_{\boldsymbol{\theta}}\| > \mathfrak{z}_\alpha) \leq \alpha.$$

Theorem 12.2.1 implies the probability bound

$$\mathbb{I}\mathcal{P}\left(\sqrt{2T} > \mathfrak{z}_\alpha + 7\Diamond(\mathbf{x})\right) \leq \alpha + 2e^{-\mathbf{x}}.$$

If the error term $7\Diamond(\mathbf{x})$ is negligible as well as the probability $e^{-\mathbf{x}}$, then we can expect a nearly exact level of the LR test $\tau = \mathbb{I}(T > \mathfrak{z}_\alpha^2/2)$. Further, the test τ is powerful if the deviation probability

$$\mathbb{I}\mathcal{P}(\|\check{\xi}_{\boldsymbol{\theta}} + \mathbf{b}_{\boldsymbol{\theta}}\| > \mathfrak{z}_\alpha)$$

is significantly positive. In particular, if $\|\mathbf{b}_\theta\| > 2\mathfrak{z}_\alpha$, then

$$\mathbb{P}(\|\check{\xi}_\theta + \mathbf{b}_\theta\| > \mathfrak{z}_\alpha) \geq 1 - \mathbb{P}(\|\check{\xi}_\theta\| > \mathfrak{z}_\alpha) \gtrsim 1 - \alpha.$$

If, in addition, one can show asymptotic normality of the vector $\check{\xi}_\theta$, the power result can be refined using non-central Gaussian probability bounds.

12.3 Semiparametric BvM approximation

This section discusses the semiparametric Bernstein - von Mises (BvM) result which describes the properties of the posterior distribution of the parameter \mathbf{v} . This distribution is a random measure on Υ given the data \mathbf{Y} and a prior Π distribution on Υ . Here we only check the case of an uninformative prior with the constant density $\Pi \equiv 1$.

The general approach to show the BvM result for the full parameter involves two big steps: posterior contraction and local quadratic approximation. The first step is to show the contraction property for the local elliptic set $\Upsilon_\circ(\mathbf{s}_0)$ for $\mathbf{s}_0^2 = \mathbf{C}(p^* + \mathbf{x})$: the posterior mass of the complement $\Upsilon \setminus \Upsilon_\circ(\mathbf{s}_0)$ is exponentially small. The second step is based on the quadratic expansion of the log-likelihood in this local vicinity with the accuracy $\diamond(\mathbf{x}) = \diamond(\mathbf{s}_0, \mathbf{x})$. However, this expansion yields the error term of order $\mathbf{s}_0 \diamond(\mathbf{s}_0, \mathbf{x})$ which is in its turn of order $\sqrt{(p^* + \mathbf{x})^3/n}$ in the i.i.d. case. If we are interested in the marginal distribution of the posterior for the target parameter, the contraction result and the error term can be refined. Namely, we show that the marginal posterior concentrates on the set $\Theta_0(\mathbf{r}_0)$ for $\mathbf{r}_0^2 \asymp p + \mathbf{x}$ and the corresponding error term in the BvM approximation is of order $\mathbf{r}_0 \diamond(\mathbf{x}) \asymp \sqrt{p + \mathbf{x}} (p^* + \mathbf{x})/\sqrt{n}$ which can be a significant improvement relative to $\sqrt{(p^* + \mathbf{x})^3/n}$.

Below we operate with integrals of the likelihood $\exp\{L(\mathbf{v})\}$ over rectangle local sets $\Upsilon(\mathbf{r}, \mathbf{h}) = \Theta_0(\mathbf{r}) \times \mathcal{H}_0(\mathbf{h})$. Isotropic rectangles with $\mathbf{r} = \mathbf{h}$ have nearly the same geometry as the elliptic sets $\Upsilon_\circ(\mathbf{s})$. More precisely, with $\mathbf{s}^2 = 2\mathbf{h}^2$, the following inclusions hold:

$$\Upsilon(\mathbf{h}, \mathbf{h}) \subset \Upsilon_\circ(\mathbf{s}) \subset \Upsilon(\mathbf{s}, \mathbf{s}).$$

Therefore, the contraction and the local approximation results can be restated for the isotropic sets $\Upsilon(\mathbf{h}, \mathbf{h})$. We assume that for a fixed $\mathbf{h}_0 = \mathbf{h}_0(\mathbf{x})$, it holds on a random set $\Omega(\mathbf{x})$ with $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - e^{-x}$

$$\int_{\Upsilon \setminus \Upsilon(\mathbf{h}_0, \mathbf{h}_0)} e^{L(\mathbf{v})} d\mathbf{v} \leq \rho(\mathbf{x}) \int_{\Upsilon(\mathbf{h}_0, \mathbf{h}_0)} e^{L(\mathbf{v})} d\mathbf{v} \quad (12.16)$$

for some small value $\rho(\mathbf{x})$, and

$$\|\mathcal{D}^{-1}\{\nabla L(\mathbf{v}) - \nabla L(\mathbf{v}^*)\} - \mathcal{D}(\mathbf{v} - \mathbf{v}^*)\| \leq \diamond(\mathbf{x}), \quad \mathbf{v} \in \Upsilon(\mathbf{h}_0, \mathbf{h}_0), \quad (12.17)$$

where $\diamond(x) = \diamond(h_0, x)$. Below we just restrict ourselves to this random set $\Omega(x)$ and assume the conditions (12.16) and (12.17).

Remind the notation E° for the conditional expectation and γ for a standard normal vector in \mathbb{R}^p given \mathbf{Y} . The next result describes the lower and upper bounds for conditional integrals $E^\circ f(\boldsymbol{\vartheta}) = E f(\boldsymbol{\vartheta} | \mathbf{Y})$, where $\boldsymbol{\vartheta}$ w.r.t. the posterior distribution.

Theorem 12.3.1 (BvM for a subvector). *Assume (12.16) and (12.17). For any non-negative bounded function $f(\mathbf{u})$ on \mathbb{R}^p with $\|f\|_\infty < \infty$, it holds on $\Omega(x)$ for $r_0 \leq h_0$*

$$\begin{aligned} E^\circ f(\mathbb{I}_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{\vartheta} - \boldsymbol{\theta}^*)) \\ \leq \|f\|_\infty \{ \rho(x) + \rho_0(x) \} + \exp\{2r_0 \diamond(x)\} E^\circ \left\{ f(\gamma + \check{\xi}_{\boldsymbol{\theta}}) \mid \|\gamma + \check{\xi}_{\boldsymbol{\theta}}\| \leq r_0 \right\}. \end{aligned}$$

and

$$\begin{aligned} E^\circ f(\mathbb{I}_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{\vartheta} - \boldsymbol{\theta}^*)) \\ \geq \exp\{-2r_0 \diamond(x) - \rho(x) - \rho_0(x)\} E^\circ \left\{ f(\gamma + \check{\xi}_{\boldsymbol{\theta}}) \mid \|\gamma + \check{\xi}_{\boldsymbol{\theta}}\| \leq r_0 \right\} \end{aligned}$$

with

$$\rho_0(x) \stackrel{\text{def}}{=} 5e^{2r_0 \diamond(x)-x}.$$

Moreover, if $r_0 \geq z(B_{\boldsymbol{\theta}}, x) + \sqrt{p} + \sqrt{2x}$ then

$$E^\circ \left\{ f(\gamma + \check{\xi}_{\boldsymbol{\theta}}) \mid \|\gamma + \check{\xi}_{\boldsymbol{\theta}}\| \leq r_0 \right\} \leq \frac{1}{1 - e^{-x}} E^\circ \left\{ f(\gamma + \check{\xi}_{\boldsymbol{\theta}}) \right\}.$$

Proof. Define the local posterior mass

$$\Pi(r, h) \stackrel{\text{def}}{=} \int_{\Upsilon(r, h)} \exp\{L(\mathbf{v})\} d\mathbf{v}.$$

Let $f(\mathbf{u})$ be any measurable function of $\mathbf{u} \in \mathbb{R}^p$. We aim at evaluating the ratio

$$\frac{1}{\Pi(r, h)} \int_{\Upsilon(r, h)} \exp\{L(\mathbf{v})\} f(\mathbb{I}_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{\vartheta} - \boldsymbol{\theta}^*)) d\mathbf{v}.$$

Lemma 12.3.1. *Assume (12.16) and (12.17). Then for any r and h with $r, h \leq h_0$, and measurable function $f(\mathbf{u})$ on \mathbb{R}^p , it holds on $\Omega(x)$*

$$\begin{aligned} & \exp\{-2r \diamond(x)\} E^\circ \left\{ f(\gamma + \check{\xi}_{\boldsymbol{\theta}}) \mid \|\gamma + \check{\xi}_{\boldsymbol{\theta}}\| \leq r \right\} \\ & \leq \frac{\int_{\Upsilon(r, h)} \exp\{L(\mathbf{v})\} f(\mathbb{I}_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{\vartheta} - \boldsymbol{\theta}^*)) d\mathbf{v}}{\int_{\Upsilon(r, h)} \exp\{L(\mathbf{v})\} d\mathbf{v}} \\ & \leq \exp\{2r \diamond(x)\} E^\circ \left\{ f(\gamma + \check{\xi}_{\boldsymbol{\theta}}) \mid \|\gamma + \check{\xi}_{\boldsymbol{\theta}}\| \leq r \right\}. \end{aligned}$$

Proof. It suffices to prove the statement for the orthogonal case $\mathbb{F}_{\theta\eta} = 0$, $\mathbb{I}_\theta = \mathbb{F}_\theta$, and $\check{\xi}_\theta = \xi_\theta$. Represent

$$L(\mathbf{v}) - L(\mathbf{v}^*) = L(\boldsymbol{\theta}, \boldsymbol{\eta}) - L(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*) = L(\boldsymbol{\theta}, \boldsymbol{\eta}) - L(\boldsymbol{\theta}^*, \boldsymbol{\eta}) + L(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - L(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$$

and use the approximation (12.13) for $L(\boldsymbol{\theta}, \boldsymbol{\eta}) - L(\boldsymbol{\theta}^*, \boldsymbol{\eta})$ which implies

$$\begin{aligned} L(\mathbf{v}) - L(\mathbf{v}^*) &\leq L(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - L(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*) + \mathbf{u}^\top \xi_\theta - \|\mathbf{u}\|^2/2 + \mathbf{r} \diamond (\mathbf{x}), \\ L(\mathbf{v}) - L(\mathbf{v}^*) &\geq L(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - L(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*) + \mathbf{u}^\top \xi_\theta - \|\mathbf{u}\|^2/2 - \mathbf{r} \diamond (\mathbf{x}), \end{aligned}$$

with $\mathbf{u} = D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ and $\diamond(\mathbf{x}) = \diamond(h_0, \mathbf{x})$. Fubini integration leads to the bounds

$$\begin{aligned} &\int_{\Upsilon(\mathbf{r}, h)} \exp\{L(\mathbf{v}) - L(\mathbf{v}^*)\} d\mathbf{v} \\ &\leq \int_{\Upsilon(\mathbf{r}, h)} \exp\{L(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - L(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*) + \mathbf{u}^\top \xi_\theta - \|\mathbf{u}\|^2/2 + \mathbf{r} \diamond (\mathbf{x})\} d\mathbf{v} \\ &\leq \exp\{\mathbf{r} \diamond (\mathbf{x})\} \det(\mathbb{F}_\theta^{-1/2}) \Pi_1(h) \int_{\|\mathbf{u}\| \leq \mathbf{r}} \exp\{\mathbf{u}^\top \xi_\theta - \|\mathbf{u}\|^2/2\} d\mathbf{u} \end{aligned}$$

with

$$\Pi_1(h) \stackrel{\text{def}}{=} \int_{\mathcal{H}_0(h)} \exp\{L(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - L(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)\} d\boldsymbol{\eta}.$$

Similarly, for any measurable function $f(\mathbf{u})$

$$\begin{aligned} &\int_{\Upsilon(\mathbf{r}, h)} \exp\{L(\mathbf{v}) - L(\mathbf{v}^*)\} f(D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)) d\mathbf{v} \\ &\geq \exp\{\mathbf{r} \diamond (\mathbf{x})\} \det(\mathbb{F}_\theta^{-1/2}) \Pi_1(h) \int_{\|\mathbf{u}\| \leq \mathbf{r}} \exp\{\mathbf{u}^\top \xi_\theta - \|\mathbf{u}\|^2/2\} f(\mathbf{u}) d\mathbf{u}. \end{aligned}$$

Therefore

$$\begin{aligned} &\frac{\int_{\Upsilon(\mathbf{r}, h)} \exp\{L(\mathbf{v})\} f(D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)) d\mathbf{v}}{\int_{\Upsilon(\mathbf{r}, h)} \exp\{L(\mathbf{v})\} d\mathbf{v}} \\ &\leq \exp\{2\mathbf{r} \diamond (\mathbf{x})\} \frac{\int_{\|\mathbf{u}\| \leq \mathbf{r}} \exp\{\mathbf{u}^\top \xi_\theta - \|\mathbf{u}\|^2/2\} f(\mathbf{u}) d\mathbf{u}}{\int_{\|\mathbf{u}\| \leq \mathbf{r}} \exp\{\mathbf{u}^\top \xi_\theta - \|\mathbf{u}\|^2/2\} d\mathbf{u}} \\ &= \exp\{2\mathbf{r} \diamond (\mathbf{x})\} \frac{\int_{\|\mathbf{u}\| \leq \mathbf{r}} \exp\{-\|\mathbf{u} - \xi_\theta\|^2/2\} f(\mathbf{u}) d\mathbf{u}}{\int_{\|\mathbf{u}\| \leq \mathbf{r}} \exp\{-\|\mathbf{u} - \xi_\theta\|^2/2\} d\mathbf{u}} \\ &= \exp\{2\mathbf{r} \diamond (\mathbf{x})\} \mathbb{E}^\circ \left\{ f(\boldsymbol{\gamma} + \xi_\theta) \mid \|\boldsymbol{\gamma} + \xi_\theta\| \leq \mathbf{r} \right\}. \end{aligned} \tag{12.18}$$

Similarly one can prove the bound from below.

Moreover, conditional expectation can be changed to unconditional one if the value \mathbf{r} is sufficiently large. Indeed, the probabilistic bound $\mathbb{P}\{\|\boldsymbol{\xi}_{\boldsymbol{\theta}}\| > z(B_{\boldsymbol{\theta}}, \mathbf{x})\} \leq e^{-\mathbf{x}}$ and the choice $\mathbf{r} = \mathbf{r}_0 = z(B_{\boldsymbol{\theta}}, \mathbf{x}) + \sqrt{p} + \sqrt{2\mathbf{x}}$ ensures

$$\mathbb{P}(\|\boldsymbol{\gamma} + \boldsymbol{\xi}_{\boldsymbol{\theta}}\| \leq \mathbf{r}_0) \geq \mathbb{P}(\|\boldsymbol{\gamma}\| \leq \sqrt{p} + \sqrt{2\mathbf{x}}) \geq 1 - e^{-\mathbf{x}}. \quad (12.19)$$

Now we show that the posterior measure concentrates on the anisotropic rectangle set $\Upsilon(\mathbf{r}_0, \mathbf{h}_0) = \Theta_0(\mathbf{r}_0) \times \mathcal{H}_0(\mathbf{h}_0)$ which can be essentially smaller than the isotropic rectangle set $\Upsilon(\mathbf{h}_0, \mathbf{h}_0)$.

Lemma 12.3.2. *Assume (12.17). If $\mathbf{r}_0 \geq z(B_{\boldsymbol{\theta}}, \mathbf{x}) + \sqrt{p} + \sqrt{2\mathbf{x}}$, then with $U_0 = \Theta_0(\mathbf{r}_0) \times \mathcal{H}_0(\mathbf{h}_0)$, it holds for $\diamond(\mathbf{x}) = \diamond(\mathbf{h}_0, \mathbf{x})$ on the set $\Omega(\mathbf{x})$*

$$\begin{aligned} \int_{\Upsilon(\mathbf{h}_0, \mathbf{h}_0)} e^{L(\boldsymbol{v})} d\boldsymbol{v} &\leq (1 + \rho_0(\mathbf{x})) \int_{U_0} e^{L(\boldsymbol{v})} d\boldsymbol{v}, \\ \rho_0(\mathbf{x}) &\stackrel{\text{def}}{=} 5e^{2\mathbf{r}_0 \diamond(\mathbf{x}) - \mathbf{x}}. \end{aligned} \quad (12.20)$$

Proof. Let $\mathbf{r}_0 < \mathbf{r}_1 \dots < \mathbf{r}_K = \mathbf{h}_0$ be a growing sequence. Define $U_k = \Theta_0(\mathbf{r}_k) \times \mathcal{H}_0(\mathbf{h}_0)$ and apply the bound (12.18) to evaluate the integral over the rectangle set $U_k \setminus U_{k-1}$. More specifically, with $\Pi_K = \Pi(\mathbf{r}_K, \mathbf{h}_0) = \Pi(\mathbf{h}_0, \mathbf{h}_0)$

$$\begin{aligned} \frac{1}{\Pi_K} \int_{U_K \setminus U_0} e^{L(\boldsymbol{v})} d\boldsymbol{v} &= \sum_{k=1}^K \frac{1}{\Pi_K} \int_{U_k \setminus U_{k-1}} e^{L(\boldsymbol{v})} d\boldsymbol{v} \\ &\leq \sum_{k=1}^K \frac{\int_{U_k} e^{L(\boldsymbol{v})} \mathbb{I}(\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| > \mathbf{r}_{k-1}) d\boldsymbol{v}}{\int_{U_k} e^{L(\boldsymbol{v})} d\boldsymbol{v}} \\ &\leq \sum_{k=1}^K e^{2\mathbf{r}_k \diamond(\mathbf{x})} \mathbb{P}(\|\boldsymbol{\gamma} + \boldsymbol{\xi}_{\boldsymbol{\theta}}\| > \mathbf{r}_{k-1} \mid \|\boldsymbol{\gamma} + \boldsymbol{\xi}_{\boldsymbol{\theta}}\| \leq \mathbf{r}_k). \end{aligned}$$

The choice $\mathbf{r}_0 = z(B_{\boldsymbol{\theta}}, \mathbf{x}) + \sqrt{p} + \sqrt{2\mathbf{x}}$ ensures by (12.19)

$$\mathbb{P}(\|\boldsymbol{\gamma} + \boldsymbol{\xi}_{\boldsymbol{\theta}}\| \leq \mathbf{r}_0) \geq 1 - e^{-\mathbf{x}}.$$

Further, define $\mathbf{r}_k = z(B_{\boldsymbol{\theta}}, \mathbf{x}) + \sqrt{p} + \sqrt{2\mathbf{x} + k^2}$, so that $\mathbf{r}_k - \mathbf{r}_0 \leq k$. It holds by similar arguments under the conditions $2\diamond(\mathbf{x}) \leq 1/2$, $\mathbf{x} \geq 2$, and $\|\boldsymbol{\xi}_{\boldsymbol{\theta}}\| \leq z(B_{\boldsymbol{\theta}}, \mathbf{x})$:

$$\begin{aligned}
\frac{1}{\Pi_K} \int_{U_K \setminus U_0} e^{L(\boldsymbol{v})} d\boldsymbol{v} &\leq (1 - e^{-x})^{-1} \sum_{k=1}^K e^{2r_k \diamond(x)} I\!\!P(\|\boldsymbol{\gamma} + \boldsymbol{\xi}_{\boldsymbol{\theta}}\| > r_{k-1}) \\
&\leq (1 - e^{-x})^{-1} \sum_{k=1}^K e^{2r_k \diamond(x)} e^{-x - (k-1)^2/2} \\
&\leq (1 - e^{-x})^{-1} e^{2r_0 \diamond(x) - x} \sum_{k=1}^{\infty} e^{k/2 - (k-1)^2/2} \\
&\leq 4(1 - e^{-x})^{-1} e^{2r_0 \diamond(x) - x} \leq 5e^{2r_0 \diamond(x) - x}.
\end{aligned}$$

Therefore, on a set of dominating probability, the posterior mass of the set $U_K \setminus U_0$ with $U_0 = \Theta_0(r_0) \times \mathcal{H}_0(h_0)$ is exponentially small provided that r_0^2 exceeds the level $C(p+x)$ and the value $r_0 \diamond(x)$ is small. The full dimensional contraction result (12.16) claims the concentration property for the set $\Upsilon(h_0, h_0)$ with $h_0^2 \geq C(p^* + x)$, leading to the error term $h_0 \diamond(x) = h_0 \diamond(h_0, x)$ in the Gaussian approximation of the posterior. The result (12.20) shows that the contraction property can be stated even for a smaller anisotropic rectangle $\Upsilon(r_0, h_0) = \Theta_0(r_0) \times \mathcal{H}_0(h_0)$. This allows to refine the error term in the semiparametric BvM result to $r_0 \diamond(x) = r_0 \diamond(h_0, x)$.

Below we shorten $\rho(x)$ and $\rho_0(x)$ to ρ and ρ_0 . By definition

$$\mathbb{E}^\circ f(\mathbb{I}_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{\vartheta} - \boldsymbol{\theta}^*)) = \frac{\int_{\Upsilon} e^{L(\boldsymbol{v})} f(\mathbb{I}_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)) d\boldsymbol{v}}{\int_{\Upsilon} e^{L(\boldsymbol{v})} d\boldsymbol{v}}$$

with $\boldsymbol{v} = (\boldsymbol{\theta}, \boldsymbol{\eta})$. Conditions (12.17) and (12.20) imply

$$\begin{aligned}
\int_{\Upsilon} e^{L(\boldsymbol{v})} d\boldsymbol{v} &\geq \int_{U_K} e^{L(\boldsymbol{v})} d\boldsymbol{v} \geq \int_{U_0} e^{L(\boldsymbol{v})} d\boldsymbol{v}, \\
\int_{\Upsilon} e^{L(\boldsymbol{v})} d\boldsymbol{v} &\leq (1 + \rho) \int_{U_K} e^{L(\boldsymbol{v})} d\boldsymbol{v} \leq (1 + \rho)(1 + \rho_0) \int_{U_0} e^{L(\boldsymbol{v})} d\boldsymbol{v} \leq e^{\rho + \rho_0} \int_{U_0} e^{L(\boldsymbol{v})} d\boldsymbol{v}.
\end{aligned}$$

Similar bounds holds for the integrals of a non-negative function f . This implies with $\boldsymbol{u} = \mathbb{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ for $\boldsymbol{v} = (\boldsymbol{\theta}, \boldsymbol{\eta})$

$$\begin{aligned}
\frac{\int_{\Upsilon} e^{L(\boldsymbol{v})} f(\boldsymbol{u}) d\boldsymbol{v}}{\int_{\Upsilon} e^{L(\boldsymbol{v})} d\boldsymbol{v}} &\leq \|f\|_\infty \left(\frac{\int_{\Upsilon \setminus U_K} e^{L(\boldsymbol{v})} d\boldsymbol{v}}{\int_{U_K} e^{L(\boldsymbol{v})} d\boldsymbol{v}} + \frac{\int_{U_K \setminus U_0} e^{L(\boldsymbol{v})} d\boldsymbol{v}}{\int_{U_K} e^{L(\boldsymbol{v})} d\boldsymbol{v}} \right) + \frac{\int_{U_0} e^{L(\boldsymbol{v})} f(\boldsymbol{u}) d\boldsymbol{v}}{\int_{U_0} e^{L(\boldsymbol{v})} d\boldsymbol{v}} \\
&\leq \|f\|_\infty (\rho + \rho_0) + \exp\{2r_0 \diamond(x)\} \mathbb{E}^\circ \left\{ f(\boldsymbol{\gamma} + \check{\boldsymbol{\xi}}_{\boldsymbol{\theta}}) \mid \|\boldsymbol{\gamma} + \check{\boldsymbol{\xi}}_{\boldsymbol{\theta}}\| \leq r_0 \right\}.
\end{aligned}$$

Similarly, as $f \geq 0$,

$$\begin{aligned}
\frac{\int_{\Upsilon} e^{L(\boldsymbol{v})} f(\boldsymbol{u}) d\boldsymbol{v}}{\int_{\Upsilon} e^{L(\boldsymbol{v})} d\boldsymbol{v}} &\geq \frac{\int_{U_0} e^{L(\boldsymbol{v})} f(\boldsymbol{u}) d\boldsymbol{v}}{\int_{\Upsilon} e^{L(\boldsymbol{v})} d\boldsymbol{v}} \geq e^{-\rho - \rho_0} \frac{\int_{U_0} e^{L(\boldsymbol{v})} f(\boldsymbol{u}) d\boldsymbol{v}}{\int_{U_0} e^{L(\boldsymbol{v})} d\boldsymbol{v}} \\
&\geq \exp\{-2r_0 \diamond(x) - \rho - \rho_0\} \mathbb{E}^\circ \left\{ f(\boldsymbol{\gamma} + \check{\boldsymbol{\xi}}_{\boldsymbol{\theta}}) \mid \|\boldsymbol{\gamma} + \check{\boldsymbol{\xi}}_{\boldsymbol{\theta}}\| \leq r_0 \right\}.
\end{aligned}$$

Remark 12.3.1. The general Wilks expansion and posterior contraction results require “ $h_0 \diamondsuit(h_0, x)$ small”, while the focus on the target component of dimension p allows to relax the condition to “ $r_0 \diamondsuit(h_0, x)$ small”. Under regularity conditions $h_0^2 \asymp p^* + x$ while $r_0^2 \asymp p + x$, so the gain in accuracy is of order $\sqrt{p/p^*}$. In particular, in the i.i.d. and regression cases $\diamondsuit(h_0, x) \asymp (p^* + x)/\sqrt{n}$, while $r_0^2 \asymp p + x$, and the low dimensional target can be estimated under “ $\sqrt{(p^* + x)^2 p/n}$ small”, while the inference on the full parameter requires “ $\sqrt{(p^* + x)^3/n}$ small”.

12.4 Sieve semiparametric inference

This section presents a semiparametric version of the Fisher, Wilks, and BvM results when the nuisance parameter is infinite dimensional. In some sense, the results are obtained by compiling the already stated results: to approximate the infinite dimensional nuisance parameter ϕ by a finite dimensional parameter η , to evaluate the bias induced by this approximation, and to state the results for the approximating finite dimensional model with parameter $v = (\theta, \eta)$. This approach leads to the sieve MLE $\tilde{\theta}_m$ and its population counterpart θ_m^*

$$\begin{aligned}\tilde{\theta}_m &= \operatorname{argmax}_{\theta \in \Theta} \max_{\phi \in \mathcal{S}_m} L(\theta, \phi), \\ \theta_m^* &= \operatorname{argmax}_{\theta \in \Theta} \max_{\phi \in \mathcal{S}_m} \mathbb{E}L(\theta, \phi).\end{aligned}$$

Below $L(\theta, \eta)$ means $L(\theta, \phi)$ for $\phi = (\eta, 0) \in \mathcal{S}_m$. The obtained results for the sieve model \mathcal{S}_m yield on a set $\Omega_m(x)$ with $\mathbb{P}(\Omega_m(x)) \geq 1 - e^{-x}$ for the target component θ

$$\|\mathbb{I}_{m,\theta}^{1/2}(\tilde{\theta}_m - \theta_m^*) - \check{\xi}_m\| \leq \diamondsuit_m(x).$$

This result yields the concentration property of the sieve estimator $\tilde{\theta}_m$ in the elliptic vicinity of θ_m^* . In combination with the approximation results of Section 12.6, we can derive similar bounds in terms of the true target θ^* and the efficient information matrix \mathbb{I}_θ .

12.5 Estimation of a nonlinear functional

Suppose that $L(v)$ is a quasi log-likelihood function of a functional parameter $v \in \Upsilon$ satisfying the regularity conditions which ensure the local linear approximation of the gradient $\nabla L(v)$. Let now the target of estimation be a smooth nonlinear functional $\theta = \psi(v)$. As usual the true value is $\theta^* = \psi(v^*)$ for v^* maximizing $\mathbb{E}L(v)$. We assume that the mapping $v \rightarrow \psi(v)$ is locally injective.

For each θ , define $\Upsilon(\theta)$ as a subset of Υ with $\psi(\mathbf{v}) = \theta$. Further, on the product $\Upsilon \times \Theta$, define

$$T(\mathbf{v}^\circ, \theta) \stackrel{\text{def}}{=} \underset{\mathbf{v}: \psi(\mathbf{v})=\theta}{\operatorname{arginf}} |\mathbb{E}L(\mathbf{v}) - \mathbb{E}L(\mathbf{v}^\circ)|.$$

may be better

$$T(\mathbf{v}^\circ, \theta) \stackrel{\text{def}}{=} \underset{\mathbf{v}: \psi(\mathbf{v})=\theta}{\operatorname{arginf}} \|\mathbf{v} - \mathbf{v}^\circ\|_{\mathcal{V}_0}.$$

Obviously $\psi(\mathbf{v}) = \theta$ yields $T(\mathbf{v}, \theta) = \mathbf{v}$. For a fixed \mathbf{v} , the mapping $\theta \rightarrow T(\mathbf{v}, \theta)$ can be viewed as a one-dimensional curve $\mathbf{v}(\theta)$ in Υ passing through \mathbf{v} for different values of the functional $\psi(\cdot)$. We assume that any two such curves either coincide or do not overlap. Moreover, we implicitly assume that locally such curves cover Υ in a unique way.

A toy example of this construction is a direct product with $\mathbf{v} = (\theta, \boldsymbol{\eta})$ and $L(\mathbf{v}) = L_1(\theta) + L_2(\boldsymbol{\eta})$. Then, for $\mathbf{v}^\circ = (\theta^\circ, \boldsymbol{\eta})$, it holds $T(\mathbf{v}^\circ, \theta) = (\theta, \boldsymbol{\eta})$.

Our first step can be viewed as decoupling. Namely, consider the product $\Theta \times \Upsilon$ and for a prior Π on Υ , define a measure $d\theta \Pi(d\mathbf{v})$ on this product.

Lemma 12.5.1. *It holds for any bounded function $f(\theta)$*

$$\frac{\iint e^{L(T(\mathbf{v}, \theta))} f(\theta) d\theta \Pi(d\mathbf{v})}{\iint e^{L(T(\mathbf{v}, \theta))} d\theta \Pi(d\mathbf{v})} \approx \frac{\int e^{L(\mathbf{v})} f(\theta) \Pi(d\mathbf{v})}{\int e^{L(\mathbf{v})} \Pi(d\mathbf{v})}.$$

Next we use the decomposition

$$L(T(\mathbf{v}, \theta)) - L(\mathbf{v}^*) = L(T(\mathbf{v}, \theta)) - L(T(\mathbf{v}, \theta^*)) + L(T(\mathbf{v}, \theta^*)) - L(\mathbf{v}^*)$$

Obviously the second difference $L(T(\mathbf{v}, \theta^*)) - L(\mathbf{v}^*)$ does not depend on θ , while the first one $L(T(\mathbf{v}, \theta)) - L(T(\mathbf{v}, \theta^*))$ is being computed along the curve $\mathbf{v}(\theta)$ passing through \mathbf{v} and only weakly depends on \mathbf{v} . Indeed, the local linear approximation of $\nabla L(\mathbf{v})$ implies for $\mathbf{v} \in \Upsilon_\circ(\mathbf{r})$

$$\left\| D^{-1} \{ \nabla L(T(\mathbf{v}, \theta)) - \nabla L(T(\mathbf{v}, \theta^*)) - \boldsymbol{\xi} + D\{T(\mathbf{v}, \theta) - T(\mathbf{v}, \theta^*)\} \} \right\| \leq \diamond(\mathbf{r})$$

and

$$\begin{aligned} & \left| L(T(\mathbf{v}, \theta)) - L(T(\mathbf{v}, \theta^*)) - \boldsymbol{\xi}^\top D\{T(\mathbf{v}, \theta) - T(\mathbf{v}, \theta^*)\} + \frac{1}{2} \|D\{T(\mathbf{v}, \theta) - T(\mathbf{v}, \theta^*)\}\|^2 \right| \\ & \leq \|D\{T(\mathbf{v}, \theta) - T(\mathbf{v}, \theta^*)\}\| \diamond(\mathbf{r}). \end{aligned}$$

Finally we need to ensure that along the curve $T(\mathbf{v}, \theta)$ for a fixed \mathbf{v} , the difference $D\{T(\mathbf{v}, \theta) - T(\mathbf{v}, \theta^*)\}$ mainly depends on $\theta - \theta^*$ and the impact of \mathbf{v} is not significant at least in the vicinity $\Upsilon_\circ(\mathbf{r})$ of \mathbf{v}^* .

Lemma 12.5.2. *It holds for each \mathbf{v}*

$$\begin{aligned} \frac{\int e^{L(T(\mathbf{v}, \theta))} f(\theta) d\theta}{\int e^{L(T(\mathbf{v}, \theta))} d\theta} &= \frac{\iint e^{L(T(\mathbf{v}, \theta), T(\mathbf{v}, \theta^*))} f(\theta) d\theta}{\iint e^{L(T(\mathbf{v}, \theta), T(\mathbf{v}, \theta^*))} d\theta} \\ &\approx \frac{\int \exp\{\boldsymbol{\xi}^\top D\{T(\mathbf{v}, \theta) - T(\mathbf{v}, \theta^*)\} - \|D\{T(\mathbf{v}, \theta) - T(\mathbf{v}, \theta^*)\}\|^2/2\} f(\theta) d\theta}{\int \exp\{\boldsymbol{\xi}^\top D\{T(\mathbf{v}, \theta) - T(\mathbf{v}, \theta^*)\} - \|D\{T(\mathbf{v}, \theta) - T(\mathbf{v}, \theta^*)\}\|^2/2\} d\theta} \end{aligned}$$

To be done: state carefully with the error term

12.6 Bias in semiparametric sieve approximation

Now we consider the case when a growing sequence of sieves $(\boldsymbol{\theta}, \boldsymbol{\eta})$ approximates the full model $(\boldsymbol{\theta}, \boldsymbol{\phi})$ with the true parameter $\mathbf{v}^* = (\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$.

Below we consider the sieve approximation $(\boldsymbol{\eta}, 0) \in \mathcal{S}_m$ of the nuisance parameter $\boldsymbol{\phi} = (\boldsymbol{\eta}, \boldsymbol{\varkappa})$. We check the difference between the target value $\boldsymbol{\theta}^*$ defined as

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \max_{\boldsymbol{\phi}} \mathbb{E} L(\boldsymbol{\theta}, \boldsymbol{\phi})$$

compared with the sieved version

$$\boldsymbol{\theta}_m^* \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \max_{\boldsymbol{\eta}} \mathbb{E} L(\boldsymbol{\theta}, \boldsymbol{\eta}, 0). \quad (12.21)$$

Consider the full dimensional Fisher operator

$$\mathbb{F} \stackrel{\text{def}}{=} \mathbb{F}(\mathbf{v}^*) = -\nabla^2 \mathbb{E} L(\mathbf{v}^*).$$

It can be written in the block form as

$$\mathbb{F} = \begin{pmatrix} \mathbb{F}_{\boldsymbol{\theta}} & \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\phi}} \\ \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\phi}}^\top & \mathbb{F}_{\boldsymbol{\phi}} \end{pmatrix} = \begin{pmatrix} \mathbb{F}_{\boldsymbol{\theta}} & \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}} & \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\varkappa}} \\ \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top & \mathbb{F}_{\boldsymbol{\eta}} & \mathbb{F}_{\boldsymbol{\eta}\boldsymbol{\varkappa}} \\ \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\varkappa}}^\top & \mathbb{F}_{\boldsymbol{\eta}\boldsymbol{\varkappa}}^\top & \mathbb{F}_{\boldsymbol{\varkappa}} \end{pmatrix},$$

and we are mainly interested in the efficient $\boldsymbol{\theta}$ -matrix $\mathbb{I}_{\boldsymbol{\theta}}$ defined via inversion or diagonalization of \mathbb{F} :

$$\mathbb{I}_{\boldsymbol{\theta}} = \mathbb{F}_{\boldsymbol{\theta}} - \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\phi}} \mathbb{F}_{\boldsymbol{\phi}}^{-1} \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\phi}}^\top.$$

The main question of this section is to quantify how this efficient information matrix is sensitive to truncation (sieve approximation) of the nuisance parameter. This requires to evaluate the difference between $\mathbb{I}_{\boldsymbol{\theta}}$ and a similar matrix $\mathbb{I}_{m,\boldsymbol{\theta}}$ coming from the sieve \mathcal{S}_m :

$$\mathbb{I}_{m,\boldsymbol{\theta}} = \mathbb{F}_{\boldsymbol{\theta}} - \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}} \mathbb{F}_{\boldsymbol{\eta}}^{-1} \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top.$$

We first explain how the general case can be reduced to the case when the nuisance block of this matrix is the unity matrix.

12.6.1 Basis transformation for the nuisance

In what follows, it is much more convenient to make a basis transformation which provides a simple unity structure for the nuisance blocks \mathbb{F}_η , \mathbb{F}_\varkappa , and $\mathbb{F}_{\eta\varkappa}$. The only additional condition we need is that the following matrix is positive:

$$\mathbb{I}_\varkappa \stackrel{\text{def}}{=} \mathbb{F}_\varkappa - \mathbb{F}_{\eta\varkappa}^\top \mathbb{F}_{\eta\varkappa}^{-1} \mathbb{F}_{\eta\varkappa} > 0.$$

For any nuisance vector $\phi = (\eta, \varkappa)$, it holds with $\check{\eta} = \eta + \mathbb{F}_\eta^{-1} \mathbb{F}_{\eta\varkappa} \varkappa$

$$\begin{aligned} \phi^\top \mathbb{F} \phi &= \eta^\top \mathbb{F}_\eta \eta + 2\eta^\top \mathbb{F}_{\eta\varkappa} \varkappa + \varkappa^\top \mathbb{F}_\varkappa \varkappa = \check{\eta}^\top \mathbb{F}_\eta \check{\eta} + \varkappa^\top \mathbb{I}_\varkappa \varkappa \\ &= \|\mathbb{F}_\eta^{1/2} \check{\eta}\|^2 + \|\mathbb{I}_\varkappa^{1/2} \varkappa\|^2. \end{aligned}$$

This means that the change of variable $\eta \rightarrow \mathbb{F}_\eta^{1/2} \check{\eta} = \mathbb{F}_\eta^{1/2} \eta + \mathbb{F}_\eta^{-1/2} \mathbb{F}_{\eta\varkappa} \varkappa$ and $\varkappa \rightarrow \mathbb{I}_\varkappa^{1/2} \varkappa$ reduces the general case to the case with the identity nuisance block $\mathbb{F}_\phi = I_\phi$:

$$\mathbb{F} = \begin{pmatrix} \mathbb{F}_\theta & \check{\mathbb{F}}_{\theta\eta} & \check{\mathbb{F}}_{\theta\varkappa} \\ \check{\mathbb{F}}_{\theta\eta}^\top & I_\eta & 0 \\ \check{\mathbb{F}}_{\theta\varkappa}^\top & 0 & I_\varkappa \end{pmatrix},$$

where

$$\begin{aligned} \check{\mathbb{F}}_{\theta\eta} &= \mathbb{F}_{\theta\eta} \mathbb{F}_\eta^{-1/2}, \\ \check{\mathbb{F}}_{\theta\varkappa} &= (\mathbb{F}_{\theta\varkappa} + \mathbb{F}_{\theta\eta} \mathbb{F}_\eta^{-1} \mathbb{F}_{\eta\varkappa}) \mathbb{I}_\varkappa^{-1/2}. \end{aligned} \tag{12.22}$$

The efficient Fisher matrices in the sieve and in the full models can be written in the form

$$\mathbb{I}_{m,\theta} = \mathbb{F}_\theta - \check{\mathbb{F}}_{\theta\eta} \check{\mathbb{F}}_{\theta\eta}^\top = \mathbb{F}_\theta - \mathbb{F}_{\theta\eta} \mathbb{F}_\eta^{-1} \mathbb{F}_{\theta\eta}^\top, \tag{12.23}$$

$$\begin{aligned} \mathbb{I}_\theta &= \mathbb{F}_\theta - \check{\mathbb{F}}_{\theta\eta} \check{\mathbb{F}}_{\theta\eta}^\top - \check{\mathbb{F}}_{\theta\varkappa} \check{\mathbb{F}}_{\theta\varkappa}^\top \\ &= \mathbb{F}_\theta - \mathbb{F}_{\theta\eta} \mathbb{F}_\eta^{-1} \mathbb{F}_{\theta\eta}^\top + (\mathbb{F}_{\theta\varkappa} + \mathbb{F}_{\theta\eta} \mathbb{F}_\eta^{-1} \mathbb{F}_{\eta\varkappa}) \mathbb{I}_\varkappa^{-1} (\mathbb{F}_{\theta\varkappa} + \mathbb{F}_{\theta\eta} \mathbb{F}_\eta^{-1} \mathbb{F}_{\eta\varkappa})^\top. \end{aligned} \tag{12.24}$$

Below we suppose that the change of the nuisance is made to ensure the identity block in the Fisher matrix which will be written as

$$\mathbb{F} = \begin{pmatrix} \mathbb{F}_\theta & \mathbb{F}_{\theta\phi} \\ \mathbb{F}_{\theta\phi}^\top & I_\phi \end{pmatrix} = \begin{pmatrix} \mathbb{F}_\theta & \mathbb{F}_{\theta\eta} & \mathbb{F}_{\theta\varkappa} \\ \mathbb{F}_{\theta\eta}^\top & I_\eta & 0 \\ \mathbb{F}_{\theta\varkappa}^\top & 0 & I_\varkappa \end{pmatrix}. \tag{12.25}$$

In the sieve model $v^s = (\theta, \eta, 0) \in \mathcal{S}_m$, we consider the Fisher submatrix \mathbb{F}_m with

$$\mathbb{F}_m = \begin{pmatrix} \mathbb{F}_\theta & \mathbb{F}_{\theta\eta} \\ \mathbb{F}_{\theta\eta}^\top & I_\eta \end{pmatrix}. \tag{12.26}$$

The efficient Fisher information matrix $\mathbb{I}_{\boldsymbol{\theta}}$ and $\mathbb{I}_{m,\boldsymbol{\theta}}$ in the full and in the sieve model are given by

$$\begin{aligned}\mathbb{I}_{\boldsymbol{\theta}} &= \mathbb{F}_{\boldsymbol{\theta}} - \mathbb{F}_{\boldsymbol{\theta}\eta}\mathbb{F}_{\boldsymbol{\theta}\eta}^\top - \mathbb{F}_{\boldsymbol{\theta}\varkappa}\mathbb{F}_{\boldsymbol{\theta}\varkappa}^\top. \\ \mathbb{I}_{m,\boldsymbol{\theta}} &= \mathbb{F}_{\boldsymbol{\theta}} - \mathbb{F}_{\boldsymbol{\theta}\eta}\mathbb{F}_{\boldsymbol{\theta}\eta}^\top.\end{aligned}\quad (12.27)$$

In general case, these formulas have to be understood in the sense of (12.23) and (12.24).

12.6.2 Smoothness conditions

First we specify the conditions ensuring that the sieve approximation of the nuisance parameter does not significantly affect the quality of estimation of the nuisance. Usually these conditions are given in term of smoothness of the nuisance component. We discuss the relation of our conditions with smoothness of the model in the nuisance parameter ϕ when discussing the structural regression in Chapter 16.

For simplicity of presentation, the conditions are stated for the case of the identity nuisance block in the information matrix as in (12.25). In the general case, one has to use $\check{\mathbb{F}}_{\boldsymbol{\theta}\eta}$ and $\check{\mathbb{F}}_{\boldsymbol{\theta}\varkappa}$ from (12.22) in place of $\mathbb{F}_{\boldsymbol{\theta}\eta}$ and $\mathbb{F}_{\boldsymbol{\theta}\varkappa}$.

The first condition is a kind of semiparametric identifiability and it allows to separate the target and the nuisance parameters. Formally it requires that the angle between two tangent subspaces for these parameters is separated away from zero:

(\mathcal{I}_s) The blocks of the Fisher operator \mathbb{F} fulfill

$$\|\mathbb{F}_{\boldsymbol{\theta}}^{-1/2}\mathbb{F}_{\boldsymbol{\theta}\phi}\mathbb{F}_{\boldsymbol{\theta}\phi}^\top\mathbb{F}_{\boldsymbol{\theta}}^{-1/2}\|_{\text{op}} = \|\mathbb{F}_{\boldsymbol{\theta}}^{-1/2}(\mathbb{F}_{\boldsymbol{\theta}\eta}\mathbb{F}_{\boldsymbol{\theta}\eta}^\top + \mathbb{F}_{\boldsymbol{\theta}\varkappa}\mathbb{F}_{\boldsymbol{\theta}\varkappa}^\top)\mathbb{F}_{\boldsymbol{\theta}}^{-1/2}\|_{\text{op}} \leq \nu < 1. \quad (12.28)$$

Further we need a kind of smoothness property of the nuisance parameter ϕ^* . This means that the projection on the target space of the cut-off parameter \varkappa^* in the decomposition $\phi^* = (\boldsymbol{\eta}^*, \varkappa^*)$ is nearly negligible.

(b_m) For some small non-negative number b_m

$$\|\mathbb{F}_{\boldsymbol{\theta}}^{-1/2}\mathbb{F}_{\boldsymbol{\theta}\varkappa}\varkappa^*\| \leq b_m. \quad (12.29)$$

In addition, we assume that the target parameter $\boldsymbol{\theta}$ is smooth in a similar sense which means that its interaction with the truncated parameter \varkappa is negligible.

($\mathbb{F}_{\boldsymbol{\theta}\varkappa}$) For some small non-negative number ρ_m

$$\|\mathbb{F}_{\boldsymbol{\theta}}^{-1/2}\mathbb{F}_{\boldsymbol{\theta}\varkappa}\mathbb{F}_{\boldsymbol{\theta}\varkappa}^\top\mathbb{F}_{\boldsymbol{\theta}}^{-1/2}\|_{\text{op}} \leq \rho_m. \quad (12.30)$$

We aim at checking how the cut-off of \varkappa can affect the efficient Fisher information $\mathbb{I}_{\boldsymbol{\theta}} = \mathbb{F}_{\boldsymbol{\theta}} - \mathbb{F}_{\boldsymbol{\theta}\phi}\mathbb{F}_{\boldsymbol{\theta}\phi}^\top$. Also we bound the scaled distance $\|\mathbb{I}_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_m^*)\|$ between two target values $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_m^*$; see (12.21).

Theorem 12.6.1. *Consider a semiparametric model with a quasi log-likelihood $L(\boldsymbol{\theta}, \boldsymbol{\phi})$. For the true value $\boldsymbol{v}^* = (\boldsymbol{\theta}^*, \boldsymbol{\phi}^*) = \operatorname{argmax} IEL(\boldsymbol{v}^*)$, let $\mathbb{F} \stackrel{\text{def}}{=} -\nabla^2 IEL(\boldsymbol{v}^*)$ be the corresponding Fisher operator. Suppose that the nuisance parameter $\boldsymbol{\phi}$ is rescaled to ensure that the corresponding $\boldsymbol{\phi}$ -block is identity. Let $(\boldsymbol{\eta}, 0)$ be a sieve approximation of the functional nuisance parameter $\boldsymbol{\phi} = (\boldsymbol{\eta}, \varkappa)$, and (12.25) be the related block representation of \mathbb{F} . Suppose the identifiability condition **(I_s)** and the smoothness condition **(b_m)** and **(F_{θ,ν})** with $\rho_m \leq \nu$. Then the efficient Fisher information matrices $\mathbb{I}_{m,\boldsymbol{\theta}} = \mathbb{F}_{\boldsymbol{\theta}} - \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}\mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top$ and $\mathbb{I}_{\boldsymbol{\theta}} = \mathbb{F}_{\boldsymbol{\theta}} - \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}\mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top - \mathbb{F}_{\boldsymbol{\theta}\varkappa}\mathbb{F}_{\boldsymbol{\theta}\varkappa}^\top$ in the sieve and in full models are related by*

$$\|\mathbb{I}_{m,\boldsymbol{\theta}}^{-1/2} (\mathbb{I}_{m,\boldsymbol{\theta}} - \mathbb{I}_{\boldsymbol{\theta}}) \mathbb{I}_{m,\boldsymbol{\theta}}^{-1/2}\|_{\text{op}} \leq \frac{1}{1-\nu} \rho_m, \quad (12.31)$$

$$\operatorname{tr} \left\{ \mathbb{I}_{m,\boldsymbol{\theta}}^{-1/2} (\mathbb{I}_{m,\boldsymbol{\theta}} - \mathbb{I}_{\boldsymbol{\theta}}) \mathbb{I}_{m,\boldsymbol{\theta}}^{-1/2} \right\}^2 \leq (1-\nu)^{-2} p \rho_m^2. \quad (12.32)$$

The target parameter $\boldsymbol{\theta}^*$ and its sieve counterpart $\boldsymbol{\theta}_m^*$ are related by

$$\|\mathbb{I}_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_m^*)\| \leq \frac{1}{1-\nu} \{b_m + 2\delta(\mathbf{r}_m)\mathbf{r}_m\}, \quad (12.33)$$

where $\mathbf{r}_m = \|\varkappa^*\|$.

Proof. The proof is done in two step. First we compare the deterministic quantities $\mathbb{I}_{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ with their sieve counterparts. Then we proceed with the random score vectors $\boldsymbol{\xi}$ and $\check{\boldsymbol{\xi}}_m$.

The identifiability condition (12.28) guarantees for $\mathbb{I}_{m,\boldsymbol{\theta}} = \mathbb{F}_{\boldsymbol{\theta}} - \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}\mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top$ that

$$\|\mathbb{F}_{\boldsymbol{\theta}}^{-1/2} \mathbb{I}_{m,\boldsymbol{\theta}} \mathbb{F}_{\boldsymbol{\theta}}^{-1/2}\|_{\text{op}} \geq 1 - \nu. \quad (12.34)$$

Due to (12.27) $\mathbb{I}_{m,\boldsymbol{\theta}} - \mathbb{I}_{\boldsymbol{\theta}} = \mathbb{F}_{\boldsymbol{\theta}\varkappa}\mathbb{F}_{\boldsymbol{\theta}\varkappa}^\top$. Therefore,

$$\begin{aligned} \mathbb{F}_{\boldsymbol{\theta}}^{-1/2} (\mathbb{I}_{m,\boldsymbol{\theta}} - \mathbb{I}_{\boldsymbol{\theta}}) \mathbb{F}_{\boldsymbol{\theta}}^{-1/2} &= \mathbb{F}_{\boldsymbol{\theta}}^{-1/2} \mathbb{F}_{\boldsymbol{\theta}\varkappa}\mathbb{F}_{\boldsymbol{\theta}\varkappa}^\top \mathbb{F}_{\boldsymbol{\theta}}^{-1/2}, \\ \mathbb{I}_{m,\boldsymbol{\theta}}^{-1/2} (\mathbb{I}_{m,\boldsymbol{\theta}} - \mathbb{I}_{\boldsymbol{\theta}}) \mathbb{I}_{m,\boldsymbol{\theta}}^{-1/2} &\leq \frac{1}{1-\nu} \mathbb{F}_{\boldsymbol{\theta}}^{-1/2} \mathbb{F}_{\boldsymbol{\theta}\varkappa}\mathbb{F}_{\boldsymbol{\theta}\varkappa}^\top \mathbb{F}_{\boldsymbol{\theta}}^{-1/2}. \end{aligned}$$

The use of the smoothness condition (12.30) implies (12.31) and (12.32). In a similar way we can compare the matrices \mathbb{F}_m^{-1} from (12.26) and the corresponding $(\boldsymbol{\theta}, \boldsymbol{\eta})$ -block \mathbb{I}_m^{-1} of \mathbb{F}^{-1} . Obviously

$$\mathbb{I}_m = \mathbb{F}_m - \operatorname{block}\{\mathbb{F}_{\boldsymbol{\theta}\varkappa}\mathbb{F}_{\boldsymbol{\theta}\varkappa}^\top, 0\}. \quad (12.35)$$

In its turn, the $\boldsymbol{\theta}$ -block of \mathbb{F}_m^{-1} is equal to $\mathbb{I}_{m,\boldsymbol{\theta}}^{-1}$, and it holds

$$\|\mathbb{F}_m^{-1/2}(\mathbb{I}_m - \mathbb{F}_m)\mathbb{F}_m^{-1/2}\|_{\text{op}} = \|\mathbb{I}_{m,\boldsymbol{\theta}}^{-1/2}\mathbb{F}_{\boldsymbol{\theta}\varkappa}\mathbb{F}_{\boldsymbol{\theta}\varkappa}^\top\mathbb{I}_{m,\boldsymbol{\theta}}^{-1/2}\|_{\text{op}} \leq \frac{1}{1-\nu}\rho_m.$$

This implies for $\rho_m \leq \nu$

$$\|\mathbb{F}_m^{-1/2}\mathbb{I}_m\mathbb{F}_m^{-1/2}\|_{\text{op}} \leq 1 + \frac{1}{1-\nu}\rho_m \leq \frac{1}{1-\nu}. \quad (12.36)$$

Next we bound the bias $\boldsymbol{\theta}^* - \boldsymbol{\theta}_m^*$ introduced in $\boldsymbol{\theta}^*$ by truncation. Consider first the Gaussian likelihood $L(\mathbf{v})$ with

$$\mathbb{E}L(\mathbf{v}^*) - \mathbb{E}L(\mathbf{v}) = \|\mathbb{F}^{1/2}(\mathbf{v} - \mathbf{v}^*)\|^2/2.$$

Then \mathbf{v}_m^* coincides with $\mathbf{v}_m^s = (\boldsymbol{\theta}_m^s, \boldsymbol{\eta}_m^s, 0)$ defined as minimizer of the quadratic function

$$f(\mathbf{v}_m) = f(\boldsymbol{\theta}, \boldsymbol{\eta}) = \|\mathbb{F}^{1/2}(\mathbf{v}_m - \mathbf{v}^*)\|^2$$

over the set of parameter $\mathbf{v}_m = (\boldsymbol{\theta}, \boldsymbol{\eta}, 0)$:

$$\begin{aligned} (\boldsymbol{\theta}_m^s, \boldsymbol{\eta}_m^s) &= \underset{\mathbf{v}_m = (\boldsymbol{\theta}, \boldsymbol{\eta}, 0)}{\operatorname{argmin}} \|\mathbb{F}^{1/2}(\mathbf{v}_m - \mathbf{v}^*)\|^2 \\ &= \underset{\boldsymbol{\theta}, \boldsymbol{\eta}}{\operatorname{argmin}} \left\{ (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \|\boldsymbol{\eta} - \boldsymbol{\eta}^*\|^2 \right. \\ &\quad \left. + 2(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}(\boldsymbol{\eta} - \boldsymbol{\eta}^*) - 2(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathbb{F}_{\boldsymbol{\theta}\varkappa}\varkappa^* \right\}. \end{aligned}$$

This implies by $\nabla_{\boldsymbol{\theta}} f(\mathbf{v}_m^s) = 0$ and $\nabla_{\boldsymbol{\eta}} f(\mathbf{v}_m^s) = 0$ that

$$\begin{aligned} \boldsymbol{\eta}_m^s - \boldsymbol{\eta}^* &= -\mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top(\boldsymbol{\theta}_m^s - \boldsymbol{\theta}^*), \\ \boldsymbol{\theta}_m^s - \boldsymbol{\theta}^* &= (\mathbb{F}_{\boldsymbol{\theta}} - \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}\mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top)^{-1}\mathbb{F}_{\boldsymbol{\theta}\varkappa}\varkappa^* = \mathbb{I}_{m,\boldsymbol{\theta}}^{-1}\mathbb{F}_{\boldsymbol{\theta}\varkappa}\varkappa^*. \end{aligned} \quad (12.37)$$

Further, (12.37) yields by simple algebra that

$$\mathbb{F}(\mathbf{v}^* - \mathbf{v}_m^s) = \left(0, 0, (I_\varkappa - \mathbb{F}_{\boldsymbol{\theta}\varkappa}^\top\mathbb{I}_{m,\boldsymbol{\theta}}^{-1}\mathbb{F}_{\boldsymbol{\theta}\varkappa})\varkappa^*\right) \quad (12.38)$$

and

$$\begin{aligned} \|\mathbb{F}^{1/2}(\mathbf{v}^* - \mathbf{v}_m^s)\|^2 &= (\mathbf{v}^* - \mathbf{v}_m^s)^\top \mathbb{F}(\mathbf{v}^* - \mathbf{v}_m^s) \\ &= \varkappa^{*\top}(I_\varkappa - \mathbb{F}_{\boldsymbol{\theta}\varkappa}^\top\mathbb{I}_{m,\boldsymbol{\theta}}^{-1}\mathbb{F}_{\boldsymbol{\theta}\varkappa})\varkappa^* \leq \|\varkappa^*\|^2. \end{aligned} \quad (12.39)$$

The identifiability condition (12.28) and the smoothness conditions (12.29), (12.30) imply

$$\begin{aligned}\|\mathbb{I}_{m,\theta}^{1/2}(\boldsymbol{\theta}_m^s - \boldsymbol{\theta}^*)\| &= \|\mathbb{I}_{m,\theta}^{-1/2}\mathbb{F}_{\boldsymbol{\theta}, \boldsymbol{\varkappa}}\boldsymbol{\varkappa}^*\| \\ &\leq (1-\nu)^{-1/2}\|\mathbb{F}_{\boldsymbol{\theta}}^{-1/2}\mathbb{F}_{\boldsymbol{\theta}, \boldsymbol{\varkappa}}\boldsymbol{\varkappa}^*\| \leq (1-\nu)^{-1/2}b_m.\end{aligned}\quad (12.40)$$

Now we consider a general non-Gaussian case. We use a linear approximation of the gradient of the expected log-likelihood. By (12.39), the point \boldsymbol{v}_m^s is a vicinity $\Upsilon_{\circ}(\mathbf{r}_m)$ of \boldsymbol{v}^* with $\mathbf{r}_m = \|\boldsymbol{\varkappa}^*\|$. Now the smoothness condition **(L₀)** implies that

$$\left\| \mathbb{F}^{-1/2} \left\{ \nabla \mathbb{E} L(\boldsymbol{v}^*) - \nabla \mathbb{E} L(\boldsymbol{v}_m^s) - \mathbb{F}(\boldsymbol{v}^* - \boldsymbol{v}_m^s) \right\} \right\| \leq \delta(\mathbf{r}_m)\mathbf{r}_m. \quad (12.41)$$

Further, $\nabla \mathbb{E} L(\boldsymbol{v}^*) = 0$ by the definition of $\boldsymbol{v}^* = (\boldsymbol{\theta}^*, \boldsymbol{\eta}^*, \boldsymbol{\varkappa}^*)$. Define

$$\boldsymbol{b} \stackrel{\text{def}}{=} \nabla \mathbb{E} L(\boldsymbol{v}_m^s) + \mathbb{F}(\boldsymbol{v}^* - \boldsymbol{v}_m^s),$$

and \boldsymbol{b}_m as its projection on the sieve subspace $(\boldsymbol{\theta}, \boldsymbol{\eta})$, that is, $\boldsymbol{b}_m = \nabla_m \mathbb{E} L(\boldsymbol{v}_m^s)$; see (12.38). The inequality (12.41) means $\|\mathbb{F}^{-1/2}\boldsymbol{b}\| \leq \delta(\mathbf{r}_m)\mathbf{r}_m$ and it implies

$$\|\mathbb{F}_m^{-1/2} \nabla_m \mathbb{E} L(\boldsymbol{v}_m^s)\| \leq \delta(\mathbf{r}_m)\mathbf{r}_m,$$

with \mathbb{F}_m from (12.35). Combining with (12.36) yields

$$\|\mathbb{F}_m^{-1/2} \nabla_m \mathbb{E} L(\boldsymbol{v}_m^s)\| \leq (1-\nu)^{-1/2}\delta(\mathbf{r}_m)\mathbf{r}_m. \quad (12.42)$$

Similar arguments in the sieve subspace $(\boldsymbol{\theta}, \boldsymbol{\eta})$ allow to derive for the expected gradient $\nabla_m \mathbb{E} L(\boldsymbol{v})$ that

$$\left\| \mathbb{F}_m^{-1/2} \left\{ \nabla_m \mathbb{E} L(\boldsymbol{v}_m^s) - \nabla_m \mathbb{E} L(\boldsymbol{v}_m^s) - \mathbb{F}_m(\boldsymbol{v}_m^s - \boldsymbol{v}_m^s) \right\} \right\| \leq \delta(\mathbf{r}_m)\mathbf{r}_m$$

and thus, in view of $\nabla_m \mathbb{E} L(\boldsymbol{v}_m^s) = 0$ and (12.42)

$$\|\mathbb{F}_m^{1/2}(\boldsymbol{v}_m^s - \boldsymbol{v}_m^s)\| \leq (1-\nu)^{-1/2}\delta(\mathbf{r}_m)\mathbf{r}_m + \delta(\mathbf{r}_m)\mathbf{r}_m.$$

Now, projecting on the $\boldsymbol{\theta}$ -subspace implies similarly to (12.37)

$$\|\mathbb{I}_{m,\theta}^{1/2}(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}_m^s) - \mathbb{I}_{m,\theta}^{-1/2}\mathbb{F}_{\boldsymbol{\theta}, \boldsymbol{\varkappa}}\boldsymbol{\varkappa}^*\| \leq \delta(\mathbf{r}_m)\mathbf{r}_m \{1 + (1-\nu)^{-1/2}\}.$$

This implies by (12.40)

$$\|\mathbb{I}_{m,\theta}^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_m^*)\| \leq (1-\nu)^{-1/2} \{b_m + 2\delta(\mathbf{r}_m)\mathbf{r}_m\},$$

which completes the proof of (12.33).

Finally we compare two standardized score vectors $\check{\boldsymbol{\xi}}$ and $\check{\boldsymbol{\xi}}_m$ which appear in the Fisher, Wilks, and BvM results. The vector $\check{\boldsymbol{\xi}}_m$ shows up in the Fisher expansion for

the sieve model, while $\check{\xi}$ is defined for the original full dimensional nuisance parameter. These vectors $\check{\xi}, \check{\xi}_m$ are defined by

$$\begin{aligned}\check{\xi} &= \mathbb{I}_{\theta}^{-1/2} \check{\nabla}_{\theta} = \mathbb{I}_{\theta}^{-1/2} (\nabla_{\theta} - \mathbb{F}_{\theta\eta} \nabla_{\eta} - \mathbb{F}_{\theta\kappa} \nabla_{\kappa}), \\ \check{\xi}_m &= \mathbb{I}_{m,\theta}^{-1/2} \check{\nabla}_{m,\theta} = \mathbb{I}_{m,\theta}^{-1/2} (\nabla_{\theta} - \mathbb{F}_{\theta\eta} \nabla_{\eta}),\end{aligned}$$

where $\nabla_{\theta} = \nabla_{\theta} L(\theta^*, \phi^*)$ and similarly $(\nabla_{\eta}, \nabla_{\kappa})$. In the classical theory both vectors $\check{\xi}$ and $\check{\xi}_m$ are nearly standard normal.

Theorem 12.6.2. Suppose the conditions of Theorem 12.6.1. If $\theta^\circ = \theta^* + \mathbb{I}_{\theta}^{-1/2} \check{\xi}$ and $\theta_m^\circ = \theta_m^* + \mathbb{I}_{m,\theta}^{-1/2} \check{\xi}_m$, then it holds with $\xi_m^s \stackrel{\text{def}}{=} \rho_m^{-1} \mathbb{F}_{\theta}^{-1/2} \mathbb{F}_{\theta\kappa} \nabla_{\kappa}$

$$\|\mathbb{I}_{\theta}^{1/2} (\theta^\circ - \theta_m^\circ)\| \leq \frac{\rho_m}{1-\nu} (\|\check{\xi}_m\| + \|\xi_m^s\|) + \frac{b_m}{1-\nu} + \delta(r_m) r_m. \quad (12.43)$$

Moreover, under **(ED₀)** and **(I_s)**, it holds on a set $\Omega(x)$ with $\mathbb{P}(\Omega(x)) \geq 1 - 4e^{-x}$

$$\|\mathbb{I}_{\theta}^{1/2} (\theta^\circ - \theta_m^\circ)\| \leq \frac{2a\rho_m}{1-\nu} (\sqrt{p} + \sqrt{2x}) + \frac{b_m}{1-\nu} + \delta(r_m) r_m. \quad (12.44)$$

Proof. Consider the stochastic part of $\mathbb{I}_{\theta}^{1/2} (\theta^\circ - \theta_m^\circ)$. The bounds (12.31) and (12.34) imply

$$\begin{aligned}\left\| \mathbb{I}_{\theta}^{1/2} \{ \mathbb{I}_{m,\theta}^{-1/2} \check{\xi}_m - \mathbb{I}_{\theta}^{-1/2} \check{\xi} \} \right\| &= \left\| \mathbb{I}_{\theta}^{1/2} \mathbb{I}_{m,\theta}^{-1} \check{\nabla}_{\theta,m} - \mathbb{I}_{\theta}^{-1/2} \check{\nabla}_{\theta} \right\| \\ &= \left\| \{ \mathbb{I}_{\theta}^{1/2} \mathbb{I}_{m,\theta}^{-1} - \mathbb{I}_{\theta}^{-1/2} \} \check{\nabla}_{\theta,m} + \mathbb{I}_{\theta}^{-1/2} \mathbb{F}_{\theta\kappa} \nabla_{\kappa} \right\| \\ &\leq \left\| \left\{ \mathbb{I}_{\theta}^{-1/2} \mathbb{I}_{m,\theta} \mathbb{I}_{\theta}^{1/2} - I_p \right\} \mathbb{I}_{\theta}^{1/2} \mathbb{I}_{m,\theta}^{-1} \check{\nabla}_{\theta,m} \right\| + \left\| \mathbb{I}_{\theta}^{-1/2} \mathbb{F}_{\theta\kappa} \nabla_{\kappa} \right\| \\ &\leq \frac{1}{1-\nu} \rho_m \|\check{\xi}_m\| + \frac{1}{1-\nu} \|\mathbb{F}_{\theta}^{-1/2} \mathbb{F}_{\theta\kappa} \nabla_{\kappa}\| \\ &= \frac{1}{1-\nu} (\|\check{\xi}_m\| + \|\xi_m^s\|) \rho_m.\end{aligned}$$

Here we have also used that $\mathbb{I}_{m,\theta} \geq \mathbb{I}_{\theta}$. This proves (12.43). It remains to check (12.44). If the full model is true then $\text{Var}(\check{\xi}_m) = I_p$ and under **(ED₀)**, it holds on a set of probability at least $1 - 2e^{-x}$

$$\|\check{\xi}_m\| \leq \sqrt{p} + \sqrt{2x}; \quad (12.45)$$

see Corollary B.1.2. Similarly, under the correct model specification, it holds $\text{Var}(\nabla_{\kappa}) = I_{\kappa}$. For $\xi_m^s = \rho_m^{-1} \mathbb{F}_{\theta}^{-1/2} \mathbb{F}_{\theta\kappa} \nabla_{\kappa}$, it holds by **(F_{θκ})**

$$\text{Var}(\xi_m^s) \leq \rho_m^{-2} \|\mathbb{F}_{\theta}^{-1/2} \mathbb{F}_{\theta\kappa} \mathbb{F}_{\theta\kappa}^T \mathbb{F}_{\theta}^{-1/2}\|_{\text{op}} \leq 1,$$

and hence, the use of **(ED₀)** implies with a dominating probability $1 - 2e^{-x}$

$$\|\xi_m^s\| \leq \sqrt{p} + \sqrt{2x}$$

similarly to (12.45). In the general situation, if PA is not exactly true, we still can use **(ED₀)**. It ensures $\text{Var}(\check{\xi}_m) \leq \alpha^2 I_p$, $\text{Var}(\xi_m^s) \leq \alpha^2 I_p$, and

$$\|\check{\xi}_m\| \leq \alpha(\sqrt{p} + \sqrt{2x}), \quad \|\xi_m^s\| \leq \alpha(\sqrt{p} + \sqrt{2x}).$$

This yields the last claim of the theorem.

Parametric i.i.d. models

The model with independent identically distributed (i.i.d.) observations is one of the most popular setups in statistical literature and in statistical applications. The essential and the most developed part of the statistical theory is designed for the i.i.d. modeling. Especially, the classical asymptotic parametric theory is almost complete including asymptotic root-n normality and efficiency of the MLE and Bayes estimators under rather mild assumptions; see e.g. Chapter 2 and 3 in [Ibragimov and Khas'minskij \(1981\)](#). So, the i.i.d. model can naturally serve as a benchmark for any extension of the statistical theory: being applied to the i.i.d. setup, the new approach should lead to essentially the same conclusions as in the classical theory. Similar reasons apply to the regression model and its extensions. Below we try demonstrate that the proposed non-asymptotic viewpoint is able to reproduce the existing brilliant and well established results of the classical parametric theory. Surprisingly, the majority of classical efficiency results can be easily derived from the obtained general non-asymptotic bounds.

13.1 Quasi MLE in an i.i.d. model

The basic i.i.d. parametric model means that the observations $\mathbf{Y} = (Y_1, \dots, Y_n)$ are independent identically distributed from a distribution P from a given parametric family $(P_\theta, \theta \in \Theta)$ on the observation space \mathcal{Y}_1 . Each $\theta \in \Theta$ clearly yields the product data distribution $\mathbb{P}_\theta = P_\theta^{\otimes n}$ on the product space $\mathcal{Y} = \mathcal{Y}_1^n$. This section illustrates how the obtained general results can be applied to this type of modeling under possible model misspecification. Different types of misspecification can be considered. Each of the assumptions, namely, data independence, identical distribution, parametric form of the marginal distribution can be violated. To be specific, we assume the observations Y_i independent and identically distributed. However, we admit that the distribution of each Y_i does not necessarily belong to the parametric family (P_θ) . The case of non-identically distributed observations can be done similarly at cost of more complicated notation.

In what follows the parametric family $(P_{\boldsymbol{\theta}})$ is supposed to be dominated by a measure μ_0 , and each density $p(y, \boldsymbol{\theta}) = dP_{\boldsymbol{\theta}}/d\mu_0(y)$ is two times continuously differentiable in $\boldsymbol{\theta}$ for all y . Denote $\ell(y, \boldsymbol{\theta}) = \log p(y, \boldsymbol{\theta})$. The parametric assumption $Y_i \sim P_{\boldsymbol{\theta}^*} \in (P_{\boldsymbol{\theta}})$ leads to the log-likelihood

$$L(\boldsymbol{\theta}) = \sum_{i=1}^n \ell(Y_i, \boldsymbol{\theta}), \quad (13.1)$$

where the summation is taken over $i = 1, \dots, n$. The quasi MLE $\tilde{\boldsymbol{\theta}}$ maximizes this sum over $\boldsymbol{\theta} \in \Theta$:

$$\tilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^n \ell(Y_i, \boldsymbol{\theta}).$$

The target of estimation $\boldsymbol{\theta}^*$ maximizes the expectation of $L(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \mathbb{E} L(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^n \mathbb{E} \ell(Y_i, \boldsymbol{\theta}).$$

Let $\zeta_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \ell(Y_i, \boldsymbol{\theta}) - \mathbb{E} \ell(Y_i, \boldsymbol{\theta})$. Then $\zeta(\boldsymbol{\theta}) = \sum_{i=1}^n \zeta_i(\boldsymbol{\theta})$. The equation $\mathbb{E} \nabla L(\boldsymbol{\theta}^*) = 0$ implies

$$\nabla \zeta(\boldsymbol{\theta}^*) = \sum_{i=1}^n \nabla \zeta_i(\boldsymbol{\theta}^*) = \sum_{i=1}^n \nabla \ell(Y_i, \boldsymbol{\theta}^*). \quad (13.2)$$

13.2 Conditions in the i.i.d. case

I.i.d. structure of the Y_i 's allows for rewriting the conditions **(E₀)**, **(E₂)**, **(I)**, **(L₀)**, and **(L)** in terms of the marginal distribution. In the following conditions the index i runs from 1 to n .

(ed₀) *There exists a positive symmetric matrix v_0^2 , such that for all $|\lambda| \leq g_1$*

$$\sup_{\gamma \in \mathbb{R}^p} \log \mathbb{E} \exp \left\{ \lambda \frac{\gamma^\top \nabla \zeta_i(\boldsymbol{\theta}^*)}{\|v_0 \gamma\|} \right\} \leq v_0^2 \lambda^2 / 2.$$

A natural candidate on v_0^2 is given by the variance of the gradient $\nabla \ell(Y_1, \boldsymbol{\theta}^*)$, that is, $v_0^2 = \operatorname{Var} \nabla \ell(Y_1, \boldsymbol{\theta}) = \operatorname{Var} \nabla \zeta_1(\boldsymbol{\theta})$. Note that **(ed₀)** is automatically fulfilled if the model is correctly specified and $P = P_{\boldsymbol{\theta}^*}$ because $E_{\boldsymbol{\theta}^*} \exp \{\ell(Y_1, \boldsymbol{\theta}) - \ell(Y_1, \boldsymbol{\theta}^*)\} \equiv 1$.

Further we restate the local regularity conditions **(E₂)** and **(L₀)** in terms of the expected value $\bar{\ell}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{E} \ell(Y_i, \boldsymbol{\theta})$ of each $\ell(Y_i, \boldsymbol{\theta})$. We suppose that $\bar{\ell}(\boldsymbol{\theta})$ is two times differentiable and define the matrix function $F(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2 \bar{\ell}(\boldsymbol{\theta})$. Define $F_0 \stackrel{\text{def}}{=} F(\boldsymbol{\theta}^*)$ and consider the local sets

$$\Theta_0(\mathbf{r}) = \{\boldsymbol{\theta} : \|\mathsf{F}_0^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}/n^{1/2}\}.$$

In the regular parametric case with $P \in (P_{\boldsymbol{\theta}})$, the matrices v_0^2 and F_0 coincide with the Fisher information matrix $\mathsf{F}_0 = \mathsf{F}(\boldsymbol{\theta}^*)$ of the family $(P_{\boldsymbol{\theta}})$ at the point $\boldsymbol{\theta}^*$. Below we suppose the log-likelihood function $\ell(y, \boldsymbol{\theta})$ to be sufficiently smooth in $\boldsymbol{\theta}$. This allows to fix the functions $\delta(\mathbf{r})$ and $\varrho(\mathbf{r})$ used in **(E₂)** and **(L₀)** proportional to \mathbf{r} .

(ed₂) *There exist a value $\omega^* > 0$ and for each $\mathbf{r} > 0$, a constant $\mathbf{g}(\mathbf{r}) > 0$ such that*

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r})} \sup_{\mathbf{u} \in \mathcal{S}_p} \log I\!\!E \exp \left\{ \frac{\lambda}{\omega^*} \mathbf{u}^\top \mathsf{F}_0^{-1/2} \nabla^2 \zeta_i(\boldsymbol{\theta}) \mathsf{F}_0^{-1/2} \mathbf{u} \right\} \leq \frac{\nu_0^2 \lambda^2}{2}, \quad |\lambda| \leq \mathbf{g}(\mathbf{r}).$$

(l₀) *The function $\bar{\ell}(\boldsymbol{\theta})$ is two times differentiable and the matrix function $\mathsf{F}(\boldsymbol{\theta}) = -\nabla^2 I\!\!E \ell(Y_1, \boldsymbol{\theta})$ fulfills for some constant δ^* :*

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r})} \left\| \mathsf{F}_0^{-1/2} \mathsf{F}(\boldsymbol{\theta}) \mathsf{F}_0^{-1/2} - I_p \right\|_{\text{op}} \leq \frac{\delta^* \mathbf{r}}{\sqrt{n}}.$$

The consistency result for $\tilde{\boldsymbol{\theta}}$ requires certain growth of the value $\bar{\ell}(\boldsymbol{\theta}^*) - \bar{\ell}(\boldsymbol{\theta})$ as $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|$ grows. The marginal version of the global condition **(L)** reads as follows:

(\bar{\ell}) *There exists $C_1 \geq 0$ such that for each $\boldsymbol{\theta} \in \Theta$, it holds with $\delta = \delta^*/\sqrt{n}$ and $\mathbf{r} = \sqrt{n} \|\mathsf{F}_0^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|$*

$$n \{ \bar{\ell}(\boldsymbol{\theta}^*) - \bar{\ell}(\boldsymbol{\theta}) \} \geq (1 - \delta) (\mathbf{r} \mathbf{r}_0 - \mathbf{r}_0^2/2) - C_1 \mathbf{r}^2.$$

Remark 13.2.1. If the parametric i.i.d. model is correct, then

$$\bar{\ell}(\boldsymbol{\theta}^*) - \bar{\ell}(\boldsymbol{\theta}) = \mathcal{K}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}^*} \log \frac{dP_{\boldsymbol{\theta}^*}}{dP_{\boldsymbol{\theta}}}(Y_1)$$

is the Kullback-Leibler divergence for the family $(P_{\boldsymbol{\theta}})$. Condition **(\bar{\ell})** is fulfilled automatically if $\bar{\ell}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) > 0$ for $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$ and Θ is a compact set. Then

$$\inf_{\boldsymbol{\theta} \in \Theta} \frac{\bar{\ell}(\boldsymbol{\theta}^*) - \bar{\ell}(\boldsymbol{\theta})}{\|\mathsf{F}_0^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2} \geq b > 0.$$

This and **(l₀)** imply **(\bar{\ell})**.

The *identifiability condition* relates the matrices v_0^2 and F_0 .

(\iota) There is a constant $a > 0$ such that $a^2 \mathsf{F}_0 \geq \mathsf{v}_0^2$.

Lemma 13.2.1. *Let Y_1, \dots, Y_n be i.i.d. Then **(ed₀)**, **(ed₂)**, **(l₀)**, **(\bar{\ell})**, and **(\iota)** imply **(E₀)**, **(E₂)**, **(L₀)**, **(L)**, **(I)**, with $V^2 = n \mathsf{v}_0^2$, $D^2 = n \mathsf{F}_0$, $\omega = \omega^*/\sqrt{n}$, $\delta(\mathbf{r}) = \delta^* \mathbf{r}/\sqrt{n}$, and the same constants ν_0 , a , $\mathbf{g} \stackrel{\text{def}}{=} g_1 \sqrt{n}$.*

Proof. The identities $V^2 = nv_0^2$, $D^2 = nF_0$ follow from the i.i.d. structure of the observations Y_i . We briefly comment on condition (E_0) . The use once again of the i.i.d. structure yields by (13.2) in view of $V^2 = nv_0^2$

$$\log I\!\!E \exp \left\{ \lambda \frac{\gamma^\top \nabla \zeta(\theta^*)}{\|V\gamma\|} \right\} = nI\!\!E \exp \left\{ \frac{\lambda}{n^{1/2}} \frac{\gamma^\top \nabla \zeta_1(\theta^*)}{\|\nu\gamma\|} \right\} \leq \nu_0^2 \lambda^2 / 2$$

as long as $\lambda \leq n^{1/2}g_1 \leq g$. Similarly one can check (E_2) . The conditions (L_0) , (L) , and (I) follow from (ℓ_0) and $(\bar{\ell})$, and (ι) due to $D^2 = nF_0$ and $I\!\!E L(\theta) = n\bar{\ell}(\theta)$.

Below we specify the obtained general results to the i.i.d. setup.

13.3 Results in the non-penalized i.i.d. case

Here we specify the general results of previous chapters to the i.i.d. case. In particular, we explicitly state the large deviation bound and show that it yields a root-n consistency of the qMLE $\tilde{\theta}$. Then we comment on the Fisher, Wilks, and the BvM theorems.

First we describe the large deviation probability for the event $\{\tilde{\theta} \notin \Theta_0(r_0)\}$ for a fixed r_0 . The next result specifies the general large deviation statement of Theorem 9.3.1 to the finite dimensional non-penalized i.i.d. case and states the inference results.

Theorem 13.3.1. *Let (ed_0) , (ed_2) , (ℓ_0) , (ι) , and $(\bar{\ell})$ hold with*

$$\{1 - \delta(r_0)\} r_0 \geq 2z(B, x), \quad C_1 \geq \varrho(r, x), \quad r > r_0, \quad (13.3)$$

where $B = F_0^{-1/2} v_0^2 F_0^{-1/2} = D^{-1} V^2 D^{-1}$, $z(B, x)$ is given by (B.22), and

$$\varrho(r, x) \stackrel{\text{def}}{=} \nu_0 \mathfrak{z}_{\mathbb{H}}(x + \log(2r/r_0)) \omega^*/\sqrt{n}$$

with $\mathfrak{z}_{\mathbb{H}}(x) \leq C\sqrt{p+x}$. Then it holds on a set $\Omega(x)$ with $I\!\!P(\Omega(x)) \geq 1 - 5e^{-x}$

$$\sqrt{n} \|F_0^{1/2}(\tilde{\theta} - \theta^*)\| \leq r_0. \quad (13.4)$$

Furthermore, on this set $\Omega(x)$, it holds

$$\begin{aligned} \|\sqrt{n}F_0(\tilde{\theta} - \theta^*) - \xi\| &\leq C\sqrt{(p+x)^2/n}, \\ \left| \sqrt{2L(\tilde{\theta}, \theta^*)} - \|\xi\| \right| &\leq C\sqrt{(p+x)^2/n}, \\ \left| 2L(\tilde{\theta}, \theta^*) - \|\xi\|^2 \right| &\leq C\sqrt{(p+x)^3/n}. \end{aligned}$$

The constant C here depends in an explicit way on the constants a_G , g_1 , and ν_0 from our conditions, and

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} (n\mathsf{F}_0)^{-1/2} \sum_{i=1}^n \nabla \ell(Y_i, \boldsymbol{\theta}^*). \quad (13.5)$$

Proof. Condition (13.4) implies $B = \mathsf{F}_0^{-1/2} v_0^2 \mathsf{F}_0^{-1/2} \leq \alpha^2 I_p$ and thus, $\text{tr}(B) \leq \alpha^2 p$. Therefore, the value $z(B, \mathbf{x})$ fulfills $z^2(B, \mathbf{x}) \leq C(p + \mathbf{x})$. The same bound holds for $\mathfrak{z}_{\mathbb{H}}^2(\mathbf{x})$. Condition (13.3) with $1 - \delta(\mathbf{r}_0) \approx 1$ yields $\mathbf{r}_0^2 \approx 4z^2(B, \mathbf{x}) \approx C(p + \mathbf{x})$. This yields in view of $\delta(\mathbf{r}_0) \leq \delta^* \mathbf{r}_0 / \sqrt{n}$ and $\omega = \omega^* n^{-1/2}$

$$\diamond(\mathbf{r}_0, \mathbf{x}) \leq \{\delta(\mathbf{r}_0) + \nu_0 \mathfrak{z}_{\mathbb{H}}(\mathbf{x}) \omega\} \mathbf{r}_0 \leq C(p + \mathbf{x}) / \sqrt{n}.$$

Similarly

$$\Delta(\mathbf{r}_0, \mathbf{x}) \leq \{\delta(\mathbf{r}_0) + \nu_0 \mathfrak{z}_{\mathbb{H}}(\mathbf{x}) \omega\} \mathbf{r}_0^2 \leq C \sqrt{(p + \mathbf{x})^3 / n}.$$

The results follow now from general theorems of Section 11.1.

For the classical asymptotic setup when n tends to infinity, the random vector $\boldsymbol{\xi}$ from (13.5) fulfills $\text{Var}(\boldsymbol{\xi}) \leq \mathsf{F}_0^{-1/2} v_0^2 \mathsf{F}_0^{-1/2} = B$ and by the central limit theorem $\boldsymbol{\xi}$ is asymptotically normal $\mathcal{N}(0, B)$. This yields by Theorem 13.3.1 that $\sqrt{n}\mathsf{F}_0(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ is asymptotically normal $\mathcal{N}(0, B)$ as well. The correct model specification implies $B \equiv I_p$ and hence $\tilde{\boldsymbol{\theta}}$ is asymptotically efficient; see [Ibragimov and Khas'minskij \(1981\)](#). Also $2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \approx \|\boldsymbol{\xi}\|^2$ which is nearly χ^2 r.v. with p degrees of freedom. This result is known as asymptotic Wilks theorem.

In the non-asymptotic framework of this paper, the error terms still depend on n and they can only be small if n is large. However, we show in explicit way how these error terms depend on the parameter dimension. It appears that the root-n consistency result (13.4) requires “ p/n small”. The Fisher and square root Wilks results apply if “ p^2/n is small”. Finally, the Wilks expansion is valid under “ p^3/n small”. Existing statistical literature addresses the issue of a growing parameter dimension in different set-ups. The classical results by [Portnoy \(1984, 1985, 1986\)](#) provide some constraints on parameter dimension for consistency and asymptotic normality of the M-estimator for regression models. Our results are consistent with the conclusion of that papers. We refer to [Andresen and Spokoiny \(2014\)](#) for a version of such result in context of semiparametric profile estimation. That paper also provides an example of an i.i.d. model in which the Fisher expansion of Theorem 13.3.1 fails for $p^2 \geq n$. The next section demonstrates how these constraints on the parameter dimension can be relaxed by using a penalization.

13.4 Roughness penalization for an i.i.d. sample

This section discusses the impact of penalization in the case of an i.i.d. model with n observations. For penalty term $\text{pen}(\boldsymbol{\theta}) = \|G\boldsymbol{\theta}\|^2/2$, the penalized log-likelihood is given by

$L_G(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \|G\boldsymbol{\theta}\|^2/2$, where $L(\boldsymbol{\theta})$ is from (13.1). With $\boldsymbol{\theta}_G^* = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \mathbb{E}L_G(\boldsymbol{\theta})$, define

$$D_G^2 = n\mathsf{F}(\boldsymbol{\theta}_G^*) + G^2, \quad V^2 = n\mathsf{v}_0^2, \quad \boldsymbol{\xi}_G = D_G^{-1} \sum_{i=1}^n \nabla \ell(Y_i, \boldsymbol{\theta}_G^*),$$

where $\mathsf{F}(\boldsymbol{\theta}) = -\nabla^2 \mathbb{E} \ell(Y_1, \boldsymbol{\theta})$, $\mathsf{v}_0^2 = \operatorname{Var}\{\ell(Y_1, \boldsymbol{\theta}_G^*)\}$. The value p_G is defined as previously by (11.4).

Note that all the introduced quantities including the parameter set Θ , the parameter dimension p , and the effective dimension p_G , may depend on n . Here we also allow a functional parameter $\boldsymbol{\theta}$ with $p = \infty$. The main goal is to show that the presented general approach yields sharp results in this special case.

Suppose that the conditions of Section 13.2 are fulfilled. One can easily check the conditions from Section 11.3 with $\delta_G(\mathbf{r}) = C\mathbf{r}/\sqrt{n}$ and $\omega = C/\sqrt{n}$; cf. Lemma 13.2.1. The large deviation bound of Theorem 11.4.1 applies for $\mathbf{r}_G \approx 2z(B_G, \mathbf{x}) \asymp \sqrt{p_G + \mathbf{x}}$. The general statements of Theorems 11.5.1 and 11.5.2 apply with $\diamond_G(\mathbf{x}) \leq C(p_G + \mathbf{x})/\sqrt{n}$ yielding the following expansions.

Theorem 13.4.1. *Suppose also that the conditions (ed₀), (ed₂), (ℓ₀), (ℓ̄), and (ι) are fulfilled, and \mathbf{r}_G and C_1 fulfill*

$$\{1 - \delta_G(\mathbf{r}_G)\} \mathbf{r}_G \geq 2z(B_G, \mathbf{x}), \quad C_1 \geq \varrho_G(\mathbf{r}, \mathbf{x}), \quad \mathbf{r} > \mathbf{r}_G,$$

with $B_G = D_G^{-1} V^2 D_G^{-1/2}$, then on a set of dominating probability $1 - 5e^{-\mathbf{x}}$, it holds

$$\begin{aligned} \|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*) - \boldsymbol{\xi}_G\| &\leq C\sqrt{(p_G + \mathbf{x})^2/n}, \\ \left| \sqrt{2L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*)} - \|\boldsymbol{\xi}_G\| \right| &\leq C\sqrt{(p_G + \mathbf{x})^2/n}, \\ \left| 2L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*) - \|\boldsymbol{\xi}_G\|^2 \right| &\leq C\sqrt{(p_G + \mathbf{x})^3/n}. \end{aligned}$$

The constant C here depends in an explicit way on the constants a_G , g_1 , and ν_0 from our conditions.

A short look at the results for non-penalized and penalized estimates indicates that the quality of the penalized MLE $\tilde{\boldsymbol{\theta}}_G$ improves relative to the non-penalized case because the matrix D_G^2 can be much larger than D^2 , the variance of the stochastic term $\boldsymbol{\xi}_G$ is of order p_G instead of p for the variance of $\boldsymbol{\xi}$, and, simultaneously, the error terms in the Fisher and Wilks expansions become smaller due to reduction of the effective dimension p_G in place of the full dimension p .

Andresen and Spokoiny (2014) provides a simple example of estimating the squared norm $\|\boldsymbol{\theta}\|^2$ which shows that the Fisher expansion fails if p/\sqrt{n} is not small. The result can be easily adjusted to the penalized case.

13.5 BvM Theorem for the i.i.d. data

Another constraint in the BvM Theorem on the dimension growth p_n can be found in [Ghosal \(1999\)](#) for linear regression models; see the condition (2.6) $p_n^{3/2}(\log p_n)^{1/2} \eta_n \rightarrow 0$ there, in which η_n is of order $(p_n/n)^{-1/2}$ in regular situations yielding a suboptimal constraint $n^{-1}p_n^4 \log p \rightarrow 0$. [Ghosal \(2000\)](#) obtained a version of the BvM result under the condition $n^{-1}p_n^3(\log p_n) \rightarrow 0$ for a class of exponential models. A forthcoming paper [Panov and Spokoiny \(2015\)](#) presents an example illustrating that the condition $p_n^3/n \rightarrow 0$ cannot be dropped or relaxed.

The setup with growing parameter dimension is naturally used in sieve nonparametric estimation when a nonparametric model is approximated by a sequence of parametric ones. We mention papers by [Shen and Wong \(1994\)](#); [Shen \(1997\)](#), [Birgé and Massart \(1993\)](#), [Van de Geer \(1993\)](#); [van de Geer \(2002\)](#). Some minimal smoothness assumptions are normally imposed on the underlying nonparametric function which ensure that the parameter dimension of a sieve is smaller in order than the sample size.

Generalized linear models

This chapter specifies the previously obtained general results to the case of a Generalized Linear Model (GLM). Such models are frequently used in many areas and applications including categorical data analysis, Poisson and Binary regression, classification and statistical learning, density estimation.

The most important feature of such models is linearity (in parameter) of the stochastic term in the log-likelihood ratio $L(\boldsymbol{\theta})$. Another important feature is concavity of the log-likelihood and of the expected log-likelihood. This allows to substantially simplify the conditions and to obtain more strong results relative to general parametric modeling. In the most of cases we will try to give a complete proof of the results in the GLM case without referring to general statements.

We begin with a short recap of the results for the linear models. Then we switch to the GLM case and try to establish a version of all general results like concentration of the MLE $\tilde{\boldsymbol{\theta}}$, Fisher and Wilks expansion, Bernstein - von Mises theorem for uninformative and Gaussian priors for the special case of generalized linear modeling. Section 14.6 comments on the case of a random design.

14.1 Linear models

In the case of a linear model $\mathbf{Y} = \Psi^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon}$ with a given design $p \times n$ matrix Ψ under the assumption of Gaussian noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$, the standard calculus leads to the log-likelihood

$$L(\boldsymbol{\theta}) = -\frac{1}{2}(\mathbf{Y} - \Psi^\top \boldsymbol{\theta})^\top \Sigma^{-1}(\mathbf{Y} - \Psi^\top \boldsymbol{\theta}) + R,$$

where the remainder R does not depend on $\boldsymbol{\theta}$. Moreover, $L(\boldsymbol{\theta})$ is quadratic in $\boldsymbol{\theta}$ and its Hessian is constant: $\nabla^2 L(\boldsymbol{\theta}) = -\Psi \Sigma^{-1} \Psi^\top$. One can summarize as follows: with $\mathbb{E}\mathbf{Y} = \mathbf{f}$

$$\begin{aligned} D^2 &\stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}) = \Psi \Sigma^{-1} \Psi^\top, & \tilde{\boldsymbol{\theta}} &= D^{-2} \Psi \Sigma^{-1} \mathbf{Y}, \\ \boldsymbol{\xi} &\stackrel{\text{def}}{=} D^{-1} \nabla L(\boldsymbol{\theta}^*) = D^{-1} \Psi \Sigma^{-1} (\mathbf{Y} - \mathbf{f}), & \boldsymbol{\theta}^* &= D^{-2} \Psi \Sigma^{-1} \mathbf{f}. \end{aligned}$$

Moreover,

$$D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \equiv \boldsymbol{\xi}, \quad L(\tilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) \equiv \|\boldsymbol{\xi}\|^2 / 2.$$

All these results are straightforward, the last one is obtained by the Tailor expansion of the second order around $\tilde{\boldsymbol{\theta}}$ with the use of $\nabla L(\tilde{\boldsymbol{\theta}}) = 0$:

$$L(\tilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) = -\frac{1}{2} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \nabla^2 L(\tilde{\boldsymbol{\theta}}) (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \frac{1}{2} \|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 = \frac{1}{2} \|\boldsymbol{\xi}\|^2.$$

The presented derivations mean that the Fisher and Wilks expansions are *identities*, they apply for *any sample size* without *any conditions*, and are only based on *quadraticity* of the likelihood function $L(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$. The *true distribution* of \mathbf{Y} can be whatever and it is not involved at all. There is *no dimensional restrictions*. However, for *inference*, the assumption about the noise is important. It only concerns the *distribution of $\boldsymbol{\xi}$* . Let $\text{Var}(\mathbf{Y}) = \Sigma_0 \neq \Sigma$. Then with $D^2 = \Psi \Sigma^{-1} \Psi^\top$

$$\text{Var}\{\nabla L(\boldsymbol{\theta}^*)\} = \text{Var}\{\Psi \Sigma^{-1} \mathbf{Y}\} = \Psi \Sigma^{-1} \Sigma_0 \Sigma^{-1} \Psi^\top \stackrel{\text{def}}{=} V^2 \neq D^2$$

which leads to the famous *sandwich formula*

$$\text{Var}(\boldsymbol{\xi}) = \text{Var}\{D^{-1} \nabla L(\boldsymbol{\theta}^*)\} = D^{-1} V^2 D^{-1} \neq I_p.$$

The general bound for quadratic forms from Theorem B.2.2 yields the following result.

Theorem 14.1.1. *Suppose that the matrix V^2 be such that it holds*

$$\log \mathbb{E} \exp\{\mathbf{u}^\top V^{-1} \nabla L(\boldsymbol{\theta}^*)\} \leq \frac{\|\mathbf{u}\|^2}{2}, \quad \mathbf{u} \in \mathbb{R}^p, \quad \|\mathbf{u}\| \leq g.$$

Then it holds for the MLE $\tilde{\boldsymbol{\theta}}$

$$\begin{aligned} D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) &= \boldsymbol{\xi} = D^{-1} \nabla L(\boldsymbol{\theta}^*), \\ L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) &= \frac{\|\boldsymbol{\xi}\|^2}{2}, \end{aligned}$$

Moreover, for any $x > 0$

$$\mathbb{P}\left(\sqrt{2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)} > z(B, x)\right) = \mathbb{P}(\|\boldsymbol{\xi}\| > z(B, x)) \lesssim 2e^{-x},$$

where $B \stackrel{\text{def}}{=} D^{-1} V^2 D^{-1}$ and $z(B, x)$ is given by (B.22) or (B.23).

14.2 Generalized linear models (GLM)

Generalized linear models (GLM) are frequently used for modeling the data with special structure: categorical data, binary data, Poissonian and exponential data, volatility models, etc. All these examples can be treated in a unified way by a GLM approach. This section specifies the results and conditions to this case.

Let $\mathbf{Y} = (Y_1, \dots, Y_n)^\top \sim \mathbb{I}\mathcal{P}$ be a sample of independent r.v.'s. The parametric GLM is given by $Y_i \sim P_{\Psi_i^\top \boldsymbol{\theta}} \in (P_v)$, where Ψ_i are given factors in \mathbb{R}^p , $\boldsymbol{\theta} \in \mathbb{R}^p$ is the unknown parameter in \mathbb{R}^p , and (P_v) is an exponential family with canonical parametrization yielding the log-density $\ell(y, v) = yv - g(v)$ for a convex function $g(v)$. Below we suppose that the function $g(v)$ is sufficiently smooth, in particular, three times differentiable.

The (quasi) log-likelihood $L(\boldsymbol{\theta})$ can be represented in the form

$$L(\boldsymbol{\theta}) = \sum_{i=1}^n \{Y_i \Psi_i^\top \boldsymbol{\theta} - g(\Psi_i^\top \boldsymbol{\theta})\} = S^\top \boldsymbol{\theta} - A(\boldsymbol{\theta}) \quad (14.1)$$

with a random p -vector S and a function $A(\boldsymbol{\theta})$ given by

$$S \stackrel{\text{def}}{=} \sum_{i=1}^n Y_i \Psi_i, \quad A(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_i g(\Psi_i^\top \boldsymbol{\theta}).$$

The MLE $\tilde{\boldsymbol{\theta}}$ and the target $\boldsymbol{\theta}^*$ for this GLM read as

$$\begin{aligned} \tilde{\boldsymbol{\theta}} &= \operatorname{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}} \{S^\top \boldsymbol{\theta} - A(\boldsymbol{\theta})\}, \\ \boldsymbol{\theta}^* &= \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E} L(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}} \{\mathbb{E} S^\top \boldsymbol{\theta} - A(\boldsymbol{\theta})\}, \end{aligned} \quad (14.2)$$

where

$$\mathbb{E} S = \sum_{i=1}^n \mathbb{E} Y_i \Psi_i.$$

The definition of $\boldsymbol{\theta}^*$ implies the identity $\nabla \mathbb{E} L(\boldsymbol{\theta}^*) = 0$ which yields

$$\mathbb{E} S = \nabla A(\boldsymbol{\theta}^*).$$

An important feature of a GLM is that the stochastic component $\zeta(\boldsymbol{\theta})$ of $L(\boldsymbol{\theta})$ is *linear in $\boldsymbol{\theta}$* : with $\varepsilon_i = Y_i - \mathbb{E} Y_i$

$$\begin{aligned} \zeta(\boldsymbol{\theta}) &= L(\boldsymbol{\theta}) - \mathbb{E} L(\boldsymbol{\theta}) = \sum_{i=1}^n \varepsilon_i \Psi_i^\top \boldsymbol{\theta}, \\ \nabla \zeta(\boldsymbol{\theta}) &= S - \mathbb{E} S = \sum_{i=1}^n \varepsilon_i \Psi_i. \end{aligned} \quad (14.3)$$

In the contrary to the linear case, the Fisher information matrix $D^2 = \mathbb{F}(\boldsymbol{\theta}^*)$ for

$$\mathbb{F}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E}L(\boldsymbol{\theta}) = \sum_{i=1}^n \Psi_i \Psi_i^\top g''(\Psi_i^\top \boldsymbol{\theta}) \quad (14.4)$$

depends on the true data distribution via the target $\boldsymbol{\theta}^*$. As $g(\cdot)$ is convex, it holds $g''(u) \geq 0$ for any u and thus $\mathbb{F}(\boldsymbol{\theta}) \geq 0$.

14.2.1 A general deviation bound for the MLE $\tilde{\boldsymbol{\theta}}$

For any GLM, the following two features are fulfilled: the stochastic component $\zeta(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$, and the deterministic part $\mathbb{E}L(\boldsymbol{\theta})$ is concave in $\boldsymbol{\theta}$. This allows to prove in a simple and straightforward way the result about concentration of the MLE $\tilde{\boldsymbol{\theta}}$. Note that these two conditions only rely on the geometric structure of the log-likelihood $L(\boldsymbol{\theta})$ and do not refer to the true data generating process.

Recall the definition of the local vicinity $\Theta_0(\mathbf{r})$ of $\boldsymbol{\theta}^*$:

$$\Theta_0(\mathbf{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} : \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}\}. \quad (14.5)$$

Also define

$$B \stackrel{\text{def}}{=} D^{-1} \text{Var}(S) D^{-1}.$$

Theorem 14.2.1. *If for some $\mathbf{r}_0 > 0$, $\mathbb{F}(\boldsymbol{\theta})$ from (14.4) fulfill for $D^2 = \mathbb{F}(\boldsymbol{\theta}^*)$*

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)} \|D^{-1} \mathbb{F}(\boldsymbol{\theta}) D^{-1} - I_p\|_{\text{op}} \leq \delta(\mathbf{r}_0) \quad (14.6)$$

with $\delta(\mathbf{r}_0) < 1$, and if S from (14.3) follows for $\mathbf{x} > 0$ the probability bound

$$\mathbb{P}\left(\|D^{-1}(S - \mathbb{E}S)\| > z(B, \mathbf{x})\right) \leq 2e^{-\mathbf{x}}, \quad (14.7)$$

then the solution $\tilde{\boldsymbol{\theta}}$ of (14.2) satisfies

$$\mathbb{P}(\tilde{\boldsymbol{\theta}} \notin \Theta_0(\mathbf{r}_0)) \leq 2e^{-\mathbf{x}}$$

provided that

$$\mathbf{r}_0\{1 - \delta(\mathbf{r}_0)\} \geq 2z(B, \mathbf{x}). \quad (14.8)$$

Proof. The function $L(\boldsymbol{\theta})$ is concave in $\boldsymbol{\theta}$ because

$$-\nabla^2 L(\boldsymbol{\theta}) = \mathbb{F}(\boldsymbol{\theta}) \geq 0. \quad (14.9)$$

If $\tilde{\boldsymbol{\theta}} \notin \Theta_0(\mathbf{r}_0)$, denote by $\check{\boldsymbol{\theta}}$ the point at which the line connecting $\boldsymbol{\theta}^*$ and $\tilde{\boldsymbol{\theta}}$ crosses the boundary of $\Theta_0(\mathbf{r}_0)$. It is easy to see that

$$\check{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = \frac{\|D(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|}{\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \frac{\mathbf{r}_0}{\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*).$$

As $L(\boldsymbol{\theta})$ is a concave function and $\tilde{\boldsymbol{\theta}}$ is its point of maximum it follows for the point $\check{\boldsymbol{\theta}}$ on the line between $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ that

$$L(\check{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) \geq 0.$$

Therefore, it suffices to check that for each $\boldsymbol{\theta}$ with $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathbf{r}_0$ that

$$L(\boldsymbol{\theta}^*) - L(\boldsymbol{\theta}) > 0$$

on a set $\Omega(\mathbf{x})$ of probability $1 - 2e^{-\mathbf{x}}$. Then the event $\tilde{\boldsymbol{\theta}} \notin \Theta_0(\mathbf{r}_0)$ is impossible on $\Omega(\mathbf{x})$. For any such $\boldsymbol{\theta}$, we apply the second order Taylor expansion of $L(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$. By definition of $\boldsymbol{\theta}^*$, it holds $\nabla \mathbb{E} L(\boldsymbol{\theta}^*) = 0$ and thus $\nabla L(\boldsymbol{\theta}^*) = \nabla \zeta(\boldsymbol{\theta}^*) = (S - \mathbb{E} S)$. The use of (14.9), (14.6) yields now for $\boldsymbol{\xi} = D^{-1}(S - \mathbb{E} S)$ and for $\boldsymbol{\theta}$ with $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathbf{r}_0$

$$\begin{aligned} L(\boldsymbol{\theta}^*) - L(\boldsymbol{\theta}) &= -(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla L(\boldsymbol{\theta}^*) + \frac{1}{2} \|\sqrt{\mathbb{F}(\boldsymbol{\theta}^*)}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \\ &\geq -(S - \mathbb{E} S)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \frac{1 - \delta(\mathbf{r}_0)}{2} \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \\ &= -\boldsymbol{\xi}^\top D(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \frac{1 - \delta(\mathbf{r}_0)}{2} \mathbf{r}_0^2 \geq -\|\boldsymbol{\xi}\| \mathbf{r}_0 + \frac{1 - \delta(\mathbf{r}_0)}{2} \mathbf{r}_0^2. \end{aligned}$$

Here $\boldsymbol{\theta}^*$ is a point from $\Omega(\mathbf{x})$ on the interval connecting $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$. If $\|\boldsymbol{\xi}\| \leq \mathbf{r}_0 \{1 - \delta(\mathbf{r}_0)\}/2$, then this and (14.8) imply $L(\boldsymbol{\theta}^*) - L(\boldsymbol{\theta}) > 0$, and the result follows.

14.2.2 Fisher and Wilks expansions for $\tilde{\boldsymbol{\theta}}$

As a corollary, we obtain Fisher and Wilks expansions for the quasi MLE $\tilde{\boldsymbol{\theta}}$ in a generalized linear model.

Theorem 14.2.2. *Suppose the conditions of Theorem 14.2.1 for some \mathbf{r}_0 . Then it holds on a set $\Omega(\mathbf{x})$ with $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - 2e^{-\mathbf{x}}$*

$$\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi}\| \leq \mathbf{r}_0 \delta(\mathbf{r}_0),$$

$$|2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\boldsymbol{\xi}\|^2| \leq 2\mathbf{r}_0^2 \delta(\mathbf{r}_0) + \mathbf{r}_0^2 \delta^2(\mathbf{r}_0).$$

$$\left| \sqrt{2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)} - \|\boldsymbol{\xi}\| \right| \leq 3\mathbf{r}_0 \delta(\mathbf{r}_0).$$

Proof. The large deviation bound of Theorem 14.2.1 allows to restrict the whole parameter space to the local vicinity $\Theta_0(\mathbf{r}_0)$. In this vicinity, the log-likelihood ratio $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$ can be well approximated by the quadratic expansion $\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$:

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= (S - \mathbb{E}S)^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \mathbb{E}S^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - A(\boldsymbol{\theta}) + A(\boldsymbol{\theta}^*), \\ \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &\stackrel{\text{def}}{=} (S - \mathbb{E}S)^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \frac{1}{2}\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2. \end{aligned}$$

Lemma 14.2.1. Suppose (14.6) for some \mathbf{r}_0 . The difference $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is deterministic and it holds for each $\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)$

$$|L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) - \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)| \leq \frac{\delta(\mathbf{r}_0)}{2}\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \leq \frac{\delta(\mathbf{r}_0)}{2}\mathbf{r}_0^2, \quad (14.10)$$

$$\|D^{-1}\{\nabla L(\boldsymbol{\theta}) - \nabla \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\}\| \leq \mathbf{r}_0 \delta(\mathbf{r}_0). \quad (14.11)$$

Proof. The linear stochastic terms $(S - \mathbb{E}S)^\top \boldsymbol{\theta}$ are the same for $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ and $\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$. For the deterministic terms $\mathbb{E}S^\top \boldsymbol{\theta} - A(\boldsymbol{\theta})$ we use the Taylor formula of the second order at $\boldsymbol{\theta}^*$, the extreme point equation $\nabla A(\boldsymbol{\theta}^*) = \mathbb{E}S$, and the definition $D^2 = \mathbb{F}(\boldsymbol{\theta}^*)$:

$$\begin{aligned} |\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbb{E}\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)| &= \left| A(\boldsymbol{\theta}) - A(\boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla A(\boldsymbol{\theta}^*) - \frac{1}{2}\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \right| \\ &= \frac{1}{2}|(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \{\mathbb{F}(\boldsymbol{\theta}^*) - \mathbb{F}(\boldsymbol{\theta}^\circ)\}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)|, \end{aligned}$$

where $\boldsymbol{\theta}^\circ$ is a point on the interval between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$. Now the condition (14.6) implies

$$\begin{aligned} |\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbb{E}\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)| &\leq \frac{\delta(\mathbf{r}_0)}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top D^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\ &= \frac{\delta(\mathbf{r}_0)}{2}\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \leq \frac{\delta(\mathbf{r}_0)}{2}\mathbf{r}_0^2 \end{aligned}$$

and the first assertion follows. The second one can be proved similarly.

With the approximation (14.11), all the statements of the theorem follow from the general results of Theorem 11.7.2.

To complete the study of a generalized linear model, we translate the general conditions of Theorem 14.2.1 into conditions on the design Ψ and on individual errors ε_i .

14.2.3 Sufficient conditions on design and errors

This section presents a list of conditions ensuring (14.6) and (14.7). We stress once again that the true data generated process is not assumed to follow the GLM assumption. It can be completely misspecified. However, the presented conditions seem to be rather natural and easy to verify in most of applications.

Let the point $\boldsymbol{\theta}^*$ be defined by (14.2) and the local elliptic set $\Theta_0(\mathbf{r})$ is given by (14.5). We also denote

$$d_i^2 \stackrel{\text{def}}{=} g''(\Psi_i^\top \boldsymbol{\theta}^*), \quad i = 1, \dots, n,$$

and

$$\mathbf{H}^2 = \frac{1}{n} \sum_{i=1}^n d_i^2 \Psi_i \Psi_i^\top.$$

Obviously $D^2 = n\mathbf{H}^2$. Conditions below implicitly assume that the d_i 's are positive and bounded away from zero. This is fulfilled automatically if $g''(\mathbf{v})$ is a positive and continuous function because each local set $\Theta_0(\mathbf{r})$ is compact.

- **Design regularity** is measured via its third directional moments: for a constant a_Ψ

$$\sup_{\mathbf{u} \in \mathcal{S}_p} \frac{1}{n} \sum_{i=1}^n |d_i \mathbf{u}^\top \mathbf{H}^{-1} \Psi_i|^3 \leq a_\Psi. \quad (14.12)$$

We impose one more technical condition that the values $|d_i \mathbf{u}^\top \mathbf{H}^{-1} \Psi_i|$ are uniformly bounded:

$$\sup_{\mathbf{u} \in \mathcal{S}_p} \max_{i=1, \dots, n} |d_i \mathbf{u}^\top \mathbf{H}^{-1} \Psi_i| \leq \delta_\Psi. \quad (14.13)$$

In the case of a regular or random design, the value δ_Ψ is of order $\sqrt{p/n}$. This condition will only be used by checking **(ED₀)** and can be relaxed.

- **Exponential moments of the errors** $\varepsilon_i = Y_i - \mathbb{E}Y_i$. Suppose that for some values \mathbf{s}_i and fixed constants $C_0, \lambda_0 > 0$

$$\mathbb{E} \exp(\lambda_0 \mathbf{s}_i^{-1} \varepsilon_i) \leq C_0, \quad i = 1, \dots, n. \quad (14.14)$$

This condition means that the errors ε_i have exponential moments. In most of cases one can use $\mathbf{s}_i^2 = \text{Var}(Y_i)$. Condition (14.14) implies that there are another constants $g_1 \leq \lambda_0$ and ν_0 such that the following condition is fulfilled:

$$\mathbb{E} \exp(\lambda \mathbf{s}_i^{-1} \varepsilon_i) \leq \frac{1}{2} \nu_0^2 \lambda^2, \quad i = 1, \dots, n, \quad |\lambda| \leq g_1. \quad (14.15)$$

This follows from the fact that each function $\log \mathbb{E} \exp(\lambda_0 \mathbf{s}_i^{-1} \varepsilon_i)$ analytic in λ in a vicinity of the point zero and can be well approximated by $\lambda^2/2$; see [Golubev and Spokoiny \(2009\)](#) for more details.

- **Noise homogeneity** is measured by the ratio of \mathbf{s}_i and d_i :

$$a_s \stackrel{\text{def}}{=} \max_{i=1, \dots, n} \mathbf{s}_i / d_i. \quad (14.16)$$

- **Smoothness of the link function** $g(v)$ can be measured by its third derivative, more precisely, by the ratio $|g'''(v)|/|g''(v)|^{3/2}$. Namely, for any $\mathbf{r} > 0$, there is a constant $a_g(\mathbf{r})$ such that

$$\max_{i=1,\dots,n} |d_i^{-3} g'''(\Psi_i^\top \boldsymbol{\theta})| \leq a_g(\mathbf{r}), \quad \boldsymbol{\theta} \in \Theta_0(\mathbf{r}), \quad i = 1, \dots, n. \quad (14.17)$$

- **Identifiability** is measured by relationship between the matrices D^2 and V^2 , where the matrix V^2 defined as

$$V^2 \stackrel{\text{def}}{=} \sum_{i=1}^n \mathbf{s}_i^2 \Psi_i \Psi_i^\top. \quad (14.18)$$

If the the observation Y_i follow the GLM assumption P_{v_i} for $v_i = \Psi_i^\top \boldsymbol{\theta}^*$, that is, the model is correctly specified, then $\text{Var}(Y_i) = g''(v_i)$ and the matrices V^2 and D^2 coincide. In the general case under a possible model misspecification, the matrices V^2 and D^2 may be different. In this case we need an identifiability condition

$$\underline{\alpha}^2 D^2 \leq V^2 \leq \alpha^2 D^2 \quad (14.19)$$

for some positive constants $\underline{\alpha} \leq \alpha$. This condition can be spelled out as

$$\underline{\alpha}^2 \sum_{i=1}^n d_i^2 \Psi_i \Psi_i^\top \leq \sum_{i=1}^n \mathbf{s}_i^2 \Psi_i \Psi_i^\top \leq \alpha^2 \sum_{i=1}^n d_i^2 \Psi_i \Psi_i^\top.$$

Theorem 14.2.3. Suppose that the conditions (14.12), (14.13), (14.15), (14.16), (14.17), and (14.19) hold and define $\mathbf{g} \stackrel{\text{def}}{=} \mathbf{g}_1 \underline{\alpha} / (a_s \delta_\Psi)$. Fix

$$\mathbf{r}_0 = 4\nu_0 z(B, \mathbf{x}) \quad (14.20)$$

for $z(B, \mathbf{x})$ from (B.22) with $B = D^{-1} V^2 D^{-1}$ and with such defined \mathbf{g} . Suppose also that a_Ψ and $a_g(\mathbf{r}_0)$ ensure

$$2a_g(\mathbf{r}_0) a_\Psi \mathbf{r}_0 < \sqrt{n}. \quad (14.21)$$

Then the conditions of Theorem 14.2.1 are fulfilled with $\delta(\mathbf{r}_0) \leq a_g(\mathbf{r}_0) a_\Psi \mathbf{r}_0 / \sqrt{n}$ and the results of this theorem continue to apply.

Proof. Let \mathbf{r}_0 be fixed by (14.20). First we bound the value $\delta(\mathbf{r}_0)$.

Lemma 14.2.2. The condition (14.6) is fulfilled with $\delta(\mathbf{r}_0) = a_g(\mathbf{r}_0) a_\Psi \mathbf{r}_0 n^{-1/2}$.

Proof. For each $\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)$, the difference $\mathbb{F}(\boldsymbol{\theta}) - \mathbb{F}(\boldsymbol{\theta}^*)$ can be written in the form

$$\mathbb{F}(\boldsymbol{\theta}) - \mathbb{F}(\boldsymbol{\theta}^*) = \sum_{i=1}^n \{g''(\Psi_i^\top \boldsymbol{\theta}) - g''(\Psi_i^\top \boldsymbol{\theta}^*)\} \Psi_i \Psi_i^\top.$$

Fix a unit vector $\mathbf{u} \in \mathbb{R}^p$ and define $\boldsymbol{\gamma} = D^{-1}\mathbf{u} = n^{-1/2}\mathsf{H}^{-1}\mathbf{u}$. Also define $\boldsymbol{\alpha} = D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$, so that $\|\boldsymbol{\alpha}\| \leq \mathbf{r}_0$. Then

$$\begin{aligned} \mathbf{u}^\top D^{-1}\{\mathbb{F}(\boldsymbol{\theta}) - \mathbb{F}(\boldsymbol{\theta}^*)\}D^{-1}\mathbf{u} &= \boldsymbol{\gamma}^\top \{\mathbb{F}(\boldsymbol{\theta}) - \mathbb{F}(\boldsymbol{\theta}^*)\}\boldsymbol{\gamma} \\ &= \sum_{i=1}^n \{g''(\Psi_i^\top \boldsymbol{\theta}) - g''(\Psi_i^\top \boldsymbol{\theta}^*)\} |\Psi_i^\top \boldsymbol{\gamma}|^2 = \sum_{i=1}^n g'''(\Psi_i^\top \boldsymbol{\theta}^\circ) \Psi_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*) |\Psi_i^\top \boldsymbol{\gamma}|^2 \end{aligned}$$

for a point $\boldsymbol{\theta}^\circ$ on the interval between $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$ (possibly depending on \mathbf{u}). The use of (14.17) and of the Hölder inequality helps to bound for $\boldsymbol{\alpha} = \mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$

$$\begin{aligned} |\mathbf{u}^\top D^{-1}\{\mathbb{F}(\boldsymbol{\theta}) - \mathbb{F}(\boldsymbol{\theta}^*)\}D^{-1}\mathbf{u}| &\leq \frac{1}{n} \sum_{i=1}^n a_g(\mathbf{r}_0) |d_i \boldsymbol{\alpha}^\top \mathsf{H}^{-1} \Psi_i| |d_i \mathbf{u}^\top \mathsf{H}^{-1} \Psi_i|^2 \leq a_g(\mathbf{r}_0) a_\Psi \|\boldsymbol{\alpha}\| \end{aligned}$$

and the result follows by $\|\boldsymbol{\alpha}\| \leq \mathbf{r}_0/\sqrt{n}$.

This lemma and (14.21) imply $\delta(\mathbf{r}_0) < 1/2$. Now we check (14.7).

Lemma 14.2.3. *Let the errors $\varepsilon_i = Y_i - \mathbb{E}Y_i$ be independent and follow (14.15). Then it holds under conditions (14.19), (14.13), and (14.16) for $\mathbf{g} = \mathbf{g}_1 \underline{a}/(\delta_\Psi a_s)$*

$$\log \mathbb{E} \exp\{\lambda \mathbf{u}^\top V^{-1}(S - \mathbb{E}S)\} \leq \frac{\nu_0^2}{2} \lambda^2, \quad \lambda \leq \mathbf{g},$$

with V^2 from (14.18).

Proof. The formula (14.3) and independence of the ε_i 's imply for any unit vector $\mathbf{u} \in \mathbb{R}^p$ and $|\lambda| \leq \mathbf{g}$

$$\log \mathbb{E} \exp\{\lambda \mathbf{u}^\top V^{-1}(S - \mathbb{E}S)\} = \sum_{i=1}^n \log \mathbb{E} \exp(\lambda_i \mathbf{s}_i^{-1} \varepsilon_i),$$

where $\lambda_i = \lambda \mathbf{s}_i \mathbf{u}^\top V^{-1} \Psi_i$. Conditions (14.19), (14.13), and (14.16) imply

$$\sup_{\mathbf{u} \in \mathcal{S}_p} |\mathbf{u}^\top V^{-1} \Psi_i| \mathbf{s}_i \leq \sup_{\mathbf{u} \in \mathcal{S}_p} \underline{a}^{-1} |\mathbf{u}^\top D^{-1} \Psi_i| \mathbf{s}_i \leq \underline{a}^{-1} a_s |d_i \mathbf{u}^\top D^{-1} \Psi_i| \leq \underline{a}^{-1} a_s \delta_\Psi$$

and hence for any $|\lambda| \leq \mathbf{g}$,

$$|\lambda_i| = \mathbf{g} \underline{a}^{-1} a_s \delta_\Psi \leq \mathbf{g}_1.$$

Therefore, by (14.15) and the definition of V^2

$$\begin{aligned} \log \mathbb{E} \exp\{\lambda \mathbf{u}^\top V^{-1}(S - \mathbb{E}S)\} &\leq \sum_{i=1}^n \frac{\nu_0^2 \lambda_i^2}{2} \\ &= \frac{\nu_0^2 \lambda^2}{2} \sum_{i=1}^n \mathbf{u}^\top V^{-1} (\Psi_i \Psi_i^\top \mathbf{s}_i^2) V^{-1} \mathbf{u} = \frac{\nu_0^2 \lambda^2}{2}, \end{aligned}$$

and the assertion follows.

The result of Lemma 14.2.3 provides exponential moments of ξ and one can apply Theorem B.2.1 from Section B.2 yielding the bound (14.7) under the condition

$$\frac{1 - \delta(\mathbf{r}_0)}{2} \mathbf{r}_0 \geq \nu_0 z(p, \mathbf{x})$$

which is obviously fulfilled for our choice of $\mathbf{r}_0 = 4\nu_0 z(B, \mathbf{x})$ for $B = D^{-1}V^2D^{-1}$ in view of $\delta(\mathbf{r}_0) < 1/2$. This will also provide (14.7). All the conditions of Theorem 14.2.1 have been checked.

14.3 Nonparametric sieve GLM estimation

Nonparametric estimation can often be reduced to generalized linear modeling for a infinite dimensional parameter space. Unfortunately, the conditions like (ED_0) or (L_0) implicitly involve the parameter dimension p , and become intractable when p grows to infinity. To get the rid of this issue, one of two options is usually applied: sieve parametric estimation restricts the estimation problem to a finite dimensional subspace of the infinite dimensional parameter set Θ , while the penalized maximum likelihood estimation is based on maximization of the properly penalized likelihood function. Both approaches yields some estimation bias, because the underlying model becomes misspecified. Below we show how the previously developed general setup can be applied to the nonparametric problems. This section discusses the sieve parametric approach.

Let $\boldsymbol{\theta}^*$ be the nonparametric true point:

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} IEL(\boldsymbol{\theta}).$$

Let m be the index for the sieve subspace Θ_m of dimension p_m , and

$$\begin{aligned} \boldsymbol{\theta}_m^* &\stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta_m} IEL(\boldsymbol{\theta}). \\ \tilde{\boldsymbol{\theta}}_m &\stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta_m} L(\boldsymbol{\theta}). \end{aligned}$$

The previous approach only operates with Θ_m . Now we specify the results using the true point $\boldsymbol{\theta}^*$ from the full nonparametric model. Define

$$\begin{aligned} \mathbb{F}_m(\boldsymbol{\theta}) &\stackrel{\text{def}}{=} -\nabla_m^2 IEL(\boldsymbol{\theta}), \\ D_m^2 &\stackrel{\text{def}}{=} \mathbb{F}_m(\boldsymbol{\theta}^*). \end{aligned} \tag{14.22}$$

Here ∇_m means differentiation along the sieve subspace, that is, D_m^2 is a $p_m \times p_m$ matrix. The only difference in comparison with the finite dimensional case is that the

Hessian matrix is computed at $\boldsymbol{\theta}^*$ instead of $\boldsymbol{\theta}_m^*$. Also define the local vicinity $\Theta_m(\mathbf{r})$ of $\boldsymbol{\theta}^*$:

$$\Theta_m(\mathbf{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} \in \Theta_m : \|D_m(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}\}.$$

This vicinity is centered at another point $\boldsymbol{\theta}^*$ and we will have to account for the bias to ensure the concentration properties of the MLE $\tilde{\boldsymbol{\theta}}_m$.

Let also S_m be the projection of the score vector S on the sieve subspace \mathcal{S}_m and

$$B_m \stackrel{\text{def}}{=} D_m^{-1} \text{Var}(S_m) D_m^{-1}.$$

Theorem 14.3.1. *If for some $\mathbf{r}_m > 0$, $\mathbb{F}_m(\boldsymbol{\theta})$ from (14.22) fulfill for $D_m^2 = \mathbb{F}_m(\boldsymbol{\theta}^*)$*

$$\sup_{\boldsymbol{\theta} \in \Theta_m(\mathbf{r}_m)} \|D_m^{-1} \mathbb{F}_m(\boldsymbol{\theta}) D_m^{-1} - I_m\|_{\text{op}} \leq \delta_m(\mathbf{r}_m)$$

with $\delta_m(\mathbf{r}_m) < 1$, and if S_m from (14.3) follows for $\mathbf{x} > 0$ the probability bound

$$\mathbb{P}\left(\|D_m^{-1}(S_m - \mathbb{E} S_m)\| > z(B_m, \mathbf{x})\right) \leq 2e^{-\mathbf{x}},$$

then the solution $\tilde{\boldsymbol{\theta}}_m$ of (14.2) satisfies

$$\mathbb{P}(\tilde{\boldsymbol{\theta}}_m \notin \Theta_m(\mathbf{r}_m)) \leq 2e^{-\mathbf{x}}$$

provided that

$$(\mathbf{r}_m - \|b_m\|)\{1 - \delta_m(\mathbf{r}_m)\} \geq 2z(B_m, \mathbf{x}), \quad (14.23)$$

where

$$b_m \stackrel{\text{def}}{=} D_m(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*). \quad (14.24)$$

Proof. Theorem 14.2.1 applies to the sieve MLE $\tilde{\boldsymbol{\theta}}_m$ and states the concentration result in the vicinity of $\boldsymbol{\theta}_m^*$. An increase of the radius by the bias term $\|b_m\|$ yields the similar result in the vicinity of $\boldsymbol{\theta}^*$.

Now we adapt the Fisher and Wilks expansions to the sieve approach.

Theorem 14.3.2. *Suppose the conditions of Theorem 14.3.1 for some \mathbf{r}_m satisfying (14.23) for the bias b_m from (14.24). Then on a set $\Omega_m(\mathbf{x})$ with $\mathbb{P}(\Omega_m(\mathbf{x})) \geq 1 - 2e^{-\mathbf{x}}$*

$$\|D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*) - \boldsymbol{\xi}_m\| \leq \mathbf{r}_m \delta_m(\mathbf{r}_m).$$

This particularly yields

$$\|D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*) - \boldsymbol{\xi}_m - b_m\| \leq \mathbf{r}_m \delta_m(\mathbf{r}_m).$$

Similarly one can state the Wilks and square-root Wilks results. We only present the square-root version.

Theorem 14.3.3. Suppose the conditions of Theorem 14.3.1 for some \mathbf{r}_m satisfying (14.23). Then it holds on a set $\Omega_m(\mathbf{x})$ with $\mathbb{P}(\Omega_m(\mathbf{x})) \geq 1 - 2e^{-\mathbf{x}}$

$$\begin{aligned} \left| \sqrt{2L(\tilde{\boldsymbol{\theta}}_m, \boldsymbol{\theta}_m^*)} - \|\boldsymbol{\xi}_m\| \right| &\leq 3\mathbf{r}_m \delta(\mathbf{r}_m), \\ \left| \sqrt{2L(\tilde{\boldsymbol{\theta}}_m, \boldsymbol{\theta}^*)} - \|\boldsymbol{\xi}_m + b_m\| \right| &\leq 5\mathbf{r}_m \delta(\mathbf{r}_m). \end{aligned} \quad (14.25)$$

To be done: please complete

14.4 Estimation for a penalized GLM

This section briefly discusses what will be changed if the GLM (14.1) is penalized by a roughness penalty term $\|G\boldsymbol{\theta}\|^2/2$. The corresponding penalized log-likelihood $L_G(\boldsymbol{\theta})$ reads as

$$L_G(\boldsymbol{\theta}) = S^\top \boldsymbol{\theta} - A(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2/2.$$

The penalized MLE and its target are defined by maximizing $L_G(\boldsymbol{\theta})$ and its expectation:

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_G &\stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \{S^\top \boldsymbol{\theta} - A(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2/2\}, \\ \boldsymbol{\theta}_G^* &\stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \{\mathbb{E} S^\top \boldsymbol{\theta} - A(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2/2\}. \end{aligned} \quad (14.26)$$

Penalization introduces some bias which measures the difference between the original target $\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta})$ and the the penalized target $\boldsymbol{\theta}_G^*$. Define also the information matrix D_G by $D_G^2 = \mathbb{F}_G(\boldsymbol{\theta}^*)$ for

$$\mathbb{F}_G(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{F}(\boldsymbol{\theta}) + G^2 = \sum_{i=1}^n \Psi_i \Psi_i^\top g''(\Psi_i^\top \boldsymbol{\theta}) + G^2. \quad (14.27)$$

One can write

$$D_G^2 = D^2 + G^2$$

and hence, the use of penalization leads to a growth of the “information matrix” D_G^2 relative to the matrix D^2 for the non-penalized case. The stochastic term $(S - \mathbb{E} S)^\top \boldsymbol{\theta}$ of $L_G(\boldsymbol{\theta})$ remains the same as in the non-penalized case, thus, the matrix V^2 from (14.18) can be used here as well and the identifiability condition (14.19) continues to hold.

In the GLM case as in the general situation the penalized MLE $\tilde{\boldsymbol{\theta}}_G$ concentrates in a local vicinity of a point $\boldsymbol{\theta}_G^*$. However, below we consider a local vicinity $\Theta_{0,G}(\mathbf{r})$ around $\boldsymbol{\theta}^*$ which is defined as

$$\Theta_{0,G}(\mathbf{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} : \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}\}.$$

This neighborhood can be much smaller than a similar non-penalized counterpart defined with matrix D in place of D_G . From the other side, we have to account for the bias

$$b_G \stackrel{\text{def}}{=} D_G(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*).$$

Theorem 14.4.1. *Let S from (14.3) follow for $\mathbf{x} > 0$ the probability bound*

$$\|D_G^{-1}(S - IES)\| \leq z(B_G, \mathbf{x})$$

on a random set $\Omega_G(\mathbf{x})$ with $I\!\!P(\Omega_G(\mathbf{x})) \geq 1 - 2e^{-\mathbf{x}}$. Here $B_G = D_G^{-1}V^2D_G^{-1}$ and $z(B_G, \mathbf{x}) \leq \sqrt{p_G} + \sqrt{2\lambda_{\max}(B_G)\mathbf{x}}$; see (B.22). Let also the radius $\mathbf{r}_G > 0$ be fixed in such a way that

$$\sup_{\boldsymbol{\theta} \in \Theta_{0,G}(\mathbf{r}_0)} \|D_G^{-1}\mathbb{F}_G(\boldsymbol{\theta})D_G^{-1} - I_p\|_{\text{op}} \leq \delta(\mathbf{r}_G)$$

for the matrix function $\mathbb{F}_G(\boldsymbol{\theta})$ from (14.27) and $\delta(\mathbf{r}_G) < 1$, and

$$(\mathbf{r}_G - \|b_G\|)\{1 - \delta(\mathbf{r}_G)\} \geq 2z(B_G, \mathbf{x}),$$

for the bias $b_G = D_G(\boldsymbol{\theta}_G^ - \boldsymbol{\theta}^*)$. Then the solution $\tilde{\boldsymbol{\theta}}_G$ of (14.26) satisfies on $\Omega_G(\mathbf{x})$*

$$\tilde{\boldsymbol{\theta}}_G \in \Theta_{0,G}(\mathbf{r}_G).$$

For $\xi_G \stackrel{\text{def}}{=} D_G^{-1}(S - IES)$, it holds on $\Omega_G(\mathbf{x})$

$$\begin{aligned} \|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*) - \xi_G - b_G\| &\leq \mathbf{r}_G \delta(\mathbf{r}_G), \\ \left| \sqrt{2L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*)} - \|\xi_G\| \right| &\leq 3\mathbf{r}_G \delta(\mathbf{r}_G), \\ \left| \sqrt{2L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*)} - \|\xi_G + b_G\| \right| &\leq 5\mathbf{r}_G \delta(\mathbf{r}_G). \end{aligned}$$

The proof of the non-penalized case applies here with obvious changes in notation. However, at one place the difference is essential. Namely, the radius \mathbf{r}_G can be much smaller and it depends on the effective dimension $p_G = \text{tr}(B_G) = \text{tr}(D_G^{-1}V^2D_G^{-1})$ rather than on the total dimension p .

14.5 BvM Theorem for a GLM

This section discusses how the general BvM result can be stated for the case of a generalized linear model. Basically we confirm the conclusion of the previous section: such a result is valid almost for free and only relies to the nice geometric structure of the log-likelihood. However, in the contrary to the frequentist estimation, dimensionality of the parameter space matters a lot. We begin with the case of a non-informative prior. Then we extend to the case of a Gaussian prior.

14.5.1 A non-informative prior

Concavity of the log-likelihood function $L(\boldsymbol{\theta})$ for a GLM can be used to establish a global large deviation bound for the MLE $\tilde{\boldsymbol{\theta}}$ without global conditions on the expected log-likelihood $\mathbb{E}L(\boldsymbol{\theta})$. An interesting question is whether the same conclusion holds for the BvM result. The question is non-trivial because the general BvM Theorem 10.1.1 uses the global condition (\mathcal{L}) on $\mathbb{E}L(\boldsymbol{\theta})$ to bound from above the integral of $\exp\{L(\boldsymbol{\theta})\}$ on the complement of the local vicinity $\Theta_0(\mathbf{r}_0)$ of $\boldsymbol{\theta}^*$. Now we want to do the same only using the local quadraticity and global concavity of $\mathbb{E}L(\boldsymbol{\theta})$. This is a bit harder to show than just check $L(\boldsymbol{\theta}) < 0$ for $\boldsymbol{\theta} \notin \Theta_0(\mathbf{r}_0)$. The main difficulty is to bound the integral of $\exp\{L(\boldsymbol{\theta})\}$ over the complement of the local vicinity $\Theta_0(\mathbf{r}_0)$, because it is a huge set in a high dimensional space. In particular, it requires to choose \mathbf{r}_0 larger by factor about 2 than in the case of estimation. The proof heavily relies to the concentration bound for the Gaussian law. Recall the notation $z(p, \mathbf{x})$ for the quantiles of a χ_p^2 random variable: $\mathbb{P}\{\|\boldsymbol{\gamma}\| \geq z(p, \mathbf{x})\} = e^{-\mathbf{x}}$, where $\boldsymbol{\gamma}$ is a standard normal vector in \mathbb{R}^p . One can use the bound $z(p, \mathbf{x}) \leq \sqrt{p} + \sqrt{2\mathbf{x}}$.

Theorem 14.5.1. Suppose (\mathcal{L}_0) and (\mathcal{I}) . Let also \mathbf{r}_0 be such that

$$\mathbf{r}_0^2 \geq \mathbf{x} + 2p + 4z^2(B, \mathbf{x}).$$

Then it holds on a random set $\Omega(B, \mathbf{x})$

$$\mathbb{P}(\boldsymbol{\vartheta} \notin \Theta_0(\mathbf{r}_0) \mid \mathbf{Y}) \leq e^{-\mathbf{x}}.$$

Also on $\Omega(B, \mathbf{x})$, it holds with $\Delta_o(\mathbf{x}) = \mathbf{r}_0 \delta(\mathbf{r}_0, \mathbf{x})$, for any measurable set $A \subset \Theta_0(\mathbf{r}_0)$

$$\begin{aligned} \mathbb{P}(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}} \in A \mid \mathbf{Y}) &\geq \exp\{-2\Delta_o(\mathbf{x}) - 2e^{-\mathbf{x}}\} \mathbb{P}(D_G^{-1}\boldsymbol{\gamma} \in A) - e^{-\mathbf{x}}, \\ \mathbb{P}(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}} \in A \mid \mathbf{Y}) &\leq \exp\{2\Delta_o(\mathbf{x}) + e^{-\mathbf{x}}\} \mathbb{P}(D_G^{-1}\boldsymbol{\gamma} \in A) + e^{-\mathbf{x}}. \end{aligned}$$

Remark 14.5.1. Again we only impose local conditions on $\Theta_0(\mathbf{r}_0)$, the convexity of A helps to handle the posterior on the complement part $\boldsymbol{\Theta} \setminus \Theta_0(\mathbf{r}_0)$. Compared to the

case of maximum likelihood estimation, the radius for contraction has to be only slightly larger.

Proof. The main benefit of a GLM is linearity of the stochastic term:

$$\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top (S - \mathbb{E}S) = \boldsymbol{\xi}^\top D(\boldsymbol{\theta} - \boldsymbol{\theta}^*).$$

This allows to apply the result of Propositions 10.4.1 and 10.4.2 with $\varrho(\mathbf{r}, \mathbf{x}) = 0$. We restate the claims for the GLM case using that $\|\boldsymbol{\xi}\| \leq z(B, \mathbf{x})$ on $\Omega(B, \mathbf{x})$; see Theorem B.2.2. The proof is started with the result describing the local properties of the posterior. It basically refines the statement of Proposition 10.4.1 to the GLM case.

Proposition 14.5.1. *Suppose (14.10) for*

$$\mathbf{r}_0 \geq z(B, \mathbf{x}) + z(p, \mathbf{x}). \quad (14.28)$$

Then for any nonnegative function $f(\cdot)$ on \mathbb{R}^p , it holds on $\Omega(B, \mathbf{x})$

$$\mathbb{E}^\circ[f(D(\boldsymbol{\vartheta} - \check{\boldsymbol{\vartheta}})) \mathbb{I}\{\boldsymbol{\vartheta} \in \Theta_0(\mathbf{r}_0)\}] \leq \exp\{\Delta_\circ^+(\mathbf{x})\} \mathbb{E}f(\boldsymbol{\gamma}), \quad (14.29)$$

where with $\Delta_\circ(\mathbf{x}) = \mathbf{r}_0 \delta(\mathbf{r}_0, \mathbf{x})$

$$\begin{aligned} \Delta_\circ^+(\mathbf{x}) &= \Delta_\circ(\mathbf{x}) + \nu(\mathbf{r}_0) \leq \Delta_\circ(\mathbf{x}) + 2e^{-\mathbf{x}}, \\ \nu(\mathbf{r}_0) &\stackrel{\text{def}}{=} -\log \mathbb{P}^\circ(\|\boldsymbol{\gamma} + \boldsymbol{\xi}\| \leq \mathbf{r}_0) \leq 2e^{-\mathbf{x}}. \end{aligned} \quad (14.30)$$

The proof repeats the similar proof of Proposition 10.4.1 and is omitted.

The next step of the proof is to show that convexity of A yields condition **(L)**.

Lemma 14.5.1. *For $C_0 \stackrel{\text{def}}{=} 1 - \delta(\mathbf{r}_0)$ and any $\boldsymbol{\theta}$ with $\mathbf{r} = \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \geq \mathbf{r}_0$, it holds*

$$\mathbb{E}L(\boldsymbol{\theta}^*) - \mathbb{E}L(\boldsymbol{\theta}) \geq C_0 \left(\mathbf{r}_0 \mathbf{r} - \frac{\mathbf{r}_0^2}{2} \right). \quad (14.31)$$

Proof. The expected log-likelihood ratio $\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ in the GLM case reads as

$$\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = A(\boldsymbol{\theta}) - A(\boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla A(\boldsymbol{\theta}^*)$$

which is obviously concave in $\boldsymbol{\theta}$ with the point of minimum at $\boldsymbol{\theta}^*$; see Lemma 14.2.1. Moreover, by this lemma, condition **(L₀)** implies for any $\boldsymbol{\theta}^\circ$ with $\|D(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*)\| = \mathbf{r}_0$

$$-\mathbb{E}L(\boldsymbol{\theta}^\circ, \boldsymbol{\theta}^*) \geq \frac{C_0}{2} \mathbf{r}_0^2. \quad (14.32)$$

Similarly

$$\|D^{-1}\nabla A(\boldsymbol{\theta}^{\circ}) - D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}_0 \delta(\mathbf{r}_0); \quad (14.33)$$

see Lemma 14.2.1. Consider $\boldsymbol{\theta} = \boldsymbol{\theta}^* + \mathbf{r}\mathbf{u}$ for $\mathbf{r} \geq \mathbf{r}_0$ and $\mathbf{u} = D(\boldsymbol{\theta}^{\circ} - \boldsymbol{\theta}^*)$. Then by (14.32) and (14.33)

$$\begin{aligned} A(\boldsymbol{\theta}) - A(\boldsymbol{\theta}^*) &\geq A(\boldsymbol{\theta}^{\circ}) - A(\boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^{\circ})^\top \nabla A(\boldsymbol{\theta}^{\circ}) \\ &= A(\boldsymbol{\theta}^{\circ}) - A(\boldsymbol{\theta}^*) + (\mathbf{r} - \mathbf{r}_0)\mathbf{u}^\top D^{-1}\nabla A(\boldsymbol{\theta}^{\circ}) \\ &\geq \frac{C_0}{2}\mathbf{r}_0^2 + \{1 - \delta(\mathbf{r}_0)\}\mathbf{r}_0(\mathbf{r} - \mathbf{r}_0) = C_0\left(\mathbf{r}_0\mathbf{r} - \frac{1}{2}\mathbf{r}_0^2\right). \end{aligned}$$

This implies (14.31).

The next important step in our analysis is to check that the posterior $\boldsymbol{\vartheta} \mid \mathbf{Y}$ concentrates in a small vicinity $\Theta_0(\mathbf{r}_0)$ of the point $\boldsymbol{\theta}^*$ with a properly selected \mathbf{r}_0 . This will be described by using the random quantity

$$\rho(\mathbf{r}_0) \stackrel{\text{def}}{=} \frac{\int_{\Theta \setminus \Theta_0(\mathbf{r}_0)} \exp\{L(\boldsymbol{\theta})\} d\boldsymbol{\theta}}{\int_{\Theta_0(\mathbf{r}_0)} \exp\{L(\boldsymbol{\theta})\} d\boldsymbol{\theta}} = \frac{\int_{\Theta \setminus \Theta_0(\mathbf{r}_0)} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} d\boldsymbol{\theta}}{\int_{\Theta_0(\mathbf{r}_0)} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} d\boldsymbol{\theta}}.$$

Proposition 14.5.2. *Let \mathbf{r}_0 fulfill (14.28) and*

$$\{1 - \delta(\mathbf{r}_0)\}\mathbf{r}_0 \geq 2\{(p + \mathbf{x})^{1/2} + (p + \mathbf{x})^{-1/2}\},$$

then with $\Delta_{\circ}^+(\mathbf{x})$ from (14.30), it holds on a set $\Omega(B, \mathbf{x})$

$$\rho(\mathbf{r}_0) \leq \exp\left\{-\frac{p + \mathbf{x}}{2} + \Delta_{\circ}^+(\mathbf{x})\right\}.$$

Proof. We follow the proof of the general BvM result in Proposition 10.4.2 and use the notations of that proof. For the denominator of $\rho(\mathbf{r}_0)$ we apply a special case of (14.29) with $f(\mathbf{u}) \equiv 1$: by definition of $\nu(\mathbf{r}_0)$, it follows

$$\int_{\Theta_0(\mathbf{r}_0)} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} d\boldsymbol{\theta} \geq \exp\{-\Delta_{\circ}(\mathbf{x})\} \frac{1}{\det(D)} \int_{\|\mathbf{u}\| \leq \mathbf{r}_0} \exp\left(-\frac{1}{2}\|\mathbf{u}\|^2 + \boldsymbol{\xi}^\top \mathbf{u}\right) d\mathbf{u};$$

It remains to bound from above the integral over the complement of the local set $\Theta_0(\mathbf{r}_0)$. Convexity of the function $-\mathbb{E}L(\boldsymbol{\theta})$ and (L₀) imply for any $\boldsymbol{\theta}$ with $\mathbf{r} = \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \geq \mathbf{r}_0$ that

$$\mathbb{E}L(\boldsymbol{\theta}) - \mathbb{E}L(\boldsymbol{\theta}^*) \geq C_0\left(\mathbf{r}_0\mathbf{r} - \frac{1}{2}\mathbf{r}_0^2\right);$$

with $C_0 = \{1 - \delta(\mathbf{r}_0)\}$; see Lemma 14.5.1. The decomposition $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \boldsymbol{\xi}^\top D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ implies

$$\int_{\Theta \setminus \Theta_0(\mathbf{r}_0)} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} d\boldsymbol{\theta} \leq \frac{1}{\det(D)} \int_{\|\mathbf{u}\| \geq \mathbf{r}_0} \exp\left(-C_0\mathbf{r}_0\|\mathbf{u}\| + \frac{1}{2}C_0\mathbf{r}_0^2 + \boldsymbol{\xi}^\top \mathbf{u}\right) d\mathbf{u}.$$

Theorem A.2.3 yields

$$\rho(\mathbf{r}_0) \leq \exp\left\{\Delta_{\circ}^{+}(\mathbf{x}) - \frac{p + \mathbf{x}}{2}\right\}.$$

This completes the proof of the BvM result; cf. the proof of Theorem 10.1.1.

14.5.2 Nonparametric BvM with a Gaussian prior

The BvM result for the non-informative prior can be extended to the case of a Gaussian prior as described in Section 11.9.1. Let the prior $\Pi(\cdot)$ be zero mean with the covariance operator G^{-2} . Remind that the use of a Gaussian prior yields some bias in the posterior distribution measured by the value

$$b_G = \|DD_G^{-2}G^2\boldsymbol{\theta}^*\| \leq \|D_G^{-1}G^2\boldsymbol{\theta}^*\| \leq \|G\boldsymbol{\theta}^*\|. \quad (14.34)$$

If the point $\boldsymbol{\theta}_G^\dagger$ minimizes the quadratic expression $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + \|G\boldsymbol{\theta}\|^2$, that is,

$$D_G^2\boldsymbol{\theta}_G^\dagger = D^2\boldsymbol{\theta}^*,$$

then

$$\|D(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^\dagger)\| = \|DD_G^{-2}G^2\boldsymbol{\theta}^*\| = b_G.$$

Let now Q be a matrix which satisfies $D^2 \leq Q^2 \leq D_G^2$ and $p_{Q|G} = \text{tr}(QD_G^{-2}Q) < \infty$. We consider the centered and rescaled posterior $Q(\boldsymbol{\vartheta}_G - \check{\boldsymbol{\theta}}_G) | \mathbf{Y}$ with

$$\check{\boldsymbol{\theta}}_G = \boldsymbol{\theta}_G^\dagger + D_G^{-1}\nabla\zeta(\boldsymbol{\theta}^*) = \boldsymbol{\theta}_G^\dagger + D_G^{-1}\boldsymbol{\xi}_G. \quad (14.35)$$

The next result specifies the BvM statement for a Gaussian prior from Theorem 11.9.2 to the GLM case.

Theorem 14.5.2. *Let Q satisfy $D^2 \leq Q^2 \leq D_G^2$ and $p_{Q|G} = \text{tr}(QD_G^{-2}Q) < \infty$. Let also the random vector $\boldsymbol{\xi}_G$ fulfill $\|\boldsymbol{\xi}_G\| \leq z(B_G, \mathbf{x})$ on a random set $\Omega(\mathbf{x})$ with $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - e^{-\mathbf{x}}$. Suppose that it holds for a \mathbf{r}_0 fixed*

$$\left| \mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \frac{1}{2}\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \right| \leq \frac{1}{2}\mathbf{r}_0^2\delta(\mathbf{r}_0), \quad \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}_0.$$

If \mathbf{r}_0 is large enough to ensure with b_G from (14.34)

$$\{1 - \delta(\mathbf{r}_0)\}\mathbf{r}_0 \geq 2\left\{\sqrt{p_{Q|G} + \mathbf{x}} + 1 + z(B_G, \mathbf{x}) + b_G\right\},$$

then it holds on $\Omega(\mathbf{x})$ with $\check{\boldsymbol{\theta}}_G$ from (14.35) and $\Delta_{\circ}(\mathbf{x}) = \mathbf{r}_0^2\delta(\mathbf{r}_0)/2$ for any measurable set $\mathcal{A} \subset \mathbb{R}^p$

$$\mathbb{P}\left(Q(\boldsymbol{\vartheta}_G - \check{\boldsymbol{\theta}}_G) \in \mathcal{A} \mid \mathbf{Y}\right) \geq e^{4\Delta_{\circ}(\mathbf{x})} \mathbb{P}^{\circ}(QD_G^{-1}\boldsymbol{\gamma} \in \mathcal{A}) - e^{-\mathbf{x}} - \rho_Q(\mathbf{x}),$$

$$\mathbb{P}\left(Q(\boldsymbol{\vartheta}_G - \check{\boldsymbol{\theta}}_G) \in \mathcal{A} \mid \mathbf{Y}\right) \leq \frac{e^{4\Delta_{\circ}(\mathbf{x})}}{1 - e^{-\mathbf{x}}} \mathbb{P}^{\circ}(QD_G^{-1}\boldsymbol{\gamma} \in \mathcal{A}) + \rho_Q(\mathbf{x}),$$

where the quantity $\rho_Q(\mathbf{r}_0)$ fulfills under condition $\|\boldsymbol{\xi}_G\| \leq z(B_G, \mathbf{x})$

$$\rho_Q(\mathbf{r}_0) \leq e^{4\Delta_{\circ}(\mathbf{x})} \left\{ e^{-\frac{1}{2}(\mathbf{p}_{Q|G} + \mathbf{x})} + \frac{e^{-\mathbf{x}}}{1 - e^{-\mathbf{x}}} \right\}.$$

14.6 GLM with random design

This section discusses the situation when the design vectors Ψ_i are random. This study continues the one started for linear models in Chapter 2.

Let $\mathbf{Y} = (Y_1, \dots, Y_n)^{\top} \sim \mathbb{P}$ be a sample of independent r.v.'s. We consider the model

$$Y_i \mid \Psi_i \sim P_{\Psi_i^{\top} \boldsymbol{\theta}} \in (P_v)$$

in which the independent variables Ψ_i are random and (P_v) is a given exponential family with canonical parametrization. The (quasi) log-likelihood $L(\boldsymbol{\theta})$ can be represented in the same form as in the case of a deterministic design:

$$L(\boldsymbol{\theta}) = \sum_{i=1}^n \{Y_i \Psi_i^{\top} \boldsymbol{\theta} - g(\Psi_i^{\top} \boldsymbol{\theta})\} = S^{\top} \boldsymbol{\theta} - A(\boldsymbol{\theta})$$

with a random p -vector

$$S \stackrel{\text{def}}{=} \boldsymbol{\Psi} \mathbf{Y} = \sum_{i=1}^n Y_i \Psi_i \quad (14.36)$$

and a random function

$$A(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{i=1}^n g(\Psi_i^{\top} \boldsymbol{\theta}).$$

The MLE $\tilde{\boldsymbol{\theta}}$ for this GLM read exactly as in the case of a deterministic design:

$$\tilde{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}} \{S^{\top} \boldsymbol{\theta} - A(\boldsymbol{\theta})\}, \quad (14.37)$$

while the definition of the target $\boldsymbol{\theta}^*$ slightly changes:

$$\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E} L(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}} \{(I\mathbb{E} S)^{\top} \boldsymbol{\theta} - I\mathbb{E} A(\boldsymbol{\theta})\},$$

where

$$\mathbb{E}S = \mathbb{E}(\Psi Y) = \sum_{i=1}^n \mathbb{E}(Y_i \Psi_i).$$

The definition of $\boldsymbol{\theta}^*$ implies the identity $\nabla \mathbb{E}L(\boldsymbol{\theta}^*) = 0$ which yields

$$\mathbb{E}S = \nabla \mathbb{E}A(\boldsymbol{\theta}^*). \quad (14.38)$$

The stochastic component $\zeta(\boldsymbol{\theta})$ of $L(\boldsymbol{\theta})$ can be written as

$$\begin{aligned} \zeta(\boldsymbol{\theta}) &= L(\boldsymbol{\theta}) - \mathbb{E}L(\boldsymbol{\theta}) \\ &= (S - \mathbb{E}S)^\top \boldsymbol{\theta} - \{A(\boldsymbol{\theta}) - \mathbb{E}A(\boldsymbol{\theta})\} \\ &= \sum_{i=1}^n \{Y_i \Psi_i - \mathbb{E}(Y_i \Psi_i)\}^\top \boldsymbol{\theta} - \sum_{i=1}^n \{g(\Psi_i^\top \boldsymbol{\theta}) - \mathbb{E}g(\Psi_i^\top \boldsymbol{\theta})\}. \end{aligned}$$

This yields for the gradient $\nabla \zeta(\boldsymbol{\theta})$

$$\begin{aligned} \nabla \zeta(\boldsymbol{\theta}) &= S - \mathbb{E}S - \nabla A(\boldsymbol{\theta}) + \nabla \mathbb{E}A(\boldsymbol{\theta}) \\ &= \sum_{i=1}^n \{Y_i \Psi_i - \mathbb{E}(Y_i \Psi_i)\} - \sum_{i=1}^n [g'(\Psi_i^\top \boldsymbol{\theta}) \Psi_i - \mathbb{E}\{g'(\Psi_i^\top \boldsymbol{\theta}) \Psi_i\}]. \end{aligned} \quad (14.39)$$

In the contrary to the case of a deterministic design, the gradient $\nabla \zeta(\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$. This makes the situation more complicated to study.

The Fisher information matrix is defined by $D^2 = \mathbb{F}(\boldsymbol{\theta}^*)$, where

$$\mathbb{F}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E}L(\boldsymbol{\theta}) = \sum_{i=1}^n \mathbb{E}\{\Psi_i \Psi_i^\top g''(\Psi_i^\top \boldsymbol{\theta})\}.$$

Its empirical counterpart $\tilde{\mathbb{F}}(\boldsymbol{\theta})$ looks similarly to the case of a deterministic design

$$\tilde{\mathbb{F}}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2 L(\boldsymbol{\theta}) = \sum_{i=1}^n \Psi_i \Psi_i^\top g''(\Psi_i^\top \boldsymbol{\theta}). \quad (14.40)$$

As $g(\cdot)$ is convex, it holds $g''(u) \geq 0$ for any u and thus $\mathbb{F}(\boldsymbol{\theta}) \geq 0$ and $\tilde{\mathbb{F}}(\boldsymbol{\theta}) \geq 0$. Define also the standardized score

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} D^{-1} \nabla L(\boldsymbol{\theta}^*) = D^{-1} \{S - \nabla A(\boldsymbol{\theta}^*)\}. \quad (14.41)$$

In view of (14.38) and (14.39), it can be represented as

$$\boldsymbol{\xi} = D^{-1} \sum_{i=1}^n \{Y_i - g'(\Psi_i^\top \boldsymbol{\theta}^*)\} \Psi_i. \quad (14.42)$$

14.6.1 Local concentration of $\tilde{\boldsymbol{\theta}}$

Concavity of $L(\boldsymbol{\theta})$ and of its deterministic part $I\!EL(\boldsymbol{\theta})$ allow for a simple and straightforward proof of the result about concentration of the MLE $\tilde{\boldsymbol{\theta}}$. Recall the definition of the local vicinity $\Theta_0(\mathbf{r})$ of $\boldsymbol{\theta}^*$: for $D^2 = \mathbb{F}(\boldsymbol{\theta}^*)$

$$\Theta_0(\mathbf{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} : \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}\}.$$

Theorem 14.6.1. *If for some $\mathbf{r}_0 > 0$, $\widetilde{\mathbb{F}}(\boldsymbol{\theta})$ from (14.40) fulfill*

$$I\!P\left\{\sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)} \|D^{-1} \widetilde{\mathbb{F}}(\boldsymbol{\theta}) D^{-1} - I_p\|_{\text{op}} > \delta(\mathbf{r}_0)\right\} \leq C_{\mathbb{F}} e^{-x} \quad (14.43)$$

with $\delta(\mathbf{r}_0) < 1$, and if $\boldsymbol{\xi}$ from (14.41) follows for $x > 0$ the probability bound

$$I\!P\left(\|\boldsymbol{\xi}\| > \frac{1 - \delta(\mathbf{r}_0)}{2} \mathbf{r}_0\right) \leq C_{\boldsymbol{\xi}} e^{-x}, \quad (14.44)$$

then the solution $\tilde{\boldsymbol{\theta}}$ of (14.37) satisfies

$$I\!P(\tilde{\boldsymbol{\theta}} \notin \Theta_0(\mathbf{r}_0)) \leq (C_{\mathbb{F}} + C_{\boldsymbol{\xi}}) e^{-x}.$$

Proof. The function $L(\boldsymbol{\theta})$ is concave in $\boldsymbol{\theta}$ because

$$-\nabla^2 L(\boldsymbol{\theta}) = \widetilde{\mathbb{F}}(\boldsymbol{\theta}) \geq 0.$$

If $\tilde{\boldsymbol{\theta}} \notin \Theta_0(\mathbf{r}_0)$, denote by $\check{\boldsymbol{\theta}}$ the point at which the line connecting $\boldsymbol{\theta}^*$ and $\tilde{\boldsymbol{\theta}}$ crosses the boundary of $\Theta_0(\mathbf{r}_0)$. It is easy to see that

$$\check{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = \frac{\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|}{\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \frac{\mathbf{r}_0}{\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*).$$

Concavity of $L(\boldsymbol{\theta})$ implies for the point of maximum $\tilde{\boldsymbol{\theta}}$ that

$$L(\tilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) \geq L(\check{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*).$$

Therefore, it suffices to check that on a set $\Omega(x)$ of probability $1 - 2e^{-x}$, it holds for each $\boldsymbol{\theta}$ with $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathbf{r}_0$ that

$$L(\boldsymbol{\theta}^*) - L(\boldsymbol{\theta}) > 0.$$

Then the event $\tilde{\boldsymbol{\theta}} \notin \Theta_0(\mathbf{r}_0)$ is impossible on $\Omega(x)$. For any such $\boldsymbol{\theta}$, we apply the second order Taylor expansion of $L(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$. By definition of $\boldsymbol{\theta}^*$, it holds $\nabla I\!EL(\boldsymbol{\theta}^*) = 0$ and thus $\nabla L(\boldsymbol{\theta}^*) = \nabla \zeta(\boldsymbol{\theta}^*)$. The use of (14.43) yields now for $\boldsymbol{\xi} = D^{-1} \nabla \zeta(\boldsymbol{\theta}^*)$ and for $\boldsymbol{\theta}$ with $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathbf{r}_0$

$$\begin{aligned}
L(\boldsymbol{\theta}^*) - L(\boldsymbol{\theta}) &= (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla L(\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \tilde{\mathbb{F}}(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\
&\geq (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla \zeta(\boldsymbol{\theta}^*) + \frac{1 - \delta(\mathbf{r}_0)}{2} \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \\
&= \boldsymbol{\xi}^\top D(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \frac{1 - \delta(\mathbf{r}_0)}{2} \mathbf{r}_0^2 \\
&\geq -\|\boldsymbol{\xi}\| \mathbf{r}_0 + \frac{1 - \delta(\mathbf{r}_0)}{2} \mathbf{r}_0^2.
\end{aligned}$$

Here $\boldsymbol{\theta}^\circ$ is a point from $\Omega(\mathbf{x})$ on the interval connecting $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$. If $\|\boldsymbol{\xi}\| \leq \mathbf{r}_0 \{1 - \delta(\mathbf{r}_0)\}/2$, then this implies $L(\boldsymbol{\theta}^*) - L(\boldsymbol{\theta}) > 0$, and the result follows.

14.6.2 Fisher and Wilks expansions

As a corollary, we obtain Fisher and Wilks expansions for the quasi MLE $\tilde{\boldsymbol{\theta}}$ in a generalized linear model with a random design.

Theorem 14.6.2. *Suppose the conditions of Theorem 14.6.1 for some \mathbf{r}_0 . Then it holds on a set $\Omega(\mathbf{x})$ with $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - (\mathbf{C}_{\mathbb{F}} + \mathbf{C}_{\boldsymbol{\xi}}) e^{-\mathbf{x}}$*

$$\begin{aligned}
\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi}\| &\leq \mathbf{r}_0 \delta(\mathbf{r}_0), \\
|2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\boldsymbol{\xi}\|^2| &\leq 2\mathbf{r}_0^2 \delta(\mathbf{r}_0) + \mathbf{r}_0^2 \delta^2(\mathbf{r}_0), \\
\left| \sqrt{2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)} - \|\boldsymbol{\xi}\| \right| &\leq 3\mathbf{r}_0 \delta(\mathbf{r}_0).
\end{aligned}$$

Proof. Denote

$$\mathbb{L}(\boldsymbol{\theta}) = \boldsymbol{\xi}^\top D(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \frac{1}{2} \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2$$

with $\boldsymbol{\xi}$ from (14.41). Similarly to the proof of Theorem 14.6.1, it suffices to bound the difference $D^{-1}\{\nabla L(\boldsymbol{\theta}) - \nabla \mathbb{L}(\boldsymbol{\theta})\}$ on a local set $\Theta_0(\mathbf{r}_0)$. The definitions of $L(\boldsymbol{\theta})$ and $\boldsymbol{\xi}$ yield

$$\begin{aligned}
\nabla L(\boldsymbol{\theta}) - \nabla \mathbb{L}(\boldsymbol{\theta}) &= S - \nabla A(\boldsymbol{\theta}) - \boldsymbol{\xi} + D^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\
&= \nabla A(\boldsymbol{\theta}^*) - \nabla A(\boldsymbol{\theta}) + D^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*),
\end{aligned}$$

The second order Taylor expansion of $A(\boldsymbol{\theta})$ implies in view of $\nabla^2 A(\boldsymbol{\theta}) = \tilde{\mathbb{F}}(\boldsymbol{\theta})$ for any vector $\mathbf{u} \in \mathbb{R}^p$

$$\mathbf{u}^\top \{\nabla L(\boldsymbol{\theta}) - \nabla \mathbb{L}(\boldsymbol{\theta})\} = \mathbf{u}^\top \{D^2 - \tilde{\mathbb{F}}(\boldsymbol{\theta}^\circ)\} (\boldsymbol{\theta} - \boldsymbol{\theta}^*),$$

where $\boldsymbol{\theta}^\circ$ is a point on the interval connecting $\boldsymbol{\theta}^*$ and $\tilde{\boldsymbol{\theta}}$. The use of $\mathbf{u} = D^{-1}\boldsymbol{\gamma}$ for a unit vector $\boldsymbol{\gamma} \in \mathbb{R}^p$ together with condition (14.43) leads to the bound

$$\left| \boldsymbol{\gamma}^\top D^{-1} \{ \nabla L(\boldsymbol{\theta}) - \nabla \mathbb{L}(\boldsymbol{\theta}) \} \right| \leq \| \boldsymbol{\gamma} \| \| D^{-1} \widetilde{\mathbb{F}}(\boldsymbol{\theta}^*) D^{-1} - I_p \|_{\text{op}} \| D(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \| \leq \delta(\mathbf{r}_0) \mathbf{r}_0.$$

This is the desirable bound, and the results of the theorem follow from Theorem 11.7.2.

14.6.3 Sufficient conditions for the case of random design

Below in this section \mathbb{Q} denotes the design measure, that is, the joint distribution of the design vectors Ψ_i . Similarly, $\mathbb{E}_{\mathbb{Q}}$ stands for the expectation w.r.t. \mathbb{Q} . Define

$$\mathsf{H}^2 \stackrel{\text{def}}{=} \mathbb{E}_{\mathbb{Q}} \left\{ \frac{1}{n} \sum_{i=1}^n g''(\Psi_i^\top \boldsymbol{\theta}^*) \Psi_i \Psi_i^\top \right\},$$

so that $D^2 = n\mathsf{H}^2$. Also write $d_i = \sqrt{g''(\Psi_i^\top \boldsymbol{\theta}^*)}$. Note that these quantities are design dependent and therefore, random. Our conditions below involve these and some other random quantities. We assume that there exists a random set $\Omega_{\Psi}(\mathbf{x})$ with $\mathbb{Q}(\Omega_{\Psi}(\mathbf{x})) \geq 1 - e^{-x}$, and all the conditions are fulfilled when restricting on this set. This allows excluding the situations with an exotic design configuration.

- **Design regularity** is measured by values a_{Ψ} and δ_{Ψ} such that it holds on $\Omega_{\Psi}(\mathbf{x})$

$$\sup_{\mathbf{u} \in \mathcal{S}_p} \frac{1}{n} \sum_{i=1}^n |d_i \mathbf{u}^\top \mathsf{H}^{-1} \Psi_i|^3 \leq a_{\Psi}, \quad (14.45)$$

$$\sup_{\mathbf{u} \in \mathcal{S}_p} \max_{i=1,\dots,n} |d_i \mathbf{u}^\top \mathsf{H}^{-1} \Psi_i| \leq \delta_{\Psi} \sqrt{n}.$$

- **Smoothness of the link function** $g(\cdot)$ can be measured by its third derivative. For each \mathbf{r} , there is a constant $a_g(\mathbf{r})$ such that it holds on $\Omega_{\Psi}(\mathbf{x})$

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r})} \max_{i=1,\dots,n} \frac{|g'''(\Psi_i^\top \boldsymbol{\theta})|}{|g''(\Psi_i^\top \boldsymbol{\theta}^*)|^{3/2}} \leq a_g(\mathbf{r}). \quad (14.46)$$

- **Exponential moments of the errors** $\varepsilon_i = Y_i - \mathbb{E}(Y_i | \Psi_i)$. Suppose that for some values $s_i = s_i(\Psi_i)$ and fixed constants $c_0, \lambda_0 > 0$ it holds on $\Omega_{\Psi}(\mathbf{x})$

$$\mathbb{E} \left\{ \exp(\lambda_0 s_i^{-1} \varepsilon_i) | \Psi_i \right\} \leq c_0, \quad i = 1, \dots, n. \quad (14.47)$$

This condition means that the errors ε_i have exponential moments given Ψ_i . In most of cases one can use $s_i^2 = \text{Var}(Y_i | \Psi_i)$. Condition (14.47) implies that there are another constants $g_1 \leq \lambda_0$ and ν_0 such that the following condition is fulfilled on $\Omega_{\Psi}(\mathbf{x})$:

$$\mathbb{E} \left\{ \exp(\lambda s_i^{-1} \varepsilon_i) | \Psi_i \right\} \leq \frac{1}{2} \nu_0^2 \lambda^2, \quad i = 1, \dots, n, \quad |\lambda| \leq g_1. \quad (14.48)$$

- **Noise homogeneity** is measured by the variability of the values \mathbf{s}_i . Namely, assume that for some fixed value $a_{\mathbf{s}}$, it holds on $\Omega_{\Psi}(\mathbf{x})$

$$\max_{i=1,\dots,n} \frac{\mathbf{s}_i}{d_i} \leq a_{\mathbf{s}}. \quad (14.49)$$

- **Identifiability** is measured by relationship between the matrices D^2 and V^2 , where the matrix V^2 is defined as

$$\begin{aligned} V^2 &\stackrel{\text{def}}{=} \mathbb{E}_{\mathbb{Q}} \tilde{V}^2, \\ \tilde{V}^2 &\stackrel{\text{def}}{=} \sum_{i=1}^n \mathbf{s}_i^2 \Psi_i \Psi_i^\top. \end{aligned} \quad (14.50)$$

If the observations Y_i follow the GLM assumption P_{v_i} for $v_i = \Psi_i^\top \boldsymbol{\theta}^*$, that is, the model is correctly specified, then with $\mathbf{s}_i^2 = \text{Var}(Y_i | \Psi_i) = g''(v_i)$, the matrices V^2 and D^2 coincide. In the general case under a possible model misspecification, the matrices V^2 and D^2 may be different. In this case we need an identifiability condition: the matrix $B_{\varepsilon} \stackrel{\text{def}}{=} D^{-1} V^2 D^{-1}$ is bounded from above and from below:

$$\underline{\alpha}^2 D^2 \leq V^2 \leq \overline{\alpha}^2 D^2 \quad (14.51)$$

for some $\underline{\alpha} > 0$. This condition can be spelled out as

$$\underline{\alpha}^2 \mathbb{E} \left(\sum_{i=1}^n g''(\Psi_i^\top \boldsymbol{\theta}^*) \Psi_i \Psi_i^\top \right) \leq \mathbb{E} \left(\sum_{i=1}^n \mathbf{s}_i^2 \Psi_i \Psi_i^\top \right) \leq \overline{\alpha}^2 \mathbb{E} \left(\sum_{i=1}^n g''(\Psi_i^\top \boldsymbol{\theta}^*) \Psi_i \Psi_i^\top \right).$$

In the case of a random design, the modeling bias comes in play. This situation differs from the case of a deterministic design, where the bias $f(X_i) - g'(\Psi_i^\top \boldsymbol{\theta}^*)$ does not matter in evaluating the accuracy of quadratic approximation of the log-likelihood.

- **Modeling bias** in the linear parametric assumption $f_i = \mathbb{E}(Y_i | \Psi_i) = g'(\Psi_i^\top \boldsymbol{\theta}^*)$ can be expressed via the vector $\mathbf{b} = (b_i)$ with

$$b_i = b_i(\Psi_i) \stackrel{\text{def}}{=} f_i - g'(\Psi_i^\top \boldsymbol{\theta}^*) = \mathbb{E}(Y_i | \Psi_i) - g'(\Psi_i^\top \boldsymbol{\theta}^*), \quad i = 1, \dots, n.$$

More precisely, we use the matrix

$$B_b \stackrel{\text{def}}{=} D^{-1} \left(\sum_{i=1}^n \mathbb{E}(b_i^2 \Psi_i \Psi_i^\top) \right) D^{-1}$$

to measure the impact of the modeling bias. It is clearly small if the values $b_i^2 / g''(\Psi_i^\top \boldsymbol{\theta}^*)$ are uniformly small.

The first condition on the third directional moments of the design in (14.45) is not restrictive and “dimension free”. Indeed, the following result holds.

Lemma 14.6.1. Let ξ_1, \dots, ξ_n be independent zero mean vectors in \mathbb{R}^p such that for some fixed $\lambda_0 > 0$ and C_0 and any unit vector \mathbf{u}

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} |\mathbf{u}^\top \xi_i|^2 \leq 1,$$

$$\max_{i=1, \dots, n} \mathbb{E} \exp(\lambda_0 \mathbf{u}^\top \xi_i) \leq C_0.$$

Then

$$\mathbb{P} \left(\sup_{\mathbf{u} \in \mathcal{S}_p} \frac{1}{n} \sum_{i=1}^n |\mathbf{u}^\top \xi_i|^3 > (a_1 + \sqrt{a_2 x})^{3/2} \right) \leq C_1 e^{-x}.$$

To be done: please check, apply e.g. Bousquet's inequality

Our first result shows that the empirical matrix $\tilde{\mathbb{F}}(\boldsymbol{\theta}^*) = \sum_{i=1}^n g''(\Psi_i^\top \boldsymbol{\theta}^*) \Psi_i \Psi_i^\top$ is close to their population counterparts $\mathbb{F}(\boldsymbol{\theta}^*)$ under conditions made above; cf. Corollary 2.2.1. The same applies to the random matrix $\tilde{V}^2 = \sum_{i=1}^n \mathbf{s}_i^2 \Psi_i \Psi_i^\top$.

Lemma 14.6.2. Let $x > 0$ be fixed. For the random matrices $\tilde{\mathbb{F}}(\boldsymbol{\theta}^*)$ and \tilde{V}^2 and their expectation $D^2 = \mathbb{E}\{\tilde{\mathbb{F}}(\boldsymbol{\theta}^*)\}$, $V^2 = \mathbb{E}(\tilde{V}^2)$, it holds on a random set $\Omega_\Psi(x)$ with $\mathbb{Q}(\Omega_\Psi(x)) \geq 1 - 2e^{-x}$

$$\|D^{-1} \tilde{\mathbb{F}}(\boldsymbol{\theta}^*) D^{-1} - I_p\|_{\text{op}} \leq \delta_{\mathbb{F}}(x),$$

$$\|V^{-1} \tilde{V}^2 V^{-1} - I_p\|_{\text{op}} \leq \delta_V(x).$$

Proof. **To be done:** show with $\delta_{\mathbb{F}} \asymp \delta_\Psi$ and with $\delta_V \asymp \delta_\Psi$ using the matrix Bernstein inequality

Theorem 14.6.3. Suppose (14.48), (14.49), (14.46), and (14.51). For $z(B, x)$ from (B.21), fix

$$\mathbf{r}_0 = 4\nu_0 z(B_\epsilon^*, x) + 4z(B_b, x), \quad (14.52)$$

with $B_\epsilon^* = (1 + \delta_V) B_\epsilon$, δ_V , δ_b from Lemma 14.6.2. Suppose also that

$$a_\Psi a_g(\mathbf{r}_0) \mathbf{r}_0 / \sqrt{n} + \delta_{\mathbb{F}} < 1/2.$$

Then the conditions of Theorem 14.6.1 are fulfilled with $\delta(\mathbf{r}_0) \leq a_\Psi a_g(\mathbf{r}_0) \mathbf{r}_0 / \sqrt{n} + \delta_{\mathbb{F}}$ and the results of this theorem continue to apply.

Proof. Let \mathbf{r}_0 be fixed by (14.52). We split the proof into few steps.

Lemma 14.6.3. *It holds on $\Omega_{\Psi}(\mathbf{x})$*

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)} \|D^{-1} \{\tilde{\mathbb{F}}(\boldsymbol{\theta}) - \tilde{\mathbb{F}}(\boldsymbol{\theta}^*)\} D^{-1} - I_p\|_{\text{op}} \leq \delta_0,$$

$$\delta_0 \stackrel{\text{def}}{=} a_{\Psi} a_g(\mathbf{r}_0) \mathbf{r}_0 / \sqrt{n}.$$

Proof. After restricting to $\Omega_{\Psi}(\mathbf{x})$ the proof follows the line of the similar result for a deterministic design; cf. Lemma 14.2.2.

Lemma 14.6.3 and 14.6.2 imply on $\Omega_{\mathbb{F}}(\mathbf{x}) \cap \Omega_{\Psi}(\mathbf{x})$

$$\begin{aligned} & \|D^{-1} \tilde{\mathbb{F}}(\boldsymbol{\theta}) D^{-1} - I_p\|_{\text{op}} \\ & \leq \|D^{-1} \{\tilde{\mathbb{F}}(\boldsymbol{\theta}) - \tilde{\mathbb{F}}(\boldsymbol{\theta}^*)\} D^{-1} - I_p\|_{\text{op}} + \|D^{-1} \tilde{\mathbb{F}}(\boldsymbol{\theta}^*) D^{-1} - I_p\|_{\text{op}} \leq \delta_0 + \delta_{\mathbb{F}}, \end{aligned}$$

and condition (14.43) follows in an obvious way with $\delta(\mathbf{r}_0) = \delta_0 + \delta_{\mathbb{F}}$.

Now we show that the presented conditions are sufficient for checking the bound (14.44) on the norm of the random vector ξ from (14.42). Define $f_i = \mathbb{E}(Y_i | \Psi_i)$ and $\varepsilon_i = Y_i - f_i$. Then ξ can be represented as

$$\xi = D^{-1} \sum_{i=1}^n \varepsilon_i \Psi_i + D^{-1} \sum_{i=1}^n \{f_i - g'(\Psi_i^\top \boldsymbol{\theta}^*)\} \Psi_i = \xi_{\varepsilon} + \xi_b,$$

where the first term ξ_{ε} is mainly responsible for the individual errors ε_i :

$$\xi_{\varepsilon} \stackrel{\text{def}}{=} D^{-1} \Psi \varepsilon = D^{-1} \sum_{i=1}^n \varepsilon_i \Psi_i.$$

The second term ξ_b describes the variability induced by the random design:

$$\xi_b \stackrel{\text{def}}{=} D^{-1} \sum_{i=1}^n \{f_i - g'(\Psi_i^\top \boldsymbol{\theta}^*)\} \Psi_i = D^{-1} \sum_{i=1}^n \{b_i \Psi_i - \mathbb{E}(b_i \Psi_i)\}.$$

Here we have used that $\mathbb{E}(\Psi b) = \mathbb{E}\{S - \nabla A(\boldsymbol{\theta}^*)\} = 0$. Now the bound for $\|\xi_{\varepsilon}\|$ and $\|\xi_b\|$ can be obtained similarly to the linear case of Section 2.5 under the conditions made.

Lemma 14.6.4. *Let the errors $\varepsilon_i = Y_i - \mathbb{E}(Y_i | \Psi_i)$ be independent and follow (14.48). Then on $\Omega_{\Psi}(\mathbf{x})$*

$$\log \mathbb{E} \left\{ \exp(\mathbf{u}^\top \tilde{V}^{-1} \Psi \varepsilon) \mid \Psi \right\} \leq \frac{\nu_0^2}{2} \|\mathbf{u}\|^2, \quad \|\mathbf{u}\| \leq g \stackrel{\text{def}}{=} \frac{g_1}{\delta_{\Psi} a_s}, \quad (14.53)$$

where \tilde{V}^2 is from (14.50) and a_s from (14.49).

Proof. The formula (14.36) and independence of the Ψ_i 's and of each ε_i given Ψ_i imply for any vector $\mathbf{u} \in I\!\!R^p$ with $\|\mathbf{u}\| \leq g$

$$\log I\!\!E \left\{ \exp(\mathbf{u}^\top \tilde{V}^{-1} \Psi \varepsilon) \mid \Psi \right\} = \sum_{i=1}^n \log I\!\!E \left\{ \exp(\lambda_i \mathbf{s}_i^{-1} \varepsilon_i) \mid \Psi_i \right\},$$

where the definitions (14.49) and (14.50) imply for $\lambda_i = \mathbf{u}^\top \tilde{V}^{-1} \Psi_i \mathbf{s}_i$

$$|\lambda_i| = |\mathbf{u}^\top \tilde{V}^{-1} \Psi_i| \mathbf{s}_i \leq g \|\tilde{V}^{-1} \Psi_i\| \mathbf{s}_i \leq g_1.$$

Therefore, by (14.48) and (14.49)

$$\begin{aligned} \log I\!\!E \left\{ \exp(\mathbf{u}^\top \tilde{V}^{-1} \Psi \varepsilon) \mid \Psi \right\} &\leq \frac{\nu_0^2}{2} \sum_{i=1}^n \lambda_i^2 \\ &= \frac{\nu_0^2}{2} \sum_{i=1}^n \mathbf{u}^\top \tilde{V}^{-1} (\Psi_i \Psi_i^\top \mathbf{s}_i^2) \tilde{V}^{-1} \mathbf{u} = \frac{\nu_0^2}{2} \|\mathbf{u}\|^2, \end{aligned}$$

and (14.53) follows.

This bound allows to apply Theorem B.2.2 from Section B.2 conditionally on Ψ . Namely, we use that

$$\xi_\varepsilon = D^{-1} \Psi \varepsilon = D^{-1} \tilde{V} (\tilde{V}^{-1} \Psi \varepsilon). \quad (14.54)$$

Then by Lemma 14.6.2 on the random set $\Omega_\Psi(\mathbf{x})$

$$D^{-1} \tilde{V}^2 D^{-1} \leq (1 + \delta_V) D^{-1} V^2 D^{-1} = (1 + \delta_V) B_\varepsilon = B_\varepsilon^*$$

and by Theorem B.2.2, the norm of ξ_ε from (14.54) is bounded by $\nu_0 z(B_\varepsilon^*, \mathbf{x})$ with a high probability, that is,

$$I\!\!P \left(\|D^{-1} \Psi \varepsilon\| \geq \nu_0 z(B_\varepsilon^*, \mathbf{x}) \mid \Psi \right) \leq 2e^{-x}.$$

Here we assume that g is sufficiently large and ignore the second term in the exponential bound of Theorem B.2.1. Combination with Lemma 14.6.2 yields a similar bound on the random set $\Omega_\Psi(\mathbf{x})$:

$$I\!\!P \left(\|\xi_\varepsilon\| \geq \nu_0 z(B_\varepsilon^*, \mathbf{x}) \mid \Psi \right) \leq 2e^{-x}.$$

This results in an unconditional bound

$$I\!\!P \left(\|\xi_\varepsilon\| \geq \nu_0 z(B_\varepsilon^*, \mathbf{x}) \right) \leq 3e^{-x}$$

The norm of $\xi_b = D^{-1} \Psi b$ can be bounded again by Theorem B.2.2: with

$$B_b \stackrel{\text{def}}{=} \text{Var}(\boldsymbol{\xi}_b) = D^{-1} \left\{ \sum_{i=1}^n I\!\!E(b_i^2 \Psi_i \Psi_i^\top) \right\} D^{-1},$$

it holds on $\Omega_\Psi(\mathbf{x})$

$$I\!\!P\left(\|\boldsymbol{\xi}_b\| \geq \nu_0 z(B_b, \mathbf{x})\right) \leq 2e^{-x}.$$

Putting all together yields

$$\|\boldsymbol{\xi}\| \leq \nu_0 z(B_\epsilon^*, \mathbf{x}) + \nu_0 z(B_b, \mathbf{x})$$

on a random set $\Omega(\mathbf{x})$ with $I\!\!P(\Omega(\mathbf{x})) \geq 1 - 5e^{-x}$. This yields the bound (14.44) under the condition

$$\frac{1 - \delta(r_0)}{2} r_0 \geq \nu_0 z(B_\epsilon^*, \mathbf{x}) + \nu_0 z(B_b, \mathbf{x})$$

which is obviously fulfilled for our choice of $r_0 = 4\nu_0 z(B_\epsilon^*, \mathbf{x}) + 4\nu_0 z(B_b, \mathbf{x})$ in view of $\delta(r_0) < 1/2$. This will also provide (14.44). All the conditions of Theorem 14.6.1 have been checked.

14.6.4 Nonparametric BvM for a Gaussian prior

Estimation of a log-density

This chapter illustrates the applicability of the obtained results to the problem of density estimation. More precisely, we consider an expansion of a log-density function with some functional basis and identify this function with the related set of coefficients. It appears that the problem of log-density estimation is very close to the generalized linear modeling. One explanation of this fact is that the density model is almost equivalent to Poisson regression which is again a GLM case.

15.1 Log-density estimation. Conditions

Suppose we are given a random sample X_1, \dots, X_n in \mathbb{R}^d . The i.i.d. model assumption means that all these random variables are independent identically distributed from some measure P with a density $f(x)$ with respect to a σ -finite measure μ_0 in \mathbb{R}^d . This density function is the target of estimation. By definition, the function f is non-negative, measurable, and integrates to one:

$$\int f(x) \mu_0(dx) = 1.$$

Here and below, the integral \int without limits means the integral over the whole space \mathbb{R}^d . If $f(\cdot)$ has a smaller support \mathcal{X} , one can restrict integration to this set. To recover f from observed data X_1, \dots, X_n , this function is usually assumed to possess some smoothness properties.

Below we parametrize the density function by a linear decomposition of the log-density function. Let $\psi_1(x), \psi_2(x), \dots$ be a collection of functions in \mathbb{R}^d satisfying

$$\int \exp\{|\psi_j(x)|\} f(x) \mu_0(dx) < \infty. \quad (15.1)$$

Denote by Θ_1 the subset in \mathbb{R}^∞ of all $\boldsymbol{\theta}$ satisfying

$$\int \exp\left\{\sum_{j=1}^{\infty} \theta_j \psi_j(x)\right\} \mu_0(dx) < \infty. \quad (15.2)$$

The Hölder inequality and condition (15.1) imply that Θ_1 is a convex set in \mathbb{R}^∞ containing the simplex set of all $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 \leq 1$. Note that orthogonality of the ψ_j 's is not assumed. Below we assume that $\boldsymbol{\theta}$ belongs to a convex subset Θ of Θ_1 .

For each $\boldsymbol{\theta} \in \Theta_1$, define a log-density function $\ell(x, \boldsymbol{\theta})$ using the expansion of the function $\log f(x)$ with respect to the $\{\psi_j(\cdot)\}$ -basis:

$$\log f(x) = \ell(x, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} \theta_j \psi_j(x) - g(\boldsymbol{\theta}), \quad (15.3)$$

where $g(\boldsymbol{\theta})$ is a constant given by

$$\int \exp \left\{ \sum_{j=1}^{\infty} \theta_j \psi_j(x) \right\} \mu_0(dx) = e^{g(\boldsymbol{\theta})}. \quad (15.4)$$

A nice feature of such representation is that the function $\log f(x)$ does not need to be non-negative, this simplifies the study of the linear decomposition (15.3). However, the main benefit of using the log-density is that the stochastic part of the corresponding log-likelihood expression has a linear structure w.r.t. the parameter $\boldsymbol{\theta}$. This enables us to apply the well developed theory similarly to the GLM case.

The true density $f(x)$ w.r.t. μ_0 is supposed to be in the form $f(x) = \exp\{\ell(x, \boldsymbol{\theta}^*)\} = \exp\{\boldsymbol{\theta}^{*\top} \Psi(x) - g(\boldsymbol{\theta}^*)\}$ for some parameter vector $\boldsymbol{\theta}^* \in \Theta_1$. The decomposition (15.3) and (15.4) can now be rewritten as

$$\begin{aligned} \int \exp \left\{ \sum_{j=1}^{\infty} (\theta_j - \theta_j^*) \psi_j(x) \right\} f(x) \mu_0(dx) &= \int \exp\{(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \otimes \Psi(x)\} f(x) \mu_0(dx) \\ &= \exp\{g(\boldsymbol{\theta}) - g(\boldsymbol{\theta}^*)\}. \end{aligned} \quad (15.5)$$

The i.i.d. assumption yields the log-likelihood $L(\boldsymbol{\theta})$ in the form

$$L(\boldsymbol{\theta}) = \log \left\{ \prod_{i=1}^n f(X_i) \right\} = \sum_{i=1}^n \ell(X_i, \boldsymbol{\theta}) = \sum_{i=1}^n \Psi_i^\top \boldsymbol{\theta} - n g(\boldsymbol{\theta}),$$

where $\Psi_i = \Psi(X_i)$ is a vector in \mathbb{R}^∞ with entries $\psi_j(X_i)$ for $j = 1, 2, \dots$. This and (15.4) imply

$$\begin{aligned} -\nabla^2 L(\boldsymbol{\theta}) &= n \nabla^2 g(\boldsymbol{\theta}) = n \nabla^2 \log \int \exp\{\boldsymbol{\theta} \otimes \Psi(x)\} \mu_0(dx) \\ &= n E_{\boldsymbol{\theta}} \left[\{\Psi(X_1) - E_{\boldsymbol{\theta}} \Psi(X_1)\} \otimes \{\Psi(X_1) - E_{\boldsymbol{\theta}} \Psi(X_1)\}^\top \right]; \end{aligned}$$

see Lemma 15.1.1 below. This yields that the matrix (operator) $-\nabla^2 L(\boldsymbol{\theta})$ is positive semidefinite, that is, $L(\boldsymbol{\theta})$ is a concave function as in the GLM case.

Now we study the properties of the MLE $\tilde{\boldsymbol{\theta}}$ maximizing $L(\boldsymbol{\theta})$. To be as concise as possible, we assume that the model is correctly specified, that is, under the true measure $\mathbb{I}\mathcal{P}$, the variables X_i are i.i.d. The results easily extend to the case of independent but not identically distributed observations.

The stochastic part $\zeta(\boldsymbol{\theta})$ of $L(\boldsymbol{\theta})$ reads exactly as in the GLM case:

$$\zeta(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - \mathbb{E}L(\boldsymbol{\theta}) = (S - \mathbb{E}S)^\top \boldsymbol{\theta}, \quad (15.6)$$

where S is the sum of random vectors $\Psi_i = \Psi(X_i) \in \mathbb{R}^\infty$:

$$S \stackrel{\text{def}}{=} \sum_{i=1}^n \Psi_i = \sum_{i=1}^n \Psi(X_i).$$

The expected log-likelihood is of the form

$$\mathbb{E}L(\boldsymbol{\theta}) = \mathbb{E}(S^\top \boldsymbol{\theta}) - n g(\boldsymbol{\theta}) = n\{\boldsymbol{\theta}^\top \bar{\Psi} - g(\boldsymbol{\theta})\}$$

with

$$\bar{\Psi} \stackrel{\text{def}}{=} \mathbb{E}\Psi(X_1).$$

Also

$$V^2 \stackrel{\text{def}}{=} \text{Var}(S) = n\{\mathbb{E}[\Psi(X_1) \otimes \Psi(X_1)] - \bar{\Psi} \otimes \bar{\Psi}\} = n\mathsf{H}^2$$

for

$$\mathsf{H}^2 \stackrel{\text{def}}{=} \text{Var}(\Psi_1) = \mathbb{E}[\Psi(X_1) \otimes \Psi(X_1)] - \bar{\Psi} \otimes \bar{\Psi}.$$

The negative Hessian $-\nabla^2 \mathbb{E}L(\boldsymbol{\theta}^*)$ of the expected log-likelihood at the true point $\boldsymbol{\theta}^*$ coincides with the n -times Hessian of $g(\boldsymbol{\theta}^*)$:

$$D^2 = -\nabla^2 \mathbb{E}L(\boldsymbol{\theta}^*) = n\nabla^2 g(\boldsymbol{\theta}^*),$$

which is a self-adjoint operator in \mathbb{R}^∞ . As the model is correctly specified, the same holds for the covariance structure of the score vector $\nabla\zeta(\boldsymbol{\theta}) = S - \mathbb{E}S$ under the true measure $\mathbb{I}\mathcal{P} = \mathbb{I}\mathcal{P}_{\boldsymbol{\theta}^*}$:

$$V^2 = D^2 = n\nabla^2 g(\boldsymbol{\theta}^*);$$

see Lemma 15.1.1 below for a direct proof of this identity.

We study the properties of the log-likelihood $L(\boldsymbol{\theta})$ under the regularity condition on the score vector χ_1 for one observation X_1 given by

$$\chi_1 \stackrel{\text{def}}{=} H^{-1}\{\Psi(X_1) - \bar{\Psi}\}.$$

In the case of a Gaussian density, the variable χ_1 is standard normal and hence, for any vector α and any unit vector u

$$\begin{aligned} I\!E \exp\left(\alpha^\top \chi_1 - \frac{\|\alpha\|^2}{2}\right) &\equiv 1, \\ I\!E \left[u^\top (\chi_1 - \alpha) \exp\left(\alpha^\top \chi_1 - \frac{\|\alpha\|^2}{2}\right) \right] &\equiv 0, \\ I\!E \left[|u^\top (\chi_1 - \alpha)|^2 \exp\left(\alpha^\top \chi_1 - \frac{\|\alpha\|^2}{2}\right) \right] &\equiv 1. \end{aligned}$$

If χ_1 has exponential moment, a similar behavior can be expected for small vectors α .

(χ_1) *The vector $\chi_1 = H^{-1}\{\Psi(X_1) - \bar{\Psi}\}$ satisfies*

$$\begin{aligned} \left| \log I\!E \exp\left(\alpha^\top \chi_1 - \frac{\|\alpha\|^2}{2}\right) \right| &\leq \rho_0(\alpha), \\ \sup_{\|u\| \leq 1} \left| I\!E \left\{ u^\top (\chi_1 - \alpha) \exp\left(\alpha^\top \chi_1 - \frac{\|\alpha\|^2}{2}\right) \right\} \right| &\leq \rho_1(\alpha), \quad (15.7) \\ \sup_{\|u\| \leq 1} \left| \log I\!E \left\{ |u^\top (\chi_1 - \alpha)|^2 \exp\left(\alpha^\top \chi_1 - \frac{\|\alpha\|^2}{2}\right) \right\} \right| &\leq \rho_2(\alpha), \end{aligned}$$

and for some $g_1 > 0$ and all α with $\|\alpha\| \leq g_1$, it holds

$$e^{\rho_0(\alpha)} \rho_1(\alpha) \leq \frac{1}{2}.$$

Remark 15.1.1. If $I\!E e^{\lambda \|\chi_1\|} < \infty$ for some positive λ , then the moment generating function function $f(\alpha) \stackrel{\text{def}}{=} I\!E e^{\alpha^\top \chi_1}$ is analytic in a vicinity of zero. Obviously $f(0) = 1$, $\nabla f(0) = I\!E \chi_1 = 0$, $\nabla^2 f(0) = \text{Var}(\chi_1)$. This yields that

$$\rho_0(\alpha) \leq w_0 \|\alpha\|^3, \quad \rho_1(\alpha) \leq w_1 \|\alpha\|^2, \quad \rho_2(\alpha) \leq w_2 \|\alpha\|,$$

where the constants w_0 , w_1 , and w_2 depend on the third moments of χ_1 . Therefore, condition **(χ_1)** is not restrictive, it mainly requires bounded exponential moments of χ_1 , which is anyway assumed to validate our log-density model; cf. (15.2).

The next lemma shows that **(χ_1)** implies **(\mathcal{L}_0)**.

Lemma 15.1.1. *Suppose that condition **(χ_1)** is fulfilled. Then*

$$\nabla g(\theta^*) = \bar{\Psi}, \quad \nabla^2 g(\theta^*) = \text{Var}\{\Psi(X_1)\} = H^2.$$

Furthermore, for each $\theta \in \Theta$ and $\alpha = H(\theta - \theta^*)$

$$\left| g(\boldsymbol{\theta}) - g(\boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \bar{\Psi} - \frac{1}{2} \|\mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \right| \leq \rho_0(\boldsymbol{\alpha}); \quad (15.8)$$

$$\|\mathsf{H}^{-1}\{\nabla g(\boldsymbol{\theta}) - \bar{\Psi}\} - \mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq e^{\rho_0(\boldsymbol{\alpha})} \rho_1(\boldsymbol{\alpha}). \quad (15.9)$$

Finally,

$$\|\mathsf{H}^{-1}\nabla^2 g(\boldsymbol{\theta})\mathsf{H}^{-1} - I\|_{\text{op}} \leq \left[e^{\rho_0(\boldsymbol{\alpha}) + \rho_2(\boldsymbol{\alpha})} - 1 + e^{2\rho_0(\boldsymbol{\alpha})} \rho_1^2(\boldsymbol{\alpha}) \right]. \quad (15.10)$$

Proof. By definition $g(\boldsymbol{\theta}) = \log \int e^{\boldsymbol{\theta}^\top \Psi(x)} \mu_0(dx)$, and

$$\begin{aligned} \nabla g(\boldsymbol{\theta}) &= \frac{\int \Psi(x) e^{\boldsymbol{\theta}^\top \Psi(x)} \mu_0(dx)}{\int e^{\boldsymbol{\theta}^\top \Psi(x)} \mu_0(dx)}, \\ \nabla^2 g(\boldsymbol{\theta}) &= \frac{\int \Psi(x) \otimes \Psi(x) e^{\boldsymbol{\theta}^\top \Psi(x)} \mu_0(dx)}{\int e^{\boldsymbol{\theta}^\top \Psi(x)} \mu_0(dx)} - \nabla g(\boldsymbol{\theta}) \otimes \nabla g(\boldsymbol{\theta}) \\ &= \frac{\int \{\Psi(x) - \nabla g(\boldsymbol{\theta})\} \otimes \{\Psi(x) - \nabla g(\boldsymbol{\theta})\} e^{\boldsymbol{\theta}^\top \Psi(x)} \mu_0(dx)}{\int e^{\boldsymbol{\theta}^\top \Psi(x)} \mu_0(dx)}. \end{aligned} \quad (15.11)$$

As $f(x) = \exp\{\Psi(x)^\top \boldsymbol{\theta}^* - g(\boldsymbol{\theta}^*)\}$, it holds for $\boldsymbol{\theta} = \boldsymbol{\theta}^*$

$$\nabla g(\boldsymbol{\theta}^*) = \frac{\int \Psi(x) e^{\boldsymbol{\theta}^{*\top} \Psi(x)} \mu_0(dx)}{\int e^{\boldsymbol{\theta}^{*\top} \Psi(x)} \mu_0(dx)} = \int \Psi(x) f(x) \mu_0(dx) = \bar{\Psi},$$

and similarly

$$\nabla^2 g(\boldsymbol{\theta}^*) = \frac{\int \Psi(x) \otimes \Psi(x) e^{\boldsymbol{\theta}^{*\top} \Psi(x)} \mu_0(dx)}{\int e^{\boldsymbol{\theta}^{*\top} \Psi(x)} \mu_0(dx)} - \bar{\Psi} \otimes \bar{\Psi} = \text{Var } \Psi(X_1) = \mathsf{H}^2.$$

Fix $\boldsymbol{\theta} \in \Theta_1$ and define $\boldsymbol{\alpha} = \mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$. Then with $\chi_1(x) = \mathsf{H}^{-1}\{\Psi(x) - \bar{\Psi}\}$, it holds by (15.5)

$$\begin{aligned} g(\boldsymbol{\theta}) - g(\boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \bar{\Psi} - \frac{1}{2} \|\mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \\ = \log \int \exp\left\{(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top [\Psi(x) - \bar{\Psi}] - \frac{1}{2} \|\mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2\right\} f(x) \mu_0(dx) \\ = \log \int \exp\left\{\boldsymbol{\alpha}^\top \chi_1(x) - \frac{\|\boldsymbol{\alpha}\|^2}{2}\right\} f(x) \mu_0(dx) \end{aligned}$$

yielding by (15.7) from (χ_1)

$$\begin{aligned} \left| g(\boldsymbol{\theta}) - g(\boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \bar{\Psi} - \frac{1}{2} \|\mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \right| \\ = \left| g(\boldsymbol{\theta}) - g(\boldsymbol{\theta}^*) - \boldsymbol{\alpha}^\top \mathsf{H}^{-1} \bar{\Psi} - \frac{\|\boldsymbol{\alpha}\|^2}{2} \right| \leq \rho_0(\boldsymbol{\alpha}). \end{aligned} \quad (15.12)$$

Further, for $\chi_1(x) = \mathsf{H}^{-1}\{\Psi(x) - \bar{\Psi}\}$ and $\boldsymbol{\alpha} = \mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$, it holds

$$\begin{aligned}
\exp\{\boldsymbol{\theta}^\top \Psi(x) - g(\boldsymbol{\theta}^*)\} &= \exp\{(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \Psi(x)\} f(x) \\
&= \exp\{\boldsymbol{\alpha}^\top \chi_1(x)\} \exp(\boldsymbol{\alpha}^\top \mathsf{H}^{-1} \bar{\Psi}) f(x) \\
&= \exp\left\{\boldsymbol{\alpha}^\top \chi_1(x) - \frac{\|\boldsymbol{\alpha}\|^2}{2}\right\} \exp\left(\boldsymbol{\alpha}^\top \mathsf{H}^{-1} \bar{\Psi} + \frac{\|\boldsymbol{\alpha}\|^2}{2}\right) f(x).
\end{aligned}$$

Now by (15.11), for any function $h(x)$

$$\begin{aligned}
\frac{\int h(x) e^{\boldsymbol{\theta}^\top \Psi(x)} \mu_0(dx)}{\int e^{\boldsymbol{\theta}^\top \Psi(x)} \mu_0(dx)} &= \int h(x) \exp\left(\boldsymbol{\alpha}^\top \chi_1(x) - \frac{\|\boldsymbol{\alpha}\|^2}{2}\right) f(x) \mu_0(dx) \\
&\quad \times \exp\left(g(\boldsymbol{\theta}^*) - g(\boldsymbol{\theta}) + \boldsymbol{\alpha}^\top \mathsf{H}^{-1} \bar{\Psi} + \frac{\|\boldsymbol{\alpha}\|^2}{2}\right). \quad (15.13)
\end{aligned}$$

Next, given a unit vector \mathbf{u} , we apply this identity with

$$h(x) \stackrel{\text{def}}{=} \mathbf{u}^\top \mathsf{H}^{-1} \{\Psi(x) - \bar{\Psi} - \mathsf{H}^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\} = \mathbf{u}^\top \{\chi_1(x) - \boldsymbol{\alpha}\}.$$

By (15.11), (15.12), and (15.7)

$$\begin{aligned}
&|\mathbf{u}^\top \mathsf{H}^{-1} \{\nabla g(\boldsymbol{\theta}) - \bar{\Psi} - \mathsf{H}^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}| \\
&\leq e^{\rho_0(\boldsymbol{\alpha})} \left| \int \mathbf{u}^\top \{\chi_1(x) - \boldsymbol{\alpha}\} \exp\left(\boldsymbol{\alpha}^\top \chi_1(x) - \frac{\|\boldsymbol{\alpha}\|^2}{2}\right) f(x) \mu_0(dx) \right| \\
&\leq e^{\rho_0(\boldsymbol{\alpha})} \rho_1(\boldsymbol{\alpha})
\end{aligned}$$

and (15.9) follows. It remains to bound the difference $\nabla^2 g(\boldsymbol{\theta}) - \nabla^2 g(\boldsymbol{\theta}^*)$. We use the identity $\nabla^2 g(\boldsymbol{\theta}) = \text{Var}\{\Psi(X_1)\}$??? which implies for any vector $\boldsymbol{\gamma}$ the representation

$$\begin{aligned}
\boldsymbol{\gamma}^\top \nabla^2 g(\boldsymbol{\theta}) \boldsymbol{\gamma} &= \frac{\int |\boldsymbol{\gamma}^\top \{\Psi(x) - \bar{\Psi} - \mathsf{H}^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}|^2 e^{\boldsymbol{\theta}^\top \Psi(x)} \mu_0(dx)}{\int e^{\boldsymbol{\theta}^\top \Psi(x)} \mu_0(dx)} \\
&\quad - |\boldsymbol{\gamma}^\top \{\nabla g(\boldsymbol{\theta}) - \bar{\Psi} - \mathsf{H}^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}|^2.
\end{aligned}$$

For the second term, it holds by (15.9) for $\boldsymbol{\gamma} = \mathsf{H}^{-1} \mathbf{u}$

$$|\mathbf{u}^\top \mathsf{H}^{-1} \{\nabla g(\boldsymbol{\theta}) - \bar{\Psi} - \mathsf{H}^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}|^2 \leq e^{2\rho_0(\boldsymbol{\alpha})} \rho_1^2(\boldsymbol{\alpha}).$$

For the first term, similarly to (15.13), we obtain

$$\begin{aligned}
&\frac{\int |\mathbf{u}^\top \mathsf{H}^{-1} \{\Psi(x) - \bar{\Psi} - \mathsf{H}^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}|^2 e^{\boldsymbol{\theta}^\top \Psi(x)} \mu_0(dx)}{\int e^{\boldsymbol{\theta}^\top \Psi(x)} \mu_0(dx)} \\
&\leq e^{\rho_0(\boldsymbol{\alpha})} \int |\mathbf{u}^\top \{\chi_1(x) - \boldsymbol{\alpha}\}|^2 \exp\left(\boldsymbol{\alpha}^\top \chi_1(x) - \frac{\|\boldsymbol{\alpha}\|^2}{2}\right) f(x) \mu_0(dx) \\
&\leq e^{\rho_0(\boldsymbol{\alpha}) + \rho_2(\boldsymbol{\alpha})}.
\end{aligned}$$

Putting these two bounds together yields for any $\|\mathbf{u}\| = 1$

$$|\mathbf{u}^\top \mathsf{H}^{-1} \nabla^2 g(\boldsymbol{\theta}) \mathsf{H}^{-1} \mathbf{u} - 1| \leq e^{\rho_0(\boldsymbol{\alpha}) + \rho_2(\boldsymbol{\alpha})} - 1 + e^{2\rho_0(\boldsymbol{\alpha})} \rho_1^2(\boldsymbol{\alpha})$$

and (15.10) follows.

These bounds and the representation of the $L(\boldsymbol{\theta})$ help to obtain the explicit bounds for a quadratic log-likelihood expansion.

Lemma 15.1.2. *Suppose that (\mathbf{x}_1) is fulfilled. Then for any $\boldsymbol{\theta} \in \Theta_0(\mathbf{r})$, it holds with $\boldsymbol{\alpha} = \mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$*

$$\left| L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) - \boldsymbol{\xi}^\top D(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \frac{\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2}{2} \right| \leq n\rho_0(\boldsymbol{\alpha}), \quad (15.14)$$

and

$$\|D^{-1}\{\nabla L(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta}^*)\} - D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \sqrt{n} e^{\rho_0(\boldsymbol{\alpha})} \rho_1(\boldsymbol{\alpha}) \quad (15.15)$$

Proof. By definition

$$L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) = S^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*) - ng(\boldsymbol{\theta}) + ng(\boldsymbol{\theta}^*)$$

and by (15.8) in view of $\mathbb{E}S = n\bar{\Psi}$

$$\begin{aligned} & \left| L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) - (S - \mathbb{E}S)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \frac{n\|\mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2}{2} \right| \\ &= n \left| g(\boldsymbol{\theta}) - g(\boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \bar{\Psi} - \frac{1}{2} \|\mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \right| \leq n\rho_0(\boldsymbol{\alpha}) \end{aligned}$$

and (15.14) follows by $\boldsymbol{\xi} = D^{-1}(S - \mathbb{E}S) = n^{-1/2} \mathsf{H}^{-1}(S - \mathbb{E}S)$. Further, in a similar way we derive using $\nabla g(\boldsymbol{\theta}^*) = \bar{\Psi}$ and (15.9)

$$\begin{aligned} & \|D^{-1}\{\nabla L(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta}^*)\} - D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \\ &= \|\sqrt{n} \mathsf{H}^{-1}\{\nabla g(\boldsymbol{\theta}) - \nabla g(\boldsymbol{\theta}^*)\} - \sqrt{n} \mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \\ &\leq \sqrt{n} e^{\rho_0(\boldsymbol{\alpha})} \rho_1(\boldsymbol{\alpha}), \end{aligned}$$

and (15.15) follows as well.

Next we check condition **(ED₀)**. Remind that **(ED₂)** is fulfilled automatically because the stochastic component of $L(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$. The identities $D^2 = n\mathsf{H}^2$ and (15.6) yield the following representation for the normalized score vector $\boldsymbol{\xi} = D^{-1}\nabla L(\boldsymbol{\theta}^*) = D^{-1}\nabla\zeta(\boldsymbol{\theta}^*)$:

$$\boldsymbol{\xi} = D^{-1}(S - \mathbb{E}S) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathsf{H}^{-1}\{\Psi(X_i) - \bar{\Psi}\}. \quad (15.16)$$

Now condition (χ_1) easily implies (ED_0) .

Lemma 15.1.3. *Suppose (χ_1) . Then (ED_0) is fulfilled with $\mathbf{g} = \sqrt{n}\mathbf{g}_1$.*

Proof. I.i.d. structure of the X_i 's implies by (15.16) and (χ_1) for any vector $\boldsymbol{\gamma}$ with $\|\boldsymbol{\gamma}\| \leq \mathbf{g}_1\sqrt{n}$ and $\boldsymbol{\alpha} = n^{-1/2}\boldsymbol{\gamma}$

$$\left| \log \mathbb{E} \exp(\boldsymbol{\gamma}^\top \boldsymbol{\xi}) - \frac{\|\boldsymbol{\gamma}\|^2}{2} \right| = \left| n \log \mathbb{E} \exp\left(\boldsymbol{\alpha}^\top \boldsymbol{\chi}_1 - \frac{\|\boldsymbol{\alpha}\|^2}{2}\right) \right| \leq n\rho_0(\boldsymbol{\alpha}).$$

Moreover, see Remark 15.1.1,

$$n\rho_0(\boldsymbol{\alpha}) \leq n\mathbf{w}_0\|\boldsymbol{\alpha}\|^3 \leq \frac{\mathbf{w}_0\|\boldsymbol{\gamma}\|}{\sqrt{n}}\|\boldsymbol{\gamma}\|^2$$

and (ED_0) follows with ν_0 close to one provided that $\mathbf{w}_0\|\boldsymbol{\gamma}\| \ll \sqrt{n}$.

Conditions (15.7) in (χ_1) can be restrictive in the nonparametric setup. In particular, in many examples the constants $\mathbf{g}_1, \mathbf{w}_{0,1,2}$ depend on the basis $\{\psi_j\}$ and they can even explode as its cardinality grows.

Exercise 15.1.1. Consider a histogram basis $\psi_j = \mathbb{I}(x \in A_j)$ for a given partition $A_j, j = 1, \dots, p$. Check that H^2 is diagonal with the diagonal entries $P_j - P_j^2$ for $P_j = P(A_j)$. Show (χ_1) for $\mathbf{g}_1^2 = \min_{j \leq p} P_j$.

Applicability of the general results to the histogram example requires $\mathbf{g} = \sqrt{n}\mathbf{g}_1 \gg p$. This condition can be violated if the partition is too fine. Below we consider two options for keeping this constant under control: sieve parametric and penalized nonparametric estimation.

15.2 Sieve nonparametric density estimation

This section explains how the sieve nonparametric approach can be applied to the problem of log-density estimation. We follow the line of Section 14.3 from the GLM case.

Let $\boldsymbol{\theta}^*$ be the nonparametric true point:

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \mathbb{E}L(\boldsymbol{\theta}).$$

Let m be the index for the sieve subspace Θ_m of dimension p_m . The sieve MLE $\tilde{\boldsymbol{\theta}}_m$ reads

$$\tilde{\boldsymbol{\theta}}_m \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta_m} L(\boldsymbol{\theta}).$$

Its target $\boldsymbol{\theta}_m^*$ can be defined as

$$\boldsymbol{\theta}_m^* \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta_m} \mathbb{E}L(\boldsymbol{\theta}).$$

Alternatively, one can optimize the quadratic approximation $\mathbb{E}L(\boldsymbol{\theta}) \approx \mathbb{E}L(\boldsymbol{\theta}^*) - \frac{1}{2}n\|\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2$ to get an explicit expression:

$$\boldsymbol{\theta}_m^* \stackrel{\text{def}}{=} \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta_m} \|\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2.$$

The solution $\boldsymbol{\theta}_m^*$ satisfies

$$\Pi_m \mathbf{H}^2 (\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*) = 0,$$

where Π_m means the projector on the sieve subspace. For the explicit expression, introduce the block representation of the matrix \mathbf{H}^2 :

$$\mathbf{H}^2 = \begin{pmatrix} \mathbf{H}_m^2 & A_{m\nu} \\ A_{m\nu}^\top & \mathbf{H}_\nu^2 \end{pmatrix}.$$

Then

$$\boldsymbol{\theta}_m^* - \Pi_m \boldsymbol{\theta}^* = \mathbf{H}_m^{-2} A_{m\nu} \boldsymbol{\theta}^*. \quad (15.17)$$

Note that $\boldsymbol{\theta}_m^*$ does not coincide with the projection $\Pi_m \boldsymbol{\theta}^*$ of the true vector $\boldsymbol{\theta}^*$ onto Θ_m unless the orthogonality conditions $A_{m\nu} = 0$ holds and the matrix \mathbf{H}^2 has block-diagonal structure $\mathbf{H}^2 = \text{block}\{\mathbf{H}_m^2, \mathbf{H}_\nu^2\}$.

The block \mathbf{H}_m^2 serves as sieve Fisher information matrix $\mathbf{H}_m^2 = \nabla_m^2 g(\boldsymbol{\theta}^*)$ for one observation, where ∇_m means derivative along the sieve subspace. Below we assume the identifiability condition

$$\|\mathbf{H}_m^{-1} A_{m\nu} \mathbf{H}_\nu^{-1}\|^2 \leq \rho_m < 1.$$

It implies

$$\mathbf{H}^2 \leq (1 + \rho_m) \text{block}\{\mathbf{H}_m^2, \mathbf{H}_\nu^2\}, \quad \mathbf{H}^{-2} \leq \frac{1}{1 - \rho_m} \text{block}\{\mathbf{H}_m^{-2}, \mathbf{H}_\nu^{-2}\}. \quad (15.18)$$

The total sieve Fisher information is n times multiple of \mathbf{H}_m^2 : $D_m^2 = n\mathbf{H}_m^2$. The local root-n vicinity $\Theta_m(r)$ of $\boldsymbol{\theta}_m^*$ can be represented as

$$\Theta_m(r) = \{\boldsymbol{\theta} \in \Theta_m : \sqrt{n}\|\mathbf{H}_m(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*)\| \leq r\}. \quad (15.19)$$

As the conditions on the model are stated in terms of the true density f corresponding to $\boldsymbol{\theta}^*$, it is convenient to link $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_m^*$ to each other. We therefore, consider another vicinity

$$\Theta_m^s(\mathbf{r}) = \{\boldsymbol{\theta} \in \Theta_m : \sqrt{n}\|\mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}\}.$$

It obviously holds

$$\Theta_m(\mathbf{r}) \subseteq \Theta_m^s(\mathbf{r} + \|\mathbf{b}_m\|),$$

where the sieve bias \mathbf{b}_m is given by

$$\mathbf{b}_m = D(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*) = \sqrt{n}\mathsf{H}(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*). \quad (15.20)$$

Both definition can be equally used under the conditions of “small bias” $\|\mathbf{b}_m\|$.

The normalized score vector $\boldsymbol{\xi}_m = D_m^{-1}\{\nabla_m L(\boldsymbol{\theta}^*) - \nabla_m \mathbb{E}L(\boldsymbol{\theta}^*)\} = D_m^{-1}\nabla_m \zeta(\boldsymbol{\theta}^*)$ reads

$$\boldsymbol{\xi}_m = D_m^{-1}(S_m - \mathbb{E}S_m) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathsf{H}_m^{-1}\{\Psi_m(X_i) - \bar{\Psi}_m\}. \quad (15.21)$$

Here $\Psi_m(X_i)$ denotes the m -subvector of $\Psi(X_i)$. The result on local concentration of $\tilde{\boldsymbol{\theta}}_m$ is a specification of the general large deviation bound for a MLE. The i.i.d. assumption on the sample X_1, \dots, X_n simplifies the formulation because the “noise” structure of the model is correctly specified and, in particular, the matrices H_m^2 and $\text{Var}\{\Psi_m(X_1)\}$ coincide. This implies $\text{Var}(\boldsymbol{\chi}_m) = I_m$ for $\boldsymbol{\chi}_m = \mathsf{H}_m^{-1}\{\Psi_m(X_1) - \bar{\Psi}_m\}$.

Theorem 15.2.1. Consider the sieve MLE $\tilde{\boldsymbol{\theta}}_m = \text{argmax}_{\boldsymbol{\theta} \in \Theta_m} L(\boldsymbol{\theta})$ for a sieve dimension $p_m = \dim(\Theta_m)$. Let condition **(C1)** hold for $\boldsymbol{\chi}_m = \mathsf{H}_m^{-1}\{\Psi_m(X_1) - \bar{\Psi}_m\}$ and let the standardized score vector $\boldsymbol{\xi}_m$ from (15.21) satisfy $\|\boldsymbol{\xi}_m\| \leq z(p_m, \mathbf{x})$ on a set $\Omega_m(\mathbf{x})$ with $\mathbb{P}(\Omega_m(\mathbf{x})) \geq 1 - 2e^{-x}$. Let also hold for a radius \mathbf{r}_m

$$\diamondsuit_m(\mathbf{r}_m) \stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta} \in \Theta_m^s(\mathbf{r}_m)} \sqrt{n}(1 + \rho_m)^{1/2} e^{\rho_0(\boldsymbol{\alpha})} \rho_1(\boldsymbol{\alpha}) \leq \frac{1}{2},$$

and let \mathbf{r}_m and \mathbf{b}_m from (15.20) fulfill

$$\{\mathbf{r}_m - \|\mathbf{b}_m\| - 2\diamondsuit_m(\mathbf{r}_m)\} \geq 2z(p_m, \mathbf{x}). \quad (15.22)$$

Then $\tilde{\boldsymbol{\theta}}_m$ concentrates in the local vicinity $\Theta_m(\mathbf{r}_m)$ from (15.19): on a set $\Omega_m(\mathbf{x})$

$$\tilde{\boldsymbol{\theta}}_m \in \Theta_m(\mathbf{r}_m).$$

Moreover, on the set $\Omega_m(\mathbf{x})$, it holds

$$\begin{aligned}\|\sqrt{n}\mathsf{H}_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*) - \boldsymbol{\xi}_m\| &\leq \diamond_m(\mathbf{r}_m), \\ \|\sqrt{n}\mathsf{H}_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*) - \boldsymbol{\xi}_m - \mathbf{b}_m\| &\leq \mathbf{r}_m \delta_m(\mathbf{r}_m),\end{aligned}\tag{15.23}$$

and

$$\begin{aligned}\left|L(\tilde{\boldsymbol{\theta}}_m) - L(\boldsymbol{\theta}_m^*) - \frac{1}{2}\|\boldsymbol{\xi}_m\|^2\right| &\leq \mathbf{r}_m \diamond(\mathbf{r}_m) + \frac{1}{2}\diamond_m^2(\mathbf{r}_m), \\ \left|\sqrt{2L(\tilde{\boldsymbol{\theta}}_m) - 2L(\boldsymbol{\theta}_m^*)} - \|\boldsymbol{\xi}_m\|\right| &\leq 3\diamond_m(\mathbf{r}_m).\end{aligned}$$

Moreover, for any $\boldsymbol{\theta}^\circ \in \Theta_m(\mathbf{r}_m)$, it holds on $\Omega(\mathbf{x})$ with $\mathbf{u}_m = D_m(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}_m^*)$

$$\begin{aligned}\left|L(\tilde{\boldsymbol{\theta}}_m) - L(\boldsymbol{\theta}^\circ) - \frac{1}{2}\|\boldsymbol{\xi}_m + \mathbf{u}_m\|^2\right| &\leq 2\mathbf{r}_m \diamond(\mathbf{r}_m) + \frac{1}{2}\diamond_m^2(\mathbf{r}_m), \\ \left|\sqrt{2L(\tilde{\boldsymbol{\theta}}_m) - 2L(\boldsymbol{\theta}^\circ)} - \|\boldsymbol{\xi}_m + \mathbf{b}_m\|\right| &\leq 5\diamond_m(\mathbf{r}_m).\end{aligned}\tag{15.24}$$

Proof. The result on concentration of $\tilde{\boldsymbol{\theta}}_m$ in $\Theta_m(\mathbf{r}_m)$ can be proved exactly as a similar result for a GLM; see Theorem 14.2.1. The proof only uses that the log-likelihood function is concave in $\boldsymbol{\theta}$, its stochastic component is linear in $\boldsymbol{\theta}$, and the expectation $\mathbb{E}L(\boldsymbol{\theta})$ fulfills **(ED₀)** on $\Theta_m(\mathbf{r}_m)$. We, however, present an independent proof. The basic step of the proof is the log-likelihood expansion on the sieve Θ_m .

First we specify the result of Lemma 15.1.1 to the sieve case.

Lemma 15.2.1. *It holds for any $\boldsymbol{\theta} \in \Theta_m$*

$$\sqrt{n}\|\mathsf{H}_m^{-1}\{\nabla_m g(\boldsymbol{\theta}) - \bar{\Psi}_m\} - \mathsf{H}_m(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*)\| \leq \diamond_m(\mathbf{r}_m),\tag{15.25}$$

$$n\|g(\boldsymbol{\theta}) - g(\boldsymbol{\theta}_m^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}_m^*)^\top \bar{\Psi}_m - \frac{1}{2}\mathsf{H}_m^2(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*)\| \leq \mathbf{r}_m \diamond_m(\mathbf{r}_m).\tag{15.26}$$

Proof. For any $\boldsymbol{\theta} \in \Theta_m$, Lemma 15.1.1 yields

$$\mathsf{H}^{-1}\{\nabla g(\boldsymbol{\theta}) - \nabla g(\boldsymbol{\theta}^*)\} - \mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) = \mathsf{H}^{-1}\{\nabla g(\boldsymbol{\theta}) - \bar{\Psi}\} - \mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) = \boldsymbol{\beta}(\boldsymbol{\alpha})$$

with $\boldsymbol{\alpha} = \mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ and $\|\boldsymbol{\beta}(\boldsymbol{\alpha})\| \leq \rho_1(\boldsymbol{\alpha})e^{\rho_0(\boldsymbol{\alpha})}$. Now we decompose

$$\mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) = \mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*) - \mathsf{H}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_m^*),$$

yielding

$$\nabla g(\boldsymbol{\theta}) - \bar{\Psi} - \mathsf{H}^2(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*) = \mathsf{H}\boldsymbol{\beta}(\boldsymbol{\alpha}) + \mathsf{H}^2(\boldsymbol{\theta}^* - \boldsymbol{\theta}_m^*).\tag{15.27}$$

For any $\boldsymbol{\theta} \in \Theta_m$, it holds $\boldsymbol{\theta} - \boldsymbol{\theta}_m^* = \Pi_m(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*) \in \Theta_m$ and by (15.17) $\Pi_m \mathsf{H}^2(\boldsymbol{\theta}^* - \boldsymbol{\theta}_m^*) = 0$. Thus, projecting the expansion (15.27) onto the sieve subspace Θ_m and scaling by H_m implies

$$\mathsf{H}_m^{-1} \{ \nabla_m g(\boldsymbol{\theta}) - \bar{\Psi}_m \} - \mathsf{H}_m(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*) = \mathsf{H}_m^{-1} \Pi_m \mathsf{H} \boldsymbol{\beta}(\boldsymbol{\alpha}).$$

The bound (15.18) implies

$$\| \mathsf{H}_m^{-1} \Pi_m \mathsf{H} \boldsymbol{\beta}(\boldsymbol{\alpha}) \| \leq (1 + \rho_m)^{1/2} \| \boldsymbol{\beta}(\boldsymbol{\alpha}) \| \leq (1 + \rho_m)^{1/2} \rho_1(\boldsymbol{\alpha}) e^{\rho_0(\boldsymbol{\alpha})}.$$

Putting altogether yields (15.25).

To check (15.26), observe that

$$\begin{aligned} g(\boldsymbol{\theta}) - g(\boldsymbol{\theta}_m^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}_m^*)^\top \bar{\Psi}_m - \frac{1}{2} \mathsf{H}_m^2(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*) \\ = (\boldsymbol{\theta} - \boldsymbol{\theta}_m^*)^\top \{ \nabla_m g(\boldsymbol{\theta}^\circ) - \bar{\Psi}_m - \mathsf{H}_m^2(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}_m^*) \} \\ = \{ \mathsf{H}_m(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*) \}^\top [\mathsf{H}_m^{-1} \{ \nabla_m g(\boldsymbol{\theta}^\circ) - \bar{\Psi}_m \} - \mathsf{H}_m(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}_m^*)], \end{aligned}$$

where $\boldsymbol{\theta}^\circ$ is a point on the line connecting $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_m^*$. It remains to note that $\sqrt{n} \| \mathsf{H}_m(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*) \| \leq \mathbf{r}_m$, and the same with $\boldsymbol{\theta}^\circ$ in place of $\boldsymbol{\theta}$.

For the gradient $\nabla L(\boldsymbol{\theta}) = S - n \nabla g(\boldsymbol{\theta})$ of the log-likelihood $L(\boldsymbol{\theta})$, (15.25) implies with $D_m^2 = n \mathsf{H}_m^2$

$$\begin{aligned} & \| D_m^{-1} \{ \nabla_m L(\boldsymbol{\theta}) - \nabla_m L(\boldsymbol{\theta}^*) \} + D_m(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*) \| \\ &= \sqrt{n} \| \mathsf{H}_m^{-1} \{ \nabla_m g(\boldsymbol{\theta}) - \nabla_m g(\boldsymbol{\theta}^*) \} - \mathsf{H}_m(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*) \| \leq \diamond_m(\mathbf{r}_m). \end{aligned} \quad (15.28)$$

One can further use that $\nabla \mathbb{E} L(\boldsymbol{\theta}^*) = 0$, and thus, $\nabla L(\boldsymbol{\theta}^*) = \nabla L(\boldsymbol{\theta}^*) - \nabla \mathbb{E} L(\boldsymbol{\theta}^*) = S - \mathbb{E} S$. Projecting onto the sieve space yields $\nabla_m L(\boldsymbol{\theta}^*) = S_m - \mathbb{E} S_m$. For $\boldsymbol{\xi}_m = D_m^{-1} \nabla_m L(\boldsymbol{\theta}^*) = D_m^{-1} (S_m - \mathbb{E} S_m)$, the expansion (15.28) implies

$$\| D_m^{-1} \nabla_m L(\boldsymbol{\theta}) - \boldsymbol{\xi}_m + D_m(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*) \| \leq \diamond_m(\mathbf{r}_m). \quad (15.29)$$

A quadratic approximation of the log-likelihood $L(\boldsymbol{\theta})$ can be obtained from the linear expansion for the gradient $\nabla L(\boldsymbol{\theta})$:

$$\begin{aligned} & L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_m^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}_m^*)^\top D_m \boldsymbol{\xi}_m + \frac{1}{2} \| D_m(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*) \|^2 \\ &= (\boldsymbol{\theta} - \boldsymbol{\theta}_m^*)^\top \{ \nabla_m L(\boldsymbol{\theta}') - D_m \boldsymbol{\xi}_m + D_m^2(\boldsymbol{\theta}' - \boldsymbol{\theta}_m^*) \} \\ &= \{ D_m(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*) \}^\top \{ D_m^{-1} \nabla_m L(\boldsymbol{\theta}') - \boldsymbol{\xi}_m - D_m(\boldsymbol{\theta}' - \boldsymbol{\theta}_m^*) \}, \end{aligned}$$

where $\boldsymbol{\theta}'$ is a point on the line connecting $\boldsymbol{\theta}_m^*$ and $\boldsymbol{\theta}$. The use of $\| D_m(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*) \| \leq \mathbf{r}_m$ and of (15.29) yields

$$\left| L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_m^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}_m^*)^\top D_m \boldsymbol{\xi}_m + \frac{1}{2} \| D_m(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*) \|^2 \right| \leq \mathbf{r}_m \diamond_m(\mathbf{r}_m). \quad (15.30)$$

Now suppose for a moment that $\tilde{\boldsymbol{\theta}}_m$ lies outside of $\Theta_m(\mathbf{r}_m)$ and, in particular, $L(\tilde{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta}_m^*)$. As the log-likelihood function $L(\boldsymbol{\theta})$ is concave in $\boldsymbol{\theta} \in \Theta_m$, the condition $L(\tilde{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta}_m^*)$ implies that there is a point $\boldsymbol{\theta}^\circ$ on the boundary of the ball $\Theta_m(\mathbf{r}_m)$ with the same property $L(\boldsymbol{\theta}^\circ) \geq L(\boldsymbol{\theta}_m^*)$. However, it is impossible under the condition $2\|\boldsymbol{\xi}_m\| \leq \mathbf{r}_m - 2\Diamond_m(\mathbf{r}_m)$, because $\|D_m(\boldsymbol{\theta} - \boldsymbol{\theta}_m^*)\| = \mathbf{r}_m$ and (15.30) implies

$$L(\boldsymbol{\theta}^\circ) - L(\boldsymbol{\theta}_m^*) \leq \mathbf{r}_m \|\boldsymbol{\xi}_m\| - \frac{1}{2}\mathbf{r}_m^2 + \mathbf{r}_m \Diamond_m(\mathbf{r}_m) < 0.$$

Let now $\Omega_m(\mathbf{x})$ mean the event of a high probability on which $\|\boldsymbol{\xi}_m\| \leq z(\mathbf{r}_m, \mathbf{x})$. Then the inequality $\mathbf{r}_m(1 - 2\Diamond_m(\mathbf{r}_m)) \geq 2z(\mathbf{r}_m, \mathbf{x})$ ensures that $\tilde{\boldsymbol{\theta}}_m \in \Theta_m(\mathbf{r}_m)$. After restricting to this set $\Omega_m(\mathbf{x})$, we can plug $\tilde{\boldsymbol{\theta}}_m$ in place of $\boldsymbol{\theta}$ in the expansion (15.29) of the gradient $\nabla_m L(\boldsymbol{\theta})$ yielding by $\nabla_m L(\tilde{\boldsymbol{\theta}}_m) = 0$

$$\|D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*) - \boldsymbol{\xi}_m\| \leq \Diamond_m(\mathbf{r}_m).$$

Similarly, the use of $\tilde{\boldsymbol{\theta}}_m$ in place of $\boldsymbol{\theta}$ in the log-likelihood expansion (15.30) yields

$$\begin{aligned} |L(\tilde{\boldsymbol{\theta}}_m) - L(\boldsymbol{\theta}_m^*) - \frac{1}{2}\|\boldsymbol{\xi}_m\|^2| &\leq \frac{1}{2}\|D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*) - \boldsymbol{\xi}_m\|^2 + \mathbf{r}_m \Diamond_m(\mathbf{r}_m) \\ &\leq \mathbf{r}_m \Diamond_m(\mathbf{r}_m) + \frac{1}{2}\Diamond_m^2(\mathbf{r}_m). \end{aligned}$$

If $\boldsymbol{\theta}^\circ$ is any other point in $\Theta(\mathbf{r}_m)$, then (15.29) implies with $\mathbf{u}_m = D_m(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^\circ)$ that

$$\|D_m^{-1}\nabla_m L(\boldsymbol{\theta}) - \boldsymbol{\xi}_m - \mathbf{u}_m + D_m(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\| \leq \Diamond_m(\mathbf{r}_m). \quad (15.31)$$

In particular, with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_m$, it holds on $\Omega(\mathbf{x})$ due to $\nabla_m L(\tilde{\boldsymbol{\theta}}_m) = 0$

$$\|D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^\circ) - \boldsymbol{\xi}_m - \mathbf{u}_m\| \leq \Diamond_m(\mathbf{r}_m). \quad (15.32)$$

Now

$$\begin{aligned} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^\circ) - (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top D_m(\boldsymbol{\xi}_m + \mathbf{u}_m) + \frac{1}{2}\|D_m(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\|^2 \\ = (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top \{\nabla_m L(\boldsymbol{\theta}') - D_m(\boldsymbol{\xi}_m + \mathbf{u}_m) + D_m^2(\boldsymbol{\theta}' - \boldsymbol{\theta}^\circ)\} \\ = \{D_m(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\}^\top \{D_m^{-1}\nabla_m L(\boldsymbol{\theta}') - \boldsymbol{\xi}_m - \mathbf{u}_m + D_m(\boldsymbol{\theta}' - \boldsymbol{\theta}^\circ)\} \\ \leq \|D_m(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\| \Diamond_m(\mathbf{r}_m), \end{aligned} \quad (15.33)$$

where $\boldsymbol{\theta}'$ is a point on the line between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\circ$. As $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\circ$ belong to $\Theta_m(\mathbf{r}_m)$, it holds $\|D_m(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\| \leq 2\mathbf{r}_m$. One can rewrite (15.33) as

$$\left| L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^\circ) - \frac{1}{2}\|\boldsymbol{\xi}_m + \mathbf{u}_m\|^2 + \frac{1}{2}\|D_m(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) - \boldsymbol{\xi}_m - \mathbf{u}_m\|^2 \right| \leq 2\mathbf{r}_m \Diamond_m(\mathbf{r}_m).$$

The use $\tilde{\boldsymbol{\theta}}_m$ in place of $\boldsymbol{\theta}$ yields on $\Omega(\mathbf{x})$ by (15.32)

$$\left| L(\tilde{\boldsymbol{\theta}}_m) - L(\boldsymbol{\theta}^\circ) - \frac{1}{2} \|\xi_m + \mathbf{u}_m\|^2 \right| \leq \frac{1}{2} \diamondsuit_m^2(\mathbf{r}_m) + 2\mathbf{r}_m \diamondsuit_m(\mathbf{r}_m).$$

In a similar way, (15.32) and (15.33) yield on $\Omega(\mathbf{x})$

$$\begin{aligned} & \left| L(\tilde{\boldsymbol{\theta}}_m) - L(\boldsymbol{\theta}^\circ) - \frac{1}{2} \|D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^\circ)\|^2 \right| \\ & \leq \|D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^\circ)\| \diamondsuit_m(\mathbf{r}_m) + \|D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^\circ)\| \|D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^\circ) - \xi_m - \mathbf{u}_m\| \\ & \leq 2 \|D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^\circ)\| \diamondsuit_m(\mathbf{r}_m). \end{aligned}$$

This also implies

$$\begin{aligned} & \left| \sqrt{2L(\tilde{\boldsymbol{\theta}}_m) - 2L(\boldsymbol{\theta}^\circ)} - \|D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^\circ)\| \right| \\ & \leq \frac{2L(\tilde{\boldsymbol{\theta}}_m) - 2L(\boldsymbol{\theta}^\circ) - \|D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^\circ)\|^2}{\|D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^\circ)\|} \leq 4 \diamondsuit_m(\mathbf{r}_m). \end{aligned}$$

This and (15.32) yields (15.24).

Remark 15.2.1. Condition (15.22) on \mathbf{r}_m is sufficient to show that the sieve MLE $\tilde{\boldsymbol{\theta}}_m$ well concentrates in the vicinity $\Theta_m(\mathbf{r}_m)$ for $\mathbf{r}_m > 2\|\mathbf{b}_m\| + 2z(p_m, \mathbf{x}) \asymp \|\mathbf{b}_m\| + \sqrt{p_m + \mathbf{x}}$. The vector ξ_m can be bounded by $z(p_m, \mathbf{x})$ on $\Omega_m(\mathbf{x})$. The bias term $\|\mathbf{b}_m\|$ does not dominate in the loss expansion (15.23) if $\|\mathbf{b}_m\|^2 \leq C p_m$.

15.3 Sieve likelihood ratio test

Sieve approach usually assumes that there is a growing sequence of sieve subsets $\Theta_1 \subset \Theta_2 \subset \dots \subset \Theta$. For notational simplicity, let $\Theta_m \subset \mathbb{R}^m$ and $\dim(\Theta_m) = m$. Inference and model selection on sieves often relies on the likelihood ratio test which compares the maximum values $L(\tilde{\boldsymbol{\theta}}_m)$ for different m . Consider $m > m^\circ$ and suppose that Θ_{m° is naturally embedded in Θ_m . The bias $\mathbf{b}_{m,m^\circ} \stackrel{\text{def}}{=} D_m(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}_{m^\circ}^*)$ is assumed not too large. The Fisher expansion

$$D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*) \approx \xi_m = D_m^{-1} \nabla_m L(\boldsymbol{\theta}^*)$$

for the sieve model m and a similar expansion for the model m° yield

$$D_m(\tilde{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_{m^\circ}) \approx \xi_{m,m^\circ} + \mathbf{b}_{m,m^\circ}$$

where

$$\begin{aligned}\xi_{m,m^\circ} &\stackrel{\text{def}}{=} \xi_m - D_m \Pi_{m^\circ}^\top D_{m^\circ}^{-1} \xi_m \\ &= D_m^{-1} \{ I_m - D_m^2 \Pi_{m^\circ}^\top D_{m^\circ}^{-2} \Pi_{m^\circ} \} \nabla_m L(\boldsymbol{\theta}^*),\end{aligned}$$

where Π_{m° is a projector from \mathbb{R}^m . Further, the Wilks expansion for $L(\tilde{\boldsymbol{\theta}}_m)$ implies

$$2L(\tilde{\boldsymbol{\theta}}_m) - 2L(\tilde{\boldsymbol{\theta}}_{m^\circ}) \approx \|D_m(\tilde{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_{m^\circ})\|^2 \approx \|\xi_{m,m^\circ} + b_{m,m^\circ}\|^2.$$

Under the null hypothesis $\boldsymbol{\theta}_{m^\circ} = \boldsymbol{\theta}_m$, we obtain

$$2L(\tilde{\boldsymbol{\theta}}_m) - 2L(\tilde{\boldsymbol{\theta}}_{m^\circ}) \approx \|\xi_{m,m^\circ}\|^2.$$

15.4 Error of estimation for the log-density function

This section discusses the implication of the error bound of Theorem 15.2.1 on the error of estimating the unknown log-density function $\log f$.

Let m be a fixed sieve index. The sieve estimator $\tilde{\boldsymbol{\theta}}_m$ of the vector of coefficients $\boldsymbol{\theta}^*$ maximizes the log-likelihood $L(\boldsymbol{\theta})$ over the sieve subset Θ_m and it yields the estimator of the log-density

$$\log \tilde{f}_m(x) = \tilde{\boldsymbol{\theta}}_m^\top \Psi(x) - g(\tilde{\boldsymbol{\theta}}_m)$$

of the log-density $\log f(x)$; cf (15.3).

Let also $\boldsymbol{\theta}_m^*$ be the related target, and $\log f_m(x) = \boldsymbol{\theta}_m^{*\top} \Psi_m(x) - g(\boldsymbol{\theta}_m^*)$, the corresponding log-density. The bias $b_m(x)$ is defined as

$$\begin{aligned}b_m(x) &= \log f_m(x) - \log f(x) \\ &= \boldsymbol{\theta}_m^{*\top} \Psi_m(x) - g(\boldsymbol{\theta}_m^*) - \boldsymbol{\theta}^{*\top} \Psi(x) - g(\boldsymbol{\theta}^*).\end{aligned}$$

Its mean and ℓ_2 -norm are defined as

$$\begin{aligned}\bar{b}_m &\stackrel{\text{def}}{=} \int b_m(x) f(x) \mu_0(dx), \\ \|b_m\|^2 &\stackrel{\text{def}}{=} \int b_m^2(x) f(x) \mu_0(dx).\end{aligned}$$

The definition implies

$$\bar{b}_m = - \int \log \frac{f(x)}{f_m(x)} f(x) \mu_0(dx) = -\mathcal{K}(f, f_m) \quad (15.34)$$

for the Kullback-Leibler divergence $\mathcal{K}(f, f_m)$ between f and f_m . Obviously $\bar{b}_m^2 \leq \|b_m\|^2$. Below in this section we study the Kullback-Leibler and integrated squared loss and risk of this estimator.

15.4.1 Kullback-Leibler divergence

Let $f(x)$ be the true density and $\tilde{f}_m(x)$ its sieve estimator. As mentioned in (15.34), a natural measure of the estimation quality is the Kullback-Leibler (KL) divergence:

$$\begin{aligned}\mathcal{K}(f, \tilde{f}_m) &= \int \left\{ \log f(x) - \log \tilde{f}_m(x) \right\} f(x) \mu_0(dx) \\ &= \int \log \left(\frac{f(x)}{\tilde{f}_m(x)} \right) f(x) \mu_0(dx).\end{aligned}$$

The result on the KL-loss of \tilde{f}_m will be stated under the same conditions as in Theorem 15.2.1.

Theorem 15.4.1. *Suppose that the conditions of Theorem 15.2.1 are fulfilled. Let also the random set $\Omega_m(\mathbf{x})$ be defined there. Then it holds on $\Omega_m(\mathbf{x})$*

$$\begin{aligned}\left| n\mathcal{K}(f, \tilde{f}_m) - \frac{n}{2} \|\mathsf{H}_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*)\|^2 - n\mathcal{K}(f, f_m) \right| &\leq \mathbf{r}_m \diamondsuit_m(\mathbf{r}_m), \\ \left| n\mathcal{K}(f, \tilde{f}_m) - \frac{1}{2} \|\boldsymbol{\xi}_m\|^2 - n\mathcal{K}(f, f_m) \right| &\leq \Delta_m^{\mathcal{K}},\end{aligned}\tag{15.35}$$

where the vector $\boldsymbol{\xi}_m$ is given by (15.21) and $\Delta_m^{\mathcal{K}}$ fulfills on $\Omega_m(\mathbf{x})$

$$\Delta_m^{\mathcal{K}} \leq \{z(p_m, \mathbf{x}) + \mathbf{r}_m\} \diamondsuit_m + \frac{1}{2} \diamondsuit_m^2$$

Proof. We use that

$$\begin{aligned}n \int \{\log f(x) - \log \tilde{f}_m(x)\} f(x) \mu_0(dx) \\ = n \int \{\log f(x) - \log f_m(x)\} f(x) \mu_0(dx) \\ + n \int \{\log f_m(x) - \log \tilde{f}_m(x)\} f(x) \mu_0(dx) \\ = n\mathcal{K}(f, f_m) - n(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*)^\top \int \Psi_m(x) f(x) \mu_0(dx) + ng(\tilde{\boldsymbol{\theta}}_m) - ng(\boldsymbol{\theta}_m).\end{aligned}$$

Further, by definition

$$\int \Psi_m(x) f(x) \mu_0(dx) = \bar{\Psi}_m.$$

On the set $\Omega_m(\mathbf{x})$, the estimate $\tilde{\boldsymbol{\theta}}_m$ belongs to the local vicinity $\Theta_m(\mathbf{r}_m)$, and (15.26) implies

$$\begin{aligned}n\mathcal{K}(f, \tilde{f}_m) &= n\mathcal{K}(f, f_m) - n(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*)^\top \bar{\Psi}_m + ng(\tilde{\boldsymbol{\theta}}_m) - ng(\boldsymbol{\theta}_m^*) \\ &\leq n\mathcal{K}(f, f_m) + \frac{n}{2} \|\mathsf{H}_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*)\|^2 + \mathbf{r}_m \diamondsuit_m(\mathbf{r}_m).\end{aligned}$$

The bound (15.23) can be written in the form

$$n^{1/2} \mathsf{H}_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*) = \boldsymbol{\xi}_m + \boldsymbol{\tau}_m$$

with $\boldsymbol{\xi}_m$ from (15.21) and $\|\boldsymbol{\tau}_m\| \leq \diamond_m$. Moreover, it holds on $\Omega_m(\mathbf{x})$ that $\|\boldsymbol{\xi}_m\| \leq z(p_m, \mathbf{x})$ and

$$\left| \|\boldsymbol{\xi}_m + \boldsymbol{\tau}_m\|^2 - \|\boldsymbol{\xi}_m\|^2 \right| \leq 2|\boldsymbol{\xi}_m^\top \boldsymbol{\tau}_m| + \|\boldsymbol{\tau}_m\|^2 \leq 2z(p_m, \mathbf{x})\diamond_m + \diamond_m^2,$$

and the assertion (15.35) follows as well.

The result (15.35) is usually referred to as ‘‘bias-variance decomposition’’ of the estimation loss $n\mathcal{K}(f, \tilde{f}_m)$. Indeed, the deterministic terms $\mathcal{K}(f, f_m)$ describes the distance between the true density f and its sieve approximation f_m , while the term $\|\boldsymbol{\xi}_m\|^2$ measures the error of statistical estimation in the sieve model Θ_m .

Remark 15.4.1. Condition **(x1)** yields the bounds $\rho_0(\mathbf{r}_m) \leq w_0(\mathbf{r}_m/\sqrt{n})^3$, $\rho_1(\mathbf{r}_m) \leq w_1 \mathbf{r}_m^2/n$, where \mathbf{r}_m should be at least of order $\sqrt{p_m + \mathbf{x}}$; cf. Remark 15.1.1. The value $z(p_m, \mathbf{x})$ is of the same order. This implies that $\diamond_m \approx w_1(p_m + \mathbf{x})/\sqrt{n}$ and

$$\Delta_m^{\mathcal{K}} \asymp \frac{(p_m + \mathbf{x})^{3/2}}{\sqrt{n}}.$$

The leading terms in the loss decomposition (15.35) are $\|\boldsymbol{\xi}_m\|^2$ and $\mathcal{K}(f, f_m)$, and $\|\boldsymbol{\xi}_m\|^2 \asymp p_m$ for a large sieve dimension m . Therefore, the error term $\Delta_m^{\mathcal{K}}$ is nearly negligible if $\Delta_m^{\mathcal{K}}/p_m \asymp \sqrt{p_m/n}$ is small.

15.4.2 Hellinger loss

Now we consider the Hellinger loss

$$\begin{aligned} \mathcal{H}(f, f_1) &\stackrel{\text{def}}{=} \frac{1}{2} \int \left(\sqrt{f(x)} - \sqrt{f_1(x)} \right)^2 \mu_0(dx) = 1 - \int \sqrt{\frac{f_1(x)}{f(x)}} f(x) \mu_0(dx) \\ &= 1 - \int \exp \left[\frac{1}{2} \{ \log f_1(x) - \log f(x) \} \right] f(x) \mu_0(dx). \end{aligned}$$

Theorem 15.4.2. Suppose that the conditions of Theorem 15.2.1 are fulfilled. Let also the random set $\Omega_m(\mathbf{x})$ be defined there. Then it holds on $\Omega_m(\mathbf{x})$

$$n \mathcal{H}(f, \tilde{f}_m) \leq \frac{1}{8} (\|\boldsymbol{\xi}_m\|^2 + \|\mathbf{b}_m\|^2) + \Delta_m^{\mathcal{H}},$$

where $\Delta_m^{\mathcal{H}}$ fulfills on $\Omega_m(\mathbf{x})$ with \diamond_m from (??) the following upper bound:

$$\Delta_m^{\mathcal{H}} \leq \frac{1}{4} z(p_m, \mathbf{x}) \diamond_m + \frac{1}{8} \diamond_m^2 + \frac{1}{4} \|\mathbf{b}_m\| (1 + \sqrt{2\mathbf{x}} + \diamond_m) + n \rho_0(\mathbf{r}_m). \quad (15.36)$$

Proof. We use the expansion

$$\log \tilde{f}_m(x) - \log f(x) = (\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*) \otimes \{\Psi(x) - \bar{\Psi}\} - \frac{1}{2} \|\mathcal{H}(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2 - \beta_m;$$

cf. (??). The value β_m here follows the bound (??) $|\beta_m| \leq \rho_0(\mathbf{r}_m)$. For the Hellinger distance $\mathcal{H}(f, \tilde{f}_m)$ this yields

$$\begin{aligned} \mathcal{H}(f, \tilde{f}_m) &= 1 - \exp\left(-\frac{1}{4} \|\mathcal{H}(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2 - \frac{\beta_m}{2}\right) \\ &\quad \times \int \exp\left[\frac{1}{2}(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*) \otimes \{\Psi(x) - \bar{\Psi}\}\right] f(x) \mu_0(dx) \end{aligned}$$

and it holds by condition **(x₁)** on the set $\Omega_m(\mathbf{x})$ due to $\tilde{\boldsymbol{\theta}}_m \in \Theta_m(\mathbf{r}_m)$

$$\begin{aligned} \mathcal{H}(f, \tilde{f}_m) &= 1 - \exp\left\{-\frac{1}{8} \|\mathcal{H}(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2 - \frac{\rho_0(\mathbf{r}_m)}{2} - \rho_0\left(\frac{\mathbf{r}_m}{2}\right)\right\} \\ &\leq \frac{1}{8} \|\mathcal{H}(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2 + \rho_0(\mathbf{r}_m). \end{aligned}$$

Here we assumed that $2\rho_0(\mathbf{r}_m/2) \leq \rho_0(\mathbf{r}_m)$. Now we can use the decomposition (??) for $\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*$:

$$n \|\mathcal{H}(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2 = \|\boldsymbol{\xi}_m + \mathbf{b}_m + \boldsymbol{\tau}_m\|^2$$

with $\|\boldsymbol{\tau}_m\| \leq \diamond_m$. It follows on $\Omega_m(\mathbf{x})$

$$\begin{aligned} \left| \|\boldsymbol{\xi}_m + \mathbf{b}_m + \boldsymbol{\tau}_m\|^2 - \|\boldsymbol{\xi}_m + \mathbf{b}_m\|^2 \right| &\leq 2 |(\boldsymbol{\xi}_m + \mathbf{b}_m)^\top \boldsymbol{\tau}_m| + \|\boldsymbol{\tau}_m\|^2 \\ &\leq 2 \{z(p_m, \mathbf{x}) + \|\mathbf{b}_m\|\} \diamond_m + \diamond_m^2, \end{aligned}$$

and

$$\left| \|\boldsymbol{\xi}_m + \mathbf{b}_m\|^2 - \|\boldsymbol{\xi}_m\|^2 - \|\mathbf{b}_m\|^2 \right| = 2 |\boldsymbol{\xi}_m^\top \mathbf{b}_m|.$$

As $\text{Var}(\boldsymbol{\xi}_m) = I_{p_m}$, it follows

$$\text{Var}(\boldsymbol{\xi}_m^\top \mathbf{b}_m) = \|\mathbf{b}_m\|^2$$

and the scalar product $\boldsymbol{\xi}_m^\top \mathbf{b}_m$ fulfills on $\Omega_m(\mathbf{x})$

$$|\boldsymbol{\xi}_m^\top \mathbf{b}_m| \leq \|\mathbf{b}_m\| (1 + \sqrt{2\mathbf{x}}).$$

Summing up all the bounds yields (15.36).

15.5 Penalized smooth density estimation

A roughness penalty approach can be naturally used in the problem of log-density estimation. For a given penalizing matrix G^2 , define the penalized MLE $\tilde{\boldsymbol{\theta}}_G$ and its target $\boldsymbol{\theta}_G^*$ as

$$\begin{aligned}\tilde{\boldsymbol{\theta}}_G &\stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} L_G(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \left\{ L(\boldsymbol{\theta}) - \frac{1}{2} \|G\boldsymbol{\theta}\|^2 \right\}, \\ \boldsymbol{\theta}_G^* &\stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \mathbb{E}L_G(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \left\{ \mathbb{E}L(\boldsymbol{\theta}) - \frac{1}{2} \|G\boldsymbol{\theta}\|^2 \right\},\end{aligned}\quad (15.37)$$

with $L(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top S - ng(\boldsymbol{\theta})$. For studying the properties of $\tilde{\boldsymbol{\theta}}_G$, introduce the penalized Fisher information matrix

$$D_G^2 = D^2 + G^2 = n\nabla^2 g(\boldsymbol{\theta}^*) + G^2.$$

Compared with the non-penalized case, it is increased by the penalty operator G^2 . Simultaneously it leads to an increase of the bias \mathbf{b}_G given by

$$\mathbf{b}_G = D_G(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*),$$

where $\boldsymbol{\theta}^*$ is the true non-penalized point, $\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E}L(\boldsymbol{\theta})$. Another way to define the bias \mathbf{b}_G is used below via the quadratic expansion of the expected log-likelihood. Minimizing over $\boldsymbol{\theta}$ the quadratic expression $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + \|G\boldsymbol{\theta}\|^2$ leads to

$$\begin{aligned}\boldsymbol{\theta}_G^\dagger &\stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \left\{ -\frac{1}{2} \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 - \frac{1}{2} \|G\boldsymbol{\theta}\|^2 \right\} = D_G^{-2} D^2 \boldsymbol{\theta}^*, \\ \mathbf{b}_G &\stackrel{\text{def}}{=} D_G(\boldsymbol{\theta}_G^\dagger - \boldsymbol{\theta}^*) = D_G^{-1} G^2 \boldsymbol{\theta}^*\end{aligned}$$

The main benefit of penalization is that the statistical complexity of the problem is reduced from the original dimension p (usually infinity) to a finite effective dimension p_G . Also penalization helps to relax the condition (χ_1) . Define

$$\mathsf{H}_G^2 \stackrel{\text{def}}{=} n^{-1} D_G^2 = \mathsf{H}^2 + n^{-1} G^2,$$

$$B_G \stackrel{\text{def}}{=} \mathsf{H}_G^{-1} \mathsf{H}^2 \mathsf{H}_G^{-1},$$

and suppose the following condition to be fulfilled.

(χ_G) The vector $\boldsymbol{\chi}_1 \stackrel{\text{def}}{=} \mathsf{H}^{-1} \{\Psi(X_1) - \bar{\Psi}\}$ satisfies

$$\begin{aligned}\left| \log \mathbb{E} \exp \left(\boldsymbol{\alpha}^\top \boldsymbol{\chi}_1 - \frac{\|\boldsymbol{\alpha}\|^2}{2} \right) \right| &\leq \rho_0(\boldsymbol{\alpha}), \\ \sup_{\mathbf{u}^\top B_G^{-1} \mathbf{u} \leq 1} \left| \mathbb{E} \left\{ \mathbf{u}^\top (\boldsymbol{\chi}_1 - \boldsymbol{\alpha}) \exp \left(\boldsymbol{\alpha}^\top \boldsymbol{\chi}_1 - \frac{\|\boldsymbol{\alpha}\|^2}{2} \right) \right\} \right| &\leq \rho_1(\boldsymbol{\alpha}),\end{aligned}$$

and for some $g_1 > 0$ and all $\boldsymbol{\alpha}$ with $\boldsymbol{\alpha}^\top B_G^{-1} \boldsymbol{\alpha} \leq g_1^2$, it holds

$$e^{\rho_0(\boldsymbol{\alpha})} \rho_1(\boldsymbol{\alpha}) \leq \frac{1}{2}.$$

The next result claims that the penalized MLE $\tilde{\boldsymbol{\theta}}_G$ concentrates in a local vicinity $\Theta_{0,G}(\mathbf{r})$ of $\boldsymbol{\theta}^*$

$$\Theta_{0,G}(\mathbf{r}) = \{\boldsymbol{\theta} : \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}\} = \{\boldsymbol{\theta} : \sqrt{n} \|\mathsf{H}_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}\} \quad (15.38)$$

with a proper choice of \mathbf{r} .

Theorem 15.5.1. Consider the penalized MLE $\tilde{\boldsymbol{\theta}}_G$ from (15.37). Let condition **(X_G)** hold, and let the standardized score vector $\boldsymbol{\xi}_G = D_G^{-1}(S - IES)$ satisfy $\|\boldsymbol{\xi}_G\| \leq z(B_G, \mathbf{x})$ on a random set $\Omega_G(\mathbf{x})$ with $\mathbb{P}(\Omega_G(\mathbf{x})) \geq 1 - 2e^{-\mathbf{x}}$. Let also hold for a radius \mathbf{r}_G that $ng_1^2 \leq C\mathbf{r}_G^2$ and

$$\diamondsuit_G(\mathbf{r}_G) \stackrel{\text{def}}{=} \sup_{\boldsymbol{\alpha}^\top B_G^{-1} \boldsymbol{\alpha} \leq \mathbf{r}_G^2/n} \sqrt{n} e^{\rho_0(\boldsymbol{\alpha})} \rho_1(\boldsymbol{\alpha}) \quad (15.39)$$

and let also \mathbf{r}_G and $\mathbf{b}_G = D_G^{-1}G^2\boldsymbol{\theta}^*$ fulfill

$$\mathbf{r}_G \geq 2z(B_G, \mathbf{x}) + 2\|\mathbf{b}_G\| + 2\diamondsuit_G(\mathbf{r}_G).$$

Then $\tilde{\boldsymbol{\theta}}_G$ concentrates in the local vicinity $\Theta_{0,G}(\mathbf{r}_G)$ from (15.38): on $\Omega_G(\mathbf{x})$, it holds

$$\tilde{\boldsymbol{\theta}}_G \in \Theta_{0,G}(\mathbf{r}_G). \quad (15.40)$$

Proof. The key argument for the proof of the concentration result (15.40) is that $L_G(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2/2$ is a concave function of $\boldsymbol{\theta}$ as in the non-penalized case. Another important step in the proof is the following result on quadratic approximation of the penalized log-likelihood.

Lemma 15.5.1. Suppose that **(X_G)** is fulfilled. Then for any \mathbf{r} and any $\boldsymbol{\theta} \in \Theta_{0,G}(\mathbf{r})$, it holds with $\boldsymbol{\alpha} = \mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$

$$\|D_G^{-1}\{\nabla L_G(\boldsymbol{\theta}) - \nabla L_G(\boldsymbol{\theta}^*)\} + D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \sqrt{n} e^{\rho_0(\boldsymbol{\alpha})} \rho_1(\boldsymbol{\alpha}) \leq \diamondsuit_G(\mathbf{r}), \quad (15.41)$$

where $\diamondsuit_G(\mathbf{r})$ is given in (15.39). Moreover,

$$\begin{aligned} & \left| L_G(\boldsymbol{\theta}) - L_G(\boldsymbol{\theta}^*) - (\boldsymbol{\xi}_G + \mathbf{b}_G)^\top D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \frac{\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2}{2} \right| \\ & \leq n\rho_0(\boldsymbol{\alpha}) \leq \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \diamondsuit_G(\mathbf{r}) \leq \mathbf{r} \diamondsuit_G(\mathbf{r}). \end{aligned} \quad (15.42)$$

$$\begin{aligned} & \left| L_G(\boldsymbol{\theta}) - L_G(\boldsymbol{\theta}^*) + \{\nabla L_G(\boldsymbol{\theta})\}^\top D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \frac{\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2}{2} \right| \\ & \leq 2n\rho_0(\boldsymbol{\alpha}) \leq 2\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \diamondsuit_G(\mathbf{r}) \leq 2\mathbf{r} \diamondsuit_G(\mathbf{r}). \end{aligned} \quad (15.43)$$

Proof. By definition

$$\begin{aligned} L_G(\boldsymbol{\theta}) - L_G(\boldsymbol{\theta}^*) &= S^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - ng(\boldsymbol{\theta}) + ng(\boldsymbol{\theta}^*) - \frac{1}{2}(\|G\boldsymbol{\theta}\|^2 - \|G\boldsymbol{\theta}^*\|^2) \\ &= S^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - ng(\boldsymbol{\theta}) + ng(\boldsymbol{\theta}^*) \\ &\quad - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top G^2 \boldsymbol{\theta}^* - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top G^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \end{aligned}$$

and by (15.8) in view of $\mathbb{E}S = n\bar{\Psi}$ for $\boldsymbol{\alpha} = \mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$

$$\begin{aligned} &\left| L_G(\boldsymbol{\theta}) - L_G(\boldsymbol{\theta}^*) - (S - \mathbb{E}S - G^2 \boldsymbol{\theta}^*)^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top(D^2 + G^2)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right| \\ &= n \left| g(\boldsymbol{\theta}) - g(\boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \bar{\Psi} - \frac{1}{2} \|\mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \right| \leq n\rho_0(\boldsymbol{\alpha}) \end{aligned}$$

so that (15.42) follows by $\boldsymbol{\xi}_G = D_G^{-1}(S - \mathbb{E}S)$ and $\mathbf{b}_G = -D_G^{-1}G^2 \boldsymbol{\theta}^*$. Further, in a similar way we derive

$$\nabla L_G(\boldsymbol{\theta}) - \nabla L_G(\boldsymbol{\theta}^*) = -G^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - n\{\nabla g(\boldsymbol{\theta}) - \nabla g(\boldsymbol{\theta}^*)\}$$

and by (15.9)

$$\begin{aligned} &\left\| D_G^{-1} \left\{ \nabla L_G(\boldsymbol{\theta}) - \nabla L_G(\boldsymbol{\theta}^*) + (D^2 + G^2)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right\} \right\| \\ &= \left\| \sqrt{n} \mathsf{H}_G^{-1} \left\{ \nabla g(\boldsymbol{\theta}) - \nabla g(\boldsymbol{\theta}^*) - \sqrt{n} \mathsf{H}^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right\} \right\| \\ &\leq \sqrt{n} e^{\rho_0(\boldsymbol{\alpha})} \rho_1(\boldsymbol{\alpha}), \end{aligned}$$

and (15.41) follows as well.

As

$$\begin{aligned} &\frac{d}{d\boldsymbol{\theta}} \left\{ L_G(\boldsymbol{\theta}) - L_G(\boldsymbol{\theta}^*) - (\boldsymbol{\xi}_G + \mathbf{b}_G)^\top D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \frac{\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2}{2} \right\} \\ &= \nabla L_G(\boldsymbol{\theta}) - \nabla L_G(\boldsymbol{\theta}^*) + D_G^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \end{aligned}$$

the first order Taylor expansion yields

$$\begin{aligned} &\left\{ L_G(\boldsymbol{\theta}) - L_G(\boldsymbol{\theta}^*) - (\boldsymbol{\xi}_G + \mathbf{b}_G)^\top D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \frac{\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2}{2} \right\} \\ &= (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \left\{ \nabla L_G(\boldsymbol{\theta}^\circ) - \nabla L_G(\boldsymbol{\theta}^*) + D_G^2(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*) \right\} \\ &= \{D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}^\top \left[D_G^{-1} \left\{ \nabla L_G(\boldsymbol{\theta}^\circ) - \nabla L_G(\boldsymbol{\theta}^*) + D_G^2(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*) \right\} \right]. \end{aligned}$$

Here $\boldsymbol{\theta}^\circ$ is a point on the line connecting $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$. This and (15.41) imply (15.42) because $\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq r$.

Now we prepared to complete the proof of the theorem. The idea is to show that the value of the penalized log-likelihood $L_G(\boldsymbol{\theta})$ on the boundary of the local set $\Theta_{0,G}(\mathbf{r}_G)$ is strictly less than the value at the point $\boldsymbol{\theta}^*$. Then we can use concavity of $L_G(\boldsymbol{\theta})$ to extend the bound $L_G(\boldsymbol{\theta}) < L_G(\boldsymbol{\theta}_G^\dagger)$ to all $\boldsymbol{\theta}$ outside of this local set.

Let \mathbf{r}_G be fixed and a point $\boldsymbol{\theta}$ be such that $\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathbf{r}_G$. The bound (15.42) of Lemma 15.5.1 implies with $\boldsymbol{\alpha} = \mathsf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ that

$$\begin{aligned} L_G(\boldsymbol{\theta}) - L_G(\boldsymbol{\theta}^*) &\leq (\boldsymbol{\xi}_G + \mathbf{b}_G)^\top D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \frac{\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2}{2} + \mathbf{r}_G \diamondsuit_G(\mathbf{r}_G) \\ &\leq (\boldsymbol{\xi}_G + \mathbf{b}_G)^\top D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \frac{\mathbf{r}_G^2}{2} + \mathbf{r}_G \diamondsuit_G(\mathbf{r}_G). \end{aligned}$$

The use of $\|\boldsymbol{\xi}_G\| \leq z(B_G, \mathbf{x})$ yields

$$(\boldsymbol{\xi}_G + \mathbf{b}_G)^\top D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \leq \{z(B_G, \mathbf{x}) + \|\mathbf{b}_G\|\}\mathbf{r}_G$$

and thus

$$L_G(\boldsymbol{\theta}) - L_G(\boldsymbol{\theta}^*) \leq \{z(B_G, \mathbf{x}) + \|\mathbf{b}_G\|\}\mathbf{r}_G - \frac{\mathbf{r}_G^2}{2} + \mathbf{r}_G \diamondsuit_G(\mathbf{r}_G) < 0 \quad (15.44)$$

under the condition $\mathbf{r}_G > 2\{z(B_G, \mathbf{x}) + \|\mathbf{b}_G\| + \diamondsuit_G(\mathbf{r}_G)\}$.

Further, if $\tilde{\boldsymbol{\theta}}_G$ lies outside of $\Theta_{0,G}(\mathbf{r}_G)$, define $\check{\boldsymbol{\theta}}$ as a point on line connecting $\boldsymbol{\theta}^*$ and $\tilde{\boldsymbol{\theta}}_G$ with $\|D_G(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| = \mathbf{r}_G$. As $L_G(\boldsymbol{\theta})$ is a concave function and $\tilde{\boldsymbol{\theta}}_G$ is its point of maximum it follows for the point $\check{\boldsymbol{\theta}}$ on the line between $\tilde{\boldsymbol{\theta}}_G$ and $\boldsymbol{\theta}^*$ that $L_G(\check{\boldsymbol{\theta}}) - L_G(\boldsymbol{\theta}^*) \geq 0$. However, it contradicts (15.44). This completes the proof.

Similarly to the sieve case, this concentration property helps to state the Fisher and Wilks expansions. The next result uses all notation of Theorem 15.5.1. All its statements follow from the log-likelihood expansion exactly as in the GLM case; cf. Theorem 14.2.2.

Theorem 15.5.2. *Suppose that the conditions of Theorem 15.5.1 are fulfilled. Then on the set $\Omega_G(\mathbf{x})$, it holds for the penalized MLE $\tilde{\boldsymbol{\theta}}_G$ and $\mathbf{b}_G \stackrel{\text{def}}{=} -D_G^{-1}G^2\boldsymbol{\theta}^*$*

$$\|\sqrt{n} \mathsf{H}_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*) - \boldsymbol{\xi}_G - \mathbf{b}_G\| \leq \diamondsuit_G(\mathbf{r}_G), \quad (15.45)$$

with

$$\diamondsuit_G(\mathbf{r}_G) \stackrel{\text{def}}{=} \sup_{\boldsymbol{\alpha}^\top B_G^{-1}\boldsymbol{\alpha} \leq \mathbf{r}_G^2/n} \sqrt{n} e^{\rho_0(\boldsymbol{\alpha})} \rho_1(\boldsymbol{\alpha}).$$

Moreover, the excess $L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}) \stackrel{\text{def}}{=} L_G(\tilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta})$ fulfills

$$\left| L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*) - \frac{1}{2} \|\boldsymbol{\xi}_G + \mathbf{b}_G\|^2 \right| \leq \Delta_G(\mathbf{r}_G), \quad (15.46)$$

$$\left| \sqrt{2L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*)} - \|\boldsymbol{\xi}_G + \mathbf{b}_G\| \right| \leq 3\diamondsuit_G(\mathbf{r}_G), \quad (15.47)$$

where

$$\Delta_G(\mathbf{r}_G) \stackrel{\text{def}}{=} \sup_{\boldsymbol{\alpha}^\top B_G^{-1} \boldsymbol{\alpha} \leq \mathbf{r}_G^2/n} n \left\{ \rho_0(\boldsymbol{\alpha}) + \frac{1}{2} e^{2\rho_0(\boldsymbol{\alpha})} \rho_1^2(\boldsymbol{\alpha}) \right\} \leq \mathbf{r}_G \diamondsuit_G(\mathbf{r}_G) + \frac{1}{2} \diamondsuit_G^2(\mathbf{r}_G).$$

Proof. We present an independent proof of this result. Theorem 15.5.1 enables us to focus on the case when $\|\boldsymbol{\xi}_G\| \leq z(B_G, \mathbf{x})$ and $\tilde{\boldsymbol{\theta}}_G \in \Theta_{0,G}(\mathbf{r}_G)$. We apply (15.41) with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_G$ and use that $\nabla L_G(\tilde{\boldsymbol{\theta}}_G) = 0$ and $\nabla L_G(\boldsymbol{\theta}^*) = \nabla L(\boldsymbol{\theta}^*) - G^2 \boldsymbol{\theta}^*$. This allows to derive for $\boldsymbol{\xi}_G = D_G^{-1} \nabla L(\boldsymbol{\theta}^*)$ and $\boldsymbol{\alpha}_G = \mathsf{H}(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)$

$$\|\boldsymbol{\xi}_G + \mathbf{b}_G - D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\| \leq \sqrt{n} e^{\rho_0(\boldsymbol{\alpha}_G)} \rho_1(\boldsymbol{\alpha}_G) \leq \diamondsuit_G(\mathbf{r}_G) \quad (15.48)$$

which proves (15.45). Now we prove (15.46). Under the condition that $\tilde{\boldsymbol{\theta}}_G \in \Theta_{0,G}(\mathbf{r}_G)$, the bound (15.42) with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_G$ can be rewritten as

$$\left| 2L_G(\tilde{\boldsymbol{\theta}}_G) - 2L_G(\boldsymbol{\theta}^*) - \|\boldsymbol{\xi}_G + \mathbf{b}_G\|^2 + \|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*) - \boldsymbol{\xi}_G - \mathbf{b}_G\|^2 \right| \leq 2\mathbf{r}_G \diamondsuit_G(\mathbf{r}_G).$$

This and (15.48) imply

$$\left| L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*) - \frac{1}{2} \|\boldsymbol{\xi}_G + \mathbf{b}_G\|^2 \right| \leq \mathbf{r}_G \diamondsuit_G(\mathbf{r}_G) + \frac{1}{2} \diamondsuit_G^2(\mathbf{r}_G) \leq \Delta_G(\mathbf{r}_G)$$

proving the bound (15.46). It remains to check the square-root expansion (15.47). The bound (15.43) with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_G$ yields in view of $\nabla L_G(\tilde{\boldsymbol{\theta}}_G) = 0$ and $\tilde{\boldsymbol{\theta}}_G \in \Theta_{0,G}(\mathbf{r}_G)$

$$\left| 2L_G(\tilde{\boldsymbol{\theta}}_G) - 2L_G(\boldsymbol{\theta}^*) - \|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|^2 \right| \leq 2\|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\| \diamondsuit_G(\mathbf{r}_G).$$

This implies

$$\begin{aligned} \left| \sqrt{2L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*)} - \|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\| \right| &\leq \frac{2L_G(\tilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*) - \|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|^2}{\|D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|} \\ &\leq 2\diamondsuit_G(\mathbf{r}_G). \end{aligned}$$

Together with (15.48), this implies (15.47).

15.5.1 Loss in penalized density estimation

The parameter estimator $\tilde{\boldsymbol{\theta}}_G$ yields the log-density estimator

$$\log \tilde{f}_G(x) = \tilde{\boldsymbol{\theta}}_G^\top \Psi(x) - g(\tilde{\boldsymbol{\theta}}_G).$$

Below we discuss the related Kullback-Leibler loss.

Theorem 15.5.3. Suppose that the conditions of Theorem 15.5.1 are fulfilled. Let also the random set $\Omega_G(\mathbf{x})$ be defined there. Then it holds on $\Omega_G(\mathbf{x})$

$$\left| n \int \left\{ \log f(x) - \log \tilde{f}_G(x) \right\} f(x) \mu_0(dx) - \frac{1}{2} (\boldsymbol{\xi}_G + \mathbf{b}_G)^\top B_G (\boldsymbol{\xi}_G + \mathbf{b}_G) \right| \leq \Delta_G^{\mathcal{K}},$$

where $\boldsymbol{\xi}_G = D_G^{-1}(S - IES)$ and $\Delta_G^{\mathcal{K}}$ fulfills on $\Omega_G(\mathbf{x})$

$$\Delta_G^{\mathcal{K}} \leq n \{ \rho_0(\mathbf{r}_G) + \dots \}$$

Proof. The definition yields

$$\begin{aligned} & \int \left\{ \log f(x) - \log \tilde{f}_G(x) \right\} f(x) \mu_0(dx) \\ &= \int \left\{ (\boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}}_G)^\top \Psi(x) \right\} f(x) \mu_0(dx) - \{ g(\boldsymbol{\theta}^*) - g(\tilde{\boldsymbol{\theta}}_G) \} \\ &= (\boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}}_G)^\top \bar{\Psi} - g(\boldsymbol{\theta}^*) + g(\tilde{\boldsymbol{\theta}}_G) \end{aligned}$$

and the bound (15.8) of Lemma 15.1.1 yields under the condition that $\tilde{\boldsymbol{\theta}}_G$ in $\Theta_{0,G}(\mathbf{r}_G)$

$$\left| \int \left\{ \log f(x) - \log \tilde{f}_G(x) \right\} f(x) \mu_0(dx) - \frac{1}{2} \|\mathsf{H}(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|^2 \right| \leq \rho_0(\boldsymbol{\alpha});$$

Now we can complete the proof by using the decomposition (15.48).

15.6 Parametric structural log-density modeling

Suppose that the density function $f(x)$ belongs to a parametric family $f(x, \mathbf{v})$ for a structural parameter $\mathbf{v} \in \Upsilon$. We also assume that each such density is smooth in x and \mathbf{v} , and its logarithm can be represented as

$$\log f(x, \mathbf{v}) = \sum_{j=1}^{\infty} \theta_j(\mathbf{v}) \psi_j(x) - g(\boldsymbol{\theta}(\mathbf{v})) = \boldsymbol{\theta}(\mathbf{v}) \otimes \Psi(x) - g(\boldsymbol{\theta}(\mathbf{v})),$$

where the coefficients $\theta_j(\mathbf{v})$ smoothly depend on \mathbf{v} .

By $\boldsymbol{\theta}^*$ we denote the true density. The case of correct specification means that $\theta_j(\mathbf{v}^*) = \theta_j^*$ for all j . Smoothness of each function $\log f(x, \mathbf{v})$ is understood in the sense that the use of sieve truncation at the index m or penalization $\|G\boldsymbol{\theta}\|^2/2$ does not introduce a significant bias. To be more specific, we consider the sieve approach and measure the bias by the norm of \mathbf{b}_m which is assumed to be not large. The roughness penalty approach can be treated as well.

We consider the maximum likelihood parametric estimator $\tilde{\mathbf{v}}$ and the sieve nonparametric sieve estimator $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}_m$:

$$\tilde{\boldsymbol{v}} = \underset{\boldsymbol{v} \in \Upsilon}{\operatorname{argmax}} \mathcal{L}(\boldsymbol{v}),$$

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta_m}{\operatorname{argmax}} L(\boldsymbol{\theta}),$$

where

$$\mathcal{L}(\boldsymbol{v}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}(\boldsymbol{v})) = \boldsymbol{\theta}(\boldsymbol{v}) \otimes S - ng(\boldsymbol{\theta}(\boldsymbol{v})) \quad (15.49)$$

for $S = \sum_{i=1}^n \Psi(X_i)$.

Under natural conditions, the sieve estimator $\tilde{\boldsymbol{\theta}}$ concentrates in a local vicinity $\Theta_0(\mathbf{r}_0)$ of $\boldsymbol{\theta}^*$. Moreover, on the set $\Omega(\mathbf{x})$, the value of the log-likelihood $L(\boldsymbol{\theta})$ outside of the elliptic set $\Theta_0(\mathbf{r}_0)$ is strictly smaller than the value $L(\boldsymbol{\theta}^*)$ at the true point. This, of course, yields that the value $\boldsymbol{\theta}(\tilde{\boldsymbol{v}})$ for the parameter estimate $\tilde{\boldsymbol{v}}$ should be inside of $\Theta_0(\mathbf{r}_0)$. As a corollary of this simple observation, we obtain the following result.

Theorem 15.6.1. *Let the family $\boldsymbol{\theta}(\boldsymbol{v})$ be smooth enough to ensure a small bias*

$$\|\mathbf{b}_m\| = \sup_{\boldsymbol{v}} \|D\{\boldsymbol{\theta}(\boldsymbol{v}) - \boldsymbol{\theta}_m(\boldsymbol{v})\}\|.$$

Under conditions of Theorem 15.2.1, it holds on the set $\Omega(\mathbf{x})$

$$\boldsymbol{\theta}(\tilde{\boldsymbol{v}}) \in \Theta_m(\mathbf{r}_0).$$

Remark 15.6.1. The way of establishing the parametric result of Theorem 15.6.1 is known as “relaxation”: we replace the nonconvex parametric constraint $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{v})$ by a more flexible quadratic constraint that $\boldsymbol{\theta}$ is a smooth vector. In this larger class of possible parameters we can apply the concentration result of Theorem 15.5.1. Note that the radius \mathbf{r}_0 corresponds to the nonparametric accuracy described via the quantity $z(B_m, \mathbf{x}) + \|\mathbf{b}\|$. However, if the parametric family $\boldsymbol{\theta}(\boldsymbol{v})$ is sufficiently smooth then there is no significant loss of accuracy by this relaxation.

The result of Theorem 15.6.1 suggests to introduce the set

$$\Upsilon_o(\mathbf{r}_0) \stackrel{\text{def}}{=} \{\boldsymbol{v}: \boldsymbol{\theta}(\boldsymbol{v}) \in \Theta_0(\mathbf{r}_0)\}.$$

This set can be viewed as smooth parametric vicinity of \boldsymbol{v}^* . We already know that $\tilde{\boldsymbol{v}}$ well concentrates on $\Upsilon_o(\mathbf{r}_0)$. In this vicinity we aim at approximating the underlying parametric model by its linearization. This will allow us to establish Fisher and Wilks expansions in terms of \boldsymbol{v} .

Note that the log-likelihood $\mathcal{L}(\boldsymbol{v})$ in (15.49) is obtained as superposition of two mappings: $\boldsymbol{v} \rightarrow \boldsymbol{\theta}(\boldsymbol{v})$ and $\mathcal{L}(\boldsymbol{v}) \rightarrow L(\boldsymbol{\theta}(\boldsymbol{v}))$. A quadratic approximation of $\mathcal{L}(\boldsymbol{v})$ in the

vicinity $\Upsilon_o(\mathbf{r}_0)$ requires a linearization of the first mapping and a quadratic approximation of the second one. The overall error of approximation is, of course, a superposition of these two sources of errors. Define

$$J_{\mathbf{v}} \stackrel{\text{def}}{=} \nabla_{\mathbf{v}} \boldsymbol{\theta}(\mathbf{v}).$$

For each $\mathbf{v} \in \Upsilon$, this is a linear operator from $\boldsymbol{\theta}$ -space to \mathbf{v} -space, and we implicitly assume that this operator is bounded uniformly in \mathbf{v} . A proper parametrization and identifiability in \mathbf{v} means that each $J_{\mathbf{v}}$ is injective. However, we allow below that the identifiability is violated and $J_{\mathbf{v}}$ can be degenerated. We will assume that the error of linear approximation $\boldsymbol{\theta}(\mathbf{v}) - \boldsymbol{\theta}(\mathbf{v}^*) \approx J_{\mathbf{v}^*}^\top (\mathbf{v} - \mathbf{v}^*)$ is sufficiently small. This error will be expressed as the vector

$$\boldsymbol{\beta}_1(\mathbf{v}) \stackrel{\text{def}}{=} D\{\boldsymbol{\theta}(\mathbf{v}) - \boldsymbol{\theta}^* - J_{\mathbf{v}^*}^\top (\mathbf{v} - \mathbf{v}^*)\}. \quad (15.50)$$

A uniform bound over $\Upsilon_o(\mathbf{r}_0)$ means that the value

$$\beta_1^*(\mathbf{r}_0) \stackrel{\text{def}}{=} \|\boldsymbol{\beta}_1(\mathbf{v})\|.$$

is sufficiently small. The error of linear approximation of the gradient of $L(\boldsymbol{\theta})$ is measured via the vector

$$\boldsymbol{\beta}_0(\boldsymbol{\theta}) \stackrel{\text{def}}{=} D^{-1}\{\nabla A(\boldsymbol{\theta}) - \nabla A(\boldsymbol{\theta}^*)\} - D\{\boldsymbol{\theta} - \boldsymbol{\theta}^*\}.$$

Lemma 15.1.1 implies that

$$\sup_{\boldsymbol{\theta} \in \Theta_o(\mathbf{r}_0)} \|\boldsymbol{\beta}_0(\boldsymbol{\theta})\| \leq \diamond(\mathbf{r}_0).$$

The parametric Fisher matrix Γ^2 can be defined as

$$\Gamma^2 \stackrel{\text{def}}{=} J_{\mathbf{v}^*} D^2 J_{\mathbf{v}^*}^\top.$$

The parametric score vector $\boldsymbol{\chi}$ reads

$$\boldsymbol{\chi} \stackrel{\text{def}}{=} \nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}^*) = J_{\mathbf{v}^*}(S - \mathbb{E} S).$$

We aim to show that $\tilde{\mathbf{v}} - \mathbf{v}^* \approx \Gamma^{-2} \boldsymbol{\chi}$ and $\mathcal{L}(\tilde{\mathbf{v}}) - \mathcal{L}(\mathbf{v}^*) \approx \frac{1}{2} \boldsymbol{\chi}^\top \Gamma^{-2} \boldsymbol{\chi}$ and evaluate the approximation errors.

Theorem 15.6.2. Suppose ... Then

$$\|\Gamma^2(\tilde{\mathbf{v}} - \mathbf{v}^*) - \boldsymbol{\chi}\| \leq \diamond(\mathbf{r}_0)$$

with

$$\diamond(\mathbf{r}_0) \stackrel{\text{def}}{=} \sup_{\mathbf{v} \in \Upsilon_o(\mathbf{r}_0)} \|\Pi_{\mathbf{v}^*} \boldsymbol{\delta}(\mathbf{v})\|.$$

Moreover,

$$\left| \mathcal{L}(\tilde{\mathbf{v}}) - \mathcal{L}(\mathbf{v}^*) - \frac{1}{2} \boldsymbol{\chi}^\top \boldsymbol{\Gamma}^{-2} \boldsymbol{\chi} \right| \leq \Delta(\mathbf{r}_0) \quad (15.51)$$

with

$$\Delta(\mathbf{r}_0) \stackrel{\text{def}}{=} n\rho_0(\mathbf{r}_0) + 2\mathbf{r}_0 \beta_1^*(\mathbf{r}_0) + \frac{1}{2} |\beta_1^*(\mathbf{r}_0)|^2 + \frac{1}{2} \Delta_0(\mathbf{r}_0).$$

Proof. The gradient of $\mathcal{L}(\mathbf{v})$ reads

$$\nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}) = \nabla_{\mathbf{v}} \boldsymbol{\theta}(\mathbf{v}) \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}(\mathbf{v})) = J_{\mathbf{v}} \{S - n \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}(\mathbf{v}))\}.$$

The main condition on the parametric family $\boldsymbol{\theta}(\mathbf{v})$ is that the gradient operator $J_{\mathbf{v}}$ does not vary much on $\Upsilon_o(\mathbf{r}_0)$. This would allow to replace $J_{\mathbf{v}}$ with $J_{\mathbf{v}^*}$ and to approximate linearly the gradient $\nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v})$. Below $J_{\mathbf{v}}^\top$ means the transpose (the dual) of $J_{\mathbf{v}}$. We aim at evaluating the quality of linear approximation

$$\nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}) - \nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}^*) \approx J_{\mathbf{v}^*} D^2 J_{\mathbf{v}^*}^\top (\mathbf{v} - \mathbf{v}^*) = \boldsymbol{\Gamma}^2 (\mathbf{v} - \mathbf{v}^*)$$

with

$$\nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}^*) = J_{\mathbf{v}^*} \nabla L(\boldsymbol{\theta}^*) = J_{\mathbf{v}^*} \{ \nabla L(\boldsymbol{\theta}^*) - \mathbb{E} \nabla L(\boldsymbol{\theta}^*) \} = J_{\mathbf{v}^*} (S - \mathbb{E} S).$$

More precisely, we aim at bounding the approximation error

$$\diamond(\mathbf{v}) \stackrel{\text{def}}{=} \nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}) - \nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}^*) - \boldsymbol{\Gamma}^2 (\mathbf{v} - \mathbf{v}^*).$$

The definition $\mathcal{L}(\mathbf{v}) = S^\top \boldsymbol{\theta}(\mathbf{v}) - A(\boldsymbol{\theta}(\mathbf{v}))$ implies

$$\nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}) - \nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}^*) = (J_{\mathbf{v}} - J_{\mathbf{v}^*}) S - \{ J_{\mathbf{v}} \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}(\mathbf{v})) - J_{\mathbf{v}^*} \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}(\mathbf{v}^*)) \}.$$

The use of $\nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}(\mathbf{v}^*)) = \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}^*) = \mathbb{E} S$ and of $\boldsymbol{\theta}^* = \boldsymbol{\theta}(\mathbf{v}^*)$ yields

$$\begin{aligned} & \nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}) - \nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}^*) \\ &= (J_{\mathbf{v}} - J_{\mathbf{v}^*})(S - \mathbb{E} S) - J_{\mathbf{v}} \{ \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}(\mathbf{v})) - \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}(\mathbf{v}^*)) \}. \end{aligned} \quad (15.52)$$

Now we use the linear approximation of $\nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$ from Lemma 15.1.1 as well as the linear approximation $\boldsymbol{\theta}(\mathbf{v}) - \boldsymbol{\theta}(\mathbf{v}^*) \approx J_{\mathbf{v}^*}^\top (\mathbf{v} - \mathbf{v}^*)$. Define the approximation errors

$$\boldsymbol{\beta}(\mathbf{v}) \stackrel{\text{def}}{=} \boldsymbol{\beta}_0(\boldsymbol{\theta}(\mathbf{v})) + \boldsymbol{\beta}_1(\mathbf{v}).$$

Then (15.52) implies in view of $\boldsymbol{\theta}^* = \boldsymbol{\theta}(\mathbf{v}^*)$

$$\nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}) - \nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}^*) + J_{\mathbf{v}^*} D^2 J_{\mathbf{v}^*}^\top (\mathbf{v} - \mathbf{v}^*) = \boldsymbol{\delta}(\mathbf{v}) \quad (15.53)$$

with

$$\begin{aligned} \boldsymbol{\delta}(\mathbf{v}) &= (J_{\mathbf{v}} - J_{\mathbf{v}^*})(S - \mathbb{E} S) + J_{\mathbf{v}} D \boldsymbol{\beta}(\mathbf{v}) \\ &= (J_{\mathbf{v}} - J_{\mathbf{v}^*}) D \{ \boldsymbol{\xi} + \boldsymbol{\beta}(\mathbf{v}) \} + J_{\mathbf{v}^*} D \boldsymbol{\beta}(\mathbf{v}). \end{aligned}$$

The concentration result allows to restrict ourselves to the case $\tilde{\mathbf{v}} \in \Upsilon_o(\mathbf{r}_0)$. The use of $\nabla_{\mathbf{v}} \mathcal{L}(\tilde{\boldsymbol{\theta}}) = 0$ in (15.53) with $\mathbf{v} = \tilde{\mathbf{v}}$ leads to the relation

$$\Gamma^2(\tilde{\mathbf{v}} - \mathbf{v}^*) - \nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}^*) = \Gamma^2(\tilde{\mathbf{v}} - \mathbf{v}^*) - \boldsymbol{\chi} = \boldsymbol{\delta}(\tilde{\mathbf{v}}).$$

If $\Pi_{\mathbf{v}^*} = J_{\mathbf{v}^*}^\top (J_{\mathbf{v}^*} J_{\mathbf{v}^*}^\top)^{-1} J_{\mathbf{v}^*}$ denotes the projector in \mathbf{v} -space corresponding to the linear operator $J_{\mathbf{v}^*}$, then $\Pi_{\mathbf{v}^*} \{ \Gamma^2(\tilde{\mathbf{v}} - \mathbf{v}^*) - \boldsymbol{\chi} \} = \Gamma^2(\tilde{\mathbf{v}} - \mathbf{v}^*) - \boldsymbol{\chi}$ and hence,

$$\| \Gamma^2(\tilde{\mathbf{v}} - \mathbf{v}^*) - \boldsymbol{\chi} \| \leq \sup_{\mathbf{v} \in \Upsilon_o(\mathbf{r}_0)} \| \Pi_{\mathbf{v}^*} \boldsymbol{\delta}(\mathbf{v}) \|.$$

This implies on $\Omega(x)$

$$\{ \Gamma^2(\tilde{\mathbf{v}} - \mathbf{v}^*) - \boldsymbol{\chi} \}^\top \Gamma^{-2} \{ \Gamma^2(\tilde{\mathbf{v}} - \mathbf{v}^*) - \boldsymbol{\chi} \} \leq \Delta_0(\mathbf{r}_0) \quad (15.54)$$

with

$$\Delta_0(\mathbf{r}_0) \stackrel{\text{def}}{=} \sup_{\mathbf{v} \in \Upsilon_o(\mathbf{r}_0)} \{ \Pi_{\mathbf{v}^*} \boldsymbol{\delta}(\mathbf{v}) \}^\top \Gamma^{-2} \Pi_{\mathbf{v}^*} \boldsymbol{\delta}(\mathbf{v}).$$

Next we consider a quadratic expansion for the log-likelihood $\mathcal{L}(\mathbf{v})$. Fix any two points \mathbf{v} and \mathbf{v}° in $\Upsilon_o(\mathbf{r}_0)$. Denote $\boldsymbol{\theta}(\mathbf{v}, \mathbf{v}^\circ) = \boldsymbol{\theta}(\mathbf{v}) - \boldsymbol{\theta}(\mathbf{v}^\circ)$. The bound (15.8) of Lemma 15.1.1 implies

$$\left| L(\boldsymbol{\theta}(\mathbf{v})) - L(\boldsymbol{\theta}(\mathbf{v}^*)) - \boldsymbol{\xi}^\top D\boldsymbol{\theta}(\mathbf{v}, \mathbf{v}^\circ) + \frac{1}{2} \| D\boldsymbol{\theta}(\mathbf{v}, \mathbf{v}^\circ) \|^2 \right| \leq n\rho_0(\mathbf{r}_0). \quad (15.55)$$

Further, the expansion (15.50) yields

$$\begin{aligned} &\boldsymbol{\xi}^\top D\boldsymbol{\theta}(\mathbf{v}, \mathbf{v}^\circ) - \frac{1}{2} \| D\boldsymbol{\theta}(\mathbf{v}, \mathbf{v}^\circ) \|^2 \\ &= \boldsymbol{\xi}^\top \{ D J_{\mathbf{v}^*}^\top (\mathbf{v} - \mathbf{v}^*) + \boldsymbol{\beta}_1(\mathbf{v}) \} - \frac{1}{2} \| D \{ J_{\mathbf{v}^*}^\top (\mathbf{v} - \mathbf{v}^*) \} + \boldsymbol{\beta}_1(\mathbf{v}) \|^2 \\ &= \boldsymbol{\xi}^\top D J_{\mathbf{v}^*}^\top (\mathbf{v} - \mathbf{v}^*) - \frac{1}{2} \| D J_{\mathbf{v}^*}^\top (\mathbf{v} - \mathbf{v}^*) \|^2 \\ &\quad + \{ \boldsymbol{\xi} - D J_{\mathbf{v}^*}^\top (\mathbf{v} - \mathbf{v}^*) \}^\top \boldsymbol{\beta}_1(\mathbf{v}) - \frac{1}{2} \| \boldsymbol{\beta}_1(\mathbf{v}) \|^2. \end{aligned} \quad (15.56)$$

Definition of the set $\Upsilon_o(\mathbf{r}_0)$ implies $\|D\mathcal{J}_{\mathbf{v}^*}^\top(\mathbf{v} - \mathbf{v}^*)\| \leq \mathbf{r}_0$, and it holds $\|\boldsymbol{\xi}\| \leq z(B, \mathbf{x})$ on $\Omega(\mathbf{x})$. It also holds

$$\begin{aligned} & \boldsymbol{\xi}^\top D\mathcal{J}_{\mathbf{v}^*}^\top(\mathbf{v} - \mathbf{v}^*) - \frac{1}{2}\|D\mathcal{J}_{\mathbf{v}^*}^\top(\mathbf{v} - \mathbf{v}^*)\|^2 \\ &= \boldsymbol{\chi}^\top(\mathbf{v} - \mathbf{v}^*) - \frac{1}{2}(\mathbf{v} - \mathbf{v}^*)^\top \Gamma^2(\mathbf{v} - \mathbf{v}^*) \\ &= \frac{1}{2}\boldsymbol{\chi}^\top \Gamma^{-2}\boldsymbol{\chi} - \frac{1}{2}(\mathbf{v} - \mathbf{v}^* - \Gamma^{-2}\boldsymbol{\chi})^\top \Gamma^2(\mathbf{v} - \mathbf{v}^* - \Gamma^{-2}\boldsymbol{\chi}) \\ &= \frac{1}{2}\boldsymbol{\chi}^\top \Gamma^{-2}\boldsymbol{\chi} - \frac{1}{2}\{\Gamma^2(\mathbf{v} - \mathbf{v}^*) - \boldsymbol{\chi}\}^\top \Gamma^{-2}\{\Gamma^2(\mathbf{v} - \mathbf{v}^*) - \boldsymbol{\chi}\}. \end{aligned}$$

By (15.55) and (15.56)

$$\begin{aligned} & |L(\boldsymbol{\theta}(\mathbf{v})) - L(\boldsymbol{\theta}(\mathbf{v}^*)) - \frac{1}{2}\boldsymbol{\chi}^\top \Gamma^{-2}\boldsymbol{\chi}| \\ &\leq n\rho_0(\mathbf{r}_0) + 2\mathbf{r}_0\|\boldsymbol{\beta}_1(\mathbf{v})\| + \frac{1}{2}\|\boldsymbol{\beta}_1(\mathbf{v})\|^2 \\ &\quad + \frac{1}{2}\{\Gamma^2(\mathbf{v} - \mathbf{v}^*) - \boldsymbol{\chi}\}^\top \Gamma^{-2}\{\Gamma^2(\mathbf{v} - \mathbf{v}^*) - \boldsymbol{\chi}\}. \end{aligned}$$

Now we use that $\tilde{\mathbf{v}} \in \Upsilon_o(\mathbf{r}_0)$ on $\Omega(\mathbf{x})$, and plug $\tilde{\mathbf{v}}$ in place of \mathbf{v} . Together with (15.54), it yields on $\Omega(\mathbf{x})$

$$|\mathcal{L}(\tilde{\mathbf{v}}) - \mathcal{L}(\mathbf{v}^*) - \frac{1}{2}\boldsymbol{\chi}^\top \Gamma^{-2}\boldsymbol{\chi}| \leq n\rho_0(\mathbf{r}_0) + 2\mathbf{r}_0\beta_1^*(\mathbf{r}_0) + \frac{1}{2}|\beta_1^*(\mathbf{r}_0)|^2 + \frac{1}{2}\Delta_0(\mathbf{r}_0),$$

and (15.51) follows.

Sieve parametric approach in nonparametric regression

This chapter focuses on the problem of nonparametric estimation in the regression model

$$Y_i = f(X_i) + \varepsilon_i$$

under the assumption that the regression function f can be well approximated by a sieve parametric family. Most of results are presented for a fixed deterministic design. However, one can easily extend them to the case of a random design under mild conditions; see Section 16.8 below.

16.1 Parametric and nonparametric regression

Below we discuss a special class of parametric models for which the parameter \boldsymbol{v} only enters in the (mean) response of the observed data. Consider a sample $\mathbf{Y} = (Y_1, \dots, Y_n)$ with independent observations Y_i . We also denote $\mathbf{f}^* \stackrel{\text{def}}{=} \mathbb{E}\mathbf{Y}$ and $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbb{E}\mathbf{Y}$. In the regression setup, each f_i is the value of the regression function $f(\cdot)$ of the d -dimensional regressor $X_i \in \mathbb{R}^d$.

The parametric approach considered below assumes that the unknown regression function is specified by the parameter $\boldsymbol{v} \in \Upsilon$, that is, $f(X_i) = f(X_i, \boldsymbol{v})$ for all $i \leq n$, where Υ is a subset of the Euclidean space \mathbb{R}^p . In this section we assume that p is finite. However, the sieve approach of Section 16.5 extends to the case of infinite dimensional parameter \boldsymbol{v} . We also write $f(X_i, \boldsymbol{v})$ in the form $f_i(\boldsymbol{v})$ to highlight dependence on the parameter \boldsymbol{v} . The parametric model can be written as

$$Y_i = f_i(\boldsymbol{v}) + \varepsilon_i, \quad i = 1, \dots, n, \tag{16.1}$$

with $\boldsymbol{v} \in \Upsilon \subseteq \mathbb{R}^p$. We focus on the least squares approach which can be viewed as (quasi) maximum likelihood estimation when the errors ε_i are independent Gaussian $\mathcal{N}(0, \sigma^2)$. The corresponding objective function reads

$$L(\boldsymbol{v}) = -\frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbf{f}(\boldsymbol{v})\|^2 + R.$$

The remainder R does not depend on \boldsymbol{v} . Expectation of $L(\boldsymbol{v})$ w.r.t. underlying data distribution satisfies

$$\mathbb{E}L(\boldsymbol{v}) = -\frac{1}{2\sigma^2} \|\mathbf{f}^* - \mathbf{f}(\boldsymbol{v})\|^2 + R_1.$$

In this model, the MLE $\tilde{\boldsymbol{v}}$ minimizes the data fit:

$$\tilde{\boldsymbol{v}} \stackrel{\text{def}}{=} \operatorname{argmin}_{\boldsymbol{v}} \|\mathbf{Y} - \mathbf{f}(\boldsymbol{v})\|^2 \quad (16.2)$$

while the target value \boldsymbol{v}^* minimizes the approximation error:

$$\boldsymbol{v}^* = \operatorname{argmin}_{\boldsymbol{v} \in \Upsilon} \|\mathbf{f}^* - \mathbf{f}(\boldsymbol{v})\|^2.$$

In analog with the case of linear regression, introduce the $p \times p$ matrix D^2 by

$$D^2 \stackrel{\text{def}}{=} \sigma^{-2} \sum_{i=1}^n \nabla f_i(\boldsymbol{v}^*) \{ \nabla f_i(\boldsymbol{v}^*) \}^\top, \quad (16.3)$$

and define for each \mathbf{r} a local vicinity $\Upsilon_\circ(\mathbf{r})$ of \boldsymbol{v}^* by

$$\Upsilon_\circ(\mathbf{r}) \stackrel{\text{def}}{=} \{ \boldsymbol{v} \in \Upsilon : \|D(\boldsymbol{v} - \boldsymbol{v}^*)\| \leq \mathbf{r} \}.$$

Remark 16.1.1. It is worth mentioning that \boldsymbol{v}^* can be replaced with any point \boldsymbol{v}° from the same local vicinity $\Upsilon_\circ(\mathbf{r}_0)$. This only affects the bias term $\|\mathbf{b}\|^2 = \|\mathbf{f}^* - \mathbf{f}(\boldsymbol{v}^\circ)\|^2$. In particular, keeping in mind an extension to the sieve nonparametric estimation, we also consider the case when \boldsymbol{v}^* is defined by maximization of the expected log-likelihood over a larger space Υ^* :

$$\boldsymbol{v}^* = \operatorname{argmin}_{\boldsymbol{v} \in \Upsilon^*} \|\mathbf{f}^* - \mathbf{f}(\boldsymbol{v})\|^2.$$

Below we study the properties of $\tilde{\boldsymbol{v}}$ including a large deviation bound, Fisher and Wilks expansions, and risk decomposition.

16.2 Conditions

This section specifies sufficient conditions on the regression function $f(\cdot)$ and on the errors ε_i which enable us to state the results on the properties of the LSE $\tilde{\boldsymbol{v}}$.

In words, we assume that the approximation $\mathbf{f}^* \stackrel{\text{def}}{=} \mathbb{E}\mathbf{Y} \approx \mathbf{f}(\boldsymbol{v}^*)$ is reasonable in the sense that the bias $\mathbf{b} = \mathbf{f}(\boldsymbol{v}^*) - \mathbf{f}^*$ is sufficiently small in the sense that the value

$$\|\mathbf{b}\|^2 = \|\mathbf{f}^* - \mathbf{f}(\mathbf{v}^*)\|^2 = \sum_{i=1}^n |f_i - f_i(\mathbf{v}^*)|^2$$

is small. Also we assume everywhere that each entry $f_i(\mathbf{v})$ of $\mathbf{f}(\mathbf{v})$ is a smooth (two times differentiable) function of the parameter \mathbf{v} . Finally we assume that the errors ε_i are zero mean Gaussian with inhomogeneous variance σ_i^2 satisfying some regularity conditions.

Define a symmetric $p \times p$ matrix H^2 by

$$\mathsf{H}^2 \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \nabla f_i(\mathbf{v}^*) \{ \nabla f_i(\mathbf{v}^*) \}^\top. \quad (16.4)$$

For a linear regression model, the second derivative $\nabla^2 f_i(\mathbf{v})$ vanishes. We show below that majority of estimation results can be extended to the case of non-linear functions $f_i(\mathbf{v})$ under regularity conditions on the second derivative $\nabla^2 f_i$. Namely, we assume the following conditions to be fulfilled for each \mathbf{r} .

(Df₂) *The functions $f_i(\mathbf{v})$ are two times differentiable in $\mathbf{v} \in \Upsilon_\circ(\mathbf{r})$, and for some constant w_2 ,*

$$\sup_{\mathbf{v} \in \Upsilon_\circ(\mathbf{r})} \sup_{\mathbf{u} \in \mathcal{S}_p} \frac{1}{n} \sum_{i=1}^n [\mathbf{u}^\top \mathsf{H}^{-1} \nabla^2 f_i(\mathbf{v}) \mathsf{H}^{-1} \mathbf{u}]^2 \leq w_2^2.$$

By the Cauchy-Schwarz inequality this implies for any $\mathbf{v} \in \Upsilon_\circ(\mathbf{r})$, and any unit vectors $\mathbf{u}, \mathbf{u}_1 \in \mathbb{R}^p$

$$\frac{1}{n} \sum_{i=1}^n [\mathbf{u}^\top \mathsf{H}^{-1} \nabla^2 f_i(\mathbf{v}) \mathsf{H}^{-1} \mathbf{u}_1]^2 \leq w_2^2. \quad (16.5)$$

In the sieve nonparametric setup, the value w_2 in (16.5) may depend on the parameter dimension p : $w_2 = w_2(p)$. To simplify our notation, we do not indicate this dependence explicitly.

Regrading the errors $\varepsilon_i = Y_i - \mathbb{E}Y_i$, we assume the following condition.

(σ_{1|n}²) *The errors $\varepsilon_i = Y_i - \mathbb{E}Y_i$ are independent Gaussian, $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, and for some constant $C_V \geq 1$*

$$\min_{i=1,\dots,n} \sigma_i^2 \geq C_V^{-2} \sigma^2, \quad \max_{i=1,\dots,n} \sigma_i^2 \leq C_V^2 \sigma^2.$$

Below we show how these conditions yield the general conditions **(ED₀)**, **(ED₂)**, and **(L₀)**.

16.2.1 Checking the local conditions (ED_0) and (ED_2)

This section discusses the conditions (ED_0) and (ED_2) on the stochastic component of the log-likelihood. For simplicity we assume that the errors ε_i are Gaussian zero mean but possibly inhomogeneous: $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$.

The stochastic component $\zeta(\mathbf{v}) = L(\mathbf{v}) - \mathbb{E}L(\mathbf{v})$ and its derivatives read as

$$\begin{aligned}\zeta(\mathbf{v}) &= \sigma^{-2} \sum_{i=1}^n \varepsilon_i f_i(\mathbf{v}) + R, \\ \nabla \zeta(\mathbf{v}) &= \sigma^{-2} \sum_{i=1}^n \varepsilon_i \nabla f_i(\mathbf{v}), \\ \nabla^2 \zeta(\mathbf{v}) &= \sigma^{-2} \sum_{i=1}^n \varepsilon_i \nabla^2 f_i(\mathbf{v}).\end{aligned}$$

Here R is a remainder term does not depending on \mathbf{v} . Under $\text{Var}(\varepsilon_i) = \sigma_i^2$, the covariance matrix $V^2 = \text{Var}\{\nabla \zeta(\mathbf{v}^*)\}$ is equal to

$$V^2 \stackrel{\text{def}}{=} \text{Var}\{\nabla \zeta(\mathbf{v}^*)\} = \sigma^{-4} \sum_{i=1}^n \sigma_i^2 \nabla f_i(\mathbf{v}^*) \{\nabla f_i(\mathbf{v}^*)\}^\top. \quad (16.6)$$

Moreover, if the errors ε_i are Gaussian then $\nabla \zeta(\mathbf{v})$ is Gaussian as well and condition (ED_0) is fulfilled automatically with the matrix V^2 . Further, for any vector $\gamma \in \mathbb{R}^p$, the sum $\gamma^\top \nabla^2 \zeta(\mathbf{v}) \gamma$ is Gaussian with zero mean and the variance

$$\text{Var}\{\gamma^\top \nabla^2 \zeta(\mathbf{v}) \gamma\} = \sigma^{-4} \sum_{i=1}^n \sigma_i^2 \{\gamma^\top \nabla^2 f_i(\mathbf{v}) \gamma\}^2.$$

This implies (ED_2) with ω given by

$$\begin{aligned}\omega^2 &= \sup_{\gamma \in \mathbb{R}^p} \frac{1}{\sigma^4 \|D\gamma\|^4} \sum_{i=1}^n \sigma_i^2 |\gamma^\top \nabla^2 f_i(\mathbf{v}) \gamma|^2 \\ &= \sup_{\mathbf{u} \in \mathbb{R}^p} \frac{1}{\sigma^4 \|\mathbf{u}\|^4} \sum_{i=1}^n \sigma_i^2 |\mathbf{u}^\top D^{-1} \nabla^2 f_i(\mathbf{v}) D^{-1} \mathbf{u}|^2 \\ &= \sup_{\mathbf{u} \in \mathbb{S}_p} \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2 |\mathbf{u}^\top H^{-1} \nabla^2 f_i(\mathbf{v}) H^{-1} \mathbf{u}|^2.\end{aligned}$$

The expression for ω can be simplified under noise regularity condition $(\sigma_{1|n}^2)$.

Lemma 16.2.1. *For the parametric regression model (16.1), suppose that the errors ε_i are Gaussian and satisfy $(\sigma_{1|n}^2)$. Then condition (ED_0) is fulfilled with the matrix V^2 from (16.6). Let also the regression function $f(\mathbf{v})$ fulfills (Df_2) . Then the matrices V^2 and D^2 from (16.3) are related as*

$$C_V^{-2} D^2 \leq V^2 \leq C_V^2 D^2$$

and **(ED₂)** is fulfilled with the value ω satisfying

$$\omega = C_V \frac{\sigma w_2}{\sqrt{n}}. \quad (16.7)$$

16.2.2 Checking the local condition **(L₀)**

Now we check **(L₀)** and establish the upper bound on the accuracy of quadratic approximation of the expected log-likelihood $EEL(\mathbf{v})$. The derivatives of $-EEL(\mathbf{v})$ read

$$\begin{aligned} -\nabla EEL(\mathbf{v}) &= \sigma^{-2} \sum_{i=1}^n \{f_i(\mathbf{v}) - f_i\} \nabla f_i(\mathbf{v}), \\ \mathbb{F}(\mathbf{v}) &\stackrel{\text{def}}{=} -\nabla^2 EEL(\mathbf{v}) \\ &= \sigma^{-2} \sum_{i=1}^n \{f_i(\mathbf{v}) - f_i\} \nabla^2 f_i(\mathbf{v}) + \sigma^{-2} \sum_{i=1}^n \nabla f_i(\mathbf{v}) \{\nabla f_i(\mathbf{v})\}^\top. \end{aligned} \quad (16.8)$$

Let D^2 be given by (16.3) and H^2 by (16.4). Given $r > 0$, consider the local vicinity $\Upsilon_o(r)$ of \mathbf{v}^* of the form

$$\Upsilon_o(r) \stackrel{\text{def}}{=} \{\mathbf{v} \in \Upsilon : \|D(\mathbf{v} - \mathbf{v}^*)\| \leq r\} = \left\{ \mathbf{v} \in \Upsilon : \|H(\mathbf{v} - \mathbf{v}^*)\| \leq \frac{\sigma r}{\sqrt{n}} \right\}; \quad (16.9)$$

Proposition 16.2.1. Suppose that condition **(Df₂)** holds on $\Upsilon_o(r)$ for a fixed r . Then

$$\sup_{\mathbf{v} \in \Upsilon_o(r)} \|D^{-1} \mathbb{F}(\mathbf{v}) D^{-1} - I_p\|_{\text{op}} \leq \delta(r) \quad (16.10)$$

with

$$\delta(r) = \frac{\|\mathbf{b}\| w_2}{\sqrt{n}} + \frac{3\sigma r w_2}{\sqrt{n}} + \frac{3\sigma^2 r^2 w_2^2}{2n}.$$

Moreover, under $w_2 \sigma r \leq 2\sqrt{n}$, one can simplify

$$\delta(r) \stackrel{\text{def}}{=} \frac{\|\mathbf{b}\| w_2}{\sqrt{n}} + \frac{6\sigma r w_2}{\sqrt{n}}. \quad (16.11)$$

Proof. We have to evaluate the modulus of continuity of $\mathbb{F}(\mathbf{v})$ over $\Upsilon_o(r)$. For this we will use the following technical statement.

Lemma 16.2.2. Let $\mathbf{g}(t)$ be a continuously differentiable vector function with values in \mathbb{R}^n for $t \in [0, 1]$. Then

$$\|\mathbf{g}(1) - \mathbf{g}(0)\|^2 \leq C_g^2, \quad (16.12)$$

with

$$C_g \stackrel{\text{def}}{=} \int_0^1 \|g'(t)\| dt \leq \sup_{t \in [0,1]} \|g'(t)\|.$$

Moreover,

$$\left| \|g(1)\|^2 - \|g(0)\|^2 \right| \leq 2 C_g \|g(0)\| + C_g^2. \quad (16.13)$$

Proof. We use the representation

$$\|g(1) - g(0)\| = \left\{ \sum_{i=1}^n |g_i(1) - g_i(0)|^2 \right\}^{1/2} = \sup_{c \in \mathbb{R}^n : \|c\|=1} \sum_{i=1}^n c_i \{g_i(1) - g_i(0)\}.$$

Further, for each unit vector $c \in \mathbb{R}^n$,

$$\sum_{i=1}^n c_i \{g_i(1) - g_i(0)\} = \sum_{i=1}^n c_i \int_0^1 g'_i(t) dt \leq \|c\| \int_0^1 \|g'(t)\| dt,$$

and (16.12) follows. It also holds

$$\left| \|g(1)\|^2 - \|g(0)\|^2 \right| \leq \|g(1) - g(0)\|^2 + 2 \|g(1) - g(0)\| \|g(0)\|$$

yielding (16.13) by (16.12).

The value σ cancels in the left hand-side of (16.10), and we can reduce the proof to the case $\sigma = 1$ in definition (16.8) by replacing r with σr . Let $v \in \Upsilon_o(r)$ so that $\|D(v - v^*)\| \leq \sigma r$. Fix an unit vector $u \in \mathbb{R}^p$ and denote $\gamma = D^{-1}u = n^{-1/2}\mathsf{H}^{-1}u$. By definition of $\mathbb{F}(v)$

$$\begin{aligned} \gamma^\top \{\mathbb{F}(v) - D^2\} \gamma &= \sum_{i=1}^n \gamma^\top \left[\nabla f_i(v) \{\nabla f_i(v)\}^\top - \nabla f_i(v^*) \{\nabla f_i(v^*)\}^\top \right] \gamma \\ &\quad + \sum_{i=1}^n \{f_i(v^*) - f_i\} \gamma^\top \nabla^2 f_i(v) \gamma \\ &\quad + \sum_{i=1}^n \{f_i(v) - f_i(v^*)\} \gamma^\top \nabla^2 f_i(v) \gamma. \end{aligned} \quad (16.14)$$

Next we bound the first sum in (16.14). For any $v \in \Upsilon_o(r)$ and $t \in [0, 1]$, denote $v(t) = v^* + t(v - v^*) = v^* + tn^{-1/2}\mathsf{H}^{-1}\alpha$ with $\alpha = n^{1/2}\mathsf{H}(v - v^*)$.

$$g_i(t) = \gamma^\top \nabla f_i(v(t)) = n^{-1/2} u^\top \mathsf{H}^{-1} \nabla f_i(v^* + t(v - v^*)), \quad 0 \leq t \leq 1.$$

It holds

$$\begin{aligned} & \sum_{i=1}^n \gamma^\top \left[\nabla f_i(\mathbf{v}) \{ \nabla f_i(\mathbf{v}) \}^\top - \nabla f_i(\mathbf{v}^*) \{ \nabla f_i(\mathbf{v}^*) \}^\top \right] \gamma \\ &= \sum_{i=1}^n |\gamma^\top \nabla f_i(\mathbf{v})|^2 - \sum_{i=1}^n |\gamma^\top \nabla f_i(\mathbf{v}^*)|^2 = \|\mathbf{g}(1)\|^2 - \|\mathbf{g}(0)\|^2. \end{aligned}$$

Obviously

$$g'_i(t) = n^{-1} \mathbf{u}^\top \mathsf{H}^{-1} \nabla^2 f_i(\mathbf{v}^* + t(\mathbf{v} - \mathbf{v}^*)) \mathsf{H}^{-1} \boldsymbol{\alpha},$$

and $\|\boldsymbol{\alpha}\| \leq \sigma \mathbf{r}$ yields by (16.5)

$$C_g^2 = \sup_{t \in [0,1]} \|\mathbf{g}'(t)\|^2 = \sup_{t \in [0,1]} \frac{1}{n^2} \sum_{i=1}^n |\mathbf{u}^\top \mathsf{H}^{-1} \nabla^2 f_i(\mathbf{v}(t)) \mathsf{H}^{-1} \boldsymbol{\alpha}|^2 \leq \frac{\sigma^2 \mathbf{r}^2 w_2^2}{n}.$$

By definition (16.3) of D^2 (with $\sigma = 1$), it holds

$$\begin{aligned} \|\mathbf{g}(0)\|^2 &= \sum_{i=1}^n |\gamma^\top \nabla f_i(\mathbf{v}^*)|^2 \\ &= \mathbf{u}^\top D^{-1} \sum_{i=1}^n \nabla f_i(\mathbf{v}^*) \{ \nabla f_i(\mathbf{v}^*) \}^\top D^{-1} \mathbf{u} = \mathbf{u}^\top \mathbf{u} = 1. \end{aligned}$$

Now Lemma 16.2.2 yields

$$\left| \|\mathbf{g}(1)\|^2 - \|\mathbf{g}(0)\|^2 \right| \leq 2C_g \|\mathbf{g}(0)\| + C_g^2 \leq \frac{2\sigma \mathbf{r} w_2}{\sqrt{n}} + \frac{\sigma^2 \mathbf{r}^2 w_2^2}{n}. \quad (16.15)$$

Further, again by Lemma 16.2.2, one can bound in a similar way

$$\begin{aligned} & \sum_{i=1}^n \{ f_i(\mathbf{v}) - f_i(\mathbf{v}^*) - (\mathbf{v} - \mathbf{v}^*)^\top \nabla f_i(\mathbf{v}^*) \}^2 \\ & \leq \frac{1}{4n^2} \sup_{t \in [0,1]} \sum_{i=1}^n |\boldsymbol{\alpha}^\top \mathsf{H}^{-1} \nabla^2 f_i(\mathbf{v}(t)) \mathsf{H}^{-1} \boldsymbol{\alpha}|^2 \leq \frac{\sigma^4 \mathbf{r}^4 w_2^2}{4n}. \end{aligned}$$

For any $\mathbf{v} \in \Upsilon_\circ(\mathbf{r})$, it holds $\|D(\mathbf{v} - \mathbf{v}^*)\| \leq \sigma \mathbf{r}$. By definition $\gamma = D^{-1} \mathbf{u} = n^{-1/2} \mathsf{H}^{-1} \mathbf{u}$ yielding

$$\gamma^\top \nabla^2 f_i(\mathbf{v}) \gamma = n^{-1} \mathbf{u}^\top \mathsf{H}^{-1} \nabla^2 f_i(\mathbf{v}) \mathsf{H}^{-1} \mathbf{u},$$

and the Cauchy-Schwarz inequality implies by (16.5)

$$\begin{aligned}
& \left| \sum_{i=1}^n \{f_i(\mathbf{v}) - f_i(\mathbf{v}^*)\} \boldsymbol{\gamma}^\top \nabla^2 f_i(\mathbf{v}) \boldsymbol{\gamma} \right| \\
& \leq \left| \sum_{i=1}^n \{f_i(\mathbf{v}) - f_i(\mathbf{v}^*) - (\mathbf{v} - \mathbf{v}^*)^\top \nabla f_i(\mathbf{v}^*)\} \boldsymbol{\gamma}^\top \nabla^2 f_i(\mathbf{v}) \boldsymbol{\gamma} \right| \\
& \quad + \left| \sum_{i=1}^n (\mathbf{v} - \mathbf{v}^*)^\top \nabla f_i(\mathbf{v}^*) \boldsymbol{\gamma}^\top \nabla^2 f_i(\mathbf{v}) \boldsymbol{\gamma} \right| \\
& \leq \left\{ \sum_{i=1}^n \{f_i(\mathbf{v}) - f_i(\mathbf{v}^*) - (\mathbf{v} - \mathbf{v}^*)^\top \nabla f_i(\mathbf{v}^*)\}^2 \right\}^{1/2} \left\{ \sum_{i=1}^n |\boldsymbol{\gamma}^\top \nabla^2 f_i(\mathbf{v}) \boldsymbol{\gamma}|^2 \right\}^{1/2} \\
& \quad + \left\{ \sum_{i=1}^n |\{D(\mathbf{v} - \mathbf{v}^*)\}^\top D^{-1} \nabla f_i(\mathbf{v}^*)|^2 \right\}^{1/2} \left\{ \sum_{i=1}^n |\boldsymbol{\gamma}^\top \nabla^2 f_i(\mathbf{v}) \boldsymbol{\gamma}|^2 \right\}^{1/2} \\
& \leq \frac{\sigma^2 \mathbf{r}^2 w_2^2}{2n} + \frac{\sigma \mathbf{r} w_2}{\sqrt{n}}. \tag{16.16}
\end{aligned}$$

Finally, for each $\mathbf{v} \in \Upsilon_\circ(\mathbf{r})$, again by the Cauchy-Schwarz inequality implies by (16.5)

$$\begin{aligned}
& \left| \sum_{i=1}^n \{f_i(\mathbf{v}^*) - f_i\} \boldsymbol{\gamma}^\top \nabla^2 f_i(\mathbf{v}) \boldsymbol{\gamma} \right|^2 \\
& \leq \sum_{i=1}^n \{f_i(\mathbf{v}^*) - f_i\}^2 \sum_{i=1}^n |\mathbf{u}^\top \mathbf{H}^{-1} \nabla^2 f_i(\mathbf{v}) \mathbf{H}^{-1} \mathbf{u}|^2 \leq \frac{\|\mathbf{b}\|^2 w_2^2}{n}. \tag{16.17}
\end{aligned}$$

Summing up all the obtained bounds (16.15), (16.16), (16.17), yields by (16.14)

$$\left| \boldsymbol{\gamma}^\top \{\mathbb{F}(\mathbf{v}) - D^2\} \boldsymbol{\gamma} \right| \leq \frac{3\sigma \mathbf{r} w_2}{\sqrt{n}} + \frac{3\sigma^2 \mathbf{r}^2 w_2^2}{2n} + \frac{\|\mathbf{b}\| w_2}{\sqrt{n}}$$

which proves (??).

Now we evaluate the quality of approximating the difference $\|f(\mathbf{v}) - f(\mathbf{v}^*)\|^2$ by a quadratic form of $\mathbf{v} - \mathbf{v}^*$.

Proposition 16.2.2. *Under conditions (Df_2) and $w_2 \sigma \mathbf{r} \leq 2n^{1/2}$, for any $\mathbf{v} \in \Upsilon_\circ(\mathbf{r})$*

$$\left| \|f(\mathbf{v}) - f(\mathbf{v}^*)\|^2 - \sigma^2 \|D(\mathbf{v} - \mathbf{v}^*)\|^2 \right| \leq \frac{3\sigma^3 \|D(\mathbf{v} - \mathbf{v}^*)\|^3 w_2}{2\sqrt{n}}, \tag{16.18}$$

and for any $\mathbf{v}, \mathbf{v}_1 \in \Upsilon_\circ(\mathbf{r})$

$$\left| \|f(\mathbf{v}) - f(\mathbf{v}_1)\|^2 - \sigma^2 \|D(\mathbf{v} - \mathbf{v}_1)\|^2 \right| \leq \frac{6\sigma^3 \mathbf{r}^3 w_2}{\sqrt{n}}.$$

Proof. Similarly to Lemma 16.2.2, one can bound for any $\mathbf{v} \in \Upsilon(\mathbf{r})$

$$\begin{aligned}
\alpha(\mathbf{v}) &\stackrel{\text{def}}{=} \|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{v}^*) - (\mathbf{v} - \mathbf{v}^*)^\top \nabla \mathbf{f}(\mathbf{v}^*)\| \\
&\leq \frac{1}{2} \sup_{\mathbf{v}^\circ \in [\mathbf{v}^*, \mathbf{v}]} \left\| (\mathbf{v} - \mathbf{v}^*)^\top \nabla^2 \mathbf{f}(\mathbf{v}^\circ) (\mathbf{v} - \mathbf{v}^*) \right\| \\
&\leq \frac{\sigma^2}{2n} \sup_{\mathbf{v}^\circ \in [\mathbf{v}^*, \mathbf{v}]} \left\| \{D(\mathbf{v} - \mathbf{v}^*)\}^\top \mathbf{H}^{-1} \nabla^2 \mathbf{f}(\mathbf{v}^\circ) \mathbf{H}^{-1} D(\mathbf{v} - \mathbf{v}^*) \right\| \\
&\leq \frac{\sigma^2 \|D(\mathbf{v} - \mathbf{v}^*)\|^2 w_2}{2\sqrt{n}}
\end{aligned}$$

and for each pair $\mathbf{v}, \mathbf{v}_1 \in \Upsilon_\circ(\mathbf{r})$

$$\begin{aligned}
\alpha(\mathbf{v}, \mathbf{v}_1) &\stackrel{\text{def}}{=} \|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{v}_1) - (\mathbf{v} - \mathbf{v}_1)^\top \nabla \mathbf{f}(\mathbf{v}^*)\| \\
&\leq \|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{v}^*) - (\mathbf{v} - \mathbf{v}^*)^\top \nabla \mathbf{f}(\mathbf{v}^*)\| \\
&\quad + \|\mathbf{f}(\mathbf{v}_1) - \mathbf{f}(\mathbf{v}^*) - (\mathbf{v}_1 - \mathbf{v}^*)^\top \nabla \mathbf{f}(\mathbf{v}^*)\| \leq \frac{\sigma^2 \mathbf{r}^2 w_2}{\sqrt{n}}.
\end{aligned}$$

This implies under $w_2 \sigma \mathbf{r} \leq 2\sqrt{n}$ in view of $\|(\mathbf{v} - \mathbf{v}^*)^\top \nabla \mathbf{f}(\mathbf{v}^*)\|^2 = \sigma^2 \|D(\mathbf{v} - \mathbf{v}^*)\|^2$

$$\begin{aligned}
&\left| \|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{v}^*)\|^2 - \sigma^2 \|D(\mathbf{v} - \mathbf{v}^*)\|^2 \right| \\
&= \left| \|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{v}^*) - (\mathbf{v} - \mathbf{v}^*)^\top \nabla \mathbf{f}(\mathbf{v}^*) + (\mathbf{v} - \mathbf{v}^*)^\top \nabla \mathbf{f}(\mathbf{v}^*)\|^2 - \sigma^2 \|D(\mathbf{v} - \mathbf{v}^*)\|^2 \right| \\
&\leq \alpha^2(\mathbf{v}) + 2\alpha(\mathbf{v}) \|(\mathbf{v} - \mathbf{v}^*)^\top \nabla \mathbf{f}(\mathbf{v}^*)\| \leq \frac{3\sigma^3 \|D(\mathbf{v} - \mathbf{v}^*)\|^3 w_2}{2\sqrt{n}}, \\
&\left| \|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{v}_1)\|^2 - \sigma^2 \|D(\mathbf{v} - \mathbf{v}_1)\|^2 \right| \\
&\leq \alpha^2(\mathbf{v}, \mathbf{v}_1) + 2\alpha(\mathbf{v}, \mathbf{v}_1) \|D(\mathbf{v} - \mathbf{v}_1)\| \leq \frac{6\sigma^3 \mathbf{r}^3 w_2}{\sqrt{n}}
\end{aligned}$$

as required.

It is worth stressing that the results of Propositions 16.2.1 and 16.2.2 are entirely based on **(Df₂)** and do not rely on the definition of \mathbf{v}^* . In other words, it applies to any central point \mathbf{v}^* provided that the bias $\mathbf{b} = \mathbf{f}(\mathbf{v}^*) - \mathbf{f}^*$ is small in ℓ_2 -norm. The next result on the quadratic approximation of the expected log-likelihood $\mathbb{E}L(\mathbf{v})$ is stated for the point \mathbf{v}^* maximizing $\mathbb{E}L(\mathbf{v}^*)$.

Proposition 16.2.3. *Let $\mathbf{v}^* = \operatorname{argmax}_{\mathbf{v}} \mathbb{E}L(\mathbf{v})$. Under condition **(Df₂)**, it holds for $\mathbf{b} = \mathbf{f}(\mathbf{v}^*) - \mathbf{f}^*$ and each $\mathbf{v} \in \Upsilon_\circ(\mathbf{r})$*

$$\left| \|\mathbf{f}(\mathbf{v}) - \mathbf{f}^*\|^2 - \|\mathbf{b}\|^2 - \sigma^2 \|D(\mathbf{v} - \mathbf{v}^*)\|^2 \right| \leq \sigma^2 \mathbf{r}^2 \delta(\mathbf{r}), \tag{16.19}$$

for $\delta(\mathbf{r})$ from (16.11). This yields in particular

$$\|\mathbf{f}(\mathbf{v}) - \mathbf{f}^*\|^2 \leq \|\mathbf{b}\|^2 + \sigma^2 \|D(\mathbf{v} - \mathbf{v}^*)\|^2 + \sigma^2 \mathbf{r}^2 \delta(\mathbf{r}). \tag{16.20}$$

Proof. By definition $\mathbb{F}(\mathbf{v}) = \nabla^2 h(\mathbf{v})$ with $h(\mathbf{v}) = (2\sigma^2)^{-1} \|\mathbf{f}(\mathbf{v}) - \mathbf{f}^*\|^2$. As \mathbf{v}^* is the extreme point of $h(\mathbf{v})$, it holds $\nabla h(\mathbf{v}^*) = 0$. The second order Taylor expansion implies

$$\begin{aligned} 2h(\mathbf{v}) - 2h(\mathbf{v}^*) - \|D(\mathbf{v} - \mathbf{v}^*)\|^2 &= (\mathbf{v} - \mathbf{v}^*)^\top \{\mathbb{F}(\mathbf{v}^\circ) - D^2\}(\mathbf{v} - \mathbf{v}^*) \\ &= \{D(\mathbf{v} - \mathbf{v}^*)\}^\top \{D^{-1}\mathbb{F}(\mathbf{v}^\circ)D^{-1} - I_p\}D(\mathbf{v} - \mathbf{v}^*), \end{aligned}$$

where \mathbf{v}° is a point on the line connecting \mathbf{v} and \mathbf{v}^* , and the statement (16.19) follows from (16.10) in view of $2\sigma^2 h(\mathbf{v}^*) = \|\mathbf{f}(\mathbf{v}^*) - \mathbf{f}^*\|^2 = \|\mathbf{b}\|^2$.

16.3 Large deviation result and Fisher expansion

With D^2 from (16.3) and V^2 from (16.6), define

$$\begin{aligned} \boldsymbol{\xi} &\stackrel{\text{def}}{=} D^{-1}\nabla L(\mathbf{v}^*) = D^{-1}\nabla\zeta(\mathbf{v}^*) = \sigma^{-2} D^{-1} \sum_{i=1}^n \varepsilon_i \nabla f_i(\mathbf{v}^*), \\ B_V &\stackrel{\text{def}}{=} D^{-1}V^2D^{-1}. \end{aligned}$$

The next result describes the radius \mathbf{r}_0 providing the concentration of $\tilde{\mathbf{v}}$ in $\Upsilon_\circ(\mathbf{r}_0)$.

Theorem 16.3.1. *Let $\Upsilon \subset \Upsilon_\circ(\mathbf{r}^*)$ for some $\mathbf{r}^* < \infty$, and conditions $(\sigma_{1|n}^2)$ and (Df_2) hold. Further, let \mathbf{r}_0 be fixed to ensure the condition*

$$\mathbf{r}_0(1 - \delta(\mathbf{r}_0)) \geq 2z(B_V, \mathbf{x}) \quad (16.21)$$

with $\delta(\mathbf{r})$ from (16.11), and let

$$1 - \delta(\mathbf{r}) \geq \sqrt{8} \nu_0 \mathfrak{z}_{\mathbb{H}}(\mathbf{x} + \log(2\mathbf{r}/\mathbf{r}_0)) C_V \frac{\sigma w_2}{\sqrt{n}}, \quad \mathbf{r} \in [\mathbf{r}_0, \mathbf{r}^*]. \quad (16.22)$$

Then there exists a random set $\Omega(\mathbf{x})$ with $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - 4e^{-\mathbf{x}}$ such that on $\Omega(\mathbf{x})$ holds

$$\tilde{\mathbf{v}} \in \Upsilon_\circ(\mathbf{r}_0), \quad \|D(\tilde{\mathbf{v}} - \mathbf{v}^*) - \boldsymbol{\xi}\| \leq \diamond(\mathbf{r}_0), \quad (16.23)$$

where the error $\diamond(\mathbf{r}_0)$ satisfies under $w_2 \sigma \mathbf{r}_0 \leq 2\sqrt{n}$

$$\begin{aligned} \diamond(\mathbf{r}_0) &\leq \mathbf{r}_0 \{\delta(\mathbf{r}_0) + C \omega \sqrt{p + \mathbf{x}}\} \\ &\leq (6\mathbf{r}_0^2 + C C_V \mathbf{r}_0 \sqrt{p + \mathbf{x}}) \frac{\sigma w_2}{\sqrt{n}} + \frac{\|\mathbf{b}\| w_2}{\sqrt{n}} \mathbf{r}_0. \end{aligned}$$

If $\|\mathbf{b}\| \leq \sigma \sqrt{p + \mathbf{x}}$, then

$$\diamond(\mathbf{r}_0) \leq C \frac{\sigma (p + \mathbf{x}) w_2}{\sqrt{n}}. \quad (16.24)$$

Proof. The result is derived from the general statement of Theorems 9.3.1 and 9.3.2. We already checked that $(\sigma_{1|n}^2)$ and (Df_2) imply the local conditions (ED_0) , (ED_2) , and (L_0) . Lemma 16.2.1 helps to upper bound the value ω ; see (16.7). Also it ensures the identifiability condition (I) . Further, (16.22) implies $-\nabla^2 \mathbb{E}L(\mathbf{v}) \geq \{1 - \delta(\mathbf{r})\} D^2$ for $\mathbf{v} \in \Upsilon_o(\mathbf{r})$ and the global condition (L) follows as well; see Remark 9.2.2. This and (16.22) yield the concentration result $\tilde{\mathbf{v}} \in \Upsilon_o(\mathbf{r}_0)$ by Theorem 9.3.1. The Fisher expansion follows by Theorem 9.3.2.

16.4 Prediction loss and bias-variance decomposition

Now we consider the error of estimating the function f by the proposed procedure. The MLE $\tilde{\mathbf{v}}$ yields the prediction $\tilde{f}_i = f_i(\tilde{\mathbf{v}})$.

Theorem 16.4.1. *Under the conditions of Theorem 16.3.1, it holds on a random set $\Omega(\mathbf{x})$ with $I\!\!P(\Omega(\mathbf{x})) \geq 1 - 4e^{-x}$*

$$\sigma^{-2} \|\mathbf{f}(\tilde{\mathbf{v}}) - \mathbf{f}^*\|^2 \leq \sigma^{-2} \|\mathbf{b}\|^2 + \{\|\boldsymbol{\xi}\| + \diamond(\mathbf{r}_0)\}^2 + \mathbf{r}_0^2 \delta(\mathbf{r}_0). \quad (16.25)$$

Proof. The bound (16.20) implies under $\tilde{\mathbf{v}} \in \Upsilon_o(\mathbf{r}_0)$

$$\sigma^{-2} \|\mathbf{f}(\tilde{\mathbf{v}}) - \mathbf{f}^*\|^2 \leq \sigma^{-2} \|\mathbf{b}\|^2 + \|D(\tilde{\mathbf{v}} - \mathbf{v}^*)\|^2 + \mathbf{r}_0^2 \delta(\mathbf{r}_0).$$

The Fisher expansion (16.23) implies on the random set $\Omega(\mathbf{x})$ from Theorem 16.3.1

$$\sigma^{-2} \|\mathbf{f}(\tilde{\mathbf{v}}) - \mathbf{f}^*\|^2 \leq \sigma^{-2} \|\mathbf{b}\|^2 + \{\|\boldsymbol{\xi}\| + \diamond(\mathbf{r}_0)\}^2 + \mathbf{r}_0^2 \delta(\mathbf{r}_0).$$

The bound (16.25) can be viewed as analog of the bias-variance decomposition for the case of linear regression. Indeed, the first term in the bound measure the distance of the true regression function f from its best parametric approximation $f(\mathbf{v}^*)$, while the second term is the energy of the stochastic noise in the estimate $\tilde{f} = f(\tilde{\mathbf{v}})$.

Note that the bound (16.24) for $\diamond(\mathbf{r}_0)$ with $\mathbf{r}_0^2 = C(p+x)$ together with $\|\boldsymbol{\xi}\|^2 \leq \mathbf{r}_0^2 \leq C(p+x)$ enable us to derive under $\tilde{\mathbf{v}} \in \Upsilon_o(\mathbf{r}_0)$

$$\begin{aligned} \sigma^{-2} \|\mathbf{f}(\tilde{\mathbf{v}}) - \mathbf{f}^*\|^2 &\leq \sigma^{-2} \|\mathbf{b}\|^2 + \|\boldsymbol{\xi}\|^2 + \Delta(\mathbf{r}_0), \\ \Delta(\mathbf{r}_0) &= 3\mathbf{r}_0 \diamond(\mathbf{r}_0) \leq C \sigma w_2 \sqrt{(p+x)^3/n}. \end{aligned} \quad (16.26)$$

As the variance term $\|\boldsymbol{\xi}\|^2$ is of order p , the remainder $\Delta(\mathbf{r}_0)$ is small relative to $\|\boldsymbol{\xi}\|^2$ and the expansion is sharp under the condition

$$\Delta(\mathbf{r}_0)/p \asymp w_2 \sigma \sqrt{(p+x)/n} \quad \text{is small.}$$

To be done: please complete

16.5 Sieve nonparametric estimation

Now we consider the case when the underlying parameter \boldsymbol{v} is of large or infinite dimension p^* . Then the result of Theorem 16.4.1 is still applicable but almost useless because the energy $I\!\!E\|\boldsymbol{\xi}\|^2$ of the standardized score vector $\boldsymbol{\xi}$ in $I\!\!R^{p^*}$ is very large or infinite. The sieve approach can be naturally incorporated in the construction. One approximates the full model by a p dimensional sub-model and applies the theory for this sub-model. The only issue to address is the definition of the central point \boldsymbol{v}^* . In the full model

$$\boldsymbol{v}^* = \operatorname{argmax}_{\boldsymbol{v} \in \Upsilon} I\!\!E L(\boldsymbol{v}).$$

Here the maximum is taken over the whole parameter set Υ . We also define its sieve counterpart

$$\boldsymbol{v}_m^* = \operatorname{argmax}_{\boldsymbol{v} \in \Upsilon_m} I\!\!E L(\boldsymbol{v}),$$

where Υ_m is a subspace of Υ of dimension p_m . The sieve MLE reads as

$$\tilde{\boldsymbol{v}}_m = \operatorname{argmax}_{\boldsymbol{v} \in \Upsilon_m} L(\boldsymbol{v}).$$

The corresponding prediction reads as $\boldsymbol{f}(\tilde{\boldsymbol{v}}_m)$. The sieve score vector $\boldsymbol{\xi}_m$ from (16.27) is defined as

$$\boldsymbol{\xi}_m \stackrel{\text{def}}{=} \sigma^{-2} D_m^{-1} \sum_{i=1}^n \varepsilon_i \nabla_m f_i(\boldsymbol{v}^*),$$

where ∇_m means the partial derivative along the sieve sub-space. The sieve Fisher matrix D_m^2 is the m -block of the full Fisher matrix $D^2 = \mathbb{F}(\boldsymbol{v}^*)$. Here we can compute all the quantities at \boldsymbol{v}^* in place of \boldsymbol{v}_m^* under the condition that these two points are in the same neighborhood.

Suppose that condition **(Df₂)** is fulfilled for the sieve submodel with the corresponding constant w_m . Now the result of Theorem 16.4.1 applied to the sieve MLE $\tilde{\boldsymbol{v}}_m$ yields with $\mathbf{r}_m = \mathbf{r}_m(p_m) \asymp \sqrt{p_m + \mathbf{x}}$ and $\mathbf{b}_m = \boldsymbol{f}(\boldsymbol{v}_m^*) - \boldsymbol{f}^*$

$$\left| \sigma^{-2} \|\boldsymbol{f}(\tilde{\boldsymbol{v}}_m) - \boldsymbol{f}^*\|^2 - \sigma^{-2} \|\mathbf{b}_m\|^2 - \|\boldsymbol{\xi}_m\|^2 \right| \leq \Delta(\mathbf{r}_m). \quad (16.27)$$

The decomposition is nearly sharp if the error term $\Delta(\mathbf{r}_m)$ is small relative to the leading term $\|\boldsymbol{\xi}_m\|^2$. If $p_m = I\!\!E\|\boldsymbol{\xi}_m\|^2$ is large, then $p_m^{-1}\|\boldsymbol{\xi}_m\|^2 \approx 1$; see Theorem B.1.1. This means that the condition “ $p_m^{-1}\Delta(\mathbf{r}_m)$ is small” ensures $\sigma^{-2}\|\boldsymbol{f}(\tilde{\boldsymbol{v}}_m) - \boldsymbol{f}^*\|^2 \approx \sigma^{-2}\|\mathbf{b}_m\|^2 + \|\boldsymbol{\xi}_m\|^2$. In view of $\Delta(\mathbf{r}_m) \leq C w_m \sigma \sqrt{(p_m + \mathbf{x})^3/n}$, see (16.26), the required condition reads

$w_m \sigma \sqrt{(p_m + x)/n}$ is small.

To be done: state a precise result

Expansion (16.27) together with $p_m^{-1} \|\xi_m\|^2 \approx 1$ suggests to pick up the sieve index m by the usual *bias-variance trade-off*:

$$\sigma^{-2} \|\mathbf{b}_m\|^2 \leq C p_m.$$

If $n^{-1} \|\mathbf{b}_m\|^2 \asymp p_m^{-2\beta}$ for the smoothness parameter β , then the proper parameter dimension $p_m \asymp (\sigma^{-2} n)^{1/(2\beta+1)}$ provides the classical rate of estimation $(\sigma^{-2} n)^{-2\beta/(2\beta+1)}$ for the quadratic risk:

$$\frac{1}{n} \|\mathbf{f}(\tilde{\mathbf{v}}_m) - \mathbf{f}^*\|^2 \leq C(\sigma^{-2} n)^{-2\beta/(2\beta+1)}.$$

To be done: state a precise result

16.6 Penalized regression

Consider the parametric regression setup (16.1) in which the dimension p of the parameter \mathbf{v} is large or even infinite but the objective function in the MLE procedure is penalized by a quadratic term $\|G\mathbf{v}\|^2/2$. The corresponding penalized estimator $\tilde{\mathbf{v}}_G$ is defined as

$$\tilde{\mathbf{v}}_G \stackrel{\text{def}}{=} \operatorname{argmax}_{\mathbf{v}} \left\{ L(\mathbf{v}) - \frac{1}{2} \|G\mathbf{v}\|^2 \right\} = \operatorname{argmin}_{\mathbf{v}} \left\{ \frac{1}{\sigma^2} \|\mathbf{Y} - \mathbf{f}(\mathbf{v})\|^2 + \|G\mathbf{v}\|^2 \right\};$$

cf. (16.2). The roughness penalty is very natural here in the cases when the second derivatives of the regression functions $f(x, \mathbf{v})$ with respect to x and \mathbf{v} are related to each other. Indeed, the roughness penalty represents the squared L_2 -norm of $\nabla^2 f(x)$, while the L_2 -norm of $\nabla_{\mathbf{v}\mathbf{v}}^2 f(\mathbf{v})$ enters in condition (Df₂).

The key issues of the general roughness penalty approach apply here as well: the penalization does not affect the stochastic component; the penalty increases the deterministic component without deteriorating its smoothness properties. As a result, all the required conditions continue to apply in the penalized case if they are checked without penalization. Moreover, the most important constraint on the parameter dimension can be relaxed by using the effective dimension instead of the full parameter dimension.

Define for D^2 from (16.3)

$$D_G^2 = D^2 + G^2.$$

For the matrix $H^2 = \sigma^2 n^{-1} D^2$ from (16.4), define in a similar way its penalized version

$$\mathsf{H}_G^2 \stackrel{\text{def}}{=} \frac{\sigma^2}{n} D_G^2.$$

Obviously $\mathsf{H}^2 \leq \mathsf{H}_G^2$. Condition **(Df2)** from the non-penalized case can be used without any change, however, one can relax it by replacing H^2 with such defined H_G^2 and the same constants w_2 . Condition **(σ_{1|n}²)** only relies to the noise and extends to the penalized case without any problem. The bias \mathbf{b} is modified to \mathbf{b}_G with

$$\|\mathbf{b}_G\|^2 \stackrel{\text{def}}{=} \|\mathbf{f}^* - \mathbf{f}(\mathbf{v}_G^*)\|^2 = \sum_{i=1}^n |f_i - f_i(\mathbf{v}_G^*)|^2$$

with the penalized target \mathbf{v}_G^* given by

$$\mathbf{v}_G^* = \underset{\mathbf{v}}{\operatorname{argmin}} \{ \sigma^{-2} \|\mathbf{f}^* - \mathbf{f}(\mathbf{v})\|^2 + \|G\mathbf{v}\|^2 \}.$$

One can also use the approximation (16.19) and replace \mathbf{v}_G^* with \mathbf{v}_G^\dagger given by

$$\mathbf{v}_G^\dagger = \underset{\mathbf{v}}{\operatorname{argmin}} \{ \|D(\mathbf{v} - \mathbf{v}^*)\|^2 + \|G\mathbf{v}\|^2 \} = D_G^{-2} D^2 \mathbf{v}^*;$$

cf. (11.42). The use of penalization allows to reduce the total parameter dimension p to the effective dimension p_G given by

$$p_G \stackrel{\text{def}}{=} \operatorname{tr}(D_G^{-2} V^2),$$

where the score covariance matrix V^2 is shown in (16.6). All the results of Theorems 16.3.1 and 16.4.1 apply to the penalized estimator $\tilde{\mathbf{v}}_G$ with all quantities just introduced and with obvious changes of ξ by ξ_G , D by D_G , and of p by p_G . In particular, the results from (16.23) read

$$\|D_G(\tilde{\mathbf{v}}_G - \mathbf{v}_G^*)\| \leq \mathbf{r}_G, \quad \|D_G(\tilde{\mathbf{v}}_G - \mathbf{v}_G^*) - \xi_G\| \leq \diamondsuit_G(\mathbf{r}_G),$$

where $\mathbf{r}_G^2 \asymp p_G + x$ and the error term $\diamondsuit_G(\mathbf{r}_0)$ in the penalized case is of order $w_2(p_G + x)\sigma n^{-1/2}$.

To be done: complete as in sieve

16.7 Semiparametric problem

Here we specify the conditions to the case when the total parameter \mathbf{v} in the regression function f is composed of the target parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ and the nuisance parameter $\boldsymbol{\eta} \in H \subset \mathbb{R}^q$. As usual, $\nabla_{\boldsymbol{\theta}}$ and $\nabla_{\boldsymbol{\eta}}$ mean the corresponding parts of the gradient vector ∇ . We use the block representation for the $p^* \times p^*$ matrix H^2 with $p^* = p + q$:

$$\mathbf{H}^2 = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\boldsymbol{v}^*) \{ \nabla f_i(\boldsymbol{v}^*) \}^\top = \begin{pmatrix} \mathbf{H}_{\boldsymbol{\theta}}^2 & \mathbf{A}_{\boldsymbol{\theta}\boldsymbol{\eta}} \\ \mathbf{A}_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top & \mathbf{H}_{\boldsymbol{\eta}}^2 \end{pmatrix}.$$

Semiparametric identifiability requires that the off-diagonal block is not too big:

$$\| \mathbf{H}_{\boldsymbol{\theta}}^{-1} \mathbf{A}_{\boldsymbol{\theta}\boldsymbol{\eta}} \mathbf{H}_{\boldsymbol{\eta}}^{-1} \|_{\text{op}} \leq \nu \quad (16.28)$$

for some $\nu < 1$.

The large deviation bound of Theorem 16.3.1 continues to apply here without any significant change. We use the same notation, in particular, $D^2 = n\sigma^{-2}\mathbf{H}^2$.

Theorem 16.7.1. *Consider a parametric regression (16.1) with the parameter $\boldsymbol{v} = (\boldsymbol{\theta}, \boldsymbol{\eta})$. Let conditions of Theorem 16.3.1 be fulfilled with **(Df₂)** replaced by (16.28), and let for any unit vectors $\mathbf{u} \in \mathbb{R}^p$ and $\mathbf{e} \in \mathbb{R}^q$*

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n [\mathbf{u}^\top \mathbf{H}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 f_i(\boldsymbol{v}) \mathbf{H}_{\boldsymbol{\theta}}^{-1} \mathbf{u}]^2 &\leq w_{\boldsymbol{\theta}}^2, \\ \frac{1}{n} \sum_{i=1}^n [\mathbf{e}^\top \mathbf{H}_{\boldsymbol{\eta}}^{-1} \nabla_{\boldsymbol{\eta}\boldsymbol{\eta}}^2 f_i(\boldsymbol{v}) \mathbf{H}_{\boldsymbol{\eta}}^{-1} \mathbf{e}]^2 &\leq w_{\boldsymbol{\eta}}^2, \\ \frac{1}{n} \sum_{i=1}^n [\mathbf{u}^\top \mathbf{H}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}\boldsymbol{\eta}}^2 f_i(\boldsymbol{v}) \mathbf{H}_{\boldsymbol{\eta}}^{-1} \mathbf{e}]^2 &\leq w_s^2, \end{aligned} \quad (16.29)$$

for some fixed constants $w_{\boldsymbol{\theta}}$, $w_{\boldsymbol{\eta}}$, and w_s . Then (16.5) is fulfilled with

$$w_2^2 = \frac{w_{\boldsymbol{\theta}}^2 + w_{\boldsymbol{\eta}}^2 + 2w_s^2}{(1-\nu)^2},$$

and the choice of \mathbf{r}_0 due to (16.21) ensures for the MLE $\tilde{\boldsymbol{v}}$ on a random set $\Omega(\mathbf{x})$ of probability at least $1 - 4e^{-x}$

$$\begin{aligned} \tilde{\boldsymbol{v}} &\in \Upsilon_o(\mathbf{r}_0), \\ \| D(\tilde{\boldsymbol{v}} - \boldsymbol{v}^*) - \boldsymbol{\xi} \| &\leq \diamond(\mathbf{r}_0) = C w_2 (p+q+x) \sigma / \sqrt{n}. \end{aligned} \quad (16.30)$$

with

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} D^{-1} \nabla L(\boldsymbol{v}^*) = D^{-1} \begin{pmatrix} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{v}^*) \\ \nabla_{\boldsymbol{\eta}} L(\boldsymbol{v}^*) \end{pmatrix}.$$

For the $\boldsymbol{\theta}$ -component of \boldsymbol{v} , this yields on $\Omega(\mathbf{x})$

$$\| \mathbb{I}^{1/2} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \check{\boldsymbol{\xi}} \| \leq \diamond(\mathbf{r}_0) \quad (16.31)$$

with

$$\begin{aligned}\mathbb{I} &\stackrel{\text{def}}{=} n\sigma^{-2}(\mathsf{H}_{\boldsymbol{\theta}}^2 - \mathsf{A}_{\boldsymbol{\theta}\boldsymbol{\eta}}\mathsf{H}_{\boldsymbol{\eta}}^{-2}\mathsf{A}_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top), \\ \check{\boldsymbol{\xi}} &\stackrel{\text{def}}{=} \mathbb{I}^{-1/2}\{\nabla_{\boldsymbol{\theta}} L(\boldsymbol{v}^*) - \mathsf{A}_{\boldsymbol{\theta}\boldsymbol{\eta}}\mathsf{H}_{\boldsymbol{\eta}}^{-2}\nabla_{\boldsymbol{\eta}} L(\boldsymbol{v}^*)\}.\end{aligned}\quad (16.32)$$

Proof. Condition (16.28) allows to replace the full matrix H^2 by the block-diagonal matrix

$$\mathsf{H}_b^2 \stackrel{\text{def}}{=} \begin{pmatrix} \mathsf{H}_{\boldsymbol{\theta}}^2 & 0 \\ 0 & \mathsf{H}_{\boldsymbol{\eta}}^2 \end{pmatrix}.$$

More precisely, it implies

$$\mathsf{H}^{-2} = \begin{pmatrix} \mathsf{H}_{\boldsymbol{\theta}}^2 & \mathsf{A}_{\boldsymbol{\theta}\boldsymbol{\eta}} \\ \mathsf{A}_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top & \mathsf{H}_{\boldsymbol{\eta}}^2 \end{pmatrix}^{-1} \leq \frac{1}{1-\nu} \begin{pmatrix} \mathsf{H}_{\boldsymbol{\theta}}^{-2} & 0 \\ 0 & \mathsf{H}_{\boldsymbol{\eta}}^{-2} \end{pmatrix} = \frac{1}{1-\nu} \mathsf{H}_b^{-2}. \quad (16.33)$$

Let $(\mathbf{u}, \mathbf{e})^\top$ denote a vector on the sphere \mathcal{S}^{p^*} in \mathbb{R}^{p^*} for $p^* = p+q$. The bound (16.33) ensures that

$$\begin{aligned}&\sup_{(\mathbf{u}, \mathbf{e})^\top \in \mathcal{S}^{p^*}} \sum_{i=1}^n \left| \begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix}^\top \mathsf{H}^{-1} \nabla^2 f_i(\boldsymbol{v}) \mathsf{H}^{-1} \begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} \right|^2 \\ &\leq \frac{1}{(1-\nu)^2} \sup_{(\mathbf{u}, \mathbf{e})^\top \in \mathcal{S}^{p^*}} \sum_{i=1}^n \left| \begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix}^\top \mathsf{H}_b^{-1} \nabla^2 f_i(\boldsymbol{v}) \mathsf{H}_b^{-1} \begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} \right|^2.\end{aligned}$$

It now follows from (16.28) and (16.29) in view of $w_{\boldsymbol{\theta}}^2 + w_{\boldsymbol{\eta}}^2 + w_s^2 = w_2^2$

$$\begin{aligned}&\sum_{i=1}^n \left| \begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix}^\top \mathsf{H}_b^{-1} \nabla^2 f_i(\boldsymbol{v}) \mathsf{H}_b^{-1} \begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} \right|^2 \\ &= \sum_{i=1}^n \left| \mathbf{u}^\top \mathsf{H}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 f_i(\boldsymbol{v}) \mathsf{H}_{\boldsymbol{\theta}}^{-1} \mathbf{u} + \mathbf{e}^\top \mathsf{H}_{\boldsymbol{\eta}}^{-1} \nabla_{\boldsymbol{\eta}\boldsymbol{\eta}}^2 f_i(\boldsymbol{v}) \mathsf{H}_{\boldsymbol{\eta}}^{-1} \mathbf{e} \right. \\ &\quad \left. + 2\mathbf{u}^\top \mathsf{H}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}\boldsymbol{\eta}}^2 f_i(\boldsymbol{v}) \mathsf{H}_{\boldsymbol{\eta}}^{-1} \mathbf{e} \right|^2 \\ &\leq \frac{w_2^2}{w_{\boldsymbol{\theta}}^2} \sum_{i=1}^n \left| \mathbf{u}^\top \mathsf{H}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 f_i(\boldsymbol{v}) \mathsf{H}_{\boldsymbol{\theta}}^{-1} \mathbf{u} \right|^2 + \frac{w_2^2}{w_{\boldsymbol{\eta}}^2} \sum_{i=1}^n \left| \mathbf{e}^\top \mathsf{H}_{\boldsymbol{\eta}}^{-1} \nabla_{\boldsymbol{\eta}\boldsymbol{\eta}}^2 f_i(\boldsymbol{v}) \mathsf{H}_{\boldsymbol{\eta}}^{-1} \mathbf{e} \right|^2 \\ &\quad + \frac{2w_2^2}{w_s^2} \sum_{i=1}^n \left| \mathbf{u}^\top \mathsf{H}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}\boldsymbol{\eta}}^2 f_i(\boldsymbol{v}) \mathsf{H}_{\boldsymbol{\eta}}^{-1} \mathbf{e} \right|^2 \\ &\leq n w_2^2 (\|\mathbf{u}\|^4 + \|\mathbf{e}\|^4 + 2\|\mathbf{u}\|^2 \|\mathbf{e}\|^2) = n w_2^2 (\|\mathbf{u}\|^2 + \|\mathbf{e}\|^2)^2 = n w_2^2\end{aligned}$$

which implies the bound in (16.5) and Theorem 16.3.1 applies.

Corollary 16.7.1. Suppose that the conditions of Theorem 16.7.1 hold. Let $\check{B} = \text{Var}(\check{\boldsymbol{\xi}})$. Then

$$\|\mathbb{I}^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \leq \|\check{\boldsymbol{\xi}}\| + \diamond(\mathbf{r}_0) \leq z(\check{B}, \mathbf{x}) + \diamond(\mathbf{r}_0) \leq \mathbf{r}_{\boldsymbol{\theta}} \quad (16.34)$$

on a set $\Omega_{\boldsymbol{\theta}}(\mathbf{x})$ with $\mathbb{P}(\Omega_{\boldsymbol{\theta}}(\mathbf{x})) \geq 1 - 6e^{-x}$.

Proof. It holds

$$\mathbb{P}(\|\check{\boldsymbol{\xi}}\| \geq z(\check{B}, \mathbf{x})) \leq 2e^{-x}.$$

This and (16.31) imply (16.34).

The bound (16.25) on the prediction loss holds in the semiparametric set-up. It can be viewed as combination of the quadratic expansion (16.19) of the expected log-likelihood and of the Fisher expansion (16.30). Now we specify the statement for its projection on the $\boldsymbol{\theta}$ -component starting from the prediction risk. It follows by (16.18) for each $\mathbf{v} = (\boldsymbol{\theta}, \boldsymbol{\eta}^*) \in \Upsilon_o(\mathbf{r}_0)$ with $D_{\boldsymbol{\theta}}^2 = n\sigma^{-2}\mathsf{H}_{\boldsymbol{\theta}}^2$

$$\left| \|f(\boldsymbol{\theta}, \boldsymbol{\eta}^*) - f(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)\|^2 - \sigma^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top D_{\boldsymbol{\theta}}^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right| \leq 3w_2 \frac{\sigma^3 \|D_{\boldsymbol{\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^3}{\sqrt{n}}.$$

Consider first a special adaptive case when the full Fisher information matrix \mathbb{F} is of block structure: $\mathsf{H}^2 = \text{block}\{\mathsf{H}_{\boldsymbol{\theta}}^2, \mathsf{H}_{\boldsymbol{\eta}}^2\}$. Then $\mathbb{I} = n\sigma^{-2}\mathsf{H}_{\boldsymbol{\theta}}^2$; cf. (16.32). After substitution $\boldsymbol{\theta}$ with $\tilde{\boldsymbol{\theta}}$ we obtain on $\Omega(\mathbf{x})$ for $\|\mathbb{I}^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}_{\boldsymbol{\theta}}$

$$\left| \|f(\tilde{\boldsymbol{\theta}}, \boldsymbol{\eta}^*) - f(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)\|^2 - \sigma^2(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \mathbb{I}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \right| \leq 3w_2 \frac{\sigma^3 \mathbf{r}_{\boldsymbol{\theta}}^3}{2\sqrt{n}}$$

and using $\|\mathbb{I}^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}_{\boldsymbol{\theta}}$

$$\|f(\tilde{\boldsymbol{\theta}}, \boldsymbol{\eta}^*) - f(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)\|^2 \leq \sigma^2 \mathbf{r}_{\boldsymbol{\theta}}^2 + 3w_2 \frac{\sigma^3 \mathbf{r}_{\boldsymbol{\theta}}^3}{2\sqrt{n}}.$$

If H^2 is not of block structure, then by (16.28) $\mathsf{H}_{\boldsymbol{\theta}}^2 - \mathsf{A}_{\boldsymbol{\theta}\boldsymbol{\eta}} \mathsf{H}_{\boldsymbol{\eta}}^{-2} \mathsf{A}_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top \geq (1-\nu)\mathsf{H}_{\boldsymbol{\theta}}^2$. Therefore,

$$\|f(\tilde{\boldsymbol{\theta}}, \boldsymbol{\eta}^*) - f(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)\|^2 \leq \frac{\sigma^2 \mathbf{r}_{\boldsymbol{\theta}}^2}{1-\nu} + 3w_2 \frac{\sigma^3 \mathbf{r}_{\boldsymbol{\theta}}^3}{2\sqrt{n}}$$

yielding by $\mathbf{r}_{\boldsymbol{\theta}}^2 \approx \mathbb{E}\|\check{\boldsymbol{\xi}}\|^2 \leq C(p+x)$

$$\|f(\tilde{\boldsymbol{\theta}}, \boldsymbol{\eta}^*) - f(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)\|^2 \leq C \frac{\sigma^2 \mathbf{r}_{\boldsymbol{\theta}}^2}{1-\nu}$$

provided that

$$w_2 \sigma \sqrt{\frac{p+x}{n}} \quad \text{is small.}$$

To be done: state the precise result.

16.8 Random design regression

This section comments what should be changed or added to our result for their applicability to regression with random design. Consider a regression model with a parametric regression function $f(x, \mathbf{v})$ and a random design $\mathbf{X} = (X_1, \dots, X_n)$. We follow below the standard path and assume that the design variables X_i are i.i.d. But this assumption can be easily relaxed. What is really important is a kind of strong ergodicity: empirical averaging w.r.t. the design measure is close to its population counterpart.

Suppose that the true distribution of the observations Y_i is described by the regression model

$$Y_i = f(X_i) + \varepsilon_i$$

where ε_i are independent zero mean with $\sigma_i^2 = \text{Var}(\varepsilon_i | \mathbf{X})$ and $f(\cdot)$ is an unknown regression function. The parametric approach assumes that $f(\cdot)$ belongs to a given parametric family $\{f(\cdot, \mathbf{v}), \mathbf{v} \in \Upsilon\}$. Consider the quasi MLE $\tilde{\mathbf{v}}$ of the parameter \mathbf{v} defined by maximization of the log-likelihood function built for parametric regression function and Gaussian independent errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$:

$$\begin{aligned} L(\mathbf{v}) &= -\frac{1}{2\sigma^2} \sum_{i=1}^n |Y_i - f(X_i, \mathbf{v})|^2 + R, \\ \tilde{\mathbf{v}} &= \underset{\mathbf{v}}{\operatorname{argmax}} L(\mathbf{v}) = \underset{\mathbf{v}}{\operatorname{argmin}} \sum_{i=1}^n |Y_i - f(X_i, \mathbf{v})|^2. \end{aligned}$$

Note that both assumptions can be violated, that is, in general, $f(\cdot)$ does not coincide with $f(\cdot, \mathbf{v})$ for any \mathbf{v} , and the errors ε_i can be non-normal and heterogeneous. Define the target value \mathbf{v}^* as

$$\mathbf{v}^* \stackrel{\text{def}}{=} \underset{\mathbf{v}}{\operatorname{argmax}} \mathbb{E}_{\mathbb{Q}} L(\mathbf{v}) = \underset{\mathbf{v}}{\operatorname{argmin}} \sum_{i=1}^n \mathbb{E}_{\mathbb{Q}} |f(X_i) - f(X_i, \mathbf{v})|^2.$$

Here the expectation $\mathbb{E}_{\mathbb{Q}}$ is taken w.r.t. the design measure. Further, define the Fisher information matrix

$$\mathsf{H}^2 \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mathbb{Q}} \left[\nabla f(X_i, \mathbf{v}^*) \{ \nabla f(X_i, \mathbf{v}^*) \}^\top \right],$$

where the symbol ∇ means differentiation w.r.t. \mathbf{v} , $\nabla = \nabla_{\mathbf{v}}$. If the design variables X_i are i.i.d. then

$$\mathsf{H}^2 = \mathbb{E}_{\mathbb{Q}} \left[\nabla f(X_1, \mathbf{v}^*) \{ \nabla f(X_1, \mathbf{v}^*) \}^\top \right].$$

Introduce

$$\begin{aligned}\tilde{V}^2 &= \frac{1}{\sigma^4} \sum_{i=1}^n \sigma_i^2 \left[\nabla f(X_i, \mathbf{v}^*) \{ \nabla f(X_i, \mathbf{v}^*) \}^\top \right] \\ V^2 &= \frac{1}{\sigma^4} \sum_{i=1}^n \sigma_i^2 \mathbb{E}_{\mathbb{Q}} \left[\nabla f(X_i, \mathbf{v}^*) \{ \nabla f(X_i, \mathbf{v}^*) \}^\top \right]\end{aligned}$$

and

$$\begin{aligned}\tilde{D}^2 &= \frac{1}{\sigma^2} \sum_{i=1}^n \left[\nabla f(X_i, \mathbf{v}^*) \{ \nabla f(X_i, \mathbf{v}^*) \}^\top \right] \\ D^2 &= \frac{1}{\sigma^2} \sum_{i=1}^n \mathbb{E}_{\mathbb{Q}} \left[\nabla f(X_i, \mathbf{v}^*) \{ \nabla f(X_i, \mathbf{v}^*) \}^\top \right] = \frac{n}{\sigma^2} \mathsf{H}^2.\end{aligned}$$

The matrix Bernstein inequality can be used to show that \tilde{D}^2 and \tilde{V}^2 are close to their population counterparts D^2 and V^2 . The bias $b(\cdot) = f(\cdot, \mathbf{v}^*) - f(\cdot)$ can be measured by its mean ℓ_2 norm

$$\mathbb{E}_{\mathbb{Q}} \|\mathbf{b}\|^2 = \sum_{i=1}^n \mathbb{E}_{\mathbb{Q}} |f(X_i, \mathbf{v}^*) - f(X_i)|^2.$$

If the design variables X_1, \dots, X_n are i.i.d. then

$$\mathbb{E}_{\mathbb{Q}} \|\mathbf{b}\|^2 = n \mathbb{E}_{\mathbb{Q}} b^2(X_1).$$

Finally, define the empirical information matrix $\tilde{\mathbb{F}}(\mathbf{v}) \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E}\{L(\mathbf{v}) \mid \mathbf{X}\}$:

$$\tilde{\mathbb{F}}(\mathbf{v}) = \frac{1}{\sigma^2} \sum_{i=1}^n F(X_i, \mathbf{v}) \quad (16.35)$$

for

$$F(X_i, \mathbf{v}) \stackrel{\text{def}}{=} \nabla f(X_i, \mathbf{v}) \{ \nabla f(X_i, \mathbf{v}) \}^\top + \{ f(X_i, \mathbf{v}) - f(X_i) \} \nabla^2 f(X_i, \mathbf{v}). \quad (16.36)$$

Also consider the standardized score vector $\boldsymbol{\xi}$

$$\begin{aligned}\boldsymbol{\xi} &= D^{-1} \nabla L(\mathbf{v}^*) = \sigma^{-2} D^{-1} \sum_{i=1}^n \{ Y_i - f(X_i, \mathbf{v}^*) \} \nabla f(X_i, \mathbf{v}^*) \\ &= \sigma^{-2} D^{-1} \sum_{i=1}^n \varepsilon_i \nabla f(X_i, \mathbf{v}^*) + \sigma^{-2} D^{-1} \sum_{i=1}^n \{ f(X_i) - f(X_i, \mathbf{v}^*) \} \nabla f(X_i, \mathbf{v}^*) \\ &= \boldsymbol{\xi}_\varepsilon + \boldsymbol{\xi}_X\end{aligned}$$

with $\varepsilon_i = Y_i - \mathbb{E}(Y_i \mid \mathbf{X}) = Y_i - f(X_i)$.

Condition **(Df2)** has to be slightly extended for the case of a random design.

($\mathbb{Q}f_2$) The function $f(x, \mathbf{v})$ are two times differentiable in $\mathbf{v} \in \Upsilon_{\circ}(\mathbf{r})$, and for some constants w , w_1 , and w_2 , for any $\mathbf{v} \in \Upsilon_{\circ}(\mathbf{r})$, and any unit vector $\mathbf{u} \in \mathbb{R}^p$

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mathbb{Q}} [\mathbf{u}^\top \mathbf{H}^{-1} \nabla f(X_i, \mathbf{v})]^2 &\leq w^2, \\ \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mathbb{Q}} [\mathbf{u}^\top \mathbf{H}^{-1} \nabla f(X_i, \mathbf{v})]^4 &\leq w_1^2, \\ \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mathbb{Q}} [\mathbf{u}^\top \mathbf{H}^{-1} \nabla^2 f(X_i, \mathbf{v}) \mathbf{H}^{-1} \mathbf{u}]^2 &\leq w_2^2. \end{aligned} \quad (16.37)$$

Also a uniform bound on each summand $f(X_i, \mathbf{v})$, $\mathbf{H}^{-1} \nabla f(X_i, \mathbf{v})$, and $\mathbf{H}^{-1} \nabla^2 f(X_i, \mathbf{v}) \mathbf{H}^{-1}$ is required:

$$\begin{aligned} \sup_{\mathbf{v} \in \Upsilon_{\circ}(\mathbf{r})} |f(x, \mathbf{v})| &\leq \bar{w}, \quad \mathbb{Q} - a.s. \\ \sup_{\mathbf{v} \in \Upsilon_{\circ}(\mathbf{r})} \|\mathbf{H}^{-1} \nabla f(x, \mathbf{v})\| &\leq \bar{w}_1, \quad \mathbb{Q} - a.s. \\ \sup_{\mathbf{v} \in \Upsilon_{\circ}(\mathbf{r})} \|\mathbf{H}^{-1} \nabla^2 f(x, \mathbf{v}) \mathbf{H}^{-1}\|_{\text{op}} &\leq \bar{w}_2, \quad \mathbb{Q} - a.s. \end{aligned}$$

The last set of conditions can be relaxed by requiring that the conditions are fulfilled on a set of a high \mathbb{Q} -probability.

16.8.1 Checking the condition (\mathcal{L}_0)

The result of Proposition 16.2.1 continues to apply in the case of a random design.

Proposition 16.8.1. Suppose that condition ($\mathbb{Q}f_2$) holds on $\Upsilon_{\circ}(\mathbf{r})$ for a fixed \mathbf{r} . Then

$$\sup_{\mathbf{v} \in \Upsilon_{\circ}(\mathbf{r})} \|D^{-1} \mathbb{F}(\mathbf{v}) D^{-1} - I_p\|_{\text{op}} \leq \delta(\mathbf{r})$$

with

$$\delta(\mathbf{r}) = \frac{b_1^2 w_2}{n} + \frac{3\sigma \mathbf{r} w_2}{\sqrt{n}} + \frac{3\sigma^2 \mathbf{r}^2 w_2^2}{2n}$$

for $b_1^2 = \mathbb{E}_{\mathbb{Q}} b^2(X_1)$. Under $w_2 \sigma \mathbf{r} \leq 2n^{1/2}$, one can simplify

$$\delta(\mathbf{r}) \stackrel{\text{def}}{=} \frac{b_1^2 w_2}{n} + \frac{6\sigma \mathbf{r} w_2}{\sqrt{n}}.$$

The proof of Proposition 16.2.1 applies without big changes, just add expectation everywhere. Only the result of Lemma 16.2.2 has to be slightly extended to random vectors \mathbf{g} .

Lemma 16.8.1. Let $\mathbf{g}(t)$ be a random continuously differentiable vector function with values in \mathbb{R}^n for $t \in [0, 1]$. Then

$$\mathbb{E}\|\mathbf{g}(1) - \mathbf{g}(0)\|^2 \leq C_g^2$$

with

$$C_g^2 \stackrel{\text{def}}{=} \int_0^1 \|\mathbf{g}'(t)\|^2 dt \leq \sup_{t \in [0,1]} \mathbb{E}\|\mathbf{g}'(t)\|^2.$$

Moreover,

$$\left| \mathbb{E}\|\mathbf{g}(1)\|^2 - \mathbb{E}\|\mathbf{g}(0)\|^2 \right| \leq 2C_g \{\mathbb{E}\|\mathbf{g}(0)\|^2\}^{1/2} + C_g^2.$$

One can also extend the result of Proposition 16.2.3.

Proposition 16.8.2. Let $\mathbf{v}^* = \operatorname{argmax}_{\mathbf{v}} \mathbb{E}L(\mathbf{v})$. Under conditions **(Df₂)** and $w_2 \sigma r \leq 2n^{1/2}$, it holds for $\mathbf{b} = \mathbf{f}(\mathbf{v}^*) - \mathbf{f}^*$ each $\mathbf{v} \in \Upsilon_r(r)$

$$\left| \mathbb{E}_{\mathbb{Q}}\|\mathbf{f}(\mathbf{v}) - \mathbf{f}^*\|^2 - \mathbb{E}_{\mathbb{Q}}\|\mathbf{b}\|^2 - \sigma^2 \|D(\mathbf{v} - \mathbf{v}^*)\|^2 \right| \leq \sigma^2 r^2 \delta(r),$$

for $\delta(r)$ from (16.11). This yields in particular

$$\mathbb{E}_{\mathbb{Q}}\|\mathbf{f}(\mathbf{v}) - \mathbf{f}^*\|^2 \leq \mathbb{E}_{\mathbb{Q}}\|\mathbf{b}\|^2 + \sigma^2 \|D(\mathbf{v} - \mathbf{v}^*)\|^2 + \sigma^2 r^2 \delta(r).$$

Moreover, under $2w_2 \sigma r \leq n^{1/2}$, for any $\mathbf{v} \in \Upsilon_r(r)$

$$\left| \mathbb{E}_{\mathbb{Q}}\|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{v}^*)\|^2 - \sigma^2 \|D(\mathbf{v} - \mathbf{v}^*)\|^2 \right| \leq \frac{3\sigma^3 \|D(\mathbf{v} - \mathbf{v}^*)\|^3 w_2}{2\sqrt{n}},$$

and for any $\mathbf{v}, \mathbf{v}_1 \in \Upsilon_r(r)$

$$\left| \mathbb{E}_{\mathbb{Q}}\|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{v}_1)\|^2 - \sigma^2 \|D(\mathbf{v} - \mathbf{v}_1)\|^2 \right| \leq \frac{6\sigma^3 r^3 w_2}{\sqrt{n}}.$$

To be done: please check

16.8.2 Checking the conditions **(ED₀)** and **(ED₂)**

The stochastic component $\zeta(\mathbf{v}) = L(\mathbf{v}) - \mathbb{E}L(\mathbf{v})$ of the log-likelihood is more involved than in the case of a deterministic design. In view of $Y_i = Y_i - \mathbb{E}(Y_i | \mathbf{X}) + f(X_i) = \varepsilon_i + f(X_i)$ with $\mathbb{E}(\varepsilon_i | \mathbf{X}) = 0$, the log-likelihood function reads as

$$L(\mathbf{v}) = -\frac{1}{2\sigma^2} \sum_{i=1}^n \{\varepsilon_i + f(X_i) - f(X_i, \mathbf{v})\}^2 + R$$

Its gradient and Hessian can be written as

$$\begin{aligned}\nabla L(\boldsymbol{v}) &= \frac{1}{\sigma^2} \sum_{i=1}^n \varepsilon_i \nabla f(X_i, \boldsymbol{v}) - \frac{1}{\sigma^2} \sum_{i=1}^n \{f(X_i, \boldsymbol{v}) - f(X_i)\} \nabla f(X_i, \boldsymbol{v}), \\ \nabla^2 L(\boldsymbol{v}) &= \frac{1}{\sigma^2} \sum_{i=1}^n \varepsilon_i \nabla^2 f(X_i, \boldsymbol{v}) - \tilde{\mathbb{F}}(\boldsymbol{v})\end{aligned}$$

with $\tilde{\mathbb{F}}(\boldsymbol{v})$ from (16.35). This implies with $b(X_i) = f(X_i, \boldsymbol{v}^*) - f(X_i)$

$$\begin{aligned}\nabla \zeta(\boldsymbol{v}^*) &= \nabla L(\boldsymbol{v}^*) = \frac{1}{\sigma^2} \sum_{i=1}^n \varepsilon_i \nabla f(X_i, \boldsymbol{v}^*) - \frac{1}{\sigma^2} \sum_{i=1}^n b(X_i) \nabla f(X_i, \boldsymbol{v}^*) \\ &= \nabla_\varepsilon + \nabla_X, \\ \nabla^2 \zeta(\boldsymbol{v}) &= \frac{1}{\sigma^2} \sum_{i=1}^n \varepsilon_i \nabla^2 f(X_i, \boldsymbol{v}) - \tilde{\mathbb{F}}(\boldsymbol{v}) + \mathbb{F}(\boldsymbol{v}).\end{aligned}$$

For the case of a random design, the score vector $\nabla \zeta(\boldsymbol{v}^*)$ can be represented as the sum of two other vectors ∇_ε and ∇_X . The second vector ∇_X reflects the impact of a random design. The i.i.d. structure of the design yields

$$\log \mathbb{E} \exp\{\lambda \mathbf{u}^\top D^{-1} \nabla_X\} = n \log \mathbb{E}_Q \exp\left\{\frac{\lambda}{\sigma\sqrt{n}} \mathbf{u}^\top U_1(X_1)\right\} \quad (16.38)$$

with

$$U_1(X_1) \stackrel{\text{def}}{=} b(X_1) \mathbf{H}^{-1} \nabla f(X_1, \boldsymbol{v}^*).$$

Definition of \boldsymbol{v}^* yields $\mathbb{E}_Q U_1(X_1) = 0$. We use the following technical fact.

Lemma 16.8.2. *Let ξ be a zero mean r.v. with $|\xi| \leq \bar{w}$. For any $n \geq 1$ and λ with $\lambda^2 \bar{w}^2 \leq n$,*

$$n \log \mathbb{E} \exp\left(\frac{\lambda \xi}{\sqrt{n}}\right) \leq \frac{\lambda^2 \mathbb{E} |\xi|^2}{2} + \frac{|\lambda|^3 \mathbb{E} |\xi|^3}{2\sqrt{n}} \leq \lambda^2 \mathbb{E} \xi^2. \quad (16.39)$$

Also

$$n \log \mathbb{E} \exp\left(\frac{\lambda^2 \xi^2}{n}\right) \leq 3\lambda^2 \mathbb{E} |\xi|^2. \quad (16.40)$$

Proof. The bound $\lambda \bar{w} \leq \sqrt{n}$ implies by the Taylor expansion of the third order

$$\exp\left(\frac{\lambda \xi}{\sqrt{n}}\right) - 1 - \frac{\lambda \xi}{\sqrt{n}} - \frac{\lambda^2 \xi^2}{2n} \leq \frac{\lambda^3 |\xi|^3}{6n^{3/2}} \exp\left(\frac{\lambda \bar{w}}{\sqrt{n}}\right) \leq \frac{\lambda^3 |\xi|^3}{2n^{3/2}}.$$

The same applies to expectations, and the use of $\mathbb{E} \xi = 0$ yields

$$\mathbb{E} \exp\left(\frac{\lambda \xi}{\sqrt{n}}\right) - 1 \leq \frac{\lambda^2 \mathbb{E} \xi^2}{2n} + \frac{\lambda^3 \mathbb{E} |\xi|^3}{2n^{3/2}} \leq \frac{\lambda^2 \mathbb{E} \xi^2}{2n} \left(1 + \frac{\lambda \bar{w}}{\sqrt{n}}\right) \leq \frac{\lambda^2 \mathbb{E} \xi^2}{n}.$$

The statement (16.39) of the lemma follows by $\log(1+x) \leq x$ for $x \geq 0$. The second bound (16.40) is obtained similarly by the Taylor expansion of the first order: for $\lambda^2 \bar{w}^2 \leq n$, it holds

$$\exp\left(\frac{\lambda^2 \xi^2}{n}\right) \leq 1 + \frac{\lambda^2 \xi^2}{n} \exp\left(\frac{\lambda^2 \bar{w}^2}{n}\right) \leq 1 + \frac{3\lambda^2 \xi^2}{n}.$$

This lemma and the bound $|b(x)| \leq \bar{b}$ yield that $\sup_x \|U_1(x)\| \leq \bar{b} w_1$ and for any λ with $\lambda \bar{b} w_1 \leq \sigma \sqrt{n}$

$$n \log I\!\!E_{\mathbb{Q}} \exp\left\{\frac{\lambda}{\sigma \sqrt{n}} \mathbf{u}^\top U_1(X_1)\right\} \leq \frac{\lambda^2}{\sigma^2} \text{Var}\{\mathbf{u}^\top U_1(X_1)\}.$$

This implies for any unit vector \mathbf{u}

$$\begin{aligned} \text{Var}\{\mathbf{u}^\top U_1(X_1)\} &\leq I\!\!E_{\mathbb{Q}} |b(X_1) \mathbf{u}^\top H^{-1} \nabla f(X_1, \mathbf{v}^*)|^2 \\ &\leq \bar{b}^2 I\!\!E_{\mathbb{Q}} \{\mathbf{u}^\top H^{-1} \nabla f(X_1, \mathbf{v}^*) \{\nabla f(X_1, \mathbf{v}^*)\}^\top H^{-1} \mathbf{u}\} = \bar{b}^2. \end{aligned}$$

This implies

$$\log I\!\!E \exp\{\lambda \mathbf{u}^\top D^{-1} \nabla_X\} \leq \frac{\lambda^2 \bar{b}^2}{\sigma^2} \ll \frac{\lambda^2}{2} \quad (16.41)$$

if \bar{b} is small relative to σ .

Now we consider the term ∇_ε . Given the design \mathbf{X} , it behaves as in the case of a deterministic design, that is, for Gaussian errors ε_i , the vector ∇_ε is also Gaussian zero mean with the covariance

$$\begin{aligned} \text{Var}(\nabla_\varepsilon \mid \mathbf{X}) &= \frac{1}{\sigma^4} \text{Var}\left\{\sum_{i=1}^n \varepsilon_i \nabla f(X_i, \mathbf{v}^*) \mid \mathbf{X}\right\} \\ &= \frac{1}{\sigma^4} \sum_{i=1}^n \sigma_i^2 \nabla f(X_i, \mathbf{v}^*) \nabla f(X_i, \mathbf{v}^*)^\top = \tilde{V}^2. \end{aligned} \quad (16.42)$$

So, if $\tilde{V}^2 \leq (1 + \delta_V) V^2$, then for any $\boldsymbol{\gamma} \in \mathbb{R}^p$

$$\begin{aligned} I\!\!E \exp\left(\boldsymbol{\gamma}^\top V^{-1} \nabla_\varepsilon\right) &= I\!\!E_{\mathbb{Q}} I\!\!E\{\exp(\boldsymbol{\gamma}^\top V^{-1} \nabla_\varepsilon) \mid \mathbf{X}\} \\ &= I\!\!E_{\mathbb{Q}} \exp\left(\frac{\boldsymbol{\gamma}^\top V^{-1} \tilde{V}^2 V^{-1} \boldsymbol{\gamma}}{2}\right) \leq \exp\left(\frac{(1 + \delta_V) \|\boldsymbol{\gamma}\|^2}{2}\right). \end{aligned}$$

This inequality and (16.41) imply **(ED₀)** with the matrix V^2 under the condition that \bar{b}/σ is small.

Now we check **(ED₂)**. Let \mathbf{u} be a unit vector in \mathbb{R}^p and $\boldsymbol{\gamma} = D^{-1} \mathbf{u}$. Similarly to (16.42), the exponential moments of the sum

$$\nabla_{\varepsilon}^2(\mathbf{v}) \stackrel{\text{def}}{=} \frac{1}{\sigma^2} \sum_{i=1}^n \varepsilon_i \nabla^2 f(X_i, \mathbf{v})$$

can be bonded by (16.40) of Lemma 16.8.2. For any $\lambda > 0$ with $\lambda^2 C_V^2 \bar{w}_2^2 \leq n$ and $\omega = \sigma/\sqrt{n}$, it follows from (16.37)

$$\begin{aligned} \log I\!E \exp\left(\frac{\lambda}{\omega} \gamma^\top \nabla_{\varepsilon}^2(\mathbf{v}) \gamma\right) &= \log I\!E_{\mathbb{Q}} I\!E \left\{ \exp\left(\frac{\lambda}{\omega} \gamma^\top \nabla_{\varepsilon}^2(\mathbf{v}) \gamma\right) \mid \mathbf{X} \right\} \\ &= \log I\!E_{\mathbb{Q}} \exp\left\{ \frac{\lambda^2}{2n} \sum_{i=1}^n \frac{\sigma_i^2}{\sigma^2} \left| \mathbf{u}^\top \mathbf{H}^{-1} \nabla^2 f(X_i, \mathbf{v}) \mathbf{H}^{-1} \mathbf{u} \right|^2 \right\} \\ &\leq \log I\!E_{\mathbb{Q}} \exp\left\{ \frac{\lambda^2 C_V^2}{2n} \sum_{i=1}^n \left| \mathbf{u}^\top \mathbf{H}^{-1} \nabla^2 f(X_i, \mathbf{v}) \mathbf{H}^{-1} \mathbf{u} \right|^2 \right\} \\ &\leq \frac{3\lambda^2 C_V^2}{2n} I\!E_{\mathbb{Q}} \sum_{i=1}^n \left| \mathbf{u}^\top \mathbf{H}^{-1} \nabla^2 f(X_i, \mathbf{v}) \mathbf{H}^{-1} \mathbf{u} \right|^2 \leq \frac{3\lambda^2 C_V^2 w_2^2}{2}. \end{aligned}$$

Further, the i.i.d. design structure and the definition (16.35) yield for any unit vector \mathbf{u} and $\gamma = D^{-1}\mathbf{u} = \sigma n^{-1/2} \mathbf{H}^{-1} \mathbf{u}$

$$\begin{aligned} \log I\!E \exp\left\{ \frac{\lambda}{\omega} \gamma^\top \{\tilde{\mathbb{F}}(\mathbf{v}) - \mathbb{F}(\mathbf{v})\} \gamma \right\} \\ = n \log I\!E \exp\left\{ \frac{\lambda}{\sqrt{n}} \mathbf{u}^\top \mathbf{H}^{-1} \left\{ F(X_1, \mathbf{v}) - I\!E_{\mathbb{Q}} F(X_1, \mathbf{v}) \right\} \mathbf{H}^{-1} \mathbf{u} \right\} \end{aligned}$$

for $F(X_1, \mathbf{v})$ from (16.36). To apply Lemma 16.8.2 we first bound the maximum of $F(x, \mathbf{v})$. Let us fix a unit vector $\mathbf{u} \in \mathcal{S}_p$ and $\mathbf{v} \in \Upsilon_{\circ}(\mathbf{r})$, so that $\|\alpha\| = \|\mathbf{H}(\mathbf{v} - \mathbf{v}^*)\| \leq \mathbf{r}\sigma/\sqrt{n}$. Let also x be a point in the design space. The bound $\|\mathbf{H}^{-1} \nabla^2 f(x, \mathbf{v}) \mathbf{H}^{-1}\|_{\text{op}} \leq \bar{w}_2$ implies

$$\begin{aligned} |\mathbf{u}^\top \mathbf{H}^{-1} \{\nabla f(x, \mathbf{v}) - \nabla f(x, \mathbf{v}^*)\}| &= |\mathbf{u}^\top \mathbf{H}^{-1} \nabla^2 f(x, \mathbf{v}^*)(\mathbf{v} - \mathbf{v}^*)| \\ &= |\mathbf{u}^\top \mathbf{H}^{-1} \nabla^2 f(x, \mathbf{v}^*) \mathbf{H}^{-1} \alpha| \leq \frac{\bar{w}_2 \mathbf{r} \sigma}{\sqrt{n}}, \end{aligned}$$

where \mathbf{v}^* is a point on the line between \mathbf{v} and \mathbf{v}^* . For the variance, similar arguments yield

$$\begin{aligned} I\!E_{\mathbb{Q}} \left[\mathbf{u}^\top \mathbf{H}^{-1} \{\nabla f(X_1, \mathbf{v}) - \nabla f(X_1, \mathbf{v}^*)\} \right]^2 &= I\!E_{\mathbb{Q}} \left[\mathbf{u}^\top \mathbf{H}^{-1} \nabla^2 f(X_1, \mathbf{v}^*)(\mathbf{v} - \mathbf{v}^*) \right]^2 \\ &= I\!E_{\mathbb{Q}} \left[\mathbf{u}^\top \mathbf{H}^{-1} \nabla^2 f(X_1, \mathbf{v}^*) \mathbf{H}^{-1} \alpha \right]^2 \leq \frac{w_2^2 \mathbf{r}^2 \sigma^2}{n}, \end{aligned}$$

and by definition of \mathbf{H}^2

$$\begin{aligned}
& \mathbb{E}_{\mathbb{Q}} \left[\mathbf{u}^\top \mathbf{H}^{-1} \nabla f(X_1, \mathbf{v}) \right]^2 \\
& \leq 2 \mathbb{E}_{\mathbb{Q}} \left[\mathbf{u}^\top \mathbf{H}^{-1} \nabla f(X_1, \mathbf{v}^*) \right]^2 + 2 \mathbb{E}_{\mathbb{Q}} \left[\mathbf{u}^\top \mathbf{H}^{-1} \{ \nabla f(X_1, \mathbf{v}) - \nabla f(X_1, \mathbf{v}^*) \} \right]^2 \\
& \leq 2 + \frac{2w_2^2 \mathbf{r}^2 \sigma^2}{n}.
\end{aligned}$$

Further,

$$\begin{aligned}
|f(x, \mathbf{v}) - f(x)| &= |b(x) + f(x, \mathbf{v}) - f(x, \mathbf{v}^*)| \\
&\leq |b(x)| + |\boldsymbol{\alpha}^\top \mathbf{H}^{-1} \nabla f(x, \mathbf{v}^*)| \\
&\leq \bar{b} + \|\boldsymbol{\alpha}\| \|\mathbf{H}^{-1} \nabla f(x, \mathbf{v}^*)\| \leq \bar{b} + \frac{\bar{w}_1 \mathbf{r} \sigma}{\sqrt{n}}.
\end{aligned} \tag{16.43}$$

Under the condition that $\bar{w}_1 \sqrt{n} \geq \bar{w}_2 \mathbf{r} \sigma$, the definition (16.36) of $F(x, \mathbf{v})$ implies now

$$\begin{aligned}
& |\mathbf{u}^\top \mathbf{H}^{-1} F(x, \mathbf{v}) \mathbf{H}^{-1} \mathbf{u}| \\
& \leq |\mathbf{u}^\top \mathbf{H}^{-1} \nabla f(x, \mathbf{v})|^2 + |f(x, \mathbf{v}) - f(x)| |\mathbf{u}^\top \mathbf{H}^{-1} \nabla^2 f(x, \mathbf{v}) \mathbf{H}^{-1} \mathbf{u}| \\
& \leq \bar{w}_1^2 + \left(\bar{b} + \frac{\bar{w}_1 \mathbf{r} \sigma}{\sqrt{n}} \right) \bar{w}_2 \leq 2\bar{w}_1^2 + \bar{b} \bar{w}_2.
\end{aligned} \tag{16.44}$$

In a similar way we bound the variance of $\mathbf{u}^\top \mathbf{H}^{-1} F(X_1, \mathbf{v}) \mathbf{H}^{-1} \mathbf{u}$. It holds similarly to (16.43)

$$\begin{aligned}
& \mathbb{E}_{\mathbb{Q}} |f(X_1, \mathbf{v}) - f(X_1)| |\mathbf{u}^\top \mathbf{H}^{-1} \nabla^2 f(X_1, \mathbf{v}) \mathbf{H}^{-1} \mathbf{u}| \\
& \leq \mathbb{E}_{\mathbb{Q}} \left\{ |b(X_1) + \boldsymbol{\alpha}^\top \mathbf{H}^{-1} \nabla f(X_1, \mathbf{v}^*)| |\mathbf{u}^\top \mathbf{H}^{-1} \nabla^2 f(X_1, \mathbf{v}) \mathbf{H}^{-1} \mathbf{u}| \right\} \\
& \leq b_1 w_2 + \frac{\mathbf{r} \sigma w w_2}{\sqrt{n}}.
\end{aligned}$$

By (16.44) and the definition (16.36) of $F(X_1, \mathbf{v})$

$$\begin{aligned}
& \mathbb{E}_{\mathbb{Q}} |\mathbf{u}^\top \mathbf{H}^{-1} F(X_1, \mathbf{v}) \mathbf{H}^{-1} \mathbf{u}|^2 \leq 2 \operatorname{Var} \{ |\mathbf{u}^\top \mathbf{H}^{-1} \nabla f(X_1, \mathbf{v})|^2 \} \\
& + 2 \mathbb{E}_{\mathbb{Q}} [\{ f(X_1, \mathbf{v}) - f(X_1) \} \mathbf{u}^\top \mathbf{H}^{-1} \nabla^2 f(X_1, \mathbf{v}) \mathbf{H}^{-1} \mathbf{u}]^2 \\
& \leq 2w_1^2 + 2 \left(b_1 w_2 + \frac{\mathbf{r} \sigma w w_2}{\sqrt{n}} \right).
\end{aligned}$$

To be done: complete and state the prrecise result

Structural regression

This chapter continues the study of parametric regression models. Below we focus on regression equations under special structural assumptions. Examples of such structures include single- and multi-index models, Projection Pursuit models, etc. We also consider regression with error-in-variables and instrumental regression.

17.1 Single-index case

This section specifies the general results for parametric and nonparametric regression to the special case of a single-index regression model

$$Y_i = f(X_i) + \varepsilon_i = g(X_i^\top \boldsymbol{\theta}^*) + \varepsilon_i,$$

where the likelihood is built for the case when the errors ε_i are supposed to be i.i.d. normal $\mathcal{N}(0, \sigma^2)$ and X_1, \dots, X_n are given design points in \mathbb{R}^p supported to a bounded set \mathfrak{X} . The results will be stated under mild assumptions on the errors admitting a inhomogeneous and non-Gaussian noise.

The unit index vector $\boldsymbol{\theta}^* \in \mathbb{R}^p$, $\|\boldsymbol{\theta}^*\| = 1$, is our target while the unknown link function $g(\cdot)$ is a nuisance parameter. We assume that $g(u)$ is sufficiently smooth and can be approximated by the sums

$$g(u) \approx g(u; \boldsymbol{\eta}) = \sum_{m=1}^q \eta_m \psi_m(u) = \boldsymbol{\eta}^\top \Psi(u)$$

for given univariate basis functions $\psi_1(\cdot), \dots, \psi_q(\cdot)$. The corresponding sieve log-likelihood reads as

$$L(\boldsymbol{v}) = -\frac{1}{2\sigma^2} \sum_{i=1}^n |Y_i - f_i(\boldsymbol{v})|^2 + R,$$

where $\boldsymbol{v} = (\boldsymbol{\theta}, \boldsymbol{\eta})$ and

$$f_i(\boldsymbol{v}) \stackrel{\text{def}}{=} \sum_{m=1}^q \eta_m \psi_m(X_i^\top \boldsymbol{\theta}) = \boldsymbol{\eta}^\top \Psi(X_i^\top \boldsymbol{\theta}) = g(X_i^\top \boldsymbol{\theta}; \boldsymbol{\eta}).$$

Here $\Psi(u) = (\psi_1(u), \dots, \psi_q(u))^\top$ is the vectors of values of basis functions $\psi_m(\cdot)$ at the point u . The definition implies

$$\begin{aligned} \nabla f_i(\boldsymbol{v}) &= \begin{pmatrix} \boldsymbol{\eta}^\top \Psi'(X_i^\top \boldsymbol{\theta}) X_i \\ \Psi(X_i^\top \boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} g'(X_i^\top \boldsymbol{\theta}; \boldsymbol{\eta}) X_i \\ \Psi(X_i^\top \boldsymbol{\theta}) \end{pmatrix}, \\ \nabla^2 f_i(\boldsymbol{v}) &= \begin{pmatrix} \boldsymbol{\eta}^\top \Psi''(X_i^\top \boldsymbol{\theta}) X_i X_i^\top & X_i \Psi'(X_i^\top \boldsymbol{\theta})^\top \\ \Psi'(X_i^\top \boldsymbol{\theta}) X_i^\top & 0 \end{pmatrix} \\ &= \begin{pmatrix} g''(X_i^\top \boldsymbol{\theta}; \boldsymbol{\eta}) X_i X_i^\top & X_i \Psi'(X_i^\top \boldsymbol{\theta})^\top \\ \Psi'(X_i^\top \boldsymbol{\theta}) X_i^\top & 0 \end{pmatrix}. \end{aligned}$$

Define

$$\mathsf{H}_{\boldsymbol{\theta}}^2 = \frac{1}{n} \sum_{i=1}^n [g'(X_i^\top \boldsymbol{\theta}^*; \boldsymbol{\eta}^*)]^2 X_i X_i^\top = \mathbb{E}_X \{ [g'(X^\top \boldsymbol{\theta}^*; \boldsymbol{\eta}^*)]^2 X X^\top \}. \quad (17.1)$$

(\mathbb{E}_X means averaging w.r.t. the design measure.) Similarly

$$\begin{aligned} \mathsf{A}_{\boldsymbol{\theta}\boldsymbol{\eta}} &= \mathbb{E}_X \{ g'(X^\top \boldsymbol{\theta}^*; \boldsymbol{\eta}^*) X \Psi(X^\top \boldsymbol{\theta}^*)^\top \}, \\ \mathsf{H}_{\boldsymbol{\eta}}^2 &= \mathbb{E}_X \{ \Psi(X^\top \boldsymbol{\theta}^*) \Psi(X^\top \boldsymbol{\theta}^*)^\top \}. \end{aligned} \quad (17.2)$$

Then one can represent the matrix H^2 in the block form

$$\mathsf{H}^2 = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\boldsymbol{v}^*) \{ \nabla f_i(\boldsymbol{v}^*) \}^\top = \begin{pmatrix} \mathsf{H}_{\boldsymbol{\theta}}^2 & \mathsf{A}_{\boldsymbol{\theta}\boldsymbol{\eta}} \\ \mathsf{A}_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top & \mathsf{H}_{\boldsymbol{\eta}}^2 \end{pmatrix}.$$

Required conditions:

(S₀) [Regular noise] The errors $\varepsilon_i = Y_i - \mathbb{E}Y_i$ are independent (sub-)Gaussian and satisfy

$$\min_i \sigma_i^2 \geq C_V^{-2} \sigma^2, \quad \max_i \sigma_i^2 \leq C_V^2 \sigma^2.$$

(S₁) [Regular design] For some fixed positive constants C_g , C_Ψ , and C_X , it holds

$$\begin{aligned} \mathsf{H}_{\boldsymbol{\theta}}^2 &= \mathbb{E}_X \{ [g'(X^\top \boldsymbol{\theta}^*; \boldsymbol{\eta}^*)]^2 X X^\top \} \geq C_g^{-2} V_X^2, \\ \mathsf{H}_{\boldsymbol{\eta}}^2 &= \mathbb{E}_X \{ \Psi(X^\top \boldsymbol{\theta}^*) \Psi(X^\top \boldsymbol{\theta}^*)^\top \} \geq C_\Psi^{-2} I_q > 0. \end{aligned}$$

with $V_X^2 = \mathbb{E}_X(X X^\top)$ and for any unit vector $\boldsymbol{u} \in \mathcal{S}_p$

$$\frac{1}{n} \sum_{i=1}^n |\mathbf{u}^\top V_X^{-1} X_i|^4 \leq C_X^4.$$

Note that this condition implies for any $\mathbf{u}, \mathbf{u}_1 \in \mathcal{S}_p$ by the Cauchy-Schwartz inequality

$$\frac{1}{n} \sum_{i=1}^n |\mathbf{u}^\top V_X^{-1} X_i|^2 |\mathbf{u}_1^\top V_X^{-1} X_i|^2 \leq C_X^4.$$

(*S*₂) [Ψ' -basis] For some constants $\mu_1 \leq \dots \leq \mu_q$

$$\|\psi'_m(\cdot)\|_\infty \leq \mu_m;$$

(*S*₃) [Smooth vector $\boldsymbol{\eta}$] The nuisance parameter $\boldsymbol{\eta}$ belongs to a set H and for each $\boldsymbol{\eta} \in H$, it holds

$$\|g^{(s)}(\cdot; \boldsymbol{\eta})\|_\infty = \|\boldsymbol{\eta}^\top \Psi^{(s)}(\cdot)\|_\infty \leq C_{g,s}, \quad s = 0, 1, 2;$$

(*S*₄) [Critical dimension] The value $\mu_q \sqrt{(p+q)/n}$ is small;

(*S*₅) [Small modeling bias] The value Δ_f defined as

$$\Delta_f^2 = \frac{1}{n} \sum_{i=1}^n |f(X_i) - g(X_i^\top \boldsymbol{\theta}^*, \boldsymbol{\eta}^*)|^2$$

satisfies the condition “ $\mu_q \Delta_f$ is small”.

(*S*₆) [Identifiability] For $\nu < 1$, it holds with $\mathsf{F}_{\boldsymbol{\theta}}$ and $\mathsf{F}_{\boldsymbol{\eta}}$ from (17.1) and (17.2)

$$\mathbb{E}_X \{ g'(X^\top \boldsymbol{\theta}^*; \boldsymbol{\eta}^*) X \Psi(X^\top \boldsymbol{\theta}^*)^\top \} \leq \nu \mathsf{H}_{\boldsymbol{\theta}} \mathsf{H}_{\boldsymbol{\eta}}.$$

Now we check the conditions from (16.29). Assumption (*S*₃) implies for any unit vector $\mathbf{u} \in \mathcal{S}_p$

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n |\mathbf{u}^\top \mathsf{H}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 f_i(\mathbf{v}) \mathsf{H}_{\boldsymbol{\theta}}^{-1} \mathbf{u}|^2 &= \frac{1}{n} \sum_{i=1}^n |\boldsymbol{\eta}^\top \Psi''(X_i^\top \boldsymbol{\theta})|^2 |\mathbf{u}^\top \mathsf{H}_{\boldsymbol{\theta}}^{-1} X_i|^4 \\ &\leq \frac{C_{g,2}^2}{n} \sum_{i=1}^n |\mathbf{u}^\top \mathsf{H}_{\boldsymbol{\theta}}^{-1} X_i|^4. \end{aligned}$$

Further, by (*S*₁)

$$\sup_{\mathbf{u} \in \mathcal{S}_p} \sum_{i=1}^n |\mathbf{u}^\top \mathsf{H}_{\boldsymbol{\theta}}^{-1} X_i|^4 \leq C_g^4 \sup_{\mathbf{u} \in \mathcal{S}_p} \sum_{i=1}^n |\mathbf{u}^\top V_X^{-1} X_i|^4 \leq n C_g^4 C_X^4. \quad (17.3)$$

This yields

$$\frac{1}{n} \sum_{i=1}^n |\mathbf{u}^\top \mathsf{H}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 f_i(\mathbf{v}) \mathsf{H}_{\boldsymbol{\theta}}^{-1} \mathbf{u}|^2 \leq C_{g,2}^2 C_g^4 C_X^4.$$

Similarly, for any unit vectors $\mathbf{u} \in \mathcal{S}_p$ and $\mathbf{w} \in \mathcal{S}_q$, it holds by **(S₁)**

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n |\mathbf{u}^\top \mathsf{H}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}\boldsymbol{\eta}}^2 f_i(\mathbf{v}) \mathsf{H}_{\boldsymbol{\eta}}^{-1} \mathbf{w}|^2 &= \frac{1}{n} \sum_{i=1}^n |\mathbf{u}^\top \mathsf{H}_{\boldsymbol{\theta}}^{-1} X_i|^2 |\mathbf{w}^\top \mathsf{H}_{\boldsymbol{\eta}}^{-1} \Psi'(X_i^\top \boldsymbol{\theta})|^2 \\ &\leq \left\{ \frac{1}{n} \sum_{i=1}^n |\mathbf{u}^\top \mathsf{H}_{\boldsymbol{\theta}}^{-1} X_i|^4 \right\}^{1/2} \left\{ \frac{1}{n} \sum_{i=1}^n |\mathbf{w}^\top \mathsf{H}_{\boldsymbol{\eta}}^{-1} \Psi'(X_i^\top \boldsymbol{\theta})|^4 \right\}^{1/2} \end{aligned}$$

and the use of **(S₁)** and **(S₂)** yields for each $i \leq n$

$$|\mathbf{w}^\top \mathsf{H}_{\boldsymbol{\eta}}^{-1} \Psi'(X_i^\top \boldsymbol{\theta})|^2 \leq C_\Psi^2 \|\mathbf{w}\|^2 \|\Psi'(X_i^\top \boldsymbol{\theta})\|^2 \leq C_\Psi^2 \mu_q^2 q.$$

Together with (17.3) this implies

$$\frac{1}{n} \sum_{i=1}^n |\mathbf{u}^\top \mathsf{H}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}\boldsymbol{\eta}}^2 f_i(\mathbf{v}) \mathsf{H}_{\boldsymbol{\eta}}^{-1} \mathbf{w}|^2 \leq q \mu_q^2 C_g^2 C_X^2 C_\Psi^2.$$

As $\nabla_{\boldsymbol{\eta}\boldsymbol{\eta}}^2 f_i(\mathbf{v}) = 0$, the condition (16.5) is fulfilled with w_2 given by

$$w_2^2 = \frac{1}{(1-\nu)^2} (C_{g,2}^2 C_g^4 C_X^4 + 2q \mu_q^2 C_g^2 C_X^2 C_\Psi^2).$$

Regularity of the design \mathbf{X} and Ψ and of the function g means that $C_{g,2}$, C_g , C_X , C_Ψ are all bounded and do not grow with the sample size, dimensions p and q etc. In the contrary, the value μ_q depends heavily on q , in many examples it is just of order q . So, the value w_2^2 is also of order $q \mu_q^2$.

The next theorem specifies the general results of Theorems 16.3.1 and 16.7.1 to the single-index case. The total Fisher information $\mathbb{F}(\mathbf{v}^*)$ is defined as in (16.8): $\mathbb{F}(\mathbf{v}^*) = -\nabla^2 \mathbb{E} L(\mathbf{v}^*)$. It coincides with $D^2 = \mathbb{F}_0$ if the model assumption $\mathbb{E} Y_i = g(X_i^\top \boldsymbol{\theta}^*, \boldsymbol{\eta}^*) = \Psi(X_i^\top \boldsymbol{\theta}^*)^\top \boldsymbol{\eta}^*$ for all i is correct. The MLE $\tilde{\mathbf{v}} = (\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}})$ is defined by maximizing $L(\mathbf{v})$ yielding the estimate $\tilde{\boldsymbol{\theta}}$ of the index vector $\boldsymbol{\theta}^*$ and the function estimate \tilde{f} with

$$\tilde{f}(X) = g(X^\top \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}) = \tilde{\boldsymbol{\eta}}^\top \Psi(X^\top \tilde{\boldsymbol{\theta}}).$$

Theorem 17.1.1. Suppose **(S₀)** through **(S₆)**. Then the choice $h_0^2 = C(q+x)$ and $r_0^2 = C(p+x)$ ensures for $\Upsilon(r_0, h_0) = \Theta_0(r_0) \times \mathcal{H}_0(h_0)$

$$\mathbb{P}(\tilde{\mathbf{v}} \notin \Upsilon(r_0, h_0)) = \mathbb{P}(\tilde{\boldsymbol{\theta}} \notin \Theta_0(r_0) \text{ or } \tilde{\boldsymbol{\eta}} \notin \mathcal{H}_0(h_0)) \leq 4e^{-x}.$$

Moreover, under $\tilde{\mathbf{v}} \in \Upsilon(r_0, h_0)$, it holds

$$\|D(\tilde{\mathbf{v}} - \mathbf{v}^*) - \boldsymbol{\xi}\| \leq \diamond(x) = C w_2 (p+q+x)\sigma/\sqrt{n}$$

with $\boldsymbol{\xi} \stackrel{\text{def}}{=} D^{-1} \nabla L(\mathbf{v}^*)$. For the $\boldsymbol{\theta}$ -component of \mathbf{v} , this yields

$$\|\mathbb{I}^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \check{\boldsymbol{\xi}}\| \leq \diamond(x) \quad (17.4)$$

with \mathbb{I} and $\check{\boldsymbol{\xi}}$ from (16.32).

Furthermore, on a set of probability at least $1 - 4e^{-x}$

$$\begin{aligned} \sigma^{-2}\|\tilde{\mathbf{f}}(\mathbf{X}) - \mathbf{f}^*\|^2 &\leq \sigma^{-2}\|g(\mathbf{X}^\top \boldsymbol{\theta}^*, \boldsymbol{\eta}^*) - \mathbf{f}^*\|^2 + \|\boldsymbol{\xi}\|^2 + \Delta_{p,q}, \\ \Delta_{p,q} &= C w_2 \sigma \sqrt{(p+q+x)^3/n}. \end{aligned}$$

Identifiability issue for the index vector

The representation $f(\mathbf{x}) = g(\mathbf{x}^\top \boldsymbol{\theta}^*)$ has a drawback of being non-identifiable. Indeed, one can rescale the vector $\boldsymbol{\theta}^*$ and the function g to get the same resulting function f . Usually one imposes an additional assumption to make the problem well specified. Note that the assumption $\|\boldsymbol{\theta}^*\| = 1$ is not sufficient for identifiability: one can multiply $\boldsymbol{\theta}^*$ by -1 and reflect the function g around the origin. A popular solution is to require that $\|\boldsymbol{\theta}\| = 1$ and the first component of $\boldsymbol{\theta}^*$ is non-negative. This leads to the parameter set Θ in the form of a subset of the unit sphere in \mathbb{R}^p . Such an approach works well in theory but is not practical because the unit sphere is not a convex set. Alternatively, one can fix $\theta_1 \equiv 1$, that is, set the first component of $\boldsymbol{\theta}$ equal to one. Then the remaining components can take any values, model becomes identifiable. This approach assumes that θ_1 in the original model is non-zero, that is, the factor X_1 is significant.

One more method to achieve identifiability is to add a penalty term $\lambda(\|\boldsymbol{\theta}\|^2 - 1)^2$ and look for a solution in the half-space $\Theta = \{\boldsymbol{\theta}: \theta_1 \geq 0\}$.

Sobolev smoothness of the link function and the use of cosine basis

For a special case of the cosine-basis $\psi_m(u) = \cos(mu)$, condition (S_2) meets with $\mu_{m,s} = m^s$. Furthermore, condition (S_3) requires that the sum $\sum_m \eta_m m^s \cos(mu)$ is bounded in sup-norm. ℓ_2 -norm convergence of this sum requires

$$\sum_m \eta_m^2 m^{2s} \leq C$$

which is equivalent to saying that $g(\cdot)$ belongs to a Sobolev class with the smoothness parameter β larger than s . Moreover, if

$$\sum_m \eta_m^2 m^{2s+1} \leq C^2$$

($\beta = s + 1/2$) then the sum $\sum_m \eta_m m^s \cos(mu)$ is bounded in the sup-norm by the Sobolev embedding theorem.

Sobolev smoothness of g of degree β implies for $\boldsymbol{\eta}^* = \boldsymbol{\eta}_q^*$

$$\begin{aligned}\|g(\cdot) - g(\cdot, \boldsymbol{\eta}^*)\| &\leq Cq^{-\beta}, \\ \|g(\cdot) - g(\cdot, \boldsymbol{\eta}^*)\|_\infty &\leq Cq^{-\beta+1/2}.\end{aligned}$$

The latter bound implies $\Delta_f \leq q^{-\beta}$. The condition “ $\mu_q \Delta_f$ small” requires $q^{-\beta+1}$ small. In view of $\psi_m''(u) = -m^2\psi_m(u)$, the condition “ $\|g^{(2)}(\cdot; \boldsymbol{\eta})\|_\infty \leq C_{g,2}$ ” is fulfilled under

$$\sum_m \eta_m^2 m^5 < \infty.$$

The root-n consistency results holds if

$$q\sqrt{(p+q)/n} \text{ is small.}$$

The Fisher expansion and root-n normality of $\mathbb{I}^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ require a stronger condition

$$q(p+q)/\sqrt{n} \text{ is small.}$$

Index-vector estimation under smoothness of g

Now we can summarize the above consideration. We assume the regularity conditions (S_0) through (S_2) and sufficient amount of smoothness of the link function g which ensures (S_3) . The main result indicates that the cut-off parameter q can be properly selected to ensure the conditions (S_4) and (S_5) simultaneously and thus, to show the properties of the index vector estimator $\tilde{\boldsymbol{\theta}}$.

Theorem 17.1.2. Suppose regularity conditions (S_0) through (S_2) . Let f follow $f(\mathbf{x}) = g(\mathbf{x}^\top \boldsymbol{\theta}^*)$, where a link function $g(\cdot)$ is Sobolev-smooth of degree $\beta = 2.5$. Fix $q = n^{1/5}$ for the sieve estimate $\tilde{\mathbf{v}}$. Then $\tilde{\boldsymbol{\theta}}$ follows the Fisher expansion (17.4) with a small term $\diamond(r_0)$. This implies that $\tilde{\boldsymbol{\theta}}$ is asymptotically normal as $n \rightarrow \infty$ and, if the errors ε_i are homogeneous, then $\tilde{\boldsymbol{\theta}}$ is asymptotically efficient. The quadratic loss $\|\mathbf{f}(\tilde{\mathbf{v}}) - \mathbf{f}^*\|^2$ satisfies

$$\sigma^{-2} \|\mathbf{f}(\tilde{\mathbf{v}}) - \mathbf{f}^*\|^2 \leq \|\boldsymbol{\xi}\|^2 + \sigma^{-2} q^{-2\beta} n + Cq \sqrt{(p+q+x)^3 \sigma/n}.$$

Proof. It has been already shown that the conditions (S_3) through (S_5) are fulfilled automatically for a smooth link function g if $q = n^{1/5}$. Also one has to check the identifiability condition (\mathcal{I}) , or, equivalently, (S_6) . Then Theorem 17.1.1 claims root-n quality of the estimate $\tilde{\boldsymbol{\theta}}$.

To be done: please complete

The result can be extended to the case when $f(X)$ is not exactly equal to a single-index function $g(X^\top \boldsymbol{\theta}^*)$ but can be reasonably well approximated by such functions.

17.2 Error-in-variable nonparametric regression

Here we consider a univariate regression model

$$Y_i = f(X_i) + \varepsilon_i,$$

$$Z_i = X_i + \nu_i,$$

which is approximated using linear expansions

$$\begin{aligned} f(X_i) &= f(X_i, \boldsymbol{\theta}) = \sum_{m=1}^p \theta_m \psi_m(X_i) = \boldsymbol{\theta}^\top \Psi(X_i), \\ \mathbf{X} &= \Phi^\top \boldsymbol{\eta} = \sum_{j=1}^q \eta_j \phi_j \end{aligned}$$

for a given functional basis system $\{\psi_m\}$ representing the target regression function f , and the given vector set $\{\phi_j\}$ in \mathbb{R}^n representing the unknown design X_1, \dots, X_n . Below we build the sieve log-likelihood $L(\mathbf{v})$ under the assumptions that the errors ε_i and ν_i are mutually independent homogeneous Gaussian:

$$L(\mathbf{v}) = L(\boldsymbol{\theta}, \boldsymbol{\eta}) = -\frac{1}{2\sigma^2} \|\mathbf{Y} - \boldsymbol{\theta}^\top \Psi(\Phi^\top \boldsymbol{\eta})\|^2 - \frac{1}{2s^2} \|\mathbf{Z} - \Phi^\top \boldsymbol{\eta}\|^2.$$

Note that the first term coincides with the similar expression for the single-index model in which the target and nuisance are exchanged and Φ stands for the design. As usual, the target $\mathbf{v}^* = (\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$ is defined by maximizing the expectation of $L(\mathbf{v})$:

$$\mathbf{v}^* \stackrel{\text{def}}{=} \underset{\mathbf{v}}{\operatorname{argmax}} \mathbb{E} L(\mathbf{v}) = \underset{\mathbf{v}=(\boldsymbol{\theta}, \boldsymbol{\eta})}{\operatorname{argmin}} \left\{ \sigma^{-2} \|\mathbb{E} \mathbf{Y} - \boldsymbol{\theta}^\top \Psi(\Phi^\top \boldsymbol{\eta})\|^2 + s^{-2} \|\mathbb{E} \mathbf{Z} - \Phi^\top \boldsymbol{\eta}\|^2 \right\}.$$

The stochastic component $\zeta(\mathbf{v})$ is given by

$$\zeta(\mathbf{v}) = \sum_{i=1}^n \left\{ \sigma^{-2} \varepsilon_i \boldsymbol{\theta}^\top \Psi(\Phi_i^\top \boldsymbol{\eta}) + s^{-2} \nu_i \Phi_i^\top \boldsymbol{\eta} \right\}.$$

This yields the gradient

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \zeta(\mathbf{v}) &= \sigma^{-2} \sum_{i=1}^n \varepsilon_i \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\Phi_i^\top \boldsymbol{\eta}) = \sigma^{-2} \sum_{i=1}^n \varepsilon_i \Psi'(\Phi_i^\top \boldsymbol{\eta}), \\ \nabla_{\boldsymbol{\eta}} \zeta(\mathbf{v}) &= \sum_{i=1}^n \left\{ \sigma^{-2} \varepsilon_i \nabla_{\boldsymbol{\eta}} f_{\boldsymbol{\theta}}(\Phi_i^\top \boldsymbol{\eta}) + s^{-2} \nu_i \Phi_i \right\} = \sum_{i=1}^n \left\{ \sigma^{-2} \varepsilon_i \boldsymbol{\theta}^\top \Psi'(\Phi_i^\top \boldsymbol{\eta}) \Phi_i + s^{-2} \nu_i \Phi_i \right\}. \end{aligned}$$

The true data distribution does not assume homogeneous Gaussian errors, we only require that pairs (ε_i, ν_i) are independent for different i but allow for the dependence between

ε_i and ν_i . If $\text{Var}(\varepsilon_i) = \sigma_i^2$, $\text{Var}(\nu_i) = s_i^2$, and $\mathbb{E}\varepsilon_i\nu_i = \rho_i$, then the corresponding covariance matrix $\mathcal{V}^2 = \text{Var}\{\nabla\zeta(\boldsymbol{v}^*)\}$ has the following blocks:

$$\begin{aligned}\mathcal{V}_{\boldsymbol{\theta}}^2 &= \text{Var}\{\nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{v}^*)\} = \sigma^{-4} \sum_{i=1}^n \sigma_i^2 \Psi(\Phi_i^\top \boldsymbol{\eta}^*) \{\Psi(\Phi_i^\top \boldsymbol{\eta}^*)\}^\top \\ \mathcal{V}_{\boldsymbol{\eta}}^2 &= \text{Var}\{\nabla_{\boldsymbol{\eta}}\zeta(\boldsymbol{v}^*)\} \\ &= \sum_{i=1}^n \left(\frac{\sigma_i^2}{\sigma^4} |\boldsymbol{\theta}^\top \Psi'(\Phi_i^\top \boldsymbol{\eta})|^2 + \frac{s_i^2}{s^4} + \frac{2\rho_i}{\sigma^2 s^2} \boldsymbol{\theta}^\top \Psi'(\Phi_i^\top \boldsymbol{\eta}) \right) \Phi_i \Phi_i^\top\end{aligned}$$

and

$$\begin{aligned}\mathcal{V}_{\boldsymbol{\theta}\boldsymbol{\eta}}^2 &= \text{Cov}\{\nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{v}^*), \nabla_{\boldsymbol{\eta}}\zeta(\boldsymbol{v}^*)\} \\ &= \sum_{i=1}^n \left\{ \frac{\sigma_i^2}{\sigma^4} \boldsymbol{\theta}^\top \Psi'(\Phi_i^\top \boldsymbol{\eta}) \Psi(\Phi_i^\top \boldsymbol{\eta}) \Phi_i^\top + \frac{\rho_i}{\sigma^2 s^2} \Psi(\Phi_i^\top \boldsymbol{\eta}^*) \Phi_i^\top \right\}.\end{aligned}$$

For the second derivative of $\zeta(\boldsymbol{v})$, it holds

$$\begin{aligned}\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 \zeta(\boldsymbol{v}) &= 0, \\ \nabla_{\boldsymbol{\theta}\boldsymbol{\eta}}^2 \zeta(\boldsymbol{v}) &= \sum_{i=1}^n \varepsilon_i \nabla_{\boldsymbol{\theta}\boldsymbol{\eta}} f_{\boldsymbol{\theta}}(\Phi_i^\top \boldsymbol{\eta}) = \sum_{i=1}^n \varepsilon_i \Psi'(\Phi_i^\top \boldsymbol{\eta}) \Phi_i, \\ \nabla_{\boldsymbol{\eta}\boldsymbol{\eta}}^2 \zeta(\boldsymbol{v}) &= \sum_{i=1}^n \varepsilon_i \nabla_{\boldsymbol{\eta}\boldsymbol{\eta}} f_{\boldsymbol{\theta}}(\Phi_i^\top \boldsymbol{\eta}) = \sum_{i=1}^n \varepsilon_i \boldsymbol{\theta}^\top \Psi''(\Phi_i^\top \boldsymbol{\eta}) \Phi_i \Phi_i^\top.\end{aligned}$$

Conditions **(ED₀)** and **(ED₂)** can be checked as in the case of a single-index model.

The expected log-likelihood reads

$$\mathbb{E}L(\boldsymbol{v}) = \mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\eta}) = -\frac{1}{2\sigma^2} \|\boldsymbol{f}^* - \boldsymbol{\theta}^\top \Psi(\Phi^\top \boldsymbol{\eta})\|^2 - \frac{1}{2s^2} \|\boldsymbol{X}^* - \Phi^\top \boldsymbol{\eta}\|^2,$$

where \boldsymbol{X}^* is the true design. The expression $\|\boldsymbol{X}^* - \Phi^\top \boldsymbol{\eta}\|^2$ is quadratic in $\boldsymbol{\eta}$ and it does not deteriorate the quality of quadratic approximation of the expected log-likelihood. For the first term the structure is similar to the single-index case, the required conditions are similar as well.

Required conditions:

(V₀) [Regular noise] The pairs $\varepsilon_i = Y_i - \mathbb{E}Y_i$ and $\nu_i = Z_i - \mathbb{E}Z_i$ are independent (sub-)Gaussian for different i and satisfy

$$\begin{aligned}\min_i \text{Var}(\varepsilon_i) &\geq C_V \sigma^2, & \max_i \text{Var}(\varepsilon_i) &\leq C_V \sigma^2, \\ \min_i \text{Var}(\nu_i) &\geq C_V s^2, & \max_i \text{Var}(\nu_i) &\leq C_V s^2,\end{aligned}$$

and for $\nu < 1$

$$|E(\varepsilon_i \nu_i)| = |\rho_i| \leq \nu \sigma_i s_i.$$

(V₁) [Regular design] For some positive constants C_f , C_Φ , and C_Ψ , it holds with $V_\Phi^2 = E_\Phi(\Phi\Phi^\top)$

$$\begin{aligned} H_\theta^2 &= E_\Phi \left[\Psi(\Phi^\top \eta^*) \{ \Psi(\Phi^\top \eta^*) \}^\top \right] \geq C_\Psi^{-2} I_p > 0, \\ H_\eta^2 &\geq (C_f^{-2} + \sigma^{-2}) V_\Phi^2, \end{aligned}$$

and for any $\mathbf{u}, \mathbf{u}_1 \in I\!\!R^q$

$$\frac{1}{n} \sum_{i=1}^n |\mathbf{u}^\top V_\Phi^{-1} \Phi_i|^2 |\mathbf{u}_1^\top V_\Phi^{-1} \Phi_i|^2 \leq C_\Phi^2 \|\mathbf{u}\|^2 \|\mathbf{u}_1\|^2.$$

(V₂) [Ψ -basis] For each $s = 0, 1, 2$, there are some fixed constants $\mu_{1,s} \leq \mu_{2,s} \leq \dots \mu_{p,s}$ such that each function $\psi_m(\cdot)$ fulfills for all $\theta \in \Theta$

$$\|\psi_m^{(s)}(\cdot)\|_\infty = \sup_u |\psi_m^{(s)}(u)| \leq \mu_{m,s}, \quad m = 1, \dots, p. \quad (17.5)$$

(V₃) [Smooth vector θ] For each $\theta \in \Theta$ and some fixed constants $C_{f,s}$ for $s = 0, 1, 2$, it holds

$$\|\theta^\top \Psi^{(s)}(\cdot)\|_\infty = \sup_u \{ \theta_1 \psi_1^{(s)}(u) + \dots + \theta_p \psi_p^{(s)}(u) \} \leq C_{f,s}, \quad s = 0, 1, 2.$$

(V₄) [Semiparametric identifiability]

$$E_\Phi \{ \theta^{*\top} \Psi'(\Phi^\top \eta^*) \Phi \Psi(\Phi^\top \eta^*)^\top \} \leq \nu H_\theta H_\eta.$$

This condition is fulfilled automatically due to the penalty term $\sigma^{-2} E_\Phi(\Phi\Phi^\top)$.

(V₅) [Critical dimension] The value $\mu_{p,1} \sqrt{(p+q)/n}$ is small;

(V₆) [Small modeling bias] The value Δ_f defined as

$$\Delta_f^2 = \frac{1}{n} \sum_{i=1}^n |f(X_i) - \theta^{*\top} \Psi(\Phi_i^\top \eta^*)|^2$$

satisfies the condition “ $\mu_{p,1} \Delta_f$ is small”.

Now we specify the general semiparametric result of Theorem 16.7.1 to the case of error-in-variable regression. We use $\mathbb{F}_0 = -\nabla^2 E\mathcal{L}(\mathbf{v}^*)$ and $\mathbb{F}_0 = nH^2$, where H^2 has the blocks H_θ^2, H_η^2 , and $A_{\theta\eta}, A_{\theta\eta}^\top$. The local vicinity $\Upsilon_\circ(\mathbf{r})$ is defined by (16.9).

Theorem 17.2.1. Suppose (V₀) through (V₆). Then the choice $r_0^2 = C(p+q+x)$ ensures

$$IP(\tilde{\mathbf{v}} \notin \Upsilon_\circ(\mathbf{r}_0)) \leq e^{-x}.$$

Moreover, for $\tilde{\mathbf{v}} \in \Upsilon_o(\mathbf{r}_0)$

$$\|D(\tilde{\mathbf{v}} - \mathbf{v}^*) - \xi\| \leq \diamond(\mathbf{r}_0) = C(C_1 + C_V)\mu_{p,1}(p + q + x)/\sqrt{n}.$$

with $C_1 = C\mu_{p,0}$ and $\xi \stackrel{\text{def}}{=} D^{-1}\nabla L(\mathbf{v}^*)$. For the $\boldsymbol{\theta}$ -component of \mathbf{v} , this yields

$$\|\mathbb{I}^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \check{\xi}\| \leq \diamond(\mathbf{r}_0)$$

with \mathbb{I} and $\check{\xi}$ from (16.32).

Cosine-basis and sieve approximation

For a special case of the cosine-basis $\psi_m(u) = \cos(mu)$, condition (17.5) meets with $\mu_{m,s} = m^s$.

Now we evaluate the distance between the underlying vector $\mathbf{f}^* = (f(X_i))$ and its sieve approximation $\Psi(\Phi^\top \boldsymbol{\eta}^*)^\top \boldsymbol{\theta}^*$. Obviously

$$\|\mathbf{f}^* - \Psi(\Phi^\top \boldsymbol{\eta}^*)^\top \boldsymbol{\theta}^*\| \leq \|\mathbf{f}^* - \Psi(\mathbf{X})^\top \boldsymbol{\theta}^*\| + \|\{\Psi(\Phi^\top \boldsymbol{\eta}^*) - \Psi(\mathbf{X})\}^\top \boldsymbol{\theta}^*\|. \quad (17.6)$$

If f is β -smooth, then the first term can be bounded as

$$\|\mathbf{f}^* - \Psi(\mathbf{X})^\top \boldsymbol{\theta}^*\|^2 \leq Cp^{-2\beta}n.$$

For the second term in (17.6), we have to use some smoothness of the basis functions ψ_1, \dots, ψ_p . The second order Taylor expansion implies for each i by (V3)

$$|\{\Psi(\Phi_i^\top \boldsymbol{\eta}^*) - \Psi(X_i)\}^\top \boldsymbol{\theta}^* - (\Phi_i^\top \boldsymbol{\eta}^* - X_i)\Psi'(X_i)\boldsymbol{\theta}^*| \leq C_{f,2}|\Phi_i^\top \boldsymbol{\eta}^* - X_i|^2/2.$$

Therefore,

$$\|\{\Psi(\Phi^\top \boldsymbol{\eta}^*) - \Psi(\mathbf{X})\}^\top \boldsymbol{\theta}^* - (\Phi^\top \boldsymbol{\eta}^* - \mathbf{X})^\top \Psi'(\mathbf{X}^*)\boldsymbol{\theta}^*\| \leq C_{f,2}\|\Phi^\top \boldsymbol{\eta}^* - \mathbf{X}\|^2/2.$$

Again by (V3)

$$\|(\Phi^\top \boldsymbol{\eta}^* - \mathbf{X})^\top \Psi'(\mathbf{X}^*)\boldsymbol{\theta}^*\| \leq C_{f,1}\|\Phi^\top \boldsymbol{\eta}^* - \mathbf{X}\|.$$

We can conclude that

$$\|\mathbf{f}^* - \Psi(\Phi^\top \boldsymbol{\eta}^*)^\top \boldsymbol{\theta}^*\| \leq C\sqrt{np^{-2\beta}} + C_{f,1}\|\Phi^\top \boldsymbol{\eta}^* - \mathbf{X}\| + C_{f,2}\|\Phi^\top \boldsymbol{\eta}^* - \mathbf{X}\|^2/2.$$

Lower bound

Consider a special choice with $p = q$, the function $f(x) = \theta_p \psi_p(x)$ and the design family corresponding to $\boldsymbol{\eta} = \boldsymbol{\eta}^* + \eta_q \boldsymbol{\phi}_q$ for the $\boldsymbol{\phi}$ -basis with $\phi_{i,q} = \psi_p(X_i)$. Then $\mathbf{f}(\mathbf{v}) = \theta_p \psi_p(\mathbf{X}^* + \eta_q \boldsymbol{\phi}_q)$. This reduces the situation to parametric family with a two dimensional parameter (θ, η) (we drop the sub-index p resp. q): $\mathbb{E}_{\theta, \eta} Y_i = \theta \psi(i/n + \eta \phi_i)$. It implies for the point \mathbf{v}^* corresponding to $\theta^* = \eta^* = 1$

$$\nabla_\theta \mathbf{f}(\mathbf{v}^*) = \boldsymbol{\psi}(\mathbf{X}^*),$$

$$\nabla_\eta \mathbf{f}(\mathbf{v}^*) = \theta^* \boldsymbol{\psi}'(\mathbf{X}^*) \cdot \boldsymbol{\phi}$$

where $X_i^* = i/n + \phi_i$. The idea of the construction is to apply a piecewise linear function $\psi(x)$ and similarly for ϕ . For instance, one can take

$$\psi(x) = (x - 1/2) \mathbb{I}(|x - 1/2| \leq h),$$

$$\phi_i = (1/2 - i/n) \mathbb{I}(|i/n - 1/2| \leq h),$$

where h is a proper bandwidth (to be selected). For $\theta^* = 1$, it holds $\nabla_\theta = -\nabla_\eta$ and we encounter an identification problem.

Impact of the design variance σ^2

The accuracy of $\tilde{\boldsymbol{\eta}}$ is of order σq , so that a small value of σ compensate a high design resolution q .

17.3 Instrumental regression

Observed: a sample from (Y, X, W) . Model

$$Y = f(X) + U, \quad \mathbb{E}[U | W] = 0.$$

where Y , an explained variable, X , an explanatory variable, W , an instrument. The target is the regression function $f(\cdot)$.

Let $\psi_1(x), \dots, \psi_j(x), \dots$ be a functional basis. Consider a finite approximation

$$f(x) = \theta_1 \psi_1(x) + \dots + \theta_p \psi_p(x)$$

or in vector form

$$f(x) = \boldsymbol{\Psi}(x)^\top \boldsymbol{\theta}$$

with $\Psi(x) = (\psi_1(x), \dots, \psi_p(x))^\top \in I\!\!R^p$ and $\theta = (\theta_1, \dots, \theta_p)^\top \in I\!\!R^p$. This leads to an approximating model

$$\mathbf{Y} = \Psi(X)^\top \theta^* + U, \quad \mathbb{E}[U | W] = 0.$$

The constraint $\mathbb{E}[U | W] = 0$ means that for any function $\phi(W)$

$$\mathbb{E}[\phi(W)\mathbf{Y}] = \mathbb{E}[\phi(W)\Psi(X)^\top]\theta^*.$$

We apply a *discretization* or *finite dimensional approximation*: for a finite collection of functions $\phi_1(w), \dots, \phi_q(w)$, it holds

$$\mathbb{E}[\Phi(W)\mathbf{Y}] = \mathbb{E}[\Psi(X)\Phi(W)^\top]^\top \theta^* = \mathbf{T}_0^\top \theta^*$$

with

$$\Phi(w) = (\phi_1(w), \dots, \phi_q(w))^\top \in I\!\!R^q,$$

$$\mathbf{T}_0 = \mathbb{E}[\Psi(X)\Phi(W)^\top] \in I\!\!R^{p \times q}.$$

Define $T_i = \Psi(X_i)\Phi(W_i)^\top$

$$\begin{aligned} \mathbf{Z} &= \mathbb{E}_n[\mathbf{Y}\Phi(W)] &= n^{-1} \sum_{i=1}^n Y_i \Phi(W_i) \in I\!\!R^q, \\ \mathbf{T}_n &= \mathbb{E}_n[\Psi(X)\Phi(W)^\top] &= n^{-1} \sum_{i=1}^n T_i \in I\!\!R^{q \times p}, \\ \boldsymbol{\varepsilon} &= \mathbb{E}_n[\Phi(W)\mathbf{U}] &= n^{-1} \sum_{i=1}^n \Phi(W_i) U_i \in I\!\!R^q. \end{aligned} \tag{17.7}$$

The original problems reduces to

$$\mathbf{Z} = \mathbf{T}_0^\top \theta^* + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} = \mathbb{E}_n[\Phi(W)\mathbf{U}]$ is the error q -vector, $\mathbf{T}_0 = \mathbb{E}[\Psi(X)\Phi(W)^\top]$ is an unknown $p \times q$ matrix and only its empirical counterpart \mathbf{T}_n is available. In such cases one speaks of an *inverse problem with error in operator*. The main problem for the analysis in this model is that \mathbf{T}_n is random and correlated with \mathbf{Z} and $\boldsymbol{\varepsilon}$. Another issue is that the errors ε_j are not independent, they are composed as linear combinations of the U_i 's. The goal is to build an estimator $\tilde{\theta}$ of the vector θ^* leading to the estimator $\tilde{f}(x) = \Psi(x)^\top \tilde{\theta}$ of the response.

Plug-in method

The natural plug-in approach suggests to replace the unknown operator \mathbf{T}_0 by its empirical counterpart \mathbf{T}_n leading to the approximating linear model

$$\mathbf{Z} = \mathbf{T}_n^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}.$$

with the random design $\mathbf{T}_n = n^{-1} \sum_{i=1}^n T_i$. The corresponding least square estimator of $\boldsymbol{\theta}^*$ reads as

$$\tilde{\boldsymbol{\theta}} = (\mathbf{T}_n \mathbf{T}_n^\top)^{-1} \mathbf{T}_n \mathbf{Z}. \quad (17.8)$$

The results of Theorem 17.3.1 justify that the random matrix $\mathbf{T}_n \mathbf{T}_n^\top$ is very close to the product $\mathbb{E}(\mathbf{T}_n) \mathbb{E}(\mathbf{T}_n^\top)$ and the theoretical study of the properties of the estimator $\tilde{\boldsymbol{\theta}}$ can be done with $\mathbf{M}_0^2 \stackrel{\text{def}}{=} \mathbf{T}_0 \mathbf{T}_0^\top$ in place of $\mathbf{T}_n \mathbf{T}_n^\top$ in (17.8). Similarly one can justify that the product $\mathbf{T}_n \mathbf{Z}$ behaves nearly as $\mathbf{T}_0 \mathbf{Z}$.

Below we assume for simplicity that all triples (Y_i, X_i, W_i) are i.i.d. so that $T_i = \Psi(X_i) \Phi(W_i)^\top$ are also i.i.d. Define

$$\mathbf{v}_1^2 = \|\mathbb{E}(T_i T_i^\top) - \mathbb{E}(T_i) \mathbb{E}(T_i^\top)\| = \|\mathbb{E}(T_i T_i^\top) - \mathbf{T}_0 \mathbf{T}_0^\top\|$$

and suppose that it holds almost surely

$$\|\mathbf{v}_1^{-1/2} (T_i - \mathbf{T}_0)\| \leq u. \quad (17.9)$$

Theorem 17.3.1. Let (Y_i, X_i, W_i) be i.i.d. and T_i from (17.7) fulfill (17.9). Let also $\mathbf{M}_0^2 \stackrel{\text{def}}{=} \mathbf{T}_0 \mathbf{T}_0^\top \geq \alpha^2 \mathbf{v}_1^2$. Then

$$\mathbb{P}\left(\sqrt{n} \|\mathbf{v}_1^{-1} (\mathbf{T}_n - \mathbf{T}_0)\| > t\right) \leq 2(p+q) \exp\left\{-\frac{t^2}{2 + 2ut/(3n^{1/2})}\right\}$$

Moreover, if $\sqrt{n} \|\mathbf{v}_1^{-1} (\mathbf{T}_n - \mathbf{T}_0)\| \leq t$, then with $\delta = \alpha t \|\mathbf{M}_0^{-1/2}\| / \sqrt{n}$

$$\|\mathbf{M}_0^{-1} \mathbf{T}_n \mathbf{T}_n^\top \mathbf{M}_0^{-1} - I_p\| \leq \delta^2 + 2\delta.$$

Semiparametric approach

Now we consider the joint optimization problem: with $\mathbf{v} = (\boldsymbol{\theta}, \mathbf{T})$, consider

$$\tilde{\mathbf{v}} = \underset{\mathbf{v}}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \|Z_i - \mathbf{T}^\top \boldsymbol{\theta}\|^2 + \sum_{i=1}^n \|\mathbf{T} - T_i\|_{\text{Fr}}^2 \right\}.$$

Here $Z_i = Y_i \Phi(W_i)$ and $T_i = \Psi(X_i) \Phi(W_i)^\top$, and $\|A\|_{\text{Fr}}^2$ means the squared Frobenius norm of A : $\|A\|_{\text{Fr}}^2 = \text{tr}(AA^\top) = \sum_{ij} a_{ij}^2$.

Represent $\tau_i = T_i - \mathbf{T}_0$ and $\boldsymbol{\varepsilon}_i = Z_i - \mathbb{E}Z_i = U_i \Phi(W_i)$. The stochastic component reads as

$$\zeta(\mathbf{v}) = \sum_{i=1}^n \boldsymbol{\varepsilon}_i^\top \mathbf{T}^\top \boldsymbol{\theta} + \sum_{i=1}^n \text{tr}(\tau_i \mathbf{T}^\top).$$

The error vectors $\boldsymbol{\varepsilon}_i \in I\!\!R^q$ are i.i.d. zero mean. Its covariance operator can be described by the variance of $\mathbf{u}^\top \boldsymbol{\varepsilon}_i$ for a vector $\mathbf{u} \in I\!\!R^q$: fulfills

$$\text{Var}(\mathbf{u}^\top \boldsymbol{\varepsilon}_i) = \text{Var}\{\mathbf{u}^\top \boldsymbol{\Phi}(W_1) U_1\} = I\!\!E\{\mathbf{u}^\top \boldsymbol{\Phi}(W_1) U_1\}^2.$$

For the error in operator $\tau_i = T_i - \mathbf{T}_0$, which is a $p \times q$ matrix, it holds in a similar way $I\!\!E\tau_i = 0$ and for any operator \mathbf{A}

$$\text{Var}\{\text{tr}(\mathbf{A}\tau_i^\top)\} = \text{Var}\{\boldsymbol{\Psi}(X_1)^\top \mathbf{A} \boldsymbol{\Phi}(W_1)\}$$

Now we check (ED_0) and (ED_2) . Direct calculus yields for $\mathbf{v} = (\boldsymbol{\theta}, \mathbf{T})$

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\zeta(\mathbf{v}) &= \mathbf{T} \sum_{i=1}^n \boldsymbol{\varepsilon}_i, \\ \nabla_{\mathbf{T}}\zeta(\mathbf{v}) &= \boldsymbol{\theta} \sum_{i=1}^n \boldsymbol{\varepsilon}_i^\top + \sum_{i=1}^n \tau_i\end{aligned}$$

Obviously the gradient $\nabla\zeta(\mathbf{v})$ has a special structure of empirical mean:

$$\nabla\zeta(\mathbf{v}) = \frac{1}{n} \sum_{i=1}^n \nabla\zeta_i(\mathbf{v}),$$

where $\zeta_i(\mathbf{v})$ are i.i.d. zero mean. Describe first the covariance structure of $\nabla\zeta(\mathbf{v}^*)$:

$$\mathbb{F}_0 \stackrel{\text{def}}{=} \text{Var}\{\nabla\zeta(\mathbf{v})\} = n \begin{pmatrix} \mathsf{F}_{\boldsymbol{\theta}} & \mathsf{F}_{\boldsymbol{\theta}\mathbf{T}} \\ \mathsf{F}_{\boldsymbol{\theta}\mathbf{T}}^\top & \mathsf{F}_{\mathbf{T}\mathbf{T}} \end{pmatrix}.$$

It holds for $\boldsymbol{\gamma} \in I\!\!R^p$ and \mathbf{A}

$$\begin{aligned}\langle \mathsf{F}_{\boldsymbol{\theta}}\boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle &\stackrel{\text{def}}{=} I\!\!E\{\boldsymbol{\gamma}^\top \mathbf{T}_0 \boldsymbol{\Phi}(W_1) U_1\}^2, \\ \langle \mathsf{F}_{\mathbf{T}\mathbf{T}}\mathbf{A}, \mathbf{A} \rangle &= \text{Var}\{\boldsymbol{\Psi}(X_1)^\top \mathbf{A} \boldsymbol{\Phi}(W_1) + \boldsymbol{\theta}^{*\top} \mathbf{A} \boldsymbol{\Phi}(W_1) U_1\}, \\ \langle \mathsf{F}_{\boldsymbol{\theta}, \mathbf{T}}\boldsymbol{\gamma}, \mathbf{A} \rangle &= I\!\!E[\boldsymbol{\gamma}^\top \mathbf{T}_0 \boldsymbol{\Phi}(W_1) U_1 \{\boldsymbol{\Psi}(X_1)^\top \mathbf{A} \boldsymbol{\Phi}(W_1) + \boldsymbol{\theta}^{*\top} \mathbf{A} \boldsymbol{\Phi}(W_1) U_1\}]\end{aligned}$$

Condition (ED_0) holds under (sub-)Gaussian behavior of $\mathbf{T}_0 \boldsymbol{\Phi}(W_i) U_i$ and of $\boldsymbol{\Psi}(X_i) \boldsymbol{\Phi}(W_i)^\top$ using the matrix Bernstein inequality.

For checking (ED_2) , compute the Hessian of $\zeta(\mathbf{v})$. It follows

$$\begin{aligned}\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2\zeta(\mathbf{v}) &= 0, \\ \langle \nabla_{\boldsymbol{\theta}\mathbf{T}}^2\zeta(\mathbf{v}), (\boldsymbol{\gamma}, \mathbf{A}) \rangle &= \sum_{i=1}^n \boldsymbol{\varepsilon}_i^\top \mathbf{A} \boldsymbol{\gamma}, \\ \nabla_{\mathbf{T}\mathbf{T}}^2\zeta(\mathbf{v}) &= 0.\end{aligned}$$

Only the off-diagonal block $\nabla_{\theta T}^2 \zeta(\mathbf{v})$ is non-zero, but it is linear in $\boldsymbol{\varepsilon}_i$ and does not depend on \mathbf{v} . This enables (ED_2) under the same conditions with $\omega \asymp n^{-1/2}$.

Now compute the expected log-likelihood. It holds

$$-2\mathbb{E}L(\mathbf{v}) = n\|\mathbf{T}_0^\top \boldsymbol{\theta}^* - \mathbf{T}^\top \boldsymbol{\theta}\|^2 + n\|\mathbf{T} - \mathbf{T}_0\|_{\text{Fr}}^2 + R$$

where the term R does not depend on \mathbf{v} . Therefore, the blocks of $\mathbb{F}(\mathbf{v}) = -\nabla^2 \mathbb{E}L(\mathbf{v})$ read for $\mathbf{v} = (\boldsymbol{\theta}, \mathbf{T})$

$$\begin{aligned} \langle \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\mathbf{v})\boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle &= n\|\mathbf{T}^\top \boldsymbol{\gamma}\|^2, \\ \langle \mathbb{F}_{\boldsymbol{\theta}\mathbf{T}}(\mathbf{v})\boldsymbol{\gamma}, \mathbf{A} \rangle &= n\boldsymbol{\gamma}^\top \mathbf{A}(\mathbf{T}_0^\top \boldsymbol{\theta}^* - \mathbf{T}^\top \boldsymbol{\theta}) + n\boldsymbol{\gamma}^\top \boldsymbol{\theta} \operatorname{tr}(\mathbf{T}\mathbf{A}^\top) \\ \langle \mathbb{F}_{\mathbf{T}\mathbf{T}}(\mathbf{v})\mathbf{A}, \mathbf{A} \rangle &= n\|\mathbf{A}^\top \boldsymbol{\theta}\|^2 + n\|\mathbf{A}\|_{\text{Fr}}^2 \end{aligned}$$

One can see that the block $\mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\mathbf{v})$ is quadratic in \mathbf{T} , the block $\mathbb{F}_{\mathbf{T}\mathbf{T}}(\mathbf{v})$ is quadratic in $\boldsymbol{\theta}$, while the block $\mathbb{F}_{\boldsymbol{\theta}\mathbf{T}}(\mathbf{v})$ is bilinear. This easily implies (\mathcal{L}_0) on a bounded set $\Upsilon_\circ(\mathbf{r})$:

$$\Upsilon_\circ(\mathbf{r}) \stackrel{\text{def}}{=} \left\{ \mathbf{v} = (\boldsymbol{\theta}, \mathbf{T}) : \|\mathbf{T}_0^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}/\sqrt{n}, \|\mathbf{T} - \mathbf{T}_0\|_{\text{Fr}} \leq \mathbf{r}/\sqrt{n} \right\}.$$

On this set, the value $\delta(\mathbf{r})$ can be bounded as

$$\delta(\mathbf{r}) \leq C\mathbf{r}/\sqrt{n}.$$

Note that the dimension $p \times q$ of the nuisance operator \mathbf{T} is much larger than the dimension p of the target $\boldsymbol{\theta}$.

Required conditions

- (I₁)** [Data] The observations (Y_i, X_i, W_i) are i.i.d.
- (I₂)** [Design set] The design (X_i, W_i) is supported on a compact set \mathfrak{X} ;
- (I₃)** [Basis Ψ, Φ] The basis functions $\psi_m(X)$ and $\phi_j(W)$ are bounded on \mathfrak{X} ;
- (I₄)** [Identifiability]
- (I₅)** [Small modeling bias] The approximation $\mathbb{E}[\Phi(W)\mathbf{Y}] = \mathbf{T}_0^\top \boldsymbol{\theta}^*$ is reasonable;
- (I₆)** [Critical dimension] It holds pq/\sqrt{n} small;

Main results

Theorem 17.3.2. Suppose (I_1) through (I_7) . Then the choice $\mathbf{r}_0^2 = C(pq + \mathbf{x})$ ensures

$$\mathbb{P}(\tilde{\mathbf{v}} \notin \Upsilon_\circ(\mathbf{r}_0)) \leq e^{-\mathbf{x}}.$$

Moreover, for $\tilde{\mathbf{v}} \in \Upsilon_\circ(\mathbf{r}_0)$

$$\|D(\tilde{\boldsymbol{v}} - \boldsymbol{v}^*) - \boldsymbol{\xi}\| \leq \diamond(\mathbf{r}_0) = \mathbf{C}(pq + \mathbf{x})/\sqrt{n}.$$

with $\boldsymbol{\xi} \stackrel{\text{def}}{=} D^{-1}\nabla L(\boldsymbol{v}^*)$. For the $\boldsymbol{\theta}$ -component of \boldsymbol{v} , this yields

$$\|\mathbb{I}^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \check{\boldsymbol{\xi}}\| \leq \diamond(\mathbf{r}_0)$$

with \mathbb{I} and $\check{\boldsymbol{\xi}}$ from (16.32).

Median and quantile regression

Generalized regression

Part III

Structural regression

Sieve Model Selection

This chapter discusses the problem of sieve model selection in the situation when no prior information about the underlying noise distribution is available. The SmA procedure from Section 4.3 requires the set of critical values \mathbf{z}_{m,m° to be fixed. Here we discuss how this can be done in a data driven way with a resampling procedure.

20.1 Sieve SmA procedure

We consider a general parametric setup $\mathbf{Y} \sim \mathbb{P} \in (\mathbb{P}_\theta)$, where the parameter θ is high or infinite dimensional. The sieve approximations assumes that there is a growing sequence of subspaces $\Theta_1 \subset \Theta_2 \subset \dots$, one fixes a proper value m and applied the MLE $\tilde{\theta}_m$ obtained by maximization of the log-likelihood $L(\theta)$ over Θ_m :

$$\tilde{\theta}_m \stackrel{\text{def}}{=} \underset{\theta \in \Theta_m}{\operatorname{argmax}} L(\theta). \quad (20.1)$$

The main issue in applying this approach is the choice of the model parameter m . This problem was discussed in details for linear models with a quadratic log-likelihood function in Chapters 4 and 5. Now we aim at extending the SmA approach to the general sieve maximum likelihood setup.

Define the sieve target

$$\theta_m^* = \underset{\theta \in \Theta_m}{\operatorname{argmax}} \mathbb{E} L(\theta).$$

The Fisher Theorem claims that the sieve MLS $\tilde{\theta}_m$ estimates θ_m^* with the parametric accuracy corresponding to the parameter dimension p_m of the subset Θ_m . The value θ_m^* differs in general from θ^* yielding some sieve bias which decreases with m . At the same time, the use of large m leads to a rather complicated parametric problem (20.1) with p_m parameters. Therefore, a proper model choice has to balance the parametric complexity within the sieve model Θ_m and the bias occurring by replacing the full model by its approximation.

Similarly to the linear case we suppose to be given by a loss weighting matrix W and measure the loss of estimation by $\|W(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|$. The quadratic risk is defined by its expectation

$$\mathbb{E}\|W(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2.$$

More generally, one can consider polynomial loss function $\|W(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^q$ for a given $q \geq 0$. For the quadratic risk, one can use the bias-variance decomposition

$$\mathbb{E}\|W(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2 = \|W(\mathbb{E}\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2 + \text{tr}\{\text{Var}(W\tilde{\boldsymbol{\theta}}_m)\}.$$

Below we use a slightly different decomposition based on the Fisher expansion. Define

$$\begin{aligned} D_m^2 &= -\nabla_m^2 \mathbb{E}L(\boldsymbol{\theta}^*), \\ \nabla &= \nabla L(\boldsymbol{\theta}^*). \end{aligned}$$

The Fisher expansion for the largest sieve $m = M$ yields a similar statement for each $m < M$: on a random set of probability at least $1 - e^{-x}$

$$\|D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*) - D_m^{-1}\nabla\| \leq \diamond(x)$$

where $\diamond(x) = \diamond(r_M, x)$ is the error of approximation in the M sieve.

Now consider the estimation loss with a weighting matrix W satisfying

$$\|WD_m^{-1}\| \leq 1.$$

Then

$$\|W(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*) - WD_m^{-2}\nabla\| = \|WD_m^{-1}\{D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*) - D_m^{-1}\nabla\}\| \leq \diamond(x)$$

For any two $m > m^\circ$,

$$\|W(\tilde{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_{m^\circ}) - W(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}_{m^\circ}^*) - W(D_m^{-2} - D_{m^\circ}^{-2})\nabla\| \leq 2\diamond(x).$$

This expansion can be rewritten as

$$\|W(\tilde{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_{m^\circ}) - b_{m,m^\circ} - \xi_{m,m^\circ}\| \leq 2\diamond(x) \quad (20.2)$$

with

$$\begin{aligned} b_{m,m^\circ} &\stackrel{\text{def}}{=} W(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}_{m^\circ}^*), \\ \xi_{m,m^\circ} &\stackrel{\text{def}}{=} W(D_m^{-2} - D_{m^\circ}^{-2})\nabla. \end{aligned} \quad (20.3)$$

In the linear case, the expansion $W(\tilde{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_{m^\circ}) - b_{m,m^\circ} - \xi_{m,m^\circ} = 0$ is exact. The formula (20.2) is an extension to a general nonlinear regular case.

Now we suppose that the error term is small enough and proceed as if this expansion is identity. The SmA procedure selects the “smallest accepted” with the acceptance rule

$$m^\circ \text{ is accepted if } \|W(\tilde{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_{m^\circ})\| \leq z_{m,m^\circ} \quad \forall m \in \mathcal{M}(m^\circ).$$

The critical values z_{m,m° have to be selected to ensure the propagation property: if there is no bias for $m > m^\circ$ then the procedure should not reject m° .

Below we discuss how these values can be selected by resampling methods.

20.2 Resampling methods for parameter tuning in generalized regression

This section explains the choice of the critical values z_{m,m° for the generalized regression model by a multiplier bootstrap procedure.

20.2.1 Generalized regression

We consider a sample $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ with independent observations Y_i .

Our parametric assumption concerns the marginal distribution of each observation Y_i and the structure of the regression function f .

We suppose that the distribution P_i of each observation Y_i belongs to a given family $\mathcal{P} = (P_\mathbf{v})$ and the value of this parameter is an unknown function f of the regressor \mathbf{X}_i . We write this relation in the form

$$Y_i \sim P_{f(\mathbf{X}_i)}. \quad (20.4)$$

Further we suppose the family \mathcal{P} to be regular and dominated by a sigma-finite measure μ_0 . By $\ell(y, \mathbf{v})$ we denote the corresponding log-density function: $\ell(y, \mathbf{v}) \stackrel{\text{def}}{=} \log \frac{dP_\mathbf{v}}{d\mu_0}(y)$.

The regression function f will be modelled by a linear expansion

$$f(\mathbf{x}) = f(\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^p \theta_j^* \psi_j(\mathbf{x}) \quad (20.5)$$

for a given basis system $\{\psi_j\}$. For simplicity this basis will be considered finite: $j \leq p$ for a finite p . We write this expansion in the vector form $\mathbf{f}^* = \Psi^\top \boldsymbol{\theta}^*$.

Our parametric assumptions (20.4) about the marginal distribution of each Y_i and (20.5) about the structure of the regression function f lead to the log-likelihood function

$$L(\boldsymbol{\theta}) = \sum_{i=1}^n \ell(Y_i, f(X_i, \boldsymbol{\theta})) = \sum_{i=1}^n \ell(Y_i, \Psi_i^\top \boldsymbol{\theta}) \quad (20.6)$$

and the MLE $\tilde{\boldsymbol{\theta}}$ maximizes $L(\boldsymbol{\theta})$ over the large set of all feasible $\boldsymbol{\theta}$ -values. The sieve approach leads to the family of estimates $\tilde{\boldsymbol{\theta}}_m$ each of them is defined by restricting the parameter set to a subset Θ_m in which only first m components of $\boldsymbol{\theta}$ are varying. This means that for $\boldsymbol{\theta} \in \Theta_m$, the expansion (20.5) reads as

$$f(\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^m \theta_j^* \psi_j(\mathbf{x}). \quad (20.7)$$

Exercise 20.2.1. Write the sieve MLE $\tilde{\boldsymbol{\theta}}_m$ for the generalized linear regression (20.7) in

- Poisson regression with canonical parameter $\nu = 1/\lambda$, where λ is the Poisson intensity parameter;
- logit (Bernoulli canonical) model

Describe in each case the Fisher matrices D_m , the score ∇_m , and the stochastic term ξ_{m,m° in expansion (20.3).

20.2.2 Multiplier bootstrap

We consider now the SmA procedure based on the family of estimates $\tilde{\boldsymbol{\theta}}_m$ and discuss how the critical values \mathbf{z}_{m,m° can be selected in a data-driven way. In what follows we suppose that the data sample \mathbf{Y} is fixed as well as the corresponding feature vectors Ψ_i . This means that the critical values will be computed given data and are in general data-dependent.

Introduce the central object of analysis - the weighted log-likelihood $L^\flat(\boldsymbol{\theta})$ defined via the family of random weights w_i^\flat :

$$L^\flat(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{i=1}^n \ell(Y_i, f(X_i, \boldsymbol{\theta})) w_i^\flat. \quad (20.8)$$

The weights w_i^\flat are assumed i.i.d. given the data \mathbf{Y} with $\mathbb{E}w_i^\flat = \text{Var } w_i^\flat = 1$. Two leading examples of such weights are Gaussian weights with $w_i^\flat \sim \mathcal{N}(1, 1)$ and the exponential weights $w_i^\flat \sim \text{Exp}(1)$. Note the the expression (20.8) looks very similar to (20.6) but there is an essential difference in the probabilistic nature of them. The original log-likelihood is a function of the random data \mathbf{Y} , its distribution is described via the unknown data distribution. In the expression (20.8) the data \mathbf{Y} are considered as fixed and non-random, the only random element there is a collection of weights w_i^\flat with

known distribution. In particular, if w_i^b are i.i.d. normal then $L^b(\boldsymbol{\theta})$ is also normal as a linear combination of normal r.v.s. One can conclude that $L(\boldsymbol{\theta})$ and $L^b(\boldsymbol{\theta})$ are living on two different probability spaces and have very different properties. The link between these two worlds (of real and of bootstrap ones) is given by a very simple observation

$$\mathbb{E}^b L^b(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}).$$

Here and everywhere below \mathbb{P}^b means the distribution of the weights w_i^b given the data \mathbf{Y} . By \mathbb{E}^b we denote the corresponding expectation. The main benefit of considering the measure \mathbb{P}^b is that it is completely known.

Now we interpret $L^b(\boldsymbol{\theta})$ as a log-likelihood process and define

$$\tilde{\boldsymbol{\theta}}^b \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}} L^b(\boldsymbol{\theta}).$$

Similarly one can define $\tilde{\boldsymbol{\theta}}_m^b$ by maximizing $L^b(\boldsymbol{\theta})$ over Θ_m :

$$\tilde{\boldsymbol{\theta}}_m^b \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta_m} L^b(\boldsymbol{\theta}). \quad (20.9)$$

The value $\tilde{\boldsymbol{\theta}}_m^b$ formally depends on both the data \mathbf{Y} and the weights w_i^b but we consider its conditional distribution given the data \mathbf{Y} with the hope that this distribution somehow reflects the original distribution of $\tilde{\boldsymbol{\theta}}_m$. In fact, for running the SmA procedure we only need to know the distribution (tail function) of stochastic parts ξ_{m,m° of considered test statistics $W(\tilde{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_{m^\circ})$. Define their analog in the bootstrap world:

$$\mathbb{T}_{m,m^\circ}^b \stackrel{\text{def}}{=} \|W(\tilde{\boldsymbol{\theta}}_m^b - \tilde{\boldsymbol{\theta}}_{m^\circ}^b)\|.$$

Now we use the great advantage of considering the bootstrap world: the distribution of the $\tilde{\boldsymbol{\theta}}_m^b$ is known, in particular, one can compute the expectation $\mathbb{E}^b(\tilde{\boldsymbol{\theta}}_m^b)$ or use the knowledge of the true value in the bootstrap model which coincides with the real world estimate $\tilde{\boldsymbol{\theta}}_m$. The corresponding stochastic component ξ_{m,m°^b is given by

$$\xi_{m,m^\circ}^b \stackrel{\text{def}}{=} W\{\tilde{\boldsymbol{\theta}}_m^b - \tilde{\boldsymbol{\theta}}_{m^\circ}^b - \mathbb{E}^b(\tilde{\boldsymbol{\theta}}_m^b - \tilde{\boldsymbol{\theta}}_{m^\circ}^b)\}.$$

Now define the tail function $z_{m,m^\circ}^b(\mathbf{x})$ by the relation

$$\mathbb{P}^b(\|\xi_{m,m^\circ}^b\| \geq z_{m,m^\circ}^b(\mathbf{x})) = e^{-\mathbf{x}} \quad (20.10)$$

and further proceed as in the case of known functions z_{m,m° 's. In particular, the multiplicity correction $q_{m^\circ} = q_{m^\circ}(\mathbf{x})$ is defined as the smallest value q satisfying the relation

$$\mathbb{I}P^{\flat} \left(\max_{m > m^{\circ}} \left\{ \|\xi_{m,m^{\circ}}^{\flat}\| - z_{m,m^{\circ}}^{\flat}(\mathbf{x} + q) \right\} \geq 0 \right) \leq e^{-x}.$$

Finally the SmA procedure is applied with

$$z_{m,m^{\circ}} \stackrel{\text{def}}{=} z_{m,m^{\circ}}^{\flat}(\mathbf{x} + q) + \beta \sqrt{p_{m,m^{\circ}}^{\flat}}.$$

20.2.3 Numerical issues

It was mentioned many times that the joint distribution of the test statistics $\mathbb{T}_{m,m^{\circ}}^{\flat}$ under $\mathbb{I}P^{\flat}$ is known. However, its analytic study is a hard task even if the weights w_i^{\flat} are normal. Similarly to linear regression case, one can use a numerical Monte Carlo procedure for evaluating the tail functions of $\mathbb{T}_{m,m^{\circ}}^{\flat}$ and the multiplicity corrections $q_{m^{\circ}}$. The approach can be described as follows:

- generate B samples of weights $\mathbf{w}^{\flat} = (w_i^{\flat})$;
- for each sample, compute and store $\tilde{\boldsymbol{\theta}}_m^{\flat} = \tilde{\boldsymbol{\theta}}_m(\mathbf{w}^{\flat})$ for all m ;
- compute the bootstrap empirical means

$$\mathbb{IE}^{\flat}(\tilde{\boldsymbol{\theta}}_m^{\flat}) \stackrel{\text{def}}{=} \frac{1}{B} \sum_{\mathbf{w}^{\flat}} \tilde{\boldsymbol{\theta}}_m^{\flat}(\mathbf{w}^{\flat});$$

- For each $m > m^{\circ}$ and all feasible \mathbf{x} , compute the tail functions $z_{m,m^{\circ}}(\mathbf{x})$ by the relations (20.10) when $\mathbb{I}P^{\flat}$ is replaced by its bootstrap empirical distribution:

$$\frac{1}{B} \sum_{\mathbf{w}^{\flat}} \mathbb{I}(\|\xi_{m,m^{\circ}}^{\flat}(\mathbf{w}^{\flat})\| \geq z_{m,m^{\circ}}^{\flat}(\mathbf{x})) \leq e^{-x}$$

with

$$\xi_{m,m^{\circ}}^{\flat}(\mathbf{w}^{\flat}) \stackrel{\text{def}}{=} W \left\{ \tilde{\boldsymbol{\theta}}_m^{\flat}(\mathbf{w}^{\flat}) - \tilde{\boldsymbol{\theta}}_{m^{\circ}}^{\flat}(\mathbf{w}^{\flat}) - \mathbb{IE}^{\flat}(\tilde{\boldsymbol{\theta}}_m^{\flat} - \tilde{\boldsymbol{\theta}}_{m^{\circ}}^{\flat}) \right\}.$$

Alternatively,

$$\xi_{m,m^{\circ}}^{\flat}(\mathbf{w}^{\flat}) \stackrel{\text{def}}{=} W \left\{ \tilde{\boldsymbol{\theta}}_m^{\flat}(\mathbf{w}^{\flat}) - \tilde{\boldsymbol{\theta}}_{m^{\circ}}^{\flat}(\mathbf{w}^{\flat}) - (\tilde{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_{m^{\circ}}) \right\}.$$

Also estimate

$$p_{m,m^{\circ}}^{\flat} = \text{Var}^{\flat}(\xi_{m,m^{\circ}}^{\flat}) \approx \text{tr} \left\{ \frac{1}{B} \sum_{\mathbf{w}^{\flat}} \xi_{m,m^{\circ}}^{\flat}(\mathbf{w}^{\flat}) \xi_{m,m^{\circ}}^{\flat}(\mathbf{w}^{\flat})^{\top} \right\}.$$

- Compute for each m° the multiplicity corrections $q_{m^{\circ}}$ as the smallest value with

$$\frac{1}{B} \sum_{\mathbf{w}^{\flat}} \mathbb{I} \left(\max_{m > m^{\circ}} \left\{ \|\xi_{m,m^{\circ}}^{\flat}(\mathbf{w}^{\flat})\| - z_{m,m^{\circ}}^{\flat}(\mathbf{x} + q) \right\} \geq 0 \right) \leq e^{-x}.$$

Complexity of this procedure is mainly determined by the complexity of computing the family of estimates $\tilde{\boldsymbol{\theta}}_m^b(\mathbf{w}^b)$ for each bootstrap sample \mathbf{w}^b . This has to be done many many times to reduce the Monte-Carlo error. Each estimate $\tilde{\boldsymbol{\theta}}_m^b$ is given implicitly via the optimization problem (20.9). This can be a hard task especially if the parameter dimension p is large. Note however one important feature of the bootstrap world: we know the true value, which the estimate $\tilde{\boldsymbol{\theta}}$ computed from the original data. This true value is actually there target of estimation from the resampled data and automatically it is a very good starting point of the estimation procedure; see further details of reducing the computational burden below in Section ??.

In the next section we discuss how the proposed procedure can be justified. First we consider the linear Gaussian case. Then we extend the results to the case of linear models with non-Gaussian errors. Finally we consider the general case of sieve model selection and show how the bootstrap validity can be derived from the Fisher expansions (20.2) in the asymptotic sense for a reasonably large sample size n .

20.3 Sieve Generalized Linear regression

In a special case of a generalized liner model, \mathcal{P} is an exponential family with canonical parametrization. Then $\ell(y, \mathbf{v}) = y\mathbf{v} - d(\mathbf{v})$ and

$$L(\boldsymbol{\theta}) = \sum_{i=1}^n \ell(Y_i, f(X_i)) = \sum_{i=1}^n \{Y_i \Psi_i^\top \boldsymbol{\theta} - d(\Psi_i^\top \boldsymbol{\theta})\} = \mathbf{Y}^\top \Psi^\top \boldsymbol{\theta} - A(\boldsymbol{\theta})$$

with

$$A(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{i=1}^n d(\Psi_i^\top \boldsymbol{\theta}).$$

Also define

$$D^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \nabla^2 A(\boldsymbol{\theta}) = \sum_i \Psi_i \Psi_i^\top d''(\Psi_i^\top \boldsymbol{\theta}).$$

The estimate $\tilde{\boldsymbol{\theta}}$ can be computed by the iterative Newton procedure: start with any $\tilde{\boldsymbol{\theta}}_0$, e.g. LSE, and then compute

$$\tilde{\boldsymbol{\theta}}_{k+1} = \tilde{\boldsymbol{\theta}}_k + D^{-2}(\tilde{\boldsymbol{\theta}}_k) \Psi(\mathbf{Y} - \Psi^\top \tilde{\boldsymbol{\theta}}_k), \quad k = 1, 2, \dots$$

Under standard conditions this procedure converges very fast after just few iterations.

The target $\boldsymbol{\theta}^*$ is described by the optimizing the value $IEL(\boldsymbol{\theta})$. In view of $IEL(\boldsymbol{\theta}) = f(\mathbf{X}_i)$

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E} L(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}} \sum_i \{f(\mathbf{X}_i) \Psi_i^\top \boldsymbol{\theta} - d(\Psi_i^\top \boldsymbol{\theta})\}$$

which leads to the normal equation

$$\sum_{i=1}^n \{f(\mathbf{X}_i) - d'(\Psi_i^\top \boldsymbol{\theta})\} \Psi_i = 0.$$

One also has

$$\mathbb{F} = D^2 = -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}^*) = \sum_i \Psi_i \Psi_i^\top d''(\Psi_i^\top \boldsymbol{\theta}^*) = \Psi \mathbf{d}''(\Psi^\top \boldsymbol{\theta}^*) \Psi^\top,$$

where $\mathbf{d}''(\mathbf{f})$ is a $n \times n$ diagonal matrix with the diagonal entries $d''(f_i)$.

Further, for the stochastic component $\zeta(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathbb{E} L(\boldsymbol{\theta})$, it holds with $\varepsilon_i = Y_i - \mathbb{E} Y_i$

$$\nabla \zeta(\boldsymbol{\theta}) = \sum_i \varepsilon_i \Psi_i = \Psi \boldsymbol{\varepsilon} \quad (20.11)$$

and its covariance matrix fulfills

$$V^2 = \operatorname{Var}\{\nabla \zeta(\boldsymbol{\theta})\} = \sum_i \operatorname{Var}(Y_i) \Psi_i \Psi_i^\top = \Psi \operatorname{Var}(\boldsymbol{\varepsilon}) \Psi^\top. \quad (20.12)$$

If the assumption on marginal distributions is correct, that is, if $Y_i \sim P_{f(\mathbf{X}_i)}$, then

$$\operatorname{Var}(Y_i) = d''(f(\mathbf{X}_i)). \quad (20.13)$$

By comparing the equations (20.12) and (20.13), one can conclude that D^2 and V^2 coincides if $d''(\Psi_i^\top \boldsymbol{\theta}^*) = d''(f(\mathbf{X}_i))$, that is, if PA is correct. Otherwise these two matrices may differ from each other. The Fisher expansion reads as

$$\|D(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi}\| \leq \diamond(x) \quad (20.14)$$

on a dominating set of probability at least $1 - e^{-x}$. Here in view of (20.11)

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} D^{-1} \nabla \zeta(\boldsymbol{\theta}^*) = D^{-1} \Psi \boldsymbol{\varepsilon}.$$

20.3.1 Sieve MLE

Now we consider the sieve MLE setup when a sequence of subsets Θ_m of growing dimension is given and for each m the MLE $\tilde{\boldsymbol{\theta}}_m$ is computed. The corresponding quadratic risk \mathcal{R}_m is

$$\mathcal{R}_m = \mathbb{E} \|W(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2.$$

The Fisher expansion (20.14) restricted to the sieve Θ_m yields

$$\|D_m(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*) - \boldsymbol{\xi}_m\| \leq \diamond(x), \quad (20.15)$$

where

$$\begin{aligned} D_m^2 &\stackrel{\text{def}}{=} -\nabla_m^2 \mathbb{E} L(\boldsymbol{\theta}^*) = \Psi_m \mathbf{d}''(\Psi_m^\top \boldsymbol{\theta}^*) \Psi_m^\top, \\ \boldsymbol{\xi}_m &\stackrel{\text{def}}{=} D_m^{-1} \nabla_m \zeta(\boldsymbol{\theta}^*) = D_m^{-1} \Psi_m \boldsymbol{\varepsilon}, \\ \boldsymbol{\theta}_m^* &\stackrel{\text{def}}{=} \underset{\boldsymbol{\theta} \in \Theta_m}{\operatorname{argmax}} \mathbb{E} L(\boldsymbol{\theta}). \end{aligned}$$

Alternatively one can define

$$\boldsymbol{\theta}_m^* = \Pi_m \boldsymbol{\theta}^*$$

the expansion (20.15) continues to apply.

The test statistic \mathbb{T}_{m,m° can be written as

$$\mathbb{T}_{m,m^\circ} = \|W(\tilde{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_{m^\circ})\|$$

and the decomposition (20.15) implies

$$\left| \mathbb{T}_{m,m^\circ} - \|b_{m,m^\circ} + \boldsymbol{\xi}_{m,m^\circ}\| \right| \leq 2\diamond(x)$$

with

$$\begin{aligned} b_{m,m^\circ} &= W(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}_{m^\circ}^*), \\ \boldsymbol{\xi}_{m,m^\circ} &= W(D_m^{-2} \nabla_m - D_{m^\circ}^{-2} \nabla_{m^\circ}) = W(D_m^{-2} \Pi_m - D_{m^\circ}^{-2} \Pi_{m^\circ}) \nabla \end{aligned} \quad (20.16)$$

for $\nabla = \Psi \boldsymbol{\varepsilon}$.

20.3.2 Bootstrap counterpart

Now we check what happens in the bootstrap world. It holds

$$L^\flat(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta}) w_i^\flat = \sum_{i=1}^n \{Y_i \Psi_i^\top \boldsymbol{\theta} - d(\Psi_i^\top \boldsymbol{\theta})\} w_i^\flat = \boldsymbol{\theta}^\top \Psi \mathcal{W}^\flat \mathbf{Y} - A^\flat(\boldsymbol{\theta})$$

with

$$A^\flat(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{i=1}^n d(\Psi_i^\top \boldsymbol{\theta}) w_i^\flat.$$

One can use that $\mathbb{E}^b A^b(\boldsymbol{\theta}) = A$ and even more, the matrix $A^b(\boldsymbol{\theta})$ is close to its expectation $A(\boldsymbol{\theta})$ for n large. This suggests to replace $A^b(\boldsymbol{\theta})$ by $A(\boldsymbol{\theta})$ in our analysis. The same applies to its Hessian $D^2(\boldsymbol{\theta})$.

This implies similarly to the linear case that the stochastic part of the bootstrap-world estimate $\tilde{\boldsymbol{\theta}}^b$ is $D^{-2}\check{\nabla}^b$ for $\check{\nabla}^b = \Psi\mathcal{E}^b\check{\boldsymbol{\varepsilon}}$ for $\check{\boldsymbol{\varepsilon}} = \mathbf{Y} - \Psi^\top\tilde{\boldsymbol{\theta}}$. One can see that the question of bootstrap validity is again reduced to comparing of two distributions: of ∇ w.r.t. the original measure \mathbb{P} and of $\check{\nabla}^b$ w.r.t. to the bootstrap measure \mathbb{P}^b given \mathbf{Y} . If the bootstrap weights w_i^b are Gaussian then the bootstrap score $\check{\nabla}^b = \Psi\mathcal{E}^b\check{\boldsymbol{\varepsilon}}$ is Gaussian as well, because it is a linear combination of the $e_i^b = w_i^b - 1$'s which are i.i.d. standard normal. Unfortunately, the real world score ∇ is a linear combination of the errors ε_i which in general are not Gaussian. Therefore, we cannot assume that the score ∇ is Gaussian. However, the desirable justification of the bootstrap procedure can be obtained by GAR arguments for the score vector ∇ .

20.3.3 Bootstrap for the SmA procedure

Now we return to the SmA procedure for the sieve GLM estimation scheme. For each m , we consider the sieve estimate $\tilde{\boldsymbol{\theta}}_m$ and its bootstrap counterpart $\tilde{\boldsymbol{\theta}}_m^b$ and try to design the procedure by its desired performance in the bootstrap world where we know the true value $\tilde{\boldsymbol{\theta}}_m$ in each sieve. We use the Fisher expansion as the main tool. Similarly to the real world case, it allows to decompose the difference $\tilde{\boldsymbol{\theta}}_m^b - \tilde{\boldsymbol{\theta}}_m$ into the deterministic and stochastic part. For the SmA procedure, we have to consider the pairs $\tilde{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_{m^\circ}$ for all $m > m^\circ$ and their bootstrap versions. The counterpart of the real world expansion (20.16) looks as

$$\xi_{m,m^\circ}^b = W(D_m^{-2}\Pi_m - D_{m^\circ}^{-2}\Pi_{m^\circ})\check{\nabla}^b$$

with $\check{\nabla}^b = \Psi\mathcal{E}^b\check{\boldsymbol{\varepsilon}}$. Again, as in the linear case, for m° fixed, all the ξ_{m,m°^b are the deterministic linear functions of the score $\check{\nabla}^b$, and this is exactly as for the real world stochastic vectors ξ_{m,m° . Therefore, it suffices to compare the distribution of ∇ and of $\check{\nabla}^b$. These two distributions do not depend on m° and they are close to each other in probability if the small modeling bias condition is fulfilled for the largest considered model p . Therefore, the bootstrap-adjusted critical values z_{m,m° can be used in the real world, and the central propagation result continues to hold: any good model will be accepted with a high probability.

Part IV

Mathematical tools

A

Some results for Gaussian law

Here we collect some simple but useful facts about the properties of the multivariate standard normal distribution. Many similar results can be found in the literature, we present the proofs to keep the presentation self-contained. Everywhere in this Chapter γ means a standard normal vector in \mathbb{R}^p .

A.1 Deviation bounds for a Gaussian vector

The next result describes the tails of the norm of a standard Gaussian vector.

Lemma A.1.1. *Let $\mu \in (0, 1)$. Then for any vector $\lambda \in \mathbb{R}^p$ with $\|\lambda\|^2 \leq p$ and any $r > 0$*

$$\log I\!\!E\{\exp(\lambda^\top \gamma) \mathbb{I}(\|\gamma\| > r)\} \leq -\frac{1-\mu}{2}r^2 + \frac{1}{2\mu}\|\lambda\|^2 + \frac{p}{2}\log(\mu^{-1}). \quad (\text{A.1})$$

Moreover, if $r^2 \geq 4p + 4x$, then

$$I\!\!E\{\exp(\lambda^\top \gamma) \mathbb{I}(\|\gamma\| \leq r)\} \geq e^{\|\lambda\|^2/2}(1 - e^{-x}). \quad (\text{A.2})$$

Proof. We use that for $\mu < 1$

$$I\!\!E\{\exp(\lambda^\top \gamma) \mathbb{I}(\|\gamma\| > r)\} \leq e^{-(1-\mu)r^2/2} I\!\!E \exp\{\lambda^\top \gamma + (1-\mu)\|\gamma\|^2/2\}.$$

It holds

$$\begin{aligned} I\!\!E \exp\{\lambda^\top \gamma + (1-\mu)\|\gamma\|^2/2\} &= (2\pi)^{-p/2} \int \exp\{\lambda^\top \gamma - \mu\|\gamma\|^2/2\} d\gamma \\ &= \mu^{-p/2} \exp(\mu^{-1}\|\lambda\|^2/2) \end{aligned}$$

and (A.1) follows.

Now we apply this result with $\mu = 1/2$. In view of $I\!\!E \exp(\lambda^\top \gamma) = e^{\|\lambda\|^2/2}$, $r^2 \geq 4p + 4x$, and $1 + \log(2) < 2$, it follows for $\|\lambda\|^2 \leq p$

$$\begin{aligned} & e^{-\|\lambda\|^2/2} I\!\!E \left\{ \exp(\lambda^\top \gamma) \mathbb{I}(\|\gamma\| \leq r) \right\} \\ & \geq 1 - \exp \left(-\frac{r^2}{4} + \frac{p + p \log(2)}{2} \right) \geq 1 - \exp(-x) \end{aligned}$$

which implies (A.2).

Lemma A.1.2. *For any $\mathbf{u} \in \mathbb{R}^p$, any unit vector $\mathbf{a} \in \mathbb{R}^p$, and any $z > 0$, it holds*

$$I\!\!P(\|\gamma - \mathbf{u}\| \geq z) \leq \exp\{-z^2/4 + p/2 + \|\mathbf{u}\|^2/2\}, \quad (\text{A.3})$$

$$I\!\!E\{|\gamma^\top \mathbf{a}|^2 \mathbb{I}(\|\gamma - \mathbf{u}\| \geq z)\} \leq (2 + |\mathbf{u}^\top \mathbf{a}|^2) \exp\{-z^2/4 + p/2 + \|\mathbf{u}\|^2/2\}. \quad (\text{A.4})$$

Proof. By the exponential Chebyshev inequality, for any $\lambda < 1$

$$\begin{aligned} I\!\!P(\|\gamma - \mathbf{u}\| \geq z) & \leq \exp(-\lambda z^2/2) I\!\!E \exp(\lambda \|\gamma - \mathbf{u}\|^2/2) \\ & = \exp\left\{-\frac{\lambda z^2}{2} - \frac{p}{2} \log(1 - \lambda) + \frac{\lambda}{2(1 - \lambda)} \|\mathbf{u}\|^2\right\}. \end{aligned}$$

In particular, with $\lambda = 1/2$, this implies (A.3). Further, for $\|\mathbf{a}\| = 1$

$$\begin{aligned} I\!\!E\{|\gamma^\top \mathbf{a}|^2 \mathbb{I}(\|\gamma - \mathbf{u}\| \geq z)\} & \leq \exp(-z^2/4) I\!\!E\{|\gamma^\top \mathbf{a}|^2 \exp(\|\gamma - \mathbf{u}\|^2/4)\} \\ & \leq (2 + |\mathbf{u}^\top \mathbf{a}|^2) \exp(-z^2/4 + p/2 + \|\mathbf{u}\|^2/2) \end{aligned}$$

and (A.4) follows.

A.2 Gaussian integrals

Let A be a linear mapping in \mathbb{R}^p with $\|A\|_{\text{op}} \leq 1$. Given r_0 and c_0 , consider the following ratio of two integrals

$$\frac{\int_{\|A\mathbf{u}\| > r_0} \exp(-c_0 r_0 \|A\mathbf{u}\| + \frac{1}{2} c_0 r_0^2 + \frac{1}{2} \|A\mathbf{u}\|^2 - \frac{1}{2} \|\mathbf{u}\|^2) d\mathbf{u}}{\int_{\|A\mathbf{u}\| \leq r_0} \exp(-\frac{1}{2} \|\mathbf{u}\|^2) d\mathbf{u}}.$$

Obviously, one can rewrite this value as ratio of two expectations

$$\frac{I\!\!E\left\{\exp(-c_0 r_0 \|A\gamma\| + \frac{1}{2} c_0 r_0^2 + \frac{1}{2} \|A\gamma\|^2) \mathbb{I}(\|A\gamma\| > r_0)\right\}}{I\!\!P(\|A\gamma\| \leq r_0)}.$$

Note that without the linear term $-c_0 r_0 \|A\gamma\|$ in the exponent, the expectation in the nominator can be infinite. We aim at describing r_0 and c_0 -values which ensure that the probability in denominator is close to one while the expectation in the nominator is small.

Theorem A.2.1. Let A be a linear operator in \mathbb{R}^p -matrix with $\|A\|_{\text{op}} \leq 1$. Define $p = \text{tr}(A^\top A)$ and $v = (p+x)^{1/2} + (p+x)^{-1/2}$. For any positive C_0, r_0 with $1/2 < C_0 \leq 1$ and $C_0 r_0 > 2v$

$$\mathbb{E}\left\{\exp\left(-C_0 r_0 \|A\gamma\| + \frac{C_0 r_0^2}{2} + \frac{1}{2}\|A\gamma\|^2\right) \mathbb{I}(\|A\gamma\| > r_0)\right\} \leq e^{-(p+x)/2} \quad (\text{A.5})$$

and

$$\mathbb{P}(\|A\gamma\| \leq r_0) \geq 1 - \exp\left\{-\frac{1}{2}(r_0 - \sqrt{p})^2\right\} \geq 1 - e^{-x}. \quad (\text{A.6})$$

Remark A.2.1. The result applies even if the full dimension p is infinite and γ is a Gaussian element in a Hilbert space, provided that $p = \text{tr}(A^\top A)$ is finite, that is, $A^\top A$ is a trace operator.

Proof. Define $r_k \stackrel{\text{def}}{=} r_0 + kp^{-1/2}$ for $k \geq 1$. Represent

$$\begin{aligned} \mathbb{E}\left\{\exp\left(-C_0 r_0 \|A\gamma\| + \frac{1}{2}\|A\gamma\|^2\right) \mathbb{I}(\|A\gamma\| > r_0)\right\} \\ = \sum_{k=1}^{\infty} \mathbb{E}\left\{\exp\left(-C_0 r_0 \|A\gamma\| + \frac{1}{2}\|A\gamma\|^2\right) \mathbb{I}(\|A\gamma\| \in [r_{k-1}, r_k])\right\}. \end{aligned}$$

If $\|Au\| \in [r_{k-1}, r_k]$, then

$$-C_0 r_0 \|Au\| + \frac{1}{2}\|Au\|^2 \leq -C_0 r_0 r_k + \frac{1}{2}r_k^2.$$

This implies

$$\begin{aligned} \mathbb{E}\left\{\exp\left(-C_0 r_0 \|A\gamma\| + \frac{1}{2}\|A\gamma\|^2\right) \mathbb{I}(\|A\gamma\| > r_0)\right\} \\ \leq \sum_{k=1}^{\infty} \exp\left(-C_0 r_0 r_k + \frac{1}{2}r_k^2\right) \mathbb{P}(\|A\gamma\| > r_{k-1}). \end{aligned} \quad (\text{A.7})$$

Now we use the deviation bound (B.3) from Theorem B.1.1: for any x

$$\mathbb{P}(\|A\gamma\| > \sqrt{p} + \sqrt{2x}) \leq e^{-x}.$$

In particular, with x defined by $\sqrt{p} + \sqrt{2x} = r_0$, we obtain (A.6). Further, define x_k by $p^{1/2} + (2x_k)^{1/2} = r_{k-1} = r_k - p^{-1/2}$. It follows by $p^{1/2} + p^{-1/2} \leq (p+x)^{1/2} + (p+x)^{-1/2}$

$$\mathbb{P}(\|A\gamma\| > r_{k-1}) \leq \exp\left\{-\frac{1}{2}(r_{k-1} - \sqrt{p})^2\right\} \leq \exp\left\{-\frac{1}{2}(r_k - v)^2\right\}. \quad (\text{A.8})$$

Plugging this bound in (A.7) yields for $C_0 r_0 \geq 2v$ in view of $r_k = r_0 + kp^{-1/2}$

$$\begin{aligned}
& \mathbb{E} \left\{ \exp \left(-C_0 r_0 \|A\gamma\| + \frac{C_0 r_0^2}{2} + \frac{1}{2} \|A\gamma\|^2 \right) \mathbb{I}(\|A\gamma\| > r_0) \right\} \\
& \leq \sum_{k=0}^{\infty} \exp \left\{ -C_0 r_0 r_k + \frac{C_0 r_0^2}{2} + \frac{1}{2} r_k^2 - \frac{1}{2} (r_k - v)^2 \right\} \\
& = \sum_{k=1}^{\infty} \exp \left\{ - \left(C_0 r_0 - v \right) r_k + \frac{C_0 r_0^2}{2} - \frac{v^2}{2} \right\} \\
& \leq \exp \left(-\frac{v^2}{2} \right) \sum_{k=1}^{\infty} e^{-k v p^{-1/2}} \leq \exp \left(-\frac{p+x}{2} \right).
\end{aligned}$$

This implies (A.5).

Now we slightly extend the result and consider the case of a shifted vector γ .

Theorem A.2.2. Let $b \in I\!\!R^p$, $1/2 < C_0 \leq 1$, and with $v = (p+x)^{1/2} + (p+x)^{-1/2}$

$$C_0 r_0 \geq 2(v + \|Ab\|).$$

Then

$$\begin{aligned}
& \frac{\mathbb{E} \left\{ \exp \left(-C_0 r_0 \|A(\gamma - b)\| + \frac{1}{2} C_0 r_0^2 + \frac{1}{2} \|A(\gamma - b)\|^2 \right) \mathbb{I}(\|A(\gamma - b)\| > r_0) \right\}}{\mathbb{P} \left\{ \|A(\gamma - b)\| \leq r_0 \right\}} \\
& \leq \exp \left(-\frac{p+x}{2} \right).
\end{aligned}$$

Proof. We follow the arguments from the proof of Theorem A.2.1 and uses the same notations. Obviously

$$\mathbb{P} \left\{ \|A(\gamma - b)\| \leq r_0 \right\} \leq \mathbb{P} \left\{ \|A\gamma\| \leq r_0 - \|Ab\| \right\}.$$

Similarly

$$\begin{aligned}
& \mathbb{E} \left\{ \exp \left(-C_0 r_0 \|A(\gamma - b)\| + \frac{1}{2} \|A(\gamma - b)\|^2 \right) \mathbb{I}(\|A(\gamma - b)\| > r_0) \right\} \\
& = \sum_{k=1}^{\infty} \mathbb{E} \left\{ \exp \left(-C_0 r_0 \|A(\gamma - b)\| + \frac{1}{2} \|A(\gamma - b)\|^2 \right) \mathbb{I}(\|A(\gamma - b)\| \in [r_{k-1}, r_k]) \right\} \\
& \leq \sum_{k=1}^{\infty} \exp \left(-C_0 r_0 r_k + \frac{1}{2} r_k^2 \right) \mathbb{P} \left(\|A(\gamma - b)\| > r_{k-1} \right) \\
& \leq \sum_{k=1}^{\infty} \exp \left(-C_0 r_0 r_k + \frac{1}{2} r_k^2 \right) \mathbb{P} \left(\|A\gamma\| \geq r_{k-1} - \|Ab\| \right) \\
& \leq \sum_{k=1}^{\infty} \exp \left\{ -C_0 r_0 r_k + \frac{1}{2} r_k^2 - \frac{1}{2} (r_k - v - \|Ab\|)^2 \right\}.
\end{aligned}$$

Now we can follow the arguments of the proof of Theorem A.2.1.

The next theorem specifies the result for an important special case with $A = I_p$.

Theorem A.2.3. Let $\mathbf{b} \in \mathbb{R}^p$, $1/2 < C_0 \leq 1$, and $C_0 r_0 \geq 2(v + \|\mathbf{b}\|)$ for $v = (p + x)^{1/2} + (p + x)^{-1/2}$. Then

$$\frac{\mathbb{E}\left\{\exp(-C_0 r_0 \|\boldsymbol{\gamma} - \mathbf{b}\| + \frac{1}{2}C_0 r_0^2 + \frac{1}{2}\|\boldsymbol{\gamma} - \mathbf{b}\|^2) \mathbb{I}(\|\boldsymbol{\gamma} - \mathbf{b}\| > r_0)\right\}}{\mathbb{P}(\|\boldsymbol{\gamma} - \mathbf{b}\| \leq r_0)} \leq \exp\left(-\frac{p+x}{2}\right).$$

Now we consider deviation bounds for exponents of Gaussian quadratic forms.

Theorem A.2.4. Let A be a linear operator in \mathbb{R}^p -matrix with $\|A\|_{\text{op}} \leq 1$. Define $p = \text{tr}(A^\top A)$ and $v = (p + x)^{1/2} + (p + x)^{-1/2}$. For any positive C_1, r_0 with $C_1 \leq 1/2$ and $(1 - C_1)r_0 \geq 2v$

$$\mathbb{E}\left\{\exp\left(\frac{C_1}{2}\|A\boldsymbol{\gamma}\|^2\right) \mathbb{I}(\|A\boldsymbol{\gamma}\| > r_0)\right\} \leq e^{-(p+x)/2}. \quad (\text{A.9})$$

Moreover, for any $\mathbf{b} \in \mathbb{R}^p$, it holds under the condition $(1 - C_1)r_0 \geq 2(v + \|\mathbf{b}\|)$

$$\mathbb{E}\left\{\exp\left(\frac{C_1}{2}\|A(\boldsymbol{\gamma} - \mathbf{b})\|^2\right) \mathbb{I}(\|A(\boldsymbol{\gamma} - \mathbf{b})\| > r_0)\right\} \leq e^{-(p+x)/2}.$$

Proof. We proceed similarly to the proof of Theorem A.2.1. Define $r_k \stackrel{\text{def}}{=} r_0 + kp^{-1/2}$ for $k \geq 1$. Represent

$$\begin{aligned} & \mathbb{E}\left\{\exp\left(\frac{C_1}{2}\|A\boldsymbol{\gamma}\|^2\right) \mathbb{I}(\|A\boldsymbol{\gamma}\| > r_0)\right\} \\ &= \sum_{k=1}^{\infty} \mathbb{E}\left\{\exp\left(\frac{C_1}{2}\|A\boldsymbol{\gamma}\|^2\right) \mathbb{I}(\|A\boldsymbol{\gamma}\| \in [r_{k-1}, r_k])\right\} \\ &\leq \sum_{k=1}^{\infty} \exp\left(\frac{C_1}{2}r_k^2\right) \mathbb{P}(\|A\boldsymbol{\gamma}\| > r_{k-1}). \end{aligned} \quad (\text{A.10})$$

Now we use the deviation bound (A.8): for any x

$$\mathbb{P}(\|A\boldsymbol{\gamma}\| > r_{k-1}) \leq \exp\left\{-\frac{1}{2}(r_k - v)^2\right\}.$$

Plugging this bound in (A.10) yields for $(1 - C_1)r_0 \geq 2v$ in view of $r_k = r_0 + kp^{-1/2}$

$$\begin{aligned} & \sum_{k=1}^{\infty} \exp\left\{\frac{C_1 r_k^2}{2} - \frac{1}{2}(r_k - v)^2\right\} \\ &= \sum_{k=1}^{\infty} \exp\left\{-\left(\frac{1-C_1}{2}r_k - v\right)r_k - \frac{v^2}{2}\right\} \\ &\leq \exp\left(-\frac{v^2}{2}\right) \sum_{k=1}^{\infty} e^{-kvp^{-1/2}} \leq \exp\left(-\frac{p+x}{2}\right). \end{aligned}$$

This implies (A.9).

B

Deviation bounds for quadratic forms

Here we collect some probability bounds for Gaussian quadratic forms.

B.1 Gaussian quadratic forms

The next result explains the concentration effect of $\gamma^\top B\gamma$ for a standard Gaussian vector γ and a symmetric matrix B . We use a version from [Laurent and Massart \(2000\)](#) with a complete proof.

Theorem B.1.1. *Let γ be a standard normal Gaussian vector and B be symmetric positively definite $p \times p$ -matrix. Then with $p = \text{tr}(B)$, $v^2 = \text{tr}(B^2)$, and $\lambda = \|B\|_{\text{op}}$, it holds for each $x \geq 0$*

$$\mathbb{P}(\gamma^\top B\gamma > z^2(B, x)) \leq e^{-x}, \quad (\text{B.1})$$

$$z(B, x) \stackrel{\text{def}}{=} \sqrt{p + 2vx^{1/2} + 2\lambda x}. \quad (\text{B.2})$$

In particular, it implies

$$\mathbb{P}(\|B^{1/2}\gamma\| > p^{1/2} + (2\lambda x)^{1/2}) \leq e^{-x}. \quad (\text{B.3})$$

Also

$$\mathbb{P}(\gamma^\top B\gamma < p - 2vx^{1/2}) \leq e^{-x}. \quad (\text{B.4})$$

If B is symmetric but non necessarily positive then

$$\mathbb{P}(|\gamma^\top B\gamma - p| > 2vx^{1/2} + 2\lambda x) \leq 2e^{-x}.$$

Proof. Normalisation by λ reduces the statement to the case with $\lambda = 1$. Further, the standard rotating arguments allow to reduce the Gaussian quadratic form $\|\gamma\|^2$ to the chi-squared form:

$$\boldsymbol{\gamma}^\top B \boldsymbol{\gamma} = \sum_{j=1}^p \lambda_j \nu_j^2$$

with independent standard normal r.v.'s ν_j . Here $\lambda_j \in [0, 1]$ are eigenvalues of B , and $p = \lambda_1 + \dots + \lambda_p$, $v^2 = \lambda_1^2 + \dots + \lambda_p^2$. One can easily compute the exponential moment of $(\boldsymbol{\gamma}^\top B \boldsymbol{\gamma} - p)/2$: for each positive $\mu < 1$

$$\log E \exp\{\mu(\boldsymbol{\gamma}^\top B \boldsymbol{\gamma} - p)/2\} = \frac{1}{2} \sum_{j=1}^p \{-\mu\lambda_j - \log(1 - \mu\lambda_j)\}. \quad (\text{B.5})$$

Lemma B.1.1. *Let $\mu\lambda_j < 1$ and $\lambda_j \leq 1$. Then*

$$\frac{1}{2} \sum_{j=1}^p \{-\mu\lambda_j - \log(1 - \mu\lambda_j)\} \leq \frac{\mu^2 v^2}{4(1 - \mu)}.$$

Proof. In view of $\mu\lambda_j < 1$, it holds for every j

$$\begin{aligned} -\mu\lambda_j - \log(1 - \mu\lambda_j) &= \sum_{k=2}^{\infty} \frac{(\mu\lambda_j)^k}{k} \\ &\leq \frac{(\mu\lambda_j)^2}{2} \sum_{k=0}^{\infty} (\mu\lambda_j)^k \leq \frac{(\mu\lambda_j)^2}{2(1 - \mu\lambda_j)} \leq \frac{(\mu\lambda_j)^2}{2(1 - \mu)}, \end{aligned} \quad (\text{B.6})$$

and thus

$$\frac{1}{2} \sum_{j=1}^p \{-\mu\lambda_j - \log(1 - \mu\lambda_j)\} \leq \sum_{j=1}^p \frac{(\mu\lambda_j)^2}{4(1 - \mu)} \leq \frac{\mu^2 v^2}{4(1 - \mu)}.$$

The next technical lemma is helpful.

Lemma B.1.2. *For each $v > 0$ and $x > 0$, it holds*

$$\inf_{\mu > 0} \left\{ -\mu(vx^{1/2} + x) + \frac{\mu^2 v^2}{4(1 - \mu)} \right\} \leq -x.$$

Proof. Let pick up

$$\mu = 1 - \frac{1}{2x^{1/2}/v + 1} = \frac{x^{1/2}}{x^{1/2} + v/2},$$

so that $\mu/(1 - \mu) = 2x^{1/2}/v$. Then

$$\begin{aligned} -\mu(vx^{1/2} + x) + \frac{\mu^2 v^2}{4(1 - \mu)} \\ &= -\mu(vx^{1/2} + x + v^2/4) + \frac{\mu v^2}{4(1 - \mu)} \\ &= -\frac{x^{1/2}}{x^{1/2} + v/2} (x^{1/2} + v/2)^2 + \frac{2x^{1/2}v}{4} = -x \end{aligned} \quad (\text{B.7})$$

and the result follows.

Now we apply the Markov inequality

$$\begin{aligned} \log I\!P(\boldsymbol{\gamma}^\top B\boldsymbol{\gamma} > p + 2vx^{1/2} + x) &= \log I\!P((\boldsymbol{\gamma}^\top B\boldsymbol{\gamma} - p)/2 > vx^{1/2} + x) \\ &\leq \inf_{\mu>0} \left\{ -\mu(vx^{1/2} + x) + \log I\!E \exp\{\mu(\boldsymbol{\gamma}^\top B\boldsymbol{\gamma} - p)/2\} \right\} \\ &\leq \inf_{\mu>0} \left\{ -\mu(vx^{1/2} + x) + \frac{\mu^2 v^2}{4(1-\mu)} \right\} \leq -x \end{aligned}$$

and the first assertion (B.1) follows. The second statement follows from the first one by $\text{tr}(B^2) \leq \|B\|_{\text{op}} \text{tr}(B) = \lambda p$.

Similarly for any $\mu > 0$

$$I\!P(\boldsymbol{\gamma}^\top B\boldsymbol{\gamma} - p < -2v\sqrt{x}) \leq \exp(-\mu v\sqrt{x}) I\!E \exp\left(-\frac{\mu}{2}(\boldsymbol{\gamma}^\top B\boldsymbol{\gamma} - p)\right).$$

By (B.5)

$$\log I\!E \exp\{-\mu(\boldsymbol{\gamma}^\top B\boldsymbol{\gamma} - p)/2\} = \frac{1}{2} \sum_{j=1}^p \{\mu\lambda_j - \log(1 + \mu\lambda_j)\}.$$

and

$$\frac{1}{2} \sum_{j=1}^p \{\mu\lambda_j - \log(1 + \mu\lambda_j)\} = \frac{1}{2} \sum_{j=1}^p \sum_{k=2}^{\infty} \frac{(-\mu\lambda_j)^k}{k} \leq \sum_{j=1}^p \frac{(\mu\lambda_j)^2}{4} = \frac{\mu^2 v^2}{4}.$$

Here the choice $\mu = 2\sqrt{x}/v$ yields (B.4).

One can put together the arguments used for obtaining the lower and the upper bound for getting a bound for a general quadratic form $\boldsymbol{\gamma}^\top B\boldsymbol{\gamma}$, where B is symmetric but not necessarily positive.

Finally we apply this result to weighted sums of centered γ_i^2 .

Corollary B.1.1. *For any unit vector $\mathbf{u} = (u_i) \in I\!R^n$ and standard normal r.v.'s γ_i , it holds with $\|\mathbf{u}\|_\infty \stackrel{\text{def}}{=} \max_i |u_i|$*

$$I\!P\left(\left|\sum_{i=1}^n u_i(\gamma_i^2 - 1)\right| \geq 2x^{1/2} + 2\|\mathbf{u}\|_\infty x\right) \leq 2e^{-x}.$$

Proof. The statement follows directly from Theorem B.1.1. It suffices to notice $v^2 = \|\mathbf{u}\|^2 = 1$.

As a special case, we present a bound for the chi-squared distribution corresponding to $B = I_p$. Then $\text{tr}(B) = p$, $\text{tr}(B^2) = p$ and $\lambda(B) = 1$.

Corollary B.1.2. Let γ be a standard normal vector in \mathbb{R}^p . Then

$$\begin{aligned} \mathbb{P}(\|\gamma\|^2 \geq p + 2\sqrt{px} + 2x) &\leq e^{-x}, \\ \mathbb{P}(\|\gamma\|^2 \leq p - 2\sqrt{px}) &\leq e^{-x}, \\ \mathbb{P}(\|\gamma\| \geq \sqrt{p} + \sqrt{2x}) &\leq e^{-x}. \end{aligned} \tag{B.8}$$

The previous results are mainly stated for a standard Gaussian vector $\gamma \in \mathbb{R}^n$. Now we extend it to the case of a zero mean Gaussian vector ξ with the $n \times n$ covariance matrix $\mathbb{V} = (\sigma_{ij})$ with $\lambda_{\max}(\mathbb{V}) \leq \lambda^*$. Given a unit vector $\mathbf{u} = (u_1, \dots, u_n)^\top \in \mathbb{R}^n$, consider the quadratic form

$$Q = \sum_{i=1}^n u_i \xi_i^2.$$

We aim at bounding $Q - \mathbb{E}Q$. To apply the result of Theorem B.1.1 represent Q as $\gamma^\top B \gamma$ with B depending on \mathbf{u} and \mathbb{V} . More precisely, let $\xi = \mathbb{V}^{1/2} \gamma$ for a standard Gaussian vector $\gamma \in \mathbb{R}^n$. Then with $\mathbf{U} = \text{diag}(u_1, \dots, u_n)$, it holds

$$S = \text{tr}(\mathbf{U} \xi \xi^\top) = \text{tr}(\mathbf{U} \mathbb{V}^{1/2} \gamma \gamma^\top \mathbb{V}^{1/2}) = \text{tr}(B \gamma \gamma^\top) = \gamma^\top B \gamma$$

with $B = \mathbb{V}^{1/2} \mathbf{U} \mathbb{V}^{1/2}$. Therefore, the bound $\|\mathbb{V}\|_{\text{op}} \leq \lambda^*$ implies

$$\begin{aligned} \lambda = \lambda(B) &= \|\mathbb{V}^{1/2} \mathbf{U} \mathbb{V}^{1/2}\|_{\text{op}} \leq \lambda^* \|\mathbf{u}\|_\infty, \\ v^2 = \text{tr}(B^2) &= \text{tr}(\mathbb{V}^{1/2} \mathbf{U} \mathbb{V} \mathbf{U} \mathbb{V}^{1/2}) \leq \lambda^* \text{tr}(\mathbf{U} \mathbb{V} \mathbf{U}) \leq \lambda^{*2} \|\mathbf{u}\|^2 = \lambda^{*2}. \end{aligned}$$

Now the general results of Theorem B.1.1 implies the result similar to Corollary B.1.1.

Corollary B.1.3. For any unit vector $\mathbf{u} = (u_i) \in \mathbb{R}^n$, $\|\mathbf{u}\| = 1$, and normal zero mean vector $\xi \sim \mathcal{N}(0, \mathbb{V})$ in \mathbb{R}^n with $\|\mathbb{V}\|_{\text{op}} \leq \lambda^*$, it holds

$$\mathbb{P}\left(\left|\sum_{i=1}^n u_i (\xi_i^2 - \mathbb{E} \xi_i^2)\right| \geq 2\lambda^* x^{1/2} + 2\lambda^* \|\mathbf{u}\|_\infty x\right) \leq 2e^{-x}.$$

It is worth noting that the identity $\|\mathbf{u}\| = 1$ implies $\|\mathbf{u}\|_\infty \leq 1$. Moreover, in typical situations, $\|\mathbf{u}\|_\infty \asymp n^{-1/2}$, and the leading term in the bounds of Corollaries B.1.1 and B.1.3 is $2\lambda^* x^{1/2}$.

B.2 Deviation bounds for non-Gaussian quadratic forms

This section presents an extension of the results obtained for Gaussian quadratic forms to the non-Gaussian case.

B.2.1 Deviation bounds for the norm of a standardized non-Gaussian vector

The bounds of Corollary B.1.2 heavily use normality of the vector ξ . This section extends the upper bound (B.8) to the case when ξ has some exponential moments. More exactly, suppose for some fixed $g > 0$ that

$$\log I\!E \exp(\gamma^\top \xi) \leq \|\gamma\|^2/2, \quad \gamma \in I\!R^p, \|\gamma\| \leq g. \quad (\text{B.9})$$

For ease of presentation, assume below that g is sufficiently large, namely, $0.3g \geq \sqrt{p}$. In typical examples of an i.i.d. sample, $g \asymp \sqrt{n}$. Define

$$\begin{aligned} x_c &\stackrel{\text{def}}{=} g^2/4, \\ z_c^2 &\stackrel{\text{def}}{=} p + \sqrt{pg^2} + g^2/2 = g^2(1/2 + \sqrt{p/g^2} + p/g^2), \\ g_c &\stackrel{\text{def}}{=} \frac{g(1/2 + \sqrt{p/g^2} + p/g^2)^{1/2}}{1 + \sqrt{p/g^2}}. \end{aligned}$$

Note that with $\alpha = \sqrt{p/g^2} \leq 0.3$, one has

$$\begin{aligned} z_c^2 &= g^2(1/2 + \alpha + \alpha^2), \\ g_c &= g \frac{(1/2 + \alpha + \alpha^2)^{1/2}}{1 + \alpha} \end{aligned}$$

so that $z_c^2/g^2 \in [1/2, 1]$ and $g_c^2/g^2 \in [1/2, 1]$.

Theorem B.2.1. *Let (B.9) hold and $0.3g \geq \sqrt{p}$. Then for each $x > 0$*

$$I\!P(\|\xi\| \geq z(p, x)) \leq 2e^{-x} + 8.4e^{-x_c} \mathbb{I}(x < x_c), \quad (\text{B.10})$$

where $z(p, x)$ is defined by

$$z(p, x) \stackrel{\text{def}}{=} \begin{cases} (p + 2\sqrt{px} + 2x)^{1/2}, & x \leq x_c, \\ z_c + 2g_c^{-1}(x - x_c), & x > x_c. \end{cases}$$

Depending on the value x , we have two types of tail behavior of the quadratic form $\|\xi\|^2$. For $x \leq x_c = g^2/4$, we have the same deviation bounds as in the Gaussian case with the extra-factor two in the deviation probability. Remind that one can use a simplified expression $(p + 2\sqrt{px} + 2x)^{1/2} \leq \sqrt{p} + \sqrt{2x}$. For $x > x_c$, we switch to the special regime driven by the exponential moment condition (B.9). Usually g^2 is a large number (of order n in the i.i.d. setup) and the second term in (B.10) can be simply ignored.

The main step of the proof is the following exponential bound.

Lemma B.2.1. Suppose (B.9). For any $\mu < 1$ with $g^2 > p\mu$, it holds

$$\mathbb{E} \exp\left(\frac{\mu\|\xi\|^2}{2}\right) \mathbb{I}(\|\xi\| \leq g/\mu - \sqrt{p/\mu}) \leq 2(1-\mu)^{-p/2}. \quad (\text{B.11})$$

Proof. Let ε be a standard normal vector in \mathbb{R}^p and $\mathbf{u} \in \mathbb{R}^p$. The bound $\mathbb{P}(\|\varepsilon\|^2 > p) \leq 1/2$ and the triangle inequality imply for any vector \mathbf{u} and any \mathbf{r} with $\mathbf{r} \geq \|\mathbf{u}\| + p^{1/2}$ that $\mathbb{P}(\|\mathbf{u} + \varepsilon\| \leq \mathbf{r}) \geq 1/2$. Let us fix some ξ with $\|\xi\| \leq g/\mu - \sqrt{p/\mu}$ and denote by \mathbb{P}_ξ the conditional probability given ξ . The previous arguments yield:

$$\mathbb{P}_\xi(\|\varepsilon + \mu^{1/2}\xi\| \leq \tau g) \geq 0.5.$$

It holds with $c_p = (2\pi)^{-p/2}$

$$\begin{aligned} c_p \int \exp\left(\gamma^\top \xi - \frac{\|\gamma\|^2}{2\mu}\right) \mathbb{I}(\|\gamma\| \leq g) d\gamma \\ = c_p \exp(\mu\|\xi\|^2/2) \int \exp\left(-\frac{1}{2}\|\tau\gamma - \mu^{1/2}\xi\|^2\right) \mathbb{I}(\tau\|\gamma\| \leq \tau g) d\gamma \\ = \mu^{p/2} \exp(\mu\|\xi\|^2/2) \mathbb{P}_\xi(\|\varepsilon + \mu^{1/2}\xi\| \leq \tau g) \\ \geq 0.5\mu^{p/2} \exp(\mu\|\xi\|^2/2), \end{aligned}$$

because $\|\mu^{1/2}\xi\| + p^{1/2} \leq \tau g$. This implies in view of $p < g^2/\mu$ that

$$\begin{aligned} \exp(\mu\|\xi\|^2/2) \mathbb{I}(\|\xi\|^2 \leq g/\mu - \sqrt{p/\mu}) \\ \leq 2\mu^{-p/2} c_p \int \exp\left(\gamma^\top \xi - \frac{\|\gamma\|^2}{2\mu}\right) \mathbb{I}(\|\gamma\| \leq g) d\gamma. \end{aligned}$$

Further, by (B.9)

$$\begin{aligned} c_p \mathbb{E} \int \exp\left(\gamma^\top \xi - \frac{1}{2\mu}\|\gamma\|^2\right) \mathbb{I}(\|\gamma\| \leq g) d\gamma \\ \leq c_p \int \exp\left(-\frac{\mu^{-1}-1}{2}\|\gamma\|^2\right) \mathbb{I}(\|\gamma\| \leq g) d\gamma \\ \leq c_p \int \exp\left(-\frac{\mu^{-1}-1}{2}\|\gamma\|^2\right) d\gamma \\ \leq (\mu^{-1}-1)^{-p/2} \end{aligned}$$

and (B.11) follows.

Due to this result, the scaled squared norm $\mu\|\xi\|^2/2$ after a proper truncation possesses the same exponential moments as in the Gaussian case. A straightforward implication is the probability bound $\mathbb{P}(\|\xi\|^2 > p+u)$ with $u = 2\sqrt{p\mathbf{x}} + 2\mathbf{x}$. Namely, given \mathbf{x} , define

$$\mu = \mu(x) = \frac{1}{1 + 0.5\sqrt{p/x}}. \quad (\text{B.12})$$

Also define for $x_c = g^2/4$

$$\mu_c \stackrel{\text{def}}{=} \mu(x_c) = \frac{1}{1 + \sqrt{p/g^2}}. \quad (\text{B.13})$$

Obviously, $\mu \leq \mu_c$ for $x \leq x_c$. Now we obtain similarly to the Gaussian case in Lemma B.1.2 for $u = 2\sqrt{px} + 2x$

$$\begin{aligned} & \mathbb{P}\left(\|\xi\|^2 > p + u, \|\xi\| \leq g/\mu - \sqrt{p/\mu}\right) \\ & \leq \exp\left\{-\frac{\mu(p+u)}{2}\right\} \mathbb{E} \exp\left(\frac{\mu\|\xi\|^2}{2}\right) \mathbb{I}\left(\|\xi\| \leq g/\mu - \sqrt{p/\mu}\right) \\ & \leq 2 \exp\left\{-\frac{1}{2}[\mu(p+u) + p \log(1-\mu)]\right\} \end{aligned} \quad (\text{B.14})$$

and by (B.7) with $v^2 = p$, it holds for μ from (B.12)

$$\mu(p + 2\sqrt{px} + 2x) + p \log(1 - \mu) \geq 2x. \quad (\text{B.15})$$

Now we show that the constraint $\|\xi\| \leq g/\mu - \sqrt{p/\mu}$ in (B.14) can be replaced by the inequality $\|\xi\| \leq z_c$.

Lemma B.2.2. *Let $0.3g \geq \sqrt{p}$, $x \leq x_c = g^2/4$, and $\mu = 1/(1 + 0.5\sqrt{p/x})$. Then*

$$\begin{aligned} & p + 2\sqrt{px} + 2x \leq p + 2\sqrt{px_c} + 2x_c, \\ & g/\mu - \sqrt{p/\mu} \geq g/\mu_c - \sqrt{p/\mu_c}, \\ & p + 2\sqrt{px_c} + 2x_c \leq (g/\mu_c - \sqrt{p/\mu_c})^2. \end{aligned} \quad (\text{B.16})$$

Proof. The definition implies $\mu \leq \mu_c$ for $x \leq x_c$ and thus the first two inequalities of the lemma are obvious. Therefore, it remains to check (B.16). Denote $\alpha^2 = p/g^2$. Then $\mu_c^{-1} = 1 + \alpha$ and

$$g/\mu_c - \sqrt{p/\mu_c} = \mu_c^{-1}g(1 - \sqrt{\mu_c\alpha^2}) = g(1 + \alpha)\{1 - \sqrt{\alpha^2/(1 + \alpha)}\}.$$

For $x_c = g^2/4$, it holds

$$p + 2\sqrt{px_c} + 2x_c = p + \sqrt{pg^2} + g^2/2 = g^2(\alpha^2 + \alpha + 1/2).$$

Direct calculus shows that for $\alpha \leq 0.3$ one can bound

$$\alpha^2 + \alpha + 1/2 \leq (1 + \alpha)^2 \left\{1 - \sqrt{\alpha^2/(1 + \alpha)}\right\}^2 \quad (\text{B.17})$$

and this proves (B.16).

We conclude from this lemma, (B.14) and (B.15) that

$$\mathbb{P}(\|\xi\|^2 > p + 2\sqrt{px} + 2x, \|\xi\| \leq z_c) \leq 2e^{-x}.$$

If (B.9) holds with $g = \infty$, then we are back in the (sub-)Gaussian case with $z_c = \infty$. In the non-Gaussian case with a finite g , we have to accompany the moderate deviation bound with a large deviation bound $\mathbb{P}(\|\xi\| > z)$ for $z \geq z_c$. This is done by combining the bound (B.11) with the standard slicing arguments.

Lemma B.2.3. Define $g_c = \mu_c z_c$; see (B.13). It holds for $z \geq z_c$

$$\mathbb{P}(\|\xi\| > z) \leq 8.4(1 - g_c/z)^{-p/2} \exp(-g_c z/2) \quad (\text{B.18})$$

$$\leq 8.4 \exp\{-x_c - g_c(z - z_c)/2\}. \quad (\text{B.19})$$

Proof. For a fixed $z \geq z_c$, consider the growing sequence (y_k) with $y_1 = z$ and

$$y_{k+1} = z + k/g_c.$$

Define also $\mu_k = g_c/y_k$. Then the sequence (μ_k) is decreasing, in particular, $\mu_k \leq \mu_1 = g_c/z \leq g_c$. Obviously

$$\mathbb{P}(\|\xi\| > z) = \sum_{k=1}^{\infty} \mathbb{P}(\|\xi\| > y_k, \|\xi\| \leq y_{k+1}).$$

Now we try to evaluate every slicing probability in this expression. We use that

$$\mu_{k+1} y_k^2 = \frac{(g_c z + k - 1)^2}{g_c z + k} \geq g_c z + k - 2.$$

Lemma B.2.2 implies $g - \sqrt{\mu_c p} \geq \mu_c z_c = g_c$. This yields $g/\mu_k - \sqrt{p/\mu_k} \geq y_k$ because

$$g/\mu_k - \sqrt{p/\mu_k} - y_k = \mu_k^{-1}(g - \sqrt{\mu_k p} - g_c) \geq \mu_k^{-1}(g - \sqrt{\mu_c p} - g_c) \geq 0.$$

Hence by (B.11)

$$\begin{aligned}
I\!P(\|\boldsymbol{\xi}\| > z) &= \sum_{k=1}^{\infty} I\!P(\|\boldsymbol{\xi}\| > y_k, \|\boldsymbol{\xi}\| \leq y_{k+1}) \\
&\leq \sum_{k=1}^{\infty} \exp\left(-\frac{\mu_{k+1}y_k^2}{2}\right) I\!E \exp\left(\frac{\mu_{k+1}\|\boldsymbol{\xi}\|^2}{2}\right) \mathbb{I}\left(\|\boldsymbol{\xi}\| \leq \frac{g}{\mu_{k+1}} - \sqrt{\frac{p}{\mu_{k+1}}}\right) \\
&\leq \sum_{k=1}^{\infty} 2(1-\mu_{k+1})^{-p/2} \exp\left(-\frac{\mu_{k+1}y_k^2}{2}\right) \\
&\leq 2(1-\mu_1)^{-p/2} \sum_{k=1}^{\infty} \exp\left(-\frac{g_c z + k - 2}{2}\right) \\
&= 2e^{1/2}(1-e^{-1/2})^{-1}(1-\mu_1)^{-p/2} \exp(-g_c z/2) \\
&\leq 8.4(1-g_c/z)^{-p/2} \exp(-g_c z/2)
\end{aligned}$$

and the assertion (B.18) follows. For $z = z_c$, it holds by (B.15)

$$g_c z_c + p \log(1 - \mu_c) = \mu_c z_c^2 + p \log(1 - \mu_c) \geq 2x_c$$

and (B.18) implies $I\!P(\|\boldsymbol{\xi}\| > z_c) \leq 8.4 \exp(-x_c)$. Now observe that the function $f(z) = g_c z / 2 + (p/2) \log(1 - g_c/z)$ fulfills $f(z_c) = x_c$ and $f'(z) \geq g_c/2$ yielding $f(z) \geq x_c + g_c(z - y_0)/2$. This implies (B.19).

Now we can conclude that for $x \geq x_c$, the choice

$$z = z(x) = 2g_c^{-1}(x - x_c) + z_c$$

implies

$$I\!P(\|\boldsymbol{\xi}\| > z(x)) \leq 8.4e^{-x}. \quad (\text{B.20})$$

The statement of the theorem is obtained by a simple combination of (B.15) and (B.20).

B.2.2 A deviation bound for a general non-Gaussian quadratic form

This section presents a bound for a quadratic form $\boldsymbol{\xi}^\top B \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ satisfies (B.9) and B is a given symmetric positive $p \times p$ matrix. Define

$$p \stackrel{\text{def}}{=} \text{tr}(B), \quad v^2 \stackrel{\text{def}}{=} \text{tr}(B^2), \quad \lambda \stackrel{\text{def}}{=} \lambda_{\max}(B).$$

For ease of presentation, suppose that $0.3g \geq \sqrt{p}$ so that $\alpha = \sqrt{p/g^2} \leq 0.3$. The other case only changes the constants in the inequalities. Define also

$$\begin{aligned} \mathbf{x}_c &\stackrel{\text{def}}{=} \mathbf{g}^2/4, \\ z_c^2 &\stackrel{\text{def}}{=} \mathbf{p} + \mathbf{v}\mathbf{g} + \lambda\mathbf{g}^2/2, \\ \mathbf{g}_c &\stackrel{\text{def}}{=} \frac{\sqrt{\mathbf{p}/\lambda + \mathbf{g}\mathbf{v}/\lambda + \mathbf{g}^2/2}}{1 + \mathbf{v}/(\lambda\mathbf{g})}. \end{aligned}$$

Theorem B.2.2. Let (B.9) hold and $0.3\mathbf{g} \geq \sqrt{\mathbf{p}/\lambda}$. Then for each $\mathbf{x} > 0$

$$\mathbb{I}\mathcal{P}(\|B^{1/2}\boldsymbol{\xi}\| \geq z(B, \mathbf{x})) \leq 2e^{-\mathbf{x}} + 8.4e^{-\mathbf{x}_c} \mathbb{I}(\mathbf{x} < \mathbf{x}_c),$$

where $z(B, \mathbf{x})$ is defined by

$$z(B, \mathbf{x}) \stackrel{\text{def}}{=} \begin{cases} \sqrt{\mathbf{p} + 2\mathbf{v}\mathbf{x}^{1/2} + 2\lambda\mathbf{x}}, & \mathbf{x} \leq \mathbf{x}_c, \\ z_c + 2\lambda(\mathbf{x} - \mathbf{x}_c)/\mathbf{g}_c, & \mathbf{x} > \mathbf{x}_c. \end{cases} \quad (\text{B.22})$$

Similarly to the Gaussian case, the upper quantile $z(B, \mathbf{x}) = \sqrt{\mathbf{p} + 2\mathbf{v}\mathbf{x}^{1/2} + 2\lambda\mathbf{x}}$ can be upper bounded by $\sqrt{\mathbf{p}} + \sqrt{2\lambda\mathbf{x}}$:

$$z(B, \mathbf{x}) \leq \begin{cases} \sqrt{\mathbf{p}} + \sqrt{2\lambda\mathbf{x}}, & \mathbf{x} \leq \mathbf{x}_c, \\ z_c + 2\lambda(\mathbf{x} - \mathbf{x}_c)/\mathbf{g}_c, & \mathbf{x} > \mathbf{x}_c. \end{cases} \quad (\text{B.23})$$

The main steps of the proof are similar to the proof of Theorem B.2.1. Normalization by λ reduces the statement to the case $\lambda = 1$ which we assume below. Moreover, the standard change-of-basis arguments allow us to reduce the problem to the case of a diagonal matrix $B = \text{diag}(a_1, \dots, a_p)$, where $1 = a_1 \geq a_2 \geq \dots \geq a_p > 0$. Note that $\mathbf{p} = a_1 + \dots + a_p$ and $\mathbf{v}^2 = a_1^2 + \dots + a_p^2$.

Lemma B.2.4. Suppose (B.9) and $\|B\|_{\text{op}} = 1$. For any $\mu < 1$ with $\mathbf{g}^2/\mu \geq \mathbf{p}$, it holds

$$\mathbb{I}\mathcal{E} \exp(\mu\|B^{1/2}\boldsymbol{\xi}\|^2/2) \mathbb{I}(\|B\boldsymbol{\xi}\| \leq \mathbf{g}/\mu - \sqrt{\mathbf{p}/\mu}) \leq 2\det(I_p - \mu B)^{-1/2}. \quad (\text{B.24})$$

Proof. With $c_p(B) = (2\pi)^{-p/2} \det(B^{-1/2})$

$$\begin{aligned} c_p(B) \int \exp\left(\boldsymbol{\gamma}^\top \boldsymbol{\xi} - \frac{1}{2\mu} \|B^{-1/2}\boldsymbol{\gamma}\|^2\right) \mathbb{I}(\|\boldsymbol{\gamma}\| \leq \mathbf{g}) d\boldsymbol{\gamma} \\ = c_p(B) \exp\left(\frac{\mu\|B^{1/2}\boldsymbol{\xi}\|^2}{2}\right) \int \exp\left(-\frac{1}{2}\|\mu^{1/2}B^{1/2}\boldsymbol{\xi} - \tau B^{-1/2}\boldsymbol{\gamma}\|^2\right) \mathbb{I}(\|\boldsymbol{\gamma}\| \leq \mathbf{g}) d\boldsymbol{\gamma} \\ = \mu^{p/2} \exp\left(\frac{\mu\|B^{1/2}\boldsymbol{\xi}\|^2}{2}\right) \mathbb{I}\mathcal{P}_{\boldsymbol{\xi}}(\|\tau B^{1/2}\boldsymbol{\varepsilon} + B^{1/2}\boldsymbol{\xi}\| \leq \mathbf{g}/\mu), \end{aligned}$$

where $\boldsymbol{\varepsilon}$ denotes a standard normal vector in \mathbb{R}^p and $\mathbb{I}\mathcal{P}_{\boldsymbol{\xi}}$ means the conditional probability given $\boldsymbol{\xi}$. Moreover, for any $\mathbf{u} \in \mathbb{R}^p$ and $\mathbf{r} \geq \mathbf{p}^{1/2} + \|\mathbf{u}\|$, it holds in view of $\mathbb{I}\mathcal{P}(\|B^{1/2}\boldsymbol{\varepsilon}\|^2 > \mathbf{p}) \leq 1/2$

$$\mathbb{P}(\|B^{1/2}\boldsymbol{\varepsilon} - \mathbf{u}\| \leq r) \geq \mathbb{P}(\|B^{1/2}\boldsymbol{\varepsilon}\| \leq \sqrt{p}) \geq 1/2.$$

This implies

$$\begin{aligned} & \exp\left(\mu\|B^{1/2}\boldsymbol{\xi}\|^2/2\right) \mathbb{I}(\|B\boldsymbol{\xi}\| \leq g/\mu - \sqrt{p}/\mu) \\ & \leq 2\mu^{-p/2}c_p(B) \int \exp\left(\boldsymbol{\gamma}^\top \boldsymbol{\xi} - \frac{1}{2\mu}\|B^{-1/2}\boldsymbol{\gamma}\|^2\right) \mathbb{I}(\|\boldsymbol{\gamma}\| \leq g) d\boldsymbol{\gamma}. \end{aligned}$$

Further, by (B.9)

$$\begin{aligned} & c_p(B)\mathbb{E} \int \exp\left(\boldsymbol{\gamma}^\top \boldsymbol{\xi} - \frac{1}{2\mu}\|B^{-1/2}\boldsymbol{\gamma}\|^2\right) \mathbb{I}(\|\boldsymbol{\gamma}\| \leq g) d\boldsymbol{\gamma} \\ & \leq c_p(B) \int \exp\left(\frac{\|\boldsymbol{\gamma}\|^2}{2} - \frac{1}{2\mu}\|B^{-1/2}\boldsymbol{\gamma}\|^2\right) d\boldsymbol{\gamma} \\ & \leq \det(B^{-1/2}) \det(\mu^{-1}B^{-1} - I_p)^{-1/2} = \mu^{p/2} \det(I_p - \mu B)^{-1/2} \end{aligned}$$

and (B.24) follows.

Now we evaluate the probability $\mathbb{P}(\|B^{1/2}\boldsymbol{\xi}\| > y)$ for moderate values of y . Given $x \leq x_c = g^2/4$, define

$$\mu = \mu(x) = \frac{1}{1 + 0.5\sqrt{x}} , \quad (\text{B.25})$$

$$\mu_c \stackrel{\text{def}}{=} \frac{1}{1 + 0.5\sqrt{x_c}} = \frac{1}{1 + \sqrt{g}} . \quad (\text{B.26})$$

Obviously $\mu \leq \mu_c$. Now we obtain similarly to the Gaussian case in Lemma B.1.2 for $u = 2\sqrt{x} + 2x$

$$\begin{aligned} & \mathbb{P}\left(\|B^{1/2}\boldsymbol{\xi}\|^2 > p + u, \|\boldsymbol{\xi}\| \leq g/\mu - \sqrt{p}/\mu\right) \\ & \leq \exp\left\{-\frac{\mu(p+u)}{2}\right\} \mathbb{E} \exp\left(\frac{\mu\|\boldsymbol{\xi}\|^2}{2}\right) \mathbb{I}(\|\boldsymbol{\xi}\| \leq g/\mu - \sqrt{p}/\mu) \\ & \leq 2 \exp\left\{-\frac{1}{2}[\mu(p+u) - \log \det(I_p - \mu B)]\right\} \end{aligned} \quad (\text{B.27})$$

and by (B.7), it holds for μ from (B.25)

$$\mu(p + 2\sqrt{x} + 2x) + \log \det(I_p - \mu B) \geq 2x.$$

Now we show that the constraint $\|\boldsymbol{\xi}\| \leq g/\mu - \sqrt{p}/\mu$ in (B.27) can be replaced by the inequality $\|\boldsymbol{\xi}\| \leq z_c$. Indeed, the definition implies $\mu \leq \mu_c$ for $x \leq x_c$ and

$$p + 2\sqrt{x} + 2x \leq p + 2\sqrt{x_c} + 2x_c,$$

$$g/\mu - \sqrt{p}/\mu \geq g/\mu_c - \sqrt{p}/\mu_c.$$

It remains to show that

$$p + 2v\sqrt{x_c} + 2x_c \leq (g/\mu_c - \sqrt{p/\mu_c})^2. \quad (\text{B.28})$$

Denote $\alpha^2 = p/g^2$. By $v^2 \leq p$ and $x_c = g^2/4$, it holds $\mu_c^{-1} = 1 + 0.5v\chi_c^{-1/2} \leq 1 + \alpha$ and

$$g/\mu_c - \sqrt{p/\mu_c} = \mu_c^{-1}g(1 - \sqrt{\mu_c\alpha^2}) \geq g(1 + \alpha)\{1 - \sqrt{\alpha^2/(1 + \alpha)}\}.$$

Also in a similar way

$$p + 2v\sqrt{x_c} + 2x_c \leq p + \sqrt{pg^2} + g^2/2 = g^2(\alpha^2 + \alpha + 1/2).$$

This and (B.17) prove (B.28) yielding

$$\mathbb{P}(\|B^{1/2}\xi\|^2 > p + 2v\sqrt{x} + 2x, \|\xi\| \leq z_c) \leq 2e^{-x}.$$

The large deviation probability $\mathbb{P}(\|B^{1/2}\xi\| > y)$ for $y > z_c$ can be bounded as in the case $B = I_p$.

Lemma B.2.5. Define $g_c = \mu_c z_c$; see (B.26). It holds for $z \geq z_c$

$$\begin{aligned} \mathbb{P}(\|B^{1/2}\xi\| > z) &\leq 8.4(1 - g_c/z)^{-p/2} \exp(-g_c z/2) \\ &\leq 8.4 \exp\{-x_c - g_c(z - z_c)/2\}. \end{aligned}$$

Proof. The arguments from the case $B \equiv I_p$ apply without changes.

B.3 Deviation probability for a normalized martingale

Consider a random p -vector M and a random symmetric positive $p \times p$ -matrix S^2 s.t.

$$\mathbb{E} \exp\left(\gamma^\top M - \frac{1}{2}\gamma^\top S^2\gamma\right) \leq 1, \quad \|\gamma\|_o \leq g. \quad (\text{B.29})$$

The aim is to evaluate the deviation probability $\mathbb{P}(\|S^{-1}M\| \geq \delta)$. Given a positive symmetric matrix S , define $r_o(S)$ by

$$\mathbb{P}(\|S^{-1}\varepsilon\|_o \leq r_o(S)) \geq 1/2, \quad \varepsilon \sim \mathcal{N}(0, I_p). \quad (\text{B.30})$$

Here are two typical examples when $r_o(S)$ can be easily evaluated. If $\|\cdot\|_o$ is the usual Euclidean norm, one can take $r_o(S) = \text{tr}(S^{-2})$. If $\|\cdot\|_o$ is the sup-norm, and S is a diagonal matrix with $S = \text{diag}(s_1 \leq \dots \leq s_p)$, define $r_o(S) = s_p^{-1}\sqrt{2\log p}$.

Theorem B.3.1. Suppose (B.29) for some $g > 0$. For any $\mu < 1$ and any deterministic positive matrices $S_- \preceq S_+$, it holds

$$\begin{aligned} \mathbb{I}P\left(\|S^{-1}\mathbf{M}\| \geq \mathfrak{z}, S_- \preceq S \preceq S_+, \|S^{-2}\mathbf{M}\|_{\circ} \leq \mu^{-1}g - \mu^{-1/2}\mathbf{r}_{\circ}(S)\right) \\ \leq \frac{2\det(S_+)}{\det(S_-)}(1-\mu)^{-p/2} \exp(-\mu\mathfrak{z}/2). \end{aligned} \quad (\text{B.31})$$

Optimizing w.r.t. μ yields

$$\begin{aligned} \mathbb{I}P\left(\|S^{-1}\mathbf{M}\|^2 \geq p + 2\sqrt{p\mathfrak{x}} + 2\mathfrak{x}, S_- \preceq S \preceq S_+, \|S^{-2}\mathbf{M}\|_{\circ} \leq \mu^{-1}g - \mu^{-1/2}\mathbf{r}_{\circ}(S)\right) \\ \leq \frac{2\det(S_+)}{\det(S_-)} \exp(-\mathfrak{x}). \end{aligned}$$

Proof. Denote $\tau = \mu^{-1/2} > 1$. Introduce random sets \mathcal{A}_1 and \mathcal{A}_2

$$\begin{aligned} \mathcal{A}_1 &\stackrel{\text{def}}{=} \{S_- \preceq S \preceq S_+\}, \\ \mathcal{A}_2 &\stackrel{\text{def}}{=} \{\|S^{-2}\mathbf{M}\|_{\circ} \leq \tau^2g - \tau\mathbf{r}_{\circ}(S)\}. \end{aligned}$$

For $\gamma \in \mathbb{R}^p$, define $\mathbf{u} = \tau S\gamma - \tau^{-1}S^{-1}\mathbf{M}$, so that $\gamma = \tau^{-1}S^{-1}\mathbf{u} + \tau^{-2}S^{-2}\mathbf{M}$. It holds on the set \mathcal{A}_2 by (B.30) with $\xi = S^{-1}\mathbf{M}$

$$\begin{aligned} c_p \det(\tau S) \int \exp\left(\gamma^{\top} \mathbf{M} - \frac{\tau^2}{2}\gamma^{\top} S^2 \gamma\right) \mathbb{I}(\|\gamma\|_{\circ} \leq g) d\gamma \\ = \exp\left(\frac{\|\xi\|^2}{2\tau^2}\right) c_p \det(\tau S) \int \exp\left(-\frac{1}{2}\|\tau^{-1}S^{-1}\mathbf{M} - \tau S\gamma\|^2\right) \mathbb{I}(\|\gamma\|_{\circ} \leq g) d\gamma \\ \geq \exp\left(\frac{\|\xi\|^2}{2\tau^2}\right) c_p \int \exp(-\|\mathbf{u}\|^2/2) \mathbb{I}\{\|S^{-1}\mathbf{u}\|_{\circ} \leq \mathbf{r}_{\circ}(S)\} d\mathbf{u} \\ \geq 0.5 \exp\left(\frac{\|\xi\|^2}{2\tau^2}\right). \end{aligned} \quad (\text{B.32})$$

Further, it holds on \mathcal{A}_1 with $\tau_0^2 \stackrel{\text{def}}{=} \tau^2 - 1$

$$\exp\left(\gamma^{\top} \mathbf{M} - \frac{\tau^2}{2}\gamma^{\top} S^2 \gamma\right) \leq \exp\left(\gamma^{\top} \mathbf{M} - \frac{1}{2}\gamma^{\top} S^2 \gamma\right) \exp\left(\frac{\tau_0^2}{2}\gamma^{\top} S_-^2 \gamma\right).$$

This implies by (B.32) and (B.29)

$$\mathbb{E}\left[\frac{\det(\tau_0 S_-)}{\det(\tau S)} \exp\left(\frac{\|\xi\|^2}{2\tau^2}\right) \mathbb{I}(\mathcal{A}_1)\right] \leq c_p \det(\tau_0 S_-) \int \exp\left(-\frac{\tau_0^2}{2}\gamma^{\top} S_-^2 \gamma\right) d\gamma = 1$$

and it follows by $\tau^2/\tau_0^2 = (1-\mu)^{-1}$ and by the Markov inequality that

$$\begin{aligned}
& \mathbb{P}\left(\|\xi\|^2 > \mathfrak{z}, \mathcal{A}_1 \cap \mathcal{A}_2\right) \\
& \leq \mathbb{P}\left(\frac{\det(\tau_0 S_-)}{\det(\tau S_+)} \exp\left(-\frac{\|\xi\|^2}{2\tau^2}\right) \mathbb{I}(\mathcal{A}_1) > (\tau/\tau_0)^p \frac{\det(S_-)}{\det(S_+)} \exp\left(-\frac{\mathfrak{z}}{2\tau^2}\right)\right) \\
& \leq 2(1-\mu)^{-p/2} \exp\left(-\frac{\mu \mathfrak{z}}{2}\right) \frac{\det(S_+)}{\det(S_-)}.
\end{aligned}$$

as required.

Note that the value $\mathbf{r}_o(S)$ from (B.30) is monotonously decreasing in S . Therefore, the inequality $\|S^{-2}M\|_o \leq \mu^{-1}g - \tau \mathbf{r}_o(S)$ follows from the stricter inequality

$$\|S_-^{-2}M\|_o \leq \mu^{-1}g - \tau \mathbf{r}_o(S_-).$$

If $\lambda_{\max}(S_-^{-1}S_+) \leq 1 + \delta$ for a fixed small constant, then on the considered random set $S_- \preceq S \preceq S_+$ one can everywhere replace S by S_- or S_+ without any significant loss of accuracy. Moreover, if $p\delta$ is small, then $\det(S_+)/\det(S_-) \approx 1$ and the bound (B.31) is nearly sharp; cf. Spokoiny and Zhilova (2013). In general, the value $\det(S_+)/\det(S_-)$ is the price for variability of the quadratic characteristic S^2 . A similar bound of Liptser and Spokoiny (2000) applies only for $g = \infty$, is not sharp and requires much more involved discretization arguments.

A typical application is given by the case when M is a martingale stopped at a time instant T and $S^2 = \langle M \rangle_T$ is its predictable quadratic characteristic at T .

Optimization of the right hand-side of (B.31) w.r.t. μ yields

$$\mathbb{P}\left(\|S^{-1}M\|^2 > p + 2\sqrt{p\mathbf{x}} + 2\mathbf{x}\right) \leq \frac{2\det(S_+)}{\det(S_-)} e^{-\mathbf{x}} + \mathbb{P}(\mathcal{A}_1) + \mathbb{P}(\mathcal{A}_2);$$

cf. Spokoiny and Zhilova (2013).

C

Sums of random matrices

Here we present a number of deviation bounds for a sum of random matrices.

C.1 Matrix Bernstein inequality

This section collects some useful facts about deviation of stochastic matrices from their mean. We mainly use the arguments from the book [Tropp \(2015\)](#). The main step of the proof is the following Master bound.

Theorem C.1.1 (Master bound). *Assume that $\mathbf{S}_1, \dots, \mathbf{S}_n$ are independent Hermitian matrices of the same size and $\mathbf{Z} = \sum_{i=1}^n \mathbf{S}_i$. Then*

$$\mathbb{E}\lambda_{\max}^+(\mathbf{Z}) \leq \inf_{\theta>0} \frac{1}{\theta} \log \text{tr} \exp \left(\sum_{i=1}^n \log \mathbb{E} e^{\theta \mathbf{S}_i} \right), \quad (\text{C.1})$$

$$\mathbb{P}\{\lambda_{\max}^+(\mathbf{Z}) \geq z\} \leq \inf_{\theta>0} e^{-\theta z} \text{tr} \exp \left(\sum_{i=1}^n \log \mathbb{E} e^{\theta \mathbf{S}_i} \right), \quad (\text{C.2})$$

where $\lambda_{\max}^+(\mathbf{Z})$ denotes the algebraically largest eigenvalue of \mathbf{Z} .

Proof. By the Markov inequality

$$\mathbb{P}\{\lambda_{\max}^+(\mathbf{Z}) \geq z\} \leq \inf_{\theta} e^{-\theta z} \mathbb{E} \exp(\theta \lambda_{\max}^+(\mathbf{Z})).$$

Recall the spectral mapping theorem: for any function $f: \mathbb{R} \rightarrow \mathbb{R}$ and Hermitian matrix A eigenvalues of $f(A)$ are equal to eigenvalues of A . Thus

$$\exp(\theta \lambda_{\max}(\mathbf{Z})) = \exp(\lambda_{\max}^+(\theta \mathbf{Z})) = \lambda_{\max}^+(\exp(\theta \mathbf{Z})) \leq \text{tr } e^{\theta \mathbf{Z}}.$$

Therefore,

$$\mathbb{P}\{\lambda_{\max}^+(\mathbf{Z}) \geq z\} \leq \inf_{\theta} e^{-\theta z} \mathbb{E} \text{tr} \exp(\theta \mathbf{Z}), \quad (\text{C.3})$$

and (C.2) follows.

To prove (C.1) fix θ . Using the spectral mapping theorem one can get that

$$\mathbb{E}\lambda_{\max}^+(\mathbf{Z}) = \frac{1}{\theta}\mathbb{E}\lambda_{\max}^+(\theta\mathbf{Z}) = \frac{1}{\theta}\log\mathbb{E}\exp(\lambda_{\max}^+(\theta\mathbf{Z})) = \frac{1}{\theta}\log\mathbb{E}\lambda_{\max}^+(\exp(\theta\mathbf{Z})).$$

Thus we get

$$\mathbb{E}\lambda_{\max}^+(\mathbf{Z}) \leq \frac{1}{\theta}\log\text{tr}\mathbb{E}\exp(\theta\mathbf{Z}). \quad (\text{C.4})$$

The final step in proving the master inequalities is to bound from above $\mathbb{E}\text{tr}\exp(\sum_{i=1}^n \mathbf{S}_i)$. To do this we use Jensen's inequality for the convex function $\text{tr}\exp(H + \log(X))$ (in matrix X), where H is deterministic Hermitian matrix. For a random Hermitian matrix X one can write

$$\mathbb{E}\text{tr}\exp(H + X) = \mathbb{E}\text{tr}\exp(H + \log e^X) \leq \text{tr}\exp(H + \log\mathbb{E}e^X). \quad (\text{C.5})$$

Denote by \mathbb{E}_i the conditional expectation with respect to random matrix X_i . To bound $\mathbb{E}\text{tr}\exp(\sum_{i=1}^n \mathbf{S}_i)$ we use (C.5) for the sum of independent Hermitian matrices by taking the conditional expectations with respect to i -th matrix:

$$\begin{aligned} \mathbb{E}\text{tr}\exp\left(\sum_{i=1}^n \mathbf{S}_i\right) &= \mathbb{E}\mathbb{E}_n\text{tr}\exp\left(\sum_{i=1}^{n-1} \mathbf{S}_i + \mathbf{S}_n\right) \\ &\leq \mathbb{E}\text{tr}\exp\left(\sum_{i=1}^{n-1} \mathbf{S}_i + \log(\mathbb{E}_n\exp(\mathbf{S}_n))\right) \\ &\leq \text{tr}\exp\left(\sum_{i=1}^n \log\mathbb{E}e^{\theta\mathbf{S}_i}\right). \end{aligned} \quad (\text{C.6})$$

To complete the prove of the Master's theorem combine (C.3) and (C.4) with (C.6).

The same result applied to $-\mathbf{Z}$ yields the bound for the norm $\|\mathbf{Z}\|_{\text{op}}$:

$$\begin{aligned} \mathbb{P}\{\|\mathbf{Z}\|_{\text{op}} \geq z\} &\leq \inf_{\theta>0} e^{-\theta z} \text{tr}\exp\left(\sum_{i=1}^n \log\mathbb{E}e^{\theta\mathbf{S}_i}\right) \\ &\quad + \inf_{\theta>0} e^{-\theta z} \text{tr}\exp\left(\sum_{i=1}^n \log\mathbb{E}e^{-\theta\mathbf{S}_i}\right). \end{aligned} \quad (\text{C.7})$$

Theorem C.1.2 (Bernstein inequality for a sum of random Hermitian matrices). Let $\mathbf{Z} = \sum_{i=1}^n \mathbf{S}_i$, where \mathbf{S}_i , $i = 1, \dots, n$ are independent, random, Hermitian matrices of the dimension $d \times d$ and

$$\lambda_{\max}^+(\mathbf{S}_i) \leq R.$$

Denote $\nu^2 = \nu^2(\mathbf{Z}) = \|\mathbb{E}(\mathbf{Z}^2)\|_{\text{op}}$. Then

$$\mathbb{E}\lambda_{\max}^+(\mathbf{Z}) \leq \sqrt{2\nu^2 \log(d)} + \frac{1}{3}R \log(d), \quad (\text{C.8})$$

$$\mathbb{P}\{\lambda_{\max}^+(\mathbf{Z}) \geq z\} \leq d \exp\left(\frac{-z^2/2}{\nu^2 + Rz/3}\right). \quad (\text{C.9})$$

Proof. Note that

$$\nu^2 = \left\| \sum_{i=1}^n \mathbb{E} \mathbf{S}_i^2 \right\|_{\text{op}}.$$

For the sake of simplicity let $\nu^2 = 1$. Denote

$$g(\theta) = \frac{\theta^2/2}{1 - R\theta/3}.$$

Apart the Master inequalities, we use the following lemma:

Lemma C.1.1. *Let \mathbf{Z} be a random Hermitian matrix $\mathbb{E}\mathbf{Z} = 0$, $\lambda_{\max}^+(\mathbf{Z}) \leq R$, then for $0 < \theta < 3/R$ the following inequalities hold*

$$\begin{aligned} \mathbb{E} e^{\theta \mathbf{Z}} &\leq \exp\left(\frac{\theta^2/2}{1 - R\theta/3} \mathbb{E}(\mathbf{Z}^2)\right), \\ \log \mathbb{E} e^{\theta \mathbf{Z}} &\leq \frac{\theta^2/2}{1 - R\theta/3} \mathbb{E}(\mathbf{Z}^2). \end{aligned}$$

Proof. Decompose the exponent in the following way

$$e^{\theta \mathbf{Z}} = I + \theta \mathbf{Z} + (e^{\theta \mathbf{Z}} - \theta \mathbf{Z} - I) = I + \theta \mathbf{Z} + \mathbf{Z} \cdot f(\mathbf{Z}) \cdot \mathbf{Z},$$

where

$$f(x) = \frac{e^{\theta x} - \theta x - 1}{x^2}, \quad \text{for } x \neq 0, \quad f(0) = \frac{\theta^2}{2}.$$

One can check that the function $f(x)$ is non-decreasing, thus for $x \leq R$, one has $f(x) \leq f(R)$. By the matrix transfer rule $f(\mathbf{Z}) \leq f(R)I$ and

$$\mathbb{E} e^{\theta \mathbf{Z}} \leq I + f(R) \mathbb{E} \mathbf{Z}^2.$$

In order to estimate $f(R)$ use $q! \geq 2 \cdot 3^{q-2}$ to get

$$f(R) = \frac{e^{\theta R} - \theta R - 1}{R^2} = \frac{1}{R^2} \sum_{q=2}^{\infty} \frac{(\theta R)^q}{q!} \leq \theta^2 \sum_{q=2}^{\infty} \frac{(R\theta)^{q-2}}{3^{q-2}} = \frac{\theta^2/2}{1 - R\theta/3}.$$

To get the result of the Lemma note that $1 + a \leq e^a$.

To prove (C.8) and (C.9) we apply the Master inequalities and Lemma C.1.1:

$$\begin{aligned}
I\!\!E \lambda_{\max}^+(\mathbf{Z}) &\leq \inf_{\theta>0} \frac{1}{\theta} \log \operatorname{tr} \exp \left(\sum_{i=1}^n \log I\!\!E \exp(\theta \mathbf{S}_i) \right) \\
&\leq \inf_{0<\theta<3/R} \frac{1}{\theta} \log \operatorname{tr} \exp \left(g(\theta) \sum_{i=1}^n I\!\!E \mathbf{S}_i^2 \right) \\
&\leq \inf_{0<\theta<3/R} \frac{1}{\theta} \log \operatorname{tr} \exp (g(\theta) I\!\!E \mathbf{Z}^2) \\
&\leq \inf_{0<\theta<3/R} \frac{1}{\theta} \log d \exp (g(\theta) \|I\!\!E \mathbf{Z}^2\|_{\text{op}}) \\
&\leq \inf_{0<\theta<3/R} \left\{ \frac{\log(d)}{\theta} + \frac{\theta/2}{1 - R\theta/3} \right\}.
\end{aligned}$$

Minimizing the right hand side in θ one can get (C.8).

The second inequality can be obtained in the same manner:

$$\begin{aligned}
I\!\!P\{\lambda_{\max}^+(\mathbf{Z}) \geq z\} &\leq \inf_{\theta>0} e^{-\theta z} \operatorname{tr} \exp \left(\sum_{i=1}^n \log I\!\!E \exp(\theta \mathbf{S}_i) \right) \\
&\leq \inf_{0<\theta<3/R} e^{-\theta z} \operatorname{tr} \exp (g(\theta) I\!\!E \mathbf{Z}^2) \\
&\leq \inf_{0<\theta<3/R} e^{-\theta z} d \exp (g(\theta) \|I\!\!E \mathbf{Z}^2\|_{\text{op}}) \\
&\leq \inf_{0<\theta<3/R} e^{-\theta z} d \exp (g(\theta)).
\end{aligned}$$

Here instead of minimizing the right hand side in θ we have used $\theta = z/(1 + Rz/3)$.

Theorem C.1.3 (Bernstein inequality for a sum of random Hermitian matrices). Let $\mathbf{Z} = \sum_{i=1}^n \mathbf{S}_i$, where \mathbf{S}_i , $i = 1, \dots, n$ are independently distributed random matrices of the size $d_1 \times d_2$ and

$$\|\mathbf{S}_i\|_{\text{op}} \leq R.$$

Denote $v^2 = v^2(\mathbf{Z}) = \max \{\|I\!\!E(\mathbf{Z}^* \mathbf{Z})\|_{\text{op}}, \|I\!\!E(\mathbf{Z} \mathbf{Z}^*)\|_{\text{op}}\}$. Then

$$I\!\!E \|\mathbf{Z}\|_{\text{op}} \leq \sqrt{2v^2 \log(d_1 + d_2)} + \frac{1}{3} R \log(d),$$

$$I\!\!P\{\|\mathbf{Z}\|_{\text{op}} \geq z\} \leq (d_1 + d_2) \exp \left(\frac{-z^2/2}{v^2 + Rz/3} \right).$$

Proof. Use the following hint: define the matrix

$$H(\mathbf{Z}) = \begin{pmatrix} 0 & \mathbf{Z} \\ \mathbf{Z}^* & 0 \end{pmatrix}.$$

It can be easily seen that $v^2 = \|H(\mathbf{Z})^2\|_{\text{op}}$, and $\|\mathbf{Z}\|_{\text{op}} = \lambda_{\max}^+(H(\mathbf{Z}))$, thus applying Proposition C.1.2 to $H(\mathbf{Z})$ the statements (C.8) and (C.9) are straightforward.

The next result provides a deviation bound for a matrix-valued quadratic forms.

Proposition C.1.1 (Deviation bound for matrix quadratic forms). *Let a $p \times n$ matrix \mathcal{U} with columns $\omega_1, \dots, \omega_n$ be such that*

$$\mathcal{U}\mathcal{U}^\top \leq I_p, \quad \|\omega_i\| \leq \delta_n \quad (\text{C.10})$$

for a fixed constant δ_n . For a random vector $\gamma = (\gamma_1, \dots, \gamma_n)^\top$ with independent standard Gaussian components, define

$$\mathbf{Z} \stackrel{\text{def}}{=} \mathcal{U} \text{diag}\{\gamma \cdot \gamma - 1\} \mathcal{U}^\top = \sum_{i=1}^n (\gamma_i^2 - 1) \omega_i \omega_i^\top.$$

Then

$$\mathbb{P}\left(\|\mathbf{Z}\|_{\text{op}} \geq 2\delta_n \sqrt{y + \log(p)} + 2\delta_n^2(y + \log p)\right) \leq 2e^{-y}. \quad (\text{C.11})$$

Proof. From the Master bound (C.7)

$$\begin{aligned} \mathbb{P}(\|\mathbf{Z}\|_{\text{op}} \geq z) &\leq \inf_{\theta > 0} e^{-\theta z} \text{tr} \exp\left(\sum_{i=1}^n \log \mathbb{E} \exp\{\theta(\gamma_i^2 - 1) \omega_i \omega_i^\top\}\right) \\ &\quad + \inf_{\theta > 0} e^{-\theta z} \text{tr} \exp\left(\sum_{i=1}^n \log \mathbb{E} \exp\{\theta(-\gamma_i^2 + 1) \omega_i \omega_i^\top\}\right). \end{aligned} \quad (\text{C.12})$$

Now we use the following general fact:

Lemma C.1.2. *If χ is a random variable and Π is a projector in \mathbb{R}^p , then*

$$\log \mathbb{E} \exp(\chi \Pi) = \log(\mathbb{E} e^\chi) \Pi. \quad (\text{C.13})$$

Proof. The result (C.13) can be easily obtained by applying twice the spectral mapping theorem.

This result yields, in particular, for any unit vector $\omega \in \mathbb{R}^p$

$$\log \mathbb{E} \exp(\chi \omega \omega^\top) = \log(\mathbb{E} e^\chi) \omega \omega^\top.$$

Moreover, for any vector $\omega \in \mathbb{R}^p$, the normalized product $\omega \omega^\top / \|\omega\|^2$ is a rank-one projector, and hence,

$$\log \mathbb{E} \exp(\chi \omega \omega^\top) = \log(\mathbb{E} e^{\chi \|\omega\|^2}) \frac{\omega \omega^\top}{\|\omega\|^2}.$$

With $\mathbf{U}_i \stackrel{\text{def}}{=} \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top / \|\boldsymbol{\omega}_i\|^2$ and $\chi_i = \theta(\gamma_i^2 - 1)$, we derive

$$\begin{aligned} \log \mathbb{E} \exp \{ \theta(\gamma_i^2 - 1) \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top \} &= \log \mathbb{E} \exp \{ \theta(\gamma_i^2 - 1) \|\boldsymbol{\omega}_i\|^2 \} \mathbf{U}_i \\ &= \log \left(\frac{\exp(-\|\boldsymbol{\omega}_i\|^2 \theta)}{\sqrt{1 - 2\|\boldsymbol{\omega}_i\|^2 \theta}} \right) \mathbf{U}_i \\ &= \left\{ -\|\boldsymbol{\omega}_i\|^2 \theta - \frac{1}{2} \log(1 - 2\theta\|\boldsymbol{\omega}_i\|^2) \right\} \mathbf{U}_i \end{aligned}$$

and

$$\begin{aligned} \log \mathbb{E} \exp \{ \theta(-\gamma_i^2 + 1) \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top \} &= \log \mathbb{E} \exp (\theta(-\gamma_i^2 + 1) \|\boldsymbol{\omega}_i\|^2) \mathbf{U}_i \\ &\leq -\|\boldsymbol{\omega}_i\|^2 \theta \mathbf{U}_i \\ &\leq \left\{ -\|\boldsymbol{\omega}_i\|^2 \theta - \frac{1}{2} \log(1 - 2\theta\|\boldsymbol{\omega}_i\|^2) \right\} \mathbf{U}_i. \end{aligned}$$

Then it follows by (C.12)

$$\begin{aligned} \mathbb{P}(\|\mathbf{Z}\|_{\text{op}} \geq z) &\leq 2 \inf_{\theta > 0} e^{-\theta z} \text{tr} \exp \left\{ \sum_{i=1}^n \frac{\boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top}{\|\boldsymbol{\omega}_i\|^2} \left\{ -\|\boldsymbol{\omega}_i\|^2 \theta - \frac{1}{2} \log(1 - 2\theta\|\boldsymbol{\omega}_i\|^2) \right\} \right\}. \end{aligned} \quad (\text{C.14})$$

Denote $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^\top$, where

$$\eta_i = -\theta - \frac{\log(1 - 2\|\boldsymbol{\omega}_i\|^2 \theta)}{2\|\boldsymbol{\omega}_i\|^2}.$$

The use of (B.6) and (C.10) yields for $\theta < (2\delta_n^2)^{-1}$

$$\begin{aligned} \eta_i &= \frac{1}{2\|\boldsymbol{\omega}_i\|^2} \left\{ 2\theta\|\boldsymbol{\omega}_i\|^2 - \log(1 - 2\theta\|\boldsymbol{\omega}_i\|^2) \right\} \\ &\leq \frac{(2\theta\|\boldsymbol{\omega}_i\|^2)^2}{4\|\boldsymbol{\omega}_i\|^2(1 - 2\theta\delta_n^2)} \leq \frac{\theta^2\delta_n^2}{(1 - 2\theta\delta_n^2)}. \end{aligned}$$

Then by (C.14) and $\mathcal{U}\mathcal{U}^\top = I_p$ using $\mu = 2\theta\delta_n^2$

$$\begin{aligned} \mathbb{P}(\|\mathbf{Z}\|_{\text{op}} \geq z) &\leq 2 \inf_{\theta > 0} e^{-\theta z} \text{tr} \exp \{ \mathcal{U} \text{diag}(\boldsymbol{\eta}) \mathcal{U}^\top \} \leq 2 \inf_{\theta > 0} e^{-\theta z} \text{tr} \exp \{ \|\boldsymbol{\eta}\|_\infty I_p \} \\ &\leq 2p \inf_{\theta > 0} \exp \left\{ -\theta z + \frac{\theta^2\delta_n^2}{1 - 2\theta\delta_n^2} \right\} = 2p \inf_{\mu > 0} \exp \left\{ -\frac{\mu z}{2\delta_n^2} + \frac{\mu^2\delta_n^{-2}}{1 - \mu} \right\}. \end{aligned}$$

Lemma B.1.2 helps to bound for $y_p = y + \log(p)$ and $z = 2\delta_n \sqrt{y_p} + 2\delta_n^2 y_p$ that

$$\inf_{\mu > 0} \exp \left\{ -\frac{\mu z}{2\delta_n^2} + \frac{\mu^2\delta_n^{-2}}{1 - \mu} \right\} = \inf_{\mu > 0} \left\{ -\mu(\delta_n^{-1} \sqrt{y_p} + y_p) + \frac{\mu^2\delta_n^{-2}}{4(1 - \mu)} \right\} \leq -y_p.$$

Therefore,

$$\mathbb{P}\left(\|\mathbf{Z}\|_{\text{op}} \geq 2\delta_n \sqrt{y + \log p} + 2\delta_n^2(y + \log p)\right) \leq 2p e^{-y - \log p} = 2e^{-y}$$

as required.

Proposition C.1.2 (Deviation bound for matrix Gaussian sums). *Let a $p \times n$ matrix \mathcal{U} with columns $\omega_1, \dots, \omega_n$ satisfy (C.10). Let γ_i be independent standard Gaussian, $i = 1, \dots, n$. For any deterministic vector $\mathbf{B} = (b_1, \dots, b_n)^\top \in \mathbb{R}^n$, consider the matrix \mathbf{Z}_1 with*

$$\mathbf{Z}_1 \stackrel{\text{def}}{=} \sum_{i=1}^n \gamma_i b_i \omega_i \omega_i^\top.$$

It holds

$$\begin{aligned} \mathbb{P}\left(\|\mathbf{Z}_1\|_{\text{op}} \geq \delta_n^2 \|\mathbf{B}\| \sqrt{2y}\right) &\leq 2e^{-y} \\ \mathbb{P}\left(\|\mathbf{Z}_1\|_{\text{op}} \geq \delta_n \|\mathbf{B}\|_\infty \sqrt{2(y + \log p)}\right) &\leq 2e^{-y}. \end{aligned}$$

Proof. As γ_i are i.i.d. standard normal and $\mathbb{E} e^{a\gamma_i} = e^{a^2/2}$ for $|a| < 1/2$, it follows from the Master inequality and Lemma C.1.2

$$\begin{aligned} \mathbb{P}(\|\mathbf{Z}_1\|_{\text{op}} \geq z) &\leq 2 \inf_{\theta > 0} e^{-\theta z} \operatorname{tr} \exp \left\{ \sum_{i=1}^n \log \mathbb{E} \exp(\theta \gamma_i b_i \omega_i \omega_i^\top) \right\} \\ &\leq 2 \inf_{\theta > 0} e^{-\theta z} \operatorname{tr} \exp \left\{ \sum_{i=1}^n \frac{\theta^2 b_i^2 \|\omega_i\|^4}{2} \frac{\omega_i \omega_i^\top}{\|\omega_i\|^2} \right\}. \end{aligned}$$

Moreover, as $\|\omega_i\| \leq \delta_n$ and $\mathbf{U}_i = \omega_i \omega_i^\top / \|\omega_i\|^2$ is a rank-one projector with $\operatorname{tr} \mathbf{U}_i = 1$, it holds

$$\operatorname{tr} \exp \left\{ \frac{\theta^2}{2} \sum_{i=1}^n b_i^2 \|\omega_i\|^4 \mathbf{U}_i \right\} \leq \exp \operatorname{tr} \left(\frac{\theta^2 \delta_n^4}{2} \sum_{i=1}^n b_i^2 \mathbf{U}_i \right) = \exp \frac{\theta^2 \delta_n^4 \|\mathbf{B}\|^2}{2}.$$

This implies for $z = \delta_n^2 \|\mathbf{B}\| \sqrt{2y}$

$$\mathbb{P}(\|\mathbf{Z}_1\|_{\text{op}} \geq z) \leq 2 \inf_{\theta > 0} \exp \left(-\theta z + \frac{1}{2} \theta^2 \delta_n^4 \|\mathbf{B}\|^2 \right) = 2e^{-y}$$

and the assertion follows. Alternatively, the definition of \mathbf{U}_i and (C.10) imply

$$\frac{\theta^2}{2} \sum_{i=1}^n b_i^2 \|\omega_i\|^4 \mathbf{U}_i \leq \frac{\theta^2 \|\mathbf{B}\|_\infty^2 \delta_n^2}{2} \sum_{i=1}^n \omega_i \omega_i^\top \leq \frac{\theta^2 \|\mathbf{B}\|_\infty^2 \delta_n^2}{2} I_p,$$

so that

$$\mathrm{tr} \exp \left(\frac{\theta^2}{2} \sum_{i=1}^n b_i^2 \|\boldsymbol{\omega}_i\|^4 \mathbf{U}_i \right) \leq p \exp \left(\frac{\theta^2 \|\mathbf{B}\|_\infty^2 \delta_n^2}{2} \right)$$

This implies for $z = \delta_n \|\mathbf{B}\|_\infty \sqrt{2(y + \log p)}$ and $\theta(z) = (\delta_n \|\mathbf{B}\|_\infty)^{-1} \sqrt{2(y + \log p)}$

$$\begin{aligned} \mathbb{P}(\|\mathbf{Z}_1\|_{\mathrm{op}} \geq z) &\leq 2 \inf_{\theta>0} \exp \left(-\theta z + \frac{\theta^2 \|\mathbf{B}\|_\infty^2 \delta_n^2}{2} + \log p \right) \\ &= 2 \exp \left(-\theta(z) z + \frac{\theta^2(z) \|\mathbf{B}\|_\infty^2 \delta_n^2}{2} + \log p \right) = 2e^{-y} \end{aligned}$$

C.2 Presmoothing and bias effects

Let $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$ for a positive symmetric matrix Σ , and let Π be a linear operator in \mathbb{R}^n . Define

$$\boldsymbol{\xi} = \Sigma^{-1/2}(\boldsymbol{\varepsilon} - \Pi \boldsymbol{\varepsilon}) = \boldsymbol{\gamma} - \Upsilon \boldsymbol{\gamma} \quad (\text{C.15})$$

where $\boldsymbol{\gamma} = \Sigma^{-1/2} \boldsymbol{\varepsilon}$ is a standard Gaussian vector, and

$$\Upsilon \stackrel{\mathrm{def}}{=} \Sigma^{-1/2} \Pi \Sigma^{1/2}.$$

For a deterministic bias vector $\mathbf{B} = (b_1, \dots, b_n)^\top$ and for a $p \times n$ matrix \mathcal{U} satisfying (C.10), we aim at bounding the Frobenius and operator norms of the matrix \mathcal{B} with

$$\begin{aligned} \mathcal{B} &\stackrel{\mathrm{def}}{=} \mathcal{U} \left[\mathrm{diag} \{ (\boldsymbol{\xi} + \mathbf{B}) \cdot (\boldsymbol{\xi} + \mathbf{B}) \} - I_n \right] \mathcal{U}^\top \\ &= \sum_{i=1}^n \{ (\xi_i + b_i)^2 - 1 \} \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top. \end{aligned} \quad (\text{C.16})$$

Proposition C.2.1. Suppose that a vector $\boldsymbol{\xi}$ can be written in the form (C.15) for a standard Gaussian vector $\boldsymbol{\gamma}$, and let the rows Υ_i^\top of $\Upsilon \stackrel{\mathrm{def}}{=} \Sigma^{-1/2} \Pi \Sigma^{1/2}$ satisfy $\|\Upsilon_i\| \leq \delta$. Further, let the matrix \mathcal{U} fulfill (C.10). Then on a random set $\Omega(y)$ with $\mathbb{P}(\Omega(y)) \geq 1 - 6e^{-y}$, it holds for \mathcal{B} from (C.16)

$$\|\mathcal{B}\|_{\mathrm{op}} \leq \Delta(y),$$

$$\|\mathcal{B}\|_{\mathrm{Fr}} \leq \Delta_{\mathrm{Fr}}(y) = \sqrt{p} \Delta(y),$$

where

$$\begin{aligned} \Delta(y) &\stackrel{\mathrm{def}}{=} \|\mathbf{B}\|_\infty^2 + \delta \|\mathbf{B}\|_\infty (\sqrt{2y + 2 \log n} + \sqrt{2y + 2 \log p}) \\ &\quad + 2\delta_n \sqrt{y + \log p} + 2\delta_n^2 (y + \log p) + (2\delta + \delta^2)(y + \log n). \end{aligned}$$

Proof. The use of $\boldsymbol{\xi} = \boldsymbol{\gamma} - \Upsilon\boldsymbol{\gamma}$ allows to decompose

$$\begin{aligned}\mathcal{B} &= \mathcal{U} \operatorname{diag}\{\mathbf{B} \cdot \mathbf{B}\} \mathcal{U}^\top &&\stackrel{\text{def}}{=} \mathcal{B}_1 \\ &+ 2\mathcal{U} \operatorname{diag}\{\boldsymbol{\gamma} \cdot \mathbf{B}\} \mathcal{U}^\top &&\stackrel{\text{def}}{=} \mathcal{B}_2 \\ &+ 2\mathcal{U} \operatorname{diag}\{\Upsilon\boldsymbol{\gamma} \cdot \mathbf{B}\} \mathcal{U}^\top &&\stackrel{\text{def}}{=} \mathcal{B}_3 \\ &+ \mathcal{U} \operatorname{diag}\{\boldsymbol{\xi} \cdot \boldsymbol{\xi} - \boldsymbol{\gamma} \cdot \boldsymbol{\gamma}\} \mathcal{U}^\top &&\stackrel{\text{def}}{=} \mathcal{B}_4 \\ &+ \mathcal{U} \{\operatorname{diag}(\boldsymbol{\gamma} \cdot \boldsymbol{\gamma}) - \mathbf{I}_n\} \mathcal{U}^\top &&\stackrel{\text{def}}{=} \mathcal{B}_5\end{aligned}$$

Obviously

$$\|\mathcal{B}\|_{\text{op}} \leq \|\mathcal{B}_1\|_{\text{op}} + \|\mathcal{B}_2\|_{\text{op}} + \|\mathcal{B}_3\|_{\text{op}} + \|\mathcal{B}_4\|_{\text{op}} + \|\mathcal{B}_5\|_{\text{op}}.$$

It is well known that the sup-norm of a standard Gaussian vector $\boldsymbol{\gamma}$ in \mathbb{R}^n can be bounded in probability

$$\mathbb{P}\left(\|\boldsymbol{\gamma}\|_\infty \geq \sqrt{2y + 2 \log n}\right) \leq e^{-y}$$

with $y_n = y + \log(n)$. Further, if each row Υ_i^\top of Υ satisfies $\|\Upsilon_i\| \leq \delta$, then the scalar product $\Upsilon_i^\top \boldsymbol{\gamma}$ is a normal zero mean r.v. with the variance

$$\operatorname{Var}(\Upsilon_i^\top \boldsymbol{\gamma}) = \|\Upsilon_i\|^2 \leq \delta^2$$

and

$$\mathbb{P}(|\Upsilon_i^\top \boldsymbol{\gamma}| > \delta z_1(y)) \leq e^{-y}$$

with $z_1(y) \leq \sqrt{2y}$ yielding

$$\mathbb{P}\left(\|\Upsilon\boldsymbol{\gamma}\|_\infty > \delta \sqrt{2y + 2 \log n}\right) \leq e^{-y}.$$

Due to this bounds, there is a random set $\Omega_\infty(y)$ with $\mathbb{P}(\Omega_\infty(y)) \geq 1 - 2e^{-y}$ such that it holds on $\Omega_\infty(y)$

$$\|\Upsilon\boldsymbol{\gamma}\|_\infty \leq \delta \sqrt{2y + 2 \log n}, \quad \|\boldsymbol{\gamma}\|_\infty \leq \sqrt{2y + 2 \log n}. \quad (\text{C.17})$$

Bounds for \mathcal{B}_1

In view of $\mathcal{U}\mathcal{U}^\top \leq \mathbf{I}_p$, the bias term \mathcal{B}_1 can be estimated as follows:

$$\|\mathcal{B}_1\|_{\text{op}} = \left\| \sum_{i=1}^n \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top b_i^2 \right\|_{\text{op}} \leq \|\mathbf{B}\|_\infty^2 \|\mathcal{U}\mathcal{U}^\top\|_{\text{op}} \leq \|\mathbf{B}\|_\infty^2.$$

Bounds for \mathcal{B}_2

Proposition C.1.2 provides a bound for a random matrix

$$\mathcal{B}_2 = 2\mathcal{U} \operatorname{diag}\{\boldsymbol{\gamma} \cdot \mathbf{B}\} \mathcal{U}^\top = 2 \sum_{i=1}^n \gamma_i b_i \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top$$

in the operator norm: on a set $\Omega_2(y)$ with $\mathbb{P}(\Omega_2(y)) \geq 1 - 2e^{-y}$

$$\|\mathcal{B}_2\|_{\text{op}} \leq \delta_n \|\mathbf{B}\|_\infty \sqrt{2y + 2 \log p}.$$

Bounds for \mathcal{B}_3

We use that $\|\Upsilon\boldsymbol{\gamma}\|_\infty \leq \delta \sqrt{2y + 2 \log n}$ on a set $\Omega_\infty(\mathbf{x})$. Then similarly to \mathcal{B}_1

$$\|\mathcal{B}_3\|_{\text{op}} = \left\| \sum_{i=1}^n \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top b_i (\Upsilon\boldsymbol{\gamma})_i \right\|_{\text{op}} \leq \|\mathbf{B}\|_\infty \|\Upsilon\boldsymbol{\gamma}\|_\infty \leq \delta \|\mathbf{B}\|_\infty \sqrt{2y + 2 \log n}.$$

Bounds for \mathcal{B}_4

The identity $\boldsymbol{\xi} = \boldsymbol{\gamma} - \Upsilon\boldsymbol{\gamma}$ implies

$$\boldsymbol{\xi} \cdot \boldsymbol{\xi} - \boldsymbol{\gamma} \cdot \boldsymbol{\gamma} = (\Upsilon\boldsymbol{\gamma}) \cdot (\Upsilon\boldsymbol{\gamma}) - 2(\Upsilon\boldsymbol{\gamma}) \cdot \boldsymbol{\gamma}$$

and

$$\|\mathcal{B}_4\|_{\text{op}} \leq \|\mathcal{U} \operatorname{diag}\{(\Upsilon\boldsymbol{\gamma}) \cdot (\Upsilon\boldsymbol{\gamma})\} \mathcal{U}^\top\|_{\text{op}} + 2\|\mathcal{U} \operatorname{diag}\{(\Upsilon\boldsymbol{\gamma}) \cdot \boldsymbol{\gamma}\} \mathcal{U}^\top\|_{\text{op}}.$$

The condition $\mathcal{U}\mathcal{U}^\top \leq I_p$ helps to bound

$$\|\mathcal{U} \operatorname{diag}\{(\Upsilon\boldsymbol{\gamma}) \cdot (\Upsilon\boldsymbol{\gamma})\} \mathcal{U}^\top\|_{\text{op}} \leq \|\Upsilon\boldsymbol{\gamma}\|_\infty^2 \|\mathcal{U}\mathcal{U}^\top\|_{\text{op}} \leq \|\Upsilon\boldsymbol{\gamma}\|_\infty^2.$$

Similarly

$$\|\mathcal{U} \operatorname{diag}\{(\Upsilon\boldsymbol{\gamma}) \cdot \boldsymbol{\gamma}\} \mathcal{U}^\top\|_{\text{op}} \leq \|\Upsilon\boldsymbol{\gamma}\|_\infty \|\boldsymbol{\gamma}\|_\infty \|\mathcal{U}\mathcal{U}^\top\|_{\text{op}} \leq \|\Upsilon\boldsymbol{\gamma}\|_\infty \|\boldsymbol{\gamma}\|_\infty.$$

By (C.17), after restricting to the set $\Omega_\infty(\mathbf{x})$, this yields the bound

$$\|\mathcal{B}_4\|_{\text{op}} \leq (2\delta + \delta^2)(y + \log n).$$

Bounds for \mathcal{B}_5

The matrix \mathcal{B}_5 can be bounded by a version of matrix Bernstein inequality (C.11) in Proposition C.1.1: on a set $\Omega_5(y)$ with $\mathbb{P}(\Omega_5(y)) \geq 1 - 2e^{-y}$

$$\|\mathcal{B}_5\|_{\text{op}} \leq 2\delta_n \sqrt{y + \log p} + 2\delta_n^2(y + \log p).$$

Gathering all the obtained bounds yields the result of the proposition about the operator norm of \mathcal{B} . The Frobenius norm is bounded by the elementary inequality $\|A\|_{\text{Fr}} \leq \sqrt{p}\|A\|_{\text{op}}$ for any $p \times p$ matrix A .

C.3 Empirical covariance matrix

Let ε_i be independent centered random vectors in \mathbb{R}^p . Consider the empirical covariance matrix

$$\widehat{S} \stackrel{\text{def}}{=} \sum_i \varepsilon_i \varepsilon_i^\top.$$

The interesting question is how well this matrix stabilizes around its expectation

$$S \stackrel{\text{def}}{=} \mathbb{E}\widehat{S} = \sum_{i=1}^n \text{Var}(\varepsilon_i) = \text{Var}\left(\sum_{i=1}^n \varepsilon_i\right). \quad (\text{C.18})$$

We implicitly assume that S is large and hence each $S^{-1/2}\varepsilon_i$ is small. Below we focus on the relative difference

$$\mathbf{Z} \stackrel{\text{def}}{=} S^{-1/2}(\widehat{S} - S)S^{-1/2} = S^{-1/2}\widehat{S}S^{-1/2} - I_p.$$

This allows to reduce the study to the case $S \equiv I_p$ considered below. Note that each vector ε_i is replaced by $S^{-1/2}\varepsilon_i$. For bounding $\|\mathbf{Z}\|_{\text{op}}$, one can apply the Bernstein inequality provided that each vector ε_i is bounded by a fixed small constant δ :

$$\|\varepsilon\|_\infty = \max_{i \leq n} \|\varepsilon_i\| \leq \delta.$$

Then $\mathbf{S}_i = \varepsilon_i \varepsilon_i^\top - \mathbb{E}\varepsilon_i \varepsilon_i^\top$ obviously fulfills

$$\|\mathbf{S}_i\|_{\text{op}} = \|\varepsilon_i \varepsilon_i^\top - \mathbb{E}\varepsilon_i \varepsilon_i^\top\|_{\text{op}} \leq \|\varepsilon_i\|^2 \leq \delta^2.$$

Define

$$\mathbf{b}^2 = \left\| \sum_{i=1}^n \mathbb{E}\mathbf{S}_i^2 \right\|_{\text{op}} = \left\| \sum_{i=1}^n \left\{ \mathbb{E}(\varepsilon_i \varepsilon_i^\top)^2 - (\mathbb{E}\varepsilon_i \varepsilon_i^\top)^2 \right\} \right\|_{\text{op}}. \quad (\text{C.19})$$

A rough bound on \mathbf{b} can be obtained by using again $\|\varepsilon_i\| \leq \delta$ and $S = I_p$:

$$\mathbf{b}^2 \leq \delta^2 \left\| \sum_{i=1}^n \mathbb{E}\varepsilon_i \varepsilon_i^\top \right\|_{\text{op}} = \delta^2 \quad (\text{C.20})$$

Now the result of Proposition C.1.2 implies for $\mathbf{Z} = \widehat{S} - I_p$ with $\mathbf{S}_i = \varepsilon_i \varepsilon_i^\top - \mathbb{E}\varepsilon_i \varepsilon_i^\top$

$$\mathbb{P}\{\|\widehat{S} - I_p\|_{\text{op}} \geq \mathbf{b} z\} \leq 2p \exp\left(\frac{-z^2/2}{1 + \delta^2 \mathbf{b}^{-1} z / 3}\right).$$

The bound is particularly informative if the ratio δ^2/\mathbf{b} is small. The use of the rough upper bound (C.20) yields for each $z > 0$

$$\mathbb{P}\{\|\widehat{S} - I_p\|_{\text{op}} \geq \delta z\} \leq 2p \exp\left(\frac{-z^2/2}{1 + z\delta/3}\right).$$

One can pick up $z \approx \sqrt{2y \log(2p)}$ to ensure a sensible deviation bound about e^{-y} for moderate values of y .

Theorem C.3.1. *Let random vectors $\varepsilon_1, \dots, \varepsilon_n$ in \mathbb{R}^p be independent zero mean, S is given by (C.18), and $\tilde{\varepsilon}_i \stackrel{\text{def}}{=} S^{-1/2} \varepsilon_i$ fulfill*

$$\|\tilde{\varepsilon}_i\| = \|S^{-1/2} \varepsilon_i\| \leq \delta \text{ a.s.}, i = 1, \dots, n,$$

for a constant $\delta < \infty$. Then with \mathbf{b}^2 defined by

$$\mathbf{b}^2 \stackrel{\text{def}}{=} \|\mathbb{E}(S^{-1/2} \widehat{S} S^{-1/2})^2\|_{\text{op}} = \left\| \sum_{i=1}^n \left\{ \mathbb{E}(\tilde{\varepsilon}_i \tilde{\varepsilon}_i^\top)^2 - (\mathbb{E} \tilde{\varepsilon}_i \tilde{\varepsilon}_i^\top)^2 \right\} \right\|_{\text{op}},$$

it holds $\mathbf{b}^2 \leq \delta^2$ and for any $z \geq 0$

$$\begin{aligned} \mathbb{P}\left\{\|S^{-1/2} \widehat{S} S^{-1/2} - I_p\|_{\text{op}} \geq \mathbf{b} z\right\} &\leq 2p \exp\left(\frac{-z^2/2}{1 + \delta^2 \mathbf{b}^{-1} z / 3}\right), \\ \mathbb{P}\left\{\|S^{-1/2} \widehat{S} S^{-1/2} - I_p\|_{\text{op}} \geq \delta z\right\} &\leq 2p \exp\left(\frac{-z^2/2}{1 + z\delta/3}\right). \end{aligned}$$

The result can be easily extended to the case when each ε_i is not bounded but can be bounded with a high probability.

Theorem C.3.2. *Let for some $y > 0$ there exists a constant $\delta(y)$ such that*

$$\mathbb{P}(\|S^{-1/2} \varepsilon\|_\infty > \delta(y)) \leq e^{-y}.$$

Then with \mathbf{b} from (C.19)

$$\mathbb{P}\{\|S^{-1/2} \widehat{S} S^{-1/2} - I_p\|_{\text{op}} \geq \mathbf{b} z\} \leq 2p \exp\left(\frac{-z^2/2}{1 + \delta^2(y) \mathbf{b}^{-1} z / 3}\right) + e^{-y}.$$

As a practical corollary, one deduces for moderate y that $z(y) \approx \sqrt{(2 + \alpha)y \log(2p)}$ for some small $\alpha > 0$ ensures

$$\mathbb{P}\left\{\|S^{-1/2} \widehat{S} S^{-1/2} - I_p\|_{\text{op}} \geq \mathbf{b} z(y)\right\} \leq 2e^{-y}.$$

To be done: The i.i.d. case

D

Gaussian comparison via KL-divergence and Pinsker's inequality

D.1 Pinsker's inequality

Suppose that two p -dimensional zero mean Gaussian vectors $\boldsymbol{\xi} \sim \mathcal{N}(0, S)$ and $\boldsymbol{\xi}^\flat \sim \mathcal{N}(0, S^\flat)$ are given. Let also T map \mathbb{R}^p to \mathbb{R}^M and $\mathbf{X} = T(\boldsymbol{\xi})$ and $\mathbf{Y} = T(\boldsymbol{\xi}^\flat)$. We aim to bound the distance between distributions of \mathbf{X} and \mathbf{Y} under the conditions

$$\begin{aligned} \|S^{-1/2}S^\flat S^{-1/2} - I_p\|_{\text{op}} &\leq \epsilon \leq 1/2, \\ \text{tr}\left(S^{-1/2}S^\flat S^{-1/2} - I_p\right)^2 &\leq \Delta^2 \end{aligned} \tag{D.1}$$

for some $\epsilon \leq 1/2$ and $\Delta \geq 0$. The next lemma bounds from above the Kullback-Leibler divergence between two normal distributions.

Lemma D.1.1. *Let $\mathbb{P}_0 = \mathcal{N}(\mathbf{b}, S)$ and $\mathbb{P}_1 = \mathcal{N}(\mathbf{b}^\flat, S^\flat)$ for some non-degenerated matrices S and S^\flat . If*

$$\begin{aligned} \|S^{-1/2}S^\flat S^{-1/2} - I_p\|_{\text{op}} &\leq 1/2, \\ \text{tr}\left\{(S^{-1/2}S^\flat S^{-1/2} - I_p)^2\right\} &\leq \Delta^2, \end{aligned}$$

then

$$\mathcal{K}(\mathbb{P}_0, \mathbb{P}_1) = -\mathbb{E}_0 \log \frac{d\mathbb{P}_1}{d\mathbb{P}_0} \leq \frac{\Delta^2}{2} + \frac{1}{2}(\mathbf{b} - \mathbf{b}^\flat)^\top S^\flat (\mathbf{b} - \mathbf{b}^\flat).$$

For any measurable set $A \subset \mathbb{R}^p$, it holds

$$|\mathbb{P}_0(A) - \mathbb{P}_1(A)| \leq \sqrt{\mathcal{K}(\mathbb{P}_0, \mathbb{P}_1)/2}.$$

Proof. The change of variables $\mathbf{u} = S^{-1/2}(\mathbf{x} - \mathbf{b})$ reduces the general case to the situation when \mathbb{P}_0 is standard normal in \mathbb{R}^p while $\mathbb{P}_1 = \mathcal{N}(\boldsymbol{\beta}, B)$ with $\boldsymbol{\beta} = S^{-1/2}(\mathbf{b}^\flat - \mathbf{b})$ and $B \stackrel{\text{def}}{=} S^{-1/2}S^\flat S^{-1/2}$

$$2 \log \frac{d\mathbb{P}_1}{d\mathbb{P}_0}(\boldsymbol{\gamma}) = \log \det(B) - (\boldsymbol{\gamma} - \boldsymbol{\beta})^\top B(\boldsymbol{\gamma} - \boldsymbol{\beta}) + \|\boldsymbol{\gamma}\|^2$$

with γ standard normal and

$$2\mathcal{K}(\mathbb{P}_0, \mathbb{P}_1) = -2\mathbb{E}_0 \log \frac{d\mathbb{P}_1}{d\mathbb{P}_0} = -\log \det(B) + \text{tr}(B - I_p) + \boldsymbol{\beta}^\top B \boldsymbol{\beta}.$$

Let a_j be the j th eigenvalue of $B - I_p$. The condition $\|B - I_p\|_{\text{op}} \leq 1/2$ yields $|a_j| \leq 1/2$ and

$$\begin{aligned} 2\mathcal{K}(\mathbb{P}_0, \mathbb{P}_1) &= \boldsymbol{\beta}^\top B \boldsymbol{\beta} + \sum_{j=1}^p \{a_j - \log(1 + a_j)\} \\ &\leq \boldsymbol{\beta}^\top B \boldsymbol{\beta} + \sum_{j=1}^p a_j^2 \\ &\leq \boldsymbol{\beta}^\top B \boldsymbol{\beta} + \text{tr}(B - I_p)^2 \leq \boldsymbol{\beta}^\top B \boldsymbol{\beta} + \Delta^2. \end{aligned}$$

This implies by Pinsker's inequality

$$\sup_A |\mathbb{P}_0(A) - \mathbb{P}_1(A)| \leq \sqrt{\frac{1}{2}\mathcal{K}(\mathbb{P}_0, \mathbb{P}_1)} \leq \frac{1}{2}\sqrt{\Delta^2 + \boldsymbol{\beta}^\top B \boldsymbol{\beta}} \quad (\text{D.2})$$

as required.

Notice that the operator norm bound

$$\|S^{-1/2} S^\flat S^{-1/2} - I_p\|_{\text{op}} \leq \epsilon \quad (\text{D.3})$$

implies for $B = S^{-1/2} S^\flat S^{-1/2}$

$$\text{tr}(B - I_p)^2 \leq p\epsilon^2, \quad \boldsymbol{\beta}^\top B \boldsymbol{\beta} \leq (1 + \epsilon)\|\boldsymbol{\beta}\|^2.$$

Corollary D.1.1. *Let $\mathbb{P}_0 = \mathcal{N}(\mathbf{b}, S)$ and $\mathbb{P}_1 = \mathcal{N}(\mathbf{b}^\flat, S^\flat)$ for some non-degenerated matrices S and S^\flat satisfying (D.3). Then with $\boldsymbol{\beta} = S^{-1/2}(\mathbf{b} - \mathbf{b}^\flat)$*

$$\sup_A |\mathbb{P}_0(A) - \mathbb{P}_1(A)| \leq \frac{1}{2}\sqrt{p\epsilon^2 + (1 + \epsilon)\|\boldsymbol{\beta}\|^2}$$

For the special case with $\boldsymbol{\beta} \equiv 0$, we bound for any Borel set $A \subset \mathbb{R}^M$

$$|\mathbb{P}(T(\boldsymbol{\xi}) \in A) - \mathbb{P}(T(\boldsymbol{\xi}^\flat) \in A)| \leq \Delta/2.$$

We state a separate corollary for the distribution of the maximum.

D.2 Gaussian comparison

Corollary D.2.1. *Let two p -dimensional zero mean Gaussian vectors $\boldsymbol{\xi} \sim \mathcal{N}(0, S)$ and $\boldsymbol{\xi}^\flat \sim \mathcal{N}(0, S^\flat)$ be given, and (D.1) holds. Then for any mapping $T: \mathbb{R}^p \rightarrow \mathbb{R}^M$ and any set of values (q_η) , the random vectors $\mathbf{X} = T(\boldsymbol{\xi})$ and $\mathbf{Y} = T(\boldsymbol{\xi}^\flat)$ fulfill*

$$|\mathbb{P}(\max_{\eta} X_{\eta} - q_{\eta} > 0) - \mathbb{P}(\max_{\eta} Y_{\eta} - q_{\eta} > 0)| \leq \Delta/2.$$

Proof. We simply apply the result of the lemma to the set $A = \{\mathbf{x} \in \mathbb{R}^p : T(\mathbf{x}) \leq z\}$.

Interestingly, this method can be used for obtaining an anti-concentration bound in the case of a homogeneous mapping $T : \mathbb{R}^p \rightarrow \mathbb{R}^M$.

Theorem D.2.1. *Let $\xi \sim \mathcal{N}(0, S)$ be a Gaussian vector in \mathbb{R}^p . For any homogeneous mapping $T : \mathbb{R}^p \rightarrow \mathbb{R}^M$, and for any $z > 0$ and Δ satisfying $0 \leq \Delta/z \leq 1$, it holds*

$$\mathbb{P}(\max_{\eta} T_{\eta}(\xi) \geq z) - \mathbb{P}(\max_{\eta} T_{\eta}(\xi) \geq z + \Delta) \leq \Delta z^{-1} \sqrt{p/2}.$$

Moreover, if $\xi^b \sim \mathcal{N}(0, S^b)$ is another Gaussian vector and (D.1) holds with $\epsilon \leq 1/2$ and some $\Delta \geq 0$, then

$$|\mathbb{P}(\max_{\eta} T_{\eta}(\xi) \geq z) - \mathbb{P}(\max_{\eta} T_{\eta}(\xi^b) \geq z + \Delta)| \leq \Delta/2 + \Delta z^{-1} \sqrt{p/2}. \quad (\text{D.4})$$

Proof. Given z and Δ , define $\xi^b = z/(z + \Delta)\xi$. It holds by homogeneity of T

$$\mathbb{P}(\max_{\eta} T_{\eta}(\xi) \geq z + \Delta) = \mathbb{P}(\max_{\eta} T_{\eta}(\xi^b) \geq z).$$

It is obvious that $\text{Var}(\xi^b) = (1 + \Delta/z)^{-2}S$. Now it holds for the KL-divergence between ξ and ξ^b

$$\mathcal{K}(\mathbb{P}_{\xi}, \mathbb{P}_{\xi^b}) = \frac{p}{2} \{2\Delta/z + (\Delta/z)^2 - 2\log(1 + \Delta/z)\} \leq p(\Delta/z)^2. \quad (\text{D.5})$$

Here we used that $\log(1 + \rho) \leq \rho - \rho^2/2$ for $\rho \leq 1$. Now Pinsker's bound (D.2) implies

$$\begin{aligned} & |\mathbb{P}(\max_{\eta} T_{\eta}(\xi) \geq z) - \mathbb{P}(\max_{\eta} T_{\eta}(\xi^b) \geq z + \Delta)| \\ & \leq \mathbb{P}(\max_{\eta} T_{\eta}(\xi) \geq z) - \mathbb{P}(\max_{\eta} T_{\eta}(\xi) \geq z + \Delta) \\ & \quad + |\mathbb{P}(\max_{\eta} T_{\eta}(\xi) \geq z + \Delta) - \mathbb{P}(\max_{\eta} T_{\eta}(\xi^b) \geq z + \Delta)| \\ & \leq \Delta/2 + \Delta z^{-1} \sqrt{p/2} \end{aligned}$$

and (D.4) follows.

We also present a simple corollary of the above result which concerns the change in the expectation $\mathbb{E}f(\xi)$ for a bounded function f .

Lemma D.2.1. *Let $\xi \sim \mathcal{N}(0, S)$ and $\xi^b \sim \mathcal{N}(0, S^b)$, where S, S^b satisfy (D.1). For any function f on \mathbb{R}^p with $|f(\mathbf{x})| \leq 1$, and any $\delta > 0$ it holds*

$$|\mathbb{E}f(\xi) - \mathbb{E}f(\xi^b)| \leq \Delta. \quad (\text{D.6})$$

Also, for any $\delta \geq 0$

$$|\mathbb{E}f(\xi) - \mathbb{E}f((1 + \delta)\xi)| \leq \delta\sqrt{2p}.$$

Proof. in view of $|f(\mathbf{x})| \leq 1$, it holds

$$|\mathbb{E}f(\xi) - \mathbb{E}f(\xi^b)| \leq \int |f(\mathbf{x})| \cdot |\phi_\xi(\mathbf{x}) - \phi_{\xi^b}(\mathbf{x})| d\mathbf{x} \leq \int |\phi_\xi(\mathbf{x}) - \phi_{\xi^b}(\mathbf{x})| d\mathbf{x}.$$

One more use of Pinsker's inequality yields

$$\int |\phi_\xi(\mathbf{x}) - \phi_{\xi^b}(\mathbf{x})| d\mathbf{x} = 2\|\mathbb{P}_\xi - \mathbb{P}_{\xi^b}\|_{TV} \leq \sqrt{2\mathcal{K}(\mathbb{P}_\xi, \mathbb{P}_{\xi^b})},$$

and the assertion (D.6) follows by $2\mathcal{K}(\mathbb{P}_\xi, \mathbb{P}_{\xi^b}) \leq \Delta^2$. It remains to note that for $S^b = (1 + \delta)^2 S$, it holds $\mathcal{K}(\mathbb{P}_\xi, \mathbb{P}_{\xi^b}) \leq \delta^2 p$; see (D.5).

E

Random multiplicity correction

E.1 Gaussian measures with random covariance

Suppose that V^\flat is a random positive symmetric $p \times p$ matrix close to a deterministic matrix V . Below we use the operator norm for quantifying the difference between V and V^\flat : namely let with probability one

$$\|V^{-1/2} V^\flat V^{-1/2} - I_p\|_{\text{op}} \leq \Delta_0. \quad (\text{E.1})$$

In what follows, $\mathbb{P} = \mathcal{N}(0, V)$ is the normal measure on \mathbb{R}^p with mean zero and covariance V . Similarly \mathbb{P}^\flat is a random measure on \mathbb{R}^p which is conditionally on V^\flat normal with $\mathbb{P}^\flat = \mathcal{N}(0, V^\flat)$. Suppose that for each m from a given set \mathcal{M} , a linear mapping $T_m: \mathbb{R}^p \rightarrow \mathbb{R}^{p_m}$ is fixed. Given \mathbf{x} , define for each $m \in \mathcal{M}$ the corresponding tail function $z_m(\mathbf{x})$ by

$$\mathbb{P}\{\mathbf{u}: \|T_m \mathbf{u}\| \geq z_m(\mathbf{x})\} = e^{-\mathbf{x}}. \quad (\text{E.2})$$

Also define a set $A(\mathbf{x})$ as

$$A(\mathbf{x}) \stackrel{\text{def}}{=} \bigcap_{m \in \mathcal{M}} \{\mathbf{u}: \|T_m \mathbf{u}\| \leq z_m(\mathbf{x})\} = \{\mathbf{u}: \|T_m \mathbf{u}\| \leq z_m(\mathbf{x}), \forall m \in \mathcal{M}\}.$$

Similarly, for each $m \in \mathcal{M}$ and $\mathbf{x} > 0$, define $z_m^\flat(\mathbf{x})$ by (E.2) with \mathbb{P}^\flat in place of \mathbb{P} , and introduce the corresponding set $A^\flat(\mathbf{x})$. Note that all these objects are random because \mathbb{P}^\flat is random. Finally, let $\bar{\mathbf{x}}_\alpha$ be the random quantity providing

$$\mathbb{P}^\flat(A^\flat(\bar{\mathbf{x}}_\alpha)) = 1 - \alpha. \quad (\text{E.3})$$

Below we try to address the question whether this random multiplicity correction based on (E.3) does a good job under \mathbb{P} . This question leads to analysis of value $\mathbb{P}(A^\flat(\bar{\mathbf{x}}_\alpha))$: the goal is in evaluating the difference

$$\mathbb{P}(A^\flat(\bar{\mathbf{x}}_\alpha)) - (1 - \alpha).$$

Theorem E.1.1. *Let a random matrix V^\flat satisfy (E.1) for a deterministic matrix V and $\Delta_0 < 1/2$. Then it holds*

$$|\mathbb{I}P(A^\flat(\bar{x}_\alpha)) - 1 + \alpha| \leq \sqrt{p} \Delta_0. \quad (\text{E.4})$$

Proof. The key property of $\mathbb{I}P^\flat = \mathcal{N}(0, V^\flat)$ is that the random matrix V^\flat concentrates around some deterministic matrix. Below we use this property in the bracketing form:

$$\begin{aligned} V^- &\leq V^\flat \leq V^+ \\ V^- &\stackrel{\text{def}}{=} (1 - \Delta_0)V, \quad V^+ \stackrel{\text{def}}{=} (1 + \Delta_0)V, \quad V^+ - V^- = 2\Delta_0V. \end{aligned} \quad (\text{E.5})$$

In other words, the random matrix V^\flat can be sandwiched in two deterministic matrices V^- and V^+ . For the proof of (E.4) we use the following well known property of the Gaussian distribution.

Lemma E.1.1. *Let $\mathbb{I}P_1 \sim \mathcal{N}(0, V_1)$ and $\mathbb{I}P_2 \sim \mathcal{N}(0, V_2)$ with $V_1 \leq V_2$. Then for any centrally symmetric star-shaped set A , it holds*

$$\mathbb{I}P_1(A) \geq \mathbb{I}P_2(A).$$

Proof. The statement is trivial in the univariate case, the general case is obtained by integration over A in polar coordinates.

Introduce two Gaussian measures $\mathbb{I}P^- = \mathcal{N}(0, V^-)$ and $\mathbb{I}P^+ = \mathcal{N}(0, V^+)$; see (E.5). Let $z_m^-(x)$ and $z_m^+(x)$ be the corresponding tail functions, and $A^-(x)$ and $A^+(x)$ - the corresponding sets. The identities (E.5) yield

$$\mathbb{I}P^+(A^+(x)) = \mathbb{I}P^-(A^-(x)). \quad (\text{E.6})$$

Lemma E.1.1 implies by (E.5) for any x

$$\mathbb{I}P^+(A(x)) \leq \mathbb{I}P^\flat(A(x)) \leq \mathbb{I}P^-(A(x)). \quad (\text{E.7})$$

The key step of the proof is given by the next lemma where we sandwich the random set $A^\flat(\bar{x})$ in two specially constructed deterministic sets.

Lemma E.1.2. *Let the deterministic values x_α^- and x_α^+ be define by*

$$\mathbb{I}P^+(A^-(x_\alpha^+)) = 1 - \alpha, \quad \mathbb{I}P^-(A^+(x_\alpha^-)) = 1 - \alpha. \quad (\text{E.8})$$

Then

$$\begin{aligned} x_\alpha^- &\leq \bar{x}_\alpha \leq x_\alpha^+ \\ A^-(x_\alpha^-) &\subseteq A^\flat(\bar{x}_\alpha) \subseteq A^+(x_\alpha^+). \end{aligned} \quad (\text{E.9})$$

Proof. By Lemma E.1.1 the following relations hold true for any \mathbf{x} :

$$\begin{aligned} z_m^-(\mathbf{x}) &\leq z_m^\flat(\mathbf{x}) \leq z_m^+(\mathbf{x}), \\ A^-(\mathbf{x}) &\subseteq A^\flat(\mathbf{x}) \subseteq A^+(\mathbf{x}). \end{aligned} \quad (\text{E.10})$$

Now by definition (E.8) in view of (E.7) and (E.10)

$$\begin{aligned} \mathbb{I}P^\flat(A^\flat(\mathbf{x}_\alpha^+)) &\geq \mathbb{I}P^+(A^\flat(\mathbf{x}_\alpha^+)) \geq \mathbb{I}P^+(A^-(\mathbf{x}_\alpha^+)) = 1 - \alpha, \\ \mathbb{I}P^\flat(A^\flat(\mathbf{x}_\alpha^-)) &\leq \mathbb{I}P^-(A^\flat(\mathbf{x}_\alpha^-)) \leq \mathbb{I}P^-(A^+(\mathbf{x}_\alpha^-)) = 1 - \alpha. \end{aligned}$$

This yields by monotonicity of $\mathbb{I}P^\flat(A^\flat(\mathbf{x}))$ in \mathbf{x} that $\bar{\mathbf{x}}_\alpha$ from (E.3) belongs to the interval $[\mathbf{x}_\alpha^-, \mathbf{x}_\alpha^+]$ and

$$A^-(\mathbf{x}_\alpha^-) \subseteq A^\flat(\mathbf{x}_\alpha^-) \subseteq A^\flat(\bar{\mathbf{x}}_\alpha) \subseteq A^\flat(\mathbf{x}_\alpha^+) \subseteq A^+(\mathbf{x}_\alpha^+).$$

This implies the result.

Now we can finalize the proof. The relations (E.9) and (E.6) imply

$$\mathbb{I}P^+(A^\flat(\bar{\mathbf{x}}_\alpha)) \leq \mathbb{I}P^+(A^+(\mathbf{x}_\alpha^+)) = \mathbb{I}P^-(A^-(\mathbf{x}_\alpha^+)).$$

Furthermore, it holds by Pinsker' inequality (Corollary D.1.1) in view of (E.1) and (E.8)

$$\mathbb{I}P^-(A^-(\mathbf{x}_\alpha^+)) \leq \mathbb{I}P^+(A^-(\mathbf{x}_\alpha^+)) + \sqrt{p} \Delta_0 \leq 1 - \alpha + \sqrt{p} \Delta_0.$$

Similarly

$$\begin{aligned} \mathbb{I}P^-(A^\flat(\bar{\mathbf{x}}_\alpha)) &\geq \mathbb{I}P^-(A^-(\mathbf{x}_\alpha^-)) = \mathbb{I}P^+(A^+(\mathbf{x}_\alpha^-)) \\ &\geq \mathbb{I}P^-(A^+(\mathbf{x}_\alpha^-)) - \sqrt{p} \Delta_0 = 1 - \alpha - \sqrt{p} \Delta_0. \end{aligned}$$

This implies (E.4) for the measure $\mathbb{I}P$.

E.2 Max-case

Now we consider a more general situation. Let T_m be a family of test statistics in the real world, and \mathbb{T}_m^\flat

F

High-dimensional inference for a Gaussian law

This section collects some useful facts about high dimensional Gaussian measures.

F.1 Stein identity, Slepian bridge, and Gaussian comparison

Below for a $M \times M$ matrix A , we denote

$$\begin{aligned}\|A\|_{\text{op}} &= \sup_{\|\mathbf{u}\|=1} \|A\mathbf{u}\|, & \|A\|_{\infty} &= \max_{i,j} |a_{i,j}|, \\ \|A\|_1 &= \sum_{i,j} |a_{i,j}|, & \|A\|_{Fr}^2 &= \sum_{ij} a_{ij}^2.\end{aligned}$$

Lemma F.1.1. *Let \mathbf{X} and \mathbf{Y} be two zero mean Gaussian vectors in \mathbb{R}^M with $\Sigma_{\mathbf{X}} = \text{Var}(\mathbf{X})$ and $\Sigma_{\mathbf{Y}} = \text{Var}(\mathbf{Y})$. Let also $f(\mathbf{x})$ be a smooth function on \mathbb{R}^M . Then*

$$\epsilon \stackrel{\text{def}}{=} |\mathbb{E}f(\mathbf{X}) - \mathbb{E}f(\mathbf{Y})| \leq \frac{1}{2} \|\Sigma_{\mathbf{X}} - \Sigma_{\mathbf{Y}}\|_{\infty} \|\nabla^2 f\|_{1,\infty}, \quad (\text{F.1})$$

where $\|\nabla^2 f\|_{1,\infty} \stackrel{\text{def}}{=} \sup_{\mathbf{x}} \|\nabla^2 f(\mathbf{x})\|_1$.

Proof. Without loss of generality assume that \mathbf{X} and \mathbf{Y} are given on the same probability space and independent. For each $t \in [0, 1]$, define

$$\begin{aligned}\mathbf{Z}(t) &\stackrel{\text{def}}{=} \sqrt{t} \mathbf{X} + \sqrt{1-t} \mathbf{Y}, \\ \Psi(t) &\stackrel{\text{def}}{=} \mathbb{E}f(\mathbf{Z}(t)) = \mathbb{E}f(\sqrt{t} \mathbf{X} + \sqrt{1-t} \mathbf{Y}).\end{aligned}$$

Obviously

$$\epsilon = |\Psi(1) - \Psi(0)| = \left| \int_0^1 \Psi'(t) dt \right|. \quad (\text{F.2})$$

Further,

$$\begin{aligned}\Psi'(t) &= \mathbb{E}[\nabla f(\mathbf{Z}(t))^\top \mathbf{Z}'(t)] \\ &= \frac{1}{2} \mathbb{E}[\{\mathbf{t}^{-1/2} \mathbf{X} - (1-t)^{-1/2} \mathbf{Y}\}^\top \nabla f(\mathbf{Z}(t))].\end{aligned}\quad (\text{F.3})$$

To compute this expectation, we apply the *Stein identity*. Let \mathbf{W} be a zero mean Gaussian vector in \mathbb{R}^M . Then for any C^1 function $s: \mathbb{R}^M \rightarrow \mathbb{R}^M$, it holds

$$\mathbb{E}[\mathbf{W} s(\mathbf{W})] = \text{Var}(\mathbf{W}) \mathbb{E}[\nabla s(\mathbf{W})]. \quad (\text{F.4})$$

Exercise F.1.1. Prove (F.4) for standard normal \mathbf{W} using integration by part:

$$\int_{\mathbb{R}^M} s(\mathbf{w}) \mathbf{w} e^{-\|\mathbf{w}\|^2/2} d\mathbf{w} = \int_{\mathbb{R}^M} \nabla s(\mathbf{w}) e^{-\|\mathbf{w}\|^2/2} d\mathbf{w}.$$

Reduce the case of a Gaussian zero mean $\mathbf{w} \sim \mathcal{N}(0, \Sigma)$ with a positive symmetric matrix Σ to the case $\Sigma = I_M$.

This results can be directly extended to any C^1 vector function $s: \mathbb{R}^M \rightarrow \mathbb{R}^q$: it holds

$$\mathbb{E}[\mathbf{W} s(\mathbf{W})^\top] = \text{Var}(\mathbf{W}) \mathbb{E}[\nabla s(\mathbf{W})^\top]. \quad (\text{F.5})$$

Here $\nabla s(\mathbf{w})^\top$ means the $p \times q$ matrix with the entries $\frac{d}{d\theta_j} s_m(\mathbf{w})$ for $j = 1, \dots, p$ and $m = 1, \dots, q$.

Exercise F.1.2. Derive (F.5) by applying (F.4) columnwise.

The identity (F.5) is used with $\mathbf{W} = (\mathbf{X}^\top, \mathbf{Y}^\top)^\top$ and $s(\mathbf{w}) = \nabla f(\mathbf{z}(t))$ for $\mathbf{z}(t) = \sqrt{t} \mathbf{x} + \sqrt{1-t} \mathbf{y}$. Independence of \mathbf{X} and \mathbf{Y} implies

$$\text{Var}(\mathbf{W}) = \begin{pmatrix} \Sigma_{\mathbf{X}} & 0 \\ 0 & \Sigma_{\mathbf{Y}} \end{pmatrix}.$$

Also $\nabla s(\mathbf{w}) = (t^{1/2} \nabla^2 f(\mathbf{z}(t)), (1-t)^{1/2} \nabla^2 f(\mathbf{z}(t)))^\top$ and by (F.5)

$$\begin{aligned}\mathbb{E}[\nabla f(\mathbf{Z}(t)) \mathbf{X}^\top] &= t^{1/2} \Sigma_{\mathbf{X}} \mathbb{E}[\nabla^2 f(\mathbf{Z}(t))] \\ \mathbb{E}[\nabla f(\mathbf{Z}(t)) \mathbf{Y}^\top] &= (1-t)^{1/2} \Sigma_{\mathbf{Y}} \mathbb{E}[\nabla^2 f(\mathbf{Z}(t))],\end{aligned}$$

This and (F.3) imply

$$\begin{aligned}|\Psi'(t)| &\leq \frac{1}{2} \left| \text{tr} \{ (\Sigma_{\mathbf{X}} - \Sigma_{\mathbf{Y}}) \mathbb{E}[\nabla^2 f(\mathbf{Z}(t))] \} \right| \\ &\leq \frac{1}{2} \|\Sigma_{\mathbf{X}} - \Sigma_{\mathbf{Y}}\|_\infty \|\mathbb{E}[\nabla^2 f(\mathbf{Z}(t))]\|_1 \leq \frac{1}{2} \|\Sigma_{\mathbf{X}} - \Sigma_{\mathbf{Y}}\|_\infty \|\nabla^2 f\|_{1,\infty}.\end{aligned}$$

Now the assertion follows from (F.2).

Now we apply the obtained bound to $f(\mathbf{x}) \stackrel{\text{def}}{=} g(\Delta^{-1}h_\beta(\mathbf{x}))$, where $g(z)$ is a smooth univariate function with bounded first and second derivatives, and the *smooth maximum* function: for some $\beta > 0$

$$h_\beta(\mathbf{x}) = \beta^{-1} \log \left(\sum_{j=1}^M e^{\beta x_j} \right). \quad (\text{F.6})$$

Lemma F.1.2. *Let \mathbf{X} and \mathbf{Y} be two zero mean Gaussian vectors in \mathbb{R}^M with $\Sigma_{\mathbf{X}} = \text{Var}(\mathbf{X})$ and $\Sigma_{\mathbf{Y}} = \text{Var}(\mathbf{Y})$. For a univariate function $g(z)$ with bounded first and second derivatives, and $h_\beta(\mathbf{x})$ from (F.6)*

$$|\mathbb{E}g(\Delta^{-1}h_\beta(\mathbf{X})) - \mathbb{E}g(\Delta^{-1}h_\beta(\mathbf{Y}))| \leq \left(\frac{\beta \|g'\|_\infty}{\Delta} + \frac{\|g''\|_\infty}{2\Delta^2} \right) \|\Sigma_{\mathbf{X}} - \Sigma_{\mathbf{Y}}\|_\infty. \quad (\text{F.7})$$

Proof. It holds for $f(\mathbf{x}) \stackrel{\text{def}}{=} g(\Delta^{-1}h_\beta(\mathbf{x}))$

$$\nabla f(\mathbf{x}) = \Delta^{-1}g'(\Delta^{-1}h_\beta(\mathbf{x}))\nabla h_\beta(\mathbf{x}),$$

$$\nabla^2 f(\mathbf{x}) = \Delta^{-1}g'(\Delta^{-1}h_\beta(\mathbf{x}))\nabla^2 h_\beta(\mathbf{x}) + \Delta^{-2}g''(\Delta^{-1}h_\beta(\mathbf{x}))\nabla h_\beta(\mathbf{x})\nabla h_\beta(\mathbf{x})^\top.$$

Also for any \mathbf{x} by direct calculus

$$\|\nabla h_\beta(\mathbf{x})\|_1 = 1,$$

$$\|\nabla^2 h_\beta(\mathbf{x})\|_1 \leq 2\beta.$$

This implies

$$\begin{aligned} \|\nabla f(\mathbf{x})\|_1 &\leq \Delta^{-1}\|g'\|_\infty \times \|\nabla h_\beta(\mathbf{x})\|_1 \leq \Delta^{-1}\|g'\|_\infty, \\ \|\nabla^2 f(\mathbf{x})\|_1 &\leq \Delta^{-1}\|g'\|_\infty \times \|\nabla^2 h_\beta(\mathbf{x})\|_1 + \Delta^{-2}\|g''\|_\infty \times \|\nabla h_\beta(\mathbf{x})\|_1^2 \\ &\leq 2\Delta^{-1}\beta\|g'\|_\infty + \Delta^{-2}\|g''\|_\infty. \end{aligned}$$

Now (F.7) follows from (F.1).

A particular choice of the function g is given by

$$g(z) \stackrel{\text{def}}{=} \begin{cases} 2u^2, & u \in [0, 1/2], \\ 1 - 2(1-u)^2, & u \in [1/2, 1]. \\ 0 & \text{otherwise.} \end{cases} \quad (\text{F.8})$$

Obviously $|g'(u)| \leq 2$, $|g''(u)| \leq 4$ for all u . Then $\|g'_\Delta\|_\infty \leq 2\Delta^{-1}$, $\|g''_\Delta\|_\infty \leq 4\Delta^{-2}$. We conclude with the following bound.

Theorem F.1.1. Let \mathbf{X} and \mathbf{Y} be two zero mean Gaussian vectors in $\mathbb{R}^{\mathbb{M}}$ with $\Sigma_{\mathbf{X}} = \text{Var}(\mathbf{X})$ and $\Sigma_{\mathbf{Y}} = \text{Var}(\mathbf{Y})$. Then with $g(\cdot)$ given by (F.8), it holds for any $\Delta > 0$ and $\beta > 0$

$$|\mathbb{E}g(\Delta^{-1}h_{\beta}(\mathbf{X})) - \mathbb{E}g(\Delta^{-1}h_{\beta}(\mathbf{Y}))| \leq 2(\beta\Delta^{-1} + \Delta^{-2})\|\Sigma_{\mathbf{X}} - \Sigma_{\mathbf{Y}}\|_{\infty}. \quad (\text{F.9})$$

F.2 Comparing of the maximum of Gaussians

Let $\mathbf{X} = (X_j)$ and $\mathbf{Y} = (Y_j)$ be two zero mean Gaussian vectors in $\mathbb{R}^{\mathbb{M}}$ with $\Sigma_{\mathbf{X}} = \text{Var}(\mathbf{X})$ and $\Sigma_{\mathbf{Y}} = \text{Var}(\mathbf{Y})$, and let $\square = \|\Sigma_{\mathbf{X}} - \Sigma_{\mathbf{Y}}\|_{\infty}$. Now we aim at comparing the distributions of $\max_j X_j$ and $\max_j Y_j$. We use that the smooth maximum h_{β} fulfills

$$\max_j x_j \leq h_{\beta}(\mathbf{x}) \leq \max_j x_j + \beta^{-1} \log(\mathbb{M}).$$

As the indicator function $\mathbb{I}(z \geq 0)$ is not differentiable, we approximate it by a smooth function g_{Δ} . Namely, select a two times differentiable function g with $g(u) = 0$ for $u \leq 0$, $g(u) = 1$ for $u \geq 1$, and $g(u)$ monotonously grows from zero to one when u grows from zero to one. Define also $g_{\Delta}(u) = g(\Delta^{-1}u)$ for $\Delta > 0$. With $\Delta = \beta^{-1} \log(\mathbb{M})$

$$g_{\Delta} \circ h_{\beta}(\mathbf{x} - \boldsymbol{\Delta}) \leq \mathbb{I}(\max_j x_j > 0) \leq g_{\Delta} \circ h_{\beta}(\mathbf{x} + \boldsymbol{\Delta}).$$

Here $\boldsymbol{\Delta}$ is the vector with all entries equal to Δ . Indeed, $g_{\Delta}(z) \in [0, 1]$ for any z . If $x_j \geq 0$ for some j , then $h_{\beta}(\mathbf{x} + \boldsymbol{\Delta}) \geq \Delta$ and hence,

$$g_{\Delta} \circ h_{\beta}(\mathbf{x} + \boldsymbol{\Delta}) \geq g(\Delta/\Delta) = g(1) = 1.$$

Similarly, if $\max_j x_j \leq 0$, then due to $\Delta = \beta^{-1} \log(\mathbb{M})$

$$h_{\beta}(\mathbf{x} - \boldsymbol{\Delta}) \leq \max_j (x_j - \Delta) + \beta^{-1} \log(\mathbb{M}) \leq 0$$

and $g_{\Delta} \circ h_{\beta}(\mathbf{x} - \boldsymbol{\Delta}) = 0$. This and (F.9) yield the bound

$$\begin{aligned} \mathbb{P}\left(\max_j X_j > 0\right) &\leq \mathbb{E}[g_{\Delta} \circ h_{\beta}(\mathbf{X} + \boldsymbol{\Delta})] \\ &\leq \mathbb{E}[g_{\Delta} \circ h_{\beta}(\mathbf{Y} + \boldsymbol{\Delta})] + 2(\beta\Delta^{-1} + \Delta^{-2})\square \\ &\leq \mathbb{P}\left(\max_j Y_j > -2\Delta\right) + 2\Delta^{-2}\{\log(\mathbb{M}) + 1\}\square. \end{aligned}$$

Similarly one can approximate any indicator $\mathbb{I}(z \geq z_0)$ by shifting the function g .

Theorem F.2.1. Let \mathbf{X} and \mathbf{Y} be two zero mean Gaussian vectors in $\mathbb{R}^{\mathbb{M}}$ with $\Sigma_{\mathbf{X}} = \text{Var}(\mathbf{X})$ and $\Sigma_{\mathbf{Y}} = \text{Var}(\mathbf{Y})$.

$$\square \stackrel{\text{def}}{=} \|\Sigma_{\mathbf{X}} - \Sigma_{\mathbf{Y}}\|_{\infty},$$

it holds for any Δ and z

$$\mathbb{P}\left(\max_j X_j > z\right) \leq \mathbb{P}\left(\max_j Y_j > z - 2\Delta\right) + 2\Delta^{-2}\{\log(\mathbb{M}) + 1\} \square. \quad (\text{F.10})$$

F.3 Anti-concentration for Gaussian maxima

This section explains how one can compare the distribution of two maxima using the anti-concentration bound. The obtained results allow to bound the probability $\mathbb{P}(\max_j X_j > 0)$ by a similar probability $\mathbb{P}(\max_j Y_j > -2\Delta)$ from above and $\mathbb{P}(\max_j Y_j > 2\Delta)$ from below up to the error term $2\Delta^{-2}\{\log(\mathbb{M}) + 1\} \square$. The next question is whether we can replace 2Δ or -2Δ by zero without an essential change of probability. In other words, we have to bound the difference

$$\mathbb{P}\left(\max_j Y_j > -2\Delta\right) - \mathbb{P}\left(\max_j Y_j > 0\right).$$

The following theorem provides bounds on the Lévy concentration function of the maximum of a Gaussian random vector in $\mathbb{R}^{\mathbb{M}}$, where the terminology is borrowed from Chernozhukov et al. (2013). The Lévy concentration function of a real valued random variable ξ is defined for $\varepsilon > 0$ as

$$\mathcal{L}(\xi, \varepsilon) = \sup_{x \in \mathbb{R}} \mathbb{P}(|\xi - x| \leq \varepsilon).$$

Theorem F.3.1 (Anti-concentration). *Let $(X_1, \dots, X_{p_n})^\top$ be a centered Gaussian random vector in $\mathbb{R}^{\mathbb{M}}$ with $\sigma_j^2 = \mathbb{E}[X_j^2] > 0$ for all $1 \leq j \leq \mathbb{M}$. Moreover, let $\underline{\sigma} = \min_{1 \leq j \leq \mathbb{M}} \sigma_j$, $\bar{\sigma} = \max_{1 \leq j \leq \mathbb{M}} \sigma_j$, and $a_{\mathbb{M}} = \mathbb{E}[\max_{1 \leq j \leq \mathbb{M}} (X_j / \sigma_j)]$.*

1. If the variances are all equal, namely $\underline{\sigma} = \bar{\sigma} = \sigma$, then for every $\epsilon > 0$,

$$\mathcal{L}\left(\max_{1 \leq j \leq \mathbb{M}} X_j, \epsilon\right) \leq 4\epsilon(a_{\mathbb{M}} + 1)/\sigma;$$

2. If the variances are not equal, namely $\underline{\sigma} < \bar{\sigma}$, then for every $\epsilon > 0$,

$$\mathcal{L}\left(\max_{1 \leq j \leq \mathbb{M}} X_j, \epsilon\right) \leq C\epsilon\{a_{\mathbb{M}} + 1 \vee \log(\underline{\sigma}/\epsilon)\}$$

where $C > 0$ depends only on $\underline{\sigma}$ and $\bar{\sigma}$.

To compare the distribution of two maxima, we use the anti-concentration bound: if $\text{Var}(Y_j) \equiv \sigma^2$

$$\mathbb{P}\left(\max_j Y_j > 0\right) - \mathbb{P}\left(\max_j Y_j > -2\Delta\right) \leq 8\Delta(a_{\mathbb{M}} + 1)/\sigma,$$

where $a_{\mathbb{M}} \stackrel{\text{def}}{=} \mathbb{E} \max_j |Y_j/\sigma| \leq (2 \log \mathbb{M})^{1/2}$. If the variances $\sigma_j^2 \stackrel{\text{def}}{=} \text{Var}(Y_j)$ are unequal then

$$\mathbb{P}\left(\max_j Y_j > 0\right) - \mathbb{P}\left(\max_j Y_j > -2\Delta\right) \leq c\Delta\sqrt{\log(\mathbb{M}/\Delta)}. \quad (\text{F.11})$$

We now apply all the inequalities with the following choice: with $\Delta = b^{-1} = \beta^{-1} \log(\mathbb{M})$ and $\mathbb{Q} = \mathbb{M}/\Delta$

$$b = \square^{-1/3} \{\log(\mathbb{Q})\}^{-1/6}.$$

It follows by (F.9)

$$\begin{aligned} & |\mathbb{P}\left(\max_j X_j > 0\right) - \mathbb{P}\left(\max_j Y_j > 0\right)| \\ & \leq (2\beta\Delta^{-1} + 2\Delta^{-2}) \square + c\Delta\sqrt{\log(\mathbb{Q})} \\ & \leq c\square^{1/3} \log^{2/3}(\mathbb{Q}) + c\square^{1/3} \log^{2/3}(\mathbb{Q}) \\ & \leq c\square^{1/3} \log^{2/3}(\mathbb{Q}). \end{aligned}$$

The definition yields $\mathbb{Q} = \mathbb{M}/\Delta \leq \mathbb{M}/\square^{1/3}$. We conclude with the following result.

Theorem F.3.2. *Let \mathbf{X} and \mathbf{Y} be two zero mean Gaussian vectors in $\mathbb{R}^{\mathbb{M}}$ with $\Sigma_{\mathbf{X}} = \text{Var}(\mathbf{X})$ and $\Sigma_{\mathbf{Y}} = \text{Var}(\mathbf{Y})$. With*

$$\square \stackrel{\text{def}}{=} \|\Sigma_{\mathbf{X}} - \Sigma_{\mathbf{Y}}\|_{\infty},$$

it holds for any Δ

$$|\mathbb{P}\left(\max_j X_j > 0\right) - \mathbb{P}\left(\max_j Y_j > 0\right)| \leq c\square^{1/3} \log^{2/3}(\mathbb{M}/\square^{1/3}).$$

G

Gaussian approximation of a vector sum

This chapter discusses an approximation of a vector sum \mathbf{S} of independent random variables ε_i by a similar sum of Gaussian random variables with the same first and second moments in different norms.

G.1 A univariate case with Lindeberg telescopic sums

Consider a sample $\varepsilon_1, \dots, \varepsilon_n$ of independent centered random variables with $v_i^2 = \text{Var}(\varepsilon_i)$. We aim at bounding the error of Gaussian approximation (GAR) of the sum $\sum_i \varepsilon_i$. Let f be a smooth non-negative function satisfying $\|f^{(3)}\|_\infty / 6 \leq C_f$. Let also $\tilde{\varepsilon}_i \sim \mathcal{N}(0, v_i^2)$ be another set of independent zero mean Gaussian random variables with the same mean and variance as the original r.v.'s ε_i . Define for $k = 2, \dots, n$ the telescopic sums

$$S_k \stackrel{\text{def}}{=} \sum_{i=1}^{k-1} \varepsilon_i + \sum_{i=k+1}^n \tilde{\varepsilon}_i.$$

It obviously holds for $S \stackrel{\text{def}}{=} \sum_i \varepsilon_i$ and $\tilde{S} \stackrel{\text{def}}{=} \sum_i \tilde{\varepsilon}_i \sim \mathcal{N}(0, V^2)$

$$f(S) - f(\tilde{S}) = \sum_{k=1}^n \{f(S_k + \varepsilon_k) - f(S_k + \tilde{\varepsilon}_k)\}$$

Next, for each k , the Taylor expansion of the third order at S_k implies

$$\begin{aligned} & |f(S_k + \varepsilon_k) - f(S_k + \tilde{\varepsilon}_k) - f'(S_k)(\varepsilon_k - \tilde{\varepsilon}_k) - f''(S_k)(\varepsilon_k^2 - \tilde{\varepsilon}_k^2)/2| \\ & \leq \|f^{(3)}\|_\infty (|\varepsilon_k|^3 + |\tilde{\varepsilon}_k|^3)/6. \end{aligned}$$

As S_k and ε_k are independent and $I\!\!E(\varepsilon_k - \tilde{\varepsilon}_k) = I\!\!E(\varepsilon_k^2 - \tilde{\varepsilon}_k^2) = 0$, we derive

$$|I\!\!Ef(S) - I\!\!Ef(\tilde{S})| = \left| I\!\!E \sum_{k=1}^n \{f(S_k + \varepsilon_k) - f(S_k + \tilde{\varepsilon}_k)\} \right| \leq \frac{\|f^{(3)}\|_\infty}{6} \delta_n \quad (\text{G.1})$$

with

$$\delta_n \stackrel{\text{def}}{=} \sum_k I\!\!E(|\varepsilon_k|^3 + |\tilde{\varepsilon}_k|^3) = \sum_k I\!\!E(|\varepsilon_k|^3 + \tilde{C}_3 v_k^3), \quad (\text{G.2})$$

where $\tilde{C}_3 = I\!\!E|\gamma_i|^3 = 2\sqrt{2/\pi} < 2$ for a standard normal γ . One typical example of such a function f approximating the indicator function $\mathbb{I}(t \geq q)$ for a given q is given by

$$f(t) = \frac{e^{\beta(t-q)}}{1 + e^{\beta(t-q)}}. \quad (\text{G.3})$$

The β -coefficient controls the quality of approximation.

Lemma G.1.1. *The function f from (G.3) fulfills for any β and q the condition $\|f^{(3)}\|_\infty \leq \beta^3/8$. Moreover, with any $\Delta > 0$, it holds*

$$f(t - \Delta) - e^{-\beta\Delta} \leq \mathbb{I}(t \geq q) \leq f(t + \Delta) + e^{-\beta\Delta}. \quad (\text{G.4})$$

Proof. With $u(t) = 1 + e^{\beta(t-q)}$, it holds $f(t) = 1 - 1/u(t)$. Moreover,

$$\begin{aligned} -\left(\frac{1}{u(t)}\right)' &= \frac{u'(t)}{u^2(t)}, \\ -\left(\frac{1}{u(t)}\right)'' &= \frac{u''(t)}{u^2(t)} - \frac{2|u'(t)|^2}{u^3(t)}, \\ -\left(\frac{1}{u(t)}\right)''' &= \frac{u'''(t)}{u^2(t)} - \frac{6u'(t)u''(t)}{u^3(t)} + \frac{6[u'(t)]^3}{u^4(t)}. \end{aligned} \quad (\text{G.5})$$

The use of $u^{(j)}(t) = \beta^j\{u(t) - 1\}$ yields $f^{(3)}(t) = g(u(t) - 1)$ with

$$g(z) = \frac{\beta^3 z}{(1+z)^2} - \frac{6\beta^3 z^2}{(1+z)^3} + \frac{6\beta^3 z^3}{(1+z)^4} = \frac{\beta^3(z - 4z^2 + z^3)}{(1+z)^4}.$$

One can see that

$$\|g\|_\infty \stackrel{\text{def}}{=} \sup_{z \geq 0} |g(z)| = \beta^3/8.$$

Indeed, $g^{(3)}(\cdot)$ has only three extreme points 1 and $5 \pm \sqrt{24}$, and the bound is obtained by the direct check at these points.

The bounds in (G.4) can be checked by direct calculus.

This and (??) yield

$$\begin{aligned} I\!\!P(S \geq q + \Delta) &\leq e^{-\beta\Delta} + I\!\!E f(S) \leq e^{-\beta\Delta} + I\!\!E f(\tilde{S}) + \frac{1}{48}\beta^3\delta_n \\ &\leq I\!\!P(\tilde{S} \geq q - \Delta) + 2e^{-\beta\Delta} + \frac{1}{48}\beta^3\delta_n. \end{aligned} \quad (\text{G.6})$$

Similarly

$$\begin{aligned} \mathbb{P}(S \geq q - \Delta) &\geq -e^{-\beta\Delta} + \mathbb{E}f(S) \geq -e^{-\beta\Delta} + \mathbb{E}f(\tilde{S}) - \frac{1}{48}\beta^3\delta_n \\ &\geq \mathbb{P}(\tilde{S} \geq q + \Delta) - 2e^{-\beta\Delta} - \frac{1}{48}\beta^3\delta_n. \end{aligned} \quad (\text{G.7})$$

Theorem G.1.1. Let ε_i be independent zero mean with $v_i^2 = \text{Var}(\varepsilon_i)$ and finite third moments. Suppose that $V^2 = \sum_i v_i^2 = 1$. For any $\Delta > 0$ and $\beta > 0$, the sum $S = \sum \varepsilon_i$ satisfies for all q

$$\begin{aligned} \mathbb{P}(S \leq q) &\leq \Phi(q - 2\Delta) + 2e^{-\beta\Delta} + \frac{\beta^3}{48}\delta_n, \\ \mathbb{P}(S \leq q) &\geq \Phi(q + 2\Delta) - 2e^{-\beta\Delta} - \frac{\beta^3}{48}\delta_n, \end{aligned} \quad (\text{G.8})$$

where δ_n is given by (G.2). Moreover, if for some $C_3 \geq 0$ and $\alpha_n \geq 0$

$$\mathbb{E}|\varepsilon_i|^3 \leq C_3 v_i^3, \quad \max_i v_i \leq \alpha_n V, \quad i = 1, \dots, n,$$

then $\delta_n \leq (C_3 + \tilde{C}_3)\alpha_n$.

With $\beta = \Delta^{-1} \log(n)$, if follows

$$|\mathbb{P}(S \leq q) - \Phi(q)| \leq 0.8\Delta + 2/n + \frac{1}{48} \log^3(n)\Delta^{-3}\delta_n.$$

Proof. The statements (G.8) have been already proved. If $\mathbb{E}|\varepsilon_i|^3 \leq C_3 v_i^3$ and $\max_i v_i \leq \alpha_n V$, then

$$\mathbb{E} \sum_k |\varepsilon_k|^3 \leq C_3 \sum_k v_i^3 \leq C_3 V^2 \max_i v_i \leq C_3 V^3 \alpha_n = C_3 \alpha_n,$$

and

$$\delta_n \leq (C_3 + \tilde{C}_3)\alpha_n V^3 = (C_3 + \tilde{C}_3)\alpha_n.$$

The last assertion follows from the bound

$$|\Phi(q) - \Phi(q + 2\Delta)| \leq 2\Delta(2\pi)^{-1/2} \leq 0.8\Delta.$$

G.2 Berry-Esseen Theorem for a univariate sum

Now we consider the Lyapunov approach based on the approximation of a characteristic function of a sum $S = \varepsilon_1 + \dots + \varepsilon_n$.

G.2.1 Characteristic functions for a univariate sum

Define for $S = \varepsilon_1 + \dots + \varepsilon_n$

$$f(t) \stackrel{\text{def}}{=} f_S(t) = \mathbb{E}e^{itS},$$

where $i = \sqrt{-1}$.

Theorem G.2.1. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent zero mean with $\text{Var}(\varepsilon_i) = v_i^2$ and $\mathbb{E}|\varepsilon_i|^3 < \infty$ such that

$$\sum_{i=1}^n v_i^2 = 1.$$

Define

$$\delta_n \stackrel{\text{def}}{=} \sum_{i=1}^n \mathbb{E}|\varepsilon_i|^3.$$

Then $f(t) = \mathbb{E}e^{itS}$ fulfills

$$\sup_{|t| \leq 1/(4\delta_n)} |f(t) - e^{-t^2/2}| \leq 4\delta_n |t|^3 e^{-t^2/3}. \quad (\text{G.9})$$

Proof. Independence of the ε_i 's implies

$$f(t) = \prod_{i=1}^n \mathbb{E}e^{it\varepsilon_i} = \prod_{i=1}^n f_i(t) \quad (\text{G.10})$$

where each function $f_i(t) = \mathbb{E}e^{it\varepsilon_i}$ is smooth in t and satisfies

$$f_i^{(m)}(t) = i^m \mathbb{E}(\varepsilon_i^m e^{it\varepsilon_i}), \quad m = 0, 1, 2, 3.$$

In particular, $f_i(0) = 1$, $f'_i(0) = i \mathbb{E}\varepsilon_i = 0$, $f''_i(0) = -\mathbb{E}\varepsilon_i^2 = -v_i^2$, and

$$|f'''_i(t)| = |\mathbb{E}(\varepsilon_i^3 e^{it\varepsilon_i})| \leq \mathbb{E}|\varepsilon_i|^3.$$

This implies

$$|f_i(t) - 1 + \frac{t^2}{2}v_i^2| \leq \frac{|t|^3}{6} \mathbb{E}|\varepsilon_i|^3. \quad (\text{G.11})$$

Let us consider the following two cases separately:

$$\left\{ |t| \leq 1/(4\delta_n), |t| \leq 1/(2\delta_n)^{1/3} \right\} \quad \text{and} \quad \left\{ |t| \leq 1/(4\delta_n), |t| > 1/(2\delta_n)^{1/3} \right\}.$$

The assertion will be proven if we show that the bound (G.9) holds for the each case. For the first case we have

$$|t| v_i \leq |t| \{I\!\!E |\varepsilon_i|^3\}^{1/3} \leq |t| \delta_n^{1/3} \leq 2^{-1/3},$$

$$|t|^3 I\!\!E |\varepsilon_i|^3 \leq 1/2,$$

and $\alpha_i \stackrel{\text{def}}{=} f_i(t) - 1$ satisfies

$$|\alpha_i| \leq \frac{t^2 v_i^2}{2} + \frac{|t|^3 I\!\!E |\varepsilon_i|^3}{6} \leq 0, 4.$$

Also,

$$\alpha_i^2 \leq 2 \left(\frac{t^2 v_i^2}{2} \right)^2 + 2 \left(\frac{|t|^3 I\!\!E |\varepsilon_i|^3}{6} \right)^2 \leq 0, 43 |t|^3 I\!\!E |\varepsilon_i|^3.$$

As $|\log(1 - \alpha_i) + \alpha_i| \leq 0, 772 \alpha_i^2$ for $|\alpha_i| \leq 1/2$, it follows by (G.11)

$$|\log f_i(t) + \frac{v_i^2 t^2}{2}| \leq |\log(1 - \alpha_i) + \alpha_i| + \frac{|t|^3}{6} I\!\!E |\varepsilon_i|^3 \leq \frac{|t|^3}{2} I\!\!E |\varepsilon_i|^3$$

yielding by (G.10)

$$|\log f(t) + \frac{t^2}{2}| \leq \sum_{i=1}^n \frac{|t|^3}{2} I\!\!E |\varepsilon_i|^3 \leq \frac{|t|^3 \delta_n}{2}.$$

Now we use $e^{|t|^3 \delta_n / 2} < 2$ for $|t| \leq 1/(2\delta_n)^{1/3}$ to bound

$$\begin{aligned} |f(t) - e^{-t^2/2}| &\leq e^{-t^2/2} |\exp\{\log f(t) + t^2/2\} - 1| \\ &\leq e^{-t^2/2} |\exp\{|t|^3 \delta_n / 2\} - 1| \leq \frac{1}{2} e^{-t^2/2} |t|^3 \delta_n e^{|t|^3 \delta_n / 2} \\ &\leq e^{-t^2/2} |t|^3 \delta_n \leq 4 e^{-t^2/3} |t|^3 \delta_n. \end{aligned}$$

For the second case, $|t| > 1/(2\delta_n)^{1/3}$ and $|t| \leq 1/(4\delta_n)$, we use the symmetrization device: the value $|f_i(t)|^2$ is the characteristic function of the sum $\varepsilon_i + \tilde{\varepsilon}_i$, where $\tilde{\varepsilon}_i$ is an independent copy of ε_i . As $\text{Var}(\varepsilon_i + \tilde{\varepsilon}_i) = 2v_i^2$, $I\!\!E |\varepsilon_i + \tilde{\varepsilon}_i|^3 \leq 8 I\!\!E |\varepsilon_i|^3$, it follows similarly to the above

$$|f_i(t)|^2 \leq 1 - t^2 v_i^2 + \frac{4}{3} |t|^3 I\!\!E |\varepsilon_i|^3 \leq \exp\left(-t^2 v_i^2 + \frac{4}{3} |t|^3 I\!\!E |\varepsilon_i|^3\right).$$

For $|t|\delta_n \leq 1/4$, this implies

$$|f(t)|^2 \leq \prod_{i=1}^n |f_i(t)|^2 \leq \exp\left(-t^2 + \frac{4}{3} |t|^3 \delta_n\right) \leq e^{-2t^2/3}$$

and hence by $2|t|^3 \delta_n \geq 1$

$$|f(t) - e^{-t^2/2}| \leq |f(t)| + e^{-t^2/2} \leq 2e^{-t^2/3} \leq 4|t|^3 \delta_n e^{-t^2/3}$$

and the result follows.

G.2.2 Characteristic function of a sum. Simmerization

Let again $S = \varepsilon_1 + \dots + \varepsilon_n$ with independent zero mean r.v.'s ε_i with $\sigma_i^2 = \mathbb{E}\varepsilon_i^2$.

Denote

$$V^2 = \sum_{i=1}^n \mathbb{E}\varepsilon_i^2 = \sum_{i=1}^n \sigma_i^2. \quad (\text{G.12})$$

For any t , consider

$$f(t) \stackrel{\text{def}}{=} \frac{1}{2} \mathbb{E}\{e^{itS} + e^{-itS}\}.$$

We aim at bounding the error of approximation of the symmetrized characteristic function $f(t)$ by the cf $\tilde{f}(t) = e^{-t^2 V^2 / 2}$ of the gaussian r.v. with mean zero and variance V^2 .

Theorem G.2.2. *Let ε_i be independent zero mean with $\mathbb{E}\varepsilon_i^4 < \infty$. Let t be such that*

$$\max_{i=1,\dots,n} \left| \frac{t^2}{2} \mathbb{E}\varepsilon_i^2 + \frac{|t|^3}{6} |\mathbb{E}\varepsilon_i^3| + \frac{t^4}{24} \mathbb{E}\varepsilon_i^4 \right| \leq \frac{1}{2}.$$

It holds with V^2 from (G.12)

$$\left| f(t) \exp\left(\frac{t^2 V^2}{2}\right) - 1 \right| \leq (2\Delta_4 + \frac{1}{2}\kappa^2 e^\kappa) e^{\Delta_2},$$

where

$$\Delta_2 \stackrel{\text{def}}{=} \frac{t^4}{4} \sum_{i=1}^n (\mathbb{E}\varepsilon_i^2)^2 = \frac{t^4}{4} \sum_{i=1}^n \sigma_i^4,$$

$$\Delta_3 \stackrel{\text{def}}{=} \frac{|t|^3}{6} \sum_{i=1}^n |\mathbb{E}\varepsilon_i^3|,$$

$$\Delta_4 \stackrel{\text{def}}{=} \frac{t^4}{24} \sum_{i=1}^n \mathbb{E}\varepsilon_i^4,$$

and

$$\kappa \stackrel{\text{def}}{=} 2\Delta_3 + 2\Delta_4.$$

Proof. The proof mainly uses independence of the ε_i 's and the Taylor expansion for each characteristic function $\mathbb{E}e^{it\varepsilon_i}$.

Lemma G.2.1. *It holds*

$$\mathbb{E}e^{itS} = \exp\left\{ \sum_{i=1}^n \log\left(1 - \frac{t^2 \sigma_i^2}{2} - \frac{it^3 \mathbb{E}\varepsilon_i^3}{6} + \frac{t^4 \delta_i(t)}{24}\right) \right\}$$

with

$$|\delta_i(t)| \leq \mathbb{E}\varepsilon_i^4.$$

Proof. Independence of the ε_i 's yields

$$\mathbb{E}e^{itS} = \exp\left\{\sum_{i=1}^n \log(\mathbb{E}e^{it\varepsilon_i})\right\}$$

Now the statement follows by the fourth order Taylor expansion of $\mathbb{E}e^{it\varepsilon_i}$ in view of

$$\left| \frac{d^4}{dt^4} \mathbb{E}e^{it\varepsilon_i} \right| = \left| \mathbb{E}\varepsilon_i^4 e^{it\varepsilon_i} \right| \leq \mathbb{E}\varepsilon_i^4.$$

Lemma G.2.2. *For any sequences (α_i) , (\varkappa_i^\pm) with $|\varkappa_i^\pm| \leq \varkappa_i$ and $|\alpha_i| + \varkappa_i \leq 1/2$, it holds*

$$\begin{aligned} & \left| \frac{1}{2} \exp \sum_{i=1}^n \{\log(1 - \alpha_i + \varkappa_i^+) + \alpha_i\} + \frac{1}{2} \exp \sum_{i=1}^n \{\log(1 - \alpha_i - \varkappa_i^-) + \alpha_i\} - 1 \right| \\ & \leq \left(R + \frac{1}{2} K^2 e^K \right) e^A \end{aligned}$$

with

$$K \stackrel{\text{def}}{=} 2 \sum_{i=1}^n \varkappa_i, \quad R \stackrel{\text{def}}{=} \sum_{i=1}^n |\varkappa_i^+ - \varkappa_i^-|, \quad A \stackrel{\text{def}}{=} \sum_{i=1}^n \alpha_i^2.$$

Proof. First consider the case with $\varkappa_i^\pm \equiv 0$. The inequality $|x - \log(1 + x)| \leq x^2$ for $|x| \leq 1$ implies

$$\exp \sum_{i=1}^n \{\log(1 - \alpha_i) + \alpha_i\} \leq \exp \left(\sum_{i=1}^n \alpha_i^2 \right) = e^A. \quad (\text{G.13})$$

Define for $s \in [-1, 1]$ the function

$$g(s) \stackrel{\text{def}}{=} \exp \sum_{i=1}^n \{\log(1 - \alpha_i + s\varkappa_i^+) + \alpha_i\}.$$

Then (G.13) implies $g(0) \leq e^A$. Furthermore,

$$g'(s) = g(s) \sum_{i=1}^n \frac{\varkappa_i^+}{1 - \alpha_i + s\varkappa_i^+}$$

and

$$|g'(s)| \leq K g(s)$$

yielding $g(s) \leq g(0) \exp(K|s|)$ for $|s| \leq 1$. Also

$$g''(s) = \left\{ \left(\sum_{i=1}^n \frac{\varkappa_i^+}{1 - \alpha_i + s\varkappa_i^+} \right)^2 - \sum_{i=1}^n \frac{|\varkappa_i^+|^2}{(1 - \alpha_i + s\varkappa_i^+)^2} \right\} g(s),$$

ensuring $|g''(s)| \leq K^2 g(s)$ for $s \in [-1, 1]$. Similarly we proceed with the function

$$g_-(s) \stackrel{\text{def}}{=} \exp \sum_{i=1}^n \{\log(1 - \alpha_i + s\kappa_i^-) + \alpha_i\}.$$

In particular, $g_-(0) = g(0)$ and

$$|g'(0) + g'_-(0)| \leq 2g(0) \sum_{i=1}^n |\kappa_i^+ - \kappa_i^-| = 2g(0)R.$$

Now it holds by the Taylor expansion of the second order

$$\begin{aligned} \left| \frac{1}{2} \{g(1) + g(-1)\} - g(0) \right| &\leq \frac{1}{2} |g'(0) + g'_-(0)| + \frac{1}{2} \sup_{s \in [-1, 1]} |g''(s)| \\ &\leq g(0)R + \frac{1}{2} K^2 \sup_{s \in [-1, 1]} g(s) \leq g(0) \left(R + \frac{1}{2} K^2 e^K \right) \end{aligned}$$

and the result follows in view of (G.13).

Now we apply this lemma with $\alpha_i = t^2 \sigma_i^2 / 2$

$$\begin{aligned} \kappa_i^+ &\stackrel{\text{def}}{=} \frac{t^3}{6} I\!E \varepsilon_i^3 + \frac{t^4}{24} \delta_i(t), \quad \kappa_i^- \stackrel{\text{def}}{=} \frac{t^3}{6} I\!E \varepsilon_i^3 + \frac{t^4}{24} \delta_i(-t), \\ \kappa_i &\stackrel{\text{def}}{=} \frac{|t|^3}{6} |I\!E \varepsilon_i^3| + \frac{t^4}{24} I\!E \varepsilon_i^4 \end{aligned}$$

yielding

$$\begin{aligned} K &= 2 \sum_{i=1}^n \kappa_i \leq \frac{|t|^3}{3} \sum_{i=1}^n |I\!E \varepsilon_i^3| + \frac{t^4}{12} \sum_{i=1}^n I\!E \varepsilon_i^4 = 2\Delta_3 + 2\Delta_4, \\ R &= \sum_{i=1}^n |\kappa_i^+ - \kappa_i^-| \leq \frac{t^4}{12} \sum_{i=1}^n I\!E \varepsilon_i^4 = 2\Delta_4, \\ A &= \frac{t^4}{4} \sum_{i=1}^n \sigma_i^4 = \Delta_2. \end{aligned}$$

Now we discuss the implications of the result of Theorem G.2.2 to the important i.i.d. case.

Theorem G.2.3. Suppose that ε_i are i.i.d. with $\sigma^2 = I\!E \varepsilon_i^2 = 1/n$. Let also $|I\!E \varepsilon_i^4| \leq \kappa^4 n^{-2}$ for some $\kappa \geq 1$. Then it holds for all t with $t\kappa \leq 4/5\sqrt{n}$:

$$\left| f(t) - \exp\left(-\frac{t^2}{2}\right) \right| \leq \frac{t^4 \kappa^4 (1 + t^2 \kappa^2)}{12n} \exp\left(-\frac{t^2}{2} + \frac{2|t|^3 \kappa^3}{5\sqrt{n}} + \frac{t^4}{4n}\right)$$

Proof. We use that in the i.i.d. case for any t

$$\Delta_2 = \frac{t^4}{4n}, \quad \Delta_3 \leq \frac{|t|^3 \kappa^3}{6n}, \quad \Delta_4 \leq \frac{t^4 \kappa^4}{24n}$$

and for $|t|\kappa \leq 4/5\sqrt{n}$

$$\kappa \leq \frac{|t|^3 \kappa^3}{3\sqrt{n}} + \frac{t^4 \kappa^4}{12n} \leq \frac{2|t|^3 \kappa^3}{5\sqrt{n}}, \quad \kappa^2 \leq \frac{t^6 \kappa^6}{6n},$$

and thus

$$\begin{aligned} (2\Delta_4 + \frac{1}{2}\kappa^2 e^\kappa) e^{\Delta_2} &\leq \left\{ \frac{t^4 \kappa^4}{12n} + \frac{t^6 \kappa^6}{12n} \exp\left(\frac{2|t|^3 \kappa^3}{5\sqrt{n}}\right) \right\} \exp\left(\frac{t^4}{4n} - \frac{t^2}{2}\right) \\ &\leq \frac{t^4 \kappa^4 (1 + t^2 \kappa^2)}{12n} \exp\left(\frac{2|t|^3 \kappa^3}{5\sqrt{n}} + \frac{t^4}{4n} - \frac{t^2}{2}\right) \end{aligned}$$

G.2.3 Methods based on the Fourier-Stieltjes transform

Let F and G be distribution functions. Denote by f and g their Fourier transforms:

$$f(t) = \int_{-\infty}^{\infty} e^{itx} dF(x), \quad g(t) = \int_{-\infty}^{\infty} e^{itx} dG(x).$$

We aim at bounding $\Delta \stackrel{\text{def}}{=} \sup_x |F(x) - G(x)|$ via $|f(t) - g(t)|$.

Theorem G.2.4. *For each $b > 1/(2\pi)$, it holds*

$$\sup_x |F(x) - G(x)| \leq b \int_{-1/\delta}^{1/\delta} \left| \frac{f(t) - g(t)}{t} \right| dt + \frac{b}{\delta} \sup_x \int_{|u| \leq c(b)\delta} |G(x+u) - G(x)| du,$$

where $c(b)$ is the root of the equation

$$\int_0^{c(b)/4} \frac{\sin^2(u)}{u^2} du = \frac{\pi}{4} + \frac{1}{8b}. \quad (\text{G.14})$$

Proof. Fix $\delta > 0$ and $a > 0$ and consider the function

$$\eta_\delta(t) \stackrel{\text{def}}{=} \begin{cases} (1 - |t|\delta)e^{ita\delta}, & |t|\delta \leq 1 \\ 0, & |t|\delta > 1. \end{cases} \quad (\text{G.15})$$

It is straightforward to see that $\eta_\delta(t)$ is the Fourier transform of the distribution with the density $p_\delta(x)$ given by

$$p_\delta(x) \stackrel{\text{def}}{=} \frac{1}{\pi\delta} \frac{1 - \cos(x/\delta - a)}{(x/\delta - a)^2}.$$

Clearly

$$p_\delta(x) = \frac{1}{2\pi\delta} \left(\frac{\sin((x/\delta - a)/2)}{(x/\delta - a)/2} \right)^2 \leq \frac{1}{2\pi\delta}.$$

Now define

$$F_\delta(x) \stackrel{\text{def}}{=} F * p_\delta(x) = \int_{-\infty}^{\infty} F(x-z) p_\delta(z) dz = \int_{-\infty}^{\infty} F(u) p_\delta(x-u) du$$

so that its characteristic function is a product of f and η_δ :

$$f_\delta(t) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} e^{itx} dF_\delta(x) = f(t) \eta_\delta(t).$$

Using the fact that $\eta_\delta(t) = 0$ for $|t| > 1/\delta$, we obtain by the inversion formula

$$F_\delta(x) - F_\delta(y) = \frac{1}{2\pi} \int_{-1/\delta}^{1/\delta} \frac{e^{-itx} - e^{-ity}}{-it} f(t) \eta_\delta(t) dt.$$

The same holds for the function G :

$$G_\delta(x) - G_\delta(y) = \frac{1}{2\pi} \int_{-1/\delta}^{1/\delta} \frac{e^{-itx} - e^{-ity}}{-it} g(t) \eta_\delta(t) dt.$$

Now take the difference of these two expression and let $y \rightarrow -\infty$. As $F_\delta(-\infty) = G_\delta(-\infty) = 0$, it follows

$$F_\delta(x) - G_\delta(x) = \frac{1}{2\pi} \int_{-1/\delta}^{1/\delta} \frac{f(t) - g(t)}{-it} e^{-itx} \eta_\delta(t) dt.$$

The definition implies $|\eta_\delta(t)| \leq 1$, and we can bound

$$\left| \int_{-\infty}^{\infty} \{F(u) - G(u)\} p_\delta(x-u) du \right| \leq \frac{1}{2\pi} \int_{-1/\delta}^{1/\delta} \left| \frac{f(t) - g(t)}{t} \right| dt.$$

Also with $\Delta = \sup_x |F(x) - G(x)|$

$$\begin{aligned} & \left| \int_{x-2a\delta}^x \{F(u) - G(u)\} p_\delta(x-u) du \right| \\ & \leq \left| \int_{-\infty}^{\infty} \{F(u) - G(u)\} p_\delta(x-u) du \right| \\ & \quad + \Delta \int_{-\infty}^{x-2a\delta} p_\delta(x-u) du + \Delta \int_x^{\infty} p_\delta(x-u) du \\ & \leq \frac{1}{2\pi} \int_{-1/\delta}^{1/\delta} \left| \frac{f(t) - g(t)}{t} \right| dt + \Delta \left(1 - \int_{x-2a\delta}^x p_\delta(u) du \right). \end{aligned}$$

Define $\gamma = \gamma_\delta(a)$ by

$$\gamma \stackrel{\text{def}}{=} \int_{x-2a\delta}^x p_\delta(x-u) du = \int_0^{2a\delta} p_\delta(u) du = \frac{2}{\pi} \int_0^{a/2} \frac{\sin^2(u)}{u^2}.$$

As $F(x)$ does not decrease and $p_\delta(u) \leq 1/(2\pi\delta)$

$$\begin{aligned}
F(x - 2a\delta) &\leq \frac{1}{\gamma} \int_{x-2a\delta}^x F(u) p_\delta(x-u) du \\
&= G(x - 2a\delta) + \frac{1}{\gamma} \int_{x-2a\delta}^x \{G(u) - G(x - 2a\delta)\} p_\delta(x-u) du \\
&\quad + \frac{1}{\gamma} \int_{x-2a\delta}^x \{F(u) - G(u)\} p_\delta(x-u) du \\
&\leq G(x - 2a\delta) + \frac{1}{2\pi\gamma\delta} \int_{x-2a\delta}^x |G(u) - G(x - 2a\delta)| du \\
&\quad + \frac{1}{2\pi\gamma} \int_{-1/\delta}^{1/\delta} \left| \frac{f(t) - g(t)}{t} \right| dt + \frac{1-\gamma}{\gamma} \Delta.
\end{aligned}$$

Let $y \stackrel{\text{def}}{=} x - 2a\delta$, so we can rewrite the inequality as

$$\begin{aligned}
F(y) - G(y) &\leq \frac{1}{2\pi\gamma\delta} \int_0^{2a\delta} |G(x+u) - G(x)| du \\
&\quad + \frac{1}{2\pi\gamma} \int_{-1/\delta}^{1/\delta} \left| \frac{f(t) - g(t)}{t} \right| dt + \frac{1-\gamma}{\gamma} \Delta.
\end{aligned}$$

The use of negative a in the definition (G.15) of p_δ helps to bound from below

$$\begin{aligned}
F(x) - G(x) &\geq -\frac{1}{2\pi\gamma\delta} \int_{-2a\delta}^0 |G(x+u) - G(x)| du \\
&\quad - \frac{1}{2\pi\gamma} \int_{-1/\delta}^{1/\delta} \left| \frac{f(t) - g(t)}{t} \right| dt - \frac{1-\gamma}{\gamma} \Delta
\end{aligned}$$

for any x . These two bounds yield

$$\Delta \leq \frac{1}{2\pi\gamma\delta} \int_{-2a\delta}^{2a\delta} |G(x+u) - G(x)| du + \frac{1}{2\pi\gamma} \int_{-1/\delta}^{1/\delta} \left| \frac{f(t) - g(t)}{t} \right| dt + \frac{1-\gamma}{\gamma} \Delta.$$

It remains to select a proper value a . We use that $\gamma_\delta(a) \rightarrow 1$ as $a \rightarrow \infty$. For each b , one can fix a by the equation $2\gamma_\delta(a) - 1 = 1/(2\pi b)$. Then for $b > 1/(2\pi)$

$$\begin{aligned}
\Delta &\leq \frac{1}{2\pi(2\gamma-1)\delta} \int_{-2a\delta}^{2a\delta} |G(x+u) - G(x)| du + \frac{1}{2\pi(2\gamma-1)} \int_{-1/\delta}^{1/\delta} \left| \frac{f(t) - g(t)}{t} \right| dt \\
&\leq \frac{b}{\delta} \sup_x \int_{|u| \leq c(b)\delta} |G(x+u) - G(x)| du + b \int_{-1/\delta}^{1/\delta} \left| \frac{f(t) - g(t)}{t} \right| dt
\end{aligned}$$

with $c(b)$ from (G.14), and the assertion follows.

G.2.4 Berry-Esseen Theorem

Now we aim at evaluating the distance between the distribution of the normalized sum S by the standard Gaussian law by putting together the result of Theorem G.2.1 on the characteristic function of S and the result of Theorem G.2.4.

Theorem G.2.5. *Let $\varepsilon_1, \dots, \varepsilon_n$ be independent zero mean with $\text{Var}(\varepsilon_i) = v_i^2$ and $\mathbb{E}|\varepsilon_i|^3 < \infty$. Consider the sum $S = \varepsilon_1 + \dots + \varepsilon_n$ and define*

$$v^2 \stackrel{\text{def}}{=} \text{Var}(S) = \sum_{i=1}^n v_i^2.$$

Then with δ_n

$$\delta_n \stackrel{\text{def}}{=} v^{-3} \sum_{i=1}^n \mathbb{E}|\varepsilon_i|^3,$$

it holds for $F(x) \stackrel{\text{def}}{=} \mathbb{P}(S/v \leq x)$ and the standard normal law $\Phi(x)$

$$\sup_x |F(x) - \Phi(x)| \leq C\delta_n.$$

Proof. Without loss of generality suppose $v = 1$. The general case is obtained by standardization of S . We apply Theorem G.2.4 with $G(x) = \Phi(x)$. The use of $b = 1/\pi$ and $\delta = 4\delta_n$ implies in view of $\Phi'(x) \leq 1/(2\pi)$

$$\begin{aligned} & \frac{b}{\delta} \sup_x \int_{|u| \leq c(b)\delta} |G(x+u) - G(x)| du \\ & \leq \frac{2b}{\delta} \int_0^{c(b)\delta} \frac{u}{2\pi} du = \frac{c^2(b)\delta}{2\pi^2} = \frac{2c^2(b)\delta_n}{\pi^2} = C_1\delta_n \end{aligned}$$

and thus

$$\sup_x |F(x) - \Phi(x)| \leq \frac{1}{\pi} \int_{|t| \leq 1/(4\delta_n)} \left| \frac{f(t) - e^{-t^2/2}}{t} \right| dt + C_1\delta_n.$$

Now by Theorem G.2.1

$$\sup_x |F(x) - \Phi(x)| \leq C_1\delta_n + \frac{\delta_n}{\pi} \int_{|t| \leq 1/(4\delta_n)} 4|t|^2 e^{-t^2/3} dt \leq C_1\delta_n + C_2\delta_n.$$

and the result follows.

G.2.5 Fourier transform for the norm of a vector

Let S be a vector sum, $S = \sum_{i=1}^n \varepsilon_i \in \mathbb{R}^p$, $p \geq 2$, with zero mean independent vectors ε_i . This section is about an approximation of the characteristic function of the norm $\|S\|$

by a similar characteristic function of a Gaussian vector $\tilde{\mathbf{S}}$ with the same covariance. Denote

$$f(t) \stackrel{\text{def}}{=} I\!\!E \exp(it\|\mathbf{S}\|).$$

The basic idea is to look at the integrals of the form

$$g(t) \stackrel{\text{def}}{=} \frac{1}{|\mathcal{S}_p|} \int_{\mathcal{S}_p} I\!\!E \exp\{it\boldsymbol{\gamma}^\top \mathbf{S}\} d\boldsymbol{\gamma}$$

where \mathcal{S}_p is the unit sphere in \mathbb{R}^p and $|\mathcal{S}_p|$ is its area:

$$|\mathcal{S}_p| \stackrel{\text{def}}{=} \int_{\mathcal{S}_p} d\boldsymbol{\gamma}.$$

It is straightforward to see that the integral $\int_{\mathcal{S}_p} I\!\!E \exp\{it\boldsymbol{\gamma}^\top \mathbf{S}\} d\boldsymbol{\gamma}$ only depends on $\|\mathbf{S}\|$. Indeed, by a rotation of \mathcal{S}_p one can reduce the situation to the case of a vector $\mathbf{S} = (\|\mathbf{S}\|, 0, \dots, 0)^\top$. At the same time, for a standard Gaussian vector $\tilde{\mathbf{S}}$, the scalar product $\boldsymbol{\gamma}^\top \tilde{\mathbf{S}}$ is standard normal, and hence, it holds

$$\tilde{g}(t) = \frac{1}{|\mathcal{S}_p|} \int_{\mathcal{S}_p} I\!\!E \exp\{it\boldsymbol{\gamma}^\top \tilde{\mathbf{S}}\} d\boldsymbol{\gamma} = e^{-t^2/2}.$$

The key step of the analysis is the following identity.

Lemma G.2.3. *It holds for any t and $p \geq 2$*

$$g(t) = \frac{|\mathcal{S}_{p-1}|}{|\mathcal{S}_p|} \int_{-1}^1 f(\rho t) (1 - \rho^2)^{\frac{p-3}{2}} d\rho. \quad (\text{G.16})$$

Moreover, for any function $h(t)$

$$\int_{-\infty}^{\infty} g(t) h(t) dt = \int_{-\infty}^{\infty} f(u) \eta(u) du \quad (\text{G.17})$$

with

$$\eta(u) \stackrel{\text{def}}{=} \frac{|\mathcal{S}_{p-1}|}{|\mathcal{S}_p|} \int_{-1}^1 (1 - \rho^2)^{\frac{p-3}{2}} \rho^{-1} h(\rho^{-1}u) d\rho.$$

Proof. W.l.g. consider a vector \mathbf{S} of the form $\mathbf{S} = (\|\mathbf{S}\|, 0, \dots, 0)^\top$. For any $\rho \in [-1, 1]$, consider all unit vectors $\boldsymbol{\gamma}$ with $\gamma_1 = \rho$. Equivalently, the angle ϕ between $\boldsymbol{\gamma}$ and the vector $\mathbf{e}_1 = (1, 0, \dots, 0)^\top$ fulfills $\cos(\phi) = \rho$. All such vectors belong to a $p-1$ dimensional sphere with radius $\sqrt{1 - \rho^2}$ whose surface is $|\mathcal{S}_{p-1}|(1 - \rho^2)^{(p-1)/2}$. The result (G.16) is obtained by reparametrization $\rho = \cos(\phi)$, where ϕ is the angle between 0 and 2π . Further, by change of integration

$$\begin{aligned}
\int_{-\infty}^{\infty} h(t) g(t) dt &= \frac{|\mathcal{S}_{p-1}|}{|\mathcal{S}_p|} \int_{\infty}^{\infty} \int_{-1}^1 f(\rho t) (1 - \rho^2)^{\frac{p-3}{2}} d\rho h(t) dt \\
&= \frac{|\mathcal{S}_{p-1}|}{|\mathcal{S}_p|} \int_{\infty}^{\infty} f(u) \int_{-1}^1 (1 - \rho^2)^{\frac{p-3}{2}} \rho^{-1} h(\rho^{-1} u) d\rho du \\
&= \int_{\infty}^{\infty} f(u) \eta(u) du
\end{aligned}$$

and (G.17) follows.

An immediate corollary of the lemma is the following integral relation between two characteristic functions $f(t)$ and $\tilde{f}(t)$: for any function $h(t)$

$$\begin{aligned}
\int_{-\infty}^{\infty} h(t) \{g(t) - \tilde{g}(t)\} dt &= \frac{|\mathcal{S}_{p-1}|}{|\mathcal{S}_p|} \int_{\infty}^{\infty} \int_{-1}^1 \{f(\rho t) - \tilde{f}(\rho t)\} (1 - \rho^2)^{\frac{p-3}{2}} d\rho h(t) dt \\
&= \frac{|\mathcal{S}_{p-1}|}{|\mathcal{S}_p|} \int_{\infty}^{\infty} \{f(u) - \tilde{f}(u)\} \int_{-1}^1 (1 - \rho^2)^{\frac{p-3}{2}} \rho^{-1} h(\rho^{-1} u) d\rho du \\
&= \int_{\infty}^{\infty} \{f(u) - \tilde{f}(u)\} \eta(u) du. \tag{G.18}
\end{aligned}$$

Introduce a linear operator \mathcal{H} on the space of functions $h(\cdot)$ by $(\mathcal{H}h)(u) = \eta(u)$. Note that $h(u) = u^a$ for $a < 0$ yields

$$\eta(u) = u^a \frac{|\mathcal{S}_{p-1}|}{|\mathcal{S}_p|} \int_{-1}^1 (1 - \rho^2)^{\frac{p-3}{2}} \rho^{-1-a} d\rho$$

that is, u^a is a eigenfunction of the operator \mathcal{H} . Moreover, with $h_m(u) = u^{-1} \log^m(u)$ and $\eta_m = \mathcal{H}h_m$, one can derive

$$\begin{aligned}
\eta_m(u) &= u^{-1} \frac{|\mathcal{S}_{p-1}|}{|\mathcal{S}_p|} \int_{-1}^1 (1 - \rho^2)^{\frac{p-3}{2}} \{\log(u) - \log(\rho)\}^m d\rho \\
&= \sum_{j=0}^m c_{j,p} h_j(u)
\end{aligned}$$

where

$$c_{j,p} \stackrel{\text{def}}{=} \binom{m}{m-j} \frac{|\mathcal{S}_{p-1}|}{|\mathcal{S}_p|} \int_{-1}^1 (1 - \rho^2)^{\frac{p-3}{2}} \{-\log(\rho)\}^{m-j} d\rho.$$

Now consider $\eta(t; x) = \frac{1}{2\pi} e^{-itx}$ and define $h(\cdot; x) = \mathcal{H}\eta(\cdot; x)$. Then by (G.18)

G.2.6 Fourier transform for the squared norm of a vector

In this section we consider the squared norm $\|\mathbf{S}\|^2$ of a vector sum $\mathbf{S} = \boldsymbol{\varepsilon}_1 + \dots + \boldsymbol{\varepsilon}_n$. The aim is to approximate the characteristic function of $\|\mathbf{S}\|^2$ by a similar expression for a Gaussian sum with the same covariance structure.

Suppose that ε_i are independent zero mean vectors in \mathbb{R}^p and denote

$$v_i^2 \stackrel{\text{def}}{=} \text{Var}(\varepsilon_i).$$

Let the sum \mathbf{S} be already standardized, that is,

$$\text{Var}(\mathbf{S}) = \mathbf{I}_p.$$

For any fixed $\gamma \in \mathbb{R}^p$, the result of Theorem G.2.1 helps to approximate $\mathbb{E}e^{i\gamma^\top \mathbf{S}}$ by $e^{-\|\gamma\|^2/2}$. The question under study is whether we can approximate the Fourier transform of $\|\mathbf{S}\|^2$ by the Fourier of a χ_p^2 random variable.

We use for $t > 0$ the identity

$$\frac{1}{(2\pi t)^{p/2}} \int_{\mathbb{R}^p} \exp\left\{-\frac{1}{2t}\|\gamma - \mathbf{u}\|^2\right\} d\gamma = 1$$

for any fixed complex vector \mathbf{u} . (Define the function $f(\mathbf{u}) = \int_{\mathbb{R}^p} \exp\left\{-\frac{1}{2t}(\gamma - \mathbf{u})^\top(\gamma - \mathbf{u})\right\} d\gamma$. It is analytic in \mathbf{u} and constant on the subspace of real vectors \mathbf{u} , thus it is constant everywhere.) In particular, with $\mathbf{u} = \mathbf{h}t\mathbf{S}$ for $\mathbf{h} = i^{1/2} = (1+i)/\sqrt{2}$, we obtain the following representation

$$\begin{aligned} f(t) &\stackrel{\text{def}}{=} e^{it\|\mathbf{S}\|^2/2} = \frac{1}{(2\pi t)^{p/2}} \int_{\mathbb{R}^p} \exp\left\{\mathbf{h}\gamma^\top \mathbf{S} - \frac{\|\gamma\|^2}{2t}\right\} d\gamma \\ &= \frac{1}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} \exp\left\{\mathbf{h}a\gamma^\top \mathbf{S} - \frac{\|\gamma\|^2}{2}\right\} d\gamma \end{aligned}$$

for $a = t^{1/2}$. Similarly define $g(t) = \mathbb{E}e^{it\|\tilde{\mathbf{S}}\|^2/2}$. Our first step is to bound the difference $f(t) - g(t)$.

Lemma G.2.4. *It holds*

Proof. Step 1 (large deviation): for $z(p, \mathbf{x}) = p^{1/2} + (2\mathbf{x})^{1/2}$

$$\mathbb{P}(\|\mathbf{S}\| > z(p, \mathbf{x})) \leq 2e^{-\mathbf{x}}$$

Step 2 (integral truncation): if $\|\mathbf{S}\| \leq z(p, \mathbf{x})$

$$\frac{1}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} \exp\left\{\mathbf{h}a\gamma^\top \mathbf{S} - \frac{\|\gamma\|^2}{2}\right\} \mathbb{I}(\|\gamma\| > cz(p, \mathbf{x})) d\gamma \leq e^{-\mathbf{x}}.$$

Step 3 (Pointwise approximation): Theorem G.2.1 does not apply to $\mathbf{h}a\gamma^\top \mathbf{S}$ because \mathbf{h} has a real part. However, the result and the proof extend to this situation.

G.3 GAR for the Euclidean norm of a vector sum

Let ξ_i be independent zero mean vectors in \mathbb{R}^p for $i = 1, \dots, n$ with $\text{Var}(\xi) = v_i^2$ and finite third moments. We are interested in a Gaussian approximation for the norm $\|\mathbf{S}\|$ of the sum $\mathbf{S} = \sum_i \xi_i$. More precisely, let $\tilde{\mathbf{S}} \sim \mathcal{N}(0, \text{Var}(\mathbf{S}))$ be a Gaussian vector with the same first and second moments as \mathbf{S} . We aim at comparing the quantiles of $\|\mathbf{S}\|$ and $\|\tilde{\mathbf{S}}\|$.

Theorem G.3.1. *Let ξ_i be independent zero mean vectors in \mathbb{R}^p , and let $\tilde{\xi}_i$ be Gaussian zero mean vectors with $\text{Var}(\tilde{\xi}_i) = \text{Var}(\xi_i)$ for all i . Then the sums $\mathbf{S} = \sum_i \xi_i$ and $\tilde{\mathbf{S}} = \sum_i \tilde{\xi}_i$ satisfy for any $\Delta \geq 0$ and $q \geq 1$*

$$\begin{aligned} \mathbb{P}(\|\mathbf{S}\| \geq q) &\leq \mathbb{P}(\|\tilde{\mathbf{S}}\| \geq q - \Delta) + C(\Delta/q)\Delta^{-3}\delta_n, \\ \mathbb{P}(\|\mathbf{S}\| \geq q) &\geq \mathbb{P}(\|\tilde{\mathbf{S}}\| \geq q + \Delta) - C(\Delta/q)\Delta^{-3}\delta_n, \end{aligned} \quad (\text{G.19})$$

where δ_n and $C(a)$ are given by

$$\delta_n \stackrel{\text{def}}{=} \sum_i (\mathbb{E}\|\xi_i\|^3 + \mathbb{E}\|\tilde{\xi}_i\|^3), \quad (\text{G.20})$$

$$C(a) \stackrel{\text{def}}{=} \frac{16}{3}(1+2a)^{3/2} + 4a(1+2a)^{1/2}. \quad (\text{G.21})$$

In particular, for $a \leq 1/3$, it holds $C(a) \leq 13.2$.

Proof. The basic idea of the proof is to approximate the discontinuous function $\mathbb{I}(\|\mathbf{x}\| \geq q)$ by a smooth function $f(\mathbf{x})$ and then to apply the Lindeberg telescopic sum device. The desired bound will be derived from the next statement.

Let $g(\cdot)$ be a function from the class \mathcal{G}_3 of univariate functions which are three times continuously differentiable with $g(u) = 0$ for $u \leq 0$, $g(u) = 1$ for $u \geq 1$, and $g(u)$ monotonously grows from zero to one when u grows from zero to one. A particular choice is given by a third order spline function g with $g'''(u) = a$ for $u \in [0, 1/4] \cup [3/4, 1]$ and $g'''(u) = -a$ for $u \in [1/4, 3/4]$ for $a = 32$. This yields

$$g(u) = 16 \begin{cases} u^3/3, & u \in [0, 1/4], \\ 1/32 + (u - 1/2)/8 - (u - 1/2)^3/3, & u \in [1/4, 3/4], \\ 1/16 + (u - 1)^3/3, & u \in [3/4, 1], \end{cases}$$

Obviously $|g'(u)| \leq 2$, $|g''(u)| \leq 8$, $|g'''(u)| \leq 32$ for all u . Define for some fixed $\Delta > 0$ and $q > 0$

$$f_\Delta(\mathbf{x}, q) = g\left(\frac{1}{2q\Delta}(\|\mathbf{x}\|^2 - q^2)\right). \quad (\text{G.22})$$

Proposition G.3.1. Let ξ_i be independent zero mean vectors in \mathbb{R}^p , $\mathbf{S} = \sum_i \xi_i$, and let $\tilde{\mathbf{S}}$ be a Gaussian zero mean vector with $\text{Var}(\tilde{\mathbf{S}}) = \text{Var}(\mathbf{S})$. For any $\Delta > 0$ and $q \geq 1$, it holds with the function $f_\Delta(\mathbf{x}, q)$ from (G.22)

$$|\mathbb{E}f_\Delta(\mathbf{S}, q) - \mathbb{E}f_\Delta(\tilde{\mathbf{S}}, q)| \leq \mathbf{C}(\Delta/q)\Delta^{-3}\delta_n. \quad (\text{G.23})$$

Proof. It is straightforward to see that for any $\mathbf{d} \in \mathbb{R}^p$ with $\|\mathbf{d}\| = 1$, the function $\psi(t) = g\left(\frac{\beta}{2}(\|\mathbf{x}\|^2 - q^2)\right)$ fulfills

$$\begin{aligned} \psi'(t) &= \beta \mathbf{d}^\top (\mathbf{x} + t\mathbf{d}) g'\left(\frac{\beta}{2}(\|\mathbf{x} + t\mathbf{d}\|^2 - q^2)\right), \\ \psi''(t) &= \beta^2 |\mathbf{d}^\top (\mathbf{x} + t\mathbf{d})|^2 g''\left(\frac{\beta}{2}(\|\mathbf{x} + t\mathbf{d}\|^2 - q^2)\right) + \beta g'\left(\frac{\beta}{2}(\|\mathbf{x} + t\mathbf{d}\|^2 - q^2)\right), \\ \psi'''(t) &= \beta^3 [\mathbf{d}^\top (\mathbf{x} + t\mathbf{d})]^3 g'''\left(\frac{\beta}{2}(\|\mathbf{x} + t\mathbf{d}\|^2 - q^2)\right) \\ &\quad + 3\beta^2 \mathbf{d}^\top (\mathbf{x} + t\mathbf{d}) g''\left(\frac{\beta}{2}(\|\mathbf{x} + t\mathbf{d}\|^2 - q^2)\right). \end{aligned}$$

Now we use that $g''(u)$ and $g'''(u)$ vanish if $u < 0$ or $u > 1$. The inequality $\beta(\|\mathbf{x} + t\mathbf{d}\|^2 - q^2) \leq 2$ implies in view of $\|\mathbf{d}\| = 1$ that

$$|\mathbf{d}^\top (\mathbf{x} + t\mathbf{d})| \leq |\mathbf{x} + t\mathbf{d}| \leq (2/\beta + q^2)^{1/2}.$$

Therefore,

$$\begin{aligned} |\psi'(t)| &\leq \beta(2/\beta + q^2)^{1/2} \|g'\|_\infty, \\ |\psi''(t)| &\leq \beta^2(2/\beta + q^2) \|g''\|_\infty + \beta \|g'\|_\infty, \\ |\psi'''(t)| &\leq \beta^3(2/\beta + q^2)^{3/2} \|g'''\|_\infty + 3\beta^2(2/\beta + q^2)^{1/2} \|g''\|_\infty. \end{aligned}$$

The choice $\beta = 1/(q\Delta)$ yields

$$\beta^2(2/\beta + q^2) = \Delta^{-2} + 2\Delta^{-1}q^{-1} = \Delta^{-2}(1 + 2\Delta q^{-1}).$$

For $q \geq 1$ and Δ small, it implies $|\psi'''(t)| \leq 6\mathbf{C}(\Delta/q)\Delta^{-3}$ with $\mathbf{C}(a)$ given by

$$\begin{aligned} 6\mathbf{C}(a) &\stackrel{\text{def}}{=} (1 + 2a)^{3/2} \|g'''\|_\infty + 3a(1 + 2a)^{1/2} \|g''\|_\infty \\ &= 32(1 + 2a)^{3/2} + 24a(1 + 2a)^{1/2}; \end{aligned}$$

cf. (G.21). This implies that the function $f(\mathbf{x}) = f_\Delta(\mathbf{x}, q)$ fulfills for any points $\mathbf{x}, \mathbf{d} \in \mathbb{R}^p$ the condition

$$|f(\mathbf{x} + \mathbf{d}) - f(\mathbf{x}) - \mathbf{d}^\top f'(\mathbf{x}) - \mathbf{d}^\top f''(\mathbf{x})\mathbf{d}/2| \leq \mathbf{C}(\Delta/q)\Delta^{-3}\|\mathbf{d}\|^3;$$

Now the Lindeberg telescopic sum device yields (G.23) similarly to (G.1).

Now we are able to complete the proof of the theorem. We use the bounds

$$\mathbb{I}(\|\mathbf{x}\| \geq q + \Delta) \leq \mathbb{I}(\|\mathbf{x}\|^2 \geq q^2 + 2\Delta q) \leq f_\Delta(\mathbf{x}, q) \leq \mathbb{I}(\|\mathbf{x}\| \geq q).$$

Therefore,

$$\begin{aligned} \mathbb{P}(\|\mathbf{S}\| \geq q) &\leq \mathbb{E}f_\Delta(\mathbf{S}, q - \Delta) \\ &\leq \mathbb{E}f_\Delta(\tilde{\mathbf{S}}, q - \Delta) + C(\Delta/q)\Delta^{-3}\delta_n \\ &\leq \mathbb{P}(\|\tilde{\mathbf{S}}\| \geq q - \Delta) + C(\Delta/q)\Delta^{-3}\delta_n \end{aligned}$$

which easily implies (G.19).

Now we slightly extend the result of Theorem G.3.1 to the case when the quantile level q is slightly misspecified.

Theorem G.3.2. Let ξ_i be independent zero mean random vectors in \mathbb{R}^p with finite third moments, $v_i^2 = \text{Var}(\xi_i)$ for $i = 1, \dots, n$, and $\Sigma = \sum_i v_i^2$. Let also $\tilde{\mathbf{S}}$ be a Gaussian vector in \mathbb{R}^p with zero mean and the variance $\text{Var}(\tilde{\mathbf{S}}) = \Sigma$. Let also values q and \tilde{q} fulfill for some $\diamond \geq 0$

$$|q - \tilde{q}| \leq \diamond.$$

Then the sum $\mathbf{S} = \sum_{i=1}^n \xi_i$ satisfies

$$\begin{aligned} |\mathbb{E}f_\Delta(\mathbf{S}, q) - \mathbb{E}f_\Delta(\tilde{\mathbf{S}}, \tilde{q})| &\leq C(\Delta/q)\Delta^{-3}\delta_n + \frac{\diamond}{q}\sqrt{2p}, \\ |\mathbb{P}(\|\mathbf{S}\| \geq q) - \mathbb{P}(\|\tilde{\mathbf{S}}\| \geq \tilde{q})| &\leq C(\Delta/q)\Delta^{-3}\delta_n + \frac{\Delta + \diamond}{q}\sqrt{p/2} \end{aligned} \tag{G.24}$$

with δ_n from (G.20). Furthermore, for $\Delta/q \leq 1/3$ and $q \geq \sqrt{p}$, it holds

$$\begin{aligned} |\mathbb{E}f_\Delta(\mathbf{S}, q) - \mathbb{E}f_\Delta(\tilde{\mathbf{S}}, \tilde{q})| &\leq 13.2\Delta^{-3}\delta_n + \diamond\sqrt{2}, \\ |\mathbb{P}(\|\mathbf{S}\| \geq q) - \mathbb{P}(\|\tilde{\mathbf{S}}\| \geq \tilde{q})| &\leq 2.6\delta_n^{1/4} + \diamond/\sqrt{2}. \end{aligned} \tag{G.25}$$

Proof. Let $\tilde{\mathbf{S}} = \sum_i \tilde{\xi}_i$ as in Theorem G.3.1. By Proposition G.3.1, it holds

$$|\mathbb{E}f_\Delta(\mathbf{S}, q) - \mathbb{E}f_\Delta(\tilde{\mathbf{S}}, q)| \leq C(\Delta/q)\Delta^{-3}\delta_n.$$

Further, the definition $f_\Delta(\mathbf{x}, q) = g\left(\frac{1}{2q\Delta}(\|\mathbf{x}\|^2 - q^2)\right)$ yields for $a = \tilde{q}/q < 1$ and any \mathbf{x} that

$$f_\Delta(\mathbf{x}, \tilde{q}) \leq f_\Delta(\mathbf{x}, q) \leq f_\Delta(a\mathbf{x}, \tilde{q}).$$

Therefore,

$$f_\Delta(\mathbf{x}, q) - f_\Delta(\mathbf{x}, \tilde{q}) \leq f_\Delta(a\mathbf{x}, \tilde{q}) - f_\Delta(\mathbf{x}, \tilde{q}).$$

Lemma D.2.1 yields

$$\mathbb{E}|f_\Delta(a\mathbf{x}, \tilde{q}) - f_\Delta(\mathbf{x}, \tilde{q})| \leq (1-a)\sqrt{p} \leq \frac{q-\tilde{q}}{q}\sqrt{2p}.$$

Now we combine the Gaussian approximation of $\|\mathbf{S}\|$ by $\|\tilde{\mathbf{S}}\|$ with an anticoncentration inequality of Theorem D.2.1:

$$\mathbb{P}(\|\tilde{\mathbf{S}}\| \geq q - \Delta) - \mathbb{P}(\|\tilde{\mathbf{S}}\| \geq \tilde{q}) \leq (\Delta + \diamondsuit)q^{-1}\sqrt{p/2}.$$

Putting altogether yields (G.24). The last result of the theorem is obtained by optimizing the value Δ . Given $A, B > 0$, a sum $A\Delta^{-3} + B\Delta$ is minimized at $\Delta^4 = 3A/B$ and with this choice of Δ , it holds

$$A\Delta^{-3} + B\Delta = \frac{4}{3}(3A)^{1/4}B^{3/4}.$$

Now, under $a = \Delta/q \leq 1/3$, one can bound with $A = C(a)\delta_n$ and $B = q^{-1}\sqrt{p/2}$ $q \geq \sqrt{p}$ and $a = \Delta/q \leq 1/3$ with $\Delta = (39.6\sqrt{2}\delta_n)^{1/4}$

$$C(a)\Delta^{-3}\delta_n + \Delta q^{-1}\sqrt{p/2} \leq 2.6\delta_n^{1/4}.$$

This yields (G.25).

G.4 GAR for the sup-norm of a vector sum

The results of this section will be improved in the next one!

Here we consider GAR for general vector sums. Let Ξ be a finite set and for each $\eta \in \Xi$ with M elements, a sequence $\xi_1(\eta), \dots, \xi_n(\eta)$ of random variables with zero mean and finite third moments is supposed to be fixed. Denote $\xi_i = \{\xi_i(\eta), \eta \in \Xi\} \in \mathbb{R}^M$.

We consider the Gaussian approximation of the vector sums $\mathbf{S} = \sum_i \xi_i$. Denote by v_i^2 the $M \times M$ covariance matrix of the vector ξ_i , while V^2 is the covariance of the sum $\mathbf{S} = \sum_i \xi_i$:

$$v_i^2 \stackrel{\text{def}}{=} \mathbb{E}(\xi_i \xi_i^\top), \quad V^2 \stackrel{\text{def}}{=} \mathbb{E}(\mathbf{S} \mathbf{S}^\top) = \sum_i v_i^2.$$

Consider another sequence of mutually independent random vectors $\tilde{\xi}_i$ with the same covariance structure: $\tilde{\xi}_i \sim \mathcal{N}(0, v_i^2)$. We aim at approximating the distribution of the sum

\mathbf{S} by a normal distribution $\mathcal{N}(0, V^2)$ which is the distribution of the sum $\tilde{\mathbf{S}} = \sum_i \tilde{\boldsymbol{\xi}}_i$. Let f be a function on \mathbb{R}^M . The main step in the construction is a bound for the difference $\mathbb{E}f(\mathbf{S}) - \mathbb{E}f(\tilde{\mathbf{S}})$. Suppose that the function f is smooth and satisfies the condition

$$|f(\mathbf{x} + \mathbf{d}) - f(\mathbf{x}) - \mathbf{d}^\top f'(\mathbf{x}) - \mathbf{d}^\top f''(\mathbf{x})\mathbf{d}/2| \leq C_f \|\mathbf{d}\|_\infty^3 \quad (\text{G.26})$$

for any points $\mathbf{x}, \mathbf{d} \in \mathbb{R}^M$. Introduce for $k = 1, \dots, n$ the telescopic sums in \mathbb{R}^M

$$\mathbf{S}^{(i)} \stackrel{\text{def}}{=} \boldsymbol{\xi}_1 + \dots + \boldsymbol{\xi}_{i-1} + \tilde{\boldsymbol{\xi}}_{i+1} + \dots + \tilde{\boldsymbol{\xi}}_n.$$

It holds by (G.26)

$$\begin{aligned} & |f(\mathbf{S}^{(i)} + \boldsymbol{\xi}_i) - f(\mathbf{S}^{(i)} + \tilde{\boldsymbol{\xi}}_i) - f'(\mathbf{S}^{(i)})^\top (\boldsymbol{\xi}_i - \tilde{\boldsymbol{\xi}}_i) - \boldsymbol{\xi}_i^\top f''(\mathbf{S}^{(i)})\boldsymbol{\xi}_i/2 + \tilde{\boldsymbol{\xi}}_i^\top f''(\mathbf{S}^{(i)})\tilde{\boldsymbol{\xi}}_i/2| \\ & \leq C_f (\|\boldsymbol{\xi}_i\|_\infty^3 + \|\tilde{\boldsymbol{\xi}}_i\|_\infty^3). \end{aligned}$$

Similarly to (??) this implies

$$|\mathbb{E}f(\mathbf{S}) - \mathbb{E}f(\tilde{\mathbf{S}})| \leq C_f \delta_n \quad (\text{G.27})$$

with

$$\delta_n \stackrel{\text{def}}{=} \sum_i (\mathbb{E}\|\boldsymbol{\xi}_i\|_\infty^3 + \mathbb{E}\|\tilde{\boldsymbol{\xi}}_i\|_\infty^3).$$

One can use that

$$\delta_n \leq C \sum_i \left\{ \log(M) \sigma_i^2 \right\}^{3/2}.$$

Now we build a proper function f which approximates the indicator of the set where the maximum of x_η exceeds zero. It can be viewed as superposition of the sigmoid $e^{\beta x}/(1 + e^{\beta x})$ used in the univariate case, and of the soft-maximum function $\beta^{-1} \log(\sum_\eta e^{\beta x_\eta})$.

Lemma G.4.1. *For each $\beta > 0$ and $B \geq 1$, the function*

$$f(\mathbf{x}) = f_{\beta, B}(\mathbf{x}) \stackrel{\text{def}}{=} \frac{\sum_\eta e^{\beta x_\eta}}{B + \sum_\eta e^{\beta x_\eta}}$$

satisfies (G.26) with $C_f = \beta^3/6$.

Proof. Fix \mathbf{x} and \mathbf{d} in \mathbb{R}^M with $\|\mathbf{d}\|_\infty = 1$ and consider a univariate function $g(t) = f(\mathbf{x} + t\mathbf{d})$. Now with $w_\eta(t) = B^{-1} \exp\{\beta[x_\eta + t d_\eta]\}$, define

$$u(t) = 1 + \sum_{\eta \in \Xi} w_\eta(t) = \sum_{\eta \in \Xi^+} w_\eta(t),$$

where $\Xi^+ \stackrel{\text{def}}{=} \Xi \cup \{0\}$ and $w_0(t) = 1$. For each t , introduce a measure π on the discrete set Ξ^+ with probabilities $\pi_{\boldsymbol{\eta}} = w_{\boldsymbol{\eta}}(t)/u(t)$. Let also U be a π -random variable with $U(\boldsymbol{\eta}) = d_{\boldsymbol{\eta}}$ so that $\mathbb{I}P_{\pi}(U = d_{\boldsymbol{\eta}}) = \pi_{\boldsymbol{\eta}}$ where $d_0 = 0$. Then

$$u^{(j)}(t) = \beta^j \sum_{\boldsymbol{\eta}} d_{\boldsymbol{\eta}}^j w_{\boldsymbol{\eta}}(t) = \beta^j u(t) \sum_{\boldsymbol{\eta}} d_{\boldsymbol{\eta}}^j \pi_{\boldsymbol{\eta}} = \beta^j u(t) \mathbb{E}_{\pi} U^j, \quad j = 0, 1, 2, 3,$$

which implies by (G.5) for $\psi(t) = \{u(t) - 1\}/u(t)$

$$\begin{aligned} \psi'(t) &= \frac{\beta}{u(t)} \mathbb{E}_{\pi} U, \\ \psi''(t) &= \frac{\beta^2}{u(t)} \{ \mathbb{E}_{\pi} U^2 - 2(\mathbb{E}_{\pi} U)^2 \}, \\ \psi'''(t) &= \frac{\beta^3}{u(t)} \{ \mathbb{E}_{\pi} U^3 - 6\mathbb{E}_{\pi} U^2 \mathbb{E}_{\pi} U + 6(\mathbb{E}_{\pi} U)^3 \} = \frac{\beta^3}{u(t)} \{ \mathbb{E}_{\pi} U^3 - 6\mathbb{E}_{\pi} U \text{Var}_{\pi}(U) \}. \end{aligned}$$

It is straightforward to see that

$$\|\psi'\| \leq \beta, \quad \|\psi''\| \leq \beta^2.$$

Moreover, the maximum of $\mathbb{E}_{\pi} U^3 - 6\mathbb{E}_{\pi} U \text{Var}_{\pi}(U)$ over the class of all random variables U bounded in absolute value by 1, is achieved on the singleton $U \equiv 1$ yielding $|\psi'''(t)| \leq \|d\|_{\infty}^3$ in view of $u(t) \geq 1$. This implies (G.26) with $C_f = \beta^3/6$.

The next step is in bounding the distance between distributions of the maximum of $X(\boldsymbol{\eta})$ by a similar construction based on $\tilde{X}(\boldsymbol{\eta})$. Let \mathbf{x} be a vector in \mathbb{R}^M with entries $x(\boldsymbol{\eta})$, and Δ be the vector with the entries Δ . We use the inequality

$$f_{\beta,B}(\mathbf{x} - \Delta) - B^{-1}\mathbb{M}e^{-\beta\Delta} \leq \mathbb{1}\left(\max_{\boldsymbol{\eta}} x(\boldsymbol{\eta}) \geq 0\right) \leq f_{\beta,B}(\mathbf{x} + \Delta) + Be^{-\beta\Delta},$$

which extends (G.4) to the case of multiple comparison. Indeed, $f_{\beta,B}$ fulfills $f_{\beta,B}(\mathbf{x}) \geq 0$ and moreover, if $x(\boldsymbol{\eta}) \geq 0$ for some $\boldsymbol{\eta}$ then for any $\Delta > 0$

$$\begin{aligned} f_{\beta,B}(\mathbf{x} + \Delta) &\geq 1 - (1 + B^{-1}e^{\beta[x(\boldsymbol{\eta}) + \Delta]})^{-1} + Be^{-\beta\Delta} \\ &\geq 1 - (1 + B^{-1}e^{\beta\Delta})^{-1} + Be^{-\beta\Delta} \geq 1. \end{aligned}$$

Similarly, $f_{\beta,B}$ fulfills $f_{\beta,B}(\mathbf{x}) \leq 1$ and moreover, if $x(\boldsymbol{\eta}) \leq 0$ for all $\boldsymbol{\eta} \in \Xi$, then

$$f_{\beta,B}(\mathbf{x} - \Delta) \leq B^{-1} \sum_{\boldsymbol{\eta}} e^{\beta[x(\boldsymbol{\eta}) - \Delta]} - B^{-1}\mathbb{M}e^{-\beta\Delta} \leq 0.$$

This suggests to apply $B = \mathbb{M}^{1/2}$ with $B = B^{-1}\mathbb{M} = \mathbb{M}^{1/2}$.

Now we apply these bounds with $x(\boldsymbol{\eta}) = S(\boldsymbol{\eta}) - q(\boldsymbol{\eta})$ for given critical values $q(\boldsymbol{\eta})$. The arguments similar to (G.6) and (G.7) imply

$$\begin{aligned}
I\!\!P\left(\max_{\boldsymbol{\eta}}\{S(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 0\right) &\leq I\!\!E f_{\beta,B}(\mathbf{S} - \mathbf{q} + \boldsymbol{\Delta}) + Be^{-\beta\Delta} \\
&\leq I\!\!E f_{\beta,B}(\tilde{\mathbf{S}} - \mathbf{q} + \boldsymbol{\Delta}) + Be^{-\beta\Delta} + \beta^3\delta_n \\
&\leq I\!\!P\left(\max_{\boldsymbol{\eta}}\tilde{S}(\boldsymbol{\eta}) - q(\boldsymbol{\eta}) \geq -2\Delta\right) + 2Be^{-\beta\Delta} + \beta^3\delta_n, \\
I\!\!P\left(\max_{\boldsymbol{\eta}}\{S(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 0\right) &\geq I\!\!E f_{\beta,B}(\mathbf{S} - \mathbf{q} - \boldsymbol{\Delta}) - Be^{-\beta\Delta} \\
&\geq I\!\!E f_{\beta,B}(\tilde{\mathbf{S}} - \mathbf{q} - \boldsymbol{\Delta}) - Be^{-\beta\Delta} - \beta^3\delta_n \\
&\geq I\!\!P\left(\max_{\boldsymbol{\eta}}\tilde{S}(\boldsymbol{\eta}) - q(\boldsymbol{\eta}) \geq 2\Delta\right) - 2Be^{-\beta\Delta} - \beta^3\delta_n.
\end{aligned}$$

Summarizing the above considerations yields the following bound.

Theorem G.4.1. *Let $\xi_i = \{\xi_i(\boldsymbol{\eta}), \boldsymbol{\eta} \in \Xi\}$ be independent zero mean random vectors in \mathbb{R}^M with finite third moments and $V_i^2 = \text{Var}(\xi_i)$ for $i = 1, \dots, n$, and $\mathbf{S} = \sum_i \xi_i$. Let also $\tilde{\mathbf{S}} = \{\tilde{S}(\boldsymbol{\eta}), \boldsymbol{\eta} \in \Xi\}$ be a zero mean Gaussian vector in \mathbb{R}^M with the same covariance structure as \mathbf{S} , that is, $\text{Var}(\mathbf{S}) = \text{Var}(\tilde{\mathbf{S}})$. Then (G.27) holds for each function $f(\cdot)$ satisfying (G.26). Moreover, for any $\beta > 0$ and $\Delta > 0$, and any collection of critical values $q(\boldsymbol{\eta})$, it holds*

$$\begin{aligned}
I\!\!P\left(\max_{\boldsymbol{\eta}}\{S(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 0\right) &\leq I\!\!P\left(\max_{\boldsymbol{\eta}}\{\tilde{S}(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq -2\Delta\right) + 2M^{1/2}e^{-\beta\Delta} + \beta^3\delta_n, \\
I\!\!P\left(\max_{\boldsymbol{\eta}}\{S(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 0\right) &\geq I\!\!P\left(\max_{\boldsymbol{\eta}}\{\tilde{S}(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 2\Delta\right) - 2M^{1/2}e^{-\beta\Delta} - \beta^3\delta_n.
\end{aligned} \tag{G.28}$$

G.5 GAR for the sup-norm of a vector sum. Improved

Now we consider a slightly different construction of the function f used in vector GAR. It is built as a superposition of the smooth maximum function h_β and a function g_Δ approximating the indicator function.

The next lemma describes some useful properties of the smooth maximum function $h_\beta(\mathbf{x})$ defined as

$$h_\beta(\mathbf{x}) = \beta^{-1} \log\left(\sum_{\boldsymbol{\eta}} e^{\beta x_{\boldsymbol{\eta}}}\right).$$

Lemma G.5.1. *For any $\mathbf{x}, \mathbf{d} \in \mathbb{R}^M$ and each t , the function $\psi(t) \stackrel{\text{def}}{=} h_\beta(\mathbf{x} + t\mathbf{d})$ fulfills*

$$|\psi'(t)| \leq \|\mathbf{d}\|_\infty, \quad |\psi''(t)| \leq \beta \|\mathbf{d}\|_\infty^2, \quad |\psi'''(t)| \leq \beta^2 \|\mathbf{d}\|_\infty^3. \tag{G.29}$$

Proof. Given $\mathbf{x} = (x_{\eta})$, and $\mathbf{d} = (d_{\eta})$, define $u(t) = \sum_{\eta} e^{\beta(x_{\eta} + t d_{\eta})}$ and

$$\psi(t) = h_{\beta}(\mathbf{x} + t\mathbf{d}) = \beta^{-1} \log u(t).$$

Now with $\pi_{\eta}(t) = u(t)^{-1} \exp\{\beta(x_{\eta} + t d_{\eta})\}$, introduce a measure π on the discrete set Ξ with probabilities $\pi_{\eta}(t)$. Let also U be a π -random variable with $U(\eta) = d_{\eta}$ so that $I\!\!E_{\pi}(U = d_{\eta}) = \pi_{\eta}(t)$. Then, for $j = 1, 2, 3$

$$u^{(j)}(t) = \beta^j \sum_{\eta} d_{\eta}^j \pi_{\eta}(t) = \beta^j u(t) \sum_{\eta} d_{\eta}^j \pi_{\eta}(t) = \beta^j u(t) I\!\!E_{\pi} U^j,$$

which implies

$$\begin{aligned} |\psi'(t)| &= \beta^{-1} \left| \frac{u'(t)}{u(t)} \right| \leq I\!\!E_{\pi} U, \\ |\psi''(t)| &= \beta^{-1} \frac{|u''(t)u(t) - u'(t)^2|}{U^2(t)} \leq \beta \text{Var}_{\pi}(U), \\ |\psi'''(t)| &= \beta^{-1} \frac{|u'''(t) - 3u''(t)S'(t) + 2u'(t)^3|}{U^3(t)} \\ &\leq \beta^2 |I\!\!E_{\pi} U^3 - 3I\!\!E_{\pi} U^2 I\!\!E_{\pi} U + 2(I\!\!E_{\pi} U)^3|. \end{aligned}$$

One can check that for any \mathbf{d} , it holds $I\!\!E_{\pi} U \leq \|\mathbf{d}\|_{\infty}$, $\text{Var}_{\pi} U \leq \|\mathbf{d}\|_{\infty}^2$, and $|I\!\!E_{\pi} U^3 - 3I\!\!E_{\pi} U^2 I\!\!E_{\pi} U + 2(I\!\!E_{\pi} U)^3| \leq \|\mathbf{d}\|_{\infty}^3$, and (G.29) follows.

Let g be a three times continuously differentiable with $g(u) = 0$ for $u \leq 0$, $g(u) = 1$ for $u \geq 1$, and $g(u)$ monotonously grows from zero to one when u grows from zero to one. A particular choice is given by a third order spline function g with $g'''(u) = a$ for $u \in [0, 1/4] \cup [3/4, 1]$ and $g'''(u) = -a$ for $u \in [1/4, 3/4]$ for $a = 32$. This yields

$$g(u) = 16 \begin{cases} u^3/3, & u \in [0, 1/4], \\ 1/32 + (u - 1/2)/8 - (u - 1/2)^3/3, & u \in [1/4, 3/4], \\ 1/16 + (u - 1)^3/3, & u \in [3/4, 1], \end{cases}$$

and $|g'(u)| \leq 2$, $|g''(u)| \leq 8$, $|g'''(u)| \leq 32$ for all u . Given $\Delta > 0$, define also

$$g_{\Delta}(u) = g(\Delta^{-1}u).$$

Lemma G.5.2. *The function $f(\mathbf{x}) = g_{\Delta} \circ h_{\beta}(\mathbf{x})$ satisfies*

$$|f(\mathbf{x} + \mathbf{d}) - f(\mathbf{x}) - \mathbf{d}^T f'(\mathbf{x}) - \mathbf{d}^T f''(\mathbf{x}) \mathbf{d}/2| \leq C(\Delta, \beta) \|\mathbf{d}\|_{\infty}^3$$

for any points $\mathbf{x}, \mathbf{d} \in I\!\!R^M$ with

$$C(\Delta, \beta) \stackrel{\text{def}}{=} (\Delta^{-3} \|g'''\|_{\infty} + \Delta^{-2} \beta \|g''\|_{\infty} + \Delta^{-1} \beta^2 \|g'\|_{\infty})/6. \quad (\text{G.30})$$

Proof. Consider a univariate function

$$\phi(t) = f(\mathbf{x} + t\mathbf{d}) = g_\Delta \circ h_\beta(\mathbf{x} + t\mathbf{d}) = g(\psi(t)/\Delta).$$

It holds

$$\begin{aligned}\phi'(t) &= \Delta^{-1} g'(\psi(t)/\Delta) \psi'(t), \\ \phi''(t) &= \Delta^{-2} g''(\psi(t)/\Delta) \psi'(t)^2 + \Delta^{-1} g'(\psi(t)/\Delta) \psi''(t), \\ \phi'''(t) &= \Delta^{-3} g'''(\psi(t)/\Delta) \psi'(t)^3 + 3\Delta^{-2} g''(\psi(t)/\Delta) \psi(t) \psi''(t) + \Delta^{-1} g'(\psi(t)/\Delta) \psi'''(t).\end{aligned}$$

In view of (G.29), we bound

$$\begin{aligned}\|\phi'''\|_\infty &\leq \Delta^{-3} \|g'''\|_\infty \|\psi'\|_\infty^3 + 3\Delta^{-2} \|g''\|_\infty \|\psi'\|_\infty \|\psi''\|_\infty + \Delta^{-1} \|g'\|_\infty \|\psi'''\|_\infty \\ &\leq \Delta^{-3} \|g'''\|_\infty + \Delta^{-2} \|g''\|_\infty + \Delta^{-1} \beta^2 \|g'\|_\infty,\end{aligned}$$

and the arguments from the proof of Lemma G.4.1 imply the condition (G.26) with $C_f = C(\Delta, \beta)$ from (G.30).

Now we approximate the indicator of the event $\{\max_{\boldsymbol{\eta}} x_{\boldsymbol{\eta}} \geq 0\}$ by a smooth function $g_\Delta \circ h_\beta(\mathbf{x})$. For this, we have to specify the relation between Δ and β . Namely, we fix $\beta = \Delta \log(M)$ yielding

$$C(\Delta, \beta) = \frac{1}{6} (\|g'''\|_\infty + \log(M) \|g''\|_\infty + \log^2(M) \|g'\|_\infty) \Delta^{-3} = C(M) \Delta^{-3} \quad (\text{G.31})$$

for $C(M) \stackrel{\text{def}}{=} \frac{1}{6} (\|g'''\|_\infty + \log(M) \|g''\|_\infty + \log^2(M) \|g'\|_\infty)$.

Lemma G.5.3. *With $\Delta = \beta^{-1} \log(M)$, it holds*

$$g_\Delta \circ h_\beta(\mathbf{x} - \boldsymbol{\Delta}) \leq \mathbb{I}(\max_{\boldsymbol{\eta}} x_{\boldsymbol{\eta}} > 0) \leq g_\Delta \circ h_\beta(\mathbf{x} + \boldsymbol{\Delta}).$$

Proof. By construction, $g_\Delta(u) \in [0, 1]$ for any u . If $x_{\boldsymbol{\eta}} \geq 0$ for some $\boldsymbol{\eta}$, then $h_\beta(\mathbf{x} + \boldsymbol{\Delta}) \geq \Delta$ and hence,

$$g_\Delta \circ h_\beta(\mathbf{x} + \boldsymbol{\Delta}) \geq g(\Delta/\Delta) = g(1) = 1.$$

Similarly, if $\max_{\boldsymbol{\eta}} x_{\boldsymbol{\eta}} \leq 0$, then due to $\Delta = \beta^{-1} \log(M)$

$$h_\beta(\mathbf{x} - \boldsymbol{\Delta}) \leq \max_{\boldsymbol{\eta}} (x_{\boldsymbol{\eta}} - \Delta) + \beta^{-1} \log(M) \leq 0$$

and $g_\Delta \circ h_\beta(\mathbf{x} - \boldsymbol{\Delta}) = 0$.

Now the arguments similar to (G.6) and (G.7) imply for $\mathbf{S} - \mathbf{q}$ and $\tilde{\mathbf{S}} - \mathbf{q}$

$$\begin{aligned} \mathbb{P}\left(\max_{\boldsymbol{\eta}}\{S(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 0\right) &\leq \mathbb{E}[g_{\Delta} \circ h_{\beta}(\mathbf{S} - \mathbf{q} + \boldsymbol{\Delta})] \\ &\leq \mathbb{E}[g_{\Delta} \circ h_{\beta}(\tilde{\mathbf{S}} - \mathbf{q} + \boldsymbol{\Delta})] + C(M)\Delta^{-3}\delta_n \\ &\leq \mathbb{P}\left(\max_{\boldsymbol{\eta}}\{\tilde{S}(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq -2\Delta\right) + C(M)\Delta^{-3}\delta_n, \\ \mathbb{P}\left(\max_{\boldsymbol{\eta}}\{S(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 0\right) &\geq \mathbb{E}[g_{\Delta} \circ h_{\beta}(\mathbf{S} - \mathbf{q} - \boldsymbol{\Delta})] \\ &\geq \mathbb{E}[g_{\Delta} \circ h_{\beta}(\tilde{\mathbf{S}} - \mathbf{q} - \boldsymbol{\Delta})] - C(M)\Delta^{-3}\delta_n \\ &\geq \mathbb{P}\left(\max_{\boldsymbol{\eta}}\{\tilde{S}(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 2\Delta\right) - C(M)\Delta^{-3}\delta_n. \end{aligned}$$

Summarizing the above considerations yields the following bound.

Theorem G.5.1. Let $\mathbf{S} = \sum_i \boldsymbol{\xi}_i$, where $\boldsymbol{\xi}_i = \{\xi_i(\boldsymbol{\eta}), \boldsymbol{\eta} \in \Xi\}$ are independent zero mean random vectors with finite third moments and $V_i^2 = \text{Var}(\boldsymbol{\xi}_i)$ for $i = 1, \dots, n$. Let also $\tilde{\mathbf{S}} = \{\tilde{S}(\boldsymbol{\eta}), \boldsymbol{\eta} \in \Xi\}$ be a zero mean Gaussian vector in \mathbb{R}^M with the same covariance structure as \mathbf{S} , that is, $\text{Var}(\mathbf{S}) = \text{Var}(\tilde{\mathbf{S}})$. Then for $f_{\Delta, M}(\mathbf{x}) = g_{\Delta} \circ h_{\beta}(\mathbf{x})$ with $\beta = \Delta^{-1} \log(M)$ and for any collection of critical values $\mathbf{z} = (q(\boldsymbol{\eta}), \boldsymbol{\eta} \in \Xi)$, it holds

$$|\mathbb{E}f_{\Delta, M}(\mathbf{S} - \mathbf{z}) - \mathbb{E}f_{\Delta, M}(\tilde{\mathbf{S}} - \mathbf{z})| \leq C(M)\Delta^{-3}\delta_n$$

with $C(M)$ from (G.31) and

$$\delta_n \stackrel{\text{def}}{=} \sum_i (\mathbb{E}\|\boldsymbol{\xi}_i\|_{\infty}^3 + \mathbb{E}\|\tilde{\boldsymbol{\xi}}_i\|_{\infty}^3)$$

Moreover, for any $\Delta > 0$, it holds

$$\begin{aligned} \mathbb{P}\left(\max_{\boldsymbol{\eta}}\{S(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 0\right) &\leq \mathbb{P}\left(\max_{\boldsymbol{\eta}}\{\tilde{S}(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq -2\Delta\right) + C(M)\Delta^{-3}\delta_n, \\ \mathbb{P}\left(\max_{\boldsymbol{\eta}}\{S(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 0\right) &\geq \mathbb{P}\left(\max_{\boldsymbol{\eta}}\{\tilde{S}(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 2\Delta\right) - C(M)\Delta^{-3}\delta_n. \end{aligned} \tag{G.32}$$

Finally,

$$\begin{aligned} \mathbb{P}\left(\max_{\boldsymbol{\eta}}\{S(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 0\right) &\leq \mathbb{P}\left(\max_{\boldsymbol{\eta}}\{\tilde{S}(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 0\right) + C(M)\Delta^{-3}\delta_n + C\Delta\sqrt{\log(M/\Delta)}, \\ \mathbb{P}\left(\max_{\boldsymbol{\eta}}\{S(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 0\right) &\geq \mathbb{P}\left(\max_{\boldsymbol{\eta}}\{\tilde{S}(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 0\right) - C(M)\Delta^{-3}\delta_n - C\Delta\sqrt{\log(M/\Delta)}. \end{aligned}$$

The last statement of the theorem is obtained by combining (G.32) and the anti-concentration bound of Theorem F.3.1.

G.6 GAR for weighted sums

Now we consider weighted sums $S(\boldsymbol{\eta}) = \sum_i c_i(\boldsymbol{\eta})\varepsilon_i$ for a finite set $\Xi = \{\boldsymbol{\eta}\}$ and the set of coefficients $\mathbf{c}_i = \{c_i(\boldsymbol{\eta})\}$ for $i = 1, \dots, n$. Here ε_i are independent zero mean with finite third moments. We denote $v_i^2 = \text{Var}(\varepsilon_i)$. Obviously

$$\mathbb{E}S(\boldsymbol{\eta}) = 0, \quad \text{Var}\{S(\boldsymbol{\eta})\} = V^2(\boldsymbol{\eta}) \stackrel{\text{def}}{=} \sum_i c_i^2(\boldsymbol{\eta})v_i^2.$$

Let $\tilde{\varepsilon}_i$'s be Gaussian analogs of ε_i 's, that is, the r.v.'s $\tilde{\varepsilon}_i$ are independent normal with $\tilde{\varepsilon}_i \sim \mathcal{N}(0, v_i^2)$. Define the Gaussian sums

$$\tilde{S}(\boldsymbol{\eta}) = \sum_i c_i(\boldsymbol{\eta})\tilde{\varepsilon}_i, \quad \boldsymbol{\eta} \in \Xi.$$

Given $\mathbf{z} = \{q(\boldsymbol{\eta}), \boldsymbol{\eta} \in \Xi\}$, we aim at approximating the distribution of sup-norm the vector $\mathbf{S} - \mathbf{z} = \{S(\boldsymbol{\eta}) - q(\boldsymbol{\eta}), \boldsymbol{\eta} \in \Xi\}$ by a similar distribution for the Gaussian vector $\tilde{\mathbf{S}} = \{\tilde{S}(\boldsymbol{\eta}), \boldsymbol{\eta} \in \Xi\}$. Let f be a function on $\mathbb{R}^\mathbb{M}$. The main step in the construction is a bound for the difference $\mathbb{E}f(\mathbf{S} - \mathbf{z}) - \mathbb{E}f(\tilde{\mathbf{S}} - \mathbf{z})$. Below we proceed for $\mathbf{z} \equiv 0$, but the general case is similar.

Suppose that the function f is smooth and satisfies the condition (G.26) for any points $\mathbf{x}, \mathbf{x} + \mathbf{d} \in \mathbb{R}^\mathbb{M}$. Introduce for $k = 1, \dots, n$ the telescopic sums in $\mathbb{R}^\mathbb{M}$

$$\mathbf{S}^{(i)} \stackrel{\text{def}}{=} \mathbf{c}_1\varepsilon_1 + \dots + \mathbf{c}_{i-1}\varepsilon_{i-1} + \mathbf{c}_{i+1}\tilde{\varepsilon}_{i+1} + \dots + \mathbf{c}_n\tilde{\varepsilon}_n.$$

It holds by (G.26)

$$\begin{aligned} & |f(\mathbf{S}^{(i)} + \mathbf{c}_i\varepsilon_i) - f(\mathbf{S}^{(i)} + \mathbf{c}_i\tilde{\varepsilon}_i) - \mathbf{c}_i^\top f''(\mathbf{S}^{(i)})(\varepsilon_i - \tilde{\varepsilon}_i) - \mathbf{c}_i^\top f''(\mathbf{S}^{(i)})\mathbf{c}_i(\varepsilon_i^2 - \tilde{\varepsilon}_i^2)/2| \\ & \leq C_f(|\varepsilon_i|^3 + |\tilde{\varepsilon}_i|^3)\|\mathbf{c}_i\|_\infty^3. \end{aligned}$$

Similarly to (G.27) this implies

$$|\mathbb{E}f(\mathbf{S}) - \mathbb{E}f(\tilde{\mathbf{S}})| \leq C_f \delta_n$$

with

$$\delta_n \stackrel{\text{def}}{=} \sum_i \mathbb{E}(|\varepsilon_i|^3 + |\tilde{\varepsilon}_i|^3)\|\mathbf{c}_i\|_\infty^3 = \sum_i \mathbb{E}(|\varepsilon_i|^3 + 3v_i^3)\|\mathbf{c}_i\|_\infty^3. \quad (\text{G.33})$$

Here we used that $\tilde{\varepsilon}_i \sim \mathcal{N}(0, v_i^2)$ fulfills $\mathbb{E}|\tilde{\varepsilon}_i|^3 = 3v_i^3$.

Theorem G.6.1. *Let ε_i be independent zero mean with finite third moments and $v_i^2 = \text{Var}(\varepsilon_i)$ for $i = 1, \dots, n$. For a finite collection $\mathbf{c}(\boldsymbol{\eta}) = (c_i(\boldsymbol{\eta}), \boldsymbol{\eta} \in \Xi)$ of weighting schemes with $\mathbb{M} \stackrel{\text{def}}{=} \#(\Xi)$, define*

$$S(\boldsymbol{\eta}) \stackrel{\text{def}}{=} \sum_{i=1}^n c_i(\boldsymbol{\eta}) \varepsilon_i.$$

Then for each function $f(\cdot)$ satisfying (G.26) and any collection of critical values $q(\boldsymbol{\eta})$, it holds

$$|\mathbb{E}f(\mathbf{S} - \mathbf{z}) - \mathbb{E}f(\tilde{\mathbf{S}} - \mathbf{z})| \leq C_f \delta_n,$$

where $\tilde{\mathbf{S}} = \{\tilde{S}(\boldsymbol{\eta})\}$ is a zero mean Gaussian vector in \mathbb{R}^M with $\text{Var}(\mathbf{S}) = \text{Var}(\tilde{\mathbf{S}})$ and δ_n is given by (G.33). Moreover, for any $\Delta > 0$, it holds with $C(M)$ from (G.31)

$$\begin{aligned} \mathbb{P}\left(\max_{\boldsymbol{\eta}}\{S(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 0\right) &\leq \mathbb{P}\left(\max_{\boldsymbol{\eta}}\{\tilde{S}(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq -2\Delta\right) + C(M)\Delta^{-3}\delta_n, \\ \mathbb{P}\left(\max_{\boldsymbol{\eta}}\{S(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 0\right) &\geq \mathbb{P}\left(\max_{\boldsymbol{\eta}}\{\tilde{S}(\boldsymbol{\eta}) - q(\boldsymbol{\eta})\} \geq 2\Delta\right) - C(M)\Delta^{-3}\delta_n. \end{aligned}$$

G.7 A uniform bound for the maximum of the norm of weighted vector sums

Let $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n$ be independent zero mean vectors in \mathbb{R}^p with finite third moments. Denote $v_i^2 = \text{Var}(\boldsymbol{\xi}_i)$. Consider weighted sums $\mathbf{S}(\boldsymbol{\eta}) = \sum_i c_i(\boldsymbol{\eta}) \boldsymbol{\xi}_i$ for a finite set $\Xi = \{\boldsymbol{\eta}\}$ of cardinality $M = |\Xi|$, and the set of coefficients $\{c_i(\boldsymbol{\eta})\}$ for $i = 1, \dots, n$ and $\boldsymbol{\eta} \in \Xi$. Introduce independent Gaussian vectors $\tilde{\boldsymbol{\xi}}_i \sim \mathcal{N}(0, v_i^2)$ and consider the Gaussian sums $\tilde{\mathbf{S}}(\boldsymbol{\eta}) = \sum_i c_i(\boldsymbol{\eta}) \tilde{\boldsymbol{\xi}}_i$. Obviously

$$\mathbb{E}\tilde{\mathbf{S}}(\boldsymbol{\eta}) = \mathbb{E}\mathbf{S}(\boldsymbol{\eta}) = 0, \quad \text{Var}\{\tilde{\mathbf{S}}(\boldsymbol{\eta})\} = \text{Var}\{\mathbf{S}(\boldsymbol{\eta})\} = \sum_i c_i^2(\boldsymbol{\eta}) v_i^2, \quad \boldsymbol{\eta} \in \Xi.$$

We aim at quantifying the accuracy of approximation of the maximum of the norm $\|\mathbf{S}(\boldsymbol{\eta})\|$ over $\boldsymbol{\eta} \in \Xi$ by a similar maximum of $\|\tilde{\mathbf{S}}(\boldsymbol{\eta})\|$. We use an identity

$$\max_{\boldsymbol{\eta}} \|\mathbf{S}(\boldsymbol{\eta})\| = \max_{\boldsymbol{\eta}} \sup_{\boldsymbol{\gamma} \in \mathcal{S}_p} \boldsymbol{\gamma}^\top \mathbf{S}(\boldsymbol{\eta}) = \sup_{\boldsymbol{\gamma} \in \mathcal{S}_p} \max_{\boldsymbol{\eta}} \boldsymbol{\gamma}^\top \mathbf{S}(\boldsymbol{\eta}),$$

where \mathcal{S}_p is the unit sphere in \mathbb{R}^p . Let $\varepsilon > 0$ be fixed later, and \mathcal{G}_ε be a ε -net in \mathcal{S}_p . We can bound

$$\max_{\boldsymbol{\eta}} \|\mathbf{S}(\boldsymbol{\eta})\| \leq \frac{1}{1-\varepsilon} \max_{\boldsymbol{\gamma} \in \mathcal{G}_\varepsilon} \max_{\boldsymbol{\eta}} \boldsymbol{\gamma}^\top \mathbf{S}(\boldsymbol{\eta}).$$

For the cardinality $|\mathcal{G}_\varepsilon|$ we can also use the upper bound $|\mathcal{G}_\varepsilon| \leq (2/\varepsilon)^{p-1}$. Each product $\boldsymbol{\gamma}^\top \mathbf{S}(\boldsymbol{\eta})$ is the sum $\sum_i c_i(\boldsymbol{\eta}) \boldsymbol{\gamma}^\top \boldsymbol{\xi}_i$ which can be well approximated in distribution by $\sum_i c_i(\boldsymbol{\eta}) \boldsymbol{\gamma}^\top \tilde{\boldsymbol{\xi}}_i$. The GAR bound (G.28) yields for any $\Delta > 0$

$$\begin{aligned} & \mathbb{P}\left(\max_{\gamma \in \mathcal{G}_\varepsilon} \max_{\eta} \gamma^\top \mathbf{S}(\eta) \geq q - \varepsilon z^*\right) \\ & \leq \mathbb{P}\left(\max_{\gamma \in \mathcal{G}_\varepsilon} \max_{\eta} \gamma^\top \tilde{\mathbf{S}}(\eta) \geq q - \varepsilon z^* - 2\Delta\right) + \Delta^{-3} C(\mathbb{G}_\varepsilon) \delta_n, \end{aligned}$$

where

$$\begin{aligned} \delta_n & \stackrel{\text{def}}{=} \sum_i \mathbb{E} \left(\max_{\eta, \gamma} |c_i(\eta) \gamma^\top \xi_i|^3 + \max_{\eta, \gamma} |c_i(\eta) \gamma^\top \tilde{\xi}_i|^3 \right) \\ & \leq \sum_i \mathbb{E} (\|\xi_i\|^3 + \|\tilde{\xi}_i\|^3) \|c_i\|_\infty^3, \end{aligned} \quad (\text{G.34})$$

$\mathbb{G}_\varepsilon = \mathbb{M} |\mathcal{G}_\varepsilon|$ and $C(\mathbb{G}_\varepsilon)$ is given by (G.31); $C(\mathbb{G}_\varepsilon) \leq C \log^2(\mathbb{G}_\varepsilon)$. The anti-concentration bound (F.11) implies

$$\begin{aligned} & \left| \mathbb{P}\left(\max_{\gamma \in \mathcal{G}_\varepsilon} \max_{\eta} \gamma^\top \tilde{\mathbf{S}}(\eta) \geq q - \varepsilon z^* - 2\Delta\right) - \mathbb{P}\left(\max_{\gamma \in \mathcal{G}_\varepsilon} \max_{\eta} \gamma^\top \tilde{\mathbf{S}}(\eta) \geq q\right) \right| \\ & \leq C(\varepsilon z^* + 2\Delta) \sqrt{\log(\mathbb{G}_\varepsilon / \Delta)}. \end{aligned}$$

Altogether yields

$$\begin{aligned} & \left| \mathbb{P}\left(\max_{\eta} \|\mathbf{S}(\eta)\| \geq q\right) - \mathbb{P}\left(\max_{\eta} \|\tilde{\mathbf{S}}(\eta)\| \geq q\right) \right| \\ & \leq \Delta^{-3} C(\mathbb{G}_\varepsilon) \delta_n + C(\varepsilon z^* + 2\Delta) \sqrt{\log(\mathbb{G}_\varepsilon / \Delta)}. \end{aligned}$$

The value ε can be selected to balance Δ and εz^* yielding $\varepsilon = \Delta/z^*$. Then optimization w.r.t. Δ leads in view of $C(\mathbb{G}_\varepsilon) \leq C \log^2(\mathbb{G}_\varepsilon)$ to the relations

$$\Delta^4 \asymp \log^{3/2}(\mathbb{G}_\varepsilon) \delta_n, \quad \Delta^{-3} C(\mathbb{G}_\varepsilon) \delta_n \asymp \Delta \sqrt{\log(\mathbb{G}_\varepsilon / \Delta)} \asymp \delta_n^{1/4} \log^{7/8}(\mathbb{G}_\varepsilon).$$

Further, $\mathbb{G}_\varepsilon = \mathbb{M} \times |\mathcal{G}_\varepsilon| \leq \mathbb{M}(2/\varepsilon)^{p-1}$ fulfills

$$\log(\mathbb{G}_\varepsilon) \approx \log \mathbb{M} + (p-1) \log(2z^*/\Delta) \leq \log \mathbb{M} + (p-1) \log(2z^*/\delta_n^{1/4}). \quad (\text{G.35})$$

The obtained bounds can be easily extended to the case of η -dependent critical values.

Theorem G.7.1. Let ξ_i be independent zero mean vectors in \mathbb{R}^p with finite third moments. Denote $v_i^2 = \text{Var}(\xi_i)$. Consider weighted sums $\mathbf{S}(\eta) = \sum_i c_i(\eta) \xi_i$ for a finite set $\Xi = \{\eta\}$ and a set of coefficients $c_i(\eta)$ for $i = 1, \dots, n$ and $\eta \in \Xi$. If $\tilde{\xi}_i \sim \mathcal{N}(0, v_i^2)$ and $\tilde{\mathbf{S}}(\eta) = \sum_i c_i(\eta) \tilde{\xi}_i$, then for any $\Delta > 0$ and $\varepsilon \in (0, 1)$, and any values $q(\eta) \geq 0$, it holds with δ_n from (G.34)

$$\begin{aligned} & \left| \mathbb{P}\left(\max_{\eta} \{\|\mathbf{S}(\eta)\| - q(\eta)\} \geq 0\right) - \mathbb{P}\left(\max_{\eta} \{\|\tilde{\mathbf{S}}(\eta)\| - q(\eta)\} \geq 0\right) \right| \\ & \leq C \Delta^{-3} \log^2(\mathbb{G}_\varepsilon) \delta_n + C(\varepsilon z^* + 2\Delta) \sqrt{\log(\mathbb{G}_\varepsilon / \Delta)} \\ & \leq C \delta_n^{1/4} \log^{7/8}(\mathbb{G}_\varepsilon), \end{aligned}$$

where $\log(\mathbb{G}_\varepsilon)$ follows (G.35).

H

Deviation bounds for random processes

This chapter presents some general results of the theory of empirical processes. We assume some exponential moment conditions on the increments of the process which allow to apply the well developed chaining arguments in Orlicz spaces; see e.g. [van der Vaart and Wellner \(1996\)](#), Chapter 2.2. We, however, follow the more recent approach inspired by the notions of generic chaining and majorizing measures due to M. Talagrand; The chaining arguments are replaced by the *pilling* device; see e.g. [Talagrand \(1996, 2001, 2005\)](#). The results are close to that of [Bednorz \(2006\)](#). We state the results in a slightly different form and present an independent and self-contained proof.

The first result states a bound for local fluctuations of the process $\mathcal{U}(\mathbf{v})$ given on a metric space Υ . Then this result will be used for bounding the maximum of the negatively drifted process $\mathcal{U}(\mathbf{v}, \mathbf{v}^*) \stackrel{\text{def}}{=} \mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}^*)$ over a vicinity $\Upsilon_\circ(\mathbf{r}_0)$ of the central point \mathbf{v}^* . The behavior of $\mathcal{U}(\mathbf{v})$ outside of the local central set $\Upsilon_\circ(\mathbf{r}_0)$ is described using the *upper function* method. Namely, we construct a deterministic function $f(\mathbf{r}, \mathbf{r}_0)$ ensuring that with probability at least $1 - e^{-x}$ it holds on a dominating set of probability at least $1 - e^{-x}$ that $\mathcal{U}(\mathbf{v}, \mathbf{v}^*) - f(d(\mathbf{v}, \mathbf{v}^*), \mathbf{r}_0) < 0$ for all $\mathbf{v} \notin \Upsilon_\circ(\mathbf{r}_0)$.

H.1 Chaining and covering numbers

An important step in the whole construction is an exponential bound on the maximum of a random process $\mathcal{U}(\mathbf{v})$ under the exponential moment conditions on its increments. Let $d(\mathbf{v}, \mathbf{v}')$ be a semi-distance on Υ . We suppose the following condition to hold:

(Ed) *There exist $g > 0$, $\mathbf{r}_0 > 0$, $\nu_0 \geq 1$, such that for any $\lambda \leq g$ and $\mathbf{v}, \mathbf{v}' \in \Upsilon$ with $d(\mathbf{v}, \mathbf{v}') \leq \mathbf{r}_0$*

$$\log I\!\!E \exp \left\{ \lambda \frac{\mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}')}{d(\mathbf{v}, \mathbf{v}')} \right\} \leq \nu_0^2 \lambda^2 / 2.$$

By $\mathcal{B}_{\mathbf{r}}(\mathbf{v})$ we denote the d -ball centered at \mathbf{v} of radius \mathbf{r} :

$$\mathcal{B}_r(\mathbf{v}) \stackrel{\text{def}}{=} \{\mathbf{v}' \in \Upsilon : d(\mathbf{v}, \mathbf{v}') \leq r\}.$$

Let Υ° be a subset of a ball in Υ with center at \mathbf{v}^* and radius r_0 , and let a sequence r_k be fixed with $r_k = r_0 2^{-k}$.

For each k , by \mathcal{M}_k we denote a r_k -net in Υ° , so that

$$\Upsilon^\circ \subseteq \bigcup_{\mathbf{v} \in \mathcal{M}_k} \mathcal{B}_{r_k}(\mathbf{v}).$$

Let also $\Pi_k \mathbf{v}$ be the closest to \mathbf{v} point from \mathcal{M}_k , so that $d(\mathbf{v}, \Pi_k \mathbf{v}) \leq r_k$. We assume that \mathcal{M}_0 consists of one point \mathbf{v}^* , that is, $\Pi_0 \mathbf{v} = \mathbf{v}^*$. Let $N_k \stackrel{\text{def}}{=} |\mathcal{M}_k|$ denote the cardinality of \mathcal{M}_k . Finally set $c_k = 2^{-k}$ for $k \geq 1$, and define the values $Q_1(\Upsilon^\circ)$ and $Q_2(\Upsilon^\circ)$ by

$$\begin{aligned} Q_1(\Upsilon^\circ) &\stackrel{\text{def}}{=} \sum_{k=1}^{\infty} c_k \sqrt{2 \log(2N_k)} = \sum_{k=1}^{\infty} 2^{-k} \sqrt{2 \log(2N_k)}, \\ Q_2(\Upsilon^\circ) &\stackrel{\text{def}}{=} \sum_{k=1}^{\infty} 2c_k \log(2N_k) = \sum_{k=1}^{\infty} 2^{-k+1} \log(2N_k). \end{aligned} \tag{H.1}$$

By the Cauchy-Schwartz inequality $Q_1^2(\Upsilon^\circ) \leq Q_2(\Upsilon^\circ)$. The inverse relation is not generally true and one can build some examples with $Q_1(\Upsilon^\circ)$ finite and $Q_2(\Upsilon^\circ)$ infinite. If the process $\mathcal{U}(\mathbf{v})$ is sub-Gaussian and **(Ed)** is fulfilled with $g = \infty$, then one can only operate with $Q_1(\Upsilon^\circ)$ which is equivalent to the Dudley integral; see (H.13) below.

Theorem H.1.1. *Let \mathcal{U} be a separable process and Υ° be a ball in Υ with center \mathbf{v}° and radius r_0 for the distance $d(\cdot, \cdot)$, i.e. $d(\mathbf{v}, \mathbf{v}^\circ) \leq r_0$ for all $\mathbf{v} \in \Upsilon^\circ$. If **(Ed)** holds with $g = \infty$ then for any $x \geq 1/2$, it holds with $Q_1 = Q_1(\Upsilon^\circ)$ and $Q_2 = Q_2(\Upsilon^\circ)$*

$$\mathbb{P}\left(\frac{1}{\nu_0 r_0} \sup_{\mathbf{v} \in \Upsilon^\circ} \mathcal{U}(\mathbf{v}, \mathbf{v}^*) \geq \mathfrak{z}_{\mathbb{H}}(x)\right) \leq e^{-x}$$

with

$$\mathfrak{z}_{\mathbb{H}}(x) \stackrel{\text{def}}{=} 2Q_1 + \sqrt{8x}.$$

If **(Ed)** holds with $g \leq \infty$, then

$$\mathbb{P}\left\{\frac{1}{\nu_0 r_0} \sup_{\mathbf{v} \in \Upsilon^\circ} \mathcal{U}(\mathbf{v}, \mathbf{v}^*) \geq \mathfrak{z}_{\mathbb{H}}(x)\right\} \leq e^{-x}, \tag{H.2}$$

where $\mathfrak{z}_{\mathbb{H}}(x)$ is given by one of the following rules:

$$\mathfrak{z}_{\mathbb{H}}(x) = 2Q_1 + \sqrt{8x} + 2g^{-1}(g^{-2}x + 1)Q_2,$$

$$\mathfrak{z}_{\mathbb{H}}(x) = \begin{cases} 2\sqrt{Q_2 + 2x}, & \text{if } Q_2 + 2x \leq g^2, \\ 2g^{-1}x + g^{-1}Q_2 + g, & \text{if } Q_2 + 2x > g^2. \end{cases} \tag{H.3}$$

Moreover, the r.v. $\mathcal{U}^*(\mathbf{r}_0) \stackrel{\text{def}}{=} \sup_{\mathbf{v} \in \Upsilon^\circ} \mathcal{U}(\mathbf{v}, \mathbf{v}^*)$ fulfills

$$\begin{aligned} \mathbb{E}\mathcal{U}^*(\mathbf{r}_0) &\leq 2\nu_0 \mathbf{r}_0 (\mathbb{Q}_1 + \mathbb{Q}_2/g + 3), \\ \{\mathbb{E}|\mathcal{U}^*(\mathbf{r}_0)|^2\}^{1/2} &\leq 2\nu_0 \mathbf{r}_0 (\mathbb{Q}_1 + \mathbb{Q}_2/g + 4). \end{aligned} \quad (\text{H.4})$$

Proof. We start the proof by stating some general facts for a convex combinations of sub-exponential r.v.'s ζ_k such that

$$\log \mathbb{E} \exp(\lambda \zeta_k) \leq \frac{q_k^2 + \lambda^2}{2}, \quad |\lambda| \leq g, \quad k = 0, 1, 2, \dots, \quad (\text{H.5})$$

where $q_k \geq 1$ are fixed numbers, and g is some positive value or infinity. We aim at bounding a sum S of the form $S = \sum_k c_k \zeta_k$ for a sequence of positive weights c_k satisfying $\sum_k c_k = 1$. We implicitly assume that the numbers q_k grow with k in a way that $\sum_k \exp(-q_k) \leq 1$. Define

$$\mathbb{H}_1 \stackrel{\text{def}}{=} \sum_k c_k q_k, \quad \mathbb{H}_2 \stackrel{\text{def}}{=} \sum_k c_k q_k^2.$$

Lemma H.1.1. Suppose that random variables ζ_k follow (H.5) with $g = \infty$ and $\sum_k \exp(-q_k) \leq 1$. Let also $\sum_k c_k = 1$. Then it holds for the sum $S = \sum_k c_k \zeta_k$

$$\log \mathbb{E} \exp(S) \leq \mathbb{H}_1$$

and for any $x \geq 1/2$,

$$\mathbb{P}(S \geq \mathbb{H}_1 + \sqrt{2x}) \leq e^{-x}. \quad (\text{H.6})$$

If (H.5) holds for $g < \infty$, then for each $\lambda > 0$ with $|\lambda| \leq g$

$$\log \mathbb{E} \exp(\lambda S) \leq (\mathbb{H}_2 + \lambda^2)/2, \quad (\text{H.7})$$

and it holds for $x \geq 1/2$

$$\mathbb{P}\{S \geq \mathfrak{z}_{\mathbb{H}}(x)\} \leq e^{-x}, \quad (\text{H.8})$$

where $\mathfrak{z}_{\mathbb{H}}(x)$ is given by (H.3). Moreover, if $g^2 \geq \mathbb{H}_2 + 1$, then

$$\mathbb{E}S \leq \mathbb{H}_1 + \mathbb{H}_2/g + 3, \quad \{\mathbb{E}S^2\}^{1/2} \leq \mathbb{H}_1 + \mathbb{H}_2/g + 4.$$

Proof. Consider first the sub-Gaussian case with $g = \infty$. Define $\alpha_k = c_k/q_k$. Obviously $\sum_k \alpha_k \leq \sum_k c_k = 1$. By the Hölder inequality and (H.5), it holds

$$\begin{aligned}\log I\!E \exp\left(\sum_k c_k \zeta_k\right) &= \log I\!E \exp\left(\sum_k \alpha_k q_k \zeta_k\right) \leq \sum_k \alpha_k \log I\!E \exp(q_k \zeta_k) \\ &\leq \frac{1}{2} \sum_k \alpha_k (q_k^2 + q_k^2) \leq \sum_k c_k q_k.\end{aligned}$$

Further, by the same arguments, it holds

$$\log I\!E \exp(\lambda S) \leq \sum_k c_k \log I\!E \exp(\lambda \zeta_k) \leq \frac{1}{2} \sum_k c_k (q_k^2 + \lambda^2)$$

and the assertion (H.7) follows as well.

Let $x \geq 1/2$ be fixed. With $z_k = q_k + \sqrt{2x}$, it follows by (H.5) for $\lambda_k = z_k$ in view of $\sum_k e^{-q_k} \leq 1$

$$\begin{aligned}I\!P\left(\sum_k c_k (\zeta_k - z_k) \geq 0\right) &\leq \sum_k I\!P(\zeta_k - z_k \geq 0) \leq \sum_k I\!E \exp\{\lambda_k (\zeta_k - z_k)\} \\ &\leq \sum_k \exp(-\lambda_k z_k + \lambda_k^2/2 + q_k^2/2) = \sum_k \exp(-z_k^2/2 + q_k^2/2) \\ &= \sum_k \exp(-x - q_k \sqrt{2x}) \leq e^{-x}.\end{aligned}\tag{H.9}$$

This implies

$$\sum_k c_k z_k = \sum_k c_k (q_k + \sqrt{2x}) = H_1 + \sqrt{2x}\tag{H.10}$$

and the assertion (H.6) follows.

Now we briefly discuss how the condition (H.5) can be relaxed to the case of a finite g . Suppose that (H.5) holds for all $\lambda \leq g < \infty$. Define $k(x)$ as the largest index k , for which $\lambda_k = q_k + \sqrt{2x} \leq g$. For $k > k(x)$, define $\lambda_k = g$ and

$$z_k = \frac{x + q_k}{g} + \frac{g}{2} + \frac{q_k^2}{2g}.\tag{H.11}$$

The above arguments yield for $k > k(x)$

$$I\!P(\zeta_k \geq z_k) \leq \exp\left(-gz_k + \frac{1}{2}(q_k^2 + g^2)\right) = \exp(-x - q_k).$$

This and (H.9) yield

$$\begin{aligned}\sum_k I\!P(\zeta_k \geq z_k) &\leq \sum_{k \leq k(x)} \exp(-x - q_k \sqrt{2x}) + \sum_{k > k(x)} \exp(-x - q_k) \\ &\leq \sum_k \exp(-x - q_k) \leq e^{-x}.\end{aligned}$$

Further, as $q_k > g$ for $k > k(x)$, it follows from the definition (H.11)

$$\begin{aligned} \sum_{k>k(x)} c_k z_k &= \frac{1}{g} \sum_{k>k(x)} c_k (x + q_k) + \frac{g}{2} \sum_{k>k(x)} c_k + \frac{1}{2g} \sum_{k>k(x)} c_k q_k^2 \\ &\leq \frac{1}{g} \sum_{k>k(x)} c_k q_k + \left(\frac{x}{g^3} + \frac{1}{g} \right) \sum_{k>k(x)} c_k q_k^2. \end{aligned}$$

This and (H.10) imply due to $g \geq 1$

$$\sum_k c_k z_k \leq \sum_k c_k q_k + \left(\frac{x}{g^3} + \frac{1}{g} \right) \sum_k c_k q_k^2 + \sqrt{2x} \leq H_1 + \left(\frac{x}{g^3} + \frac{1}{g} \right) H_2 + \sqrt{2x}.$$

In particular, if $x \leq g^2$, then

$$\sum_k c_k z_k \leq H_1 + \frac{2}{g} H_2 + \sqrt{2x}.$$

Now (H.8) with $\zeta(x) = H_1 + \sqrt{2x} + g^{-1}(g^{-2}x + 1)H_2$ follows similarly to (H.6). Further, if $\zeta(x) = \sqrt{H_2 + 2x} \leq g$, then (H.7) with $\lambda = \zeta(x)$ and the exponential Chebyshev inequality implies again

$$\mathbb{P}(S \geq \zeta(x)) \leq \exp\left(-\lambda \zeta(x) + \frac{H_2 + \lambda^2}{2}\right) = \exp\left(\frac{-\zeta^2(x) + H_2}{2}\right) = \exp(-x).$$

Similarly one can check the case with $\lambda = g$ and $\zeta(x) = x/g + (H_2/g + g)/2 > g$.

To bound the moments of S , we apply the following technical result: if

$$\mathbb{P}(S \geq \zeta(x)) \leq e^{-x}$$

for all $x \geq x_0$ and if $\zeta(\cdot)$ is absolutely continuous, then

$$\begin{aligned} \mathbb{E}S &\leq \zeta(x_0) + \int_{x_0}^{\infty} \zeta'(x) e^{-x} dx, \\ \mathbb{E}S^2 &\leq \zeta^2(x_0) + 2 \int_{x_0}^{\infty} \zeta(x) \zeta'(x) e^{-x} dx. \end{aligned}$$

For $\zeta(x) = H_1 + \sqrt{2x} + g^{-1}(g^{-2}x + 1)H_2$, it holds $\zeta'(x) \leq 1 + g^{-3}$. In view of $g^2 \geq H_2 + 1$

$$\mathbb{E}S \leq H_1 + 1 + (H_2 + 1/2)/g + \int_{1/2}^{\infty} (1 + g^{-3}) e^{-x} dx \leq H_1 + H_2/g + 3.$$

Similarly one can bound

$$\mathbb{E}S^2 \leq (H_1 + H_2/g + 3/2)^2 + 2 \int_{1/2}^{\infty} \left(\frac{1}{\sqrt{2x}} + g^{-3} \right) \zeta(x) e^{-x} dx \leq (H_1 + H_2/g + 4)^2$$

as required.

Now we show how the statement of the theorem can be reduced to the bounds of Lemma H.1.1. Denote for $i < k$ by Π_i^k the product $\Pi_i^k = \Pi_i \Pi_{i+1} \dots \Pi_k$. As $\Pi_0 \mathbf{v} \equiv \mathbf{v}^*$, the telescopic sum devices yields

$$|\mathcal{U}(\Pi_k \mathbf{v}) - \mathcal{U}(\mathbf{v}^*)| \leq \sum_{i=1}^k |\mathcal{U}(\Pi_{i-1}^k \mathbf{v}) - \mathcal{U}(\Pi_i^k \mathbf{v})|.$$

Separability of $\mathcal{U}(\cdot)$ implies that

$$\lim_{k \rightarrow \infty} \mathcal{U}(\Pi_k \mathbf{v}) = \mathcal{U}(\mathbf{v}).$$

Therefore, it holds for any $\mathbf{v} \in \Upsilon^\circ$

$$|\mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}^*)| = \lim_{k \rightarrow \infty} |\mathcal{U}(\Pi_k \mathbf{v}) - \mathcal{U}(\mathbf{v}^*)| \leq \sum_{k=1}^{\infty} \xi_k^*,$$

where

$$\xi_k^* \stackrel{\text{def}}{=} \max_{\mathbf{v} \in \mathcal{M}_k} |\mathcal{U}(\mathbf{v}) - \mathcal{U}(\Pi_{k-1} \mathbf{v})|.$$

For each $\mathbf{v} \in \mathcal{M}_k$, it holds $d(\mathbf{v}, \Pi_{k-1} \mathbf{v}) \leq \mathbf{r}_{k-1}$ and

$$|\mathcal{U}(\mathbf{v}) - \mathcal{U}(\Pi_{k-1} \mathbf{v})| \leq \mathbf{r}_{k-1} \frac{|\mathcal{U}(\mathbf{v}) - \mathcal{U}(\Pi_{k-1} \mathbf{v})|}{d(\mathbf{v}, \Pi_{k-1} \mathbf{v})}.$$

This implies by the Jensen inequality and (Ed) in view of $e^{|x|} \leq e^x + e^{-x}$ for each $k \geq 1$ and $|\lambda| \leq g$

$$\mathbb{E} \exp\left(\frac{\lambda}{\mathbf{r}_{k-1}} \xi_k^*\right) \leq 2 \sum_{\mathbf{v} \in \mathcal{M}_k} \mathbb{E} \exp\left(\lambda \frac{|\mathcal{U}(\mathbf{v}) - \mathcal{U}(\Pi_{k-1} \mathbf{v})|}{d(\mathbf{v}, \Pi_{k-1} \mathbf{v})}\right) \leq 2N_k \exp(\lambda^2/2). \quad (\text{H.12})$$

For $k \geq 1$, define $q_k^2/2 = \log(2N_k)$, $c_k = 2^{-k}$, and $\zeta_k = \xi_k^*/\mathbf{r}_{k-1} = c_k^{-1} \xi_k^*/(2\mathbf{r}_0)$. Then (H.12) implies by $\mathbf{r}_{k-1} = 2^{-k+1} \mathbf{r}_0$

$$\log \mathbb{E} \exp(\lambda \zeta_k) \leq \log(2N_k) + \lambda^2/2 = \frac{q_k^2 + \lambda^2}{2}.$$

Now we apply Lemma H.1.1 with $c_k = 2^{-k}$. By construction

$$\sum_{k=1}^{\infty} c_k \zeta_k = \frac{1}{2\mathbf{r}_0} \sum_{k=1}^{\infty} \xi_k^*$$

and the results follow with $\mathbb{H}_1 = \mathbb{Q}_1(\Upsilon^\circ)$, $\mathbb{H}_2 = \mathbb{Q}_2(\Upsilon^\circ)$.

H.2 Entropy and Dudley's integral

The quantities $\mathbb{Q}_1(\Upsilon^\circ)$ and $\mathbb{Q}_2(\Upsilon^\circ)$ from (H.1) can be upper bounded by integrals over the interval $[0, 1]$. Let $\varepsilon \in (0, 1)$. Denote by $\mathcal{M}(\mathbf{r}_0, \varepsilon\mathbf{r}_0)$ an \mathbf{r}_ε -net in the \mathbf{r}_0 -ball Υ° for $\mathbf{r}_\varepsilon = \mathbf{r}_0\varepsilon$. Obviously $\mathbf{r}_k = \mathbf{r}_{\varepsilon_k}$ for $\varepsilon_k = 2^{-k}$, and the cardinality $\mathbb{N}(\varepsilon) = |\mathcal{M}(\varepsilon)|$ monotonously increases with ε . This allows to rewrite the deviation bound (H.2) in a form which only involves the cardinality $\mathbb{N}(\varepsilon)$.

Theorem H.2.1. *It holds for the values $\mathbb{Q}_1(\Upsilon^\circ)$ and $\mathbb{Q}_2(\Upsilon^\circ)$ from (H.1)*

$$\begin{aligned}\mathbb{Q}_1(\Upsilon^\circ) &= \sum_{k=1}^{\infty} 2^{-k} \sqrt{2 \log(2\mathbb{N}_k)} \leq 1.2 + \sqrt{8} \int_0^1 \sqrt{\log \mathbb{N}(\mathbf{r}_0, \varepsilon\mathbf{r}_0)} d\varepsilon. \\ \mathbb{Q}_2(\Upsilon^\circ) &= \sum_{k=1}^{\infty} 2^{-k+1} \log(2\mathbb{N}_k) \leq 1.4 + 4 \int_0^1 \log \mathbb{N}(\mathbf{r}_0, \varepsilon\mathbf{r}_0) d\varepsilon.\end{aligned}\quad (\text{H.13})$$

Proof. **To be done:**

The integral in the right hand-side is usually called *Dudley's integral*.

H.3 A local bound with generic chaining

Here we present a slightly different technique which is often called the *majorizing measure* and used in the *generic chaining* device; see [Talagrand \(2005\)](#). Formulation of the result involves a sigma-finite measure π on the space Υ . A typical example of choosing π is the Lebesgue measure on \mathbb{R}^p . Let Υ° be a subset of Υ , a sequence \mathbf{r}_k be fixed with $2\mathbf{r}_0 = \text{diam}(\Upsilon^\circ)$ and $\mathbf{r}_k = \mathbf{r}_0 2^{-k}$. Let also $\mathcal{B}_k(\mathbf{v})$ mean $\mathcal{B}_{\mathbf{r}_k}(\mathbf{v})$, that is, the d -ball centered at \mathbf{v} of radius \mathbf{r}_k and $\pi_k(\mathbf{v})$ denote its π -measure:

$$\pi_k(\mathbf{v}) \stackrel{\text{def}}{=} \int_{\mathcal{B}_k(\mathbf{v})} \pi(d\mathbf{v}') = \int_{\Upsilon^\circ} \mathbb{I}(d(\mathbf{v}, \mathbf{v}') \leq \mathbf{r}_k) \pi(d\mathbf{v}').$$

Denote also

$$\mathbb{N}_k \stackrel{\text{def}}{=} \int_{\Upsilon^\circ} \frac{\pi(d\mathbf{v})}{\pi_k(\mathbf{v})}, \quad k \geq 0. \quad (\text{H.14})$$

Finally set $c_0 = 1/3$, $c_k = 2^{-k+1}/3$ for $k \geq 1$, and define the values $\mathbb{Q}_1(\Upsilon^\circ)$ and $\mathbb{Q}_2(\Upsilon^\circ)$ by

$$\begin{aligned}\mathbb{G}_1(\Upsilon^\circ) &\stackrel{\text{def}}{=} \sum_{k=0}^{\infty} c_k \sqrt{2 \log(2\mathbb{N}_k)} = \frac{1}{3} \sqrt{2 \log(2\mathbb{N}_0)} + \frac{2}{3} \sum_{k=1}^{\infty} 2^{-k} \sqrt{2 \log(2\mathbb{N}_k)}, \\ \mathbb{G}_2(\Upsilon^\circ) &\stackrel{\text{def}}{=} 2 \sum_{k=0}^{\infty} c_k \log(2\mathbb{N}_k) = \frac{2}{3} \log(2\mathbb{N}_0) + \frac{4}{3} \sum_{k=1}^{\infty} 2^{-k} \log(2\mathbb{N}_k).\end{aligned}$$

By the Cauchy-Schwartz inequality $\mathbb{G}_1^2(\Upsilon^\circ) \leq \mathbb{G}_2(\Upsilon^\circ)$. The inverse relation is not generally true and one can build some examples with $\mathbb{G}_1(\Upsilon^\circ)$ finite and $\mathbb{G}(\Upsilon^\circ)$ infinite.

Theorem H.3.1. *Let \mathcal{U} be a separable process following to $(\mathcal{E}\mathbf{d})$. If Υ° is a d -ball in Υ with the center \mathbf{v}° and the radius \mathbf{r}_0 , i.e. $d(\mathbf{v}, \mathbf{v}^\circ) \leq \mathbf{r}_0$ for all $\mathbf{v} \in \Upsilon^\circ$, then for any $\mathbf{x} \geq 1/2$, it holds with $\mathbb{H}_1 = \mathbb{G}_1(\Upsilon^\circ)$ and $\mathbb{H}_2 = \mathbb{G}_2(\Upsilon^\circ)$*

$$\mathbb{P}\left(\frac{1}{3\nu_0\mathbf{r}_0} \sup_{\mathbf{v} \in \Upsilon^\circ} \mathcal{U}(\mathbf{v}, \mathbf{v}^*) \geq \mathbb{H}_1 + \sqrt{2\mathbf{x}}\right) \leq e^{-\mathbf{x}}.$$

If $\mathbf{g} \leq \infty$, then

$$\mathbb{P}\left\{\frac{1}{3\nu_0\mathbf{r}_0} \sup_{\mathbf{v} \in \Upsilon^\circ} \mathcal{U}(\mathbf{v}, \mathbf{v}^*) \geq \mathfrak{z}_{\mathbb{H}}(\mathbf{x})\right\} \leq e^{-\mathbf{x}},$$

where $\mathfrak{z}_{\mathbb{H}}(\mathbf{x})$ is given by (H.3). Moreover, the r.v. $\mathcal{U}^*(\mathbf{r}_0) \stackrel{\text{def}}{=} \sup_{\mathbf{v} \in \Upsilon^\circ} \mathcal{U}(\mathbf{v}, \mathbf{v}^*)$ fulfills

$$\mathbb{E}\mathcal{U}^*(\mathbf{r}_0) \leq 3\nu_0\mathbf{r}_0 (\mathbb{H}_1 + \mathbb{H}_2/\mathbf{g} + 3),$$

$$\{\mathbb{E}|\mathcal{U}^*(\mathbf{r}_0)|^2\}^{1/2} \leq 3\nu_0\mathbf{r}_0 (\mathbb{H}_1 + \mathbb{H}_2/\mathbf{g} + 4).$$

Proof. A simple change $\mathcal{U}(\cdot)$ with $\nu_0^{-1}\mathcal{U}(\cdot)$ and \mathbf{g} with $\mathbf{g}_0 = \nu_0\mathbf{g}$ allows to reduce the result to the case with $\nu_0 = 1$ which we assume below. Consider for $k \geq 0$ the smoothing operator \mathbb{S}_k defined as

$$\mathbb{S}_k f(\mathbf{v}^\circ) = \frac{1}{\pi_k(\mathbf{v}^\circ)} \int_{\mathcal{B}_k(\mathbf{v}^\circ)} f(\mathbf{v}) \pi(d\mathbf{v}).$$

Further, define

$$\mathbb{S}_{-1}\mathcal{U}(\mathbf{v}) \equiv \mathcal{U}(\mathbf{v}^\circ)$$

so that $\mathbb{S}_{-1}\mathcal{U}$ is a constant function and the same holds for $\mathbb{S}_k\mathbb{S}_{k-1}\dots\mathbb{S}_{-1}\mathcal{U}$ with any $k \geq 0$. If $f(\cdot) \leq g(\cdot)$ for two non-negative functions f and g , then $\mathbb{S}_k f(\cdot) \leq \mathbb{S}_k g(\cdot)$. Separability of the process \mathcal{U} implies that $\lim_k \mathbb{S}_k \mathcal{U}(\mathbf{v}) = \mathcal{U}(\mathbf{v})$. We conclude that for each $\mathbf{v} \in \Upsilon^\circ$

$$\begin{aligned} |\mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}^\circ)| &= \lim_{k \rightarrow \infty} |\mathbb{S}_k \mathcal{U}(\mathbf{v}) - \mathbb{S}_k \dots \mathbb{S}_{-1} \mathcal{U}(\mathbf{v})| \\ &\leq \lim_{k \rightarrow \infty} \sum_{i=0}^k |\mathbb{S}_k \dots \mathbb{S}_i (I - \mathbb{S}_{i-1}) \mathcal{U}(\mathbf{v})| \leq \sum_{i=0}^{\infty} \xi_i^*. \end{aligned}$$

Here $\xi_k^* \stackrel{\text{def}}{=} \sup_{\mathbf{v} \in \Upsilon^\circ} \xi_k(\mathbf{v})$ for $k \geq 0$ with

$$\xi_0(\mathbf{v}) \equiv |\mathbb{S}_0 \mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}^\circ)|, \quad \xi_k(\mathbf{v}) \stackrel{\text{def}}{=} |\mathbb{S}_k (I - \mathbb{S}_{k-1}) \mathcal{U}(\mathbf{v})|, \quad k \geq 1.$$

For a fixed point \mathbf{v}^\sharp and $k \geq 1$, it holds

$$\xi_k(\mathbf{v}^\sharp) \leq \frac{1}{\pi_k(\mathbf{v}^\sharp)} \int_{\mathcal{B}_k(\mathbf{v}^\sharp)} \frac{1}{\pi_{k-1}(\mathbf{v})} \int_{\mathcal{B}_{k-1}(\mathbf{v})} |\mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}')| \pi(d\mathbf{v}') \pi(d\mathbf{v}).$$

For each $\mathbf{v}' \in \mathcal{B}_{k-1}(\mathbf{v})$, it holds $d(\mathbf{v}, \mathbf{v}') \leq \mathbf{r}_{k-1} = 2\mathbf{r}_k$ and

$$|\mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}')| \leq \mathbf{r}_{k-1} \frac{|\mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}')|}{d(\mathbf{v}, \mathbf{v}')}.$$

This implies for each $\mathbf{v}^\sharp \in \Upsilon^\circ$ and $k \geq 1$ by the Jensen inequality and (H.14)

$$\begin{aligned} \exp\left\{\frac{\lambda}{\mathbf{r}_{k-1}} \xi_k(\mathbf{v}^\sharp)\right\} &\leq \int_{\mathcal{B}_k(\mathbf{v}^\sharp)} \left(\int_{\mathcal{B}_{k-1}(\mathbf{v})} \exp \frac{\lambda |\mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}')|}{d(\mathbf{v}, \mathbf{v}')} \frac{\pi(d\mathbf{v}')}{\pi_{k-1}(\mathbf{v})} \right) \frac{\pi(d\mathbf{v})}{\pi_k(\mathbf{v}^\sharp)} \\ &\leq \int_{\Upsilon^\circ} \left(\int_{\mathcal{B}_{k-1}(\mathbf{v})} \exp \frac{\lambda |\mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}')|}{d(\mathbf{v}, \mathbf{v}')} \frac{\pi(d\mathbf{v}')}{\pi_{k-1}(\mathbf{v})} \right) \frac{\pi(d\mathbf{v})}{\pi_k^\circ(\mathbf{v})}. \end{aligned}$$

As the right hand-side does not depend on \mathbf{v}^\sharp , this yields for $\xi_k^* \stackrel{\text{def}}{=} \sup_{\mathbf{v} \in \Upsilon^\circ} \xi_k(\mathbf{v})$ by condition **(Ed)** in view of $e^{|x|} \leq e^x + e^{-x}$

$$\begin{aligned} \mathbb{E} \exp\left(\frac{\lambda}{\mathbf{r}_{k-1}} \xi_k^*\right) &\leq \int_{\Upsilon^\circ} \left(\int_{\mathcal{B}_{k-1}(\mathbf{v})} \mathbb{E} \exp \frac{\lambda |\mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}')|}{d(\mathbf{v}, \mathbf{v}')} \frac{\pi(d\mathbf{v}')}{\pi_{k-1}(\mathbf{v})} \right) \frac{\pi(d\mathbf{v})}{\pi_k^\circ(\mathbf{v})} \\ &\leq 2 \exp(\lambda^2/2) \int_{\Upsilon^\circ} \left(\int_{\mathcal{B}_{k-1}(\mathbf{v})} \frac{\pi(d\mathbf{v}')}{\pi_{k-1}(\mathbf{v})} \right) \frac{\pi(d\mathbf{v})}{\pi_k^\circ(\mathbf{v})} \\ &= 2\mathbb{N}_k \exp(\lambda^2/2), \quad k \geq 1. \end{aligned}$$

Further, the use of $d(\mathbf{v}, \mathbf{v}^\circ) \leq \mathbf{r}_0$ for all $\mathbf{v} \in \Upsilon^\circ$ yields by **(Ed)**

$$\mathbb{E} \exp\left\{\frac{\lambda}{\mathbf{r}_0} |\mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}^\circ)|\right\} \leq 2 \exp(\lambda^2/2) \tag{H.15}$$

and thus

$$\begin{aligned} \mathbb{E} \exp\left\{\frac{\lambda}{\mathbf{r}_0} |\mathbb{S}_0 \mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}^\circ)|\right\} &\leq \frac{1}{\pi_0(\mathbf{v})} \int_{\mathcal{B}_0(\mathbf{v})} \mathbb{E} \exp\left\{\frac{\lambda}{\mathbf{r}_0} |\mathcal{U}(\mathbf{v}') - \mathcal{U}(\mathbf{v}^\circ)|\right\} \pi(d\mathbf{v}') \\ &\leq \frac{M_0}{\pi(\Upsilon^\circ)} \int_{\Upsilon^\circ} \mathbb{E} \exp\left\{\frac{\lambda}{\mathbf{r}_0} |\mathcal{U}(\mathbf{v}') - \mathcal{U}(\mathbf{v}^\circ)|\right\} \pi(d\mathbf{v}'). \end{aligned}$$

This implies by (H.15) for $\xi_0^* \equiv \sup_{\mathbf{v} \in \Upsilon^\circ} |\mathbb{S}_0 \mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}^\circ)|$

$$\mathbb{E} \exp\left(\frac{\lambda}{\mathbf{r}_0} \xi_0^*\right) \leq 2M_0 \exp(\lambda^2/2).$$

Denote $c_0 = 1/3$ and $c_k = \mathbf{r}_{k-1}/(3\mathbf{r}_0) = 2^{-k+1}/3$ for $k \geq 1$. Then $\sum_{k=0}^\infty c_k = 1$ and the results follow from Lemma H.1.1 below with $\zeta_k = \xi_k^*/r_{k-1}$ and $q_k = \sqrt{2\mathbb{N}_k}$.

H.4 Generic chaining with partitioning

This section presents a slightly modified construction based on the notion of *partition*. Let $\mathbf{r}_k \rightarrow 0$ be a given sequence, and let for each k , the set Υ° is split into a collection \mathcal{F}_k of subsets C . A collection of partitions \mathcal{F}_k is called *admissible* if $\text{diam}(C) \leq \mathbf{r}_k$ for all $C \in \mathcal{F}_k$ and every next partition \mathcal{F}_{k+1} is a refinement of the previous one \mathcal{F}_k :

$$\forall C_{k+1} \in \mathcal{F}_{k+1} \quad \exists C_k \in \mathcal{F}_k \text{ such that } C_{k+1} \subset C_k.$$

The generic chaining approach can be formulated via such partitions. Let $\pi(\cdot)$ be a given measure. Define

$$\mathbb{N}_k \stackrel{\text{def}}{=} \sup_{C \in \mathcal{F}_k} \frac{\pi(\Upsilon^\circ)}{\pi(C)}. \quad (\text{H.16})$$

Now define $\mathbb{Q}_1(\Upsilon^\circ)$ and $\mathbb{Q}_2(\Upsilon^\circ)$ by (H.1).

Theorem H.4.1. *Let \mathcal{U} be a separable process following to (Ed) and Υ° be a d -ball in Υ with the center \mathbf{v}° and the radius \mathbf{r}_0 , i.e. $d(\mathbf{v}, \mathbf{v}^\circ) \leq \mathbf{r}_0$ for all $\mathbf{v} \in \Upsilon^\circ$. Let $\mathbb{H}_1 = \mathbb{Q}_1(\Upsilon^\circ)$ and $\mathbb{H}_2 = \mathbb{Q}_2(\Upsilon^\circ)$ be defined in (H.1) with \mathbb{N}_k given by (H.16) for an admissible partitioning (\mathcal{F}_k) . Then all the statements of Theorem H.1.1 continue to apply.*

H.5 A large deviation bound

Due to the result of Theorem H.1.1, the bound for the maximum of $\mathcal{U}(\mathbf{v}, \mathbf{v}^*)$ over $\mathbf{v} \in \mathcal{B}_r(\mathbf{v}^*)$ grows linearly in r . So, its applications to situations with $r \gg \mathbb{Q}_1(\Upsilon^\circ)$ are limited. The next result shows that introducing a negative drift helps to state a uniform in r local probability bound. Namely, the bound for the process $\mathcal{U}(\mathbf{v}, \mathbf{v}^*) - f(d(\mathbf{v}, \mathbf{v}^*))$ for some function $f(r)$ over a ball $\mathcal{B}_r(\mathbf{v}^*)$ around the point \mathbf{v}^* does not depend on r . Here the generic chaining arguments are accomplished with the slicing technique. The idea is for a given $r^* > 1$ to split the ball $\mathcal{B}_{r^*}(\mathbf{v}^*)$ into the slices $\mathcal{B}_{r_k}(\mathbf{v}^*) \setminus \mathcal{B}_{r_{k-1}}(\mathbf{v}^*)$ and to apply Theorem H.1.1 to each slice separately.

Theorem H.5.1. *Let r^* be such that (Ed) holds on $\mathcal{B}_{r^*}(\mathbf{v}^*)$. Let also $\mathbb{Q}_1(\mathcal{B}_r(\mathbf{v}^*)) \leq \mathbb{H}_1$ and $\mathbb{Q}_2(\mathcal{B}_r(\mathbf{v}^*)) \leq \mathbb{H}_2$ for $r \leq r^*$. Given $\mathbf{r}_0 < r^*$, let a monotonous function $f(r, \mathbf{r}_0)$ fulfill for some $\rho < 1$*

$$f(r, \mathbf{r}_0) \geq \nu_0 \mathbf{r} \mathfrak{z}_{\mathbb{H}}(x + \log(r/\mathbf{r}_0)), \quad \mathbf{r}_0 \leq r \leq r^*, \quad (\text{H.17})$$

where the function $\mathfrak{z}_{\mathbb{H}}(\cdot)$ is given by (H.3). Then it holds

$$\mathbb{I}P\left(\sup_{r_0 \leq r \leq r^*} \sup_{v \in \mathcal{B}_r(v^*)} \{\mathcal{U}(v, v^*) - f(\rho^{-1}r, r_0)\} \geq 0\right) \leq \frac{\rho}{1-\rho} e^{-x}.$$

Remark H.5.1. Formally the bound applies even with $r^* = \infty$ provided that **(Ed)** is fulfilled on the whole set Υ° .

Remark H.5.2. If $g = \infty$, then $\mathfrak{z}_{\mathbb{H}}(x) = 2\mathbb{H}_1 + \sqrt{8x}$ and the condition (H.17) on the drift simplifies to $(\nu_0 r)^{-1} f(r, r_0) \geq 2\mathbb{H}_1 + \sqrt{8x + 8 \log(2r/r_0)}$.

Proof. By (H.17) and Theorem H.1.1 for any $r > r_0$

$$\begin{aligned} & \mathbb{I}P\left(\sup_{v \in \mathcal{B}_r(v^*) \setminus \mathcal{B}_{\rho r}(v^*)} \{\mathcal{U}(v, v^*) - f(r, r_0)\} \geq 0\right) \\ & \leq \mathbb{I}P\left(\frac{1}{\nu_0 r} \sup_{v \in \mathcal{B}_r(v^*)} \mathcal{U}(v, v^*) \geq \mathfrak{z}(x + \log(r/r_0))\right) \leq \frac{r_0}{r} e^{-x}. \end{aligned} \quad (\text{H.18})$$

Now defined $r_k = r_0 \rho^{-k}$ for $k = 0, 1, 2, \dots$. Define also $k^* \stackrel{\text{def}}{=} \log(r^*/r_0) + 1$. It follows from (H.18) that

$$\begin{aligned} & \mathbb{I}P\left(\sup_{v \in \mathcal{B}_{r^*}(v^*) \setminus \mathcal{B}_{r_0}(v^*)} \{\mathcal{U}(v, v^*) - f(\rho^{-1}d(v, v^*), r_0)\} \geq 0\right) \\ & \leq \sum_{k=1}^{k^*} \mathbb{I}P\left(\frac{1}{r_k} \sup_{v \in \mathcal{B}_{r_k}(v^*) \setminus \mathcal{B}_{r_{k-1}}(v^*)} \{\mathcal{U}(v, v^*) - f(r_k, r_0)\} \geq 0\right) \\ & \leq e^{-x} \sum_{k=1}^{k^*} \rho^k \leq \frac{\rho}{1-\rho} e^{-x} \end{aligned}$$

as required.

H.6 Finite-dimensional smooth case

Here we discuss the special case when Υ is an open subset in \mathbb{R}^p , the stochastic process $\mathcal{U}(v)$ is Fréchet differentiable and its gradient $\nabla \mathcal{U}(v) \stackrel{\text{def}}{=} d\mathcal{U}(v)/dv$ has bounded exponential moments.

(ED) *There exist $g > 0$, $\nu_0 \geq 1$, and for each $v \in \Upsilon$, a symmetric non-negative matrix $\mathbb{V}(v)$ such that for any $\lambda \leq g$ and any unit vector $\gamma \in \mathbb{R}^p$, it holds*

$$\log \mathbb{E} \exp\left\{\lambda \frac{\gamma^\top \nabla \mathcal{U}(v)}{\|\mathbb{V}(v)\gamma\|}\right\} \leq \frac{\nu_0^2 \lambda^2}{2}.$$

A natural candidate for $\mathbb{V}^2(v)$ is the covariance matrix $\text{Var}(\nabla \mathcal{U}(v))$ provided that this matrix is well posed. Then the constant ν_0 can be taken close to one by reducing the value g .

In what follows we fix a subset Υ° of Υ and establish a bound for the maximum of the process $\mathcal{U}(\mathbf{v}, \mathbf{v}^\circ) = \mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}^\circ)$ on Υ° for a fixed point \mathbf{v}° . We assume existence of a matrix $\mathbb{V} = \mathbb{V}(\Upsilon^\circ)$ such that $\mathbb{V}(\mathbf{v}) \preceq \mathbb{V}$ for all $\mathbf{v} \in \Upsilon^\circ$. We also assume that π is the Lebesgue measure on Υ . First we show that the differentiability condition **(ED)** implies **(Ed)**.

Lemma H.6.1. *Assume that **(ED)** holds with some g and $\mathbb{V}(\mathbf{v}) \preceq \mathbb{V}$ for $\mathbf{v} \in \Upsilon^\circ$. Consider any $\mathbf{v}, \mathbf{v}^\circ \in \Upsilon^\circ$. Then it holds for $|\lambda| \leq g$*

$$\log I\!\!E \exp \left\{ \lambda \frac{\mathcal{U}(\mathbf{v}, \mathbf{v}^\circ)}{\|\mathbb{V}(\mathbf{v} - \mathbf{v}^\circ)\|} \right\} \leq \frac{\nu_0^2 \lambda^2}{2}.$$

Proof. Denote $\delta = \|\mathbf{v} - \mathbf{v}^\circ\|$, $\boldsymbol{\gamma} = (\mathbf{v} - \mathbf{v}^\circ)/\delta$. Then

$$\mathcal{U}(\mathbf{v}, \mathbf{v}^\circ) = \delta \boldsymbol{\gamma}^\top \int_0^1 \nabla \mathcal{U}(\mathbf{v}^\circ + t\delta \boldsymbol{\gamma}) dt$$

and $\|\mathbb{V}(\mathbf{v} - \mathbf{v}^\circ)\| = \delta \|\mathbb{V} \boldsymbol{\gamma}\|$. Now the Hölder inequality and **(ED)** yield

$$\begin{aligned} & I\!\!E \exp \left\{ \lambda \frac{\mathcal{U}(\mathbf{v}, \mathbf{v}^\circ)}{\|\mathbb{V}(\mathbf{v} - \mathbf{v}^\circ)\|} - \frac{\nu_0^2 \lambda^2}{2} \right\} \\ &= I\!\!E \exp \left\{ \int_0^1 \left[\lambda \frac{\boldsymbol{\gamma}^\top \nabla \mathcal{U}(\mathbf{v}^\circ + t\delta \boldsymbol{\gamma})}{\|\mathbb{V} \boldsymbol{\gamma}\|} - \frac{\nu_0^2 \lambda^2}{2} \right] dt \right\} \\ &\leq \int_0^1 I\!\!E \exp \left\{ \lambda \frac{\boldsymbol{\gamma}^\top \nabla \mathcal{U}(\mathbf{v}^\circ + t\delta \boldsymbol{\gamma})}{\|\mathbb{V} \boldsymbol{\gamma}\|} - \frac{\nu_0^2 \lambda^2}{2} \right\} dt \leq 1 \end{aligned}$$

as required.

The result of Lemma H.6.1 enables us to define $d(\mathbf{v}, \mathbf{v}') = \|\mathbb{V}(\mathbf{v} - \mathbf{v}')\|$ so that the corresponding d -ball coincides with the following ellipsoidal set $\mathcal{B}(\mathbf{r}, \mathbf{v}^\circ)$:

$$\mathcal{B}(\mathbf{r}, \mathbf{v}^\circ) \stackrel{\text{def}}{=} \{ \mathbf{v} : \|\mathbb{V}(\mathbf{v} - \mathbf{v}^\circ)\| \leq \mathbf{r} \}.$$

H.6.1 Covering and entropy for Euclidean distance

Now we bound the value $\mathbb{Q}(\Upsilon^\circ)$ for $\Upsilon^\circ = \mathcal{B}(\mathbf{r}, \mathbf{v}^\circ)$. Note that by change of variable one can reduce the study to the case $\mathbb{V} = I_p$ and consider the entropy of the unit ball in \mathbb{R}^p w.r.t. the Euclidean distance. We use the following general result which allows to upperbound the covering number of a convex set in \mathbb{R}^p for the Euclidean metric.

Lemma H.6.2. *Let Υ° be a convex set in \mathbb{R}^p , $\delta > 0$, and B be the unit ball in \mathbb{R}^p . Then the covering number $\mathbb{N}(\Upsilon^\circ, \delta)$ fulfills*

$$\mathbb{N}(\Upsilon^\circ, \delta) \leq \frac{\text{vol}(\Upsilon^\circ + (\delta/2)B)}{\text{vol}(B)} (2/\delta)^p.$$

Proof. Let $(\mathbf{v}^{(i)}, i = 1, \dots, \mathbb{N})$ be a maximal subset of Υ° such that $\|\mathbf{v}^{(i)} - \mathbf{v}^{(j)}\| \geq \delta$ for all $i \neq j$. By maximality, $(\mathbf{v}^{(i)})$ is a δ -net of Υ° . Let also B be the unit ball in \mathbb{R}^p . Note that the balls $\mathbf{v}^{(i)} + (\delta/2)B$ are disjoint and included in $\Upsilon^\circ + (\delta/2)B$. Therefore,

$$\sum_{i \leq \mathbb{N}} \text{vol}\left(\mathbf{v}^{(i)} + \frac{\delta}{2}B\right) \leq \text{vol}\left(\Upsilon^\circ + \frac{\delta}{2}B\right),$$

where $\text{vol}(A)$ means the Lebesgue measure of the set A . This yields

$$\mathbb{N}(\delta/2)^p \text{vol}(B) \leq \text{vol}(\Upsilon^\circ + (\delta/2)B)$$

and the claim of the lemma follows.

Lemma H.6.3 (Entropy of a ball). *Let $\Upsilon^\circ = \mathcal{B}(\mathbf{r}_o, \mathbf{v}^*)$ and $\mathbf{r}_k = 2^{-k}\mathbf{r}_o$. Then the covering numbers \mathbb{N}_k fulfill with $\delta = \mathbf{r}_k/\mathbf{r}_o = 2^{-k}$*

$$\mathbb{N}_k \leq (1 + 2/\delta)^p = (1 + 2^{k+1})^p.$$

Moreover, with $\mathfrak{c}_2 = 4.67$,

$$\mathbb{Q}_2(\Upsilon^\circ) \leq 2 \log 2 + \mathfrak{c}_2 p \leq 6p,$$

$$\mathbb{Q}_1(\Upsilon^\circ) \leq \sqrt{2 \log 2 + \mathfrak{c}_2 p} \leq \sqrt{6p}.$$

Proof. A change of variable reduces the statement to the case $\mathbb{V} = I_p$ and $\mathbf{r}_o = 1$. For $\delta = 2^{-k}$, this implies by Lemma H.6.2 in view of $\Upsilon^\circ = B$

$$\text{vol}\left(\Upsilon^\circ + \frac{\delta}{2}B\right) = (1 + \delta/2)^p \text{vol}(B),$$

that $\mathbb{N}_k \leq (1 + 2/\delta)^p$ as claimed. Now we derive

$$\begin{aligned} \mathbb{Q}_2(\Upsilon^\circ) &\leq \sum_{k=1}^{\infty} 2^{-k+1} \log(2\mathbb{N}_k) \leq \sum_{k=1}^{\infty} 2^{-k+1} \{\log 2 + 2p \log(1 + 2^{k+1})\} \\ &\leq 2 \log 2 + p \sum_{k=0}^{\infty} 2^{-k+1} \log(1 + 2^k) \\ &\leq 2 \log 2 + \mathfrak{c}_2 p \end{aligned}$$

as required.

Now we specify the local bounds of Theorem H.1.1 to the smooth case. We consider the local sets of the elliptic form $\Upsilon_o(\mathbf{r}) \stackrel{\text{def}}{=} \{\mathbf{v} : \|\mathbb{V}(\mathbf{v} - \mathbf{v}^*)\| \leq \mathbf{r}\}$, where \mathbb{V} dominates $\mathbb{V}(\mathbf{v})$ on this set: $\mathbb{V}(\mathbf{v}) \preceq \mathbb{V}$.

Theorem H.6.1. *Let **(ED)** hold with some $\mathfrak{g} > 0$, and matrices $\mathbb{V}(\mathbf{v})$ such that $\mathbb{V}(\mathbf{v}) \preceq \mathbb{V}$ for all $\mathbf{v} \in \Upsilon_o(\mathbf{r})$ and a fixed \mathbf{r} . For any $\mathbf{x} \geq 1/2$*

$$\mathbb{P}\left\{\frac{1}{\nu_0 \mathbf{r}} \sup_{\mathbf{v} \in \Upsilon^\circ(\mathbf{r})} |\mathcal{U}(\mathbf{v}, \mathbf{v}^*)| \geq \mathfrak{z}_{\mathbb{H}}(\mathbf{x})\right\} \leq e^{-\mathbf{x}},$$

where $\mathfrak{z}_{\mathbb{H}}(\mathbf{x})$ is given for $\mathbf{g} = \infty$ by

$$\mathfrak{z}_{\mathbb{H}}(\mathbf{x}) \stackrel{\text{def}}{=} 2\sqrt{6p} + \sqrt{8\mathbf{x}}.$$

and for \mathbf{g} finite by any of two formulas

$$\begin{aligned} \mathfrak{z}_{\mathbb{H}}(\mathbf{x}) &= 2\sqrt{6p} + \sqrt{8\mathbf{x}} + 12\mathbf{g}^{-1}(\mathbf{g}^{-2}\mathbf{x} + 1)p, \\ \mathfrak{z}_{\mathbb{H}}(\mathbf{x}) &= \begin{cases} 2\sqrt{6p + 2\mathbf{x}}, & \text{if } 6p + 2\mathbf{x} \leq \mathbf{g}^2, \\ 2\mathbf{g}^{-1}\mathbf{x} + 6\mathbf{g}^{-1}p + \mathbf{g}, & \text{if } 6p + 2\mathbf{x} > \mathbf{g}^2. \end{cases} \end{aligned} \quad (\text{H.19})$$

Proof. Lemma H.6.3 implies **(Ed)** with $d(\mathbf{v}, \mathbf{v}^*) = \|\mathbb{V}(\mathbf{v} - \mathbf{v}^*)\|$. Now the result follows from Theorem H.1.1.

H.6.2 Generic chaining

Lemma H.6.4 (Entropy for generic chaining). *Let $\Upsilon^\circ = \mathcal{B}(\mathbf{r}_\circ, \mathbf{v}^\circ)$. Under the conditions of Lemma H.6.1, it holds $\mathbb{G}_2(\Upsilon^\circ) \leq \mathfrak{c}_1 p$, where $\mathfrak{c}_1 = 2$ for $p \geq 2$, and $\mathfrak{c}_1 = 2.7$ for $p = 1$.*

Proof. The set Υ° coincides with the ellipsoid $\mathcal{B}(\mathbf{r}_\circ, \mathbf{v}^\circ)$ while the d -ball $\mathcal{B}_k(\mathbf{v})$ coincides with the ellipsoid $\mathcal{B}(\mathbf{r}_k, \mathbf{v})$ for each $k \geq 0$. By change of variables, the study can be reduced to the case with $\mathbf{v}^\circ = 0$, $\mathbb{V} \equiv I_p$, $\mathbf{r}_0 = 1$, so that $\mathcal{B}(\mathbf{r}, \mathbf{v})$ is the usual Euclidean ball in \mathbb{R}^p of radius \mathbf{r} . It is obvious that the measure of the overlap of two balls $\mathcal{B}(1, 0)$ and $\mathcal{B}(2^{-k}, \mathbf{v})$ for $\|\mathbf{v}\| \leq 1$ is minimized when $\|\mathbf{v}\| = 1$, and this value is the same for all such \mathbf{v} . For $k = 0$, the overlap of $\mathcal{B}(1, 0)$ and $\mathcal{B}(1, \mathbf{v})$ contains the ball $\mathcal{B}(1/2, \mathbf{v}/2)$ of radius $1/2$, so that $M_0 \leq 2^p$. For $k \geq 1$, we use the following observation. Fix \mathbf{v}^\sharp with $\|\mathbf{v}^\sharp\| = 1$. Let $\mathbf{r} \leq 1$, $\mathbf{v}^\flat = (1 - \mathbf{r}^2/2)\mathbf{v}^\sharp$ and $\mathbf{s} = \mathbf{r} - \mathbf{r}^2/2$. If $\mathbf{v} \in \mathcal{B}(\mathbf{s}, \mathbf{v}^\flat)$, then $\mathbf{v} \in \mathcal{B}(\mathbf{r}, \mathbf{v}^\sharp)$ because

$$\|\mathbf{v}^\sharp - \mathbf{v}\| \leq \|\mathbf{v}^\sharp - \mathbf{v}^\flat\| + \|\mathbf{v}^\flat - \mathbf{v}\| \leq \mathbf{r}^2/2 + \mathbf{r} - \mathbf{r}^2/2 \leq \mathbf{r}.$$

Moreover, for each $\mathbf{v} \in \mathcal{B}(\mathbf{s}, \mathbf{v}^\flat)$, it holds with $\mathbf{u} = \mathbf{v} - \mathbf{v}^\flat$

$$\|\mathbf{v}\|^2 = \|\mathbf{v}^\flat\|^2 + \|\mathbf{u}\|^2 + 2\mathbf{u}^\top \mathbf{v}^\flat \leq (1 - \mathbf{r}^2/2)^2 + |\mathbf{s}|^2 + 2\mathbf{u}^\top \mathbf{v}^\flat \leq 1 + 2\mathbf{u}^\top \mathbf{v}^\flat.$$

This means that either $\mathbf{v} = \mathbf{v}^\flat + \mathbf{u}$ or $\mathbf{v}^\flat - \mathbf{u}$ belongs to the ball $\mathcal{B}(\mathbf{r}_0, \mathbf{v}^\circ)$ and thus, $\pi(\mathcal{B}(1, 0) \cap \mathcal{B}(\mathbf{r}, \mathbf{v})) \geq \pi(\mathcal{B}(\mathbf{s}, \mathbf{v}^\flat))/2$. We conclude that

$$\frac{\pi(\mathcal{B}(1, 0))}{\pi(\mathcal{B}(1, 0) \cap \mathcal{B}(\mathbf{r}, \mathbf{v}^\sharp))} \leq \frac{2\pi(\mathcal{B}(1, 0))}{\pi(\mathcal{B}(\mathbf{s}, 0))} = 2(\mathbf{r} - \mathbf{r}^2/2)^{-p}.$$

This implies for $k \geq 0$ and $\mathbf{r}_k = 2^{-k}$ that $2\mathbb{N}_k \leq 2^{2+kp}(1 - 2^{-k-1})^{-p}$. The quantity $\mathbb{Q}_2(\Upsilon^\circ)$ can now be evaluated as

$$\begin{aligned} \frac{1}{2}\mathbb{Q}_2(\Upsilon^\circ) &\leq \frac{1}{3}\log(2^{1+p}) + \frac{2}{3}\sum_{k=1}^{\infty} 2^{-k}\log(2^{2+kp}) - \frac{2p}{3}\sum_{k=1}^{\infty} 2^{-k}\log(1 - 2^{-k-1}) \\ &= \frac{\log 2}{3}\left[1 + p + 2\sum_{k=1}^{\infty}(2+kp)2^{-k}\right] - \frac{2p}{3}\sum_{k=1}^{\infty} 2^{-k}\log(1 - 2^{-k-1}) \leq \mathfrak{c}_1 p, \end{aligned}$$

where $\mathfrak{c}_1 = 2$ for $p \geq 2$, and $\mathfrak{c}_1 = 2.7$ for $p = 1$, and the result follows.

H.7 Entropy of an ellipsoid

Let H be a positive self adjoint operator in \mathbb{R}^∞ . We are interested to describe the entropy of the elliptic set

$$\mathcal{E}_H(\mathbf{r}_\circ) \stackrel{\text{def}}{=} \{\mathbf{v} : \|H(\mathbf{v} - \mathbf{v}^\circ)\| \leq \mathbf{r}_\circ\} \quad (\text{H.20})$$

for given $\mathbf{v}^\circ \in \mathbb{R}^\infty$ and $\mathbf{r}_\circ > 0$ with respect to the usual Euclidean distance in \mathbb{R}^∞ . Below we evaluate the entropy of this set assuming that $\|H^{-1}\|_{\text{op}} = 1$ and H^{-2} is a trace operator, i.e., $h_1 = 1$ and

$$p_H \stackrel{\text{def}}{=} \text{tr}(H^{-2}) = \sum_{j=1}^{\infty} h_j^{-2} < \infty, \quad (\text{H.21})$$

where $h_1 \leq h_2 \leq \dots$ are the ordered eigenvalues of H .

Theorem H.7.1. *Suppose that for some $\alpha > 1$*

$$p_H(\alpha) \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} h_j^{-2} \log^\alpha(h_j^2) < \infty. \quad (\text{H.22})$$

Then for $\mathcal{E} = \mathcal{E}_H(\mathbf{r}_\circ)$

$$\mathbb{Q}_1(\mathcal{E}) \leq C(\alpha - 1)^{-1/2} \sqrt{p_H(\alpha)}, \quad (\text{H.23})$$

where C is an absolute constant. Furthermore,

$$\mathbb{Q}_2(\mathcal{E}) \leq Cp_H^* = C \sum_{j=1}^{\infty} h_j^{-1}.$$

Remark H.7.1. The log-factor in the definition of $p_H(\alpha)$ can be removed by using a more advanced generic chaining and majorising measure technique. However, in most of situations, the bound in terms of $p_H(\alpha)$ is also sharp.

The term p_H^* only appears in the sub-exponential case when $g < \infty$. In this case we need the condition $p_H^* < \infty$ which requires $\sum_j h_j^{-1} < \infty$, that is, a more rapid growth of the values h_j is necessary than in (H.22).

Proof. We begin by a general lemma which bounds the covering numbers for the elliptic set \mathcal{E} for the Euclidean distance.

Lemma H.7.1 (Entropy of the ellipsoid). *Let $\mathcal{E} = \mathcal{E}_H(\mathbf{r}_\circ)$ be an elliptic set from (H.20) with $\|H^{-1}\|_{\text{op}} = 1$ and $\text{tr}(H^{-2}) < \infty$. Let also $d(\mathbf{v}, \mathbf{v}') = \|\mathbf{v} - \mathbf{v}'\|$. Then for $\mathbf{r}_k = 2^{-k}\mathbf{r}_\circ$, the value $\mathbb{Q}_1(\mathcal{E})$ from (H.1) satisfies*

$$\mathbb{Q}_1(\mathcal{E}) \leq \sum_{k=1}^{\infty} 2^{-k} \sqrt{\log 2 + 2L_H(m_k)}, \quad (\text{H.24})$$

where m_k is the index j for which $h_{m_k}^2 = 2^{2k+1}$ and hence,

$$h_j^2 \leq 2^{2k+1}, \quad j \leq m_k, \quad (\text{H.25})$$

and

$$L_H(m) \stackrel{\text{def}}{=} \sum_{j=1}^m \log(3h_m/h_j).$$

Remark H.7.2. For the ease of presentation, we supposed in the lemma that for each $k \geq 1$, there exists some m_k with $h_{m_k} = 2^{k+1/2}$. The results easily extend to the case when this equality is approximate.

Proof. Without loss of generality assume $\mathbf{v}^\circ = 0$. A basis transform reduces the study to the case when H is diagonal:

$$H = \text{diag}\{h_1, h_2, \dots\}.$$

We only have to evaluate the covering numbers \mathbb{N}_k . Let us fix $k \geq 1$ and let m_k be given by (H.25). For any point $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2, \dots)^\top$ in \mathcal{E} , it holds

$$\begin{aligned} \sum_{j=m_k+1}^{\infty} \mathbf{v}_j^2 &= \sum_{j=m_k+1}^{\infty} h_j^{-2} h_j^2 \mathbf{v}_j^2 \\ &\leq h_{m_k+1}^{-2} \sum_{j=m_k+1}^{\infty} h_j^2 \mathbf{v}_j^2 \\ &\leq h_{m_k+1}^{-2} \sum_{j=1}^{\infty} h_j^2 \mathbf{v}_j^2 \leq 2^{-2k-1} \mathbf{r}_\circ^2 \leq \mathbf{r}_k^2/2. \end{aligned} \quad (\text{H.26})$$

Consider the elliptic set \mathcal{E}_k in \mathbb{R}^{m_k} obtained by projection Π_k of \mathcal{E} on the first m_k coordinates:

$$\mathcal{E}_k \stackrel{\text{def}}{=} \{(\mathbf{v}_1, \dots, \mathbf{v}_{m_k})^\top : \sum_{j=1}^{m_k} h_j^2 \mathbf{v}_j^2 \leq \mathbf{r}_\circ^2\}.$$

Let \mathcal{M}_k be a ϵ_k -net in \mathcal{E}_k for $\epsilon_k^2 = \mathbf{r}_k^2/2$. A \mathbf{r}_k -net in \mathcal{E} can be constructed from \mathcal{M}_k in a simple way: just fix to zero the remaining coordinates $\mathbf{v}_j = 0$ for $j > m_k$. If \mathbf{v}° is constructed in this way, then $\|H\mathbf{v}^\circ\| = \|H\Pi_k\mathbf{v}^\circ\| \leq 1$, that is, $\mathbf{v}^\circ \in \mathcal{E}$. Moreover, for any other point $\mathbf{v} \in \mathcal{E}$, take \mathbf{v}° such that their projections satisfy $\|\Pi_k(\mathbf{v} - \mathbf{v}^\circ)\| \leq \epsilon_k$. Then by (H.26)

$$\|\mathbf{v} - \mathbf{v}^\circ\|^2 = \|\Pi_k(\mathbf{v} - \mathbf{v}^\circ)\|^2 + \|(I - \Pi_k)\mathbf{v}\|^2 \leq \mathbf{r}_k^2/2 + \mathbf{r}_k^2/2 = \mathbf{r}_k^2.$$

Therefore, the covering number $\mathbb{N}(\mathcal{E}, \mathbf{r}_k)$ of the infinite dimensional elliptic set \mathcal{E} does not exceed the covering number $\mathbb{N}(\mathcal{E}_k, \epsilon_k)$ for the m_k -dimensional ellipsoid \mathcal{E}_k . By Lemma H.6.2 with $\delta = \epsilon_k$,

$$\mathbb{N}(\mathcal{E}_k, \epsilon_k) \leq \frac{\text{vol}(\mathcal{E}_k + (\epsilon_k/2)B_k)}{\text{vol}(B_k)} (2/\epsilon_k)^{m_k},$$

where B_k is the unit ball in \mathbb{R}^{m_k} . The bound $h_j^{-2} \geq 2^{-2k-1}$ for $j \leq m_k$ implies that $\mathcal{E}_k + (\epsilon_k/2)B_k$ is contained in the elliptic set $(3/2)\mathcal{E}_k$.

The definition implies due to $h_{m_k}^2 = 2^{2k+1}$

$$\begin{aligned} \mathbb{N}(\mathcal{E}, \mathbf{r}_k) &\leq \log \frac{\text{vol}((3/2)\mathcal{E}_k)}{(\epsilon_k/2)^{m_k} \text{vol}(B_k)} \\ &\leq \sum_{j=1}^{m_k} \log \frac{3h_j^{-1}}{\epsilon_k} \leq \sum_{j=1}^{m_k} \log(3h_{m_k}/h_j) = L_H(m_k). \end{aligned} \quad (\text{H.27})$$

Now the result (H.24) follows by the definition of $\mathbb{Q}_1(\mathcal{E})$.

Denote $\mathbb{N}_k = \mathbb{N}(\mathcal{E}, \mathbf{r}_k)$. By the Cauchy-Schwartz inequality for $\alpha > 1$

$$\mathbb{Q}_1(\mathcal{E}) = \sum_{k=1}^{\infty} 2^{-k} \sqrt{2 \log(2\mathbb{N}_k)} \leq \left\{ \sum_{k=1}^{\infty} k^{-\alpha} \sum_{k=1}^{\infty} k^{\alpha} 2^{-2k} 2 \log(2\mathbb{N}_k) \right\}^{1/2}. \quad (\text{H.28})$$

The use of $h_{m_\ell}^{-2} = 2^{2\ell+1}$ and $h_j^2 \geq 4h_{m_{\ell-1}}^2$ for $j \in (m_{\ell-1}, m_\ell]$ yields by (H.27) with $n_\ell \stackrel{\text{def}}{=} m_\ell - m_{\ell-1}$

$$2 \log(2\mathbb{N}_k) = \sum_{\ell=1}^k \sum_{j=m_{\ell-1}+1}^{m_\ell} \log \frac{9h_{m_k}^2}{h_j^2} \leq \sum_{\ell=1}^k \{k - \ell + \log(36)\} n_\ell$$

Further, in view of $h_{m_k} = 2^k$

$$\begin{aligned}
\sum_{k=1}^{\infty} k^{\alpha} 2^{-2k} 2\mathbb{N}_k &\leq \sum_{k=1}^{\infty} k^{\alpha} 2^{-2k} \sum_{\ell=1}^k \{k - \ell + \log(36)\} n_{\ell} \\
&= \sum_{\ell=1}^{\infty} \sum_{k \geq \ell} k^{\alpha} 2^{-2k} \{k - \ell + \log(36)\} n_{\ell} \\
&= \sum_{\ell=1}^{\infty} n_{\ell} 2^{-2\ell} \sum_{k \geq \ell} k^{\alpha} 2^{-2(k-\ell)} \{k - \ell + \log(36)\} \\
&= C \sum_{\ell=1}^{\infty} n_{\ell} 2^{-2\ell} \ell^{\alpha}.
\end{aligned}$$

It remains to note that $2^{2\ell-1} \leq h_j^2 \leq 2^{2\ell+1}$ for $m_{\ell-1} < j \leq m_{\ell}$ and

$$\sum_{\ell=1}^{\infty} n_{\ell} 2^{-2\ell} \ell^{\alpha} \leq \sum_{\ell=1}^{\infty} \sum_{j=m_{\ell-1}+1}^{m_{\ell}} h_j^{-2} \log^{\alpha}(h_j^2) = \sum_{j=1}^{\infty} h_j^{-2} \log^{\alpha}(h_j^2) = p_H(\alpha). \quad (\text{H.29})$$

The assertion (H.23) now follows from (H.28) in view of $\sum_{k \geq 1} k^{-\alpha} \leq C(\alpha - 1)^{-1}$.

The result on $\mathbb{Q}_2(\mathcal{E})$ requires to bound the sum of $2^{-k} \log \mathbb{N}_k$. Similarly to the above, one easily derives

$$\begin{aligned}
\sum_{k=1}^{\infty} 2^{-k} \mathbb{N}_k &\leq \sum_{k=1}^{\infty} 2^{-k} \sum_{\ell=1}^k \{k - \ell + \log(36)\} n_{\ell} \\
&= \sum_{\ell=1}^{\infty} \sum_{k \geq \ell} 2^{-k} \{k - \ell + \log(36)\} n_{\ell} \\
&= \sum_{\ell=1}^{\infty} n_{\ell} 2^{-\ell} \sum_{k \geq \ell} 2^{-(k-\ell)} \{k - \ell + \log(36)\} \\
&= C \sum_{\ell=1}^{\infty} n_{\ell} 2^{-\ell} \leq C \sum_{j=1}^{\infty} h_j^{-1} = Cp_H^*.
\end{aligned}$$

Theorem is proved.

Now we present a special case for which the entropy can be bounded via the effective dimension p_H of Υ° defined in (H.21).

Theorem H.7.2. *Let $h_j^2 = f(j)$ for a monotonously increasing smooth function $f(x) > 0$. If $xf'(x)/f(x) \leq \beta$, then*

$$\begin{aligned}
\mathbb{Q}_2(\mathcal{E}) &\leq C\beta p_H, \\
\mathbb{Q}_1(\mathcal{E}) &\leq C\sqrt{\beta p_H},
\end{aligned} \quad (\text{H.30})$$

where the effective dimension p_H is defined in (H.21).

Proof. Obviously

$$\sum_{j=1}^m \log\left(\frac{h_m^2}{h_j^2}\right) \leq \int_0^m \log\left(\frac{f(m)}{f(t)}\right) dt.$$

Now we note that the function

$$F(x) \stackrel{\text{def}}{=} \int_0^x \log\left(\frac{f(x)}{f(t)}\right) dt$$

fulfills $F(0) = 0$ and

$$F'(x) = \frac{x f'(x)}{f(x)}.$$

yielding

$$\sum_{j=1}^m \log\left(\frac{h_m^2}{h_j^2}\right) \leq \int_0^m \log\left(\frac{f(m)}{f(t)}\right) dt = \int_0^m \frac{x f'(x)}{f(x)} dx.$$

Moreover, In particular, if $F'(x) \leq \beta$, then $F(x) \leq \beta x$ and thus, $L_H(m_k) \leq \beta m_k$. Now it holds similarly to (H.29)

$$\sum_{k=1}^{\infty} 2^{-k} m_k = \sum_{k=1}^{\infty} 2^{-k} \sum_{\ell=1}^k n_{\ell} \leq \sum_{\ell=1}^{\infty} n_{\ell} \sum_{k \geq \ell}^{\infty} 2^{-k} = \sum_{\ell=1}^{\infty} n_{\ell} 2^{-\ell} \leq 2 \sum_{j=1}^{\infty} h_j^{-2} = 2 p_H,$$

and the statement (H.23) follows.

Now we evaluate the entropy for the cases when h_j grow polynomially.

Theorem H.7.3. Let $h_j^2 = 1 + \varkappa^2 j^{2\beta}$ for $\beta > 1/2$ and some small value \varkappa . Then

$$\mathbb{Q}_1(\mathcal{E}) \leq C(2\beta - 1)^{-1/2} \varkappa^{-1/(2\beta)},$$

$$\mathbb{Q}_2(\mathcal{E}) \leq C(2\beta - 1)^{-1} \varkappa^{-1/\beta},$$

where C is an absolute constant.

Proof. For $f(x) = 1 + \varkappa^2 x^{2\beta}$, it holds $x f'(x)/f(x) \leq 2\beta$ and we can apply the result of Theorem H.7.2. With $\beta > 1/2$, the effective dimension p_H from (H.21) fulfills

$$p_H \leq \sum_{j=1}^{\infty} h_j^{-2} = \sum_{j=1}^{\infty} \frac{1}{1 + \varkappa^2 j^{2\beta}} \leq \int_0^{\infty} \frac{1}{1 + \varkappa^2 x^{2\beta}} dx = C \varkappa^{-1/\beta} \frac{1}{2\beta - 1}$$

and the result follows by (H.30).

H.8 Roughness constraints for dimension reduction

The local bounds of Theorems H.1.1 and H.5.1 can be extended in several directions. Here we briefly discuss one extension related to the use of a smoothness condition on the parameter \mathbf{v} . Let $\text{pen}(\mathbf{v})$ be a non-negative *penalty* function on Υ . A particular example of such penalty function is the *roughness penalty* $\text{pen}(\mathbf{v}) = \|G\mathbf{v}\|^2$ for a given p -matrix G^2 . Let \mathbf{r} be fixed. Consider the intersection of the ball $\mathcal{B}_{\mathbf{r}}(\mathbf{v}^\circ)$ with the set Υ given by the constraint $\text{pen}(\mathbf{v}) \leq 1$:

$$\Upsilon_{\text{pen}}(\mathbf{r}) = \{\mathbf{v} \in \Upsilon : d(\mathbf{v}, \mathbf{v}^\circ) \leq \mathbf{r}; \text{pen}(\mathbf{v}) \leq 1\},$$

for a fixed central point \mathbf{v}° and the radius \mathbf{r} . Here and below we assume that the central point \mathbf{v}° is “smooth” in the sense that $\text{pen}(\mathbf{v}^\circ) < 1$. One can easily check that the results of Theorems H.1.1 and H.5.1 and their corollaries extend to this situation without any change. The only difference is in the definition of the values $\mathbb{Q}_1(\Upsilon_\circ)$ and $\mathbb{Q}_2(\Upsilon_\circ)$ for $\Upsilon_\circ = \Upsilon_{\text{pen}}(\mathbf{r})$. Examples below show that the use of the penalization can substantially reduce these values relative to the non-penalized case.

We consider the case of a smooth process \mathcal{U} given on a local set $\Upsilon_G(\mathbf{r})$ of the form

$$\Upsilon_G(\mathbf{r}) = \{\mathbf{v} \in \Upsilon : \|\mathbb{V}(\mathbf{v} - \mathbf{v}^\circ)\| \leq \mathbf{r}; \|G\mathbf{v}\| \leq 1\}, \quad (\text{H.31})$$

with the distance $d(\mathbf{v}, \mathbf{v}^\circ) = \|\mathbb{V}(\mathbf{v} - \mathbf{v}^\circ)\|$ and a smoothness constraint $\|G\mathbf{v}\|^2 \leq 1$. Then the set $\Upsilon_G(\mathbf{r})$ is contained in an elliptic set

$$\Upsilon_\circ \stackrel{\text{def}}{=} \{\boldsymbol{\theta} : \|G\mathbf{v}\|^2 + \|\mathbb{V}(\mathbf{v} - \mathbf{v}^\circ)\|^2 \leq 1 + \mathbf{r}^2\}. \quad (\text{H.32})$$

Define

$$\mathbb{V}_G^2 = \mathbb{V}^2 + G^2,$$

$$\mathbf{v}_G = \mathbb{V}_G^{-2} \mathbb{V}^2 \mathbf{v}^\circ.$$

Then

$$\mathbf{v}^\circ - \mathbf{v}_G = (I_p - \mathbb{V}_G^{-2} \mathbb{V}^2) \mathbf{v}^\circ = \mathbb{V}_G^{-2} G^2 \mathbf{v}^\circ,$$

and one can get by simple algebra

$$\begin{aligned} \|G\mathbf{v}\|^2 + \|\mathbb{V}(\mathbf{v} - \mathbf{v}^\circ)\|^2 &= \|\mathbb{V}_G(\mathbf{v} - \mathbf{v}_G)\|^2 + \|G\mathbf{v}_G\|^2 + \|\mathbb{V}(\mathbf{v}_G - \mathbf{v}^\circ)\|^2 \\ &= \|\mathbb{V}_G(\mathbf{v} - \mathbf{v}_G)\|^2 + \mathbf{v}^\circ \mathbb{V}_G^{-2} G^2 \mathbb{V}_G^{-2} \mathbb{V}^2 \mathbf{v}^\circ = \|\mathbb{V}_G(\mathbf{v} - \mathbf{v}^\circ)\|^2 + d_G \end{aligned}$$

with $d_G = \mathbf{v}^{\circ\top} G^2 \mathbb{V}_G^{-2} \mathbb{V}^2 \mathbf{v}^{\circ} \leq \|G\mathbf{v}^{\circ}\|^2 < 1$. A change of variables $\mathbf{v} \rightarrow \mathbb{V}(\mathbf{v} - \mathbf{v}_G)$ allows us to reduce the study to the case of an ellipsoid considered in Section H.7. For H defined by $H^{-2} = \mathbb{V} \mathbb{V}_G^{-2} \mathbb{V}$, the set Υ_{\circ} from (H.32) is transferred into the elliptic set

$$\Upsilon_H(\mathbf{r}) = \{\mathbf{v}: \|H\mathbf{v}\|^2 \leq 1 + \mathbf{r}^2 - d_G\},$$

whose entropy for the Euclidean distance is given via the effective dimension $p_H = \text{tr}(H^{-2})$.

Now we are prepared to state the penalized bound for the process $\mathcal{U}(\cdot)$ over Υ_{\circ} which naturally generalizes the result of Theorem H.6.1 to the penalized case.

Theorem H.8.1. *Let $\Upsilon_{\circ} = \Upsilon_{\text{pen}}(\mathbf{r})$ be given by (H.31) and $\|G\mathbf{v}^{\circ}\| \leq 1$. Let also **(ED)** hold with some \mathbf{g} and a matrix $\mathbb{V}(\mathbf{v}) \preceq \mathbb{V}$ for all $\mathbf{v} \in \Upsilon_{\circ}$. For H defined by $H^{-2} = \mathbb{V} \mathbb{V}_G^{-2} \mathbb{V}$, let the entropy values $\mathbb{Q}_1(\Upsilon^{\circ})$ and $\mathbb{Q}_2(\Upsilon^{\circ})$ for the elliptic set $\Upsilon_H(\mathbf{r})$ from (H.32) be given in Section H.7. Then for any $\mathbf{x} \geq 1/2$*

$$I\!\!P\left\{\frac{1}{\nu_0 \mathbf{r}} \sup_{\mathbf{v} \in \Upsilon_{\text{pen}}(\mathbf{r})} |\mathcal{U}(\mathbf{v}, \mathbf{v}^{\circ})| \geq \mathfrak{z}_{\mathbb{H}}(\mathbf{x})\right\} \leq e^{-\mathbf{x}},$$

where $\mathfrak{z}_{\mathbb{H}}(\mathbf{x})$ is from (H.3) with these values $\mathbb{Q}_1(\Upsilon^{\circ})$ and $\mathbb{Q}_2(\Upsilon^{\circ})$.

H.9 Bound for a bivariate process

Consider a smooth bivariate process $\mathcal{U}(\mathbf{v}) = \mathcal{U}(\mathbf{v}_1, \mathbf{v}_2)$ over a product set $\Upsilon = \Upsilon_1 \times \Upsilon_2$, where $\Upsilon_j \subseteq \mathbb{R}^{p_j}$ for $j = 1, 2$. We suppose that partial derivatives of \mathcal{U} have uniform exponential moments.

(ED_p) *There exist $\mathbf{g} > 0$, $\nu_0 \geq 1$, and for each $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2) \in \Upsilon = \Upsilon_1 \times \Upsilon_2$, symmetric non-negative $p_j \times p_j$ matrices \mathbb{V}_j , $j = 1, 2$, such that for any $\lambda \leq \mathbf{g}$ and any unit vector $\boldsymbol{\gamma} \in \mathbb{R}^p$, it holds*

$$\log I\!\!E \exp\left\{\lambda \frac{\boldsymbol{\gamma}^{\top} \nabla_j \mathcal{U}(\mathbf{v})}{\|\mathbb{V}_j \boldsymbol{\gamma}\|}\right\} \leq \frac{\nu_0^2 \lambda^2}{2}, \quad j = 1, 2.$$

Here $\nabla_j \mathcal{U}$ denotes the partial derivative $\partial \mathcal{U} / \partial \mathbf{v}_j$ for $j = 1, 2$.

This allows to establish an exponential bound for the process $\mathcal{U}(\mathbf{v})$. Let us fix the central point $\mathbf{v}^{\circ} = (\mathbf{v}_1^{\circ}, \mathbf{v}_2^{\circ})$ and a radius \mathbf{r} . As usual,

$$\Upsilon_j(\mathbf{r}) = \{\mathbf{v}_j \in \Upsilon_j: \|\mathbb{V}_j(\mathbf{v}_j - \mathbf{v}_j^{\circ})\| \leq \mathbf{r}\}$$

denotes the ball in Υ_j with this radius.

Theorem H.9.1. *Let a bivariate random process $\mathcal{U}(\mathbf{v})$ on $\Upsilon = \Upsilon_1 \times \Upsilon_2$ satisfy (\mathcal{ED}_p) . Then for any \mathbf{r}_o and $\mathbf{x} \geq 1/2$, it holds on the product set $\Upsilon_o = \Upsilon_1(\mathbf{r}_o) \times \Upsilon_2(\mathbf{r}_o)$*

$$\mathbb{P}\left\{\frac{1}{\sqrt{8}\nu_0\mathbf{r}_o} \sup_{\mathbf{v} \in \Upsilon_o} |\mathcal{U}(\mathbf{v}, \mathbf{v}^o)| \geq \mathfrak{z}_{\mathbb{H}}(\mathbf{x})\right\} \leq e^{-\mathbf{x}},$$

with $\mathfrak{z}_{\mathbb{H}}(\mathbf{x})$ from (H.3) for $\mathbb{Q}_1(\Upsilon^o) = \mathbb{Q}_1(\Upsilon_1) + \mathbb{Q}_1(\Upsilon_2)$ and $\mathbb{Q}_2(\Upsilon^o) = \mathbb{Q}_2(\Upsilon_1) + \mathbb{Q}_2(\Upsilon_2)$.

Proof. By the Hölder inequality, (H.35), and (H.34), it holds for $\|\gamma_1\| = \|\gamma_2\| = 1$ and $\mathbf{v} \in \Upsilon$.

$$\begin{aligned} & \log \mathbb{E} \exp \left\{ \frac{\lambda}{2} (\gamma_1, \gamma_2)^\top \nabla \mathcal{U}(\mathbf{v}) \right\} \\ & \leq \frac{1}{2} \log \mathbb{E} \exp \left\{ \lambda \gamma_1^\top \nabla_1 \mathcal{U}(\mathbf{v}) \right\} + \frac{1}{2} \log \mathbb{E} \exp \left\{ \lambda \gamma_2^\top \nabla_2 \mathcal{U}(\mathbf{v}) \right\} \\ & \leq \frac{1}{2} \log \mathbb{E} \exp \left\{ \lambda \gamma_1^\top \nabla_1 \mathcal{U}(\mathbf{v}) \right\} + \frac{1}{2} \log \mathbb{E} \exp \left\{ \lambda \gamma_2^\top \nabla_2 \mathcal{U}(\mathbf{v}) \right\} \\ & \leq \frac{\nu_0^2 \lambda^2}{2}, \quad |\lambda| \leq g. \end{aligned}$$

This means that the bivariate process $\mathcal{U}(\mathbf{v})/2$ fulfills the full dimensional condition (\mathcal{ED}) with $\mathbb{V} = \text{block}(\mathbb{V}_1, \mathbb{V}_2)$. Let $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)$ and $\mathbf{v}^o = (\mathbf{v}_1^o, \mathbf{v}_2^o)$ be a couple of points in Υ such that $\|\mathbb{V}_j(\mathbf{v}_j - \mathbf{v}_j^o)\| \leq \varepsilon$ for $j = 1, 2$. Then obviously

$$\|\mathbb{V}(\mathbf{v} - \mathbf{v}^o)\|^2 \leq 2\varepsilon^2. \quad (\text{H.33})$$

Therefore, the direct product of two ε -nets $\mathcal{M}_j(\varepsilon)$ in Υ_j for $j = 1, 2$ yield a $\sqrt{2}\varepsilon$ -net $\mathcal{M}(\varepsilon) = \mathcal{M}_1(\varepsilon) \times \mathcal{M}_2(\varepsilon)$ in the product space Υ .

Due to (H.33), the product set $\Upsilon_o \stackrel{\text{def}}{=} \Upsilon_1(\mathbf{r}_o) \times \Upsilon_2(\mathbf{r}_o)$ has the radius \mathbf{r}_o . Now we can easily bound the entropy of the product set Υ_o via the entropy of Υ_1 and Υ_2 . Indeed, it holds with $\mathbf{r}_k = 2^{-k}\mathbf{r}_o$ for the cardinality \mathbb{N}_k of $\mathcal{M}_k = \mathcal{M}(\mathbf{r}_k)$

$$\mathbb{N}_k = \mathbb{N}_k(\Upsilon_1) \mathbb{N}_k(\Upsilon_2)$$

and

$$\begin{aligned} \mathbb{Q}_2(\Upsilon_o) & \leq \sum_{k=1}^{\infty} 2^{-k+1} \log(2\mathbb{N}_k) \\ & \leq \sum_{k=1}^{\infty} 2^{-k+1} \log(2\mathbb{N}_k(\Upsilon_1)) + \sum_{k=1}^{\infty} 2^{-k+1} \log(2\mathbb{N}_k(\Upsilon_2)) \leq \mathbb{Q}_2(\Upsilon_1) + \mathbb{Q}_2(\Upsilon_2). \end{aligned}$$

Similarly

$$\begin{aligned} \mathbb{Q}_1(\Upsilon_\circ) &\leq \sum_{k=1}^{\infty} 2^{-k} \sqrt{2 \log(2\mathbb{N}_k)} \\ &\leq \sum_{k=1}^{\infty} 2^{-k} \sqrt{2 \log(2\mathbb{N}_k(\Upsilon_1)) + 2 \log(2\mathbb{N}_k(\Upsilon_2))} \leq \mathbb{Q}_1(\Upsilon_1) + \mathbb{Q}_1(\Upsilon_2). \end{aligned}$$

Now we just apply the assertion of Theorem H.6.1 to the process $\mathcal{U}(\mathbf{v})/2$ and account for the fact that by (H.33) the radius of Υ_\circ is $\sqrt{2}\mathbf{r}_\circ$.

H.10 A bound for the norm of a vector random process

Let $\mathcal{Y}(\mathbf{v})$, $\mathbf{v} \in \Upsilon$, be a smooth centered random vector process with values in \mathbb{R}^q , where $\Upsilon \subseteq \mathbb{R}^p$. Let also $\mathcal{Y}(\mathbf{v}^*) = 0$ for a fixed point $\mathbf{v}^* \in \Upsilon$. Without loss of generality assume $\mathbf{v}^* = 0$. We aim to bound the maximum of the norm $\|\mathcal{Y}(\mathbf{v})\|$ over a vicinity Υ_\circ of \mathbf{v}^* . By $\nabla \mathcal{U}(\mathbf{v})$ we denote the $p \times q$ matrix with entries $\nabla_{\mathbf{v}_i} \mathcal{U}_j$, $i \leq p$, $j \leq q$. Suppose that $\mathcal{Y}(\mathbf{v})$ satisfies for each $\gamma_1 \in \mathbb{R}^p$ and $\gamma_2 \in \mathbb{R}^q$ with $\|\gamma_1\| = \|\gamma_2\| = 1$

$$\sup_{\mathbf{v} \in \Upsilon} \log \mathbb{E} \exp \left\{ \lambda \gamma_1^\top \nabla \mathcal{Y}(\mathbf{v}) \gamma_2 \right\} \leq \frac{\nu_0^2 \lambda^2}{2}, \quad |\lambda| \leq g. \quad (\text{H.34})$$

Condition (H.34) implies for any $\mathbf{v} \in \Upsilon_\circ$ with $\|\mathbf{v}\| \leq \mathbf{r}$ and $\gamma \in \mathbb{R}^q$ with $\|\gamma\| = 1$ in view of $\mathcal{Y}(\mathbf{v}^*) = 0$ by Lemma H.6.1

$$\log \mathbb{E} \exp \left\{ \frac{\lambda}{\mathbf{r}} \mathcal{Y}(\mathbf{v})^\top \gamma \right\} \leq \frac{\nu_0^2 \lambda^2 \|\mathbf{v}\|^2}{2\mathbf{r}^2}, \quad |\lambda| \leq g. \quad (\text{H.35})$$

In what follows, we use the representation

$$\|\mathcal{Y}(\mathbf{v})\| = \sup_{\|\mathbf{u}\| \leq \mathbf{r}} \frac{1}{\mathbf{r}} \mathbf{u}^\top \mathcal{Y}(\mathbf{v}). \quad (\text{H.36})$$

This implies for $\Upsilon_\circ(\mathbf{r}) = \{\mathbf{v} \in \Upsilon : \|\mathbf{v} - \mathbf{v}^*\| \leq \mathbf{r}\}$

$$\sup_{\mathbf{v} \in \Upsilon_\circ(\mathbf{r})} \|\mathcal{Y}(\mathbf{v})\| = \sup_{\mathbf{v} \in \Upsilon_\circ(\mathbf{r})} \sup_{\|\mathbf{u}\| \leq \mathbf{r}} \frac{1}{\mathbf{r}} \mathbf{u}^\top \mathcal{Y}(\mathbf{v}).$$

Consider a bivariate process $\mathbf{u}^\top \mathcal{Y}(\mathbf{v})$ of $\mathbf{u} \in \mathbb{R}^q$ and $\mathbf{v} \in \Upsilon \subset \mathbb{R}^p$. By definition $\mathbb{E} \mathbf{u}^\top \mathcal{Y}(\mathbf{v}) = 0$. Further, $\nabla_{\mathbf{u}} [\mathbf{u}^\top \mathcal{Y}(\mathbf{v})] = \mathcal{Y}(\mathbf{v})$ while $\nabla_{\mathbf{v}} [\mathbf{u}^\top \mathcal{Y}(\mathbf{v})] = \mathbf{u}^\top \nabla \mathcal{Y}(\mathbf{v}) = \|\mathbf{u}\| \gamma^\top \nabla \mathcal{Y}(\mathbf{v})$ for $\gamma = \mathbf{u}/\|\mathbf{u}\|$. Suppose that $\mathbf{u} \in \mathbb{R}^q$ and $\mathbf{v} \in \Upsilon$ are such that $\|\mathbf{u}\| \leq \mathbf{r}$ and $\|\mathbf{v}\| \leq \mathbf{r}$. By (H.34), it holds for $\gamma \in \mathbb{R}^p$ with $\|\gamma\| = 1$ and $\mathbf{v} \in \Upsilon_\circ(\mathbf{r})$

$$\log \mathbb{E} \exp \left\{ \frac{\lambda}{\mathbf{r}} \nabla_{\mathbf{v}} [\mathbf{u}^\top \mathcal{Y}(\mathbf{v})] \gamma \right\} \leq \log \mathbb{E} \exp \left\{ \frac{\lambda}{\mathbf{r}} \mathbf{u}^\top \nabla \mathcal{Y}(\mathbf{v}) \gamma \right\} \leq \frac{\nu_0^2 \lambda^2}{2},$$

and by (H.35) for a unit vector $\gamma \in \mathbb{R}^q$

$$\log \mathbb{E} \exp \left\{ \frac{\lambda}{r} \nabla_{\mathbf{u}} [\mathbf{u}^\top \mathcal{Y}(\mathbf{v})] \boldsymbol{\gamma} \right\} \leq \log \mathbb{E} \exp \left\{ \frac{\lambda}{r} \mathcal{Y}(\mathbf{v}) \boldsymbol{\gamma} \right\} \leq \frac{\nu_0^2 \lambda^2}{2}.$$

Therefore, (\mathcal{ED}_p) is fulfilled for $\mathbf{u}^\top \mathcal{Y}(\mathbf{v})$ and Theorem H.6.1 applies. We summarize our findings in the following theorem.

Theorem H.10.1. *Let a random p -vector process $\mathcal{Y}(\mathbf{v})$ for $\mathbf{v} \in \Upsilon \subseteq \mathbb{R}^p$ fulfill $\mathcal{Y}(\mathbf{v}^*) = 0$, $\mathbb{E} \mathcal{Y}(\mathbf{v}) \equiv 0$, and the condition (H.34) be satisfied. Then for each r and any $x \geq 1/2$, it holds for $\Upsilon_o = \Upsilon_o(r)$*

$$\mathbb{P} \left\{ \sup_{\mathbf{v} \in \Upsilon_o(r)} \|\mathcal{Y}(\mathbf{v})\| \geq \sqrt{8} \nu_0 r \mathfrak{z}_{\mathbb{H}}(x) \right\} \leq e^{-x}, \quad (\text{H.37})$$

where $\mathfrak{z}_{\mathbb{H}}(x)$ is given by (H.3) with $\mathbb{Q}_1 = \mathbb{Q}_1(\Upsilon_o) + \sqrt{6q}$ and $\mathbb{Q}_1 = \mathbb{Q}_1(\Upsilon_o) + 6q$.

H.11 A bound for a family of quadratic forms

Now we consider an extension of the previous result with a quadratic form $\|A \mathcal{Y}(\mathbf{v})\|^2$ to be bounded under the conditions (H.34) and (H.35) on $\mathcal{Y}(\mathbf{v})$ for $\mathbf{v} \in \Upsilon \subset \mathbb{R}^p$. Here $\mathcal{Y}(\cdot)$ is a vector process with values in \mathbb{R}^q and A is a $q \times q$ matrix with $\|A^\top A\|_{\text{op}} \leq 1$. The idea is to use the representation (H.36) in which we replace \mathbf{u} with $A\mathbf{u}$. The bound (H.37) implies for any r

$$\mathbb{P} \left\{ \sup_{\mathbf{v} \in \Upsilon_o(r), \|A\mathbf{u}\| \leq r} \mathbf{u}^\top A \mathcal{Y}(\mathbf{v}) > \sqrt{8} \nu_0 r \mathfrak{z}_{\mathbb{H}}(x) \right\} \leq e^{-x},$$

where $\mathfrak{z}_{\mathbb{H}}(x)$ corresponds to $\mathbb{Q}_1 = \sqrt{\mathbb{Q}_2} = \sqrt{6p + \mathbb{Q}_2(\Upsilon_o)}$.

Now we discuss how this bound can be refined if A is a smoothing operator. For simplicity assume that A fulfills the condition of Theorem H.7.2. One can expect that the dimension q can be replaced by the effective dimension p_A . The arguments similar to the above yield

$$\|A \mathcal{Y}(\mathbf{v})\| = \sup_{\mathbf{u} \in \mathbb{R}^q : \|\mathbf{u}\| \leq r} \frac{1}{r} \mathbf{u}^\top A \mathcal{Y}(\mathbf{v}),$$

and we again consider a bivariate process $\mathbf{u}^\top A \mathcal{Y}(\mathbf{v})$ of $\mathbf{u} \in \mathbb{R}^q$ and $\mathbf{v} \in \Upsilon \subset \mathbb{R}^p$. The conditions (H.34) and (H.35) imply for any two unit vectors $\boldsymbol{\gamma}_1 \in \mathbb{R}^q$ and $\boldsymbol{\gamma}_2 \in \mathbb{R}^p$ and any points $\mathbf{u} \in \mathbb{R}^q$ with $\|A\mathbf{u}\| \leq r$ and $\mathbf{v} \in \Upsilon_o(r)$, it holds

$$\log \mathbb{E} \exp \left\{ \frac{\lambda}{r} \nabla_{\mathbf{v}} [\mathbf{u}^\top A \mathcal{Y}(\mathbf{v})] \boldsymbol{\gamma}_2 \right\} = \log \mathbb{E} \exp \left\{ \frac{\lambda}{r} \mathbf{u}^\top A \nabla \mathcal{Y}(\mathbf{v}) \boldsymbol{\gamma}_2 \right\} \leq \frac{\nu_0^2 \lambda^2}{2},$$

and by (H.35) with $\mathbb{V}_1^2 = A^\top A$

$$\log I\!\!E \exp \left\{ \frac{\lambda}{\|\mathbb{V}_1 \gamma_1\|} \gamma_1^\top \nabla_{\mathbf{u}} [\mathbf{u}^\top A \mathcal{Y}(\mathbf{v})] \right\} \leq \log I\!\!E \exp \left\{ \frac{\lambda}{\|\mathbb{V}_1 \gamma_1\|} (A \gamma_1)^\top \mathcal{Y}(\mathbf{v}) \right\} \leq \frac{\nu_0^2 \lambda^2}{2}.$$

Therefore, (\mathcal{ED}_p) is fulfilled for $\mathbf{u}^\top A \mathcal{Y}(\mathbf{v})$. Now we apply the bound from Theorem H.9.1 and the entropy bound for the elliptic set $\|A\mathbf{u}\| \leq \mathbf{r}$ from Theorem H.7.2.

Theorem H.11.1. *Let a random vector process $\mathcal{Y}(\mathbf{v}) \in \mathbb{R}^q$ for $\mathbf{v} \in \Upsilon \subseteq \mathbb{R}^p$ fulfill $\mathcal{Y}(\mathbf{v}^*) = 0$, $I\!\!E \mathcal{Y}(\mathbf{v}) \equiv 0$, and the condition (H.34) be satisfied. Let A fulfill $1/2 \leq \|AA^\top\|_{\text{op}} \leq 1$. Then for each \mathbf{r} , it holds*

$$I\!\!P \left\{ \sup_{\mathbf{v} \in \Upsilon_\circ(\mathbf{r})} \|A \mathcal{Y}(\mathbf{v})\| > \sqrt{8} \nu_0 \mathbf{r} \mathfrak{z}_{\mathbb{H}}(\mathbf{x}) \right\} \leq e^{-\mathbf{x}},$$

where $\mathfrak{z}_{\mathbb{H}}(\mathbf{x})$ is given by (H.3) with $\mathbb{Q}_2 = C p_A + \mathbb{Q}_2(\Upsilon_\circ(\mathbf{r}))$ and $\mathbb{Q}_1 = C \sqrt{p_A} + \mathbb{Q}_1(\Upsilon_\circ(\mathbf{r}))$.

H.12 A bound for a smooth quadratic field

Let $\boldsymbol{\varepsilon}(\mathbf{v})$ be a vector random process with values in \mathbb{R}^n , $\mathbf{v} \in \Upsilon$, with independent entries $\varepsilon_i(\mathbf{v})$. We suppose that $\boldsymbol{\varepsilon}(\mathbf{v})$ satisfies the exponential moment condition (B.9).

Let also B be a fixed symmetric non-negative $n \times n$ matrix. Define

$$\mathcal{U}(\mathbf{v}) \stackrel{\text{def}}{=} \boldsymbol{\varepsilon}(\mathbf{v})^\top B \boldsymbol{\varepsilon}(\mathbf{v}).$$

The problem is to bound the variability of such process over $\mathbf{v} \in \Upsilon_\circ(\mathbf{r})$ in the form

$$I\!\!P \left(\sup_{\mathbf{v} \in \Upsilon_\circ(\mathbf{r})} |\mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}^*)| > z(\mathbf{x}) \right) \leq e^{-\mathbf{x}}.$$

Represent B as $B = A^\top A$. Then

$$\mathcal{U}(\mathbf{v}) = \|A \boldsymbol{\varepsilon}(\mathbf{v})\|^2$$

and

$$\begin{aligned} |\mathcal{U}(\mathbf{v}) - \mathcal{U}(\mathbf{v}^*)| &= \|A \boldsymbol{\varepsilon}(\mathbf{v})\|^2 - \|A \boldsymbol{\varepsilon}(\mathbf{v}^*)\|^2 \\ &\leq \|A\{\boldsymbol{\varepsilon}(\mathbf{v}) - \boldsymbol{\varepsilon}(\mathbf{v}^*)\}\|^2 + 2\|A \boldsymbol{\varepsilon}(\mathbf{v}^*)\| \cdot \|A\{\boldsymbol{\varepsilon}(\mathbf{v}) - \boldsymbol{\varepsilon}(\mathbf{v}^*)\}\|. \end{aligned}$$

The quadratic form $\|A \boldsymbol{\varepsilon}(\mathbf{v}^*)\|$ can be bounded as in Theorem B.2.2:

$$I\!\!P \left(\|A \boldsymbol{\varepsilon}(\mathbf{v}^*)\| > z(B, \mathbf{x}) \right) \leq 2e^{-\mathbf{x}}.$$

Here we ignore a small term related to \mathbf{x}_c in Theorem B.2.2. Further, the deviations $\|A\{\boldsymbol{\varepsilon}(\mathbf{v}) - \boldsymbol{\varepsilon}(\mathbf{v}^*)\}\|$ can be bounded on a ball $\Upsilon_\circ(\mathbf{r})$ by Theorem H.11.1:

$$\mathbb{P} \left\{ \sup_{v \in \Upsilon_\circ(r)} \|A\{\varepsilon(v) - \varepsilon(v^*)\}\| > \sqrt{8} \nu_0 r q_{\mathbb{H}}(x) \right\} \leq e^{-x},$$

where $q_{\mathbb{H}}(x)$ is given by (H.3). One can summarize that on a set $\Omega(x)$ with $\mathbb{P}(\Omega(x)) \geq 1 - 3e^{-x}$, it holds

$$|\mathcal{U}(v) - \mathcal{U}(v^*)| \leq 8|\nu_0 r q_{\mathbb{H}}(x)|^2 + 4\sqrt{8} z(B, x) \nu_0 r q_{\mathbb{H}}(x).$$

References

- Andresen, A. and Spokoiny, V. (2014). Critical dimension in profile semiparametric estimation. *Electronic J. Statist.*, 8(2):3077–3125. Manuscript. arXiv:1303.4640.
- Arlot, S. (2009). Model selection by resampling penalization. *Electron. J. Statist.*, 3:557–624.
- Barron, A., Birgé, L., and Massart, P. (1999a). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413.
- Barron, A., Birgé, L., and Massart, P. (1999b). Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields*, 113(3):301–413.
- Bednorz, W. (2006). A theorem on majorizing measures. *Ann. Probab.*, 34(5):1771–1781.
- Belloni, A. and Chernozhukov, V. (2009). On the computational complexity of MCMC-based estimators in large samples. *Ann. Statist.*, 37(4):2011–2055.
- Beran, R. (1986). Discussion: Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14(4):1295–1298.
- Bickel, P. J. and Kleijn, B. J. K. (2012). The semiparametric Bernstein-von Mises theorem. *Ann. Statist.*, 40(1):206–237.
- Birgé, L. (2001). *An alternative point of view on Lepski’s method*, volume Volume 36 of *Lecture Notes–Monograph Series*, pages 113–133. Institute of Mathematical Statistics, Beachwood, OH.
- Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Relat. Fields*, 97(1-2):113–150.
- Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375.
- Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268.
- Birgé, L. and Massart, P. (2007a). Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, 138(1-2):33–73.

- Birgé, L. and Massart, P. (2007b). Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields*, 138(1-2):33–73.
- Bontemps, D. (2011). Bernstein-von Mises theorems for Gaussian regression with increasing number of regressors. *Ann. Statist.*, 39(5):2557–2584.
- Boucheron, S. and Gassiat, E. (2009). A Bernstein-von Mises theorem for discrete probability distributions. *Electron. J. Stat.*, 3:114–148.
- Boucheron, S. and Massart, P. (2011). A high-dimensional Wilks phenomenon. *Probability Theory and Related Fields*, 150:405–433. 10.1007/s00440-010-0278-7.
- Bunke, O. and Milhaud, X. (1998). Asymptotic behavior of Bayes estimates under possibly incorrect models. *Ann. Statist.*, 26(2):617–644.
- Cai, T. T. and Low, M. G. (2003). A note on nonparametric estimation of linear functionals. *Ann. Statist.*, 31(4):1140–1153.
- Cai, T. T. and Low, M. G. (2005). On adaptive estimation of linear functionals. *Ann. Statist.*, 33(5):2311–2343.
- Castillo, I. (2012). A semiparametric Bernstein - von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields*, 152:53–99. 10.1007/s00440-010-0316-5.
- Castillo, I. and Nickl, R. (2013). Nonparametric Bernstein-von Mises theorems in Gaussian white noise. *Ann. Statist.*, 41(4):1999–2028.
- Castillo, I. and Rousseau, J. (2013). A general bernstein-von mises theorem in semiparametric models.
- Cavalier, L., Golubev, G. K., Picard, D., and Tsybakov, A. B. (2002). Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3):843–874.
- Cavalier, L. and Golubev, Y. (2006). Risk hull method and regularization by projections of ill-posed inverse problems. *Ann. Statist.*, 34(4):1653–1677.
- Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.
- Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.*, 42(5):1787–1818.
- Dalalyan, A. S. and Salmon, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.*, 40(4):2327–2355.
- Fan, J., Zhang, C., and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Stat.*, 29(1):153–193.
- Freedman, D. (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Stat.*, 27(4):1119–1140.

- Gach, F., Nickl, R., and Spokoiny, V. (2013). Spatially Adaptive Density Estimation by Localised Haar Projections. *Annales de l'Institut Henri Poincaré - Probability and Statistics*, 49(3):900–914. DOI: 10.1214/12-AIHP485; arXiv:1111.2807.
- Ghosal, S. (1999). Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli*, 5(2):315–331.
- Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.*, 74(1):49–68.
- Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.
- Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35:192.
- Gine, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.*, 38(2):1122–1170.
- Goldenshluger, A. (2009). A universal procedure for aggregating estimators. *Ann. Statist.*, 37(1):542–568.
- Golubev, Y. and Spokoiny, V. (2009). Exponential bounds for minimum contrast estimators. *Electron. J. Statist.*, 3:712–746.
- Green, P. J. and Silverman, B. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. London: Chapman & Hall.
- Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 221–233 (1967).
- Ibragimov, I. and Khas'minskij, R. (1981). *Statistical estimation. Asymptotic theory. Transl. from the Russian by Samuel Kotz*. New York - Heidelberg -Berlin: Springer-Verlag .
- Johnstone, I. M. (2010). High dimensional Bernstein–von Mises: simple examples. In *Borrowing strength: theory powering applications—a Festschrift for Lawrence D. Brown*, volume 6 of *Inst. Math. Stat. Collect.*, pages 87–98. Inst. Math. Statist., Beachwood, OH.
- Kim, Y. (2006). The Bernstein-von Mises theorem for the proportional hazard model. *Ann. Statist.*, 34(4):1678–1700.
- Kleijn, B. J. K. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.*, 34(2):837–877.
- Kleijn, B. J. K. and van der Vaart, A. W. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic J. Statist.*, 6:354–381.
- Kneip, A. (1994). Ordered linear smoothers. *Ann. Statist.*, 22(2):835–866.

- Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, 81(4):673–680.
- Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338.
- Lepski, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3):459–470.
- Lepski, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4):645–659.
- Lepski, O. V. (1992). Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 37(3):468–481.
- Lepski, O. V., Mammen, E., and Spokoiny, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics*, 25(3):929–947.
- Lepski, O. V. and Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.*, 25(6):2512–2546.
- Liptser, R. and Spokoiny, V. (2000). Deviation probability bound for martingales with applications to statistical estimation. *Statist. Probab. Lett.*, 46(4):347–357.
- Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Stat.*, 21(1):255–285.
- Mammen, E. (1996). Empirical process of residuals for high-dimensional linear models. *Ann. Stat.*, 24(1):307–335.
- Massart, P. (2007). *Concentration inequalities and model selection*. Number 1896 in Ecole d'Eté de Probabilités de Saint-Flour. Springer.
- Panov, M. and Spokoiny, V. (2015). Finite Sample Bernstein - von Mises Theorem for Semiparametric Problems. *Bayesian Analysis*, 10(3):665–710.
- Portnoy, S. (1984). Asymptotic behavior of M -estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.*, 12(4):1298–1309.
- Portnoy, S. (1984). Asymptotic behavior of M -estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Stat.*, 12:1298–1309.
- Portnoy, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.*, 13(4):1403–1417.
- Portnoy, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II: Normal approximation. *Ann. Stat.*, 13:1403–1417.
- Portnoy, S. (1986). Asymptotic behavior of the empiric distribution of M -estimated residuals from a regression model with many parameters. *Ann. Stat.*, 14:1152–1170.

- Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.*, 16(1):356–366.
- Rivoirard, V. and Rousseau, J. (2012). Bernstein–von mises theorem for linear functionals of the density. *Ann. Stat.*, 40(3):1489–1523.
- Shen, X. (1997). On methods of sieves and penalization. *Ann. Stat.*, 25(6):2555–2591.
- Shen, X. (2002). Asymptotic normality of semiparametric and nonparametric posterior distributions. *J. Amer. Statist. Assoc.*, 97(457):222–235.
- Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Ann. Stat.*, 22(2):580–615.
- Spokoiny, V. (2012). Parametric estimation. Finite sample theory. *Ann. Statist.*, 40(6):2877–2909. arXiv:1111.3029.
- Spokoiny, V. and Vial, C. (2009). Parameter tuning in pointwise adaptation using a propagation approach. *Ann. Statist.*, 37:2783–2807.
- Spokoiny, V., Wang, W., and Härdle, W. (2013). Local quantile regression (with rejoinder). *J. of Statistical Planing and Inference*, 143(7):1109–1129. ArXiv:1208.5384.
- Spokoiny, V. and Zhilova, M. (2013). Sharp deviation bounds for quadratic forms. *Mathematical Methods of Statistics*, 22(2):100–113. arXiv:1302.1699; doi:10.3103/S1066530713020026.
- Spokoiny, V. and Zhilova, M. (2015). Bootstrap confidence sets under model misspecification. *Annals of Statist.* in print. arXiv:1410.0347.
- Talagrand, M. (1996). Majorizing measures: the generic chaining. *Ann. Probab.*, 24(3):1049–1103.
- Talagrand, M. (2001). Majorizing measures without measures. *Ann. Probab.*, 29(1):411–417.
- Talagrand, M. (2005). *The generic chaining*. Springer Monographs in Mathematics. Springer-Verlag, Berlin. Upper and lower bounds of stochastic processes.
- Tropp, J. A. (2015). Found. Trends Mach. Learning. to appear.
- Tsybakov, A. (2000). On the best rate of adaptive estimation in some inverse problems. *Comptes Rendus de l'Academie des Sciences - Series I - Mathematics*, 330(9):835 – 840.
- Van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Stat.*, 21(1):14–44.
- van de Geer, S. (2002). M-estimation using penalties or sieves. *J. Stat. Plann. Inference*, 108(1-2):55–69.
- Van de Geer, S. A. (2000). *Applications of empirical process theory*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

- van der Vaart, A. and Wellner, J. A. (1996). *Weak convergence and empirical processes. With applications to statistics.* Springer Series in Statistics. New York, Springer.
- Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14(4):1261–1295.
- Zaitsev, A., Burnaev, E., and Spokoiny, V. (2013). Properties of the posterior distribution of a regression model based on gaussian random fields. *Automation and Remote Control*, 74(10):1645–1655.