Vladimir Spokoiny

Nonparametric estimation: parametric view

February 25, 2019

Springer

Berlin Heidelberg New York
Hong Kong London
Milan Paris Tokyo

Contents

Part I Model selection for linear methods

Part II Mathematical tools

Part I

Model selection for linear methods

1

Quasi maximum likelihood estimation in linear models

This chapter discusses the estimation problem for the regression model. First a linear regression model is considered, then a generalized linear modeling is discussed. We also mention median and quantile regression.

1.1 Regression model

The (mean) *regression model* can be written in the form $I\!\!E(Y|X) = f(X)$, or equivalently,

$$Y = f(X) + \varepsilon, \tag{1.1}$$

where Y is the dependent (explained) variable and X is the explanatory variable (regressor) which can be multidimensional. The target of analysis is the systematic dependence of the explained variable Y from the explanatory variable X. The *regression function* f describes the dependence of the mean of Y as a function of X. The value ε can be treated as an individual deviation (error). It is usually assumed to be random with zero mean. Below we discuss in more detail the components of the regression model (1.1).

1.1.1 Observations

In almost all practical situations, regression analysis is performed on the basis of available data (observations) given in the form of a sample of pairs (X_i, Y_i) for $i = 1, \ldots, n$, where n is the *sample size*. Here $Y_1, \ldots, Y_n$ are observed values of the regression variable Y and $X_1, \ldots, X_n$ are the corresponding values of the explanatory variable X. For each observation Y_i, the regression model reads as:

$$Y_i = f(X_i) + \varepsilon_i$$

where ε_i is the individual ith error.

1.1.2 Design

The set $X_1, \ldots, X_n$ of the regressor's values is called a *design*. The set $\mathcal{X}$ of all possible values of the regressor X is called the *design space*. If this set $\mathcal{X}$ is compact, then one speaks of a *compactly supported design*.

The nature of the design can be different for different statistical models. However, it is important to mention that the design is always observable. Two kinds of design assumptions are usually used in statistical modeling. A *deterministic* design assumes that the points $X_1, \ldots, X_n$ are nonrandom and given in advance. Here are typical examples:

Example 1.1.1 (Time series). Let $Y_{t_0}, Y_{t_0+1}, \ldots, Y_T$ be a time series. The time points $t_0, t_0 + 1, \ldots, T$ build a regular deterministic design. The regression function f explains the trend of the time series Y_t as a function of time.

Example 1.1.2 (Imaging). Let Y_{ij} be the observed grey value at the pixel (i, j) of an image. The coordinate X_{ij} of this pixel is the corresponding design value. The regression function $f(X_{ij})$ gives the true image value at X_{ij} which is to be recovered from the noisy observations Y_{ij}.

If the design is supported on a cube in $I\!R^d$ and the design points X_i form a grid in this cube, then the design is called *equidistant*. An important feature of such a design is that the number N_A of design points in any "massive" subset A of the unit cube is nearly the volume of this subset V_A multiplied by the sample size n: $N_A \approx nV_A$. *Design regularity* means that the value N_A is nearly proportional to nV_A, that is, $N_A \approx cnV_A$ for some positive constant c which may depend on the set A.

In some applications, it is natural to assume that the design values X_i are randomly drawn from some design distribution. Typical examples are given by sociological studies. In this case one speaks of a *random* design. The design values $X_1, \ldots, X_n$ are assumed to be independent and identically distributed from a law P_X on the design space $\mathcal{X}$ which is a subset of the Euclidean space $I\!R^d$. The design variables X are also assumed to be independent of the observations Y.

One special case of random design is the *uniform* design when the design distribution is uniform on the unit cube in $I\!R^d$. The uniform design possesses a similar, important property to an equidistant design: the number of design points in a "massive" subset of the unit cube is on average close to the volume of this set multiplied by n. The random design is called *regular* on $\mathcal{X}$ if the design distribution is absolutely continuous with respect to the Lebesgue measure and the design density $f(x) = dP_X(x)/d\lambda$ is positive and continuous on $\mathcal{X}$. This again ensures with a probability close to one the regularity property $N_A \approx cnV_A$ with $c = f(x)$ for some $x \in A$.

It is worth mentioning that the case of a random design can be reduced to the case of a deterministic design by considering the conditional distribution of the data given the design variables $X_1, \ldots, X_n$.

1.1.3 Errors

The decomposition of the observed response variable Y into the systematic component $f(x)$ and the error $\boldsymbol{\varepsilon}$ in the model equation (1.1) is not formally defined and cannot be done without some assumptions on the errors ε_i. The standard approach is to assume that the mean value of every ε_i is zero. Equivalently this means that the expected value of the observation Y_i is just the regression function $f(X_i)$. This case is called *mean regression* or simply regression. It is usually assumed that the errors ε_i have finite second moments. *Homogeneous errors* case means that all the errors ε_i have the same variance $\sigma^2 = \operatorname{Var} \varepsilon_i^2$. The variance of *heterogeneous errors* ε_i may vary with i. In many applications not only the systematic component $f(X_i) = \mathbb{E} Y_i$ but also the error variance $\operatorname{Var} Y_i = \operatorname{Var} \varepsilon_i$ depend on the regressor (location) X_i. Such models are often written in the form

$$Y_i = f(X_i) + \sigma(X_i)\varepsilon_i .$$

The observation (noise) variance $\sigma^2(x)$ can be the target of analysis similarly to the mean regression function.

The assumption of zero mean noise, $\mathbb{E} \varepsilon_i = 0$, is very natural and has a clear interpretation. However, in some applications, it can cause trouble, especially if data are contaminated by outliers. In this case, the assumption of a zero mean can be replaced by a more robust assumption of a zero median. This leads to the *median regression* model which assumes $\mathbb{P}(\varepsilon_i \leq 0) = 1/2$, or, equivalently

$$\mathbb{P}\bigl(Y_i - f(X_i) \leq 0\bigr) = 1/2.$$

A further important assumption concerns the joint distribution of the errors ε_i. In the majority of applications the errors are assumed to be independent. However, in some situations, the dependence of the errors is quite natural. One example can be given by time series analysis. The errors ε_i are defined as the difference between the observed values Y_i and the trend function f_i at the ith time moment. These errors are often serially correlated and indicate short or long range dependence. Another example comes from imaging. The neighbor observations in an image are often correlated due to the imaging technique used for recoding the images. The correlation particularly results from the automatic movement correction.

For theoretical study one often assumes that the errors ε_i are not only independent but also identically distributed. This, of course, yields a homogeneous noise. The theoretical study can be simplified even further if the error distribution is normal. This case is called *Gaussian regression* and is denoted as $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. This assumption is very useful and greatly simplifies the theoretical study. The main advantage of Gaussian noise is that the observations and their linear combinations are also normally distributed. This is an exclusive property of the normal law which helps to simplify the exposition and avoid technicalities.

Under the given distribution of the errors, the joint distribution of the observations Y_i is determined by the regression function $f(\cdot)$.

1.1.4 Regression function

By the equation (1.1), the regression variable Y can be decomposed into a systematic component and a (random) error ε. The systematic component is a deterministic function f of the explanatory variable X called the *regression* function. Classical regression theory considers the case of *linear* dependence, that is, one fits a linear relation between Y and X:

$$f(x) = a + bx$$

leading to the model equation

$$Y_i = \theta_1 + \theta_2 X_i + \varepsilon_i .$$

Here θ_1 and θ_2 are the parameters of the linear model. If the regressor x is multidimensional, then θ_2 is a vector from $I\!R^d$ and $\theta_2 x$ becomes the scalar product of two vectors. In many practical examples the assumption of linear dependence is too restrictive. It can be extended by several ways. One can try a more sophisticated functional dependence of Y on X, for instance polynomial. More generally, one can assume that the regression function f is known up to the finite-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top \in I\!R^p$. This situation is called *parametric regression* and denoted by $f(\cdot) = f(\cdot, \boldsymbol{\theta})$. If the function $f(\cdot, \boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$ linearly, that is, $f(x, \boldsymbol{\theta}) = \theta_1 \psi_1(x) + \ldots + \theta_p \psi_p(x)$ for some given functions $\psi_1, \ldots, \psi_p$, then the model is called *linear regression.* An important special case is given by polynomial regression when $f(x)$ is a polynomial function of degree $p - 1$: $f(x) = \theta_1 + \theta_2 x + \ldots + \theta_p x^{p-1}$.

In many applications a parametric form of the regression function cannot be justified. Then one speaks of *nonparametric regression.*

1.2 Linear Modeling

A linear model assumes that the regression function $f(\cdot)$ can be represented as a linear combinations of *factors* or *features* $\Psi_i = \Psi(X_i) \in I\!\!R^p$:

$$f(x) = \theta_1 \Psi_1(x) + \ldots + \theta_p \Psi_p(x), \qquad x \in I\!\!R^d.$$

One can say, that we first map the original d-dimensional design $X_1, \ldots, X_n$ into a p-dimensional feature space and then use a linear representation of the response function $f(\cdot)$ in terms of features Ψ_j for $j \le p$. Then the observations Y_i follow the equation:

$$Y_i = \Psi_i^\top \boldsymbol{\theta}^* + \varepsilon_i \tag{1.2}$$

for $i = 1, \ldots, n$, where $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_p^*)^\top \in I\!\!R^p$ is an unknown parameter vector, Ψ_i are given vectors in $I\!\!R^p$ 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 $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots \varepsilon_n)^\top$ 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 $\sigma^2 \boldsymbol{I}_n$ with $\boldsymbol{I}_n$ being the unit matrix in $I\!\!R^n$.
2. the errors are independent but not homogeneous, that is, $I\!\!E \varepsilon_i^2 = \sigma_i^2$. Then the matrix Σ is diagonal: $\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)$.
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 $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ (resp. $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$) the vector of observations (resp. of errors) in $I\!\!R^n$ and by Ψ the $p \times n$ matrix with columns Ψ_i. Let also $\Psi^\top$ denote its transpose. Then the model equation can be rewritten as:

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

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

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

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:

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

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

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}\in I\!R^p} L(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}\in I\!R^p}(\boldsymbol{Y} - \Psi^{\top}\boldsymbol{\theta})^{\top}\Sigma^{-1}(\boldsymbol{Y} - \Psi^{\top}\boldsymbol{\theta}). \tag{1.3}$$

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 (1.3) is a quadratic optimization problem which has a closed form solution. Differentiating the right hand-side of (1.3) w.r.t. $\boldsymbol{\theta}$ yields the *normal equation*

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

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

$$\widetilde{\boldsymbol{\theta}} = \bigl(\Psi\Sigma^{-1}\Psi^{\top}\bigr)^{-1}\Psi\Sigma^{-1}\boldsymbol{Y} = \mathcal{S}\boldsymbol{Y}, \tag{1.4}$$

where

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

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

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 (1.4) 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* $\widetilde{\boldsymbol{\theta}}$. Note that due to the model (1.2), the product $\widetilde{\boldsymbol{f}} = \Psi^{\top}\widetilde{\boldsymbol{\theta}}$ is an estimate of the mean $\boldsymbol{f}^{*} \stackrel{\mathrm{def}}{=} I\!\!E\boldsymbol{Y}$ of the vector of observations $\boldsymbol{Y}$:

$$\widetilde{\boldsymbol{f}} = \Psi^{\top}\widetilde{\boldsymbol{\theta}} = \Psi^{\top}\bigl(\Psi\Sigma^{-1}\Psi^{\top}\bigr)^{-1}\Psi\Sigma^{-1}\boldsymbol{Y} = \Pi\boldsymbol{Y},$$

where

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

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

Below we study the properties of the estimates $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{f}}$. 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

$$Y_i = f_i + \varepsilon_i, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma_0). \tag{1.5}$$

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 (1.2) can be viewed as an approximation of (1.5) while Σ is an approximation of the true covariance matrix Σ_0. If $\boldsymbol{f}^{*}$ is indeed equal to $\Psi^{\top}\boldsymbol{\theta}^{*}$ and $\Sigma = \Sigma_0$, then $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{f}}$ are MLEs, otherwise quasi MLEs. In our study we mostly restrict ourselves to the case 1 assumption about the noise $\boldsymbol{\varepsilon}$: $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$. The general case can be reduced to this one by a simple data transformation, namely, by multiplying the equation (1.5) $\boldsymbol{Y} = \boldsymbol{f}^{*} + \boldsymbol{\varepsilon}$ with the matrix $\Sigma^{-1/2}$, see Section 1.6 for more detail.

1.2.1 Estimation under homogeneous noise assumption

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

$$\widetilde{\boldsymbol{\theta}} = \bigl(\Psi\Psi^{\top}\bigr)^{-1}\Psi\boldsymbol{Y} = \mathcal{S}\boldsymbol{Y}$$

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

$$\widetilde{\boldsymbol{f}} = \Psi^{\top}\bigl(\Psi\Psi^{\top}\bigr)^{-1}\Psi\boldsymbol{Y} = \Pi\boldsymbol{Y}$$

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

Exercise 1.2.1. Derive the formulae for $\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{f}}$ 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.

1.2.2 * Linear basis transformation

Denote by $\boldsymbol{\psi}_1^\top, \ldots, \boldsymbol{\psi}_p^\top$ the rows of the matrix Ψ. Then the $\boldsymbol{\psi}_i$'s are vectors in $\mathbb{R}^n$ and we call them *the basis vectors.* In the linear regression case the $\boldsymbol{\psi}_i$'s are obtained as the values of the basis functions at the design points. Our linear parametric assumption simply means that the underlying vector $\boldsymbol{f}^*$ can be represented as a linear combination of the vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p$:

$$\boldsymbol{f}^* = \theta_1^* \boldsymbol{\psi}_1 + \ldots + \theta_p^* \boldsymbol{\psi}_p .$$

In other words, $\boldsymbol{f}^*$ belongs to the linear subspace in $\mathbb{R}^n$ spanned by the vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p$. 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 $\boldsymbol{f}^* = \Psi^\top \boldsymbol{\theta}^*$ can be rewritten as

$$\boldsymbol{f}^* = \Psi^\top U U^\top \boldsymbol{\theta}^* = \breve{\Psi}^\top \boldsymbol{u}^*$$

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

$$\boldsymbol{f}^* = u_1^* \breve{\boldsymbol{\psi}}_1 + \ldots + u_p^* \breve{\boldsymbol{\psi}}_p .$$

The natural question is how the expression for the MLEs $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{f}}$ 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

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

yielding the solutions

$$\widetilde{\boldsymbol{u}} = \bigl(\breve{\Psi}\breve{\Psi}^\top\bigr)^{-1} \breve{\Psi} \boldsymbol{Y} = \breve{\mathcal{S}} \boldsymbol{Y}, \qquad \widetilde{\boldsymbol{f}} = \breve{\Psi}^\top \bigl(\breve{\Psi}\breve{\Psi}^\top\bigr)^{-1} \breve{\Psi} \boldsymbol{Y} = \breve{\Pi} \boldsymbol{Y},$$

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

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

This yields

$$
\widetilde{\boldsymbol{u}} = U^{\top}\widetilde{\boldsymbol{\theta}}
$$

and moreover, the estimate $\widetilde{\boldsymbol{f}}$ is not changed for any linear transformation of the basis. The first statement can be expected in view of $\boldsymbol{\theta}^{*} = U\boldsymbol{u}^{*}$, 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 1.2.2. 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

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

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

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

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 $\boldsymbol{u} = A\boldsymbol{\theta}$ for $p = 1$.
- Describe the transformation A such that $\boldsymbol{u} = A\boldsymbol{\theta}$ for $p > 1$.

Hint: use the formula

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

to identify the coefficient u_j^{*} via $\theta_j^{*},\ldots,\theta_p^{*}$.

1.2.3 Orthogonal and orthonormal design

Orthogonality of the design matrix Ψ means that the basis vectors $\psi_1,\ldots,\psi_p$ are orthonormal in the sense

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

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

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

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

$$\widetilde{\boldsymbol{\theta}} = \Lambda^{-1}\Psi\boldsymbol{Y}, \qquad \widetilde{\boldsymbol{f}} = \Psi^\top\widetilde{\boldsymbol{\theta}} = \Psi^\top\Lambda^{-1}\Psi\boldsymbol{Y}$$

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

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

that is, $\widetilde{\theta}_j$ is the scalar product of the data and the basis vector $\boldsymbol{\psi}_j$ for $j = 1, \ldots, p$. The estimate of the response $\boldsymbol{f}$ reads as

$$\widetilde{\boldsymbol{f}} = \widetilde{\theta}_1\boldsymbol{\psi}_1 + \ldots + \widetilde{\theta}_p\boldsymbol{\psi}_p.$$

Below we will prove the following result.

Theorem 1.2.1. *Consider the model* $\boldsymbol{Y} = \Psi^\top\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ *with homogeneous errors* $\boldsymbol{\varepsilon}$ *:* $\mathbb{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top = \sigma^2\boldsymbol{I}_n$ *. If the design* Ψ *is orthogonal, that is, if* $\Psi\Psi^\top = \Lambda$ *for a diagonal matrix* Λ *, then the estimated coefficients* $\widetilde{\theta}_j$ *are uncorrelated:* $\operatorname{Var}(\widetilde{\boldsymbol{\theta}}) = \sigma^2\Lambda^{-1}$ *. Moreover, if* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}_n)$ *, then* $\widetilde{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}^*, \sigma^2\Lambda^{-1})$ *.*

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 $\widetilde{\theta}_j$ independently on the remaining coefficients.

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

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

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

$$\operatorname{Var}\bigl(\widetilde{\boldsymbol{\theta}}\bigr) = \sigma^2 I_p\,.$$

1.2.4 * Spectral representation

Consider a linear model

$$\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{1.6}$$

with homogeneous errors $\boldsymbol{\varepsilon}$: $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$. The rows of the matrix Ψ can be viewed as basis vectors in $I\!\!R^n$ and the product $\Psi^\top \boldsymbol{\theta}$ is a linear combinations of these vectors with the coefficients $(\theta_1, \ldots, \theta_p)$. 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 $U : I\!\!R^p \to I\!\!R^p$ ensuring the orthogonality of the transformed basis. This means that the rows of the matrix $\breve{\Psi} = U^\top \Psi$ are orthogonal and the matrix $\breve{\Psi}\breve{\Psi}^\top$ is diagonal:

$$\breve{\Psi}\breve{\Psi}^\top = U^\top \Psi \Psi^\top U = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_p).$$

The original model reads after this transformation in the form

$$\boldsymbol{Y} = \breve{\Psi}^\top \boldsymbol{u} + \boldsymbol{\varepsilon}, \qquad \breve{\Psi}\breve{\Psi}^\top = \Lambda,$$

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

$$\boldsymbol{Z} = \Lambda \boldsymbol{u} + \Lambda^{1/2} \boldsymbol{\xi} \tag{1.7}$$

where $\boldsymbol{Z} = \breve{\Psi}\boldsymbol{Y} = U^\top \Psi \boldsymbol{Y}$ is a vector in $I\!\!R^p$ and $\boldsymbol{\xi} = \Lambda^{-1/2}\breve{\Psi}\boldsymbol{\varepsilon} = \Lambda^{-1/2} U^\top \Psi \boldsymbol{\varepsilon} \in I\!\!R^p$. The equation (1.7) is called the *spectral representation* of the linear model (1.6). 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 4.7.

Theorem 1.2.2. *Consider the model (1.6) with homogeneous errors $\boldsymbol{\varepsilon}$, that is, $I\!\!E \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top = \sigma^2 I_n$. Then there exists an orthogonal transform $U : I\!\!R^p \to I\!\!R^p$ leading to the spectral representation (1.7) with homogeneous uncorrelated errors $\boldsymbol{\xi}$: $I\!\!E \boldsymbol{\xi}\boldsymbol{\xi}^\top = \sigma^2 I_p$. If $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$, then the vector $\boldsymbol{\xi}$ is normal as well: $\boldsymbol{\xi} = \mathcal{N}(0, \sigma^2 \boldsymbol{I}_p)$.*

Exercise 1.2.3. Prove the result of Theorem 1.2.2.

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

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

A special case of the spectral representation corresponds to the orthonormal design with $\Psi \Psi^\top = \boldsymbol{I}_p$. In this situation, the spectral model reads as $\boldsymbol{Z} = \boldsymbol{u} + \boldsymbol{\xi}$, that is, we simply observe the target $\boldsymbol{u}$ corrupted with a homogeneous noise $\boldsymbol{\xi}$. Such an equation is often called the *sequence space model* and it is intensively used in the literature for the theoretical study; cf. Section 4 below.

1.3 Properties of the response estimate $\widetilde{\boldsymbol{f}}$

This section discusses some properties of the estimate $\widetilde{\boldsymbol{f}} = \Psi^\top \widetilde{\boldsymbol{\theta}} = \Pi \boldsymbol{Y}$ of the response vector $\boldsymbol{f}^*$. 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 $\boldsymbol{f} = \Psi^\top \boldsymbol{\theta}^*$ and $\operatorname{Cov}(\boldsymbol{\varepsilon}) = \Sigma = \sigma^2 \boldsymbol{I}_n$. The real model only appears when studying the risk of estimation. We will comment on the cases of misspecified $\boldsymbol{f}$ and Σ.

When $\Sigma = \sigma^2 \boldsymbol{I}_n$, the operator Π in the representation $\widetilde{\boldsymbol{f}} = \Pi \boldsymbol{Y}$ of the estimate $\widetilde{\boldsymbol{f}}$ reads as

$$\Pi = \Psi^\top \bigl(\Psi \Psi^\top\bigr)^{-1} \Psi. \tag{1.8}$$

First we make use of the linear structure of the model (1.2) and of the estimate $\widetilde{\boldsymbol{f}}$ to derive a number of its simple but important properties.

1.3.1 Decomposition into a deterministic and a stochastic component

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

$$\widetilde{\boldsymbol{f}} = \Pi \boldsymbol{Y} = \Pi(\boldsymbol{f}^* + \boldsymbol{\varepsilon}) = \Pi \boldsymbol{f}^* + \Pi \boldsymbol{\varepsilon}. \tag{1.9}$$

The first element of this sum, $\Pi \boldsymbol{f}^*$ is purely deterministic, but it depends on the unknown response vector $\boldsymbol{f}^*$. Moreover, it will be shown in the next lemma that $\Pi \boldsymbol{f}^* = \boldsymbol{f}^*$ if the parametric assumption holds and the vector $\boldsymbol{f}^*$ indeed can be represented as $\Psi^\top \boldsymbol{\theta}^*$. The second element is stochastic as a linear transformation of the stochastic vector $\boldsymbol{\varepsilon}$ but is independent of the model response $\boldsymbol{f}^*$. The properties of the estimate $\widetilde{\boldsymbol{f}}$ heavily rely on the properties of the linear operator Π from (1.8) which we collect in the next section.

1.3.2 Properties of the operator $\boldsymbol{\Pi}$

Let $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p$ be the columns of the matrix $\Psi^\top$. These are the vectors in $I\!\!R^n$ also called *the basis vectors.*

Lemma 1.3.1. *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* $I\!\!R^n$, *i.e.* $\Pi^\top \Pi = \Pi^2 = \Pi$ *and* $\Pi(\boldsymbol{1}_n - \Pi) = 0$, *where* $\boldsymbol{1}_n$ *means the unity operator in* $I\!\!R^n$.

(iii) For an arbitrary vector v *from* $I\!\!R^n$, *it holds* $\|v\|^2 = \|\Pi v\|^2 + \|v - \Pi v\|^2$.

(iv) The trace of Π *is equal to the dimension of its image,* $\operatorname{tr} \Pi = p$.

(v) Π *projects the linear space* $I\!\!R^n$ *on the linear subspace* $\mathrm{L}_p = \langle \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p \rangle$, *which is spanned by the basis vectors* $\boldsymbol{\psi}_1, \ldots \boldsymbol{\psi}_p$, *that is,*

$$\|\boldsymbol{f}^* - \Pi \boldsymbol{f}^*\| = \inf_{\boldsymbol{g} \in \mathrm{L}_p} \|\boldsymbol{f}^* - \boldsymbol{g}\|.$$

(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:*

$$\Lambda_p = \operatorname{diag}\{\underbrace{1, \ldots, 1}_{p}, \underbrace{0, \ldots, 0}_{n-p}\}.$$

Proof. It holds

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

and

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

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

$$\operatorname{tr} \Pi = \operatorname{tr}\, \Psi^\top \left(\Psi\Psi^\top\right)^{-1} \Psi = \operatorname{tr}\, \Psi\Psi^\top \left(\Psi\Psi^\top\right)^{-1} = \operatorname{tr}\, \boldsymbol{I}_p = p.$$

The second property means that Π is a projector in $I\!\!R^n$ and the fourth one means that the dimension of its image space is equal to p. The basis vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p$ are the rows of the matrix Ψ. It is clear that

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

Therefore, the vectors $\boldsymbol{\psi}_j$ are invariants of the operator Π and in particular, all these vectors belong to the image space of this operator. If now $\boldsymbol{g}$ is a vector in L_p, then it can be represented as $\boldsymbol{g} = c_1 \boldsymbol{\psi}_1 + \ldots + c_p \boldsymbol{\psi}_p$ and therefore, $\Pi \boldsymbol{g} = \boldsymbol{g}$ and $\Pi \mathrm{L}_p = \mathrm{L}_p$. Finally, the non-singularity of the matrix $\Psi \Psi^{\top}$ means that the vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p$ forming the rows of Ψ are linearly independent. Therefore, the space L_p spanned by the vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p$ 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 1.3.1. Consider the case of an orthogonal design with $\Psi \Psi^{\top} = I_p$. Specify the projector Π of Lemma 1.3.1 for this situation, particularly its decomposition from (vi).

1.3.3 Quadratic loss and risk of the response estimation

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

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

$$\wp(\widetilde{\boldsymbol{f}}, \boldsymbol{f}^{*}) = \| \widetilde{\boldsymbol{f}} - \boldsymbol{f}^{*} \|^2 = \sum_{i=1}^{n} | f_i - \widetilde{f}_i |^2 .$$

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

$$\mathcal{R}(\widetilde{\boldsymbol{f}}) = \mathbb{E} \wp(\widetilde{\boldsymbol{f}}, \boldsymbol{f}^{*}) = \mathbb{E} \bigl[(\widetilde{\boldsymbol{f}} - \boldsymbol{f}^{*})^{\top} (\widetilde{\boldsymbol{f}} - \boldsymbol{f}^{*}) \bigr].$$

The next result describes the decomposition of the loss. It is important to note that the result is purely geometric and does not rely on the properties of the noise. We distinguish between two cases: when the parametric assumption $\boldsymbol{f}^{*} = \Psi^{\top} \boldsymbol{\theta}^{*}$ is correct and in the general case.

Theorem 1.3.1. *Let* $\widetilde{\boldsymbol{f}} = \Pi \boldsymbol{Y}$ *be the linear response estimator of the vector* $\boldsymbol{f}$. *Then the loss* $\wp(\widetilde{\boldsymbol{f}}, \boldsymbol{f}^{*}) = \| \Pi \boldsymbol{Y} - \boldsymbol{f}^{*} \|^2$ *fulfills*

$$\wp(\widetilde{\boldsymbol{f}}, \boldsymbol{f}^{*}) = \| \boldsymbol{f}^{*} - \Pi \boldsymbol{f}^{*} \|^2 + \| \Pi \boldsymbol{\varepsilon} \|^2 . \tag{1.10}$$

Moreover, if the LPA is correct then $\Pi \boldsymbol{f}^{*} = \boldsymbol{f}^{*}$ *and*

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

Proof. We apply the decomposition (1.9) of the estimate $\widetilde{\boldsymbol{f}}$. It follows

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

This implies the decomposition for the loss of $\widetilde{\boldsymbol{f}}$ by Lemma 1.3.1, (ii).

For the risk, we also need to specify the noise covariance.

Theorem 1.3.2. *Suppose that the errors ε_i from (1.2) are independent with $I\!\!E\,\varepsilon_i = 0$ and $I\!\!E\,\varepsilon_i^2 = \sigma^2$, i.e. $\Sigma = \sigma^2 I_n$. Then the risk $\mathcal{R}(\widetilde{\boldsymbol{f}})$ of the LSE $\widetilde{\boldsymbol{f}}$ fulfills*

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

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

$$\mathcal{R}(\widetilde{\boldsymbol{f}}) = p\sigma^2.$$

Proof. The first term in the decomposition (1.10) is deterministic, and we only need to compute the mean of $\|\Pi\boldsymbol{\varepsilon}\|^2$ applying again Lemma 1.3.1. It holds

$$\begin{aligned}I\!\!E\|\Pi\boldsymbol{\varepsilon}\|^2 &= I\!\!E(\Pi\boldsymbol{\varepsilon})^\top\Pi\boldsymbol{\varepsilon} = I\!\!E\operatorname{tr}\bigl\{\Pi\boldsymbol{\varepsilon}(\Pi\boldsymbol{\varepsilon})^\top\bigr\} = I\!\!E\operatorname{tr}\bigl(\Pi\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\Pi^\top\bigr) \\ &= \operatorname{tr}\bigl\{\Pi I\!\!E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)\Pi\bigr\} = \sigma^2\operatorname{tr}(\Pi^2) = p\sigma^2.\end{aligned}$$

Now consider the case when $\boldsymbol{f}^* = \Psi^\top\boldsymbol{\theta}^*$. By Lemma 1.3.1 $\boldsymbol{f}^* = \Pi\boldsymbol{f}^*$ and and the last two statements of the theorem clearly follow.

1.3.4 * Misspecified "colored noise"

Here we briefly comment on the case when $\boldsymbol{\varepsilon}$ 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 $\widetilde{\boldsymbol{f}} = \Pi\boldsymbol{Y}$ which are simply based on the linearity of the model (1.2) and of the estimate $\widetilde{\boldsymbol{f}}$ itself continue to apply. In particular, the loss $\wp\bigl(\widetilde{\boldsymbol{f}}, \boldsymbol{f}^*\bigr) = \|\widetilde{\boldsymbol{f}} - \boldsymbol{f}^*\|^2$ can again be decomposed as

$$\|\widetilde{\boldsymbol{f}} - \boldsymbol{f}^*\|^2 = \|\boldsymbol{f}^* - \Pi\boldsymbol{f}^*\|^2 + \|\Pi\boldsymbol{\varepsilon}\|^2.$$

Theorem 1.3.3. *Suppose that $I\!\!E\boldsymbol{\varepsilon} = 0$ and $\operatorname{Var}(\boldsymbol{\varepsilon}) = \Sigma_0$. Then the loss $\wp(\widetilde{\boldsymbol{f}}, \boldsymbol{f})$ and the risk $\mathcal{R}(\widetilde{\boldsymbol{f}})$ of the LSE $\widetilde{\boldsymbol{f}}$ fulfill*

$$\begin{aligned} \wp(\widetilde{\boldsymbol{f}}, \boldsymbol{f}^*) &= \|\boldsymbol{f}^* - \Pi \boldsymbol{f}^*\|^2 + \|\Pi \boldsymbol{\varepsilon}\|^2, \\ \mathcal{R}(\widetilde{\boldsymbol{f}}) &= \|\boldsymbol{f}^* - \Pi \boldsymbol{f}^*\|^2 + \operatorname{tr}\bigl(\Pi \Sigma_0 \Pi\bigr). \end{aligned}$$

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

$$\begin{aligned} \wp(\widetilde{\boldsymbol{f}}, \boldsymbol{f}^*) &= \|\Pi \boldsymbol{\varepsilon}\|^2, \\ \mathcal{R}(\widetilde{\boldsymbol{f}}) &= \operatorname{tr}\bigl(\Pi \Sigma_0 \Pi\bigr). \end{aligned}$$

Proof. The decomposition of the loss from Theorem 1.3.2 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

$$I\!E \|\Pi \boldsymbol{\varepsilon}\|^2 = I\!E \operatorname{tr}\bigl[\Pi \boldsymbol{\varepsilon} (\Pi \boldsymbol{\varepsilon})^\top\bigr] = \operatorname{tr}\bigl[\Pi I\!E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) \Pi\bigr] = \operatorname{tr}\bigl(\Pi \Sigma_0 \Pi\bigr)$$

as required.

1.4 Properties of the MLE $\widetilde{\boldsymbol{\theta}}$

In this section we focus on the properties of the quasi MLE $\widetilde{\boldsymbol{\theta}}$ built for the idealized linear Gaussian model $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$. As in the previous section, we do not assume the parametric structure of the underlying model and consider a more general model $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$ with an unknown vector $\boldsymbol{f}^*$ and errors $\boldsymbol{\varepsilon}$ with zero mean and covariance matrix Σ_0 . Due to (1.4), it holds $\widetilde{\boldsymbol{\theta}} = \mathcal{S}\boldsymbol{Y}$ with $\mathcal{S} = \bigl(\Psi\Psi^\top\bigr)^{-1}\Psi$. An important feature of this estimate is its linear dependence on the data. The linear model equation $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$ and linear structure of the estimate $\widetilde{\boldsymbol{\theta}} = \mathcal{S}\boldsymbol{Y}$ allow us for decomposing the vector $\widetilde{\boldsymbol{\theta}}$ into a deterministic and stochastic terms:

$$\widetilde{\boldsymbol{\theta}} = \mathcal{S}\boldsymbol{Y} = \mathcal{S}\bigl(\boldsymbol{f}^* + \boldsymbol{\varepsilon}\bigr) = \mathcal{S}\boldsymbol{f}^* + \mathcal{S}\boldsymbol{\varepsilon}. \tag{1.11}$$

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

1.4.1 Properties of the stochastic component

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

Theorem 1.4.1. *Assume* $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$ *with* $I\!E\boldsymbol{\varepsilon} = 0$ *and* $\operatorname{Var}(\boldsymbol{\varepsilon}) = \Sigma_0$ *. The stochastic component* $\boldsymbol{\delta} = \mathcal{S}\boldsymbol{\varepsilon}$ *in* (1.11) *fulfills*

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

Moreover, if $\Sigma = \Sigma_0 = \sigma^2 \boldsymbol{I}_n$ *, then*

$$W^2 = \sigma^2\bigl(\Psi\Psi^\top\bigr)^{-1}, \qquad \mathbb{E}\|\boldsymbol{\delta}\|^2 = \operatorname{tr}(W^2) = \sigma^2 \operatorname{tr}\bigl[\bigl(\Psi\Psi^\top\bigr)^{-1}\bigr]. \tag{1.12}$$

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

$$\mathbb{E}\widetilde{\boldsymbol{\theta}} = \mathcal{S}\boldsymbol{f}^*, \qquad \operatorname{Var}(\widetilde{\boldsymbol{\theta}}) = W^2.$$

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

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

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

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

Next we use that $\mathbb{E}\|\boldsymbol{\delta}\|^2 = \mathbb{E}\boldsymbol{\delta}^\top\boldsymbol{\delta} = \mathbb{E}\operatorname{tr}(\boldsymbol{\delta}\boldsymbol{\delta}^\top) = \operatorname{tr} W^2$. If $\Sigma = \Sigma_0 = \sigma^2\boldsymbol{I}_n$, then (1.12) follows by simple algebra.

If $\boldsymbol{\varepsilon}$ is a Gaussian vector, then $\boldsymbol{\delta}$ as its linear transformation is Gaussian as well. The properties of $\widetilde{\boldsymbol{\theta}}$ follow directly from the decomposition (1.11).

With $\Sigma_0 \neq \sigma^2\boldsymbol{I}_n$, the variance W^2 can be represented as

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

Exercise 1.4.1. Let $\boldsymbol{\delta}$ be the stochastic component of $\widetilde{\boldsymbol{\theta}}$ built for the misspecified linear model $\boldsymbol{Y} = \Psi^\top\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with $\operatorname{Var}(\boldsymbol{\varepsilon}) = \Sigma$. Let also the true noise variance is Σ_0 . Then $\operatorname{Var}(\widetilde{\boldsymbol{\theta}}) = W^2$ with

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

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

1.4.2 Properties of the deterministic component

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

Theorem 1.4.2. *Let* $\boldsymbol{f}^{*} = \Psi^{\top}\boldsymbol{\theta}^{*}$ *. Then* $\widetilde{\boldsymbol{\theta}} = \mathcal{S}\boldsymbol{Y}$ *with* $\mathcal{S} = \bigl(\Psi\Psi^{\top}\bigr)^{-1}\Psi$ *is unbiased, that is,* $I\!\!E\widetilde{\boldsymbol{\theta}} = \mathcal{S}\boldsymbol{f}^{*} = \boldsymbol{\theta}^{*}$ *.*

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

Now we briefly discuss what happens when the linear parametric assumption is not fulfilled, that is, $\boldsymbol{f}^{*}$ cannot be represented as $\Psi^{\top}\boldsymbol{\theta}^{*}$. In this case it is not yet clear what $\widetilde{\boldsymbol{\theta}}$ 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 $L(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^{*} = \operatorname*{argmax}_{\boldsymbol{\theta}} I\!\!E L(\boldsymbol{\theta}). \tag{1.13}$$

Theorem 1.4.3. *The solution* $\boldsymbol{\theta}^{*}$ *of the optimization problem* (1.13) *is given by*

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

Moreover,

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

In particular, if $\boldsymbol{f}^{*} = \Psi^{\top}\boldsymbol{\theta}^{*}$ *, then* $\boldsymbol{\theta}^{*}$ *follows* (1.13)*.*

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

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

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

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

and the solution $\boldsymbol{\theta}^{*} = \bigl(\Psi\Psi^{\top}\bigr)^{-1}\Psi\boldsymbol{f}^{*}$ which is exactly the expected value of $\widetilde{\boldsymbol{\theta}}$ by Theorem 1.4.1.

Exercise 1.4.2. State the result of Theorems 1.4.2 and 1.4.3 for the MLE $\widetilde{\boldsymbol{\theta}}$ built in the model $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ with $\operatorname{Var}(\boldsymbol{\varepsilon}) = \Sigma$.
Hint: check that the statements continue to apply with $\mathcal{S} = \bigl(\Psi\Sigma^{-1}\Psi^{\top}\bigr)^{-1}\Psi\Sigma^{-1}$.

The last results and the decomposition (1.11) explain the behavior of the estimate $\widetilde{\boldsymbol{\theta}}$ in a very general situation. The considered model is $\boldsymbol{Y} = \boldsymbol{f}^{*} + \boldsymbol{\varepsilon}$. We assume a linear parametric structure and independent homogeneous noise. The estimation procedure means in fact a kind of projection of the data $\boldsymbol{Y}$ on a p-dimensional linear subspace in $I\!\!R^{n}$ spanned by the given basis vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p$. This projection, as a linear operator, can be decomposed into a projection of the deterministic vector $\boldsymbol{f}^{*}$ and a projection of the random noise $\boldsymbol{\varepsilon}$. If the linear parametric assumption $\boldsymbol{f}^{*} \in \langle \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p \rangle$ is correct, that is, $\boldsymbol{f}^{*} = \theta_1^{*}\boldsymbol{\psi}_1 + \ldots + \theta_p^{*}\boldsymbol{\psi}_p$, then this projection keeps $\boldsymbol{f}^{*}$ unchanged and only the random noise is reduced via this projection. If $\boldsymbol{f}^{*}$ cannot be exactly expanded using the basis $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p$, then the procedure recovers the projection of $\boldsymbol{f}^{*}$ onto this subspace. The latter projection can be written as $\Psi^{\top}\boldsymbol{\theta}^{*}$ and the vector $\boldsymbol{\theta}^{*}$ can be viewed as the target of estimation.

1.4.3 * Risk of estimation. R-efficiency

This section briefly discusses how the obtained properties of the estimate $\widetilde{\boldsymbol{\theta}}$ can be used to evaluate the risk of estimation. A particularly important question is the optimality of the MLE $\widetilde{\boldsymbol{\theta}}$. The main result of the section claims that $\widetilde{\boldsymbol{\theta}}$ 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 $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$, that is, the linear parametric assumption $\boldsymbol{f}^{*} = \Psi^{\top}\boldsymbol{\theta}^{*}$ is exactly fulfilled and the noise $\boldsymbol{\varepsilon}$ is homogeneous: $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Later we extend the result to the case when the LPA $\boldsymbol{f}^{*} = \Psi^{\top}\boldsymbol{\theta}^{*}$ 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 $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$, the estimate $\widetilde{\boldsymbol{\theta}}$ is also normal with mean $\boldsymbol{\theta}^{*}$ and the variance $W^2 = \sigma^2 \mathcal{S}\mathcal{S}^{\top} = \sigma^2 \bigl(\Psi\Psi^{\top}\bigr)^{-1}$. 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 $\widetilde{\boldsymbol{\theta}}$ 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 $\mathbb{F}(\boldsymbol{\theta})$ for the log-likelihood $L(\boldsymbol{\theta})$ is defined as the variance of $\nabla L(\boldsymbol{\theta})$ under $I\!\!P_{\boldsymbol{\theta}}$.

Theorem 1.4.4 (Gauss-Markov). *Let* $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. *Then* $\widetilde{\boldsymbol{\theta}}$ *is R-efficient estimate of* $\boldsymbol{\theta}^{*}$ *:* $I\!\!E\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^{*}$,

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

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

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

Proof. Theorems 1.4.1 and 1.4.2 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 $\boldsymbol{\theta}$

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

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

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

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

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

and in view of $\operatorname{Var}(\boldsymbol{Y}) = \Sigma = \sigma^{2}\boldsymbol{I}_{n}$, it holds

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

as required.

The R-efficiency $\tilde{\boldsymbol{\theta}}$ follows from the Cramér-Rao efficiency bound because $\bigl\{\operatorname{Var}(\tilde{\boldsymbol{\theta}})\bigr\}^{-1} = \operatorname{Var}\bigl\{\nabla L(\boldsymbol{\theta})\bigr\}$. 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 $\mathbb{E}\bigl\{\tilde{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})^{\top}\bigr\} = 0$. This, in turns, implies $\operatorname{Var}(\hat{\boldsymbol{\theta}}) = \operatorname{Var}(\tilde{\boldsymbol{\theta}}) + \operatorname{Var}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) \geq \operatorname{Var}(\tilde{\boldsymbol{\theta}})$. So, it remains to check the orthogonality of $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}$. Let $\hat{\boldsymbol{\theta}} = A\boldsymbol{Y}$ 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 $\mathbb{E}\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*}$ 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 $\tilde{\boldsymbol{\theta}} = \mathcal{S}\boldsymbol{Y}$ yielding $\mathcal{S}\Psi^{\top} = I_{p}$. Next, in view of $\mathbb{E}\hat{\boldsymbol{\theta}} = \mathbb{E}\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^{*}$

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

and the assertion follows.

Exercise 1.4.3. Check the details of the proof of the theorem. Show that the statement $\operatorname{Var}(\hat{\boldsymbol{\theta}}) \geq \operatorname{Var}(\tilde{\boldsymbol{\theta}})$ only uses that $\hat{\boldsymbol{\theta}}$ is unbiased and that $\mathbb{E}\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*}$ and $\operatorname{Var}(\boldsymbol{Y}) = \sigma^{2}\boldsymbol{I}_{n}$.

Exercise 1.4.4. 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

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

A colored noise

The majority of the presented results continue to apply in the case of heterogeneous and even dependent noise with $\operatorname{Var}(\boldsymbol{\varepsilon}) = \Sigma_0$. The key facts behind this extension are the decomposition (1.11) and the properties of the stochastic component $\boldsymbol{\delta}$ from Section 1.4.1: $\boldsymbol{\delta} \sim \mathcal{N}(0, W^2)$. 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 1.4.5. State and prove the analog of Theorem 1.4.4 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 $\boldsymbol{f}^* = \Psi^\top \boldsymbol{\theta}^*$ and for the case when the identity $\boldsymbol{f}^* = \Psi^\top \boldsymbol{\theta}$ 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 $\boldsymbol{f}^*$ by $\Psi^\top \boldsymbol{\theta}$. We already know from the results of Section 1.4.1 and 1.4.2 that the estimate $\widetilde{\boldsymbol{\theta}}$ is also normal with mean $\boldsymbol{\theta}^* = \mathcal{S}\boldsymbol{f}^* = \bigl(\Psi\Psi^\top\bigr)^{-1}\Psi\boldsymbol{f}^*$ and the variance $W^2 = \sigma^2 \mathcal{S}\mathcal{S}^\top = \sigma^2\bigl(\Psi\Psi^\top\bigr)^{-1}$.

Theorem 1.4.5. *Assume $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Let $\boldsymbol{\theta}^* = \mathcal{S}\boldsymbol{f}^*$. Then $\widetilde{\boldsymbol{\theta}}$ is R-efficient estimate of $\boldsymbol{\theta}^*$: $I\!\!E\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$,*

$$I\!\!E\bigl[\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr)\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr)^\top\bigr] = \operatorname{Var}\bigl(\widetilde{\boldsymbol{\theta}}\bigr) = D^{-2},$$

and for any unbiased linear estimate $\widehat{\boldsymbol{\theta}}$ satisfying $I\!\!E_{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}} \equiv \boldsymbol{\theta}$, it holds

$$\operatorname{Var}\bigl(\widehat{\boldsymbol{\theta}}\bigr) \ge \operatorname{Var}\bigl(\widetilde{\boldsymbol{\theta}}\bigr) = D^{-2}.$$

Proof. The proofs only utilize that $\widetilde{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}^*, W^2)$ with $W^2 = D^{-2}$. The only small remark concerns the equality $\operatorname{Var}\bigl[\nabla L(\boldsymbol{\theta})\bigr] = D^2$ from Theorem 1.4.4.

Exercise 1.4.6. Check the identity $\operatorname{Var}\bigl[\nabla L(\boldsymbol{\theta})\bigr] = D^2$ from Theorem 1.4.4 for $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma_0)$.

1.4.4 * The case of a misspecified noise

Here we again consider the linear parametric assumption $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$. However, contrary to the previous section, we admit that the noise $\boldsymbol{\varepsilon}$ is not homogeneous normal: $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma_0)$ while our estimation procedure is the quasi MLE based on the assumption of noise homogeneity $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. We already know that the estimate $\widetilde{\boldsymbol{\theta}}$ is unbiased with mean $\boldsymbol{\theta}^{*}$ and variance $W^2 = \mathcal{S}\Sigma_0\mathcal{S}^{\top}$, where $\mathcal{S} = \bigl(\Psi\Psi^{\top}\bigr)^{-1}\Psi$. This gives

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

The question is whether the estimate $\widetilde{\boldsymbol{\theta}}$ based on the misspecified distributional assumption is efficient. The Cramér-Rao result delivers the lower bound for the quadratic risk in form of $\operatorname{Var}(\widetilde{\boldsymbol{\theta}}) \geq \bigl[\operatorname{Var}\bigl(\nabla L(\boldsymbol{\theta})\bigr)\bigr]^{-1}$. We already know that the use of the correctly specified covariance matrix of the errors leads to an R-efficient estimate $\widetilde{\boldsymbol{\theta}}$. 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 1.4.6. *Let* $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma_0)$. *Then*

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

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

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

$$\operatorname{Var}(\widetilde{\boldsymbol{\theta}}) = \operatorname{Var}(\widetilde{\boldsymbol{\theta}}_0) + \operatorname{Var}(\widetilde{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}_0);$$

cf. with the proof of Gauss-Markov-Theorem 1.4.4.

Exercise 1.4.7. Compare directly the variances of $\widetilde{\boldsymbol{\theta}}$ and of $\widetilde{\boldsymbol{\theta}}_0$.

1.5 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 its Hessian is a deterministic symmetric matrix D^2 :

$$\nabla^2 L(\boldsymbol{\theta}) \equiv -D^2.$$

The second order Taylor expansion implies for any $\boldsymbol{\theta}^\circ, \boldsymbol{\theta}$

$$\begin{aligned} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^\circ) &= (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top \nabla L(\boldsymbol{\theta}^\circ) - (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top D^2 (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)/2, \\ \nabla L(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta}^\circ) &= -D^2(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ). \end{aligned} \tag{1.14}$$

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

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

The next result describes some properties of the estimate $\widetilde{\boldsymbol{\theta}}$ 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 $\boldsymbol{\zeta} = \nabla L(\boldsymbol{\theta}^*)$.

Theorem 1.5.1. *Let $L(\boldsymbol{\theta})$ be quadratic for a matrix $D^2 > 0$. Then for any $\boldsymbol{\theta}^\circ$*

$$\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ = D^{-2} \nabla L(\boldsymbol{\theta}^\circ). \tag{1.15}$$

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

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

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

$$\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = D^{-2} \boldsymbol{\zeta} \tag{1.16}$$

with $\boldsymbol{\zeta} \stackrel{\text{def}}{=} \nabla L(\boldsymbol{\theta}^*)$.

Proof. The extremal point equation $\nabla L(\widetilde{\boldsymbol{\theta}}) = 0$ together with the expansion (1.14) for the quadratic function $L(\boldsymbol{\theta})$ yields (1.15).

Now we study the moments of the qMLE $\widetilde{\boldsymbol{\theta}}$.

Theorem 1.5.2. *Let $L(\boldsymbol{\theta})$ be quadratic for a matrix $D^2 > 0$, and $\boldsymbol{\zeta} \stackrel{\text{def}}{=} \nabla L(\boldsymbol{\theta}^*)$. It holds with $V^2 = \operatorname{Var}(\boldsymbol{\zeta}) = I\!\!E \boldsymbol{\zeta} \boldsymbol{\zeta}^\top$*

$$\begin{aligned} I\!\!E \widetilde{\boldsymbol{\theta}} &= \boldsymbol{\theta}^* \\ \operatorname{Var}(\widetilde{\boldsymbol{\theta}}) &= D^{-2} V^2 D^{-2}. \end{aligned}$$

Proof. The equation (1.14) with $\boldsymbol{\theta}^{\circ} = \boldsymbol{\theta}^{*}$ implies for any $\boldsymbol{\theta}$

$$\nabla L(\boldsymbol{\theta}) = \nabla L(\boldsymbol{\theta}^{*}) - D^{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{*}) = \boldsymbol{\zeta} - D^{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{*}). \tag{1.17}$$

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 $\mathbb{E}\boldsymbol{\zeta} = 0$.

Finally we study the properties of the maximum log-likelihood $L(\widetilde{\boldsymbol{\theta}}) = \sup_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$. In particular, we compare the maximum log-likelihood with its value at the true point $\boldsymbol{\theta}^{*}$. The difference $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^{*})$ is called the *excess*, it plays an important role in statistical inference.

Theorem 1.5.3. *Let $L(\boldsymbol{\theta})$ be quadratic for a matrix $D^{2} > 0$. For any $\boldsymbol{\theta}$, it holds*

$$L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}) = (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\top} D^{2} (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})/2 = \| D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \|^{2}/2. \tag{1.18}$$

Moreover, it holds for the excess $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{*}) \overset{\text{def}}{=} L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^{*})$ fulfills

$$2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{*}) = (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{*})^{\top} D^{2} (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{*}) = \boldsymbol{\zeta}^{\top} D^{-2} \boldsymbol{\zeta} = \| \boldsymbol{\xi} \|^{2} \tag{1.19}$$

with the normalized score

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

Proof. To show (1.18), apply again the property (1.14) with $\boldsymbol{\theta}^{\circ} = \widetilde{\boldsymbol{\theta}}$:

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

Here we used that $\nabla L(\widetilde{\boldsymbol{\theta}}) = 0$ because $\widetilde{\boldsymbol{\theta}}$ is an extreme point of $L(\boldsymbol{\theta})$. The last result (1.19) is a special case with $\boldsymbol{\theta} = \boldsymbol{\theta}^{*}$ in view of (1.16).

This theorem delivers an important message: the main properties of the MLE $\widetilde{\boldsymbol{\theta}}$ 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 (1.17) with $\boldsymbol{\theta}^{\circ} = \boldsymbol{\theta}^{*}$ can be rewritten as

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

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

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

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

$$\boldsymbol{Y} = D\boldsymbol{\theta}^* + \boldsymbol{\xi}, \tag{1.20}$$

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

Theorem 1.5.4. *Let $L(\boldsymbol{\theta})$ be quadratic with a non-degenerated matrix D^2. Then $\boldsymbol{Y} \stackrel{\mathrm{def}}{=} D^{-1}\bigl\{\nabla L(\boldsymbol{\theta}) + D^2\boldsymbol{\theta}\bigr\}$ does not depend on $\boldsymbol{\theta}$ and $L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$ is the quasi log-likelihood ratio for the linear Gaussian model (1.20) with $\boldsymbol{\xi}$ standard normal. It is the true log-likelihood if and only if $\boldsymbol{\zeta} \sim \mathcal{N}(0, D^2)$.*

Proof. The model (1.20) with $\boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{I}_p)$ leads to the log-likelihood ratio

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

in view of the definition of $\boldsymbol{Y}$. The definition (1.14) 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 $\boldsymbol{\xi} = D^{-1}\boldsymbol{\zeta}$ is standard normal.

1.6 Inference in linear regression model based on the maximum likelihood

All the results presented above for linear models were based on the explicit representation of the (quasi) MLE $\widetilde{\boldsymbol{\theta}}$. 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* $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$. We also show how these results can be used to study the accuracy of the MLE $\widetilde{\boldsymbol{\theta}}$, 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 $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ claims that the distribution of $2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ is chi-squared with p degrees of freedom χ^2_p and it does not depend on the noise covariance; see Section 1.6.

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

$$L(\boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{Y} - \Psi^{\top}\boldsymbol{\theta})^{\top}\Sigma^{-1}(\boldsymbol{Y} - \Psi^{\top}\boldsymbol{\theta}) + R, \tag{1.21}$$

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 1.5 which are only based on the geometric (quadratic) structure of the log-likelihood $L(\boldsymbol{\theta})$: the true data distribution can be arbitrary. Therefore, our study of the properties of the maximum likelihood allows a possible model misspecification. We only assume that $\boldsymbol{Y} = \boldsymbol{f}^{*} + \boldsymbol{\varepsilon}$ with $I\!\!E\boldsymbol{\varepsilon} = 0$. The "true" parameter $\boldsymbol{\theta}^{*}$ is defined by maximization of the expectation $I\!\!E L(\boldsymbol{\theta})$: $\boldsymbol{\theta}^{*} \overset{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}} I\!\!E L(\boldsymbol{\theta})$.

Theorem 1.6.1. *Consider $L(\boldsymbol{\theta})$ from (1.21). For any $\boldsymbol{\theta}$, it holds with $D^2 = \Psi\Sigma^{-1}\Psi^{\top}$*

$$L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\top} D^2 (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})/2. \tag{1.22}$$

The "target" $\boldsymbol{\theta}^{*}$ fulfills $\boldsymbol{\theta}^{*} = D^{-2}\Psi\Sigma^{-1}\boldsymbol{f}^{*}$ for $\boldsymbol{f}^{*} = I\!\!E\boldsymbol{Y}$, and

$$2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{*}) = \boldsymbol{\zeta}^{\top} D^{-2}\boldsymbol{\zeta} = \|\boldsymbol{\xi}\|^2 \tag{1.23}$$

with

$$\begin{aligned} \boldsymbol{\zeta} &\overset{\text{def}}{=} \nabla L(\boldsymbol{\theta}^{*}) = \Psi\Sigma^{-1}\boldsymbol{\varepsilon}, \\ \boldsymbol{\xi} &\overset{\text{def}}{=} D^{-1}\boldsymbol{\zeta} = D^{-1}\Psi\Sigma^{-1}\boldsymbol{\varepsilon}. \end{aligned} \tag{1.24}$$

Proof. The results (1.22) and (1.23) follow from Theorem 1.5.1; see (1.18) and (1.19). Further, by direct calculus

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

Now we use that $I\!\!E\boldsymbol{\zeta} = 0$; see Theorem 1.5.2. Therefore, $\boldsymbol{\zeta} = \boldsymbol{\zeta} - I\!\!E\boldsymbol{\zeta} = \Psi\Sigma^{-1}(\boldsymbol{Y} - I\!\!E\boldsymbol{Y})$, and the result (1.24) follows.

Exercise 1.6.1. If $\Sigma = \sigma^2 I_n$ then the fitted log-likelihood is proportional to the quadratic loss $\|\widetilde{\boldsymbol{f}} - \boldsymbol{f}_{\boldsymbol{\theta}}\|^2$ for $\widetilde{\boldsymbol{f}} = \Psi^{\top}\widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{f}_{\boldsymbol{\theta}} = \Psi^{\top}\boldsymbol{\theta}$:

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

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

Theorem 1.6.2. *Let $L(\boldsymbol{\theta})$ from (1.21) be the log-likelihood for the model $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$. Then $\boldsymbol{\xi} = D^{-1}\boldsymbol{\zeta} \sim \mathcal{N}(0, I_p)$ and $2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{*}) \sim \chi^2_p$ is chi-squared with p degrees of freedom.*

Proof. By (1.24), $\boldsymbol{\zeta}$ is a linear transformation of a Gaussian vector $\boldsymbol{Y}$ and thus it is Gaussian as well. By Theorem 1.5.1, $I\!\!E\boldsymbol{\zeta} = 0$. Moreover, $\operatorname{Var}(\boldsymbol{\varepsilon}) = \Sigma$ implies

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

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

The last result $2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{*}) \sim \chi^2_p$ 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 1.6.2. Check that the linear transformation $\check{\boldsymbol{Y}} = \Sigma^{-1/2}\boldsymbol{Y}$ of the data does not change the value of the log-likelihood ratio $L(\boldsymbol{\theta}, \boldsymbol{\theta}^{*})$ and hence, of the maximum likelihood $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{*})$.
Hint: use the representation

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

and check that the transformed data $\check{\boldsymbol{Y}}$ is described by the model $\check{\boldsymbol{Y}} = \check{\Psi}^{\top}\boldsymbol{\theta}^{*} + \check{\boldsymbol{\varepsilon}}$ 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 1.6.3. Assume homogeneous noise in (1.21) with $\Sigma = \sigma^2 I_n$. Then it holds

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

where $\Pi = \Psi^{\top}\bigl(\Psi\Psi^{\top}\bigr)^{-1}\Psi$ is the projector in $I\!\!R^n$ on the subspace spanned by the vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p$.
Hint: use that $\boldsymbol{\zeta} = \sigma^{-2}\Psi\boldsymbol{\varepsilon}$, $D^2 = \sigma^{-2}\Psi\Psi^{\top}$, and

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

We write the result of Theorem 1.6.1 in the form $2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{*}) \sim \chi^2_p$, where χ^2_p 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

$$\mathcal{E}(\mathfrak{z}) = \bigl\{\boldsymbol{\theta} : L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \le \mathfrak{z}\bigr\} = \Bigl\{\boldsymbol{\theta} : \sup_{\boldsymbol{\theta}'} L(\boldsymbol{\theta}') - L(\boldsymbol{\theta}) \le \mathfrak{z}\Bigr\}. \tag{1.25}$$

Theorem 1.6.3. *Assume* $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$ *and consider the MLE* $\widetilde{\boldsymbol{\theta}}$. *Define* $\mathfrak{z}_{\alpha}$ *by* $P\bigl(\chi^2_p > 2\mathfrak{z}_{\alpha}\bigr) = \alpha$. *Then* $\mathcal{E}(\mathfrak{z}_{\alpha})$ *from* (1.25) *is an* α*-confidence set for* $\boldsymbol{\theta}^{*}$.

Exercise 1.6.4. Let $D^2 = \Psi\Sigma^{-1}\Psi^{\top}$. Check that the likelihood-based CS $\mathcal{E}(\mathfrak{z}_{\alpha})$ and estimate-based CS $E(z_{\alpha}) = \{\boldsymbol{\theta} : \|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\| \le z_{\alpha}\}$, $z_{\alpha}^2 = 2\mathfrak{z}_{\alpha}$, coincide in the case of the linear modeling:

$$\mathcal{E}(\mathfrak{z}_{\alpha}) = \bigl\{\boldsymbol{\theta} : \bigl\|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\bigr\|^2 \le 2\mathfrak{z}_{\alpha}\bigr\}.$$

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 $L(\widetilde{\theta}, \theta^{*})$ 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 1.6.1 states that the distribution of $2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{*})$ 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 1.6.4. *Assume* $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$. *Then for every* $\mathtt{x} > 0$

$$\begin{aligned} & I\!\!P\bigl(2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{*}) > p + 2\sqrt{\mathtt{x}\, p} + 2\mathtt{x}\bigr) \\ & \quad = I\!\!P\bigl(\bigl\|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{*})\bigr\|^2 > p + 2\sqrt{\mathtt{x}\, p} + 2\mathtt{x}\bigr) \le \exp(-\mathtt{x}). \end{aligned} \tag{1.26}$$

Proof. Define $\boldsymbol{\xi} \overset{\text{def}}{=} D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{*})$. By Theorem 1.4.4 $\boldsymbol{\xi}$ is standard normal vector in $I\!\!R^p$ and by Theorem 1.6.1 $2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{*}) = \|\boldsymbol{\xi}\|^2$. Now the statement (1.26) follows from the general deviation bound for the Gaussian quadratic forms; see Corollary C.1.2.

The main message of this result can be explained as follows: the deviation probability that the estimate $\widetilde{\boldsymbol{\theta}}$ does not belong to the elliptic set $E(z) = \{\boldsymbol{\theta} : \|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\| \le z\}$ starts to vanish when z^2 exceeds the dimensionality p of the parameter space. Similarly, the coverage probability that the true parameter $\boldsymbol{\theta}^{*}$ is not covered by the confidence set $\mathcal{E}(\mathfrak{z})$ starts to vanish when $2\mathfrak{z}$ exceeds p.

Corollary 1.6.1. *Assume* $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$. *Then for every* $\mathtt{x} > 0$, *it holds with* $2\mathfrak{z} = p + 2\sqrt{\mathtt{x}\, p} + 2\mathtt{x}$

$$I\!\!P\bigl(\mathcal{E}(\mathfrak{z}) \not\ni \boldsymbol{\theta}^{*}\bigr) \le \exp(-\mathtt{x}).$$

Exercise 1.6.5. Compute $\mathfrak{z}$ ensuring the covering of 95% in the dimension $p = 1, 2, 10, 20$.

1.6.1 * A misspecified LPA

Now we discuss the behavior of the fitted log-likelihood for the misspecified linear parametric assumption $\mathbb{E}\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^*$. Let the response function $\boldsymbol{f}^*$ not be linearly expandable as $\boldsymbol{f}^* = \Psi^\top \boldsymbol{\theta}^*$. Following to Theorem 1.4.3, define $\boldsymbol{\theta}^* = \mathcal{S}\boldsymbol{f}^*$ with $\mathcal{S} = \bigl(\Psi\Sigma^{-1}\Psi^\top\bigr)^{-1}\Psi\Sigma^{-1}$. This point provides the best approximation of the nonlinear response $\boldsymbol{f}^*$ by a linear parametric fit $\Psi^\top\boldsymbol{\theta}$.

Theorem 1.6.5. *Assume* $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$. *Let* $\boldsymbol{\theta}^* = \mathcal{S}\boldsymbol{f}^*$. *Then* $\widetilde{\boldsymbol{\theta}}$ *is an R-efficient estimate of* $\boldsymbol{\theta}^*$ *and*

$$2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \boldsymbol{\zeta}^\top D^{-2}\boldsymbol{\zeta} = \|\boldsymbol{\xi}\|^2 \sim \chi^2_p,$$

where $D^2 = \Psi\Sigma^{-1}\Psi^\top$, $\boldsymbol{\zeta} = \nabla L(\boldsymbol{\theta}^*) = \Psi\Sigma^{-1}\boldsymbol{\varepsilon}$, $\boldsymbol{\xi} = D^{-1}\boldsymbol{\zeta}$ *is standard normal vector in* $\mathbb{R}^p$ *and* χ^2_p *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* $\boldsymbol{\theta}^*$ *and the bound of Corollary 1.6.1 applies.*

Exercise 1.6.6. Prove the result of Theorem 1.6.5.

1.6.2 * 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 1.6.1 and 1.6.3 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 1.6.6. *Let* $\widetilde{\boldsymbol{\theta}}$ *be built for the model* $\boldsymbol{Y} = \Psi^\top\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$, *while the true noise covariance is* Σ_0: $\mathbb{E}\boldsymbol{\varepsilon} = 0$ *and* $\operatorname{Var}(\boldsymbol{\varepsilon}) = \Sigma_0$. *Then*

$$\begin{aligned} \mathbb{E}\widetilde{\boldsymbol{\theta}} &= \boldsymbol{\theta}^*, \\ \operatorname{Var}\bigl(\widetilde{\boldsymbol{\theta}}\bigr) &= D^{-2}W^2D^{-2}, \end{aligned}$$

where

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

Further,

$$2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \| D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \|^2 = \| \boldsymbol{\xi} \|^2, \tag{1.27}$$

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

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

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

Proof. The moments of $\widetilde{\boldsymbol{\theta}}$ have been computed in Theorem 1.5.1 while the equality $2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \| D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \|^2 = \| \boldsymbol{\xi} \|^2$ is given in Theorem 1.6.1. Next, $\boldsymbol{\zeta} = \nabla L(\boldsymbol{\theta}^*) = \Psi \Sigma^{-1} \boldsymbol{\varepsilon}$ and

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

This implies that

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

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 $\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = D^{-2} \boldsymbol{\zeta}$ are Gaussian as well.

Exercise 1.6.7. 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 (1.27) is zero mean and its norm has a similar behavior as in the case of the correct noise specification: the probability $\mathbb{P}\bigl(\| \boldsymbol{\xi} \| > z\bigr)$ starts to degenerate when z^2 exceeds $\mathbb{E} \| \boldsymbol{\xi} \|^2$. A general bound from Theorem C.1.1 in Section B.1 implies the following bound for the coverage probability.

Corollary 1.6.2. *Under the conditions of Theorem 1.6.6, for every* $\mathtt{x} > 0$ *, it holds with* $\mathtt{p} = \operatorname{tr}(B)$ *,* $\mathtt{v}^2 = \operatorname{tr}(B^2)$ *, and* $\lambda = \| B \|_{\infty}$

$$\mathbb{P}\bigl(2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) > \mathtt{p} + 2\mathtt{v}\mathtt{x}^{1/2} + 2\lambda\mathtt{x}\bigr) \le \exp(-\mathtt{x}).$$

Exercise 1.6.8. Show that an overestimation of the noise in the sense $\Sigma \ge \Sigma_0$ 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^2_p, then $\mathbb{P}\bigl(\mathcal{E}(\mathfrak{z}_\alpha) \not\ni \boldsymbol{\theta}^*\bigr) \le \alpha$.

1.7 * Approximation spaces

This section discusses popular methods of approximating the unknown function $f(\cdot)$ using a linear feature representation.

1.7.1 Polynomial approximation

It is well known that any smooth function $f(\cdot)$ can be approximated by a polynomial. Moreover, the larger smoothness of $f(\cdot)$ is the better the accuracy of approximation. The Taylor expansion yields an approximation in the form

$$f(x) \approx \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_m x^m. \tag{1.28}$$

Such an approximation is very natural, however, it is rarely used in statistical applications. The main reason is that the different power functions $\psi_m(x) = x^m$ are highly correlated between each other. This makes difficult to identify the corresponding coefficients. Instead one can use different polynomial systems which fulfill certain orthogonality conditions.

We say that $f(x)$ is a polynomial of degree m, if it can be represented in the form (1.28) with $\theta_m \neq 0$. Any sequence $1, \psi_1(x), \ldots, \psi_m(x)$ of such polynomials yields a basis in the vector space of polynomials of degree m.

Exercise 1.7.1. Let for each $j \leq m$ a polynomial of degree j be fixed. Then any polynomial $P_m(x)$ of degree m can be represented in a unique way in the form

$$P_m(x) = c_0 + c_1\psi_1(x) + \ldots + c_m\psi_m(x)$$

Hint: define $c_m = P_m^{(m)}/\psi_m^{(m)}$ and apply induction to $P_m(x) - c_m\psi_m(x)$.

1.7.2 Orthogonal polynomials

Let μ be any measure on the real line satisfying the condition

$$\int x^m \mu(dx) < \infty, \tag{1.29}$$

for any integer m. This enables us to define the scalar product for two polynomial functions f, g by

$$\langle f, g \rangle \overset{\text{def}}{=} \int f(x)g(x)\mu(dx).$$

With such a Hilbert structure we aim to define an orthonormal polynomial system of polynomials ψ_m of degree m for $m = 0, 1, 2, \ldots$ such that

$$\langle \psi_j, \psi_m \rangle = \delta_{j,m} = \mathrm{1\!I}(j = m), \qquad j, m = 0, 1, 2, \ldots.$$

Theorem 1.7.1. *Given a measure μ satisfying the condition (1.29) there exists unique orthonormal polynomial system $\psi_1, \psi_2, \ldots$. Any polynomial P_m of degree m can be represented as*

$$P_m(x) = a_0 + a_1\psi_1(x) + \ldots + a_m\psi_m(x)$$

with

$$a_j = \bigl\langle P_m, \psi_j \bigr\rangle. \tag{1.30}$$

Proof. We construct the function ψ_m successively. The function ψ_0 is a constant defined by

$$\psi_0^2 \int \mu(dx) = 1.$$

Suppose now that the orthonormal polynomials $\psi_1, \ldots, \psi_{m-1}$ have been already constructed. Define the coefficients

$$a_j \overset{\text{def}}{=} \int x^m \psi_j(x)\mu(dx), \qquad j = 0, 1, \ldots, m-1,$$

and consider the function

$$g_m(x) \overset{\text{def}}{=} x^m - a_0\psi_0 - a_1\psi_1(x) - \ldots - a_{m-1}\psi_{m-1}(x).$$

This is obviously a polynomial of degree m. Moreover, by orthonormality of the ψ_j 's for $j < m$

$$\int g_m(x)\psi_j(x)\mu(dx) = \int x^m \psi_j(x)\mu(dx) - a_j \int \psi_j^2(x)\mu(dx) = 0.$$

So, one can define ψ_m by normalization of g_m:

$$\psi_m(x) \overset{\text{def}}{=} \bigl\langle g_m, g_m \bigr\rangle^{-1/2} g_m(x).$$

One can also easily see that such defined ψ_m is only polynomial of degree m which is orthogonal to ψ_j for $j < m$ and fulfills $\bigl\langle \psi_m, \psi_m \bigr\rangle = 1$, because the number of constraints is equal to the number of coefficients $\theta_0, \ldots, \theta_m$ of $\psi_m(x)$.

Let now P_m be a polynomial of degree m. Define the coefficient a_m by (1.30). Similarly to above one can show that

$$P_m(x) - \bigl\{ a_0 + a_1\psi_1(x) + \ldots + a_m\psi_m(x) \bigr\} \equiv 0$$

which implies the second claim.

Exercise 1.7.2. Let $\{\psi_m\}$ be an orthonormal polynomial system. Show that for any polynomial $P_j(x)$ of degree $j < m$, it holds

$$\langle P_j, \psi_m \rangle = 0.$$

Finite approximation and the associated kernel

Let f be a function satisfying

$$\int f^2(x)\mu(dx) < \infty. \tag{1.31}$$

Then the scalar product $a_j = \langle f, \psi_j \rangle$ is well defined for all $j \geq 0$ leading for each $m \geq 1$ to the following approximation:

$$\begin{aligned} f_m(x) &\overset{\text{def}}{=} \sum_{j=0}^{m} a_j \psi_j(x) = \sum_{j=0}^{m} \int f(u)\psi_j(u)\mu(du)\psi_j(x) \\ &= \int f(u)\Phi_m(x,u)\mu(du) \end{aligned} \tag{1.32}$$

with

$$\Phi_m(x,u) = \sum_{j=0}^{m} \psi_j(x)\psi_j(u).$$

Completeness

The accuracy of approximation of f by f_m with m growing is one of the central questions in the approximation theory. The answer depends on the regularity of the function f and on choice of the system $\{\psi_m\}$. Let $\mathcal{F}$ be a linear space of functions f on the real line satisfying (1.31). We say that the basis system $\{\psi_m(x)\}$ is *complete* in $\mathcal{F}$ if the identities $\langle f, \psi_m \rangle = 0$ for all $m \geq 0$ imply $f \equiv 0$. As $\psi_m(x)$ is a polynomial of degree m, this definition is equivalent to the condition

$$\langle f, x^m \rangle = 0, \quad m = 0, 1, 2, \ldots \Longleftrightarrow f \equiv 0.$$

Squared bias and accuracy of approximation

Let $f \in \mathcal{F}$ be a function in L_2 satisfying (1.31), and let $\{\psi_m\}$ be a complete basis. Consider the error $f(x) - f_m(x)$ of the finite approximation $f_m(x)$ from (1.32). The Parseval identity yields

$$\int f^2(x)\mu(dx) = \sum_{m=0}^{\infty} a_m^2.$$

This yields that the finite sums of $\sum_{j=0}^{m} a_j^2$ converge to the infinite sum $\sum_{m=0}^{\infty} a_m^2$ and the remainder $b_m = \sum_{j=m+1}^{\infty} a_j^2$ tends to zero with m:

$$b_m \overset{\text{def}}{=} \langle f - f_m \rangle = \int \bigl| f(x) - f_m(x) \bigr|^2 \mu(dx) = \sum_{j=m+1}^{\infty} a_j^2 \to 0$$

as $m \to \infty$. The value b_m is often called the *squared bias.*

1.8 Piecewise methods and splines

This section discusses piecewise polynomial methods of approximation of the univariate regression functions.

1.8.1 Piecewise constant estimation

Any continuous function can be locally approximated by a constant. This naturally leads to the basis consisting of piecewise constant functions. Let $A_1, \ldots, A_K$ be a *non-overlapping partition* of the design space $\mathcal{X}$:

$$\mathcal{X} = \bigcup_{k=1,\ldots,K} A_k \, , \qquad A_k \cap A_{k'} = \emptyset, \; k \neq k'. \tag{1.33}$$

We approximate the function f by a finite sum

$$f(x) \approx f(x, \boldsymbol{\theta}) = \sum_{k=1}^{K} \theta_k \, \mathbb{1}(x \in A_k). \tag{1.34}$$

Here $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ with $p = K$. A nice feature of this approximation is that the basis indicator functions $\psi_1, \ldots, \psi_K$ are orthogonal because they have non-overlapping supports. For the case of independent errors, this makes the computation of the qMLE $\widetilde{\boldsymbol{\theta}}$ very simple. In fact, every coefficient $\widetilde{\theta}_k$ can be estimated independently of the others. Indeed, the general formula yields

$$\begin{aligned} \widetilde{\boldsymbol{\theta}} &= \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} L(Y_i - f(X_i, \boldsymbol{\theta})) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta} = (\theta_k)} \sum_{k=1}^{K} \sum_{X_i \in A_k} L(Y_i - \theta_k). \end{aligned} \tag{1.35}$$

Exercise 1.8.1. Show that $\widetilde{\theta}_k$ can be obtained by the constant approximation of the data Y_i for $X_i \in A_k$:

$$\widetilde{\theta}_k = \operatorname*{argmax}_{\theta_k} \sum_{X_i \in A_k} L(Y_i - \theta_k), \qquad k = 1, \ldots, K. \tag{1.36}$$

A similar formula can be obtained for the target $\boldsymbol{\theta}^{*} = (\theta_k^{*}) = \operatorname{argmax}_{\boldsymbol{\theta}} I\!\!E L(\boldsymbol{\theta})$:

$$\theta_k^{*} = \operatorname*{argmax}_{\theta_k} \sum_{X_i \in A_k} I\!\!E L(Y_i - \theta_k), \qquad m = 1, \ldots, K.$$

The estimator $\widetilde{\boldsymbol{\theta}}$ can be computed explicitly in some special cases. In particular, if f corresponds a density of a normal distribution, then the resulting estimator $\widetilde{\theta}_k$ is nothing but the mean of observations Y_i over the piece A_k. For the Laplacian errors, the solution is the median of the observations over A_k. First we consider the case of Gaussian likelihood.

Theorem 1.8.1. *Let* $L(y) = -y^2/(2\sigma^2) + R$ *be a log-density of a normal law. Then for every* $k = 1, \ldots, K$

$$\widetilde{\theta}_k = \frac{1}{N_k} \sum_{X_i \in A_k} Y_i \,,$$

$$\theta_k^{*} = \frac{1}{N_k} \sum_{X_i \in A_k} I\!\!E Y_i \,,$$

where N_k *stands for the number of design points* X_i *within the piece* A_k *:*

$$N_k \overset{\text{def}}{=} \sum_{X_i \in A_k} 1 = \#\bigl\{i \colon X_i \in A_k\bigr\}.$$

Exercise 1.8.2. Check the statements of Theorem 1.8.1.

The properties of each estimator $\widetilde{\theta}_k$ repeats ones of the MLE for the sample retracted to A_k.

Theorem 1.8.2. *Let* $\widetilde{\boldsymbol{\theta}}$ *be defined by* (1.35) *for a normal density* $f(y)$. *Then with* $\boldsymbol{\theta}^{*} = (\theta_1^{*}, \ldots, \theta_K)^{\top} = \operatorname{argmax}_{\boldsymbol{\theta}} I\!\!E L(\boldsymbol{\theta})$, *it holds*

$$I\!\!E \widetilde{\theta}_k = \theta_k^{*}, \qquad \operatorname{Var}(\widetilde{\theta}_k) = \frac{1}{N_k^2} \sum_{X_i \in A_k} \operatorname{Var}(Y_i).$$

Moreover,

$$L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{*}) = \sum_{k=1}^{K} \frac{N_k}{2\sigma^2} (\widetilde{\theta}_k - \theta_k^{*})^2.$$

The statements follow by direct calculus on each interval separately.

If the errors $\varepsilon_i = Y_i - I\!\!E Y_i$ are normal and homogeneous, then the distribution of the maximum likelihood $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{*})$ is available.

Theorem 1.8.3. *Consider a Gaussian regression* $Y_i \sim \mathcal{N}(f(X_i), \sigma^2)$ *for* $i = 1, \ldots, n$. *Then* $\widetilde{\theta}_k \sim \mathcal{N}(\theta_k^*, \sigma^2 / N_k)$ *and*

$$L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \sum_{m=1}^{K} \frac{N_k}{2\sigma^2} \bigl(\widetilde{\theta}_k - \theta_k^* \bigr)^2 \sim \chi_K^2 ,$$

where χ_K^2 *stands for the chi-squared distribution with* K *degrees of freedom.*

This statement is a specification of the results from Section 1.6. It is worth mentioning once again that the regression function $f(\cdot)$ is not assumed to be piecewise constant, it can be whatever function. Each $\widetilde{\theta}_k$ estimates the mean θ_k^* of $f(\cdot)$ over the design points X_i within A_k.

The results on the behavior of the maximum likelihood $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ are often used for studying the properties of the chi-squared test.

A choice of the partition is an important issue in the piecewise constant approximation. The presented results indicate that the accuracy of estimation of θ_k^* by $\widetilde{\theta}_k$ is inversely proportional to the number of points N_k within each piece A_k. In the univariate case one usually applies the equidistant partition: the design interval is split into p equal intervals A_k leading to approximately equal values N_k. Sometimes, especially if the design is irregular, a non-uniform partition can be preferable. In general it can be recommended to split the whole design space into intervals with approximately the same number N_k of design points X_i.

A constant approximation is often not accurate enough to expand a regular regression function. One often uses a linear or polynomial approximation. The next sections explain this approach for the case of a univariate regression.

1.8.2 Piecewise linear univariate estimation

The piecewise constant approximation can be naturally extended to piecewise linear and piecewise polynomial construction. The starting point is again a non-overlapping partition of $\mathcal{X}$ into intervals A_k for $k = 1, \ldots, K$. First we explain the idea for the linear approximation of the function f on each interval A_k. Any linear function on A_k can be represented in the form $a_k + c_k x$ with some coefficients a_k, c_k. This yields in total $p = 2K$ coefficients: $\boldsymbol{\theta} = (a_1, c_1, \ldots, a_K, c_K)^\top$. The corresponding function $f(\cdot, \boldsymbol{\theta})$ can be represented as

$$f(x) \approx f(x, \boldsymbol{\theta}) = \sum_{k=1}^{K} (a_k + c_k x) \, \mathbb{I}(x \in A_k).$$

The non-overlapping structure of the sets A_k yields orthogonality of basis functions for different pieces. As a corollary, one can optimize the linear approximation on every interval A_k independently of the others.

Exercise 1.8.3. Show that $\widetilde{a}_k, \widetilde{c}_k$ can be obtained by the linear approximation of the data Y_i for $X_i \in A_k$:

$$(\widetilde{a}_k, \widetilde{c}_k) = \operatorname*{argmax}_{(a_k, c_k)} \sum L(Y_i - a_k - c_k X_i)\, \mathrm{I\!I}(X_i \in A_k), \qquad k = 1, \ldots, K.$$

On every piece A_k, the constant and the linear function x are not orthogonal except some very special situation. However, one can easily achieve orthogonality by a shift of the linear term.

Exercise 1.8.4. For each $k \le K$, there exists a point x_k such that

$$\sum (X_i - x_k)\, \mathrm{I\!I}(X_i \in A_k) = 0. \tag{1.37}$$

Introduce for each $k \le K$ two basis functions $\phi_{j-1}(x) = \mathrm{I\!I}(x \in A_k)$ and $\phi_j(x) = (x - x_k)\, \mathrm{I\!I}(x \in A_k)$ with $j = 2k$.

Exercise 1.8.5. Assume (1.37) for each $k \le K$. Check that any piecewise linear function can be uniquely represented in the form

$$f(x) = \sum_{j=1}^{p} \theta_j \phi_j(x)$$

with $p = 2K$ and the functions ϕ_j are orthogonal in the sense that for $j \ne j'$

$$\sum_{i=1}^{n} \phi_j(X_i)\phi_{j'}(X_i) = 0.$$

In addition, for each $k \le K$

$$\|\phi_j\|^2 \overset{\text{def}}{=} \sum_{i=1}^{n} \phi_j^2(X_i) = \begin{cases} N_k, & j = 2k-1, \\ V_k^2 & j = 2k. \end{cases}$$

$$N_k^2 \overset{\text{def}}{=} \sum_{X_i \in A_k} 1, \qquad V_k^2 \overset{\text{def}}{=} \sum_{X_i \in A_k} (X_i - x_k)^2.$$

In the case of Gaussian regression, orthogonality of the basis helps to gain a simple closed form for the estimators $\widetilde{\boldsymbol{\theta}} = (\widetilde{\theta}_j)$:

$$\widetilde{\theta}_j = \frac{1}{\|\psi_j\|^2} \sum_{i=1}^{n} Y_i \psi_j(X_i) = \begin{cases} \frac{1}{N_k} \sum_{X_i \in A_k} Y_i\,, & j = 2k-1, \\ \frac{1}{V_k^2} \sum_{X_i \in A_k} Y_i (X_i - x_k), & j = 2k. \end{cases}$$

see Section 1 in the next chapter for a comprehensive study.

1.8.3 Piecewise polynomial estimation

Local linear expansion of the function $f(x)$ on each piece A_k can be extended to a piecewise polynomial case. The basic idea is to apply a polynomial approximation of a certain degree q on each piece A_k independently. One can use for each piece a basis of the form $(x - x_k)^m \, \mathrm{I\!I}(x \in A_k)$ for $m = 0, 1, \ldots, q$ with x_k from (1.37) yielding the approximation

$$\begin{aligned} f(x) \, \mathrm{I\!I}(x \in A_k) &= f(x, \boldsymbol{a}_k) \, \mathrm{I\!I}(x \in A_k) \\ &= \sum_{k=1}^{K} \bigl\{ a_{0,k} + a_{1,k}(x - x_k) + \ldots + a_{q,k}(x - x_k)^q \bigr\} \, \mathrm{I\!I}(x \in A_k) \end{aligned}$$

for $\boldsymbol{a}_k = (a_{0,k}, a_{1,k}, \ldots, a_{q,k})^\top$. This involves $q + 1$ parameter for each piece and $p = K(q + 1)$ parameters in total. A nice feature of the piecewise approach is that the coefficients $\boldsymbol{a}_k$ of the piecewise polynomial approximation can be estimated independently for each piece. Namely,

$$\widetilde{\boldsymbol{a}}_k = \operatorname*{argmax}_{\boldsymbol{a}} \sum_{X_i \in A_k} \bigl\{ Y_i - f(X_i, \boldsymbol{a}) \bigr\}^2$$

1.8.4 Spline estimation

The main drawback of the piecewise polynomial approximation is that the resulting function f is discontinuous at the edge points between different pieces. A natural way of improving the boundary effect is to force some conditions on the boundary behavior. One important special case is given by the spline system. Let $\mathcal{X}$ be an interval on the real line, perhaps infinite. Let also $t_0 < t_1 < \ldots < t_K$ be some ordered points in $\mathcal{X}$ such that t_0 is the left edge and t_K the right edge of $\mathcal{X}$. Such points are called *knots*. We say that a function f is a *spline* of degree q at knots (t_k) if is polynomial on each *span* (t_{k-1}, t_k) for $k = 1, \ldots, K$ and satisfies the boundary conditions

$$f^{(m)}(t_k-) = f^{(m)}(t_k+), \qquad m = 0, \ldots, q - 1, \quad k = 1, \ldots, K - 1.$$

Here $f^{(m)}(t-)$ stands for the left derivative of f at t. In words, the function f and its first $q - 1$ derivatives are continuous on $\mathcal{X}$ and only the qth derivatives may have discontinuities at the knots t_k. It is obvious that the q derivative $f^{(q)}(t)$ of the spline of degree q is a piecewise constant functions on the spans $A_k = [t_{k-1}, t_k)$.

The spline is called *uniform* if the knots are equidistant, or, in other words, if all the spans A_k have equal length. Otherwise it is *non-uniform*.

Lemma 1.8.1. *The set of all splines of degree q at knots (t_k) is a linear space, that is, any linear combination of such splines is again a spline. Any function having a continuous m th derivative for $m < K$ and piecewise constant q th derivative is a q-spline.*

Splines of degree zero are just piecewise constant functions studied in Section 1.8.1. Linear splines are particularly transparent: this is the set of all piecewise linear continuous functions on $\mathcal{X}$. Each of them can be easily constructed from left to right or from right to left: start with a linear function $a_1 + c_1 x$ on the piece $A_1 = [t_0, t_1]$. Then $f(t_1) = a_1 + c_1 t_1$. On the piece A_2 the slope of f can be changed for c_2 leading to the function $f(x) = f(t_1) + c_2(x - t_1)$ for $x \in [t_1, t_2]$. Similarly, at t_2 the slop of $f s$ can change for c_3 yielding $f(x) = f(t_2) + c_3(x - t_2)$ on A_3, and so on. Splines of higher order can be constructed similarly step by step: one fixes the polynomial form on the very first piece A_1 and then continues the spline function to every next piece A_k using the boundary conditions and the value of the q th derivative of f on A_k. This construction explains the next result.

Lemma 1.8.2. *Each spline f of degree q and knots (t_k) is uniquely described by the vector of coefficients $\boldsymbol{a}_1$ on the first span and the values $f^{(q)}(x)$ for each span $A_1, \ldots, A_K$.*

This result explains that the parameter dimension of the linear spline space is $q + K$. One possible basis in this space is given by polynomials x^{m-1} of degree $m = 0, 1, \ldots, q$ and the functions $\phi_k(x) \overset{\text{def}}{=} (x - t_k)_+^q$ for $k = 1, \ldots, K - 1$.

Exercise 1.8.6. Check that $\phi_j(x)$ for $j = 1, \ldots, q + K$ form a basis in the linear spline space, and any q-spline f can be represented as

$$f(x) = \sum_{m=0}^{q} \alpha_m x^m + \sum_{k=1}^{K-1} \theta_k \phi_k(x). \tag{1.38}$$

Hint: check that the functions $\phi_j(x)$ are linearly independent and that each q th derivative $\phi_j^{(q)}(x)$ is piecewise constant.

B–splines

Unfortunately, the basis functions $\{\phi_k(x)\}$ with $\phi_k(x) = (x - t_k)_+^q$ are only useful for theoretical study. The main problem is that the functions $\phi_j(x)$ are strongly correlated, and the recovering the coefficients θ_j in the expansion (1.38) is a hard numerical task. by this reason, one often uses another basis called B–splines. The idea is to build splines of the given degree with the minimal support. Each B–spline basis function $b_{k,q}(x)$ is only non-zero on the q neighbor spans A_k, A_{k+1}, … A_{k+q-1} for $k = 1, \ldots, K - q$.

Exercise 1.8.7. Let $f(x)$ be a q-spline with the support on $q' < q$ neighbor spans A_k, A_{k+1}, … $A_{k+q'-1}$. Then $f(x) \equiv 0$.
Hint: consider any spline of the form $f(x) = \sum_{j=k}^{k+q'-1} c_j \phi_j(x)$. Show that the boundary conditions $f^{(m)}(t_{k+q'}) = 0$ for $m = 0, 1, \ldots, q$ yield $c_j \equiv 0$.

The basis B–spline functions can be constructed successfully. For $q = 0$, the B–splines $b_{k,0}(x)$ coincide with the functions $\phi_k(x) = \mathbb{1}(x \in A_k)$, $k = 1, \ldots, K$. Each linear B–spline $b_{k,1}(x)$ has a triangle shape on the two connected intervals A_k and A_{k+1}. It can be defined by the formula

$$b_{k,1}(x) \stackrel{\text{def}}{=} \frac{x - t_{k-1}}{t_k - t_{k-1}} b_{k,0}(x) + \frac{t_{k+1} - x}{t_{k+1} - t_k} b_{k+1,0}(x), \quad k = 1, \ldots, K-1.$$

One can continue this way leading to the *Cox–de Boor recursion formula*

$$b_{k,m}(x) \stackrel{\text{def}}{=} \frac{x - t_{k-1}}{t_{k+m-1} - t_{k-1}} b_{k,m-1}(x) + \frac{t_{k+m} - x}{t_{k+m} - t_k} b_{k+1,m-1}(x)$$

for $k = 1, \ldots, K - m$.

Exercise 1.8.8. Check by induction for each function $b_{k,m}(x)$ the following conditions:

1. $b_{k,m}(x)$ a polynomial of degree m on each span $A_k, \ldots, A_{k+m-1}$ and zero outside;
2. $b_{k,m}(x)$ can be uniquely represented as a sum $b_{k,m}(x) = \sum_{l=0}^{m-1} c_{l,k} \phi_{k+l}(x)$;
3. $b_{k,m}(x)$ is a m-spline.

The formulas simplify for the uniform splines with equal span length $\Delta = |A_k|$:

$$b_{k,m}(x) \stackrel{\text{def}}{=} \frac{x - t_{k-1}}{m\Delta} b_{k,m-1}(x) + \frac{t_{k+m} - x}{m\Delta} b_{k+1,m-1}(x)$$

for $k = 1, \ldots, K - m$.

Exercise 1.8.9. Check that

$$b_{k,m}(x) = \sum_{l=0}^{m} \omega_{l,m} \phi_{k+l}(x)$$

with

$$\omega_{l,m} \stackrel{\text{def}}{=} \frac{(-1)^l}{\Delta^m l!(m-l)!}$$

Smoothing splines

Such a spline system naturally arises as a solution of a penalized maximum likelihood problem. Suppose we are given the regression data (Y_i, X_i) with the univariate design $X_1 \le X_2 \le \ldots \le X_n$. Consider the mean regression model $Y_i = f(X_i) + \varepsilon_i$ with zero mean errors ε_i. The assumption of independent homogeneous Gaussian errors leads to the Gaussian log-likelihood

$$L(f) = -\sum_{i=1}^{n} \bigl| Y_i - f(X_i) \bigr|^2 / (2\sigma^2) \tag{1.39}$$

Maximization of this expression w.r.t. all possible functions f or, equivalently, all vectors $\bigl(f(X_1), \ldots, f(X_n)\bigr)^\top$ results in the trivial solution: $f(X_i) = Y_i$. This mean that the full dimensional maximum likelihood perfectly reproduces the original noisy data. Some additional assumptions are needed to force any desirable feature of the reconstructed function. One popular example is given by smoothness of the function f. Degree of smoothness (or, inversely, degree of roughness) can be measured by the value

$$\mathcal{R}_q(f) \stackrel{\text{def}}{=} \int_{\mathcal{X}} \bigl| f^{(q)}(x) \bigr|^2 dx. \tag{1.40}$$

One can try to optimize the fit (1.39) subject to the constraint on the amount of roughness from (1.40). Equivalently, one can optimize the penalized log-likelihood

$$L_\lambda(f) \stackrel{\text{def}}{=} L(f) - \lambda \mathcal{R}_q(f) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \bigl| Y_i - f(X_i) \bigr|^2 - \lambda \int_{\mathcal{X}} \bigl| f^{(q)}(x) \bigr|^2 dx,$$

where $\lambda > 0$ is a Lagrange multiplier. The corresponding maximizer is the penalized maximum likelihood estimator:

$$\widetilde{f}_\lambda = \operatorname*{argmax}_f L_\lambda(f), \tag{1.41}$$

where the maximum is taken over the class of all measurable functions. It is remarkable that the solution of this optimization problem is a spline of degree q with the knots $X_1, \ldots, X_n$.

Theorem 1.8.4. *For any $\lambda > 0$ and any integer q, the problem (1.41) has a unique solution which is a q-spline with knots at design points (X_i).*

For the proof we refer to Green and Silverman (1994). Due to this result, one can simplify the problem and look for a spline f which minimizes the objective $L_\lambda(f)$. A solution to (1.41) is called a *smoothing spline.* If f is a q-spline, the integral $\mathcal{R}_q(f)$ can be easily computed. Indeed, $f^{(q)}(x)$ is piecewise constant, that is, $f^{(q)}(x) = c_k$ for $x \in A_k$, and

$$\mathcal{R}_q(f) = \sum_{k=1}^{K} c_k^2 \, |t_k - t_{k-1}|.$$

For the uniform design, the formula simplifies even more, and by change of the multiplier λ, one can use $\mathcal{R}_q(f) = \sum_k c_k^2$. The use of any parametric representation of a spline function f allows to represent the optimization problem (1.41) as a penalized least squares problem. Estimation and inference in such problems are studied below in Section 4.

2

* Parametric modeling. Method of substitution and M-estimation

This chapter discusses a more general situation when the regression function $f(x)$ is not assumed to be linear in the parameter $\boldsymbol{\theta}$. The more general (quasi) parametric approach is based on a parametric assumption $f(\cdot) = f(\cdot, \boldsymbol{\theta}^{*}) \in \bigl(f(\cdot, \boldsymbol{\theta}),\, \boldsymbol{\theta} \in \Theta \subset I\!R^{p}\bigr)$.

Observe that the parametric regression equation can be rewritten as

$$\varepsilon_i = Y_i - f(X_i, \boldsymbol{\theta}^{*}).$$

If $\widetilde{\boldsymbol{\theta}}$ is an estimate of the parameter $\boldsymbol{\theta}^{*}$, then the *residuals* $\widetilde{\varepsilon}_i = Y_i - f(X_i, \widetilde{\boldsymbol{\theta}})$ are estimates of the individual errors ε_i. So, the idea of the method is to select the parameter estimate $\widetilde{\boldsymbol{\theta}}$ in a way that the empirical distribution P_n of the residuals $\widetilde{\varepsilon}_i$ mimics as well as possible certain prescribed features of the error distribution. We consider one approach called minimum contrast or M-estimation. Let $\psi(y)$ be an *influence* or *contrast* function. The main condition on the choice of this function is that

$$I\!E \psi(\varepsilon_i + z) \geq I\!E \psi(\varepsilon_i)$$

for all $i = 1, \ldots, n$ and all z. Then the true value $\boldsymbol{\theta}^{*}$ clearly minimizes the expectation of the sum $\sum_i \psi\bigl(Y_i - f(X_i, \boldsymbol{\theta})\bigr)$:

$$\boldsymbol{\theta}^{*} = \operatorname*{argmin}_{\boldsymbol{\theta}} I\!E \sum_i \psi\bigl(Y_i - f(X_i, \boldsymbol{\theta})\bigr).$$

This leads to the *M-estimate*

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \sum_i \psi\bigl(Y_i - f(X_i, \boldsymbol{\theta})\bigr).$$

This estimation method can be treated as replacing the true expectation of the errors by the empirical distribution of the residuals.

We specify this approach for regression estimation by the classical examples of least squares, least absolute deviation and maximum likelihood estimation corresponding to

$\psi(x) = x^2$, $\psi(x) = |x|$ and $\psi(x) = -\log p(x)$, where $p(x)$ is the error density. All these examples belong within framework of M-estimation and the quasi maximum likelihood approach.

2.1 Mean regression. Least squares estimate

The observations Y_i are assumed to follow the model

$$Y_i = f(X_i, \boldsymbol{\theta}^*) + \varepsilon_i\,, \qquad I\!\!E\varepsilon_i = 0 \tag{2.1}$$

with an unknown target $\boldsymbol{\theta}^*$. Suppose in addition that $\sigma_i^2 = I\!\!E\varepsilon_i^2 < \infty$. Then for every $\boldsymbol{\theta} \in \Theta$ and every $i \le n$ due to (2.1)

$$\begin{aligned} I\!\!E_{\boldsymbol{\theta}^*}\bigl\{Y_i - f(X_i, \boldsymbol{\theta})\bigr\}^2 &= I\!\!E_{\boldsymbol{\theta}^*}\bigl\{\varepsilon_i + f(X_i, \boldsymbol{\theta}^*) - f(X_i, \boldsymbol{\theta})\bigr\}^2 \\ &= \sigma_i^2 + \bigl|f(X_i, \boldsymbol{\theta}^*) - f(X_i, \boldsymbol{\theta})\bigr|^2. \end{aligned}$$

This yields for the whole sample

$$I\!\!E_{\boldsymbol{\theta}^*} \sum \bigl\{Y_i - f(X_i, \boldsymbol{\theta})\bigr\}^2 = \sum \bigl\{\sigma_i^2 + \bigl|f(X_i, \boldsymbol{\theta}^*) - f(X_i, \boldsymbol{\theta})\bigr|^2\bigr\}.$$

This expression is clearly minimized at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. This leads to the idea of estimating the parameter $\boldsymbol{\theta}^*$ by maximizing its empirical counterpart. The resulting estimate is called the (ordinary) *least squares estimate* (LSE):

$$\widetilde{\boldsymbol{\theta}}_{LSE} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \sum \bigl\{Y_i - f(X_i, \boldsymbol{\theta})\bigr\}^2.$$

This estimate is very natural and requires minimal information about the errors ε_i . Namely, one only needs $I\!\!E\varepsilon_i = 0$ and $I\!\!E\varepsilon_i^2 < \infty$.

2.2 Median regression. Least absolute deviation estimate

Consider the same regression model as in (2.1), but the errors ε_i are not zero-mean. Instead we assume that their median is zero:

$$Y_i = f(X_i, \boldsymbol{\theta}^*) + \varepsilon_i\,, \qquad \operatorname{med}(\varepsilon_i) = 0.$$

As previously, the target of estimation is the parameter $\boldsymbol{\theta}^*$. Observe that $\varepsilon_i = Y_i - f(X_i, \boldsymbol{\theta}^*)$ and hence, the latter r.v. has median zero. We now use the following simple fact: if $\operatorname{med}(\varepsilon) = 0$, then for any $z \ne 0$

$$I\!\!E|\varepsilon + z| \ge I\!\!E|\varepsilon|. \tag{2.2}$$

Exercise 2.2.1. Prove (2.2).

The property (2.2) implies for every $\boldsymbol{\theta}$

$$I\!\!E_{\boldsymbol{\theta}^*} \sum \bigl| Y_i - f(X_i, \boldsymbol{\theta}) \bigr| \ge I\!\!E_{\boldsymbol{\theta}^*} \sum \bigl| Y_i - f(X_i, \boldsymbol{\theta}^*) \bigr|,$$

that is, $\boldsymbol{\theta}^*$ minimizes over $\boldsymbol{\theta}$ the expectation under the true measure of the sum $\sum \bigl| Y_i - f(X_i, \boldsymbol{\theta}) \bigr|$. This leads to the empirical counterpart of $\boldsymbol{\theta}^*$ given by

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \sum \bigl| Y_i - f(X_i, \boldsymbol{\theta}) \bigr|.$$

This procedure is usually referred to as *least absolute deviations* regression estimate.

2.3 Maximum likelihood regression estimation

Let the density function $p(\cdot)$ of the errors ε_i be known. The regression equation (2.1) implies $\varepsilon_i = Y_i - f(X_i, \boldsymbol{\theta}^*)$. Therefore, every Y_i has the density $p(y - f(X_i, \boldsymbol{\theta}^*))$. Independence of the Y_i 's implies the product structure of the density of the joint distribution:

$$\prod p(y_i - f(X_i, \boldsymbol{\theta})),$$

yielding the log-likelihood

$$L(\boldsymbol{\theta}) = \sum \ell(Y_i - f(X_i, \boldsymbol{\theta}))$$

with $\ell(t) = \log p(t)$. The maximum likelihood estimate (MLE) is the point of maximum of $L(\boldsymbol{\theta})$:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \sum \ell(Y_i - f(X_i, \boldsymbol{\theta})).$$

A closed form solution for this equation exists only in some special cases like linear Gaussian regression. Otherwise this equation has to be solved numerically.

Consider an important special case corresponding to the i.i.d. Gaussian errors when $p(y)$ is the density of the normal law with mean zero and variance σ^2. Then

$$L(\boldsymbol{\theta}) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum \bigl| Y_i - f(X_i, \boldsymbol{\theta}) \bigr|^2.$$

The corresponding MLE maximizes $L(\boldsymbol{\theta})$ or, equivalently, minimizes the sum $\sum \bigl| Y_i - f(X_i, \boldsymbol{\theta}) \bigr|^2$:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \sum \bigl| Y_i - f(X_i, \boldsymbol{\theta}) \bigr|^2.$$

This estimate has already been introduced as the *ordinary least squares* estimate (oLSE).

An extension of the previous example is given by inhomogeneous Gaussian regression, when the errors ε_i are independent Gaussian zero-mean but the variances depend on i: $I\!\!E \varepsilon_i^2 = \sigma_i^2$. Then the log-likelihood $L(\boldsymbol{\theta})$ is given by the sum

$$L(\boldsymbol{\theta}) = \sum \Bigl\{ -\frac{\bigl|Y_i - f(X_i, \boldsymbol{\theta})\bigr|^2}{2\sigma_i^2} - \frac{1}{2} \log(2\pi\sigma_i^2) \Bigr\}.$$

Maximizing this expression w.r.t. $\boldsymbol{\theta}$ is equivalent to minimizing the weighted sum $\sum \sigma_i^{-2} \bigl|Y_i - f(X_i, \boldsymbol{\theta})\bigr|^2$:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \sum \sigma_i^{-2} \bigl|Y_i - f(X_i, \boldsymbol{\theta})\bigr|^2.$$

Such an estimate is also-called the *weighted least squares* (wLSE).

Another example corresponds to the case when the errors ε_i are i.i.d. double exponential, so that $I\!\!P(\pm\varepsilon_1 > t) = e^{-t/\sigma}$ for some given $\sigma > 0$. Then $p(y) = (2\sigma)^{-1} e^{-|y|/\sigma}$ and

$$L(\boldsymbol{\theta}) = -n \log(2\sigma) - \sigma^{-1} \sum \bigl|Y_i - f(X_i, \boldsymbol{\theta})\bigr|.$$

The MLE $\widetilde{\boldsymbol{\theta}}$ maximizes $L(\boldsymbol{\theta})$ or, equivalently, minimizes the sum $\sum \bigl|Y_i - f(X_i, \boldsymbol{\theta})\bigr|$:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \sum \bigl|Y_i - f(X_i, \boldsymbol{\theta})\bigr|.$$

So the maximum likelihood regression with Laplacian errors leads back to the *least absolute deviations* (LAD) estimate.

2.4 Quasi Maximum Likelihood approach

This section very briefly discusses an extension of the maximum likelihood approach. A more detailed discussion will be given in context of linear modeling in Chapter **??**. To be specific, consider a regression model

$$Y_i = f(X_i) + \varepsilon_i.$$

The maximum likelihood approach requires to specify the two main ingredients of this model: a parametric class $\{f(x, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ of regression functions and the distribution of the errors ε_i. Sometimes such information is lacking. One or even both modeling assumptions can be misspecified. In such situations one speaks of a *quasi maximum likelihood* approach, where the estimate $\widetilde{\boldsymbol{\theta}}$ is defined via maximizing over $\boldsymbol{\theta}$ the random function $L(\boldsymbol{\theta})$ even through it is not necessarily the real log-likelihood. Some examples of this approach have already been given.

Below we distinguish between misspecification of the first and second kind. The first kind corresponds to the parametric assumption about the regression function: assumed is the equality $f(X_i) = f(X_i, \boldsymbol{\theta}^*)$ for some $\boldsymbol{\theta}^* \in \Theta$. In reality one can only expect a reasonable quality of approximating $f(\cdot)$ by $f(\cdot, \boldsymbol{\theta}^*)$. A typical example is given by linear (polynomial) regression. The linear structure of the regression function is useful and tractable but it can only be a rough approximation of the real relation between Y and X. The quasi maximum likelihood approach suggests to ignore this misspecification and proceed as if the parametric assumption is fulfilled. This approach raises a number of questions: what is the target of estimation and what is really estimated by such quasi ML procedure? In Below we show in the context of linear modeling that the target of estimation can be naturally defined as the parameter $\boldsymbol{\theta}^*$ providing the best approximation of the true regression function $f(\cdot)$ by its parametric counterpart $f(\cdot, \boldsymbol{\theta})$.

The second kind of misspecification concerns the assumption about the errors ε_i. In the most of applications, the distribution of errors is unknown. Moreover, the errors can be dependent or non-identically distributed. Assumption of a specific i.i.d. structure leads to a model misspecification and thus, to the quasi maximum likelihood approach. We illustrate this situation by few examples.

Consider the regression model $Y_i = f(X_i, \boldsymbol{\theta}^*) + \varepsilon_i$ and suppose for a moment that the errors ε_i are i.i.d. normal. Then the principal term of the corresponding log-likelihood is given by the negative sum of the squared residuals: $\sum \bigl| Y_i - f(X_i, \boldsymbol{\theta}) \bigr|^2$, and its maximization leads to the least squares method. So, one can say that the LSE method is the quasi MLE when the errors are assumed to be i.i.d. normal. That is, the LSE can be obtained as the MLE for the imaginary Gaussian regression model when the errors ε_i are not necessarily i.i.d. Gaussian.

If the data are contaminated or the errors have heavy tails, it could be unwise to apply the LSE method. The LAD method is known to be more robust against outliers and data contamination. At the same time, it has already been shown in Section 2.3 that the LAD estimates is the MLE when the errors are Laplacian (double exponential). In other words, LAD is the quasi MLE for the model with Laplacian errors.

2.5 Generalized regression

Let the response Y_i be observed at the design point $X_i \in I\!\!R^d$, $i = 1, \ldots, n$. A (mean) regression model assumes that the observed values Y_i are independent and can be decomposed into the systematic component $f(X_i)$ and the individual centered stochastic error ε_i. In some cases such a decomposition is questionable. This especially concerns the case when the data Y_i are categorical, e.g. binary or discrete. Another striking example is given by nonnegative observations Y_i. In such cases one usually assumes that the distribution of Y_i belongs to some given parametric family $(P_\upsilon, \upsilon \in \mathcal{U})$ and only the parameter of this distribution depends on the design point X_i. We denote this parameter value as $f(X_i) \in \mathcal{U}$ and write the model in the form

$$Y_i \sim P_{f(X_i)} .$$

As previously, $f(\cdot)$ is called a *regression function* and its values at the design points X_i completely specify the joint data distribution:

$$\boldsymbol{Y} \sim \prod_i P_{f(X_i)}.$$

Below we assume that (P_υ) is a univariate exponential family with the log-density $\ell(y, \upsilon)$.

The parametric modeling approach assumes that the regression function f can be specified by a finite-dimensional parameter $\boldsymbol{\theta} \in \Theta \subset I\!\!R^p$: $f(x) = f(x, \boldsymbol{\theta})$. As usual, by $\boldsymbol{\theta}^*$ we denote the true parameter value. The log-likelihood function for this model reads

$$L(\boldsymbol{\theta}) = \sum_i \ell\bigl(Y_i, f(X_i, \boldsymbol{\theta})\bigr).$$

The corresponding MLE $\widetilde{\boldsymbol{\theta}}$ maximizes $L(\boldsymbol{\theta})$:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_i \ell\bigl(Y_i, f(X_i, \boldsymbol{\theta})\bigr).$$

The estimating equation $\nabla L(\boldsymbol{\theta}) = 0$ reads as

$$\sum_i \ell'\bigl(Y_i, f(X_i, \boldsymbol{\theta})\bigr) \nabla f(X_i, \boldsymbol{\theta}) = 0$$

where $\ell'(y, \upsilon) \overset{\text{def}}{=} \partial \ell(y, \upsilon) / \partial \upsilon$.

The approach essentially depends on the parametrization of the considered EF. Usually one applies either the natural or canonical parametrization. In the case of the natural parametrization, $\ell(y, \upsilon) = C(\upsilon) y - B(\upsilon)$, where the functions $C(\cdot), B(\cdot)$ satisfy $B'(\upsilon) = \upsilon C'(\upsilon)$. This implies $\ell'(y, \upsilon) = y C'(\upsilon) - B'(\upsilon) = (y - \upsilon) C'(\upsilon)$ and the estimating equation reads as

$$\sum_i \bigl(Y_i - f(X_i, \boldsymbol{\theta})\bigr) C'\bigl(f(X_i, \boldsymbol{\theta})\bigr) \nabla f(X_i, \boldsymbol{\theta}) = 0$$

Unfortunately, a closed form solution for this equation exists only in very special cases. Even the questions of existence and uniqueness of the solution cannot be studied in whole generality. Some numerical algorithms are usually applied to solve the estimating equation.

Exercise 2.5.1. Specify the estimating equation for generalized EFn regression and find the solution for the case of the constant regression function $f(X_i, \theta) \equiv \theta$.
Hint: If $f(X_i, \theta) \equiv \theta$, then the Y_i are i.i.d. from P_θ.

The equation can be slightly simplified by using the canonical parametrization. If (P_υ) is an EFc with the log-density $\ell(y, \upsilon) = y\upsilon - d(\upsilon)$, then the log-likelihood $L(\boldsymbol{\theta})$ can be represented in the form

$$L(\boldsymbol{\theta}) = \sum_i \bigl\{ Y_i f(X_i, \boldsymbol{\theta}) - d\bigl(f(X_i, \boldsymbol{\theta})\bigr) \bigr\}.$$

The corresponding estimating equation is

$$\sum_i \bigl\{ Y_i - d'\bigl(f(X_i, \boldsymbol{\theta})\bigr) \bigr\} \nabla f(X_i, \boldsymbol{\theta}) = 0.$$

Exercise 2.5.2. Specify the estimating equation for generalized EFc regression and find the solution for the case of constant regression with $f(X_i, \upsilon) \equiv \upsilon$. Relate the natural and canonical representation.

A generalized regression with a canonical link is often applied in combination with linear modeling of the regression function considered in the next section.

2.5.1 Generalized linear models

Consider the generalized regression model

$$Y_i \sim P_{f(X_i)} \in \mathcal{P}.$$

In addition we assume a linear (in parameters) structure of the regression function $f(X)$. Such modeling is particularly useful to combine with the canonical parametrization of the considered EF with the log-density $\ell(y, \upsilon) = y\upsilon - d(\upsilon)$; see Chapter A. The reason is that the stochastic part in the log-likelihood of an EFc linearly depends on the parameter. So, below we assume that $\mathcal{P} = (P_\upsilon, \upsilon \in \mathcal{U})$ is an EFc.

Linear regression $f(X_i) = \Psi_i^\top \boldsymbol{\theta}$ with given feature vectors $\Psi_i \in I\!\!R^p$ leads to the model with the log-likelihood

$$L(\boldsymbol{\theta}) = \sum_i \bigl\{ Y_i \Psi_i^\top \boldsymbol{\theta} - d\bigl(\Psi_i^\top \boldsymbol{\theta}\bigr) \bigr\}.$$

Such a setup is called *generalized linear model* (GLM). Note that the log-likelihood can be represented as

$$L(\boldsymbol{\theta}) = S^{\top}\boldsymbol{\theta} - A(\boldsymbol{\theta}),$$

where

$$S = \sum_i Y_i \Psi_i, \qquad A(\boldsymbol{\theta}) = \sum_i d\bigl(\Psi_i^{\top}\boldsymbol{\theta}\bigr).$$

The corresponding MLE $\widetilde{\boldsymbol{\theta}}$ maximizes $L(\boldsymbol{\theta})$. Again, a closed form solution only exists in special cases. However, an important advantage of the GLM approach is that the solution always exists and is unique. The reason is that the log-likelihood function $L(\boldsymbol{\theta})$ is concave in $\boldsymbol{\theta}$.

Lemma 2.5.1. *The MLE $\widetilde{\boldsymbol{\theta}}$ solves the following estimating equation:*

$$\nabla L(\boldsymbol{\theta}) = S - \nabla A(\boldsymbol{\theta}) = \sum_i \Bigl(Y_i - d'(\Psi_i^{\top}\boldsymbol{\theta})\Bigr)\Psi_i = 0. \tag{2.3}$$

The solution exists and is unique.

Proof. Define the matrix

$$B(\boldsymbol{\theta}) = \sum_i d''(\Psi_i^{\top}\boldsymbol{\theta})\Psi_i\Psi_i^{\top}. \tag{2.4}$$

Since $d''(v)$ is strictly positive for all u, the matrix $B(\boldsymbol{\theta})$ is positively defined as well. It holds

$$\nabla^2 L(\boldsymbol{\theta}) = -\nabla^2 A(\boldsymbol{\theta}) = -\sum_i d''(\Psi_i^{\top}\boldsymbol{\theta})\Psi_i\Psi_i^{\top} = -B(\boldsymbol{\theta}).$$

Thus, the function $L(\boldsymbol{\theta})$ is strictly concave w.r.t. $\boldsymbol{\theta}$ and the estimating equation $\nabla L(\boldsymbol{\theta}) = S - \nabla A(\boldsymbol{\theta}) = 0$ has the unique solution $\widetilde{\boldsymbol{\theta}}$.

The solution of (2.3) can be easily obtained numerically by the Newton-Raphson algorithm: select the initial estimate $\boldsymbol{\theta}^{(0)}$. Then for every $k \ge 1$ apply

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \bigl\{B(\boldsymbol{\theta}^{(k)})\bigr\}^{-1}\bigl\{S - \nabla A(\boldsymbol{\theta}^{(k)})\bigr\} \tag{2.5}$$

until convergence.

Below we consider two special cases of GLMs for binary and Poissonian data.

2.5.2 Logit regression for binary data

Suppose that the observed data Y_i are independent and binary, that is, each Y_i is either zero or one, $i = 1, \ldots, n$. Such models are often used in e.g. sociological and medical study, two-class classification, binary imaging, among many other fields. We treat each Y_i as a Bernoulli r.v. with the corresponding parameter $f_i = f(X_i)$. This is a special case of generalized regression also called *binary response models*. The parametric modeling assumption means that the regression function $f(\cdot)$ can be represented in the form $f(X_i) = f(X_i, \boldsymbol{\theta})$ for a given class of functions $\{f(\cdot, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \in I\!\!R^p\}$. Then the log-likelihood $L(\boldsymbol{\theta})$ reads as

$$L(\boldsymbol{\theta}) = \sum_i \ell(Y_i, f(X_i, \boldsymbol{\theta})), \tag{2.6}$$

where $\ell(y, \upsilon)$ is the log-density of the Bernoulli law. For linear modeling, it is more useful to work with the canonical parametrization. Then $\ell(y, \upsilon) = y\upsilon - \log(1 + e^{\upsilon})$, and the log-likelihood reads

$$L(\boldsymbol{\theta}) = \sum_i \bigl[Y_i f(X_i, \boldsymbol{\theta}) - \log\bigl(1 + \mathrm{e}^{f(X_i, \boldsymbol{\theta})}\bigr)\bigr].$$

In particular, if the regression function $f(\cdot, \boldsymbol{\theta})$ is linear, that is, $f(X_i, \boldsymbol{\theta}) = \Psi_i^\top \boldsymbol{\theta}$, then

$$L(\boldsymbol{\theta}) = \sum_i \bigl[Y_i \Psi_i^\top \boldsymbol{\theta} - \log(1 + \mathrm{e}^{\Psi_i^\top \boldsymbol{\theta}})\bigr]. \tag{2.7}$$

The corresponding estimate reads as

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_i \bigl[Y_i \Psi_i^\top \boldsymbol{\theta} - \log(1 + \mathrm{e}^{\Psi_i^\top \boldsymbol{\theta}})\bigr]$$

This modeling is usually referred to as *logit regression*.

Exercise 2.5.3. Specify the estimating equation for the case of logit regression.

Exercise 2.5.4. Specify the step of the Newton-Raphson procedure for the case of logit regression.

2.5.3 Parametric Poisson regression

Suppose that the observations Y_i are nonnegative integer numbers. The Poisson distribution is a natural candidate for modeling such data. It is supposed that the underlying Poisson parameter depends on the regressor X_i. Typical examples arise in different types of imaging including medical positron emission and magnet resonance tomography, satellite and low-luminosity imaging, queueing theory, high frequency trading, etc. The regression equation reads

$$Y_i \sim \text{Poisson}(f(X_i)).$$

The *Poisson regression* function $f(X_i)$ is usually the target of estimation. The parametric specification $f(\cdot) \in \bigl\{f(\cdot, \boldsymbol{\theta}), \, \boldsymbol{\theta} \in \Theta\bigr\}$ reduces this problem to estimating the parameter $\boldsymbol{\theta}$. Under the assumption of independent observations Y_i, the corresponding maximum likelihood $L(\boldsymbol{\theta})$ is given by

$$L(\boldsymbol{\theta}) = \sum \bigl[Y_i \log\bigl\{f(X_i, \boldsymbol{\theta})\bigr\} - f(X_i, \boldsymbol{\theta})\bigr] + R,$$

where the remainder R does not depend on $\boldsymbol{\theta}$ and can be omitted. Obviously, the constant function family $f(\cdot, \theta) \equiv \theta$ leads back to the case of i.i.d. modeling studied in Section A. A further extension is given by linear Poisson regression: $f(X_i, \boldsymbol{\theta}) = \Psi_i^\top \boldsymbol{\theta}$ for some given factors Ψ_i. The regression equation reads

$$L(\boldsymbol{\theta}) = \sum_i \bigl[Y_i \log(\Psi_i^\top \boldsymbol{\theta}) - \Psi_i^\top \boldsymbol{\theta}\bigr]. \tag{2.8}$$

Exercise 2.5.5. Specify the estimating equation and the Newton-Raphson procedure for the linear Poisson regression (2.8).

An obvious problem of linear Poisson modeling is that it requires all the values $\Psi_i^\top \boldsymbol{\theta}$ to be positive. The use of canonical parametrization helps to avoid this problem. The linear structure is assumed for the canonical parameter leading to the representation $f(X_i) = \exp(\Psi_i^\top \boldsymbol{\theta})$. Then the general log-likelihood process $L(\boldsymbol{\theta})$ from (2.6) translates into

$$L(\boldsymbol{\theta}) = \sum_i \bigl[Y_i \Psi_i^\top \boldsymbol{\theta} - \exp(\Psi_i^\top \boldsymbol{\theta})\bigr]; \tag{2.9}$$

cf. with (2.7).

Exercise 2.5.6. Specify the estimating equation and the Newton-Raphson procedure for the canonical link linear Poisson regression (2.9).

If the factors Ψ_i are properly scaled then the scalar products $\Psi_i^\top \boldsymbol{\theta}$ for all i and all $\boldsymbol{\theta} \in \Theta_0$ belong to some bounded interval. For the matrix $B(\boldsymbol{\theta})$ from (2.4), it holds

$$B(\boldsymbol{\theta}) = \sum_i \exp(\Psi_i^\top \boldsymbol{\theta}) \Psi_i \Psi_i^\top.$$

Initializing the ML optimization problem with $\boldsymbol{\theta} = 0$ leads to the oLSE

$$\widetilde{\boldsymbol{\theta}}^{(0)} = \Bigl(\sum_i \Psi_i \Psi_i^\top \Bigr)^{-1} \sum_i \Psi_i Y_i \, .$$

The further steps of the algorithm (2.5) can be done as weighted LSE with the weights $\exp\bigl(\Psi_i^\top \widetilde{\boldsymbol{\theta}}^{(k)}\bigr)$ for the estimate $\widetilde{\boldsymbol{\theta}}^{(k)}$ obtained at the previous step.

2.5.4 Piecewise constant methods in generalized regression

Consider a generalized regression model

$$Y_i \sim P_{f(X_i)} \in (P_\upsilon)$$

for a given exponential family (P_υ). Further, let $A_1, \ldots, A_K$ be a *non-overlapping partition* of the design space $\mathcal{X}$; see (1.33). A piecewise constant approximation (1.34) of the regression function $f(\cdot)$ leads to the additive log-likelihood structure: for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\top$

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\theta_1,\ldots,\theta_K} \sum_{k=1}^{K} \sum_{X_i \in A_k} \ell(Y_i, \theta_k);$$

cf. (1.35). Similarly to the mean regression case, the global optimization w.r.t. the vector $\boldsymbol{\theta}$ can be decomposed into K separated simple optimization problems:

$$\widetilde{\theta}_k = \operatorname*{argmax}_{\theta} \sum_{X_i \in A_k} \ell(Y_i, \theta);$$

cf. (1.36). The same decomposition can be obtained for the target $\boldsymbol{\theta}^* = (\theta_1, \ldots, \theta_K)^\top$:

$$\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta}} I\!\!E L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\theta_1,\ldots,\theta_K} \sum_{k=1}^{K} \sum_{X_i \in A_k} I\!\!E \ell(Y_i, \theta_k).$$

The properties of each estimator $\widetilde{\theta}_k$ repeats ones of the qMLE for a univariate EFn; see Section A.

Theorem 2.5.1. *Let $L(y, \theta) = C(\theta) y - B(\theta)$ be a density of an EFn, so that the functions $B(\theta)$ and $C(\theta)$ satisfy $B'(\theta) = \theta C'(\theta)$. Then for every $k = 1, \ldots, K$*

$$\widetilde{\theta}_k = \frac{1}{N_k} \sum_{X_i \in A_k} Y_i \, ,$$

$$\theta_k^* = \frac{1}{N_k} \sum_{X_i \in A_k} I\!\!E Y_i \, ,$$

where N_k *stands for the number of design points* X_i *within the piece* A_k *:*

$$N_k \stackrel{\mathrm{def}}{=} \sum_{X_i \in A_k} 1 = \#\bigl\{i \colon X_i \in A_k\bigr\}.$$

Moreover, it holds

$$I\!\!E \widetilde{\theta}_k = \theta_k^*$$

and

$$L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \sum_{k=1}^{K} N_k \mathcal{K}(\widetilde{\theta}_k, \theta_k^*) \tag{2.10}$$

where $\mathcal{K}(\theta, \theta') \stackrel{\mathrm{def}}{=} E_\theta \bigl\{\ell(Y_i, \theta) - \ell(Y_i, \theta')\bigr\}$.

These statements follow from Theorem 2.5.1 and Theorem A.1.1 of Section A. For the presented results, the true regression function $f(\cdot)$ can be of arbitrary structure, the true distribution of each Y_i can differ from $P_{f(X_i)}$.

Exercise 2.5.7. Check the statements of Theorem 2.5.1.

If PA is correct, that is, if f is indeed piecewise constant and the distribution of Y_i is indeed $P_{f(X_i)}$, the deviation bound for the excess $L(\widetilde{\theta}_k, \theta_k^*)$ from Theorem A.3.1 can be applied to each piece A_k yielding the following result.

Theorem 2.5.2. *Let* (P_θ) *be a EFn and let* $Y_i \sim P_{\theta_k}$ *for* $X_i \in A_k$ *and* $k = 1, \ldots, K$ *. Then for any* $\mathfrak{z} > 0$

$$I\!\!P\bigl(L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) > K\mathfrak{z}\bigr) \le 2K \mathrm{e}^{-\mathfrak{z}}.$$

Proof. By (2.10) and Theorem A.3.1

$$\begin{aligned} I\!\!P\bigl(L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) > K\mathfrak{z}\bigr) &= I\!\!P\Bigl(\sum_{k=1}^{K} N_k \mathcal{K}(\widetilde{\theta}_k, \theta_k^*) > K\mathfrak{z}\Bigr) \\ &\le \sum_{k=1}^{K} I\!\!P\Bigl(N_k \mathcal{K}(\widetilde{\theta}_k, \theta_k^*) > \mathfrak{z}\Bigr) \le 2K \mathrm{e}^{-\mathfrak{z}} \end{aligned}$$

and the result follows.

A piecewise linear generalized regression can be treated in a similar way. The main benefit of piecewise modeling remains preserved: a global optimization over the vector $\boldsymbol{\theta}$ can be decomposed into a set of small optimization problems for each piece A_k . However, a closed form solution is available only in some special cases like Gaussian regression.

2.5.5 Smoothing splines for generalized regression

Consider again the generalized regression model

$$Y_i \sim P_{f(X_i)} \in \mathcal{P}$$

for an exponential family $\mathcal{P}$ with canonical parametrization. Now we do not assume any specific parametric structure for the function f. Instead, the function f is supposed to be smooth and its smoothness is measured by the roughness $\mathcal{R}_q(f)$ from (1.40). Similarly to the regression case of Section 1.8.4, the function f can be estimated directly by optimizing the penalized log-likelihood $L_\lambda(f)$:

$$\begin{aligned} \widetilde{f}_\lambda &= \operatorname*{argmax}_f L_\lambda(f) = \operatorname*{argmax}_f \bigl\{ L(f) - \mathcal{R}_q(f) \bigr\} \\ &= \operatorname*{argmax}_f \sum_i \bigl\{ Y_i f(X_i) - d\bigl(f(X_i)\bigr) \bigr\} - \int_{\mathcal{X}} \bigl| f^{(q)}(x) \bigr|^2 dx. \end{aligned} \tag{2.11}$$

The maximum is taken over the class of all regular q-times differentiable functions. In the regression case, the function $d(\cdot)$ is quadratic and the solution is a spline functions with knots X_i. This conclusion can be extended to the case of any convex function $d(\cdot)$, thus, the problem (2.11) yields a *smoothing spline* solution. Numerically this problem is usually solved by iterations. One starts with a quadratic function $d(\upsilon) = \upsilon^2/2$ to obtain an initial approximation $\widetilde{f}^{(0)}(\cdot)$ of $f(\cdot)$ by a standard smoothing spline regression. Further, at each new step $k+1$, the use of the estimate $\widetilde{f}^{(k)}(\cdot)$ from the previous step k for $k \geq 0$ helps to approximate the problem (2.11) by a weighted regression. The corresponding iterations can be written in the form (2.5).

3

* Linear regression with random design

3.1 Random design linear regression

Consider the linear regression equation

$$\boldsymbol{Y} = \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\boldsymbol{\theta}^{*} \in I\!R^{p}$ is the target parameter, $\boldsymbol{Y}$ is the n-vector of responses, $\boldsymbol{\varepsilon}$ is the n-vector of errors, and $\boldsymbol{\Psi} = (\Psi_1, \ldots, \Psi_n)$ is a $p \times n$ design matrix with columns $\Psi_i \in I\!R^{p}$. The assumption of homogeneous Gaussian errors $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_p)$ yields the corresponding Gaussian log-likelihood

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

and the MLE

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

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

$$I\!E\boldsymbol{\varepsilon} = 0, \qquad \operatorname{Var}(\boldsymbol{\varepsilon}) = \Sigma = \operatorname{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}. \tag{3.2}$$

For the case of a random design $\boldsymbol{\Psi}$, analysis of the MLE $\widetilde{\boldsymbol{\theta}}$ 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.

3.2 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.

3.2.1 Design with independent measurements

Let the feature vectors $\Psi_1, \ldots, \Psi_n$ in (3.1) be independent. We assume that the design is non degenerate, so that the matrix $\boldsymbol{M}^2 \overset{\text{def}}{=} \mathbb{E}\bigl(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\bigr)$ is positive. Consider

$$\boldsymbol{M}^{-1}\bigl\{\boldsymbol{\Psi}\boldsymbol{\Psi}^\top - \mathbb{E}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)\bigr\}\boldsymbol{M}^{-1} = A_1 + \ldots + A_n \,,$$

where

$$A_i = \boldsymbol{M}^{-1}\bigl\{\Psi_i\Psi_i^\top - \mathbb{E}(\Psi_i\Psi_i^\top)\bigr\}\boldsymbol{M}^{-1} \tag{3.3}$$

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

$$S_n^2 \overset{\text{def}}{=} \bigl\|\mathbb{E}(A_1^2 + \ldots + A_n^2)\bigr\|_{\text{op}} . \tag{3.4}$$

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

$$\|A_i\|_{\text{op}} \le u_n \qquad a.s. \tag{3.5}$$

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

$$\begin{aligned} M_1^2 &\overset{\text{def}}{=} \mathbb{E}(\Psi_1\Psi_1^\top), \\ \sigma_1^2 &\overset{\text{def}}{=} \mathbb{E}\bigl(M_1^{-1}\Psi_1\Psi_1^\top M_1^{-1} - \boldsymbol{I}_p\bigr)^2 . \end{aligned}$$

Also suppose that with probability one

$$\bigl\|M_1^{-1}\Psi_1\Psi_1^\top M_1^{-1} - \boldsymbol{I}_p\bigr\|_{\text{op}} \le u^*.$$

Then it holds

$$\begin{aligned} \boldsymbol{M}^2 &= n\, M_1^2, \\ S_n^2 &= n^{-1}\sigma_1^2, \\ u_n &\le n^{-1}u^* \end{aligned} \tag{3.6}$$

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

Theorem 3.2.1. *Suppose that Ψ_i are independent and A_i from (3.3) fulfill (3.5). Then with $\boldsymbol{M}^2 = \mathbb{E}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)$ and S_n^2 defined by (3.4), it holds for all $z > 0$*

$$\mathbb{P}\Bigl(\bigl\|\boldsymbol{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\boldsymbol{M}^{-1} - I_p\bigr\|_{\text{op}} > z\Bigr) \le 2p\exp\Bigl\{-\frac{z^2}{2S_n^2 + 2u_n z/3}\Bigr\}.$$

If Ψ_i are i.i.d. and (3.6) holds then

$$I\!\!P\Bigl(n^{1/2} \bigl\| \boldsymbol{M}^{-1} \boldsymbol{\Psi} \boldsymbol{\Psi}^{\top} \boldsymbol{M}^{-1} - \boldsymbol{I}_p \bigr\|_{\mathrm{op}} > z \Bigr) \le 2p \exp\Bigl\{ - \frac{z^2}{2\sigma_1^2 + 2n^{-1/2} u^* z/3} \Bigr\}.$$

Proof. (please check)

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

$$n \ge \bigl(2\sigma_1^2 \delta^{-2} + 2u^* \delta^{-1}/3 \bigr) \bigl\{ \mathtt{x} + \log(2p) \bigr\} \tag{3.7}$$

to ensure

$$I\!\!P\bigl(\| \boldsymbol{M}^{-1} \boldsymbol{\Psi} \boldsymbol{\Psi}^{\top} \boldsymbol{M}^{-1} - \boldsymbol{I}_p \|_{\mathrm{op}} > \delta \bigr) \le \mathrm{e}^{-\mathtt{x}}. \tag{3.8}$$

If n and $\mathtt{x}$ are fixed, then one can check (3.8) for

$$\delta = \delta(p, \mathtt{x}) = \sqrt{ \frac{2\sigma_1^2}{n} \bigl(\mathtt{x} + \log(2p) \bigr) } + \frac{2u^*}{3n} \bigl(\mathtt{x} + \log(2p) \bigr). \tag{3.9}$$

Corollary 3.2.1. *Suppose Ψ_i are i.i.d. and (3.6) holds. If n fulfills (3.7) for some fixed δ and $\mathtt{x}$, then (3.8) holds true. Similarly, if n and $\mathtt{x}$ are fixed, then δ from (3.9) ensures (3.8).*

Proof. (please check).

The result (3.8) guarantees that on a set $\Omega(\mathtt{x})$ with $I\!\!P\bigl(\Omega(\mathtt{x})\bigr) \ge 1 - \mathrm{e}^{-\mathtt{x}}$

$$\| \boldsymbol{M}^{-1} \boldsymbol{\Psi} \boldsymbol{\Psi}^{\top} \boldsymbol{M}^{-1} - \boldsymbol{I}_p \|_{\mathrm{op}} \le \delta. \tag{3.10}$$

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

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

(please check).

3.2.2 Aggregated random design

Here we discuss another random design setup for the regression model (3.1). Namely, assume that the design matrix $\boldsymbol{\Psi}$ can be represented as a sum of independent $p \times q$ matrices $\Psi_1, \ldots, \Psi_n$:

$$\boldsymbol{\Psi} = \Psi_1 + \ldots + \Psi_n. \tag{3.11}$$

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

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

and

$$S_n^2 \overset{\text{def}}{=} \bigl\| \boldsymbol{M}^{-1}\bigl(V_1^2 + \ldots + V_n^2\bigr)\boldsymbol{M}^{-1} \bigr\|_{\text{op}}.$$

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

$$\bigl\| \boldsymbol{M}^{-1}(\Psi_i - \mathbb{E}\Psi_i) \bigr\|_{\text{op}} \le u_n \qquad a.s.$$

for a small constant u_n. Now consider

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

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

$$\mathbb{P}\bigl(\|\boldsymbol{M}^{-1}(\boldsymbol{\Psi} - \mathbb{E}\boldsymbol{\Psi})\|_{\text{op}} \ge z\bigr) \le (p+q)\exp\Bigl\{-\frac{z^2}{2S_n^2 + 2u_n z/3}\Bigr\} \tag{3.12}$$

In the case with i.i.d. Ψ_i, define

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

and suppose that

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

Then it holds

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

and

$$\mathbb{P}\Bigl(n^{1/2}\bigl\|\boldsymbol{M}^{-1}\bigl(\boldsymbol{\Psi} - \mathbb{E}\boldsymbol{\Psi}\bigr)\bigr\|_{\text{op}} \ge z\Bigr) \le (p+q)\exp\Bigl\{-\frac{z^2}{2\sigma_1^2 + 2u_1 z/(3n^{1/2})}\Bigr\}.$$

The result (3.12) implies that $\boldsymbol{\Psi}$ is close to $I\!\!E\boldsymbol{\Psi}$. The MLE $\widetilde{\boldsymbol{\theta}}$ also involves the product $\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}$ which has to be close to the corresponding product $\boldsymbol{M}^2 = I\!\!E\boldsymbol{\Psi}\, I\!\!E(\boldsymbol{\Psi}^{\top})$. Below we assume that $\boldsymbol{M}$ is sufficiently large and bound the difference $\boldsymbol{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}\boldsymbol{M}^{-1} - I_p$.

Theorem 3.2.2. *Let* $\|\boldsymbol{M}^{-1}\bigl(\boldsymbol{\Psi} - I\!\!E\boldsymbol{\Psi}\bigr)\|_{\mathrm{op}} \le \delta$ *for some* $\delta > 0$. *Then*

$$\bigl\|\boldsymbol{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}\boldsymbol{M}^{-1} - I_p\bigr\|_{\mathrm{op}} \le \delta^2 + 2\delta.$$

Proof. One can bound

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

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

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

Therefore,

$$\|\boldsymbol{M}^{-1}(\boldsymbol{\Psi} - I\!\!E\boldsymbol{\Psi})\, I\!\!E(\boldsymbol{\Psi}^{\top})\boldsymbol{M}^{-1}\boldsymbol{\gamma}\| \le \|\boldsymbol{M}^{-1}(\boldsymbol{\Psi} - I\!\!E\boldsymbol{\Psi})\|_{\mathrm{op}}$$

thus

$$\|\boldsymbol{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}\boldsymbol{M}^{-1} - I_p\|_{\mathrm{op}} \le \|\boldsymbol{M}^{-1}(\boldsymbol{\Psi} - I\!\!E\boldsymbol{\Psi})\|_{\mathrm{op}}^2 + 2\|\boldsymbol{M}^{-1}(\boldsymbol{\Psi} - I\!\!E\boldsymbol{\Psi})\|_{\mathrm{op}},$$

and the result follows.

Theorem 3.3.2 applies in this situation without any change.

3.3 Fisher and Wilks expansions for the MLE under random design

Now we apply this result to the MLE in the regression model (3.1) under possible misspecification of the error variance. The corresponding log-likelihood ratio can be written in the form

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

Introduce also an approximating quadratic log-likelihood defined by

$$\mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \frac{1}{\sigma^2}(\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}^*)^{\top}\boldsymbol{\Psi}^{\top}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \frac{1}{2\sigma^2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\top}\boldsymbol{M}^2(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \qquad (3.13)$$

with $\boldsymbol{M}^2 = I\!\!E\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}$. Note that two expressions $L(\boldsymbol{\theta},\boldsymbol{\theta}^*)$, $\mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ differ only in the quadratic term. Moreover, we already know that $\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}$ is close to it expectation $\boldsymbol{M}^2 = I\!\!E\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}$; see (3.10).

Theorem 3.3.1. *Let* (3.10) *hold on a set* $\Omega(\mathtt{x})$ *with* $I\!P\bigl(\Omega(\mathtt{x})\bigr) \ge 1 - \mathrm{e}^{-\mathtt{x}}$. *Then with*

$$D^2 = \sigma^{-2} \boldsymbol{M}^2 = \sigma^{-2} I\!E(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top), \tag{3.14}$$

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

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

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

$$\begin{aligned} \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= \frac{1}{2\sigma^2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \bigl(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top - \boldsymbol{M}^2 \bigr) (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\ &= \frac{1}{2} \bigl\{ D(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \bigr\}^\top \bigl(\boldsymbol{M}^{-1} \boldsymbol{\Psi}\boldsymbol{\Psi}^\top \boldsymbol{M}^{-1} - I_p \bigr) \bigl\{ D(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \bigr\}. \end{aligned} \tag{3.15}$$

Therefore, it holds for $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}$ on $\Omega(\mathtt{x})$ by (3.10)

$$\bigl| L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \bigr| \le \delta \mathtt{r}^2 / 2.$$

Differentiating the identity (3.15) in $\boldsymbol{\theta}$ yields for the gradients $\nabla L(\boldsymbol{\theta})$ and $\nabla \mathbb{L}(\boldsymbol{\theta})$

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

and thus, for any $\boldsymbol{\theta}$ with $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}$, (3.10) implies on $\Omega(\mathtt{x})$

$$\bigl\| D^{-1} \bigl\{ \nabla L(\boldsymbol{\theta}) - \nabla \mathbb{L}(\boldsymbol{\theta}) \bigr\} \bigr\| \le \delta \mathtt{r}.$$

One can represent the approximating process $\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ from (3.13) 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

$$\boldsymbol{\xi} \stackrel{\mathrm{def}}{=} \sigma^{-2} D^{-1} \boldsymbol{\Psi} (\boldsymbol{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}^*) = \sigma^{-2} D^{-1} \boldsymbol{\Psi} \boldsymbol{\varepsilon} = \sigma^{-1} \boldsymbol{M}^{-1} \boldsymbol{\Psi} \boldsymbol{\varepsilon}. \tag{3.16}$$

For studying the properties of the MLE $\widetilde{\boldsymbol{\theta}}$ and of the maximum log-likelihood $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta})$, one can use the expansions from Theorem 3.3.1. However, we present a direct proof based on quadraticity of $L(\boldsymbol{\theta})$.

Theorem 3.3.2. *Consider the model* (3.1) *and suppose*

$$\|\boldsymbol{M}^{-1} \boldsymbol{\Psi}\boldsymbol{\Psi}^\top \boldsymbol{M}^{-1} - I_p\|_{\mathrm{op}} \le \delta \tag{3.17}$$

for $\boldsymbol{M}^2 = I\!E(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)$ *and some* $\delta < 1/2$ *on a dominating set* $\Omega(\mathtt{x})$. *Then the MLE* $\widetilde{\boldsymbol{\theta}}$ *fulfills on* $\Omega(\mathtt{x})$ *for* D^2 *from* (3.14) *and* $\boldsymbol{\xi}$ *from* (3.16)

$$\begin{aligned} \bigl\| D\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) - \boldsymbol{\xi} \bigr\| \ &\leq \frac{\delta}{1-\delta} \|\boldsymbol{\xi}\|, && (3.18) \\ \bigl| 2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\boldsymbol{\xi}\|^2 \bigr| \ &\leq \frac{\delta}{1-\delta} \|\boldsymbol{\xi}\|^2, && (3.19) \\ \left| \sqrt{2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)} - \|\boldsymbol{\xi}\| \right| &\leq \frac{\delta}{1-\delta} \ \|\boldsymbol{\xi}\|. && (3.20) \end{aligned}$$

Proof. The bound (3.17) also implies

$$\| \boldsymbol{M} (\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)^{-1} \boldsymbol{M} - \boldsymbol{I}_p \|_{\mathrm{op}} \leq \frac{\delta}{1-\delta} \, . \qquad (3.21)$$

By using quadraticity of $L(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$, one obtains (cf. Theorem 1.5.1 in Section 1.5)

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

Further, the model equation (3.1) implies with $\boldsymbol{\xi} = \sigma^{-1} \boldsymbol{M}^{-1} \boldsymbol{\Psi} \boldsymbol{\varepsilon}$

$$D\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) = D (\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)^{-1} \boldsymbol{\Psi} \boldsymbol{\varepsilon} = \boldsymbol{M} (\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)^{-1} \boldsymbol{M} \, \boldsymbol{\xi} = A_{\Psi} \boldsymbol{\xi}$$

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

$$\bigl\| D\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) - \boldsymbol{\xi} \bigr\| = \bigl\| \bigl(A_{\Psi} - \boldsymbol{I}_p \bigr) \boldsymbol{\xi} \bigr\|$$

so that (3.18) follows from (3.21). Similarly

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

yielding (3.19) on $\Omega(\mathtt{x})$. Finally, as $L\bigl(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*\bigr) \geq 0$, it holds

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

yielding (3.20) by (3.19).

3.4 A deviation bound for $\boldsymbol{\xi}$

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

One can easily compute the moments of $\boldsymbol{\xi}$ conditioned on the design: by (3.2) $I\!\!E\bigl(\boldsymbol{\xi} \bigm| \boldsymbol{\Psi}\bigr) = 0$, $\operatorname{Var}\bigl(\boldsymbol{\varepsilon} \bigm| \boldsymbol{\Psi}\bigr) = \Sigma$, and

$$
\begin{aligned}
\mathcal{A}_{\boldsymbol{\xi}} &\stackrel{\text{def}}{=} \operatorname{Var}\bigl(\boldsymbol{\xi} \bigm| \boldsymbol{\Psi}\bigr) = \sigma^{-2}\boldsymbol{M}^{-1}(\boldsymbol{\Psi}\Sigma\boldsymbol{\Psi}^{\top})\boldsymbol{M}^{-1} \\
&= \sigma^{-2}\boldsymbol{M}^{-1}\left(\sum_{i=1}^{n}\sigma_i^2\,\Psi_i\,\Psi_i^{\top}\right)\boldsymbol{M}^{-1}.
\end{aligned}
\tag{3.22}
$$

Define

$$
B_{\boldsymbol{\xi}} \stackrel{\text{def}}{=} \operatorname{Var}(\boldsymbol{\xi}) = I\!\!E\mathcal{A}_{\boldsymbol{\xi}} = \sigma^{-2}\boldsymbol{M}^{-1}I\!\!E\left(\sum_{i=1}^{n}\sigma_i^2\,\Psi_i\,\Psi_i^{\top}\right)\boldsymbol{M}^{-1}. \tag{3.23}
$$

Similarly to Theorem 3.2.1, the conditional variance $\operatorname{Var}\bigl(\boldsymbol{\xi} \bigm| \boldsymbol{\Psi}\bigr)$ in (3.22) is close to its expectation $B_{\boldsymbol{\xi}}$. This allows to state the following deviation bound for $\|\boldsymbol{\xi}\|$.

Theorem 3.4.1. *Let the design $\boldsymbol{\Psi}$ and the noise variance Σ be such that*

$$
\|B_{\boldsymbol{\xi}}^{-1/2}\mathcal{A}_{\boldsymbol{\xi}}B_{\boldsymbol{\xi}}^{-1/2} - I_p\|_{\mathrm{op}} \le \delta_1 \tag{3.24}
$$

with $\delta_1 = \delta_1(\mathtt{x})$ on a set $\Omega_1(\mathtt{x})$ with $I\!\!P\bigl(\Omega_1(\mathtt{x})\bigr) \ge 1 - \mathrm{e}^{-\mathtt{x}}$. Let also the errors ε_i be conditionally on $\boldsymbol{\Psi}$ Gaussian zero mean with $\sigma_i^2 = \operatorname{Var}(\varepsilon_i \bigm| \Psi_i)$. Then it holds for the vector $\boldsymbol{\xi} = \sigma^{-1}\boldsymbol{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\varepsilon}$

$$
I\!\!P\Bigl\{\|\boldsymbol{\xi}\| \ge (1+\delta_1)\,z(B_{\boldsymbol{\xi}},\mathtt{x})\Bigr\} \le 2\mathrm{e}^{-\mathtt{x}}; \tag{3.25}
$$

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

Proof. Let (3.24) hold on a set $\Omega_1(\mathtt{x})$ with $I\!\!P\bigl(\Omega_1(\mathtt{x})\bigr) \ge 1 - \mathrm{e}^{-\mathtt{x}}$ for some value $\delta_1 = \delta_1(\mathtt{x})$. It helps to bound the moment generating function of $\boldsymbol{\xi}$. If $\boldsymbol{\varepsilon}$ is Gaussian conditioned on $\boldsymbol{\Psi}$, then , it holds for any $\lambda > 0$ and any vector $\boldsymbol{\gamma} \in I\!\!R^p$

$$
\log I\!\!E\bigl\{\exp\bigl(\lambda\,\boldsymbol{\gamma}^{\top}\boldsymbol{\xi}\bigr) \bigm| \boldsymbol{\Psi}\bigr\} = \frac{\lambda^2\|\mathcal{A}_{\boldsymbol{\xi}}^{1/2}\boldsymbol{\gamma}\|^2}{2}.
$$

This implies by (3.24) on $\Omega_1(\mathtt{x})$

$$
\begin{aligned}
\log I\!\!E\Bigl\{\exp\Bigl(\lambda\frac{\boldsymbol{\gamma}^{\top}\boldsymbol{\xi}}{\|B_{\boldsymbol{\xi}}^{1/2}\boldsymbol{\gamma}\|}\Bigr)\,\mathbb{1}\bigl(\Omega_1(\mathtt{x})\bigr)\Bigr\} &\le \log I\!\!E\exp\left\{\frac{\lambda^2\|\mathcal{A}_{\boldsymbol{\xi}}^{1/2}\boldsymbol{\gamma}\|^2}{2\|B_{\boldsymbol{\xi}}^{1/2}\boldsymbol{\gamma}\|^2}\,\mathbb{1}\bigl(\Omega_1(\mathtt{x})\bigr)\right\} \\
&\le \frac{(1+\delta_1)^2\lambda^2}{2}.
\end{aligned}
$$

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

One can combine the expansions from previous section with the bound (3.25). In particular, putting all together yields on the set $\Omega_2(\mathtt{x}) = \Omega(\mathtt{x}) \cup \Omega_1(\mathtt{x})$ with $I\!\!P\bigl(\Omega_2(\mathtt{x}) \geq 1 - 3\mathrm{e}^{-\mathtt{x}}\bigr)$

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

(please check).

3.5 Misspecified linear modeling assumption

The derivations of previous sections explicitly used the linear modeling assumption (3.1). This section discusses what changes if this assumption is not fulfilled. Namely, we only assume that the response variable Y and the feature vector Ψ are correlated. In this situation, the errors $\boldsymbol{\varepsilon}$ can be defined via conditional expectation $\boldsymbol{\varepsilon} = \boldsymbol{Y} - I\!\!E\bigl(\boldsymbol{Y} \bigm| \boldsymbol{\Psi}\bigr)$, and the parameter vector $\boldsymbol{\theta}^{*}$ can be naturally associated with the canonical correlation coefficients:

$$\boldsymbol{\theta}^{*} \stackrel{\mathrm{def}}{=} \bigl\{ I\!\!E(\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}) \bigr\}^{-1} I\!\!E(\boldsymbol{\Psi}\boldsymbol{Y}). \tag{3.26}$$

It is instructive to check that this definition becomes identity if (3.1) holds. In a slightly different way, one can define $\boldsymbol{\theta}^{*}$ by projection:

$$\boldsymbol{\theta}^{*} = \operatorname*{arginf}_{\boldsymbol{\theta}} I\!\!E \|\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}\|^{2}.$$

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

$$I\!\!E\bigl\{\boldsymbol{\Psi}(\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}^{*})\bigr\} = I\!\!E\bigl\{\boldsymbol{\Psi}\bigl(\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}^{*}\bigr)\bigr\} = 0.$$

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

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

The corresponding score vector

$$\boldsymbol{\xi} = D^{-1} \nabla L(\boldsymbol{\theta}^{*}) = D^{-1}\bigl\{\nabla L(\boldsymbol{\theta}^{*}) - \nabla I\!\!E L(\boldsymbol{\theta}^{*})\bigr\}$$

can be written as

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

Define a bias function

$$b(\boldsymbol{\Psi}) \overset{\text{def}}{=} \mathbb{E}\bigl(\boldsymbol{Y} \bigm| \boldsymbol{\Psi}\bigr) - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}^{*} \tag{3.27}$$

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

$$\boldsymbol{Y} = \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}^{*} + b(\boldsymbol{\Psi}) + \boldsymbol{\varepsilon} \tag{3.28}$$

for $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \mathbb{E}\bigl(\boldsymbol{Y} \bigm| \boldsymbol{\Psi}\bigr)$; cf. with the model equation (3.1) under correct model specification. Moreover, the definition of $\boldsymbol{\varepsilon}$ implies $\mathbb{E}(\boldsymbol{\Psi}\boldsymbol{\varepsilon}) = 0$, and the definitions (3.26) and (3.27) imply $\mathbb{E}\bigl[\boldsymbol{\Psi}\, b(\boldsymbol{\Psi})\bigr] = 0$. For the vector $\boldsymbol{\xi}$, we obtain the decomposition

$$\boldsymbol{\xi} = \sigma^{-2}D^{-1}\bigl\{\boldsymbol{\Psi}\, b(\boldsymbol{\Psi}) + \boldsymbol{\Psi}\boldsymbol{\varepsilon}\bigr\} = \boldsymbol{\xi}_{0} + \boldsymbol{\xi}_{1}\,, \tag{3.29}$$

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

$$\begin{aligned} \boldsymbol{\xi}_{0} &\overset{\text{def}}{=} \sigma^{-1}\boldsymbol{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\varepsilon} &&= \sigma^{-1}\boldsymbol{M}^{-1}\boldsymbol{\Psi}\bigl\{\boldsymbol{Y} - \mathbb{E}\bigl(\boldsymbol{Y} \bigm| \boldsymbol{\Psi}\bigr)\bigr\}, \\ \boldsymbol{\xi}_{1} &\overset{\text{def}}{=} \sigma^{-1}\boldsymbol{M}^{-1}\boldsymbol{\Psi}\, b(\boldsymbol{\Psi}) &&= \sigma^{-1}\boldsymbol{M}^{-1}\boldsymbol{\Psi}\bigl\{\mathbb{E}(\boldsymbol{Y} \bigm| \boldsymbol{\Psi}) - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}^{*}\bigr\}. \end{aligned} \tag{3.30}$$

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 $\boldsymbol{Y}$. The term $\boldsymbol{\xi}_{1}$ appears only if the model assumption $\mathbb{E}(\boldsymbol{Y} \bigm| \boldsymbol{\Psi}) = \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}^{*}$ is not correct. It only relies to the random design distribution. By construction $\mathbb{E}(\boldsymbol{\xi}_{0} \bigm| \boldsymbol{\Psi}) = 0$, in particular, $\boldsymbol{\xi}_{0}$ is zero mean. The same is true for $\boldsymbol{\xi}_{1}$.

For the LSE $\widetilde{\boldsymbol{\theta}} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top})^{-1}\boldsymbol{\Psi}\boldsymbol{Y}$, it holds from (3.28) in the case of a non-degenerated design matrix $\boldsymbol{M} = \boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}$

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

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

$$\begin{aligned} D\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{*}\bigr) &= \boldsymbol{M}(\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top})^{-1}\boldsymbol{M}\boldsymbol{\xi}_{0} + \boldsymbol{M}(\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top})^{-1}\boldsymbol{M}\boldsymbol{\xi}_{1} \\ &= A_{\Psi}\bigl(\boldsymbol{\xi}_{0} + \boldsymbol{\xi}_{1}\bigr) = A_{\Psi}\boldsymbol{\xi}\,. \end{aligned} \tag{3.31}$$

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 (3.27). The origin of this term is that the target $\boldsymbol{\theta}^{*} = \mathbb{E}(\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top})^{-1}\mathbb{E}(\boldsymbol{\Psi}\boldsymbol{Y})$ is defined under the design measure while the construction of the estimate $\widetilde{\boldsymbol{\theta}}$ is based on its empirical counterpart. In particular, as $I\!\!E\bigl(\boldsymbol{\varepsilon} \,\big|\, \boldsymbol{\Psi}\bigr) = 0$, it holds

$$\begin{aligned} I\!\!E\Bigl(\|D\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{*}\bigr)\|^{2} \,\Big|\, \boldsymbol{\Psi}\Bigr) &= I\!\!E\bigl(\|A_{\Psi}\boldsymbol{\xi}_{0}\|^{2} \,\big|\, \boldsymbol{\Psi}\bigr) + \bigl\|A_{\Psi}\boldsymbol{\xi}_{1}\bigr\|^{2} \\ &= A_{\Psi}\operatorname{Var}\bigl(\boldsymbol{\xi}_{0} \,\big|\, \boldsymbol{\Psi}\bigr)A_{\Psi} + \sigma^{-2}\bigl\|A_{\Psi}\boldsymbol{M}^{-1}\boldsymbol{\Psi}\, b(\boldsymbol{\Psi})\bigr\|^{2}. \end{aligned}$$

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 3.3.1 and 3.3.2 apply with such defined $\boldsymbol{\theta}^{*}$ and $\boldsymbol{\xi}$ even under model misspecification.

Theorem 3.5.1. *Consider the model* (3.1) *and suppose*

$$\|\boldsymbol{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}\boldsymbol{M}^{-1} - I_{p}\|_{\mathrm{op}} \le \delta \tag{3.32}$$

for $\boldsymbol{M}^{2} = I\!\!E(\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top})$ *and some* $\delta < 1/2$ *on a dominating set* $\Omega(\mathtt{x})$. *Then the MLE* $\widetilde{\boldsymbol{\theta}}$ *fulfills on* $\Omega(\mathtt{x})$ *for* $\boldsymbol{\xi}$ *from* (3.29)

$$\bigl\|D\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{*}\bigr) - \boldsymbol{\xi}\bigr\| \le \frac{\delta}{1-\delta}\|\boldsymbol{\xi}\|, \tag{3.33}$$

$$\bigl|2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{*}) - \|\boldsymbol{\xi}\|^{2}\bigr| \le \frac{\delta}{1-\delta}\|\boldsymbol{\xi}\|^{2}. \tag{3.34}$$

Proof. The bound (3.32) implies $\|I_{p} - A_{\Psi}\|_{\mathrm{op}} \le \delta/(1-\delta)$, and (3.33) follows from the decomposition (3.31). Further, for any $\boldsymbol{\theta}$, quadraticity of $L(\boldsymbol{\theta})$ implies

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

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

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

Now (3.34) follows from (3.31).

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

$$\|\boldsymbol{\xi}\| \le \|\boldsymbol{\xi}_{0}\| + \|\boldsymbol{\xi}_{1}\| = \|\sigma^{-1}\boldsymbol{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\varepsilon}\| + \|\sigma^{-1}\boldsymbol{M}^{-1}\boldsymbol{\Psi}\, b(\boldsymbol{\Psi})\|.$$

The term $\boldsymbol{\xi}_{0}$ from (3.30) can be bounded by Theorem 3.4.1 provided that the conditional variance $\mathcal{A}_{\boldsymbol{\xi}} = \operatorname{Var}\bigl(\boldsymbol{\xi} \,\big|\, \boldsymbol{\Psi}\bigr)$ of $\boldsymbol{\xi}$ is close to its unconditional counterpart $B_{\boldsymbol{\xi}} = I\!\!E\mathcal{A}_{\boldsymbol{\xi}}$ and the errors $\varepsilon_{i} = Y_{i} - I\!\!E\bigl(Y_{i} \,\big|\, \Psi_{i}\bigr)$ 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 3.5.2. *Let, given* $\mathtt{x}$, *condition* (3.24) *hold with* $\delta_1 = \delta_1(\mathtt{x})$ *on a set* $\Omega_1(\mathtt{x})$ *with* $I\!P\bigl(\Omega_1(\mathtt{x})\bigr) \geq 1 - \mathrm{e}^{-\mathtt{x}}$. *Let also* $\varepsilon_i \bigm| \Psi_i \sim \mathcal{N}(0, \sigma_i^2)$ *and the matrix* $B_{\boldsymbol{\xi}}$ *is given by* (3.23). *Let, for the vector* $b(\boldsymbol{\Psi})$, *the matrix* B_b *be defined by*

$$B_b \stackrel{\mathrm{def}}{=} \operatorname{Var}\bigl\{\boldsymbol{M}^{-1}\boldsymbol{\Psi}\, b(\boldsymbol{\Psi})\bigr\} = \boldsymbol{M}^{-1} I\!E\biggl\{\sum_{i=1}^{n} b_i^2\, \Psi_i\, \Psi_i^{\top}\biggr\}\boldsymbol{M}^{-1}.$$

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

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

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

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

Proof. The bound for $\boldsymbol{\xi}_0$ is already proved in Theorem 3.4.1. It remains to show that

$$I\!P\left(\bigl\|\boldsymbol{M}^{-1}\boldsymbol{\Psi}\, b(\boldsymbol{\Psi})\bigr\| \geq z(B_b, \mathtt{x})\right) \leq 2\mathrm{e}^{-\mathtt{x}}. \tag{3.35}$$

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

$$\operatorname{Var}\bigl\{\boldsymbol{M}^{-1}\boldsymbol{\Psi}\, b(\boldsymbol{\Psi})\bigr\} = \boldsymbol{M}^{-1} I\!E\bigl\{\boldsymbol{\Psi}\, b(\boldsymbol{\Psi})\, b(\boldsymbol{\Psi})^{\top}\boldsymbol{\Psi}^{\top}\bigr\}\boldsymbol{M}^{-1}.$$

Now the result (3.35) follows by Theorem C.1.1.

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

$$\operatorname{Var}\bigl\{\boldsymbol{M}^{-1}\boldsymbol{\Psi}\, b(\boldsymbol{\Psi})\bigr\} \leq b_{\infty}^2 \boldsymbol{M}^{-1} I\!E\bigl(\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}\bigr)\boldsymbol{M}^{-1} = b_{\infty}^2 I_p\,,$$

and by Theorem C.2.1

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

3.6 Application to instrumental regression

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

$$Y = f(X) + U, \qquad I\!E\bigl[U \mid W\bigr] = 0.$$

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

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

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

or in vector form

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

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

$$\boldsymbol{Y} = \boldsymbol{\psi}(X)^\top \boldsymbol{\theta}^* + U, \qquad I\!\!E\bigl[U \mid W\bigr] = 0.$$

The constraint $I\!\!E\bigl[U \mid W\bigr] = 0$ means that for any function $\phi(W)$

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

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

$$I\!\!E\bigl[Y\boldsymbol{\phi}(W)\bigr] = I\!\!E\bigl[\boldsymbol{\psi}(X)\boldsymbol{\phi}(W)^\top\bigr]^\top \boldsymbol{\theta}^* = \boldsymbol{T}^\top\boldsymbol{\theta}^*$$

with

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

Define

$$\begin{aligned} \boldsymbol{Z} &= I\!\!E_n\bigl[\boldsymbol{Y}\,\boldsymbol{\phi}(W)\bigr] && = n^{-1}\textstyle\sum_i Y_i\boldsymbol{\phi}(W_i) && \in I\!\!R^q, \\ \boldsymbol{T}_n &= I\!\!E_n\bigl[\boldsymbol{\psi}(X)\,\boldsymbol{\phi}(W)^\top\bigr] && = n^{-1}\textstyle\sum_i \boldsymbol{\psi}(X_i)\,\boldsymbol{\phi}(W_i)^\top && \in I\!\!R^{q\times p}, \\ \boldsymbol{\varepsilon} &= I\!\!E_n\bigl[\boldsymbol{\phi}(W)\,\boldsymbol{U}\bigr] && = n^{-1}\textstyle\sum_i \boldsymbol{\phi}(W_i)\,U_i && \in I\!\!R^q. \end{aligned}$$

The original problems reduces to

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

where $\boldsymbol{\varepsilon}$ is the error q-vector, $\boldsymbol{T} = I\!\!E\bigl[\boldsymbol{\psi}(X)\,\boldsymbol{\phi}(W)^\top\bigr]$ is an unknown $p\times q$ matrix and only its empirical counterpart $\boldsymbol{T}_n$ 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 $\boldsymbol{T}_n$ is random and correlated with $\boldsymbol{Z}$ and $\boldsymbol{\varepsilon}$. The goal is to build an estimator $\widetilde{\boldsymbol{\theta}}$ of the vector $\boldsymbol{\theta}^*$ leading to the estimator $\widetilde{f}(x) = \boldsymbol{\psi}(x)^\top\widetilde{\boldsymbol{\theta}}$ of the response.

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

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

with the random design $\boldsymbol{T}_n = n^{-1}\sum_i \boldsymbol{\psi}(X_i)\,\boldsymbol{\phi}(W_i)^\top$ so that the setup (3.11) applies. The corresponding least square estimator of $\boldsymbol{\theta}^*$ reads as

$$\widetilde{\boldsymbol{\theta}} = \bigl(\boldsymbol{T}_n\boldsymbol{T}_n^\top\bigr)^{-1}\boldsymbol{T}_n\boldsymbol{Z}. \tag{3.36}$$

The results of Theorem 3.2.2 justify that the random matrix $\boldsymbol{T}_n\boldsymbol{T}_n^\top$ is very close to the product $I\!E(\boldsymbol{T}_n)I\!E(\boldsymbol{T}_n^\top) = \boldsymbol{T}\boldsymbol{T}^\top$ and the theoretical study of the properties of the estimator $\widetilde{\boldsymbol{\theta}}$ can be done with $\boldsymbol{T}\boldsymbol{T}^\top$ in place of $\boldsymbol{T}_n\boldsymbol{T}_n^\top$ in (3.36). Similarly one can justify that the product $\boldsymbol{T}_n\boldsymbol{Z}$ behaves nearly as $\boldsymbol{T}\boldsymbol{Z}$.

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

$$\sigma_1^2 = \bigl\|\boldsymbol{M}^{-1}I\!E(T_iT_i^\top)\boldsymbol{M}^{-1} - \boldsymbol{I}_p\bigr\|_{\mathrm{op}}$$

and suppose that it holds almost surely

$$\|\boldsymbol{M}^{-1}(T_i - \boldsymbol{T})\|_{\mathrm{op}} \le u.$$

Now Theorem 3.2.2 implies for any $z > 0$

$$I\!P\Bigl(\sqrt{n}\bigl\|\boldsymbol{M}^{-1}\bigl(\boldsymbol{T}_n - \boldsymbol{T}\bigr)\bigr\|_{\mathrm{op}} > z\Bigr) \le 2(p+q)\exp\Bigl\{-\frac{z^2}{2\sigma_1^2 + 2uz/(3n^{1/2})}\Bigr\}.$$

Moreover, if $\|\boldsymbol{M}^{-1}\bigl(\boldsymbol{T}_n - \boldsymbol{T}\bigr)\|_{\mathrm{op}} \le \delta$, then

$$\bigl\|\boldsymbol{M}^{-1}\,\boldsymbol{T}_n\boldsymbol{T}_n^\top\,\boldsymbol{M}^{-1} - \boldsymbol{I}_p\bigr\|_{\mathrm{op}} \le \delta^2 + 2\delta.$$

4

Linear smoothers

Here we discuss the important situation when the number of predictors $\boldsymbol{\psi}_j$ and hence the number of parameters p in the linear model $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ 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 $\widetilde{\boldsymbol{\theta}}$ 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 $\boldsymbol{Y}$ follows the model

$$\boldsymbol{Y} = \boldsymbol{f}^{*} + \boldsymbol{\varepsilon} \tag{4.1}$$

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

Furthermore, we assume a basis or a collection of basis vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p$ is given with p large. This allows for approximating the response vector $\boldsymbol{f}^{*} = \mathbb{E}\boldsymbol{Y}$ in the form $\boldsymbol{f}^{*} = \Psi^{\top}\boldsymbol{\theta}^{*}$, or, equivalently,

$$\boldsymbol{f}^{*} = \theta_1^{*}\boldsymbol{\psi}_1 + \ldots + \theta_p^{*}\boldsymbol{\psi}_p \,.$$

In many cases we will assume that the basis is already orthogonalized: $\Psi\Psi^\top = \boldsymbol{I}_p$. The model (4.1) can be rewritten as

$$\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}, \qquad \operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n .$$

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

$$\widetilde{\boldsymbol{\theta}} = \bigl(\Psi\Psi^\top\bigr)^{-1}\Psi\boldsymbol{Y}, \qquad \widetilde{\boldsymbol{f}} = \Psi^\top\widetilde{\boldsymbol{\theta}} = \Psi^\top\bigl(\Psi\Psi^\top\bigr)^{-1}\Psi\boldsymbol{Y}.$$

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

4.1 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 $\bigl(\Psi\Psi^\top\bigr)^{-1}$ by $\bigl(\Psi\Psi^\top + R\bigr)^{-1}$ leading to the regularized least squares estimate $\widetilde{\boldsymbol{\theta}}_R$ of the parameter vector $\boldsymbol{\theta}$ and the corresponding response estimate $\widetilde{\boldsymbol{f}}_R$:

$$\widetilde{\boldsymbol{\theta}}_R \stackrel{\text{def}}{=} \bigl(\Psi\Psi^\top + R\bigr)^{-1}\Psi\boldsymbol{Y}, \qquad \widetilde{\boldsymbol{f}}_R \stackrel{\text{def}}{=} \Psi^\top\bigl(\Psi\Psi^\top + R\bigr)^{-1}\Psi\boldsymbol{Y}. \tag{4.2}$$

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

$$\widetilde{\boldsymbol{\theta}}_\alpha \stackrel{\text{def}}{=} \bigl(\Psi\Psi^\top + \alpha\boldsymbol{I}_p\bigr)^{-1}\Psi\boldsymbol{Y}, \qquad \widetilde{\boldsymbol{f}}_\alpha \stackrel{\text{def}}{=} \Psi^\top\bigl(\Psi\Psi^\top + \alpha\boldsymbol{I}_p\bigr)^{-1}\Psi\boldsymbol{Y}. \tag{4.3}$$

A proper choice of the matrix R for the ridge regression method (4.2) or the parameter α for the Tikhonov regularization (4.3) is an important issue. Below we discuss several approaches which lead to the estimate (4.2) with a specific choice of the matrix R. The properties of the estimates $\widetilde{\boldsymbol{\theta}}_R$ and $\widetilde{\boldsymbol{f}}_R$ will be studied in context of penalized likelihood estimation in the next section.

4.2 Penalized likelihood. Bias and variance

The estimate (4.2) 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:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname*{arginf}_{\boldsymbol{\theta}} \|\boldsymbol{Y} - \Psi^{\top}\boldsymbol{\theta}\|^2$$

with $L(\boldsymbol{\theta}) = \sigma^{-2}\|\boldsymbol{Y} - \Psi^{\top}\boldsymbol{\theta}\|^2/2$. (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*

$$\begin{aligned} L_G(\boldsymbol{\theta}) &\overset{\text{def}}{=} L(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2/2 \\ &= -(2\sigma^2)^{-1}\|\boldsymbol{Y} - \Psi^{\top}\boldsymbol{\theta}\|^2 - \|G\boldsymbol{\theta}\|^2/2 - (n/2)\log(2\pi\sigma^2). \end{aligned} \tag{4.4}$$

The penalized MLE reads as

$$\widetilde{\boldsymbol{\theta}}_G = \operatorname*{argmax}_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \bigl\{(2\sigma^2)^{-1}\|\boldsymbol{Y} - \Psi^{\top}\boldsymbol{\theta}\|^2 + \|G\boldsymbol{\theta}\|^2/2\bigr\}.$$

A straightforward calculus leads to the expression (4.2) for $\widetilde{\boldsymbol{\theta}}_G$ with $R = \sigma^2 G^2$:

$$\widetilde{\boldsymbol{\theta}}_G \overset{\text{def}}{=} \bigl(\Psi\Psi^{\top} + \sigma^2 G^2\bigr)^{-1}\Psi\boldsymbol{Y}. \tag{4.5}$$

We see that $\widetilde{\boldsymbol{\theta}}_G$ is again a linear estimate: $\widetilde{\boldsymbol{\theta}}_G = \mathcal{S}_G\boldsymbol{Y}$ with $\mathcal{S}_G = \bigl(\Psi\Psi^{\top} + \sigma^2 G^2\bigr)^{-1}\Psi$. The results of Section 1.4 explains that $\widetilde{\boldsymbol{\theta}}_G$ in fact estimates the value $\boldsymbol{\theta}_G$ defined by

$$\begin{aligned} \boldsymbol{\theta}_G &= \operatorname*{argmax}_{\boldsymbol{\theta}} I\!\!E L_G(\boldsymbol{\theta}) \\ &= \operatorname*{arginf}_{\boldsymbol{\theta}} I\!\!E\bigl\{\|\boldsymbol{Y} - \Psi^{\top}\boldsymbol{\theta}\|^2 + \sigma^2\|G\boldsymbol{\theta}\|^2\bigr\} \\ &= \bigl(\Psi\Psi^{\top} + \sigma^2 G^2\bigr)^{-1}\Psi\boldsymbol{f}^* = \mathcal{S}_G\boldsymbol{f}^*. \end{aligned} \tag{4.6}$$

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

$$\boldsymbol{\theta}_G = \bigl(\Psi\Psi^{\top} + \sigma^2 G^2\bigr)^{-1}\Psi\Psi^{\top}\boldsymbol{\theta}^* \tag{4.7}$$

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

Exercise 4.2.1. Check that $I\!\!E\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\alpha}} = \boldsymbol{\theta}_{\boldsymbol{\alpha}}$ for $\boldsymbol{\theta}_{\alpha} = \bigl(\Psi\Psi^{\top} + \alpha I_p\bigr)^{-1}\Psi\Psi^{\top}\boldsymbol{\theta}^*$, the bias $\|\boldsymbol{\theta}_{\alpha} - \boldsymbol{\theta}^*\|$ grows with the regularization parameter α.

The penalized MLE $\widetilde{\boldsymbol{\theta}}_G$ leads to the response estimate $\widetilde{\boldsymbol{f}}_G = \Psi^{\top}\widetilde{\boldsymbol{\theta}}_G$. It can be written as

$$\widetilde{\boldsymbol{f}}_G = \Psi^{\top}\widetilde{\boldsymbol{\theta}}_G = \Psi^{\top}\bigl(\Psi\Psi^{\top} + \sigma^2 G^2\bigr)^{-1}\Psi\boldsymbol{Y} = \Pi_G\boldsymbol{Y} \tag{4.8}$$

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

Exercise 4.2.2. Show that Π_G is a sub-projector in the sense that $\|\Pi_G \boldsymbol{u}\| \le \|\boldsymbol{u}\|$ for any $\boldsymbol{u} \in \mathbb{R}^n$.

Exercise 4.2.3. Let Ψ be orthonormal: $\Psi\Psi^\top = I_p$. Then the penalized MLE $\widetilde{\boldsymbol{\theta}}_G$ can be represented as

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

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

4.3 Bias-variance decomposition for the quadratic risk

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 show that penalization decreases the variance of estimation and thus, makes the procedure more stable. First we consider the quadratic risk of response estimation.

Theorem 4.3.1. *Let $\widetilde{\boldsymbol{f}}_G$ be a penalized estimator from (4.8). Then $\mathbb{E}\widetilde{\boldsymbol{f}}_G = \Psi^\top \boldsymbol{\theta}_G$; see (4.7). Moreover, under noise homogeneity $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$, it holds*

$$\mathbb{E}\|\widetilde{\boldsymbol{f}}_G - \boldsymbol{f}^*\|^2 = \|\Pi_G \boldsymbol{f}^* - \boldsymbol{f}^*\|^2 + \sigma^2 \operatorname{tr}(\Pi_G^2). \tag{4.9}$$

Proof. The model equation $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$ with $\mathbb{E}\boldsymbol{\varepsilon} = 0$ yields the first assertion for $\widetilde{\boldsymbol{f}}_G = \Pi_G \boldsymbol{Y}$. For the squared loss $\wp(\widetilde{\boldsymbol{f}}_G, \boldsymbol{f}^*) = \|\widetilde{\boldsymbol{f}}_G - \boldsymbol{f}^*\|^2$, we obtain the decomposition

$$\wp(\widetilde{\boldsymbol{f}}_G, \boldsymbol{f}^*) = \|\Pi_G(\boldsymbol{f}^* + \boldsymbol{\varepsilon}) - \boldsymbol{f}^*\|^2 = \|\Pi_G \boldsymbol{f}^* - \boldsymbol{f}^*\|^2 + \|\Pi_G \boldsymbol{\varepsilon}\|^2 + 2(\Pi_G \boldsymbol{f}^* - \boldsymbol{f}^*)^\top \Pi_G \boldsymbol{\varepsilon}.$$

Unfortunately, the cross term does not vanish unless Π_G is a projector. However, it is linear in the error term $\boldsymbol{\varepsilon}$ and hence, its expectation vanishes. Further, in view of $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$

$$\mathbb{E}\|\Pi_G \boldsymbol{\varepsilon}\|^2 = \mathbb{E}(\Pi_G \boldsymbol{\varepsilon})^\top \Pi_G \boldsymbol{\varepsilon} = \operatorname{tr} \mathbb{E}\bigl\{\Pi_G \boldsymbol{\varepsilon}\bigl(\Pi_G \boldsymbol{\varepsilon}\bigr)^\top\bigr\} = \operatorname{tr}\bigl\{\Pi_G \operatorname{Var}(\boldsymbol{\varepsilon})\Pi_G\bigr\} = \sigma^2 \operatorname{tr}(\Pi_G^2).$$

This implies (4.9).

Exercise 4.3.1. Show that the variance term $\sigma^2 \operatorname{tr}(\Pi_G^2)$ decreases with G^2: if $G_1^2 \ge G_2^2$, then $\operatorname{tr}(\Pi_{G_1}^2) \le \operatorname{tr}(\Pi_{G_2}^2)$.

Now we check the loss and risk of the penalized parameter estimator $\widetilde{\boldsymbol{\theta}}_G$.

Theorem 4.3.2. *Let $\widetilde{\boldsymbol{\theta}}_G = \mathcal{S}_G \boldsymbol{Y}$ be a penalized MLE from (4.5). Then $I\!\!E\widetilde{\boldsymbol{\theta}}_G = \boldsymbol{\theta}_G$; see (4.7). Under noise homogeneity* $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$ *, it holds*

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

with $D_G^2 = \sigma^{-2}\Psi\Psi^\top + G^2$. In particular, $\operatorname{Var}(\widetilde{\boldsymbol{\theta}}_G) \le D_G^{-2}$ *. If $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$, then $\widetilde{\boldsymbol{\theta}}_G$ is also normal: $\widetilde{\boldsymbol{\theta}}_G \sim \mathcal{N}(\boldsymbol{\theta}_G, D_G^{-2}D^2D_G^{-2})$.*

Proof. The first two moments of $\widetilde{\boldsymbol{\theta}}_G$ are computed from $\widetilde{\boldsymbol{\theta}}_G = \mathcal{S}_G \boldsymbol{Y}$. Monotonicity of the variance of $\widetilde{\boldsymbol{\theta}}_G$ is proved below in Exercise 4.3.3.

Exercise 4.3.2. Let Ψ be orthonormal: $\Psi\Psi^\top = \boldsymbol{I}_p$. Describe $\operatorname{Var}(\widetilde{\boldsymbol{\theta}}_G)$.

Exercise 4.3.3. Show that the variance of $\widetilde{\boldsymbol{\theta}}_G$ decreases with the penalization G^2 in the sense that $G_1^2 \ge G^2$ for two matrices G^2 and G_1^2 implies $\operatorname{Var}(\widetilde{\boldsymbol{\theta}}_{G_1}) \le \operatorname{Var}(\widetilde{\boldsymbol{\theta}}_G)$. Hint: with $D^2 = \sigma^{-2}\Psi\Psi^\top$, show that for any vector $\boldsymbol{w} \in I\!\!R^p$ and $\boldsymbol{u} = D^{-1}\boldsymbol{w}$, it holds

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

and this value decreases with G^2 because $I_p + D^{-1}G^2D^{-1}$ increases.

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

Theorem 4.3.3. *Assume the model $\boldsymbol{Y} = \Psi^\top\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with* $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}_n$ *. Then the estimate $\widetilde{\boldsymbol{\theta}}_G$ fulfills*

$$I\!\!E\|\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*\|^2 = \|\boldsymbol{\theta}_G - \boldsymbol{\theta}^*\|^2 + \operatorname{tr}\bigl(D_G^{-2}D^2D_G^{-2}\bigr).$$

This results of Theorem 4.3.1 and Theorem 4.3.3 are 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.*

Exercise 4.3.4. Let G^2 be a symmetric matrix and $\widetilde{\boldsymbol{\theta}}_G$ the corresponding penalized MLE. Show that the variance term $\operatorname{tr}\operatorname{Var}(\widetilde{\boldsymbol{\theta}}_G) = \operatorname{tr}\bigl(D_G^{-2}D^2D_G^{-2}\bigr)$ decreases in G^2 .

Exercise 4.3.5. Let $\Psi\Psi^\top = \boldsymbol{I}_p$ and let $G = \operatorname{diag}(g_1, \ldots, g_p)$ 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 $j = 1, \ldots, p$.

4.4 Inference for the penalized MLE

Here we discuss some properties of the penalized MLE $\widetilde{\boldsymbol{\theta}}_G$. In particular, we focus on the construction of confidence and concentration sets based on the penalized log-likelihood. We know that the regularized estimate $\widetilde{\boldsymbol{\theta}}_G$ is the empirical counterpart of the value $\boldsymbol{\theta}_G$ which solves the regularized deterministic problem (4.6). 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 1.6.1 to the penalized likelihood. It is entirely based on quadraticity of $L_G(\boldsymbol{\theta})$. Define for any $\boldsymbol{\theta}$

$$L_G(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}) \overset{\text{def}}{=} L_G(\widetilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}).$$

Theorem 4.4.1. *Let $L_G(\boldsymbol{\theta})$ be the penalized log-likelihood from (4.4). Then for any $\boldsymbol{\theta}$*

$$2L_G(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}) = \bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}\bigr)^\top D_G^2 \bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}\bigr). \tag{4.10}$$

Moreover, for $\boldsymbol{\theta}_G = \mathcal{S}_G \boldsymbol{f}^*$

$$\begin{aligned} 2L_G(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G) &= \bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G\bigr)^\top D_G^2 \bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G\bigr) \\ &= \sigma^{-2} \boldsymbol{\varepsilon}^\top \Pi_G \, \boldsymbol{\varepsilon} \end{aligned} \tag{4.11}$$

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

Proof. Use that $L_G(\boldsymbol{\theta})$ is quadratic in $\boldsymbol{\theta}$ with $\nabla^2 L_G(\boldsymbol{\theta}) = \sigma^2 \Psi\Psi^\top + G^2 = D_G^2$. Now we can apply Theorem 1.5.3.

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 4.4.1. Prove (4.10).
Hint: apply the Taylor expansion to $L_G(\boldsymbol{\theta})$ at $\widetilde{\boldsymbol{\theta}}_G$. Use that $\nabla L_G(\widetilde{\boldsymbol{\theta}}_G) = 0$ and $-\nabla^2 L_G(\boldsymbol{\theta}) \equiv \sigma^{-2}\Psi\Psi^\top + G^2$.

Exercise 4.4.2. Prove (4.11).
Hint: use that $\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G = \mathcal{S}_G \boldsymbol{\varepsilon}$ with $\mathcal{S}_G = \bigl(\Psi\Psi^\top + \sigma^2 G^2\bigr)^{-1} \Psi$.

The straightforward corollaries of Theorem 4.4.1 are the concentration and confidence probabilities. Define the confidence set $\mathcal{E}_G(\mathfrak{z})$ for $\boldsymbol{\theta}_G$ as

$$\mathcal{E}_G(\mathfrak{z}) \overset{\text{def}}{=} \bigl\{\boldsymbol{\theta} : L_G(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}) \le \mathfrak{z}\bigr\}.$$

The definition implies the following result for the coverage probability:

$$I\!P\bigl(\mathcal{E}_G(\mathfrak{z}) \not\ni \boldsymbol{\theta}_G\bigr) \le I\!P\bigl(L_G(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G) > \mathfrak{z}\bigr).$$

Now the representation (4.11) for $L_G(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G)$ reduces the problem to a deviation bound for a quadratic form. We apply the general result of Theorem C.1.1 in Section C.

Theorem 4.4.2. *Let* $L_G(\boldsymbol{\theta})$ *be the penalized log-likelihood from* (4.4) *and let* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$ *. Then it holds with* $\mathtt{p}_G = \operatorname{tr}(\Pi_G)$ *and* $\mathtt{v}_G^2 = \operatorname{tr}(\Pi_G^2)$ *that*

$$I\!P\bigl(2L_G(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G) > \mathtt{p}_G + 2\mathtt{v}_G \mathtt{x}^{1/2} + 2\mathtt{x}\bigr) \le \exp(-\mathtt{x}).$$

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

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

and the result of Theorem 4.4.2 can be restated as the concentration bound:

$$I\!P\bigl(\|D_G(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G)\|^2 > \mathtt{p}_G + 2\mathtt{v}_G \mathtt{x}^{1/2} + 2\mathtt{x}\bigr) \le \exp(-\mathtt{x}).$$

In other words, $\widetilde{\boldsymbol{\theta}}_G$ concentrates on the set $\mathcal{A}(\mathfrak{z}, \boldsymbol{\theta}_G) = \bigl\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_G\|^2 \le 2\mathfrak{z}\bigr\}$ for $2\mathfrak{z} > \mathtt{p}_G$.

4.5 Projection and shrinkage estimates

Consider a linear model $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$, where $\boldsymbol{f}^* = \Psi^\top \boldsymbol{\theta}^*$ is a decomposition of the function $\boldsymbol{f}^*$ using a large system of basis functions ψ_j , $j = 1, \ldots, p$. We also assume that the matrix Ψ is orthonormal in the sense $\Psi\Psi^\top = \boldsymbol{I}_p$. Then the multiplication with Ψ maps this model in the sequence space model $\boldsymbol{Z} = \boldsymbol{\theta}^* + \boldsymbol{\xi}$, where $\boldsymbol{Z} = \Psi \boldsymbol{Y} = (z_1, \ldots, z_p)^\top$ is the vector of empirical Fourier coefficients $z_j = \boldsymbol{\psi}_j^\top \boldsymbol{Y}$. 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 $z_1, \ldots, z_m$ and drop the others. The corresponding parameter estimate $\widetilde{\boldsymbol{\theta}}_m$ reads as

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

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

$$\widetilde{\boldsymbol{f}}_m = z_1 \boldsymbol{\psi}_1 + \ldots + z_m \boldsymbol{\psi}_m$$

with $z_j = \boldsymbol{\psi}_j^\top \boldsymbol{Y}$. 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 $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\alpha}}$ with

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

Here $\boldsymbol{\alpha}$ stands for the vector of coefficients α_j for $j = 1,\ldots,p$. 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

$$\alpha_j = (1 - j/m)^{\beta} \mathbf{1}(j \le m) \tag{4.12}$$

for some $\beta > 0$ and $m \le p$. 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 $\Psi\Psi^{\top} = \boldsymbol{I}_p$ yields again that the estimation risk $\mathbb{E}\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}^*\|^2$ coincides with the prediction risk $\mathbb{E}\|\widetilde{\boldsymbol{f}}_{\boldsymbol{\alpha}} - \boldsymbol{f}^*\|^2$.

Exercise 4.5.1. Let $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_p$. The risk $\mathcal{R}(\widetilde{\boldsymbol{f}}_{\boldsymbol{\alpha}})$ of the shrinkage estimate $\widetilde{\boldsymbol{f}}_{\boldsymbol{\alpha}}$ fulfills

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

Specify the cases of $\boldsymbol{\alpha} = \boldsymbol{\alpha}(m,\beta)$ from (4.12). 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:

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

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 $\boldsymbol{Z} = \boldsymbol{\theta}^* + \boldsymbol{\xi}$ is by using a roughness penalization. Let G be a symmetric matrix. Consider the regularized estimate $\widetilde{\boldsymbol{\theta}}_G$ from (4.2). The next result claims that if G is a diagonal matrix, then $\widetilde{\boldsymbol{\theta}}_G$ is a shrinkage estimate. Moreover, a general penalized MLE can be represented as shrinkage by an orthogonal basis transformation.

Theorem 4.5.1. *Let G be a diagonal matrix, $G = \operatorname{diag}(g_1,\ldots,g_p)$. The penalized MLE $\widetilde{\boldsymbol{\theta}}_G$ in the sequence space model $\boldsymbol{Z} = \boldsymbol{\theta}^* + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 I_p)$ coincides with the shrinkage estimate $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\alpha}}$ for $\alpha_j = (1 + \sigma^2 g_j^2)^{-1} \le 1$. Moreover, a penalized MLE $\widetilde{\boldsymbol{\theta}}_G$ 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 $\widetilde{\boldsymbol{\theta}}_G = (I_p + \sigma^2 G^2)^{-1} \boldsymbol{Z}$. Next, let U be an orthogonal transform leading to the diagonal representation $G^2 = U^\top D^2 U$ with $D^2 = \operatorname{diag}(g_1, \ldots, g_p)$. Then

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

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

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

Exercise 4.5.2. Let $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\alpha}}$ is a shrinkage estimate for a vector $\boldsymbol{\alpha} = (\alpha_j)$. Then there is a diagonal penalty matrix G such that $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\alpha}} = \widetilde{\boldsymbol{\theta}}_G$.
Hint: define the j th diagonal entry g_j by the equation $\alpha_j = (1 + \sigma^2 g_j^2)^{-1}$.

4.6 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 nonlinear 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 $\boldsymbol{f}^*$ has a bounded L_2-norm. A general Sobolev ball can be defined using a diagonal matrix G:

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

Now we consider a constrained ML problem:

$$\widetilde{\boldsymbol{\theta}}_{G,R} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \mathcal{B}_G(R)} L(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta \colon \|G\boldsymbol{\theta}\| \le R} \|\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}\|^2. \tag{4.13}$$

The Lagrange multiplier method leads to an unconstrained problem

$$\widetilde{\boldsymbol{\theta}}_{G,\lambda} = \operatorname*{argmin}_{\boldsymbol{\theta}} \bigl\{ \|\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}\|^2 + \lambda \|G\boldsymbol{\theta}\|^2 \bigr\}.$$

A proper choice of λ ensures that the solution $\widetilde{\boldsymbol{\theta}}_{G,\lambda}$ belongs to $\mathcal{B}_G(R)$ and solves also the problem (4.13). So, the approach based on a Sobolev smoothness assumption, leads back to regularization and shrinkage.

4.7 * 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

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

where A is a given linear operator (matrix) and $\boldsymbol{f}^*$ is the target of analysis. With the obvious change of notation this problem can be put back in the general linear setup $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon}$. 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.

4.8 * 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: $\boldsymbol{Z} = \Lambda \boldsymbol{u} + \Lambda^{1/2}\boldsymbol{\xi}$ with a diagonal matrix $\Lambda = \operatorname{diag}\{\lambda_1, \ldots, \lambda_p\}$ and a homogeneous noise $\operatorname{Var}(\boldsymbol{\xi}) = \sigma^2 \boldsymbol{I}_p$; see Section 1.2.4. 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 $j = 1, \ldots, p$. The LSE or qMLE estimate of the spectral parameter $\boldsymbol{u}$ is given by

$$\widetilde{\boldsymbol{u}} = \Lambda^{-1}\boldsymbol{Z} = (\lambda_1^{-1} z_1, \ldots, \lambda_p^{-1} z_p)^\top.$$

Exercise 4.8.1. Consider the spectral representation $\boldsymbol{Z} = \Lambda \boldsymbol{u} + \Lambda^{1/2}\boldsymbol{\xi}$. The LSE $\widetilde{\boldsymbol{u}}$ reads as $\widetilde{\boldsymbol{u}} = \Lambda^{-1}\boldsymbol{Z}$.

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 $u_1, \ldots, u_m$ and to set all the remaining ones to zero. The corresponding estimate $\widetilde{\boldsymbol{u}}_m = (\widetilde{u}_{m,1}, \ldots, \widetilde{u}_{m,p})^\top$ reads as

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

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

Exercise 4.8.2. Consider the linear model $\boldsymbol{Y} = A\boldsymbol{f}^* + \boldsymbol{\varepsilon}$. Let U be an orthogonal transform in $I\!\!R^p$ providing $UA^\top AU^\top = \Lambda$ with a diagonal matrix Λ leading to the spectral representation for $\boldsymbol{Z} = UA\boldsymbol{Y}$. Write the corresponding spectral cut-off estimate $\widetilde{\boldsymbol{f}}_m$ for the original vector $\boldsymbol{f}^*$. 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 $\widetilde{\boldsymbol{u}}_{\boldsymbol{\alpha}}$ with $\widetilde{u}_{\boldsymbol{\alpha},j} = \alpha_j \lambda_j^{-1} z_j$. Here $\boldsymbol{\alpha}$ stands for the vector of coefficients α_j for $j = 1, \ldots, p$. A spectral cut-off method is a special case of this shrinkage with α_j equal to one or zero.

Exercise 4.8.3. Specify the spectral shrinkage $\widetilde{\boldsymbol{u}}_{\boldsymbol{\alpha}}$ with a given vector $\boldsymbol{\alpha}$ for the situation of Exercise 4.8.2.

The spectral cut-off method can be described as follows. Let $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$ be the intrinsic orthonormal basis of the problem composed of the standardized eigenvectors of $A^\top A$ and leading to the spectral representation $\boldsymbol{Z} = \Lambda \boldsymbol{u} + \Lambda^{1/2}\boldsymbol{\xi}$ with the target vector $\boldsymbol{u}$. In terms of the original target $\boldsymbol{f}^*$, one is looking for a solution or an estimate in the form $\boldsymbol{f} = \sum_j u_j \boldsymbol{\psi}_j$. The design orthogonality allows to estimate every coefficient u_j independently of the others using the empirical Fourier coefficient $\boldsymbol{\psi}_j^\top \boldsymbol{Y}$. Namely, $\widetilde{u}_j = \lambda_j^{-1} \boldsymbol{\psi}_j^\top \boldsymbol{Y} = \lambda_j^{-1} z_j$. The LSE procedure tries to recover $\boldsymbol{f}$ as the full sum $\widetilde{\boldsymbol{f}} = \sum_j \widetilde{u}_j \boldsymbol{\psi}_j$. The projection method suggests to cut this sum at the index m: $\widetilde{\boldsymbol{f}}_m = \sum_{j \le m} \widetilde{u}_j \boldsymbol{\psi}_j$, while the shrinkage procedure is based on downweighting the empirical coefficients $\widetilde{u}_j$: $\widetilde{\boldsymbol{f}}_{\boldsymbol{\alpha}} = \sum_j \alpha_j \widetilde{u}_j \boldsymbol{\psi}_j$.

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

Theorem 4.8.1. *Let* $\boldsymbol{Z} = \Lambda \boldsymbol{u}^* + \Lambda^{1/2}\boldsymbol{\xi}$ *with* $\operatorname{Var}(\boldsymbol{\xi}) = \sigma^2 I_p$. *It holds for the shrinkage estimate* $\widetilde{\boldsymbol{u}}_{\boldsymbol{\alpha}}$

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

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

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

as required.

Risk minimization leads to the oracle choice of the vector $\boldsymbol{\alpha}$ or

$$\boldsymbol{\alpha}^* = \operatorname*{argmin}_{\boldsymbol{\alpha}} \mathcal{R}(\widetilde{\boldsymbol{u}}_{\boldsymbol{\alpha}})$$

where the minimum is taken over the set of all admissible vectors $\boldsymbol{\alpha}$.

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

Exercise 4.8.4. The risk of the spectral cut-off estimate $\widetilde{\boldsymbol{u}}_m$ fulfills

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

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

4.9 * Roughness penalty and random design

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

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

(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

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

Quadraticity of L_G implies

$$\widetilde{\boldsymbol{\theta}}_G = \bigl(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2\bigr)^{-1} \boldsymbol{\Psi}\boldsymbol{Y},$$
$$\boldsymbol{\theta}_G^* = \Bigl\{\mathbb{E}\bigl(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2\bigr)\Bigr\}^{-1} \mathbb{E}(\boldsymbol{\Psi}\boldsymbol{Y}) = D_G^{-2} \mathbb{E}\bigl(\boldsymbol{\Psi}\boldsymbol{Y}\bigr)$$

with

$$D_G^2 \overset{\text{def}}{=} \sigma^{-2} I\!E \bigl(\mathbf{\Psi} \mathbf{\Psi}^{\top} + \sigma^2 G^2 \bigr).$$

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

$$\boldsymbol{M}_G^2 \overset{\text{def}}{=} I\!E \bigl(\mathbf{\Psi} \mathbf{\Psi}^{\top} + \sigma^2 G^2 \bigr) = \sigma^2 D_G^2 .$$

The key step in the analysis is the same as in the non-penalized case: to show that the empirical matrix $\mathbf{\Psi} \mathbf{\Psi}^{\top} + \sigma^2 G^2$ is close to its expectation. We already know that the empirical design matrix $\mathbf{\Psi} \mathbf{\Psi}^{\top}$ 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 $I\!E G^2$.

Lemma 4.9.1. *Let a set* $\Omega(\mathtt{x})$ *be such that* $I\!P\bigl(\Omega(\mathtt{x})\bigr) \ge 1 - \mathrm{e}^{-\mathtt{x}}$*, and it holds on* $\Omega(\mathtt{x})$ *for some* $\delta = \delta(\mathtt{x})$

$$\begin{aligned} \bigl\| \boldsymbol{M}^{-1} (\mathbf{\Psi} \mathbf{\Psi}^{\top}) \boldsymbol{M}^{-1} - I_p \bigr\|_{\mathrm{op}} &\le \delta, \\ \bigl\| (I\!E G^2)^{-1/2} \, G^2 \, (I\!E G^2)^{-1/2} - I_p \bigr\|_{\mathrm{op}} &\le \delta. \end{aligned} \tag{4.14}$$

Then

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

Proof. Conditions (4.14) imply

$$\begin{aligned} -\delta \boldsymbol{M}^2 &\le \mathbf{\Psi} \mathbf{\Psi}^{\top} - \boldsymbol{M}^2 \le \delta \boldsymbol{M}^2, \\ -\delta I\!E G^2 &\le G^2 - I\!E G^2 \le \delta I\!E G^2 \end{aligned}$$

and thus

$$-\delta \bigl(\boldsymbol{M}^2 + \sigma^2 I\!E G^2 \bigr) \le \mathbf{\Psi} \mathbf{\Psi}^{\top} + \sigma^2 G^2 - \bigl(\boldsymbol{M}^2 + \sigma^2 I\!E G^2 \bigr) \le \delta \bigl(\boldsymbol{M}^2 + \sigma^2 I\!E G^2 \bigr)$$

as required.

The standardized score $\boldsymbol{\xi}_G$ can be written as

$$\begin{aligned} \boldsymbol{\xi}_G &\overset{\text{def}}{=} D_G^{-1} \nabla L_G(\boldsymbol{\theta}_G^*) \\ &= \sigma^{-2} D_G^{-1} \mathbf{\Psi} \bigl(\boldsymbol{Y} - \mathbf{\Psi}^{\top} \boldsymbol{\theta}_G^* \bigr) - D_G^{-1} G^2 \, \boldsymbol{\theta}_G^* \\ &= \sigma^{-1} \boldsymbol{M}_G^{-1} \mathbf{\Psi} \bigl\{ \boldsymbol{Y} - I\!E(\boldsymbol{Y} \mid \mathbf{\Psi}) \bigr\} + \sigma^{-1} \boldsymbol{M}_G^{-1} \Bigl\{ \mathbf{\Psi} I\!E \bigl(\boldsymbol{Y} \mid \mathbf{\Psi} \bigr) - \bigl(\mathbf{\Psi} \mathbf{\Psi}^{\top} + \sigma^2 G^2 \bigr) \, \boldsymbol{\theta}_G^* \Bigr\} \\ &= \sigma^{-1} \boldsymbol{M}_G^{-1} \mathbf{\Psi} \boldsymbol{\varepsilon} + \delta_G(\mathbf{\Psi}), \end{aligned} \tag{4.15}$$

where

$$\begin{aligned} \boldsymbol{\varepsilon} &\stackrel{\mathrm{def}}{=} \boldsymbol{Y} - \mathbb{E}\bigl(\boldsymbol{Y} \bigm| \boldsymbol{\Psi}\bigr) \\ \delta_G(\boldsymbol{\Psi}) &\stackrel{\mathrm{def}}{=} \sigma^{-1} \boldsymbol{M}_G^{-1} \Bigl\{ \boldsymbol{\Psi} \mathbb{E}\bigl(\boldsymbol{Y} \bigm| \boldsymbol{\Psi}\bigr) - \bigl(\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top} + \sigma^2 G^2\bigr) \boldsymbol{\theta}_G^{*} \Bigr\}. \end{aligned} \tag{4.16}$$

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

$$\mathbb{E}\delta_G(\boldsymbol{\Psi}) = \sigma^{-1} \boldsymbol{M}_G^{-1} \Bigl\{ \mathbb{E}\bigl(\boldsymbol{\Psi}\boldsymbol{Y}\bigr) - \mathbb{E}\bigl(\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top} + \sigma^2 G^2\bigr) \boldsymbol{\theta}_G^{*} \Bigr\} = 0.$$

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

Theorem 4.9.1. *Consider the model* (3.1) *and suppose*

$$\bigl\| \boldsymbol{M}_G^{-1} \bigl(\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top} + \sigma^2 G^2\bigr) \boldsymbol{M}_G^{-1} - I_p \bigr\|_{\mathrm{op}} \le \delta \tag{4.17}$$

for some $\delta < 1/2$ *on a dominating set* $\Omega(\mathtt{x})$. *Then the penalized MLE* $\widetilde{\boldsymbol{\theta}}_G$ *fulfills on* $\Omega(\mathtt{x})$ *for* $\boldsymbol{\xi}_G$ *from* (4.15)

$$\bigl\| D_G\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^{*}\bigr) - \boldsymbol{\xi}_G \bigr\| \le \frac{\delta}{1-\delta} \|\boldsymbol{\xi}_G\|, \tag{4.18}$$

$$\bigl| 2L_G(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^{*}) - \|\boldsymbol{\xi}_G\|^2 \bigr| \le \frac{\delta}{1-\delta} \|\boldsymbol{\xi}_G\|^2. \tag{4.19}$$

Proof. The bound (4.17) also implies

$$\bigl\| \boldsymbol{M}_G \, (\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top} + \sigma^2 G^2)^{-1} \boldsymbol{M}_G - I_p \bigr\|_{\mathrm{op}} \le \frac{\delta}{1-\delta}. \tag{4.20}$$

By using quadraticity of $L_G(\boldsymbol{\theta})$, one obtains (see Theorem 1.5.1 in Section 1.5)

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

Further, the model equation (3.1) and the decomposition $\boldsymbol{Y} = \boldsymbol{\varepsilon} + \mathbb{E}\bigl(\boldsymbol{Y} \bigm| \boldsymbol{\Psi}\bigr)$ imply for $\delta_G(\boldsymbol{\Psi})$ from (4.16) and $\boldsymbol{\xi}_G$ from (4.15)

$$\begin{aligned} D_G\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^{*}\bigr) &= D_G \, (\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top} + \sigma^2 G^2)^{-1} \Bigl\{ \boldsymbol{\Psi}\boldsymbol{\varepsilon} + \boldsymbol{\Psi}\mathbb{E}(\boldsymbol{Y} \bigm| \boldsymbol{\Psi}) - (\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top} + \sigma^2 G^2) \boldsymbol{\theta}_G^{*} \Bigr\} \\ &= \boldsymbol{M}_G \, (\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top} + \sigma^2 G^2)^{-1} \boldsymbol{M}_G \, \boldsymbol{\xi}_G = A_G \, \boldsymbol{\xi}_G \end{aligned}$$

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

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

so that (4.18) follows from (4.20). Similarly

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

yielding (4.19) on $\Omega(\mathtt{x})$.

Similarly to Sections 3.4 and 3.5, one can establish a deviation bound for the norm $\|\boldsymbol{\xi}_G\|$ entering in the error bound. Definition (4.15) can be rewritten as

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

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

$$\begin{aligned} \operatorname{Var}\bigl(\boldsymbol{\xi}_{G,0} \bigm| \boldsymbol{\Psi} \bigr) &= \operatorname{Var}\bigl(\sigma^{-1} \boldsymbol{M}_G^{-1} \boldsymbol{\Psi} \boldsymbol{\varepsilon} \bigm| \boldsymbol{\Psi} \bigr) = \sigma^{-2} \boldsymbol{M}_G^{-1} (\boldsymbol{\Psi} \Sigma \boldsymbol{\Psi}^\top) \boldsymbol{M}_G^{-1}, \\ \operatorname{Var}\bigl(\boldsymbol{\xi}_{G,0} \bigr) &= B_{G,0} \stackrel{\mathrm{def}}{=} \sigma^{-2} \boldsymbol{M}_G^{-1} \, I\!\!E (\boldsymbol{\Psi} \Sigma \boldsymbol{\Psi}^\top) \, \boldsymbol{M}_G^{-1}. \end{aligned}$$

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 $\Omega_1(\mathtt{x})$ of dominating measure $1 - 2\mathrm{e}^{-\mathtt{x}}$ on which the conditional expectation $\operatorname{Var}\bigl(\boldsymbol{\xi}_{G,0} \bigm| \boldsymbol{\Psi} \bigr)$ is close to the unconditional one up the value $\delta_1 = \delta_1(\mathtt{x})$. Then we can apply the deviation bound of Theorem C.2.2 conditionally on $\boldsymbol{\Psi}$ on the set $\Omega_1(\mathtt{x})$ yielding the bound for $\|\boldsymbol{\xi}_{G,0}\|$ similar to (3.25) with $B_{G,0}$ in place of B:

$$\|\boldsymbol{\xi}_{G,0}\| \le \sqrt{\operatorname{tr}(B_{G,0})} + \sqrt{2\mathtt{x}}$$

with a high probability.

Theorem 4.9.2. *Suppose* (4.17) *on a set* $\Omega(\mathtt{x})$. *Let the error vector* $\boldsymbol{\varepsilon}$ *satisfy the exponential moment conditions ... Then*

$$I\!\!P \Bigl\{ \|\boldsymbol{\xi}_{G,0}\| \ge (1 + \delta_1) \, z(B_{G,0}, \mathtt{x}) \Bigr\} \le 2\mathrm{e}^{-\mathtt{x}}.$$

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 $\boldsymbol{b}_G \in I\!\!R^n$

$$\boldsymbol{b}_G \stackrel{\mathrm{def}}{=} I\!\!E \bigl(\boldsymbol{Y} \bigm| \boldsymbol{\Psi} \bigr) - \boldsymbol{\Psi}^\top \boldsymbol{\theta}_G^*.$$

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 $b_i = I\!\!E\bigl(Y_i \bigm| \Psi_i\bigr) - \Psi_i^\top \boldsymbol{\theta}_G^*$ does not exceed some small value b_∞.

Theorem 4.9.3. *Suppose that* $\bigl\| I\!\!E\bigl(\boldsymbol{Y} \bigm| \boldsymbol{\Psi}\bigr) - \boldsymbol{\Psi}^\top \boldsymbol{\theta}_G^* \bigr\|_\infty \le b_\infty$. *Then*

$$I\!\!P\Bigl\{ \|\delta_G(\boldsymbol{\Psi})\| \ge \sigma\, b_\infty\, z(B_G, \mathtt{x}) \Bigr\} \le 2\mathrm{e}^{-\mathtt{x}}$$

with $B_G \overset{\text{def}}{=} \boldsymbol{M}_G^{-1}\, I\!\!E(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)\, \boldsymbol{M}_G^{-1}$.

Proof. Denote $f_i = I\!\!E\bigl(Y_i \bigm| \Psi_i\bigr)$ and $b_i = f_i - \Psi_i^\top \boldsymbol{\theta}_G^*$. We use the representation

$$\begin{aligned}
\delta_G(\boldsymbol{\Psi}) &= D_G^{-1}\bigl\{ \boldsymbol{\Psi} I\!\!E\bigl(\boldsymbol{Y} \bigm| \boldsymbol{\Psi}\bigr) - (\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2 G^2)\boldsymbol{\theta}_G^* \bigr\} \\
&= D_G^{-1} \sum_i \Psi_i \bigl(f_i - \Psi_i^\top \boldsymbol{\theta}_G^*\bigr) - \sigma^2 D_G^{-1} G^2 \boldsymbol{\theta}_G^* \\
&= D_G^{-1} \sum_i \Psi_i b_i - \sigma^2 D_G^{-1} G^2 \boldsymbol{\theta}_G^* = D_G^{-1} \sum_i \bigl\{ \Psi_i b_i - I\!\!E(\Psi_i b_i) \bigr\}.
\end{aligned}$$

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

$$\operatorname{Var}\bigl\{\delta_G(\boldsymbol{\Psi})\bigr\} = D_G^{-1} \sum_i \operatorname{Var}\bigl(\Psi_i b_i\bigr)\, D_G^{-1} \le \sigma^2\, b_\infty^2\, \boldsymbol{M}_G^{-1} \sum_i I\!\!E(\Psi_i \Psi_i^\top)\, \boldsymbol{M}_G^{-1} = \sigma^2\, b_\infty^2\, B_G.$$

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

5

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.

5.1 Projection estimation and the model choice problem

In this section we consider the linear model $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ 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 submodes corresponding to the reduced subset of parameters.

We associate the rows of the design matrix Ψ with basis vectors in $I\!R^{n}$. By Ψ_{m} we denote a sub matrix of Ψ composed of the first m rows $\boldsymbol{\psi}_{1}, \ldots, \boldsymbol{\psi}_{m}$. It corresponds to the reduced regression model

$$\boldsymbol{Y} = \Psi_{m}^{\top}\boldsymbol{\theta}_{m} + \boldsymbol{\varepsilon}$$

with the parameter $\boldsymbol{\theta}_{m}$ from $I\!R^{m}$. The corresponding estimate $\widetilde{\boldsymbol{\theta}}_{m}$ and the predictor $\widetilde{\boldsymbol{f}}_{m}$ read as

$$\begin{aligned}
\widetilde{\boldsymbol{\theta}}_{m} &= \bigl(\Psi_{m}\Psi_{m}^{\top}\bigr)^{-1}\Psi_{m}\boldsymbol{Y}, \\
\widetilde{\boldsymbol{f}}_{m} &= \Psi_{m}\bigl(\Psi_{m}\Psi_{m}^{\top}\bigr)^{-1}\Psi_{m}\boldsymbol{Y} = \Pi_{m}\boldsymbol{Y}
\end{aligned}$$

where Π_{m} is a projector in $I\!R^{n}$ on the subspace spanned by the basis vectors $\boldsymbol{\psi}_{1}, \ldots, \boldsymbol{\psi}_{m}$.

In the case of an orthonormal design, one just considers the first m empirical coefficients $z_{1}, \ldots, z_{m}$ and drop the others. The corresponding parameter estimate $\widetilde{\boldsymbol{\theta}}_{m}$ reads as

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

with $z_j = \boldsymbol{\psi}_j^\top \boldsymbol{Y}$. The response vector $\boldsymbol{f}^* = I\!\!E \boldsymbol{Y}$ is estimated by $\Psi^\top \widetilde{\boldsymbol{\theta}}_m$ leading to the representation

$$\widetilde{\boldsymbol{f}}_m = z_1 \boldsymbol{\psi}_1 + \ldots + z_m \boldsymbol{\psi}_m .$$

In other words, $\widetilde{\boldsymbol{f}}_m$ is just a projection of the observed vector $\boldsymbol{Y}$ onto the subspace L_m spanned by the first m basis vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_m$: $\mathrm{L}_m = \bigl\langle \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_m \bigr\rangle$. This explains the name of the method. Clearly one can study the properties of $\widetilde{\boldsymbol{\theta}}_m$ or $\widetilde{\boldsymbol{f}}_m$ 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 $\mathcal{R}(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*) = I\!\!E \|\widetilde{\boldsymbol{f}}_m - \boldsymbol{f}^*\|^2$. Below we focus on the case of a homogeneous noise with $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_p$. An extension to the colored noise is possible. Recall that $\widetilde{\boldsymbol{f}}_m$ effectively estimates the vector $\boldsymbol{f}_m = \Pi_m \boldsymbol{f}^*$, where Π_m is the projector on L_m; see Section 1.3.3. Moreover, the quadratic risk $\mathcal{R}(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*)$ can be decomposed as

$$\mathcal{R}(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*) = \|\boldsymbol{f}^* - \Pi_m \boldsymbol{f}^*\|^2 + \sigma^2 m. \tag{5.1}$$

Obviously the squared bias $\|\boldsymbol{f}^* - \Pi_m \boldsymbol{f}^*\|^2$ 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 $\mathcal{R}(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*)$ over all possible m:

$$m^* \stackrel{\text{def}}{=} \operatorname*{argmin}_m \mathcal{R}(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*) = \operatorname*{argmin}_m \bigl\{ \|\boldsymbol{f}^* - \Pi_m \boldsymbol{f}^*\|^2 + \sigma^2 m \bigr\}. \tag{5.2}$$

Unfortunately this choice requires some information about the bias $\|\boldsymbol{f}^* - \Pi_m \boldsymbol{f}^*\|$ which depends on the unknown vector $\boldsymbol{f}^*$. 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 $I\!\!E \|\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\|^2$. Indeed, the basis orthogonality $\Psi \Psi^\top = \boldsymbol{I}_p$ implies for $\boldsymbol{f}^* = \Psi^\top \boldsymbol{\theta}^*$

$$\|\widetilde{\boldsymbol{f}}_m - \boldsymbol{f}^*\|^2 = \|\Psi^\top \widetilde{\boldsymbol{\theta}}_m - \Psi^\top \boldsymbol{\theta}^*\|^2 = \|\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\|^2$$

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 $\widehat{\boldsymbol{\theta}}$ whose risk is close to the risk of the oracle $\widetilde{\boldsymbol{\theta}}_{m^{*}}$. The problem of *model selection* mainly focuses choosing a proper model $\widehat{m}$ 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: $\mathcal{R}_{m} \equiv \mathtt{C}$. Then any model choice yields the same risk and one is 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 $\boldsymbol{\theta}^{*}$ and of the oracle model m^{*}.

5.2 Bias-variance trade-off under smoothness assumptions. Rate of estimation for smoothness classes

This section discusses the oracle choice of the complexity parameter m for a class of model with fixed smoothness properties. Remind that the risk of projection estimation $\widetilde{\boldsymbol{f}}_{m}$ is given by

$$\mathcal{R}(\widetilde{\boldsymbol{f}}_{m}) = \|\Pi_{m}\boldsymbol{f}^{*} - \boldsymbol{f}^{*}\|^{2} + \sigma^{2}m.$$

Risk minimization over m yields the “oracle” choice. However, it cannot be implemented in practice because the bias term $\|\Pi_{m}\boldsymbol{f}^{*} - \boldsymbol{f}^{*}\|^{2}$ depends on the unknown function $\boldsymbol{f}^{*}$. A popular method of treating this problem is based on the so called *smoothness* assumption.

5.2.1 Smoothness classes

One assumes that the function f belongs to a *smoothness class* $\mathcal{F}$ and the basis system $\{\psi_{m}\}$ ensures a good quality of approximation of any function f from this class by its projections on the subspace L_{m} spanned by $\boldsymbol{\psi}_{1},\ldots,\boldsymbol{\psi}_{m}$. More precisely, one uses that

$$\sup_{\boldsymbol{f}\in\mathcal{F}} \frac{1}{n}\|\Pi_{m}\boldsymbol{f} - \boldsymbol{f}\|^{2} \le b^{2}(m)$$

for a fixed monotonously decreasing sequence $b(m)$. Note that this bias depends not only on the properties of the function class $\mathcal{F}$ but also on the design $X_{1},\ldots,X_{n}$. Usually one supposes that the design is either random from some distribution with a continuous density $h_X(x)$ or deterministic *regular*. Here we only discuss the later case. However, these two approaches are closely linked. Namely, a regular design is usually viewed as an i.i.d. sample from a continuous design distribution with a density $h_X(x)$. Design regulatrity means that any sum of the form $n^{-1}\sum_{i=1}^n f(X_i)$ can be approximated by an integral $\int \psi_j(x)\, f(x)\, h_X(x)\, dx$ uniformly over functions f from the class of $\mathcal{F}$ with a very good accuracy, usually of order $1/n$:

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \int f(x)\, h_X(x)\, dx \right| \le \frac{\mathtt{C}}{n}\,. \tag{5.3}$$

Such a condition allows to replace the sum by the integral in the condition on the accuracy of approximation:

$$\sup_{f\in\mathcal{F}} \bigl\langle \mathrm{P}_m f - f \bigr\rangle \le b^2(m)\,.$$

Here $\langle f\rangle \overset{\mathrm{def}}{=} \langle f, f\rangle$ and P_m is an integral version of the projector on the subspace L_m corresponding to the scalar product

$$\langle f, g\rangle \overset{\mathrm{def}}{=} \int f(x)\, g(x)\, h_X(x)\, dx\,.$$

This projector P_m is given by quadratic programming

$$\mathrm{P}_m f \overset{\mathrm{def}}{=} \inf_{f_m\in \mathrm{L}_m} \langle f - f_m\rangle.$$

Exercise 5.2.1. Show that P_m is a projector in the space of squared integrable functions f with $\langle f\rangle < \infty$ on the subspace $\mathrm{L}_m = \langle \psi_1,\ldots,\psi_m\rangle$.

Usually smoothness classes $\mathcal{F}$ are defined via some conditions on the derivative of the function f. A *Hölder* smoothness class $\mathcal{H}(s,R)$ of order $s = k+\alpha$ for $k\ge 0$ and $0<\alpha\le 1$ is defined by the condition

$$\bigl| f^{(k)}(x) - f^{(k)}(x') \bigr| \le R\bigl| x - x'\bigr|^{\alpha}.$$

In particular, for $s=1$, we get a Lipschitz class with

$$\bigl| f(x) - f(x')\bigr| \le R\bigl| x - x'\bigr|.$$

A *Sobolev* smoothness class $\mathcal{S}(s,R)$ is defined via the squared integral (energy) of the s-derivative $f^{(s)}$ of the function f:

$$\int \bigl| f^{(s)}(x)\bigr|^2 h_X(x)\, dx \le R^2. \tag{5.4}$$

More general definitions include *Nikolskii*, *Besov*, and other classes.

In is well known from the theory of approximation that a proper choice of the basis functions yields that the bias $b^2(m)$ decreases polynomially with the index m:

$$b^2(m) \overset{\text{def}}{=} \sup_{f \in \mathcal{F}(s,R)} \langle \mathrm{P}_m f - f \rangle \lesssim \frac{R^2}{m^{2s}} . \tag{5.5}$$

A symbol $\lesssim$ means that this bound is fulfilled up to a fixed multiplicative constant $\mathtt{C}$ or, equivalently, $b(m) \le \mathtt{C} R m^{-s}$. Standard choices of the functional basis $\psi_1, \psi_2, \ldots$ for such smoothness classes are orthogonal polynomials, piecewise polynomials, splines, wavelets, trigonometric functions.

5.2.2 Orthogonal basis and Sobolev bodies

Suppose that $\psi_1, \psi_2, \ldots$ is a functional basis. In addition, assume orthonormality of the functions ψ_j:

$$\langle \psi_j, \psi_{j'} \rangle = \delta_{j,j'} .$$

Then the coefficients of the expansion

$$f = \sum_{j=1}^{p} \theta_j \psi_j$$

can be obtained by scalar product

$$\theta_j = \langle f, \psi_j \rangle$$

Moreover, projection P_m is given by a partial sum

$$\mathrm{P}_m f = \sum_{j=1}^{m} \theta_j \psi_j$$

and by the Parseval identity

$$\langle f - \mathrm{P}_m \rangle = \langle \sum_{j=m+1}^{p} \theta_j \psi_j \rangle = \sum_{j=m+1}^{p} \theta_j^2 .$$

5.2.3 Fourier transform and Sobolev bodies

In particular, the prominent *Fourier* basis on the interval $[0,1]$ is built by the functions $\cos(2\pi m x)$ and $\sin(2\pi m x)$ for $m = 0, 1, 2, \ldots$. It ensures the smoothness condition (5.5) for Hölder and Sobolev classes of periodic functions on the interval $[0,1]$. This basis is frequently used in signal processing and time series analysis. The Fourier decomposition can be written as

$$f(x) = \sum_{j=0}^{\infty} \bigl\{ \theta_{2j} \cos(2\pi jx) + \theta_{2j+1} \sin(2\pi jx) \bigr\}$$

The main benefit of using the Fourier basis is that orthogonal w.r.t. to the usual scalar product in $L_2[0,1]$ and at the same time w.r.t. the equidistant design: for $m \neq m'$

$$\bigl\langle \psi_m, \psi_{m'} \bigr\rangle = 0, \tag{5.6}$$

$$\sum_{i=1}^{n} \psi_m(i/n) \psi_{m'}(i/n) = 0. \tag{5.7}$$

This property is very useful and allows to compute the Fourier transform very quickly by the numerical procedure known as *fast Fourier transform.*

Exercise 5.2.2. Check (5.6) and (5.7).

The cosine basis is built by the functions $\cos(\pi mx)$, $m \geq 0$. It is used for functions which are not necessarily periodic.

Exercise 5.2.3. Check orthogonality of the cosine basis for the usual scalar product and for the equidistant design $X_i = i/n$.

It is known from the approximation theory that the use of Fourier basis ensures the polynomial decay $b(m) \leq \mathtt{C} R\, m^{-s}$ over the class of periodic functions from Hölder class $\mathcal{H}(s,R)$ and the Sobolev class $\mathcal{S}(s,R)$. The use of the cosine basis allows to drop the periodicity condition.

5.2.4 Accuracy of approximation by projection

Now we suppose that the underlying function $\boldsymbol{f}^*$ belongs to a smoothness class $\mathcal{F}(s,R)$. This allows to state a prescribed accuracy of approximation of this function by a considered basis $\bigl\{\psi_j\bigr\}$ in the L_2-sense. However, for the risk optimization, we need a bound for the squared bias corresponding to the design $X_1, \ldots, X_n$. Suppose that the design is regular in the sense (5.3). Suppose also that the error of numerical approximation in (5.3) is smaller than the squared bias in (5.5). Then one can easily extend the bound $\bigl\langle \mathrm{P}_m f - f \bigr\rangle$ on the value $n^{-1} \| \Pi_m \boldsymbol{f}^* - \boldsymbol{f}^* \|^2$.

Theorem 5.2.1. *Let $X_1, \ldots, X_n$ be a regular design in the sense (5.3). Let also f belong to a smoothness class $\mathcal{F}$ yielding the error of approximation $b(m)$ in (5.5). Then it holds*

$$\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \| \Pi_m \boldsymbol{f} - \boldsymbol{f} \|^2 \ \leq b^2(m) + \frac{\mathtt{C}}{n}.$$

Proof. The function $\mathrm{P}_m f$ belongs to the linear subspace $\mathrm{L}_m = \bigl\langle \psi_1, \ldots, \psi_m \bigr\rangle$:

$$\mathrm{P}_m f(x) = c_1 \psi_1(x) + \ldots + c_m \psi_m(x) .$$

This also applies at any design point X_i . In vector form one can write

$$\mathrm{P}_m \boldsymbol{f} = c_1 \boldsymbol{\psi}_1 + \ldots + c_m \boldsymbol{\psi}_m .$$

Therefore,

$$\|\Pi_m \boldsymbol{f} - \boldsymbol{f}\|^2 \leq \|\mathrm{P}_m \boldsymbol{f} - \boldsymbol{f}\|^2.$$

Now the result follows from (5.3) applied to the function $\bigl(\mathrm{P}_m f - f\bigr)^2$.

5.2.5 Choice of the model by smoothness assumptions

Now we come back to the problem of choosing the order of projection estimator $\widetilde{\boldsymbol{f}}_m$. The approach is based on the assumption of smoothness: we suppose that the underlying function f belongs to a smoothness class $\mathcal{F}(s, R)$, this allows to bound the error of approximation $n^{-1}\|\Pi_m \boldsymbol{f} - \boldsymbol{f}\|^2$ using Theorem 5.2.1. Ignoring the error of numerical integration, we write the condition as

$$\sup_{\boldsymbol{f} \in \mathcal{F}} n^{-1} \|\Pi_m \boldsymbol{f} - \boldsymbol{f}\|^2 \leq b^2(m).$$

Now instead of minimizing the exact squared risk $\mathcal{R}(\widetilde{\boldsymbol{f}}_m) = \|\Pi_m \boldsymbol{f} - \boldsymbol{f}\|^2 + \sigma^2 m$ we minimize its upper bound:

$$m^* \overset{\text{def}}{=} \operatorname*{argmin}_m \bigl\{ \|\Pi_m \boldsymbol{f} - \boldsymbol{f}\|^2 + \sigma^2 m \bigr\} = \operatorname*{argmin}_m \Bigl\{ b^2(m) + \frac{\sigma^2 m}{n} \Bigr\}. \tag{5.8}$$

This definition enables us to control the risk uniformly over the smoothness class $\mathcal{F}$.

Theorem 5.2.2. *Let $\mathcal{F}$ be a smoothness function class with the corresponding bias function $b(m)$:*

$$\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \|\Pi_m \boldsymbol{f} - \boldsymbol{f}\|^2 \leq b^2(m) .$$

Then the choice m^* by (5.8) yields the optimal uniform risk bound for the class $\mathcal{F}$:

$$\sup_{\boldsymbol{f} \in \mathcal{F}} I\!\!E \|\widetilde{\boldsymbol{f}}_{m^*} - \boldsymbol{f}\|^2 \leq n b^2(m^*) + \sigma^2 m^* = \min_m \bigl\{ n b^2(m) + \sigma^2 m \bigr\}.$$

For the most typical situation of polynomial decay of the bias function $b(m)$, one can get a closed form solution for m^* .

Theorem 5.2.3. *Let $\mathcal{F}(s,R)$ be a smoothness function class with the corresponding bias function $b(m) = Rm^{-s}$. Then*

$$m^{*} = \operatorname*{argmin}_{m} \bigl\{ nR^2 m^{-2s} + \sigma^2 m \bigr\} \asymp \Bigl(\frac{nR^2}{\sigma^2} \Bigr)^{1/(2s+1)} . \tag{5.9}$$

For the corresponding risk $\mathcal{R}(\widetilde{\boldsymbol{f}}_{m^*})$ holds

$$\sup_{\boldsymbol{f} \in \mathcal{F}(s,R)} \frac{1}{n} I\!E \bigl\| \widetilde{\boldsymbol{f}}_{m^*} - \boldsymbol{f} \bigr\|^2 \lesssim R^{2/(2s+1)} \Bigl(\frac{\sigma^2}{n} \Bigr)^{2s/(2s+1)} .$$

The choice of m^* due to (5.9) does not depend on a particular model function $\boldsymbol{f}$. However, it is still an “oracle” choice: we have to know the parameter of the smoothness class $\mathcal{F}(s,R)$.

Exercise 5.2.4. Check that for any m and n there exists a combination smoothness class parameters (s,R) such that $m^*(s,R) = m$.

The result of Theorem 5.2.3 provides a theoretical upper bound for the risk, however, it is not practical. It does not allow to select the parameter m in practical situations.

5.3 Smoothness constraint and penalization

Another way of using the smoothness condition on the underlying regression function f is to consider a restricted or penalized risk minimization over a function class. Namely, one can define a constraint nonparametric maximum likelihood estimator

$$\widetilde{\boldsymbol{f}}_{s,R} \stackrel{\text{def}}{=} \operatorname*{argmax}_{f \in \mathcal{F}(s,R)} L(f) = \operatorname*{argmin}_{f \in \mathcal{F}(s,R)} \frac{1}{\sigma^2} \bigl\| \boldsymbol{Y} - \boldsymbol{f} \bigr\|^2$$

Here we identify a function $f \in \mathcal{F}(s,R)$ and the corresponding vector $\boldsymbol{f} \in I\!R^n$ of its values at design points X_i. A special case which we consider is given by the roughness constraint (5.4) on the proper derivative of the function f.

Unfortunately the constraint optimization is hard to implement. The Lagrangian multiplier approach allows to replace this problem by the penalized optimization

$$\widetilde{f}_{s,\lambda} \stackrel{\text{def}}{=} \operatorname*{argmin}_{f} \Biggl\{ \frac{1}{n\sigma^2} \sum_{i=1}^{n} |Y_i - f(X_i)|^2 + \lambda \bigl\langle f^{(s)} \bigr\rangle \Biggr\} .$$

In the univariate case with $X_i \in I\!R^1$, any solution to this optimization problem is a natural spline of order s with knots at points X_i, that is, the function $f^{(s)}(x)$ is constant at each piece between any two neighbor design points.

This problem can be parametrized using an orthonormal basis $\{\psi_j\}$ with

$$\langle \psi_j, \psi_{j'} \rangle = b_{j,j'} .$$

In the matrix form with the $p \times p$ matrix $B = (b_{j,j'})$

$$\langle \Psi, \Psi \rangle = B. \tag{5.10}$$

We also suppose that each ψ_j is smooth (s times differentiable) function such that

$$\psi_j^{(s)} = \sum_{\ell=1}^{p} a_{j\ell} \psi_\ell .$$

In the matrix form, this relation reads

$$\Psi^{(s)} = A\Psi$$

for the $p \times p$ matrix A with entries $a_{j\ell}$. The derivative $f^{(s)}$ of a function $f = \sum_{j \le p} \theta_j \psi_j$ can be represented as

$$f^{(s)} = \sum_{j \le p} \theta_j \psi_j^{(s)} = \boldsymbol{\theta}^\top \Psi^{(s)} = \boldsymbol{\theta}^\top A\Psi$$

The use of (5.10) implies

$$\langle f^{(s)} \rangle = \langle \boldsymbol{\theta}^\top A\Psi \rangle = \boldsymbol{\theta}^\top A \langle \Psi, \Psi \rangle A^\top \boldsymbol{\theta} = \boldsymbol{\theta}^\top ABA^\top \boldsymbol{\theta}.$$

Therefore, the penalty term is quadratic in the parameter $\boldsymbol{\theta}$:

$$\langle f^{(s)} \rangle = \boldsymbol{\theta}^\top ABA^\top \boldsymbol{\theta} = \| G\boldsymbol{\theta} \|^2$$

for $G^2 = ABA^\top$. Now the penalized optimization problem can be restated as quadratic programming:

$$\widetilde{\boldsymbol{\theta}}_{G,\lambda} = \operatorname*{arginf}_{\boldsymbol{\theta} \in I\!R^p} \left\{ \frac{1}{2\sigma^2} \| \boldsymbol{Y} - \boldsymbol{\Psi}^\top \boldsymbol{\theta} \|^2 + \lambda \| G\boldsymbol{\theta} \|^2 \right\} .$$

One often uses a double orthogonal basis which simultaneously fulfills two orthogonality conditions:

$$\boldsymbol{\Psi}\boldsymbol{\Psi}^\top = I_p , \qquad G^2 = \operatorname{diag}\bigl(g_1^2, \ldots, g_p^2 \bigr).$$

Such a construction is often called Demmler-Reisch basis. It is especially useful for linear inverse problems. Under the double orthogonality, the solution $\widetilde{\boldsymbol{\theta}}_{G,\lambda}$ is easy to compute:

$$\widetilde{\boldsymbol{\theta}}_{G,\lambda} = \bigl(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \lambda\sigma^2 G^2 \bigr)^{-1} \boldsymbol{\Psi}\boldsymbol{Y} = \bigl(I_p + \lambda\sigma^2 G^2 \bigr)^{-1} \boldsymbol{Z} = \Bigl(\frac{z_j}{1 + \lambda\sigma^2 g_j^2} \Bigr)_{j=1,\ldots,p}$$

with the p-vector $\boldsymbol{Z} = \boldsymbol{\Psi}\boldsymbol{Y}$, whose entries are z_j.

5.4 Unbiased risk estimation in projection estimation

The "oracle" choice m^* cannot be implemented because the bias term $\|\boldsymbol{f}^* - \Pi_m \boldsymbol{f}^*\|^2$ depends on the target object $\boldsymbol{f}^*$. 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) $\|\boldsymbol{Y} - \widetilde{\boldsymbol{f}}_m\|^2$ in which we replace the function $\boldsymbol{f}^*$ by the data $\boldsymbol{Y}$. Unfortunately, this rule leads to the trivial solution

$$\widehat{m} = \operatorname*{argmin}_m \|\boldsymbol{Y} - \widetilde{\boldsymbol{f}}_m\|^2 = p.$$

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

Lemma 5.4.1. *Consider the projection estimator* $\widetilde{\boldsymbol{f}}_m = \Pi_m \boldsymbol{Y}$. *For two different values* $m' > m$, *the following statements hold:*

- $\Pi_{m',m} \stackrel{\text{def}}{=} \Pi_{m'} - \Pi_m$ *is a projector in* $I\!\!R^n$.
- *If* Ψ *is orthogonal then* $\Pi_{m',m}$ *projects onto subspace generated by* $\boldsymbol{\psi}_{m+1}, \ldots, \boldsymbol{\psi}_{m'}$.
- *The next identity is fulfilled:*

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

Proof. It obviously holds

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

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 $\widetilde{\boldsymbol{f}}_m = \boldsymbol{Y}$, that is, the estimate coincides with the data. This is formally correct but the corresponding squared risk is equal to $\sigma^2 n$ 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 5.4.2. *It holds under homogeneous errors* $\boldsymbol{\varepsilon}$:

- *For each* m

$$I\!\!E \|\boldsymbol{Y} - \Pi_m \boldsymbol{Y}\|^2 = \|\boldsymbol{f}^* - \Pi_m \boldsymbol{f}^*\|^2 + \sigma^2 (n - m). \tag{5.11}$$

- *For any* $m' > m$

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

Proof. Obvious.

The first term in the statement (5.11) 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

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

Then it holds

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

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

$$\widehat{m} \overset{\mathrm{def}}{=} \operatorname*{argmin}_m \widetilde{\mathcal{R}}_m = \operatorname*{argmin}_m \bigl(\|\boldsymbol{Y} - \Pi_m \boldsymbol{Y}\|^2 + 2\sigma^2 m\bigr). \tag{5.12}$$

This rule is known as Akaiki information criteria (AIC) and it is very popular in practical applications. It suggests to balance the data fit measured by $\|\boldsymbol{Y} - \Pi_m \boldsymbol{Y}\|^2$ 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 $\|\boldsymbol{Y} - \Pi_m \boldsymbol{Y}\|^2$.

Exercise 5.4.1. Consider the projection estimator $\widetilde{\boldsymbol{f}}_m = \Pi_m \boldsymbol{Y}$ for the model $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$ with $\Pi_m = \Psi_m^\top \bigl(\Psi_m \Psi_m^\top\bigr)^{-1} \Psi_m$. For two different values $m' > m$:

- Check that $\Pi_{m',m} \overset{\mathrm{def}}{=} \Pi_{m'} - \Pi_m$ 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

$$\begin{aligned}\|\boldsymbol{Y} - \Pi_m \boldsymbol{Y}\|^2 - \|\boldsymbol{Y} - \Pi_{m'} \boldsymbol{Y}\|^2 &= \|\widetilde{\boldsymbol{f}}_{m'} - \widetilde{\boldsymbol{f}}_m\|^2 = \|\Pi_{m',m} \boldsymbol{f}^* + \Pi_{m',m} \boldsymbol{\varepsilon}\|^2, \\ \|\widetilde{\boldsymbol{f}}_{m'} - \boldsymbol{f}^*\|^2 - \|\widetilde{\boldsymbol{f}}_m - \boldsymbol{f}^*\|^2 &= -\|\Pi_{m',m} \boldsymbol{f}^*\|^2 + \|\Pi_{m',m} \boldsymbol{\varepsilon}\|^2.\end{aligned}$$

- compute $\mathbb{E}\|\widetilde{\boldsymbol{f}}_{m'} - \widetilde{\boldsymbol{f}}_m\|^2$ and $\mathbb{E}\bigl[\|\widetilde{\boldsymbol{f}}_{m'} - \boldsymbol{f}^*\|^2 - \|\widetilde{\boldsymbol{f}}_m - \boldsymbol{f}^*\|^2\bigr]$.

5.4.1 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 $\widehat{m}$ 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 $\bigl\{\widehat{m} > m^*\bigr\}$.

The definition of m^* (5.2) implies for $m > m^*$ with $\mathcal{R}_m \stackrel{\text{def}}{=} \mathcal{R}(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*)$

$$\begin{aligned} \mathcal{R}_m - \mathcal{R}_{m^*} &= \|\boldsymbol{f}^* - \Pi_m \boldsymbol{f}^*\|^2 - \|\boldsymbol{f}^* - \Pi_{m^*}\boldsymbol{f}^*\|^2 + \sigma^2(m - m^*) \\ &= -\|b_{m,m^*}\|^2 + \sigma^2(m - m^*) \geq 0, \end{aligned} \tag{5.13}$$

where $b_{m,m^*} \stackrel{\text{def}}{=} \Pi_{m,m^*}\boldsymbol{f}^*$.

Exercise 5.4.2. Consider the model $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$ with homogeneous errors $\sigma_i \equiv \sigma$. Let m^* be the oracle choice from (5.2).

- check (5.13);
- check that for $m < m^*$, it holds

$$\|b_{m^*,m}\|^2 = \|\Pi_{m^*,m}\boldsymbol{f}^*\|^2 \geq \sigma^2(m^* - m). \tag{5.14}$$

- check that for $m > m^*$, it holds

$$\|b_{m,m^*}\|^2 = \|\Pi_{m,m^*}\boldsymbol{f}^*\|^2 \leq \sigma^2(m - m^*) \tag{5.15}$$

In words, due to (5.14), it is reasonable to increase the model complexity towards m^*, the gain in the quality of approximation is larger than the additional complexity. However, (5.15) 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 $\widehat{m}$ reproduces this situation. The selected model $\widehat{m}$ is a winner in a pairwise competition with all other models, in particular, in competition with the "oracle" choice m^*. This means that $\widetilde{\mathcal{R}}_{\widehat{m}} \leq \widetilde{\mathcal{R}}_{m^*}$. If the value $\widetilde{\mathcal{R}}_m$ 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 $\widetilde{\mathcal{R}}_m \leq \widetilde{\mathcal{R}}_{m^*}$ is very small. So, one can expect that the selected model $\widehat{m}$ 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

$$\begin{aligned}
\widetilde{\mathcal{R}}_m - \widetilde{\mathcal{R}}_{m^*} &= \|\boldsymbol{Y} - \Pi_m \boldsymbol{Y}\|^2 + 2\sigma^2 m - \|\boldsymbol{Y} - \Pi_{m^*}\boldsymbol{Y}\|^2 - 2\sigma^2 m^* \\
&= -\|\Pi_{m,m^*}\boldsymbol{Y}\|^2 + 2\sigma^2(m - m^*) \\
&= -\|\Pi_{m,m^*}\boldsymbol{\varepsilon} + b_{m,m^*}\|^2 + 2\sigma^2(m - m^*) \\
&= \mathcal{R}_m - \mathcal{R}_{m^*} - \bigl\{\|\Pi_{m,m^*}\boldsymbol{\varepsilon}\|^2 - \sigma^2(m - m^*)\bigr\} - 2b_{m,m^*}^\top \Pi_{m,m^*}\boldsymbol{\varepsilon}. \qquad (5.16)
\end{aligned}$$

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 5.4.3. *Let $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Then it holds for $b_{m,m^*} = \Pi_{m,m^*}\boldsymbol{f}^*$*

$$\begin{aligned}
I\!P\Bigl(\sigma^{-2}\|\Pi_{m,m^*}\boldsymbol{\varepsilon}\|^2 > \mathfrak{z}^+(m - m^*, \mathtt{x})\Bigr) &\le \frac{1}{2}\mathrm{e}^{-\mathtt{x}}, \\
I\!P\Bigl(\sigma^{-2}\|\Pi_{m,m^*}\boldsymbol{\varepsilon}\|^2 < \mathfrak{z}^-(m - m^*, \mathtt{x})\Bigr) &\le \frac{1}{2}\mathrm{e}^{-\mathtt{x}}, \\
I\!P\Bigl(\sigma^{-2}\bigl|b_{m,m^*}^\top \Pi_{m,m^*}\boldsymbol{\varepsilon}\bigr| > \sigma^{-1}\|b_{m,m^*}\| z_1(\mathtt{x})\Bigr) &\le \mathrm{e}^{-\mathtt{x}}.
\end{aligned}$$

where $\mathfrak{z}^+(k, \mathtt{x})$ is the upper $1 - 0.5\mathrm{e}^{-\mathtt{x}}$ quantile of χ_k^2, $\mathfrak{z}^-(k, \mathtt{x})$ is its lower $0.5\mathrm{e}^{-\mathtt{x}}$ quantile, $z_1(\mathtt{x})$ is the quantile of $|\xi|$ for a standard normal r.v. $\xi \sim \mathcal{N}(0,1)$:

$$I\!P\bigl(|\xi| > z_1(\mathtt{x})\bigr) \le \mathrm{e}^{-\mathtt{x}}$$

It holds for any $k \ge 1$ and $\mathtt{x} > 0$ with $\mathtt{x}_1 = \mathtt{x} + \log(2)$

$$\begin{aligned}
\mathfrak{z}^+(k, \mathtt{x}) &\le k + 2\sqrt{k\,\mathtt{x}_1} + 2\mathtt{x}_1, \\
\mathfrak{z}^-(k, \mathtt{x}) &\ge k - 2\sqrt{k\,\mathtt{x}_1}. \qquad (5.17)
\end{aligned}$$

The proof only uses that $\sigma^{-1}\boldsymbol{\varepsilon}$ is a standard Gaussian vector in $I\!R^n$ and thus, $\sigma^{-1}\Pi_{m,m^*}\boldsymbol{\varepsilon}$ is standard normal in $I\!R^{m-m^*}$, while $\bigl(\sigma\|b_{m,m^*}\|\bigr)^{-1} b_{m,m^*}^\top \Pi_{m,m^*}\boldsymbol{\varepsilon}$ is a standard normal r.v. if the bias b_{m,m^*} does not vanish.

The presented bounds show that for moderate values of $\mathtt{x}$

$$z^{\pm}(k, \mathtt{x}) \stackrel{\mathrm{def}}{=} \mathfrak{z}^+(k, \mathtt{x}) - \mathfrak{z}^-(k, \mathtt{x}) \le \mathtt{C}\sqrt{k\,\mathtt{x}}.$$

for a fixed constant $\mathtt{C}$. Therefore, for large k, the interquartile range $z^{\pm}(k, \mathtt{x}) = \mathfrak{z}^+(k, \mathtt{x}) - \mathfrak{z}^-(k, \mathtt{x})$ is small relative to k which is the expectation of $\sigma^{-2} I\!E\|\Pi_k \boldsymbol{\varepsilon}\|^2$. 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.

5.4.2 Pairwise analysis

Now we make a more precise analysis of the term $\widetilde{\mathcal{R}}_{m} - \widetilde{\mathcal{R}}_{m^{*}}$ in (5.16). It is based on the following general property of the Gaussian distribution.

Lemma 5.4.4. *Let $\boldsymbol{\xi}$ be standard Gaussian vector in $I\!\!R^{k}$ and $\boldsymbol{\delta}$ be a deterministic vector in $I\!\!R^{k}$ with $\|\boldsymbol{\delta}\|^{2} = \Delta$. Then*

- *the distribution of $\|\boldsymbol{\xi} + \boldsymbol{\delta}\|^{2}$ only depends on k and Δ;*
- *let, for a given $\mathtt{x}$, the quantiles $\mathfrak{z}^{+}(k, \Delta; \mathtt{x})$ and $\mathfrak{z}^{-}(k, \Delta, \mathtt{x})$ be defined as*

$$I\!\!P\bigl(\|\boldsymbol{\xi} + \boldsymbol{\delta}\|^{2} \geq \mathfrak{z}^{+}(k, \Delta; \mathtt{x})\bigr) = \mathrm{e}^{-\mathtt{x}},$$

$$I\!\!P\bigl(\|\boldsymbol{\xi} + \boldsymbol{\delta}\|^{2} \leq \mathfrak{z}^{-}(k, \Delta; \mathtt{x})\bigr) = \mathrm{e}^{-\mathtt{x}}.$$

Then

$$\mathfrak{z}^{+}(k, \Delta; \mathtt{x}) \leq \Delta + \mathfrak{z}^{+}(k, \mathtt{x}) + 2\Delta^{1/2} z_{1}(\mathtt{x}),$$

$$\mathfrak{z}^{-}(k, \Delta; \mathtt{x}) \geq \Delta + \mathfrak{z}^{-}(k, \mathtt{x}) - 2\Delta^{1/2} z_{1}(\mathtt{x}),$$

Proof. Use the decomposition

$$\|\boldsymbol{\xi} + \boldsymbol{\delta}\|^{2} = \Delta + \|\boldsymbol{\xi}\|^{2} + 2\boldsymbol{\xi}^{\top}\boldsymbol{\delta}$$

and Lemma 5.4.3.

We apply this result to $\pm\sigma^{-2}\|\Pi_{m,m^{*}}\boldsymbol{Y}\|^{2}$ entering in the difference $\widetilde{\mathcal{R}}_{m} - \widetilde{\mathcal{R}}_{m^{*}}$. The bound $\widetilde{\mathcal{R}}_{m} - \widetilde{\mathcal{R}}_{m^{*}}$ can be rewritten for $m > m^{*}$ as $\sigma^{-2}\|\Pi_{m,m^{*}}\boldsymbol{Y}\|^{2} > 2(m - m^{*})$ which is directly related to the upper quantile of non-central chi-squared. Define the value of non-centrality parameter Δ to have $\mathfrak{z}^{\pm}(k, \Delta^{\pm}; \mathtt{x})$ exactly equal to $2k$:

$$\mathfrak{z}^{+}(k, \Delta^{+}(k, \mathtt{x}); \mathtt{x}) = 2k, \qquad \mathfrak{z}^{-}(k, \Delta^{-}(k, \mathtt{x}); \mathtt{x}) = 2k. \tag{5.18}$$

This definition can be rewritten as follows:

$$I\!\!P\bigl(\|\boldsymbol{\xi} + \boldsymbol{\delta}^{+}\|^{2} > 2k\bigr) \leq \mathrm{e}^{-\mathtt{x}}, \qquad \text{if } \|\boldsymbol{\delta}^{+}\|^{2} \leq \Delta^{+}(k, \mathtt{x}), \tag{5.19}$$

$$I\!\!P\bigl(\|\boldsymbol{\xi} + \boldsymbol{\delta}^{-}\|^{2} < 2k\bigr) \leq \mathrm{e}^{-\mathtt{x}}, \qquad \text{if } \|\boldsymbol{\delta}^{-}\|^{2} \geq \Delta^{-}(k, \mathtt{x}). \tag{5.20}$$

Exercise 5.4.3. For the quantities $\Delta^{+}(k, \mathtt{x}), \Delta^{-}(k, \mathtt{x})$ from (5.18) and $\mathtt{x} \geq 1$

- show that $\Delta^{+}(k, \mathtt{x}) < k$, $\Delta^{-}(k, \mathtt{x}) > k$;
- check that Lemma 5.4.4 implies

$$\Delta^{+}(k, \mathtt{x}) \geq \mathfrak{z}^{-}(k, \mathtt{x}) - 2k^{1/2} z_{1}(\mathtt{x}), \tag{5.21}$$

$$\Delta^{-}(k, \mathtt{x}) \leq \mathfrak{z}^{+}(k, \mathtt{x}) + 2k^{1/2} z_{1}(\mathtt{x}), \tag{5.22}$$

We conclude with the following statement.

Proposition 5.4.1. *Let the errors* $\boldsymbol{\varepsilon}$ *be normal and homogeneous:* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. *Then the inequalities*

$$\begin{aligned} \sigma^{-2}\|b_{m,m^*}\|^2 &\leq \Delta^+(m - m^*, \mathtt{x}), \qquad m > m^* \qquad (5.23)\\ \sigma^{-2}\|b_{m^*,m}\|^2 &\geq \Delta^-(m^* - m, \mathtt{x}), \qquad m < m^* \end{aligned}$$

ensures

$$I\!\!P\bigl(\widetilde{\mathcal{R}}_m \leq \widetilde{\mathcal{R}}_{m^*}\bigr) \leq \mathrm{e}^{-\mathtt{x}}.$$

In particular, if the bias term $\|b_m\|$ *is uniformly bounded for all* $m \geq m^*$ *by a fixed constant* $\mathtt{C}(\mathcal{F})$, *then* $\|b_{m,m^*}\| \leq \mathtt{C}(\mathcal{F})$ *and the inequality* (5.23) *is fulfilled if*

$$m - m^* > \bigl(2\sigma^{-1}\mathtt{C}(\mathcal{F}) + \mathtt{C}\,\mathtt{x}\bigr)^2 \tag{5.24}$$

for $\mathtt{C} \geq 3$.

Proof. Consider the case $m > m^*$. We apply the decomposition

$$\sigma^{-2}\bigl(\widetilde{\mathcal{R}}_m - \widetilde{\mathcal{R}}_{m^*}\bigr) = -\|\sigma^{-1}\Pi_{m,m^*}\boldsymbol{\varepsilon} + \sigma^{-1}b_{m,m^*}\|^2 + 2(m - m^*).$$

Further, $\boldsymbol{\xi} \overset{\text{def}}{=} \sigma^{-1}\Pi_{m,m^*}\boldsymbol{\varepsilon}$ is standard normal in $I\!\!R^k$ for $k = m - m^*$. The condition (5.19) with $\boldsymbol{\delta}^+ = \sigma^{-1}b_{m,m^*}$ implies

$$I\!\!P\bigl(\widetilde{\mathcal{R}}_m \leq \widetilde{\mathcal{R}}_{m^*}\bigr) = I\!\!P\bigl(\|\boldsymbol{\xi} + \boldsymbol{\delta}^+\|^2 > 2k\bigr) \leq \mathrm{e}^{-\mathtt{x}}.$$

The case $m < m^*$ can be done in a similar way using (5.20) in place of (5.19).

The inequality $\|b_m\| \leq \mathtt{C}(\mathcal{F})$ implies $\|b_{m,m^*}\| \leq \mathtt{C}(\mathcal{F})$ for all $m > m^*$; see (5.25). Now it remains to check by (5.21) and (5.17) that (5.24) implies (5.23).

To be done: complete the proof

Exercise 5.4.4. Show that the value $\|b_m\| = \|\boldsymbol{f}^* - \Pi_m \boldsymbol{f}^*\|$ monotonously decreases with m. Moreover, for any $m' > m$, the relative bias $b_{m',m} = \Pi_{m',m}\boldsymbol{f}^*$ satisfies

$$\|b_{m',m}\| \leq \|b_m\|, \quad m' > m. \tag{5.25}$$

Check whether also holds

$$\|b_{m',m}\| \leq \|b_{m'}\|, \quad m' > m.$$

5.4.3 Uniform bounds and the zone of insensitivity

This section introduces the *set of insensitivity* $\mathcal{M}^{\circ}(\mathtt{x})$ which describes the quality of model selection. Namely, we aim at describing the set $\mathcal{M}^{\circ}(\mathtt{x})$ which contains all possible values of $\widehat{m}$ with a high probability. The ideal situation would be $\mathcal{M}^{\circ}(\mathtt{x}) = \{m^{*}\}$, but it is rare the case. Usually $\mathcal{M}^{\circ}(\mathtt{x})$ 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 (5.13).

The study of the previous section quantifies the pairwise relation $\widetilde{\mathcal{R}}_{m} - \widetilde{\mathcal{R}}_{m^{*}} \le 0$: under $\sigma^{-2}\|b_{m,m^{*}}\|^{2} \le \Delta^{+}(m - m^{*}, \mathtt{x})$; see (5.23), it holds

$$\mathbb{P}\bigl(\widetilde{\mathcal{R}}_{m} \le \widetilde{\mathcal{R}}_{m^{*}}\bigr) \le \mathrm{e}^{-\mathtt{x}}. \tag{5.26}$$

Now we need its uniform version over the complement of $\mathcal{M}^{\circ}(\mathtt{x})$:

$$\mathbb{P}\biggl(\max_{m \notin \mathcal{M}^{\circ}(\mathtt{x})} \{\widetilde{\mathcal{R}}_{m} - \widetilde{\mathcal{R}}_{m^{*}}\} \ge 0\biggr) \le \mathrm{e}^{-\mathtt{x}}.$$

One can use a uniform adjustment in each bound (5.26) by increasing the value $\mathtt{x}$ to another slightly larger level $\mathtt{x}_{\mathtt{s}}$. A simple way is based on the so called Bonferroni correction: $\mathtt{x}_{\mathtt{s}} \equiv \mathtt{x} + \log(|\mathcal{M}|)$.

Proposition 5.4.2. *Let* $\varepsilon \sim \mathcal{N}(0, \sigma^{2} I_{n})$. *If* $\mathcal{M}^{\circ}(\mathtt{x})$ *is the set of indices* m *such that*

$$\mathcal{M}^{\circ}(\mathtt{x}) \stackrel{\mathrm{def}}{=} \begin{cases} \sigma^{-2}\|b_{m,m^{*}}\|^{2} \ge \Delta^{+}(m - m^{*}, \mathtt{x}_{\mathtt{s}}), & m > m^{*}, \\ \sigma^{-2}\|b_{m^{*},m}\|^{2} \le \Delta^{-}(m^{*} - m, \mathtt{x}_{\mathtt{s}}), & m < m^{*}, \end{cases} \tag{5.27}$$

for $\mathtt{x}_{\mathtt{s}} = \mathtt{x} + \log(|\mathcal{M}|)$, *then*

$$\mathbb{P}\bigl(\widehat{m} \notin \mathcal{M}^{\circ}(\mathtt{x})\bigr) \le \mathrm{e}^{-\mathtt{x}}. \tag{5.28}$$

Proof. By definition of $\mathcal{M}^{\circ}(\mathtt{x})$ and Proposition 5.4.1

$$\mathbb{P}\bigl(\widehat{m} \notin \mathcal{M}^{\circ}(\mathtt{x})\bigr) \le \sum_{m \notin \mathcal{M}^{\circ}(\mathtt{x})} \mathbb{P}\bigl(\widehat{m} = m\bigr) \le \sum_{m \notin \mathcal{M}^{\circ}(\mathtt{x})} \mathbb{P}\bigl(\widetilde{\mathcal{R}}_{m} \le \widetilde{\mathcal{R}}_{m^{*}}\bigr) \le \sum_{m \in \mathcal{M}} \mathrm{e}^{-\mathtt{x}_{\mathtt{s}}} \le \mathrm{e}^{-\mathtt{x}}.$$

One can conclude that if the m lies beyond the *insensitivity zone* $\mathcal{M}^{\circ}(\mathtt{x})$ around m^{*}, on which the difference $\mathcal{R}_{m} - \mathcal{R}_{m^{*}}$ is not sufficiently large, then the event $\widehat{m} = m$ is very unlikely. The result (5.28) can be stated in the form that there exists a random set $\Omega(\mathtt{x})$ such that $\mathbb{P}\bigl(\Omega(\mathtt{x})\bigr) \ge 1 - \mathrm{e}^{-\mathtt{x}}$, and on this set, it holds

$$\widetilde{\mathcal{R}}_{m} > \widetilde{\mathcal{R}}_{m^{*}}, \qquad m \notin \mathcal{M}^{\circ}(\mathtt{x})$$

and hence $\widehat{m} \in \mathcal{M}^{\circ}(\mathtt{x})$ on $\Omega(\mathtt{x})$.

5.4.4 A bound on the excess

Introduce another random set $\Omega_0(\mathtt{x})$ such that

$$\begin{aligned} \|\sigma^{-1}\Pi_{m,m^*}\boldsymbol{\varepsilon}\|^2 &\le \mathfrak{z}^+(m-m^*,\mathtt{x}_\mathtt{s}), \qquad && m>m^*, \\ \|\sigma^{-1}\Pi_{m^*,m}\boldsymbol{\varepsilon}\|^2 &\ge \mathfrak{z}^-(m^*-m,\mathtt{x}_\mathtt{s}), \qquad && m<m^*. \end{aligned} \tag{5.29}$$

Lemma 5.4.3 implies that $I\!\!P(\Omega_0(\mathtt{x})) \ge 1-\mathrm{e}^{-\mathtt{x}}$.

The next question is what happens if $\widehat{m} \in \mathcal{M}^{\circ}(\mathtt{x})$ and how big this set is. We will try to bound the loss difference $\varrho(\widetilde{\boldsymbol{f}}_m,\boldsymbol{f}^*) - \varrho(\widetilde{\boldsymbol{f}}_{m^*},\boldsymbol{f}^*)$. It follows from (**??**) that for $m>m^*$

$$\sigma^{-2}\bigl\{\varrho(\widetilde{\boldsymbol{f}}_m,\boldsymbol{f}^*) - \varrho(\widetilde{\boldsymbol{f}}_{m^*},\boldsymbol{f}^*)\bigr\} = -\|\sigma^{-1}b_{m,m^*}\|^2 + \|\sigma^{-1}\Pi_{m,m^*}\boldsymbol{\varepsilon}\|^2.$$

This implies for $m \in \mathcal{M}^+(\mathtt{x})$ by (5.27) and (5.21) on $\Omega_0(\mathtt{x})$

$$\begin{aligned} &\sigma^{-2}\bigl\{\varrho(\widetilde{\boldsymbol{f}}_m,\boldsymbol{f}^*) - \varrho(\widetilde{\boldsymbol{f}}_{m^*},\boldsymbol{f}^*)\bigr\} \\ &\quad \le -\Delta^+(m-m^*,\mathtt{x}_\mathtt{s}) + \mathfrak{z}^+(m-m^*,\mathtt{x}_\mathtt{s}) \\ &\quad \le \mathfrak{z}^+(m-m^*,\mathtt{x}_\mathtt{s}) - \mathfrak{z}^-(m-m^*,\mathtt{x}_\mathtt{s}) + 2z_1(\mathtt{x}_\mathtt{s})\sqrt{m-m^*} \\ &\quad = z^{\pm}(m-m^*,\mathtt{x}_\mathtt{s}) + 2z_1(\mathtt{x}_\mathtt{s})\sqrt{m-m^*}; \end{aligned} \tag{5.30}$$

here $z^{\pm}(k,\mathtt{x}) = \mathfrak{z}^+(k,\mathtt{x}) - \mathfrak{z}^-(k,\mathtt{x})$.

Now we consider the similar difference for the parameter m from the insensitivity zone $\mathcal{M}^-(\mathtt{x})$ with $m<m^*$. It holds

$$\sigma^{-2}\bigl\{\varrho(\widetilde{\boldsymbol{f}}_m,\boldsymbol{f}^*) - \varrho(\widetilde{\boldsymbol{f}}_{m^*},\boldsymbol{f}^*)\bigr\} = \|\sigma^{-1}b_{m^*,m}\|^2 - \|\sigma^{-1}\Pi_{m^*,m}\boldsymbol{\varepsilon}\|^2.$$

This implies for $m \in \mathcal{M}^{\circ}(\mathtt{x})$ by (5.27) and (5.22) on $\Omega_0(\mathtt{x})$

$$\begin{aligned} &\sigma^{-2}\bigl\{\varrho(\widetilde{\boldsymbol{f}}_m,\boldsymbol{f}^*) - \varrho(\widetilde{\boldsymbol{f}}_{m^*},\boldsymbol{f}^*)\bigr\} \\ &\quad \le \|\sigma^{-1}b_{m^*,m}\|^2 - \|\sigma^{-1}\Pi_{m^*,m}\boldsymbol{\varepsilon}\|^2 \\ &\quad \le \Delta^-(m^*-m,\mathtt{x}_\mathtt{s}) - \mathfrak{z}^-(m^*-m,\mathtt{x}_\mathtt{s}) \\ &\quad \le \mathfrak{z}^+(m^*-m,\mathtt{x}_\mathtt{s}) - \mathfrak{z}^-(m^*-m,\mathtt{x}_\mathtt{s}) + 2z_1(\mathtt{x}_\mathtt{s})\sqrt{m^*-m} \\ &\quad = z^{\pm}(m^*-m,\mathtt{x}_\mathtt{s}) + 2z_1(\mathtt{x}_\mathtt{s})\sqrt{m^*-m}. \end{aligned} \tag{5.31}$$

Now we can summarize. Define the radius $R = R(\mathcal{M}^{\circ}(\mathtt{x}))$ of the set $\mathcal{M}^{\circ}(\mathtt{x})$:

$$R \stackrel{\mathrm{def}}{=} \max_{m \in \mathcal{M}^{\circ}(\mathtt{x})} |m-m^*|.$$

Theorem 5.4.1. *Let $\widehat{m}$ be defined by (5.12) and m^{*} be the oracle choice from (5.2). Suppose that $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^{2} I_{n})$. Let $\mathtt{x}_{\mathtt{s}} = \mathtt{x} + \log(|\mathcal{M}|)$. For the insensitivity zone $\mathcal{M}^{\circ}(\mathtt{x})$ from (5.27), it holds $\widehat{m} \in \mathcal{M}^{\circ}(\mathtt{x})$ on a random set $\Omega(\mathtt{x})$ with $I\!\!P\bigl(\Omega(\mathtt{x})\bigr) \geq 1 - \mathrm{e}^{-\mathtt{x}}$. Moreover, on a random set $\Omega_{0}(\mathtt{x})$ with $I\!\!P\bigl(\Omega_{0}(\mathtt{x})\bigr) \geq 1 - 2\mathrm{e}^{-\mathtt{x}}$*

$$\sigma^{-2}\bigl\{\varrho(\widetilde{\boldsymbol{f}}_{\widehat{m}}, \boldsymbol{f}^{*}) - \varrho(\widetilde{\boldsymbol{f}}_{m^{*}}, \boldsymbol{f}^{*})\bigr\} \leq z^{\pm}(R, \mathtt{x}_{\mathtt{s}}) + 2 z_{1}(\mathtt{x}_{\mathtt{s}})\sqrt{R}.$$

Proof. The result follows from (5.30) and (5.31) by monotonicity of the function $z^{\pm}(k, \mathtt{x})$ in k and $\mathtt{x}$.

In view of $z^{\pm}(R, \mathtt{x}_{\mathtt{s}}) \asymp \sqrt{R\,\mathtt{x}_{\mathtt{s}}}$, we conclude that the data-driven choice of the parameter m leads to additional loss of order $\sigma^{2}\sqrt{R\,\mathtt{x}_{\mathtt{s}}}$. One can say that the model selection based on unbiased risk estimation works well if the size of the zone of insensitivity $R = R(\mathcal{M}^{\circ}(\mathtt{x}))$ is not too large compared with the loss and risk of the oracle $\widetilde{\boldsymbol{f}}_{m^{*}}$.

Note that the loss $\varrho(\widetilde{\boldsymbol{f}}_{m^{*}}, \boldsymbol{f}^{*})$ fulfills

$$\varrho(\widetilde{\boldsymbol{f}}_{m^{*}}, \boldsymbol{f}^{*}) = \|b_{m^{*}}\|^{2} + \|\Pi_{m^{*}}\boldsymbol{\varepsilon}\|^{2} \geq \|\Pi_{m^{*}}\boldsymbol{\varepsilon}\|^{2}.$$

By Lemma 5.4.3, it is of order m^{*}.

Exercise 5.4.5. Let $\Omega_{0}(\mathtt{x})$ be a random set on which (5.29) holds. Show that $I\!\!P\bigl(\Omega_{0}(\mathtt{x})\bigr) \geq 1 - \mathrm{e}^{-\mathtt{x}}$ and for every m, the loss $\varrho(\widetilde{\boldsymbol{f}}_{m}, \boldsymbol{f}^{*})$ satisfies on $\Omega_{0}(\mathtt{x})$

$$\sigma^{-2}\varrho(\widetilde{\boldsymbol{f}}_{m}, \boldsymbol{f}^{*}) \geq \mathfrak{z}^{-}(m, \mathtt{x}).$$

For the case when the value $R = \max_{m \in \mathcal{M}^{\circ}(\mathtt{x})} |m - m^{*}|$ is small relative to m^{*}, the loss of the estimate $\widehat{\boldsymbol{f}} = \widetilde{\boldsymbol{f}}_{\widehat{m}}$ corresponding to the data-driven selector $\widehat{m}$ is not significantly larger than the loss of the oracle estimate $\widetilde{\boldsymbol{f}}_{m^{*}}$. Unfortunately, the set $\mathcal{M}^{\circ}(\mathtt{x})$ 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 (5.1). One can easily check that

$$\|b_{m,m^{*}}\|^{2} \equiv \sigma^{2}(m - m^{*}), \qquad m > m^{*},$$
$$\|b_{m^{*},m}\|^{2} \equiv \sigma^{2}(m^{*} - m), \qquad m < m^{*}.$$

This implies that the risk function $\mathcal{R}_{m}$ is constant in m, and therefore, the set $\mathcal{M}^{\circ}(\mathtt{x})$ coincides with the whole set $\mathcal{M}$.

Exercise 5.4.6. Build an example in which the radius R is twice as large as the oracle risk $\mathcal{R}_{m^{*}}$.

6

Unbiased risk estimation for linear models

6.1 Model and problem

Consider the following regression model:

$$Y_i = f(X_i) + \varepsilon_i \qquad I\!\!E\varepsilon_i = 0, \qquad i = 1, \ldots, n,$$

Here $\varepsilon_1, \ldots, \varepsilon_n$ are individual zero mean errors with finite variance, and $X_1, \ldots, X_n$ are given design points. Below we assume a deterministic design, otherwise one can understand the results conditioned on the design. We also write this equation in vector form $\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\varepsilon} \in I\!\!R^n$. Given a set of basis functions $\psi_1, \ldots, \psi_p$, define the feature vectors $\Psi_1, \ldots, \Psi_n$ in $I\!\!R^p$ with $\Psi_i = \bigl(\psi_1(X_i), \ldots, \psi_p(X_i)\bigr)^\top$. Linear modeling assumes a linear expansion $f = \sum_{j=1}^{p} \theta_j^* \psi_j$. This relation in vector form reads as $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon} \in I\!\!R^n$, where Ψ is a $p \times n$ design matrix.

The dimension p can be large, even $p = \infty$ can be incorporated. For notational simplicity, we proceed with p finite. The proposed approach can also be extended to the case when the linear parametric assumption $I\!\!E\boldsymbol{Y} = \boldsymbol{f} = \Psi^\top \boldsymbol{\theta}^*$ is not precisely fulfilled. Then, as usual, the target of estimation $\boldsymbol{\theta}^*$ can be defined as the vector of coefficients for the best approximation of the true response $\boldsymbol{f} \overset{\text{def}}{=} I\!\!E\boldsymbol{Y} = (f_1, \ldots, f_n)^\top$ by linear combinations of the features ψ_j which are the rows of the matrix Ψ. We write the underlying model in the vector form

$$\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\varepsilon} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}\,. \tag{6.1}$$

Let $\bigl\{\widetilde{\boldsymbol{\theta}}_m, m \in \mathcal{M}\bigr\}$ be a finite family of linear estimators $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$ of $\boldsymbol{\theta}^*$. 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:

$$\varrho_m \stackrel{\mathrm{def}}{=} \|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2, \qquad \mathcal{R}_m \stackrel{\mathrm{def}}{=} I\!\!E\|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2.$$

Alternatively one can say that

$$\widetilde{\boldsymbol{\phi}}_m \stackrel{\mathrm{def}}{=} W\widetilde{\boldsymbol{\theta}}_m = W\mathcal{S}_m\boldsymbol{Y} = \mathcal{K}_m\boldsymbol{Y}$$

with $\mathcal{K}_m = W\mathcal{S}_m$ is an estimator of the target $\boldsymbol{\phi}^* = W\boldsymbol{\theta}^*$ and $\varrho_m = \|\widetilde{\boldsymbol{\phi}}_m - \boldsymbol{\phi}^*\|^2$. Typical examples of W are as follows.

Estimation of the whole vector $\boldsymbol{\theta}^*$

Let W be the identity matrix $W = \boldsymbol{I}_p$ with $q = p$. This means that the *estimation loss* is measured by the usual squared Euclidean distance $\|\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\|^2$.

Prediction

Let W be the square root of the $p \times p$ matrix $I\!\!F = \Psi\Psi^\top$, that is, $W^2 = I\!\!F$. The loss $\|W(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \|\Psi(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|$ is usually referred to as *prediction loss* because it measures the prediction ability of the true model by the model with the parameter $\boldsymbol{\theta}$.

Semiparametric estimation

Suppose that the target of estimation is not the whole vector $\boldsymbol{\theta}^*$ but some subvector $\boldsymbol{\theta}_0^*$ of dimension q. The matrix W can be defined as the projector Π_0 on the $\boldsymbol{\theta}_0^*$ subspace. The estimate $\Pi_0\widetilde{\boldsymbol{\theta}}_m$ is called the *profile maximum likelihood estimate.* The corresponding loss is equal to the squared Euclidean distance in this subspace:

$$\varrho_m = \bigl\|\Pi_0\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\bigr)\bigr\|^2.$$

Alternatively, one can select W^2 as the efficient information matrix defined by relation $W^2 = \bigl(\Pi_0 I\!\!F^{-1}\Pi_0^\top\bigr)^-$, where A^- means a pseudo-inverse of A.

Linear functional estimation

The choice of the weighting matrix W of rank one can be adjusted to address the problem of estimating some functionals of the whole parameter $\boldsymbol{\theta}^*$, for instance, the first coefficient θ_1^* or the sum of the θ_j^*'s. For instance, in the regression problem $I\!\!E Y_i = f(X_i)$ with the Fourier expansion the target function f can be represented as

$$f(x) = \sum_{j \ge 0} \theta_j^* \psi_j(x) = \sum_{j \ge 0} \bigl\{\theta_{2j}^* \cos(2\pi jx) + \theta_{2j+1}^* \sin(2\pi jx)\bigr\}.$$

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 $f'(0) = 2\pi \sum_{j \ge 0} j\theta_{2j+1}^*$.

In all cases, the most important feature of the estimators $\widetilde{\boldsymbol{\phi}}_m = \mathcal{K}_m \boldsymbol{Y}$ is *linearity.* It greatly simplifies the study of their properties including the prominent bias-variance decomposition of the risk of $\widetilde{\boldsymbol{\phi}}_m$. Namely, for the model (7.1) with $I\!\!E\boldsymbol{\varepsilon} = 0$ and $\boldsymbol{f} = I\!\!E\boldsymbol{Y}$, it holds

$$\begin{aligned} I\!\!E\widetilde{\boldsymbol{\phi}}_m &= \boldsymbol{\phi}_m^* \stackrel{\text{def}}{=} \mathcal{K}_m \boldsymbol{f}, \\ \mathcal{R}_m &= \|\boldsymbol{\phi}_m^* - \boldsymbol{\phi}^*\|^2 + \operatorname{tr}\bigl\{\mathcal{K}_m \operatorname{Var}(\boldsymbol{\varepsilon})\, \mathcal{K}_m^\top\bigr\} = \|\boldsymbol{b}_m\|^2 + \mathtt{p}_m , \end{aligned} \tag{6.2}$$

where $\|\boldsymbol{b}_m\|^2 \stackrel{\text{def}}{=} \|\boldsymbol{\phi}_m^* - \boldsymbol{\phi}^*\|^2$ is the squared bias term and $\mathtt{p}_m \stackrel{\text{def}}{=} \operatorname{tr}\operatorname{Var}(\widetilde{\boldsymbol{\phi}}_m)$ is the variance term. This is the usual "bias-variance" decomposition of the squared risk $\mathcal{R}_m$. The optimal choice of m is often defined by risk minimization:

$$m_{\text{opt}} \stackrel{\text{def}}{=} \operatorname*{argmin}_{m \in \mathcal{M}} \mathcal{R}_m = \operatorname*{argmin}_{m \in \mathcal{M}} \bigl(\|\boldsymbol{b}_m\|^2 + \mathtt{p}_m\bigr). \tag{6.3}$$

Alternatively one can define the best choice m^* via the "bias-variance trade-off"; see the definition below in (7.9). The *model selection* problem can be described as a data-based choice $\widehat{m}$ which leads to essentially the same quality of the adaptive estimator $\widetilde{\boldsymbol{\theta}}_{\widehat{m}}$ as for the optimal choice m^*.

6.2 Ordered case

This section discusses the *ordered* case. For simplicity of presentation, we assume that $\mathcal{M}$ is a finite set of positive numbers, although the approach can be extended to situations with a countable and/or continuous and even unbounded set $\mathcal{M}$. $|\mathcal{M}|$ stands for the cardinality of $\mathcal{M}$. Typical examples of the parameter m are given by a chosen dimension (number of basis vectors) in projection estimation or by the bandwidth in kernel smoothing. In general, complexity can be naturally expressed via the variance of the stochastic term of the estimator $\widetilde{\boldsymbol{\phi}}_m$: the larger m, the larger is the variance term $\mathtt{p}_m = \operatorname{tr}\bigl\{\operatorname{Var}(\widetilde{\boldsymbol{\phi}}_m)\bigr\}$. In the case of projection estimation and a homogeneous noise $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$, this variance term is linear in m: $\mathtt{p}_m = \sigma^2 m$; see Section 7.3.3 for details. In general, dependence of the variance term on m is more complicated but monotonicity of $\mathtt{p}_m$ in m should be preserved. The related condition can be written as

$$\operatorname{tr}\bigl(\mathcal{K}_m \operatorname{Var}(\boldsymbol{\varepsilon})\, \mathcal{K}_m^\top\bigr) \le \operatorname{tr}\bigl(\mathcal{K}_{m'} \operatorname{Var}(\boldsymbol{\varepsilon})\, \mathcal{K}_{m'}^\top\bigr), \qquad m' > m. \tag{6.4}$$

Further, it is implicitly assumed that the bias term $\|\boldsymbol{b}_m\|^2 = \|\boldsymbol{\phi}_m^* - \boldsymbol{\phi}^*\|^2$ becomes smaller as m increases. The smallest index $m_0 \in \mathcal{M}$ corresponds to the simplest (zero) model, usually with a large bias, while a large m ensures a good approximation quality $\boldsymbol{\phi}_m^* \approx \boldsymbol{\phi}^*$ 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 (7.2) describes the accuracy of approximating the response $\boldsymbol{f}$ 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 squared bias $\|\boldsymbol{b}_m\|^2$ monotonously decreases with m. An example is given below.

Example 6.2.1. Suppose that a signal $\boldsymbol{\theta}^*$ is observed with noise: $Y_i = \theta_i^* + \varepsilon_i$. Consider the set of projection estimates $\widetilde{\boldsymbol{\theta}}_m$ on the first m coordinates and the target is $\phi^* \stackrel{\text{def}}{=} W\boldsymbol{\theta}^* = \sum_j \theta_j^*$. If $\boldsymbol{\theta}^*$ is composed of alternating blocks of 1's and -1's with equal length, then the bias $|\phi^* - \phi_m^*|$ for $\phi_m^* = \sum_{j \le m} \theta_j^*$ is not monotonous in m.

6.3 Partially ordered case and the largest model

The ordering assumption is natural if the tuning parameter is one dimensional. This can be a bandwidth, size of the model, parameter of regularization. However, if we have to tune two or more parameters, it is rare the case that the ordering condition is fulfilled. Then one can introduce a partial ordering relation using a stronger condition than (6.5). Namely, $m' > m$ if the related variance matrix $\operatorname{Var}(\widetilde{\phi}_{m'})$ is larger than $\operatorname{Var}(\widetilde{\phi}_m)$

$$\mathcal{K}_m \operatorname{Var}(\boldsymbol{\varepsilon})\, \mathcal{K}_m^\top < \mathcal{K}_{m'} \operatorname{Var}(\boldsymbol{\varepsilon})\, \mathcal{K}_{m'}^\top. \tag{6.5}$$

Here $A > B$ for two symmetric matrices A, B means that the matrix $A - B$ is non-negative definite and not equal to zero. One example is given by the feature selection problem when m is a subset of the index set $\{1, \ldots, p\}$ for a p-dimensional linear model $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon}$.

Exercise 6.3.1. Consider a feature selection problem when $\boldsymbol{\theta} \in I\!\!R^p$ and m is an index subset of $\{1, \ldots, p\}$. Check that $m' \ge m$ if $m \subseteq m'$ whatever W is.

In spite of partial ordering assumption, we always assume that there exists the largest model M, so that $m \le M$ for all $m \in \mathcal{M}$. This largest model is a proxy for the data themselves. If the prediction problem of the estimating the response $\boldsymbol{f}^*$, the largest model can be taken as “no smoothing case” yielding the observation vector $\boldsymbol{Y}$ an an estimator $\widetilde{\boldsymbol{f}}_M$.

Exercise 6.3.2. Consider a feature selection problem. Check that $M = \{1, \ldots, p\}$ is the largest model whatever W is.

Exercise 6.3.3. Consider a piecewise polynomial approximation for a univariate regression function and for a given partition of the interval $[0, 1]$. Check that complexity grows with the degree of piecewise polynomials.

We also suppose that this large model yields a kind of "overfitting", that is, the related bias is small, however, the variance is large. The "small bias" condition allows to write $\|\boldsymbol{b}_M\|^2 = \|\boldsymbol{\phi}^*_M - \boldsymbol{\phi}^*\|^2 \approx 0$, and we will just ignore this bias.

6.4 Unbiased risk estimation

Our study heavily uses the pairwise comparison for two different models m' and m :

$$\varrho_{m',m} \stackrel{\text{def}}{=} \|W(\widetilde{\boldsymbol{\theta}}_{m'} - \widetilde{\boldsymbol{\theta}}_m)\|^2 = \|\widetilde{\boldsymbol{\phi}}_{m'} - \widetilde{\boldsymbol{\phi}}_m\|^2 = \|\mathcal{K}_{m'}\boldsymbol{Y} - \mathcal{K}_m\boldsymbol{Y}\|^2.$$

The corresponding relative risk is given by

$$\mathcal{R}_{m',m} \stackrel{\text{def}}{=} \mathbb{E}\|\widetilde{\boldsymbol{\phi}}_{m'} - \widetilde{\boldsymbol{\phi}}_m\|^2 = \mathbb{E}\|\mathcal{K}_{m'}\boldsymbol{Y} - \mathcal{K}_m\boldsymbol{Y}\|^2.$$

Lemma 6.4.1. *It holds under* $\operatorname{Var}(\boldsymbol{\varepsilon}) = \Sigma^2$

$$\mathcal{R}_{m',m} = \|\boldsymbol{b}_{m',m}\|^2 + \mathtt{p}_{m',m} \tag{6.6}$$

with

$$\begin{aligned}\boldsymbol{b}_{m',m} &\stackrel{\text{def}}{=} (\mathcal{K}_{m'} - \mathcal{K}_m)\boldsymbol{f}^* = \mathcal{K}_{m',m}\boldsymbol{f}^*,\\ \mathtt{p}_{m',m} &\stackrel{\text{def}}{=} \operatorname{tr}\operatorname{Var}\bigl(\widetilde{\boldsymbol{\phi}}_{m'} - \widetilde{\boldsymbol{\phi}}_m\bigr) = \operatorname{tr}\bigl\{\mathcal{K}_{m',m}\operatorname{Var}(\boldsymbol{\varepsilon})\,\mathcal{K}_{m',m}^{\top}\bigr\} = \operatorname{tr}\bigl\{\mathcal{K}_{m',m}\,\Sigma\,\mathcal{K}_{m',m}^{\top}\bigr\}.\end{aligned}$$

Suppose now that the collection $\mathcal{M}$ contains the largest model M with vanishing bias $\boldsymbol{b}_M$. The formula (6.6) for the relative risk $\mathcal{R}_{m',m}$ applied with $m' = M$ yields

$$\mathcal{R}_{M,m} = \|\boldsymbol{b}_{M,m}\|^2 + \mathtt{p}_{M,m} .$$

A good news here is that the bias term $\|\boldsymbol{b}_{M,m}\|^2$ is nearly the same as in the formula for the risk $\mathcal{R}_m$. Now it becomes clear how to repair the empirical value $\mathcal{R}_{M,m}$ to get a nearly unbiased estimate of the risk:

$$\widehat{\mathcal{R}}_m \stackrel{\text{def}}{=} \bigl\|\widetilde{\boldsymbol{\phi}}_M - \widetilde{\boldsymbol{\phi}}_m\bigr\|^2 + \mathtt{p}_m - \mathtt{p}_{M,m} . \tag{6.7}$$

Exercise 6.4.1. Compute $\mathtt{p}_m - \mathtt{p}_{M,m}$ for the case with $\Sigma = \sigma^2 \boldsymbol{I}_n$:

$$\mathtt{p}_m - \mathtt{p}_{M,m} = 2\sigma^2 \operatorname{tr}(\mathcal{K}_m\mathcal{K}_M^{\top}) - \sigma^2 \operatorname{tr}(\mathcal{K}_M\mathcal{K}_M^{\top}).$$

Argue that the last term can be omitted in the definition of $\widehat{m}$.

Exercise 6.4.2. Check that the formula (6.7) yields the Akaiki criterion if the estimator $\widetilde{\boldsymbol{\phi}}_m = W\widetilde{\boldsymbol{\theta}}_m$ is a projector on a linear subspace.

Theorem 6.4.1. *Suppose that* $\operatorname{Var}(\boldsymbol{\varepsilon}) = \Sigma$ *and* $b_M = 0$. *Then for any* $m \in \mathcal{M}$

$$I\!E\widehat{\mathcal{R}}_m = \mathcal{R}_m .$$

This result suggests to define the empirical analog of m^* by minimization of $\widehat{\mathcal{R}}_m$:

$$\widehat{m} \stackrel{\text{def}}{=} \operatorname*{argmin}_{m \in \mathcal{M}} \widehat{\mathcal{R}}_m .$$

The main question for practical applications is to understand whether the empirical selector $\widehat{m}$ does a good job in each particular situation. Indeed, so far we only know that the mean value of $\widehat{\mathcal{R}}_m$ coincides with the risk $\mathcal{R}_m$. If the variability of $\widehat{\mathcal{R}}_m$ around its mean is larger or comparable with the value of the risk itself, we cannot expect a reasonable or stable behavior of the selector $\widehat{m}$. This is similar to our study of the AIC. We again apply a pairwise comparison and try to exclude situations when $\mathcal{R}_m$ is significantly larger than $\mathcal{R}_{m^*}$ but $\widehat{\mathcal{R}}_m < \widehat{\mathcal{R}}_{m^*}$. Therefore, it would be very desirable to show that the stochastic part of this difference is significantly smaller than its systematic part $\mathcal{R}_m - \mathcal{R}_{m^*}$:

$$\bigl| \widehat{\mathcal{R}}_m - \widehat{\mathcal{R}}_{m^*} - (\mathcal{R}_m - \mathcal{R}_{m^*}) \bigr| \ll \mathcal{R}_m - \mathcal{R}_{m^*} .$$

Then the minimization of the empirical risk $\widehat{\mathcal{R}}_m$ estimate will be similar to minimization of the true risk $\mathcal{R}_m$. The key step in this study is the following decomposition of the difference $\widehat{\mathcal{R}}_m - \widehat{\mathcal{R}}_{m^*}$.

Lemma 6.4.2. *It holds*

$$\widehat{\mathcal{R}}_m - \widehat{\mathcal{R}}_{m^*} = \mathcal{R}_m - \mathcal{R}_{m^*} + \boldsymbol{\varepsilon}^\top A_{m,m^*} \boldsymbol{\varepsilon} - \bigl(\mathtt{p}_{M,m} - \mathtt{p}_{M,m^*} \bigr) + 2 \boldsymbol{\varepsilon}^\top \boldsymbol{a}_{m,m^*} ,$$

where

$$\begin{aligned} A_{m,m^*} &\stackrel{\text{def}}{=} \mathcal{K}_{M,m} \mathcal{K}_{M,m}^\top - \mathcal{K}_{M,m^*} \mathcal{K}_{M,m^*}^\top , \\ \boldsymbol{a}_{m,m^*} &\stackrel{\text{def}}{=} \mathcal{K}_{M,m} \boldsymbol{b}_{M,m} - \mathcal{K}_{M,m^*} \boldsymbol{b}_{M,m^*} . \end{aligned}$$

Proof. The result follows directly from the definition (6.7) and the model equation $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$:

$$\begin{aligned} \widehat{\mathcal{R}}_m - \mathcal{R}_m &= \bigl\| \mathcal{K}_{M,m} \boldsymbol{Y} \bigr\|^2 - I\!E \bigl\| \mathcal{K}_{M,m} \boldsymbol{Y} \bigr\|^2 \\ &= \bigl\| \boldsymbol{b}_{M,m} + \mathcal{K}_{M,m} \boldsymbol{\varepsilon} \bigr\|^2 - I\!E \bigl\| \boldsymbol{b}_{M,m} + \mathcal{K}_{M,m} \boldsymbol{\varepsilon} \bigr\|^2 . \end{aligned}$$

It is only to check that $I\!E\, \boldsymbol{\varepsilon}^\top A_{m,m^*} \boldsymbol{\varepsilon} = \mathtt{p}_{M,m} - \mathtt{p}_{M,m^*}$.

It becomes clear from this expansion that we have to bound a quadratic in $\boldsymbol{\varepsilon}$ term $\boldsymbol{\varepsilon}^\top A_{m,m^*}\boldsymbol{\varepsilon}$ and the linear term $2\boldsymbol{\varepsilon}^\top \boldsymbol{a}_{m,m^*}$. In the case of Gaussian errors $\boldsymbol{\varepsilon}$, we can use the general results for linear and quadratic forms of Gaussian vectors.

Lemma 6.4.3. *Let* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0,\Sigma)$. *Then it holds for any* $\mathtt{x} > 0$ *and for any symmetric matrix* A

$$\mathbb{P}\Bigl(\bigl|\boldsymbol{\varepsilon}^\top A\boldsymbol{\varepsilon} - \mathtt{p}(A)\bigr| \ge 2\mathsf{v}(A)\sqrt{\mathtt{x}} + 2\lambda(A)\mathtt{x}\Bigr) \le \mathrm{e}^{-\mathtt{x}},$$

where

$$\begin{aligned}\mathtt{p}(A) &\stackrel{\mathrm{def}}{=} \operatorname{tr}(A\Sigma),\\ \mathsf{v}^2(A) &\stackrel{\mathrm{def}}{=} \operatorname{tr}(A\Sigma)^2,\\ \lambda(A) &\stackrel{\mathrm{def}}{=} \|\Sigma^{1/2}A\Sigma^{1/2}\|_{\mathrm{op}}\,.\end{aligned}$$

Also, for any vector $\boldsymbol{a}$ *and any* $\mathtt{x} > 0$

$$\mathbb{P}\left(\bigl|\boldsymbol{a}^\top\boldsymbol{\varepsilon}\bigr| > \|\Sigma^{1/2}\boldsymbol{a}\|z_1(\mathtt{x})\right) \le \mathrm{e}^{-\mathtt{x}}.$$

Proof. Define $\boldsymbol{\gamma} = \Sigma^{-1/2}\boldsymbol{\varepsilon}$. Then $\boldsymbol{\gamma}$ is standard normal in $\mathbb{R}^n$, and with $B \stackrel{\mathrm{def}}{=} \Sigma^{1/2}A\Sigma^{1/2}$

$$\boldsymbol{\varepsilon}^\top A\boldsymbol{\varepsilon} = \boldsymbol{\gamma}^\top B\boldsymbol{\gamma}.$$

Now we can apply Theorem C.1.1 with $\operatorname{tr}(B) = \operatorname{tr}(A\Sigma)$, $\mathsf{v}^2 = \operatorname{tr}(B^2) = \operatorname{tr}(A\Sigma)^2$, $\lambda = \|B\|_{\mathrm{op}} = \|\Sigma^{1/2}A\Sigma^{1/2}\|_{\mathrm{op}}$.

Exercise 6.4.3. Let $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$. Restate the result of Lemma 6.4.3 for this case.

We can summarise that on a set of probability at least $1 - 2\mathrm{e}^{-\mathtt{x}}$, it holds

$$\bigl|\widehat{\mathcal{R}}_m - \widehat{\mathcal{R}}_{m^*} - (\mathcal{R}_m - \mathcal{R}_{m^*})\bigr| \le 2\sigma^2\mathsf{v}(A_{m,m^*})\sqrt{\mathtt{x}} + 2\sigma^2\lambda(A_{m,m^*})\mathtt{x} + 2\sigma\|\boldsymbol{a}_{m,m^*}\|z_1(\mathtt{x})$$

7

* Smallest accepted approach in ordered model selection for linear smoothers

7.1 Introduction

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. (1999), Massart (2007)). This is also the case in Lepski's method, Lepski (1990), Lepski (1991), Lepski (1992), Lepski and Spokoiny (1997), Lepski et al. (1997a), 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 the 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 the unbiased risk estimation method. One can also 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. (1999), Birgé and Massart (2007). 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 involve 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 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 paper 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. Our setup focuses on linear Gaussian regression and it 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. The procedure does not use any prior information about the variance structure of the noise, 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.

Consider a linear model $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ in $I\!R^{n}$ for an unknown parameter vector $\boldsymbol{\theta}^{*} \in I\!R^{p}$ and a given $p \times n$ design matrix Ψ. Suppose a family of linear smoothers $\widetilde{\boldsymbol{\theta}}_{m} = \mathcal{S}_{m}\boldsymbol{Y}$, $m \in \mathcal{M}$, to be fixed, where $\mathcal{M}$ is a set indexing the models considered and $\mathcal{S}_{m}$ 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 $\widehat{m}$ 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 a 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.* Unfortunately, the proposed approach requires the distribution of the errors $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \mathbb{E}\boldsymbol{Y}$ to be precisely known which is unrealistic in practical applications. Section 7.2.6 explains how the proposed procedure can be tuned in the case of Gaussian noise with unknown variance structure using a bootstrap method.

The paper presents a rigorous theoretical study of the proposed procedure for two cases. The first one corresponds to an idealistic situation that the noise distribution is precisely known; see Section 7.3.1. In particular, Theorem 7.3.1 presents finite sample results on the behavior of the proposed selector $\widehat{m}$ and the corresponding estimator $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{m}}$. It also describes a concentration set for the selected index $\widehat{m}$ and states a probabilistic oracle bound for the resulting estimator $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{m}}$. Usual rate results can be easily derived from these statements. Further results address the important quantity $\mathfrak{z}_{m^*}$ called “the payment for adaptation” which can be defined as the gap between oracle and adaptive bounds. Theorem 7.3.3 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 data-driven model selection; Lepski (1992). An extension of the obtained probabilistic bounds to the case of a polynomial loss function is given in Section A of the supplement [SW2016]. The results are also specified to the particular problems of projection and linear functional estimation.

All the obtained results will be extended to regression models with unknown heterogeneous Gaussian noise; Section 7.4.3. Our main results about model selection in Gaussian regression with unknown heterogeneous noise are based on Theorem 7.4.2 which provides a kind of “bootstrap validity” statement: the bootstrap distribution mimics the unknown error variance with explicit error terms which can be controlled under usual regularity assumptions. This allows to extend the results obtained for the case of a known error distribution to the bootstrap calibrated procedure.

The paper is structured as follows. Section 7.2.1 explains our setup of ordered model selection, then Section 7.2.3 and Section 7.2.4 link the proposed approach to the multiple testing problem. The formal definition of the procedure is given in Section 7.2.5 for known noise and in Section 7.2.6 for the case of Gaussian errors with unknown variance. Section 7.3 states the main results, Section 7.6 illustrates the performance of the methods by numerical examples, while the proofs are gathered in the Appendix. The proofs of some technical results as well as some useful bounds for Gaussian quadratic forms and sums of random matrices are collected in the supplement [SW2016].

7.2 SmA procedure

This section presents the proposed model selector. First we specify our setup.

7.2.1 Model and problem

Consider the following linear regression model:

$$Y_i = \Psi_i^\top \boldsymbol{\theta}^* + \varepsilon_i\,, \qquad I\!\!E\varepsilon_i = 0, \qquad i = 1, \ldots, n,$$

with given design $\Psi_1, \ldots, \Psi_n$ in $I\!\!R^p$. Below we assume a deterministic design, otherwise one can understand the results conditioned on the design. Further, $\boldsymbol{\theta}^*$ is an unknown vector in $I\!\!R^p$, and $\varepsilon_1, \ldots, \varepsilon_n$ are individual zero mean errors with finite variance. Our main results are stated under the assumption that individual errors $\varepsilon_i \stackrel{\text{def}}{=} Y_i - I\!\!E Y_i$ are independent normal and possibly heterogeneous, $I\!\!E\varepsilon_i^2 = \sigma_i^2$. However, Section 7.2.5 also discusses the case of an arbitrary but known error distribution.

The dimension p can be large, even $p = \infty$ can be incorporated. For notational simplicity, we proceed with p finite. The proposed approach can also be extended to the case when the linear parametric assumption $I\!\!E\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^*$ is not precisely fulfilled. Then, as usual, the target of estimation $\boldsymbol{\theta}^*$ can be defined as the vector of coefficients for the best approximation of the true response $\boldsymbol{f} \stackrel{\text{def}}{=} I\!\!E\boldsymbol{Y} = (f_1, \ldots, f_n)^\top$ by linear combinations of the feature vectors ψ_i which are the rows of the matrix Ψ. For the ease of notation, below we assume the linear parametric structure $f_i = \Psi_i^\top \boldsymbol{\theta}^*$. We write the underlying model in the vector form

$$\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\varepsilon} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}\,. \tag{7.1}$$

Let $\bigl\{\widetilde{\boldsymbol{\theta}}_m\,, m \in \mathcal{M}\bigr\}$ be a finite family of linear estimators $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$ of $\boldsymbol{\theta}^*$. 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 $q \ge 1$ and define quadratic loss and risk with this weighting matrix W:

$$\varrho_m \stackrel{\text{def}}{=} \|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2, \qquad \mathcal{R}_m \stackrel{\text{def}}{=} I\!\!E\|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2.$$

Alternatively one can say that

$$\widetilde{\boldsymbol{\phi}}_m \stackrel{\text{def}}{=} W\widetilde{\boldsymbol{\theta}}_m = W\widetilde{\boldsymbol{\theta}}_m = W\mathcal{S}_m\boldsymbol{Y} = \mathcal{K}_m\boldsymbol{Y}$$

with $\mathcal{K}_m = W\mathcal{S}_m$ is an estimator of the target $\boldsymbol{\phi}^* = W\boldsymbol{\theta}^*$ and $\varrho_m = \|\widetilde{\boldsymbol{\phi}}_m - \boldsymbol{\phi}^*\|^2$. Typical examples of W are as follows.

Estimation of the whole vector $\boldsymbol{\theta}^*$

Let W be the identity matrix $W = \boldsymbol{I}_p$ with $q = p$. This means that the *estimation loss* is measured by the usual squared Euclidean distance $\|\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\|^2$.

Prediction

Let W be the square root of the $p \times p$ matrix $\mathbb{F} = \Psi\Psi^\top$, that is, $W^2 = \mathbb{F}$. The loss $\|W(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \|\Psi(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|$ is usually referred to as *prediction loss* because it measures the prediction ability of the true model by the model with the parameter $\boldsymbol{\theta}$.

Semiparametric estimation

Suppose that the target of estimation is not the whole vector $\boldsymbol{\theta}^*$ but some subvector $\boldsymbol{\theta}_0^*$ of dimension q. The matrix W can be defined as the projector Π_0 on the $\boldsymbol{\theta}_0^*$ subspace. The estimate $\Pi_0\widetilde{\boldsymbol{\theta}}_m$ is called the *profile maximum likelihood estimate*. The corresponding loss is equal to the squared Euclidean distance in this subspace:

$$\varrho_m = \bigl\|\Pi_0\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\bigr)\bigr\|^2.$$

Alternatively, one can select W^2 as the efficient information matrix defined by relation $W^2 = \bigl(\Pi_0\mathbb{F}^{-1}\Pi_0^\top\bigr)^-$, where A^- means a pseudo-inverse of A.

Linear functional estimation

The choice of the weighting matrix W of rank one can be adjusted to address the problem of estimating some functionals of the whole parameter $\boldsymbol{\theta}^*$, for instance, the first coefficient θ_1^* or the sum of the θ_j^*'s. For instance, in the regression problem $I\!\!E Y_i = f(X_i)$ with the Fourier expansion the target function f can be represented as

$$f(x) = \sum_{j\ge 0} \theta_j^* \psi_j(x) = \sum_{j\ge 0} \bigl\{\theta_{2j}^* \cos(2\pi jx) + \theta_{2j+1}^* \sin(2\pi jx)\bigr\}.$$

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 $f'(0) = 2\pi \sum_{j\ge 0} j\theta_{2j+1}^*$.

In all cases, the most important feature of the estimators $\widetilde{\boldsymbol{\phi}}_m = \mathcal{K}_m\boldsymbol{Y}$ is *linearity*. It greatly simplifies the study of their properties including the prominent bias-variance decomposition of the risk of $\widetilde{\boldsymbol{\phi}}_m$. Namely, for the model (7.1) with $\mathbb{E}\boldsymbol{\varepsilon} = 0$ and $\boldsymbol{f} = \mathbb{E}\boldsymbol{Y}$, it holds

$$\begin{aligned} \mathbb{E}\widetilde{\boldsymbol{\phi}}_m &= \boldsymbol{\phi}_m^* \overset{\text{def}}{=} \mathcal{K}_m \boldsymbol{f}, \\ \mathcal{R}_m &= \|\boldsymbol{\phi}_m^* - \boldsymbol{\phi}^*\|^2 + \operatorname{tr}\bigl\{\mathcal{K}_m \operatorname{Var}(\boldsymbol{\varepsilon})\,\mathcal{K}_m^\top\bigr\} = \|\boldsymbol{b}_m\|^2 + \mathtt{p}_m , \end{aligned} \tag{7.2}$$

where $\|\boldsymbol{b}_m\|^2 \overset{\text{def}}{=} \|\boldsymbol{\phi}_m^* - \boldsymbol{\phi}^*\|^2$ is the squared bias term and $\mathtt{p}_m \overset{\text{def}}{=} \operatorname{tr}\operatorname{Var}(\widetilde{\boldsymbol{\phi}}_m)$ is the variance term. This is the usual "bias-variance" decomposition of the squared risk $\mathcal{R}_m$. The optimal choice of m is often defined by risk minimization:

$$m_{\text{opt}} \overset{\text{def}}{=} \operatorname*{argmin}_{m\in\mathcal{M}} \mathcal{R}_m = \operatorname*{argmin}_{m\in\mathcal{M}} \bigl(\|\boldsymbol{b}_m\|^2 + \mathtt{p}_m\bigr). \tag{7.3}$$

Alternatively one can define the best choice m^* via the "bias-variance trade-off"; see the definition below in (7.9). The *model selection* problem can be described as a data-based choice $\widehat{m}$ which leads to essentially the same quality of the adaptive estimator $\widetilde{\boldsymbol{\theta}}_{\widehat{m}}$ as for the optimal choice m^*.

7.2.2 Ordered case

Below we discuss the *ordered* case. For simplicity of presentation, we assume that $\mathcal{M}$ is a finite set of positive numbers, although the approach can be extended to situations with a countable and/or continuous and even unbounded set $\mathcal{M}$. $|\mathcal{M}|$ stands for the cardinality of $\mathcal{M}$. Typical examples of the parameter m are given by a chosen dimension (number of basis vectors) in projection estimation or by the bandwidth in kernel smoothing. In general, complexity can be naturally expressed via the variance of the stochastic term of the estimator $\widetilde{\boldsymbol{\phi}}_m$: the larger m, the larger is the variance term $\mathtt{p}_m = \operatorname{tr}\bigl\{\operatorname{Var}(\widetilde{\boldsymbol{\phi}}_m)\bigr\}$. In the case of projection estimation and a homogeneous noise $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$, this variance term is linear in m: $\mathtt{p}_m = \sigma^2 m$; see Section 7.3.3 for details. In general, dependence of the variance term on m is more complicated but monotonicity of $\mathtt{p}_m$ in m should be preserved. The related condition can be written as

$$\operatorname{tr}\bigl(\mathcal{K}_m \operatorname{Var}(\boldsymbol{\varepsilon})\,\mathcal{K}_m^\top\bigr) \le \operatorname{tr}\bigl(\mathcal{K}_{m'} \operatorname{Var}(\boldsymbol{\varepsilon})\,\mathcal{K}_{m'}^\top\bigr), \qquad m' > m, \tag{7.4}$$

where $A \ge B$ for two symmetric matrices A, B means $A - B$ positive semidefinite. Further, it is implicitly assumed that the bias term $\|\boldsymbol{b}_m\|^2 = \|\boldsymbol{\phi}_m^* - \boldsymbol{\phi}^*\|^2$ becomes smaller as m increases. The smallest index $m_0 \in \mathcal{M}$ corresponds to the simplest (zero) model, usually with a large bias, while a large m ensures a good approximation quality $\boldsymbol{\phi}_m^* \approx \boldsymbol{\phi}^*$ 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 (7.2) describes the accuracy of approximating the response $\boldsymbol{f}$ 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 squared bias $\|\boldsymbol{b}_m\|^2$ monotonously decreases with m. An example is given below.

Example 7.2.1. Suppose that a signal $\boldsymbol{\theta}^*$ is observed with noise: $Y_i = \theta_i^* + \varepsilon_i$. Consider the set of projection estimates $\widetilde{\boldsymbol{\theta}}_m$ on the first m coordinates and the target is $\phi^* \overset{\text{def}}{=} W\boldsymbol{\theta}^* = \sum_j \theta_j^*$. If $\boldsymbol{\theta}^*$ is composed of alternating blocks of 1's and -1's with equal length, then the bias $|\phi^* - \phi_m^*|$ for $\phi_m^* = \sum_{j \le m} \theta_j^*$ is not monotonous in m.

7.2.3 Smallest accepted (SmA) method in ordered model selection

This section presents the basics of the SmA procedure, in particular, relations to the multiple testing problem.

Suppose we are given an ordered set of linear estimators $\widetilde{\boldsymbol{\phi}}_m$ of the q-dimensional target of estimation $\boldsymbol{\phi}^* = W\boldsymbol{\theta}^*$, that is, $\widetilde{\boldsymbol{\phi}}_m = W\,\mathcal{S}_m\boldsymbol{Y} = \mathcal{K}_m\boldsymbol{Y}$ with $q \times n$ matrices $\mathcal{K}_m = W\mathcal{S}_m$ for $m \in \mathcal{M}$, and (7.4) holds . Below we present a general approach to model selection problems based on multiple testing. In the problem of choosing m, we face a usual dilemma: an increase of complexity of the method m yields an increase of the variance term but probably improves the approximation quality measured by the bias term $\|\boldsymbol{b}_m\|^2$. Thus, we aim at picking up a possibly small index $m^\circ \in \mathcal{M}$ for which a further increase of the index m over m° only increases the complexity of the method without real gain in the quality of approximation. The 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°. Introduce for each pair $m > m^\circ$ from $\mathcal{M}$ a hypothesis H_{m,m° of "no significant difference between the models m° and m"; see the next section for a precise formulation. Let τ_{m,m° be the corresponding test. The model m° is *accepted* if $\tau_{m,m^\circ} = 0$ for all $m > m^\circ$. This can be viewed a multiple test of the set of hypotheses $\mathcal{H}_{m^\circ} = \bigl\{H_{m,m^\circ}, m > m^\circ\bigr\}$. Finally, the selected model is the "smallest accepted":

$$\widehat{m} \overset{\text{def}}{=} \operatorname{argmin}\bigl\{m^\circ \in \mathcal{M}\colon \tau_{m,m^\circ} = 0, \, \forall m > m^\circ\bigr\}.$$

Usually the test τ_{m,m° can be written in the form

$$\tau_{m,m^\circ} = \mathbb{I}\bigl\{\mathbb{T}_{m,m^\circ} > \mathfrak{z}_{m,m^\circ}\bigr\} \tag{7.5}$$

for some *test statistics* $\mathbb{T}_{m,m^\circ}$ and for *critical values* $\mathfrak{z}_{m,m^\circ}$. The information-based criteria like AIC or BIC use the likelihood ratio test statistics $\mathbb{T}_{m,m^\circ} = \sigma^{-2}\|\Psi^\top(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ})\|^2$. A great advantage of such tests is that the test statistic $\mathbb{T}_{m,m^\circ}$ is pivotal (χ^2 with $m - m^\circ$

degrees of freedom) under the null hypothesis $I\!\!E\widetilde{\boldsymbol{\theta}}_m = I\!\!E\widetilde{\boldsymbol{\theta}}_{m^\circ}$ and homogeneous Gaussian noise with known variance σ^2, this makes it simple to compute the corresponding critical values. However, under more general assumptions on the noise distribution, and it is more convenient to apply another choice corresponding to Lepski-type procedure and based on the norm of differences $\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}$:

$$\mathbb{T}_{m,m^\circ} = \|\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}\| = \|\mathcal{K}_m \boldsymbol{Y} - \mathcal{K}_{m^\circ}\boldsymbol{Y}\| = \|\mathcal{K}_{m,m^\circ}\boldsymbol{Y}\|,$$

where $\mathcal{K}_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{K}_m - \mathcal{K}_{m^\circ}$. The main issue for such a method is a proper choice of the critical values $\mathfrak{z}_{m,m^\circ}$ in (7.5). One can say that the procedure is specified by a way of selecting these critical values. Below we fix these values by imposing a so-called *propagation property*: a "good" model m° for which all H_{m,m° with $m > m^\circ$ are true, has to be accepted with a high probability. This rule can be seen as an analogue of the family-wise error rate condition in a multiple testing problem.

7.2.4 A "good" model

This section aims at formalizing the above mentioned relations between model selection and multiple testing. We use below for each pair $m > m^\circ$ from $\mathcal{M}$ the decomposition of the test statistic $\mathbb{T}_{m,m^\circ}$

$$\begin{aligned}\mathbb{T}_{m,m^\circ} &= \|\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}\| = \|\mathcal{K}_{m,m^\circ}\boldsymbol{Y}\| \\ &= \|\mathcal{K}_{m,m^\circ}(\boldsymbol{f}+\boldsymbol{\varepsilon})\| = \|\boldsymbol{b}_{m,m^\circ} + \boldsymbol{\xi}_{m,m^\circ}\|, \qquad (7.6)\end{aligned}$$

with $\mathcal{K}_{m,m^\circ} = \mathcal{K}_m - \mathcal{K}_{m^\circ}$, where $\boldsymbol{b}_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{K}_{m,m^\circ}\boldsymbol{f} \in I\!\!R^q$ is the deterministic *bias* vector, while $\boldsymbol{\xi}_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{K}_{m,m^\circ}\boldsymbol{\varepsilon} \in I\!\!R^q$ is the *stochastic* component. It obviously holds $I\!\!E\boldsymbol{\xi}_{m,m^\circ} = 0$. Introduce the $q \times q$-matrix $\mathbb{V}_{m,m^\circ}$ as the variance of $\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ} = \mathcal{K}_{m,m^\circ}\boldsymbol{Y}$:

$$\mathbb{V}_{m,m^\circ} \stackrel{\text{def}}{=} \operatorname{Var}\bigl(\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}\bigr) = \operatorname{Var}\bigl(\mathcal{K}_{m,m^\circ}\boldsymbol{Y}\bigr) = \mathcal{K}_{m,m^\circ}\operatorname{Var}(\boldsymbol{\varepsilon})\,\mathcal{K}_{m,m^\circ}^\top.$$

If the noise $\boldsymbol{\varepsilon}$ is homogeneous with $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$, it holds

$$\mathbb{V}_{m,m^\circ} = \sigma^2\,\mathcal{K}_{m,m^\circ}\,\mathcal{K}_{m,m^\circ}^\top.$$

Further,

$$\begin{aligned} I\!\!E\,\mathbb{T}_{m,m^\circ}^2 &= \|\boldsymbol{b}_{m,m^\circ}\|^2 + I\!\!E\|\boldsymbol{\xi}_{m,m^\circ}\|^2 = \|\boldsymbol{b}_{m,m^\circ}\|^2 + \mathtt{p}_{m,m^\circ}, \qquad (7.7) \\ \mathtt{p}_{m,m^\circ} &\stackrel{\text{def}}{=} I\!\!E\|\boldsymbol{\xi}_{m,m^\circ}\|^2 = \operatorname{tr}(\mathbb{V}_{m,m^\circ}).\end{aligned}$$

For a fixed $m^\circ \in \mathcal{M}$, let

$$\mathcal{M}^{+}(m^{\circ}) \stackrel{\text{def}}{=} \bigl\{ m \in \mathcal{M} \colon m > m^{\circ} \bigr\}.$$

A "good" choice m° can be defined by the condition that, for each $m \in \mathcal{M}^{+}(m^{\circ})$, the bias term $\|\boldsymbol{b}_{m,m^{\circ}}\|^2$ is not significantly larger than the variance term $\mathtt{p}_{m,m^{\circ}}$. This condition can be quantified in the following "bias-variance trade-off" relation:

$$H_{m,m^{\circ}} \colon \quad \|\boldsymbol{b}_{m,m^{\circ}}\|^2 \le \beta^2 \, \mathtt{p}_{m,m^{\circ}} \, ,$$

with a given parameter β. This can be viewed as a null hypothesis of "no significant difference" between models with parameters m° and m. For each candidate model m°, define a set of hypotheses

$$\mathcal{H}_{m^{\circ}} = \bigl\{ H_{m,m^{\circ}} \colon \|\boldsymbol{b}_{m,m^{\circ}}\|^2 \le \beta^2 \, \mathtt{p}_{m,m^{\circ}} \, , \quad m \in \mathcal{M}^{+}(m^{\circ}) \bigr\}. \tag{7.8}$$

A "good" model m° is one with all hypotheses in this set $\mathcal{H}_{m^{\circ}}$ fulfilled. Below this set of hypotheses will be considered for each m° separately. Now define the *oracle* m^{*} as the minimal m° under (7.8):

$$m^{*} \stackrel{\text{def}}{=} \min\Bigl\{ m^{\circ} \colon \max_{m \in \mathcal{M}^{+}(m^{\circ})} \bigl\{ \|\boldsymbol{b}_{m,m^{\circ}}\|^2 - \beta^2 \, \mathtt{p}_{m,m^{\circ}} \bigr\} \le 0 \Bigr\}. \tag{7.9}$$

Clearly the notion of a "good" model depends on the value β, in particular, $m^{*} = m^{*}(\beta)$. Also m^{*} does not coincide with the risk minimizer m_{opt} from (7.3). However, both definitions exhibit bias-variance trade-off in (7.7).

7.2.5 Calibration of the SmA procedure for the known noise distribution

This section explains the choice of the critical values $\mathfrak{z}_{m,m^{\circ}}$ for the idealistic case when the noise distribution is precisely known. This greatly helps to explain the essence of the approach. Section 7.2.6 presents a data-driven procedure for the unknown noise variance using a resampling technique.

For a fixed m°, the related set of critical values $\mathfrak{z}_{m,m^{\circ}}$ should be fixed to ensure a prescribed family-wise error rate (FWER) $\mathrm{e}^{-\mathtt{x}}$ of the family of tests $\mathbb{1}\bigl(\mathbb{T}_{m,m^{\circ}} > \mathfrak{z}_{m,m^{\circ}}\bigr)$ for $m \in \mathcal{M}$ with $m > m^{\circ}$. In the terminology of Romano and Wolf (2005), this is a *weak* FWER control.

Let us start with $\beta = 0$ corresponding to the set $\mathcal{H}_{m^{\circ}}$ of hypotheses $H_{m,m^{\circ}} \colon \boldsymbol{b}_{m,m^{\circ}} = 0$ for all $m > m^{\circ}$. In this situation, the test statistic $\mathbb{T}_{m,m^{\circ}}$ coincides under $H_{m,m^{\circ}}$ with the norm of the stochastic term $\boldsymbol{\xi}_{m,m^{\circ}}$ whose distribution is precisely known under given noise. For instance, if errors $\boldsymbol{\varepsilon}$ are Gaussian, then the stochastic component $\boldsymbol{\xi}_{m,m^{\circ}}$ is a normal zero mean vector with the covariance matrix $\mathbb{V}_{m,m^{\circ}}$. Introduce for each pair $m > m^{\circ}$ from $\mathcal{M}$ a *tail function* $z_{m,m^{\circ}}(t)$ of the argument t such that

$$I\!P\Bigl(\|\boldsymbol{\xi}_{m,m^{\circ}}\| > z_{m,m^{\circ}}(t)\Bigr) = \mathrm{e}^{-t}. \tag{7.10}$$

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 deviation probability is smaller than e^{-t}. For multiple testing, we need a uniform in $m > m^{\circ}$ version of the probability bound (7.10). To guarantee the prescribed FWER for the set of hypotheses $\mathcal{H}_{m^{\circ}}$, introduce, given $\mathtt{x}$, the multiplicity correction $q_{m^{\circ}} = q_{m^{\circ}}(\mathtt{x})$:

$$I\!P\Bigl(\bigcup_{m\in\mathcal{M}^{+}(m^{\circ})} \bigl\{\|\boldsymbol{\xi}_{m,m^{\circ}}\| \ge z_{m,m^{\circ}}(\mathtt{x}+q_{m^{\circ}})\bigr\}\Bigr) \le \mathrm{e}^{-\mathtt{x}}. \tag{7.11}$$

A simple way of fixing the value $q_{m^{\circ}}$ is based on the Bonferroni bound: $q_{m^{\circ}} = \log(|\mathcal{M}^{+}(m^{\circ})|)$; cf. Spokoiny (1996) in context of adaptive testing. However, it is well known that the Bonferroni correction is very conservative and results in a large $q_{m^{\circ}}$; see e.g. Baraud et al. (2003). This is especially striking if the random vectors $\boldsymbol{\xi}_{m,m^{\circ}}$ are strongly correlated, which is exactly the case under consideration. As the joint distribution of the $\boldsymbol{\xi}_{m,m^{\circ}}$'s is precisely known, one can define the correction $q_{m^{\circ}}$ just as the smallest value ensuring (7.11); cf. (5) in Baraud et al. (2003). This choice $z_{m,m^{\circ}}(\mathtt{x}+q_{m^{\circ}})$ of the critical values yields automatically the weak FWER bound for the set of hypotheses $\mathcal{H}_{m^{\circ}} = \bigl\{H_{m,m^{\circ}}, m > m^{\circ}\bigr\}$ with $\beta = 0$. Moreover, the FWER control would fail for any other uniformly smaller set of critical values.

In the case of β positive, we define the critical values $\mathfrak{z}_{m,m^{\circ}} = \mathfrak{z}_{m,m^{\circ}}(\mathtt{x})$ by one more correction for the bias term $\|\boldsymbol{b}_{m,m^{\circ}}\|$:

$$\mathfrak{z}_{m,m^{\circ}} \overset{\mathrm{def}}{=} z_{m,m^{\circ}}(\mathtt{x}+q_{m^{\circ}}) + \beta\sqrt{\mathtt{p}_{m,m^{\circ}}} \tag{7.12}$$

for $\mathtt{p}_{m,m^{\circ}} = \operatorname{tr}(\mathbb{V}_{m,m^{\circ}})$. The bound (7.11) automatically ensures the desired *propagation property*: any good model m° in the sense (7.8) will be *rejected* with probability at most $\mathrm{e}^{-\mathtt{x}}$ in the following sense:

$$\begin{aligned} I\!P\bigl(m^{\circ} \text{ is rejected}\bigr) &\overset{\mathrm{def}}{=} I\!P\Bigl(\bigcup_{m\in\mathcal{M}^{+}(m^{\circ})} \bigl\{\|\mathbb{T}_{m,m^{\circ}}\| \ge \mathfrak{z}_{m,m^{\circ}}(\mathtt{x})\bigr\}\Bigr) \\ &\le I\!P\Bigl(\bigcup_{m\in\mathcal{M}^{+}(m^{\circ})} \bigl\{\|\boldsymbol{\xi}_{m,m^{\circ}}\| \ge z_{m,m^{\circ}}(\mathtt{x}+q_{m^{\circ}})\bigr\}\Bigr) \le \mathrm{e}^{-\mathtt{x}}. \end{aligned} \tag{7.13}$$

The last inequality follows from (7.6). One can say, this is a built-in property of the procedure. By definition, the oracle m^{*} is also the smallest "good" choice, this yields due to (7.13)

$$I\!P\bigl(m^{*} \text{ is rejected}\bigr) \le \mathrm{e}^{-\mathtt{x}}. \tag{7.14}$$

Definition (7.12) still involves two numerical constants $\mathtt{x}$ and β. It is quite common in the model selection literature to define the optimal choice of tuning parameters by minimization of the risk of the resulting procedure. Unfortunately, it does not apply in our setup which is based on multiple testing. Note however, that these values are not tuning parameters of the method, they rather serve to fix some expected features of the method. The value $\mathtt{x}$ defines the nominal FWER $\mathrm{e}^{-\mathtt{x}}$. Similarly to the testing problem, there is no unique choice for $\mathtt{x}$, a usual choice of $\mathtt{x}$ in the range between 2 and 3 can be recommended. The value β controls the amount of admissible bias in the definition of a good model; cf. (7.8) and (7.9). The natural choice for β is $\beta = 1$ which balances the bias and variance terms in (7.8). Note, however, that the procedure and the theoretical results hold for any combination of these parameters. We only require that the value β is the same in the definition of a good model and in the formula (7.12) for the critical values $\mathfrak{z}_{m,m^{\circ}}$. Our default choice is $\mathtt{x} = 2$, $\beta = 1$. An intensive numerical study indicates a very minor change in the estimation results for moderate deviations of these parameters around the mentioned default choice.

Define the selector $\widehat{m}$ by the "smallest accepted" (SmA) rule. Namely, with $\mathfrak{z}_{m,m^{\circ}}$ from (7.12), the acceptance rule reads as follows:

$$\bigl\{ m^{\circ} \text{ is accepted} \bigr\} = \Bigl\{ \max_{m \in \mathcal{M}^{+}(m^{\circ})} \bigl\{ \mathbb{T}_{m,m^{\circ}} - \mathfrak{z}_{m,m^{\circ}} \bigr\} \le 0 \Bigr\}. \tag{7.15}$$

The SmA choice is defined by the "smallest accepted" rule:

$$\widehat{m} \stackrel{\text{def}}{=} \min \Bigl\{ m^{\circ} \colon \max_{m \in \mathcal{M}^{+}(m^{\circ})} \bigl\{ \mathbb{T}_{m,m^{\circ}} - \mathfrak{z}_{m,m^{\circ}} \bigr\} \le 0 \Bigr\}. \tag{7.16}$$

Our study mainly focuses on the behavior of the selector $\widehat{m}$. The performance of the resulting estimator $\widehat{\boldsymbol{\phi}} = \widetilde{\boldsymbol{\phi}}_{\widehat{m}}$ is a kind of corollary from the statements about $\widehat{m}$. The desired solution would be $\widehat{m} \equiv m^{*}$, then the adaptive estimator $\widehat{\boldsymbol{\phi}}$ coincides with the oracle estimator $\widetilde{\boldsymbol{\phi}}_{m^{*}}$.

Remark 7.2.1. The SmA procedure originates from Lepski (1990). However, Lepski's acceptance rule for a candidate m° is a bit stronger: it requires that each larger model $m > m^{\circ}$ is accepted as well, that is, all hypotheses $H_{m',m}$ for $m' > m \ge m^{\circ}$ are accepted. This allows to efficiently implement the procedure as a top-down algorithm: start from the largest model index m and check acceptance by the criterion (7.15) until rejection. Our acceptance rule is similar to Birgé (2001) and it can be implemented as a bottom-up algorithm: start from the smallest model and check each new candidate m° by rule (7.15) until the first acceptance. Note that the way of computing the critical values by multiplicity arguments can be used for the original Lepski's rule as well. It however requires an additional correction due to the more strict acceptance rule. More precisely, define for each t and each m° the correction $q_{m^{\circ}}(t)$ similarly to (7.11):

$$\mathbb{P}\Bigl(\bigcup_{m \in \mathcal{M}^{+}(m^{\circ})} \Bigl\{ \| \boldsymbol{\xi}_{m,m^{\circ}} \| \geq z_{m,m^{\circ}}(t + q_{m^{\circ}}(t)) \Bigr\} \Bigr) \leq \mathrm{e}^{-t}.$$

Further, given $\mathtt{x}$, define an additional correction $q^{+} = q^{+}(\mathtt{x})$ by

$$\mathbb{P}\Bigl(\bigcup_{m^{\circ} \in \mathcal{M}} \bigcup_{m \in \mathcal{M}^{+}(m^{\circ})} \Bigl\{ \| \boldsymbol{\xi}_{m,m^{\circ}} \| \geq z_{m^{\circ},m}(\mathtt{x} + q_{m^{\circ}}(\mathtt{x}) + q^{+}) \Bigr\} \Bigr) \leq \mathrm{e}^{-\mathtt{x}}. \tag{7.17}$$

Finally define the critical values $\mathfrak{z}^{L}_{m,m^{\circ}} = \mathfrak{z}^{L}_{m,m^{\circ}}(\mathtt{x})$ in the form

$$\mathfrak{z}^{L}_{m,m^{\circ}} = z_{m,m^{\circ}}\bigl(\mathtt{x} + q_{m^{\circ}}(\mathtt{x}) + q^{+}\bigr) + \beta \sqrt{\mathtt{p}_{m,m^{\circ}}}.$$

Again, this construction allows to build a set of critical values which guarantees the propagation property for Lepski's procedure. The correction (7.17) can be viewed as a special case of a classical proposal for simultaneous structured testing; see e.g. Marcus et al. (1976) or Romano and Wolf (2005) and of a sequential rejection principle from **?**.

7.2.6 Bootstrap tuning

This section explains how the proposed SmA procedure can be applied in the case of Gaussian *heterogeneous* noise with *unknown* covariance matrix $\Sigma = \mathrm{Var}(\boldsymbol{\varepsilon}) = \mathrm{diag}\bigl(\sigma_1^2, \ldots, \sigma_n^2\bigr)$. Let the observed data $\boldsymbol{Y}$ follow the model $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$. Assume to be given an ordered family of linear estimators $\widetilde{\boldsymbol{\phi}}_m = \mathcal{K}_m \boldsymbol{Y} = W \mathcal{S}_m \boldsymbol{Y}$ of the target $\boldsymbol{\phi}^{*} = W\boldsymbol{\theta}^{*}$, $m \in \mathcal{M}$. For each pair $m > m^{\circ}$ from $\mathcal{M}$, we consider the test statistic $\mathbb{T}_{m,m^{\circ}}$ and its decomposition from (7.6):

$$\mathbb{T}_{m,m^{\circ}} = \| \widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^{\circ}} \| = \| \boldsymbol{b}_{m,m^{\circ}} + \boldsymbol{\xi}_{m,m^{\circ}} \|.$$

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 bootstrap procedure to approximate this distribution. The proposed procedure relates to the concept of the *wild* bootstrap, Wu (1986), Beran (1986), or **?**. 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 weighting schemes see, for example, Mammen (1993).

Suppose we are given a pilot estimator (presmoothing) $\widetilde{\boldsymbol{f}}$ of the response vector $\boldsymbol{f} = \mathbb{E}\boldsymbol{Y} \in \mathbb{R}^{n}$. Define the residuals:

$$\breve{\boldsymbol{Y}} \stackrel{\mathrm{def}}{=} \boldsymbol{Y} - \widetilde{\boldsymbol{f}}.$$

About this pilot it is supposed that the related bias is negligible and the variance of $\breve{\boldsymbol{Y}}$ is close to Σ. This presmoothing assumes some minimal regularity of the response $\boldsymbol{f}$ (usually expressed via minimal smoothness of the underlying 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 $\widetilde{\boldsymbol{f}}$ is a linear predictor, $\widetilde{\boldsymbol{f}} = \Pi \boldsymbol{Y}$, where Π is a sub-projector in the space $\mathbb{R}^n$. For example, one can take $\Pi = \Pi_{m^\dagger}$, where $\Pi_m = \Psi_m^\top \bigl(\Psi_m \Psi_m^\top \bigr)^- \Psi_m$, Ψ_m is the rank m feature matrix corresponding to the first m features, and $m^\dagger \in \mathcal{M}$ corresponds to a model with a possibly small bias, e.g. the largest model M in our collection $\mathcal{M}$. The wild bootstrap proposes to resample from the heteroscedastic Gaussian noise with the covariance matrix

$$\breve{\Sigma} = \operatorname{diag}(\breve{\boldsymbol{Y}} \cdot \breve{\boldsymbol{Y}}) = \operatorname{diag}\bigl(\breve{Y}_1^2, \ldots, \breve{Y}_n^2 \bigr),$$

where $\breve{\boldsymbol{Y}} \cdot \breve{\boldsymbol{Y}}$ denotes the coordinate-wise product of the vector $\breve{\boldsymbol{Y}}$ with itself and $\operatorname{diag}(\breve{\boldsymbol{Y}} \cdot \breve{\boldsymbol{Y}})$ denotes the diagonal matrix with entries $\breve{Y}_i^2$. These entries depend on $\boldsymbol{Y}$ and thus are random. Therefore, the bootstrap distribution 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 $\boldsymbol{Y}$. Clearly $\breve{\Sigma} = \operatorname{diag}(\breve{\boldsymbol{Y}} \cdot \breve{\boldsymbol{Y}})$ is a very poor estimator of Σ. However, under realistic conditions on the pilot $\widetilde{\boldsymbol{f}}$ and on the model, it allows to obtain essentially the same results as in the case of known Σ.

Let $\boldsymbol{w}^\flat$ denote the n-vector of bootstrap standard Gaussian weights, $\boldsymbol{w}^\flat \sim \mathcal{N}(0, \boldsymbol{I}_n)$. Clearly the product $\boldsymbol{\varepsilon}^\flat = \operatorname{diag}(\breve{\boldsymbol{Y}}) \boldsymbol{w}^\flat$ is conditionally on $\boldsymbol{Y}$ normal zero mean:

$$\boldsymbol{\varepsilon}^\flat = \operatorname{diag}(\breve{\boldsymbol{Y}}) \boldsymbol{w}^\flat \bigm|_{\boldsymbol{Y}} \sim \mathbb{P}^\flat \stackrel{\text{def}}{=} \mathcal{N}(0, \breve{\Sigma}).$$

The bootstrap analogue of $\boldsymbol{\xi}_{m,m^\circ} = \mathcal{K}_{m,m^\circ} \boldsymbol{\varepsilon}$ reads

$$\boldsymbol{\xi}_{m,m^\circ}^\flat = \mathcal{K}_{m,m^\circ} \boldsymbol{\varepsilon}^\flat = \mathcal{K}_{m,m^\circ} \operatorname{diag}(\breve{\boldsymbol{Y}}) \boldsymbol{w}^\flat . \tag{7.19}$$

The idea is to calibrate the SmA procedure under the bootstrap measure $\mathbb{P}^\flat$ using $\|\boldsymbol{\xi}_{m,m^\circ}^\flat\|$ in place of $\|\boldsymbol{\xi}_{m,m^\circ}\|$. The bootstrap quantiles $z_{m,m^\circ}^\flat(t)$ are given by the analogue of (7.10):

$$\mathbb{P}^\flat \Bigl(\|\boldsymbol{\xi}_{m,m^\circ}^\flat\| \ge z_{m,m^\circ}^\flat(t) \Bigr) = \mathrm{e}^{-t}. \tag{7.20}$$

The use of a continuous distribution for the bootstrap weights $w_i^\flat$ allows to uniquely define the values $z_{m,m^\circ}^\flat(t)$. If a discrete distribution of the weights is used, then, as usual, $z_{m,m^\circ}^\flat(t)$ is the minimal value for which the probability in the left hand-side of (7.20) does not exceed e^{-t}. The multiplicity correction $q_{m^\circ}^\flat = q_{m^\circ}^\flat(\mathtt{x})$ is specified by the condition

$$
I\!P^{\flat}\Biggl(\bigcup_{m \in \mathcal{M}^{+}(m^{\circ})} \bigl\{ \| \boldsymbol{\xi}^{\flat}_{m,m^{\circ}} \| \ge z^{\flat}_{m,m^{\circ}}(\mathtt{x} + q^{\flat}_{m^{\circ}}) \bigr\} \Biggr) = \mathrm{e}^{-\mathtt{x}}. \tag{7.21}
$$

Finally, the bootstrap critical values are fixed by the analogue of (7.12):

$$
\mathfrak{z}^{\flat}_{m,m^{\circ}} \stackrel{\mathrm{def}}{=} z^{\flat}_{m,m^{\circ}}(\mathtt{x} + q^{\flat}_{m^{\circ}}) + \beta^{\flat} \sqrt{\mathtt{p}^{\flat}_{m,m^{\circ}}} \, , \tag{7.22}
$$

where $\beta^{\flat}$ is a given positive constant and $\mathtt{p}^{\flat}_{m,m^{\circ}} = I\!E^{\flat} \| \boldsymbol{\xi}^{\flat}_{m,m^{\circ}} \|^{2}$ is the conditional expectation of $\| \boldsymbol{\xi}^{\flat}_{m,m^{\circ}} \|^{2}$ w.r.t. the bootstrap measure:

$$
\mathtt{p}^{\flat}_{m,m^{\circ}} \stackrel{\mathrm{def}}{=} \operatorname{tr}\bigl\{ \mathcal{K}^{\top}_{m,m^{\circ}} \operatorname{diag}(\check{\boldsymbol{Y}} \cdot \check{\boldsymbol{Y}}) \, \mathcal{K}_{m,m^{\circ}} \bigr\}.
$$

Now we apply the SmA procedure (7.16) with the data-driven critical values $\mathfrak{z}^{\flat}_{m,m^{\circ}}$ from (7.22).

7.3 Theoretical properties

This section contains the main theoretical properties of the proposed SmA procedure. We start again from the case of known noise. Then the results are extended to the bootstrap procedure.

7.3.1 Known noise

Let m^{*} be the oracle choice from (7.9), and let $\widehat{m}$ be the SmA selector from (7.16). Our study focuses on the properties of $\widehat{m}$. As a byproduct, we describe some oracle bounds on the loss of the corresponding adaptive procedure $\widehat{\boldsymbol{\phi}} = \widetilde{\boldsymbol{\phi}}_{\widehat{m}}$. The construction of $\widehat{m}$ ensures that the oracle m^{*} is accepted with high probability; see (7.14). Therefore, the selector $\widehat{m}$ with probability at least $1 - \mathrm{e}^{-\mathtt{x}}$ takes its value in the set

$$
\mathcal{M}^{-} = \mathcal{M}^{-}(m^{*}) = \bigl\{ m \in \mathcal{M} \colon \; m \le m^{*} \bigr\}
$$

of all models in $\mathcal{M}$ not greater than m^{*}. It remains to check the performance of the method in this region. The next step is to specify a subset of $\mathcal{M}^{-}$ which contains $\widehat{m}$-values with a high probability. By definition (7.9), m^{*} is the smallest index for which the bias terms $\| \boldsymbol{b}_{m,m^{\circ}} \|$ are uniformly bounded by $\beta \, \mathtt{p}^{1/2}_{m,m^{\circ}}$, $m > m^{\circ}$. Therefore, for each $m^{\circ} < m^{*}$, there is at least one $m > m^{\circ}$ with $\| \boldsymbol{b}_{m,m^{\circ}} \| > \beta \, \mathtt{p}^{1/2}_{m,m^{\circ}}$. The next result shows that the test $\tau_{m,m^{\circ}}$ based on $\mathbb{T}_{m,m^{\circ}}$ rejects $H_{m,m^{\circ}}$ with high probability if the condition $\| \boldsymbol{b}_{m,m^{\circ}} \| \le \beta \, \mathtt{p}^{1/2}_{m,m^{\circ}}$ is significantly violated (the "large bias" case). This observation allows us to describe the so called *zone of insensitivity* $\mathcal{M}_{\mathrm{in}}$, where $\widehat{m}$ concentrates. The results in this subsection hold for a general noise distribution with zero mean and finite variance. In the subsequent subsections we will then again assume Gaussianity of the errors.

Theorem 7.3.1. *For the linear model* $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ *with arbitrary but known distribution of* $\boldsymbol{\varepsilon}$ *, suppose to be given a family of smoothers* $\widetilde{\boldsymbol{\phi}}_{m} = \mathcal{K}_{m}\boldsymbol{Y}$ *,* $m \in \mathcal{M}$ *, ordered by their variance due to* (7.4)*. Let* $z_{m,m^{\circ}}(\cdot)$ *be the tail function from* (7.10) *for each pair* $m > m^{\circ} \in \mathcal{M}$ *. Given* $\mathtt{x}$ *and* β *, let* $\mathfrak{z}_{m,m^{\circ}}$ *be due to* (7.11) *and* (7.12)*, and let the oracle* m^{*} *be defined in* (7.9)*. Then the property* (7.14) *is fulfilled for the SmA rule* $\widehat{m}$ *. Let also* $\mathcal{M}_{\boldsymbol{b}}^{-}$ *be the subset of* $\mathcal{M}^{-}$ *defined by the "large bias" condition:*

$$\mathcal{M}_{\boldsymbol{b}}^{-} \stackrel{\text{def}}{=} \bigl\{ m \in \mathcal{M}^{-} : \|\boldsymbol{b}_{m^{*},m}\| > \mathfrak{z}_{m^{*},m} + z_{m^{*},m}(\bar{\mathtt{x}}) \bigr\},$$

where $\bar{\mathtt{x}} \stackrel{\text{def}}{=} \mathtt{x} + \log|\mathcal{M}^{-}|$ *. Then it holds with* $\mathcal{M}_{\text{in}} \stackrel{\text{def}}{=} \mathcal{M}^{-} \setminus \mathcal{M}_{\boldsymbol{b}}^{-}$

$$I\!\!P\bigl(\widehat{m} \in \mathcal{M}_{\text{in}}\bigr) \geq 1 - 2\mathrm{e}^{-\mathtt{x}}.$$

Remark 7.3.1. The *set of insensitivity* $\mathcal{M}_{\text{in}} = \mathcal{M}^{-} \setminus \mathcal{M}_{\boldsymbol{b}}^{-}$ contains all indices $m^{\circ} < m^{*}$ for which the squared bias $\|\boldsymbol{b}_{m,m^{\circ}}\|^{2}$ exceeds at some point $m > m^{\circ}$ the value $\beta^{2}\,\mathtt{p}_{m,m^{\circ}}$ but not essentially. Therefore, we cannot guarantee that the related test $\tau_{m,m^{\circ}}$ is powerful. The worst case setup corresponds to a flat bias profile with $\|\boldsymbol{b}_{m,m^{\circ}}\| \approx \beta\,\mathtt{p}_{m,m^{\circ}}^{1/2}$. Then the set of insensitivity $\mathcal{M}_{\text{in}}$ can coincide with the whole range $\mathcal{M}^{-}$.

The next result describes the properties of the SmA estimator $\widehat{\boldsymbol{\phi}} = \widetilde{\boldsymbol{\phi}}_{\widehat{m}}$.

Theorem 7.3.2. *Under conditions of Theorem 7.3.1, the SmA estimator* $\widehat{\boldsymbol{\phi}} = \widetilde{\boldsymbol{\phi}}_{\widehat{m}}$ *satisfies the following bound:*

$$I\!\!P\Bigl(\bigl\|\widehat{\boldsymbol{\phi}} - \widetilde{\boldsymbol{\phi}}_{m^{*}}\bigr\| > \bar{\mathfrak{z}}_{m^{*}}\Bigr) \leq 2\mathrm{e}^{-\mathtt{x}}\,, \tag{7.23}$$

where $\bar{\mathfrak{z}}_{m^{*}}$ *is defined as*

$$\bar{\mathfrak{z}}_{m^{*}} \stackrel{\text{def}}{=} \max_{m \in \mathcal{M}_{\text{in}}} \mathfrak{z}_{m^{*},m}\,. \tag{7.24}$$

This implies the probabilistic oracle bound: with probability at least $1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\bigl\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^{*}\bigr\| \leq \bigl\|\widetilde{\boldsymbol{\phi}}_{m^{*}} - \boldsymbol{\phi}^{*}\bigr\| + \bar{\mathfrak{z}}_{m^{*}}. \tag{7.25}$$

Remark 7.3.2. The result (7.25) is called the *oracle bound* because it compares the loss of the data-driven selector $\widehat{m}$ and of the oracle choice m^{*} . The discrepancy $\bar{\mathfrak{z}}_{m^{*}}$ in (7.23) or (7.25) can be viewed as a price for a data-driven model choice or "payment for adaptation"; cf. Lepski et al. (1997b). 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 $\boldsymbol{f}$ and the corresponding oracle choice m^{*}. The bound (7.25) is nearly sharp if the value $\bar{\mathfrak{z}}_{m^{*}}$ is smaller in order than $p_{m^{*}}^{1/2}$.

Remark 7.3.3. The usual Lepski's risk upper bound is very similar to (7.25); cf. Lepski et al. (1997b). However, the related "payment for adaptation" $\mathfrak{z}$ is evaluated by rather crude Bonferroni type arguments for the worst case, and it can be significantly larger than $\bar{\mathfrak{z}}_{m^{*}}$ from (7.24).

The procedure and the results can be extended to the case of polynomial loss, see Section A in the supplement [SW2016].

7.3.2 Analysis of the payment for adaptation $\bar{\mathfrak{z}}_{m^{*}}$

We now return to the setting of Gaussian errors ε_{i}. The benefit of considering the Gaussian case is that each vector $\boldsymbol{\xi}_{m,m^{\circ}}$ is Gaussian as well, which simplifies the analysis of the tail function $z_{m,m^{\circ}}(\cdot)$. The bounds can be easily extended to sub-Gaussian errors.

With $\mathbb{V}_{m} \overset{\text{def}}{=} \operatorname{Var}\bigl(\widetilde{\boldsymbol{\phi}}_{m}\bigr) = \mathcal{K}_{m} \operatorname{Var}(\boldsymbol{\varepsilon})\,\mathcal{K}_{m}^{\top}$, denote for $m^{\circ} < m$

$$\mathtt{p}_{m} = \operatorname{tr}(\mathbb{V}_{m}), \qquad \lambda_{m} = \|\mathbb{V}_{m}\|_{\mathrm{op}}\,,$$

$$\mathtt{p}_{m,m^{\circ}} = \operatorname{tr}(\mathbb{V}_{m,m^{\circ}}), \qquad \lambda_{m,m^{\circ}} = \|\mathbb{V}_{m,m^{\circ}}\|_{\mathrm{op}}\,.$$

Theorem 7.3.3. *Let the conditions of Theorem 7.3.1 be fulfilled, and let the errors ε_{i} be normal zero mean. Then the critical values $\mathfrak{z}_{m,m^{\circ}}$ given by (7.12) satisfy for all pairs $m > m^{\circ}$ in $\mathcal{M}$*

$$\mathfrak{z}_{m,m^{\circ}} \le (1+\beta)\sqrt{\mathtt{p}_{m,m^{\circ}}} + \sqrt{2\lambda_{m,m^{\circ}}\bigl(\mathtt{x} + \log|\mathcal{M}|\bigr)}\,.$$

Suppose also that

$$\mathtt{p}_{m^{*},m} \le \mathtt{p}_{m^{*}}\,, \qquad \lambda_{m^{*},m} \le \lambda_{m^{*}}\,, \qquad \forall m \in \mathcal{M}^{-}.$$

Then the value $\bar{\mathfrak{z}}_{m^{*}}$ follows the bound

$$\bar{\mathfrak{z}}_{m^{*}} \le (1+\beta)\sqrt{\mathtt{p}_{m^{*}}} + \sqrt{2\lambda_{m^{*}}\bigl(\mathtt{x} + \log|\mathcal{M}|\bigr)}\,.$$

Remark 7.3.4. The presented results help to understand the relation between the oracle risk $\mathcal{R}_{m^{*}}$ and the term $\bar{\mathfrak{z}}_{m^{*}}$. We know that $\mathcal{R}_{m^{*}} = \|\boldsymbol{b}_{m^{*}}\|^{2} + \mathtt{p}_{m^{*}} \ge \mathtt{p}_{m^{*}}$. Consider separately two cases: $\mathtt{p}_{m^{*}} \gg \lambda_{m^{*}}$ and $\mathtt{p}_{m^{*}} \asymp \lambda_{m^{*}}$. In the first case which is the typical situation in model selection, it also holds $2\lambda_{m^{*}}\bigl(\mathtt{x} + \log|\mathcal{M}|\bigr) \ll \mathtt{p}_{m^{*}}$ and the payment for adaptation is not essentially larger than the oracle risk. In fact, for the case with a narrow zone of insensitivity $\mathcal{M}_{\mathrm{in}} = \mathcal{M}^{-} \setminus \mathcal{M}_{\boldsymbol{b}}^{-}$, the value $\bar{\mathfrak{z}}_{m^{*}}$ is much smaller than

$\mathtt{p}_{m^*}^{1/2}$; see Section 7.3.3 for details. The second case $\mathtt{p}_{m^*} \asymp \lambda_{m^*}$ is somewhat extreme and it corresponds to estimation of a linear functional or estimation for severely ill-posed problems; see Section 7.4.1 below. In this case, the squared payment for adaptation $\bar{\mathfrak{z}}_{m^*}^2$ can be larger than the oracle risk by a factor $\log|\mathcal{M}|$.

7.3.3 Application to projection estimation

This section discusses the case of projection estimation in the linear model $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with homogeneous errors ε_i : $\operatorname{Var}(\varepsilon_i) = \sigma^2$. 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 that the features in Ψ are ordered and for each $m \in \mathbb{N}$, denote by Ψ_m the $p \times n$ matrix corresponding to the first m features and obtained from Ψ by letting to zero all the entries for the remaining features. The related estimator $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$ is the standard LSE with $\mathcal{S}_m = \bigl(\Psi_m \Psi_m^\top\bigr)^{-} \Psi_m$ and the prediction problem with $W = \Psi^\top$ yields $\mathcal{K}_m \boldsymbol{Y} = \Psi^\top \mathcal{S}_m \boldsymbol{Y} = \Pi_m \boldsymbol{Y}$, where $\Pi_m = \Psi_m^\top \bigl(\Psi_m \Psi_m^\top\bigr)^{-} \Psi_m$ is the projector in $I\!\!R^n$ onto the corresponding m -dimensional subspace. For homogeneous errors ε_i with $\operatorname{Var}(\varepsilon_i) = \sigma^2$, the variance $\mathbb{V}_m = \operatorname{Var}\bigl(\Pi_m \boldsymbol{Y}\bigr)$ satisfies

$$\mathtt{p}_m = \operatorname{tr}\bigl\{\operatorname{Var}\bigl(\Pi_m \boldsymbol{Y}\bigr)\bigr\} = \sigma^2 \operatorname{tr}\bigl(\Pi_m\bigr) = \sigma^2 m.$$

Moreover, for each pair $m > m^\circ$, it holds

$$\Psi^\top \bigl(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ}\bigr) = \bigl(\Pi_m - \Pi_{m^\circ}\bigr) \boldsymbol{Y}.$$

Corollary 7.3.1. *Consider the problem of projection estimation with homogeneous Gaussian errors* ε_i *and probabilistic loss. Then* $\mathtt{p}_{m,m^\circ} = \sigma^2 (m - m^\circ)$, $\lambda_{m,m^\circ} = \sigma^2$, *and*

$$\begin{aligned} \mathfrak{z}_{m,m^\circ} &\le \sigma(1+\beta)\sqrt{m - m^\circ} + \sigma\sqrt{2\mathtt{x} + 2\log|\mathcal{M}|}, \qquad (7.26) \\ \bar{\mathfrak{z}}_{m^*} &\le \sigma(1+\beta)\sqrt{m^*} + \sigma\sqrt{2\mathtt{x} + 2\log|\mathcal{M}|}. \end{aligned}$$

Remark 7.3.5. The first term in the expression for $\bar{\mathfrak{z}}_{m^*}$ 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. (1997a). Then $\log|\mathcal{M}| \approx \log\log n$ and $\bar{\mathfrak{z}}_{m^*} \approx \sigma\sqrt{m^*}$ for $m^* \gg \log\log n$. For the oracle risk $\mathcal{R}_{m^*}$, it holds $\mathcal{R}_{m^*} = \mathtt{p}_{m^*} + \|\boldsymbol{b}_{m^*}\|^2 \ge \sigma^2 m^*$. Therefore, the payment for adaptation $\bar{\mathfrak{z}}_{m^*}$ is not larger in order than the square root of the oracle risk, and the result of Theorem 7.3.3 has a surprising corollary: if the oracle dimension m^* is significantly larger than $\log\log n$, then the data-driven SmA estimator provides nearly the same accuracy as the oracle one.

Remark 7.3.6. The payment for adaptation can be drastically reduced in the situations with a narrow zone of insensitivity. Suppose that the "large bias set" $\mathcal{M}_{\boldsymbol{b}}^{-}$ contains all indices $m \leq m^{\circ}$ for a fixed $m^{\circ} < m^{*}$. For instance, this is the case when $\|\boldsymbol{b}_{m^{*},m}\|^{2} \geq \mathtt{C}\sigma^{2}\bigl(m^{*}-m+2\mathtt{x}+2\log|\mathcal{M}|\bigr)$ for some fixed sufficiently large constant $\mathtt{C}$ and all $m \leq m^{\circ}$. Then by (7.26)

$$\bar{\mathfrak{z}}_{m^{*}} = \max_{m \in \mathcal{M}_{\mathrm{in}}} \mathfrak{z}_{m^{*},m} \leq \sigma(1+\beta)\sqrt{m^{*}-m^{\circ}} + \sigma\sqrt{2\mathtt{x}+2\log|\mathcal{M}|}.$$

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 7.3.7. The popular Akaike criterion (AIC) defines $\widehat{m}$ as

$$\widehat{m} = \operatorname*{argmin}_{m}\bigl\{\|\boldsymbol{Y}-\Pi_{m}\boldsymbol{Y}\|^{2} + 2\sigma^{2}m\bigr\}.$$

One can easily see that this rule is equivalent to the SmA rule (7.16) with $\mathfrak{z}_{m,m^{\circ}}^{2} = 2\sigma^{2}(m-m^{\circ})$. For this choice, one can prove a risk oracle bound under rather general conditions (see e.g. Kneip (1994)), however, it does not deliver any information about the behavior of $\widehat{m}$, in particular, it does not guarantee the propagation property (7.14).

7.4 * Oracle accuracy and asymptotic minimax risk

Here we briefly discuss the relation between the oracle bound (7.25) and minimax rates of estimation in regression with regular design and homogeneous noise. Suppose that the mean response vector $\boldsymbol{f}$ corresponds to the values of a smooth regression function $f(X_{i})$ at some regular design points $X_{1},\ldots,X_{n} \in [0,1]$. Let $\psi_{1},\ldots,\psi_{m},\ldots,$ be a set of basis functions on $[0,1]$ like cosine, Demmler-Reinsch, or B-splines basis. We identify the function ψ_{m} with the vector $\boldsymbol{\psi}_{m}$ of its values at design points, $\boldsymbol{\psi}_{m} = \bigl(\psi_{m}(X_{1},\ldots,\psi_{m}(X_{n})\bigr)^{\top} \in I\!\!R^{n}$. The operator Π_{m} projects onto the subspace spanned by the first m vectors $\boldsymbol{\psi}_{1},\ldots,\boldsymbol{\psi}_{m}$. Then under the standard Sobolev smoothness condition on f, the bias $\boldsymbol{b}_{m} = \boldsymbol{f}-\Pi_{m}\boldsymbol{f}$ satisfies $n^{-1}\|\boldsymbol{b}_{m}\|^{2} \leq \mathtt{C}m^{-2s}$ and similarly $n^{-1}\|\boldsymbol{b}_{m',m}\|^{2} \leq \mathtt{C}m^{-2s}$ for $m' > m$. Together with $p_{m} = \sigma^{2}m$ this yields that the conditions (7.9) are fulfilled with any fixed β and $m^{*} \approx \mathtt{C}(\beta)n^{1/(2s+1)}$ and $n^{-1}\mathcal{R}_{m^{*}} \leq \mathtt{C}(\beta)n^{-2s/(2s+1)}$, where the constant $\mathtt{C}(\beta)$ only depensd on β. This is the optimal accuracy over the class of smooth function of the Sobolev degree s; see e.g. Ibragimov and Khas'minskij (1981) or **?**. In view of Remark 7.3.5 the proposed selector ensures the optimal estimation rate over a Sobolev smoothness class without knowing the parameters of the class up to the additive payment for adaptation $\bar{\mathfrak{z}}_{m^*}$. This easily implies the classical results on adaptive estimation: the SmA estimator is rate-adaptive over a wide range of smoothness classes such of degree s under the constraint $n^{1/(2s+1)} \geq \mathtt{C} \log\log n$.

7.4.1 Linear functional estimation

In this section, we discuss the problem of linear functional estimation. As previously, we assume a family of estimators $\widetilde{\phi}_m = \mathcal{K}_m \boldsymbol{Y}$, $m \in \mathcal{M}$, 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

$$\mathtt{p}_m = \operatorname{Var}\bigl(\mathcal{K}_m \boldsymbol{Y}\bigr) = \mathcal{K}_m \operatorname{Var}(\boldsymbol{\varepsilon}) \, \mathcal{K}_m^{\top}$$

which grows with m. Further, each stochastic component $\xi_{m,m^\circ} = \mathcal{K}_{m,m^\circ} \boldsymbol{\varepsilon}$ is one-dimensional, and it holds for $m > m^\circ$

$$\lambda_{m,m^\circ} = \mathtt{p}_{m,m^\circ} = \mathcal{K}_{m,m^\circ} \operatorname{Var}(\boldsymbol{\varepsilon}) \, \mathcal{K}_{m,m^\circ}^{\top}.$$

Note that in the case of Gaussian errors, ξ_{m,m° is also Gaussian: $\xi_{m,m^\circ} \sim \mathcal{N}(0, \mathtt{p}_{m,m^\circ})$. The tail function $z_{m,m^\circ}(\mathtt{x})$ of ξ_{m,m° can be upper-bounded by $\sqrt{2\mathtt{x} \, \mathtt{p}_{m,m^\circ}}$ yielding

$$\mathfrak{z}_{m,m^\circ} \leq \mathtt{p}_{m,m^\circ}^{1/2} \Bigl(\beta + \sqrt{2\mathtt{x} + 2 \log |\mathcal{M}|} \Bigr), \tag{7.27}$$

$$\bar{\mathfrak{z}}_{m^*} \leq \mathtt{p}_{m^*}^{1/2} \Bigl(\beta + \sqrt{2\mathtt{x} + 2 \log |\mathcal{M}|} \Bigr). \tag{7.28}$$

Theorem 7.4.1. *Let the errors ε_i be Gaussian zero mean. Consider a problem of linear functional estimation of $\phi^* = W \boldsymbol{\theta}^*$ by a given family $\widetilde{\phi}_m = \mathcal{K}_m \boldsymbol{Y}$ with* $\operatorname{rank}(\mathcal{K}_m) = 1$, *$m \in \mathcal{M}$. Then the critical values $\mathfrak{z}_{m,m^\circ}$ from (7.12) fulfill (7.27) and the oracle inequality (7.25) holds with the payment for adaptation $\bar{\mathfrak{z}}_{m^*}$ obeying (7.28).*

Remark 7.4.1. For the problem of linear functional estimation with probabilistic loss, the squared payment for adaptation $\bar{\mathfrak{z}}_{m^*}^2$ is by a factor $\log |\mathcal{M}|$ larger than the oracle variance $\mathtt{p}_{m^*}$. 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. This factor appears to be unavoidable; see e.g. Spokoiny and Vial (2009) in the context of estimating a linear functional.

7.4.2 Validity of the bootstrap procedure. Conditions

This and the next sections extend the results obtained for the case of known error distribution to the bootstrap procedure which does not use any information about the noise variance. The main result claims that the bootstrap choice still ensures the condition (7.11) and therefore, all the obtained results including the oracle bounds, apply for this choice as well; see Theorem 7.4.3. Moreover, we evaluate the distance between the unknown underlying distribution $\mathbb{Q}$ of the set of random vectors $\boldsymbol{\xi}_{m,m^\circ}$ and their bootstrap counterpart $\mathbb{Q}^\flat$. The latter is random, however, we show that with high probability, it is close to $\mathbb{Q}$. In what follows we assume the model (7.1) with a heterogeneous Gaussian noise $\boldsymbol{\varepsilon}$. The results presented below rely on the following quantities.

Design regularity is measured by the value δ_Ψ

$$\delta_\Psi \overset{\text{def}}{=} \max_{i=1,\ldots,n} \|S^{-1/2}\Psi_i\|\sigma_i\,, \quad \text{where} \quad S \overset{\text{def}}{=} \sum_{i=1}^{n} \Psi_i \Psi_i^\top \sigma_i^2\,; \tag{7.29}$$

Obviously

$$\sum_{i=1}^{n} \|S^{-1/2}\Psi_i\|^2 \sigma_i^2 = \operatorname{tr}\Bigl(\sum_{i=1}^{n} S^{-1}\Psi_i\Psi_i^\top \sigma_i^2\Bigr) = \operatorname{tr} \boldsymbol{I}_p = p,$$

and therefore in typical situations the value δ_Ψ is of order $\sqrt{p/n}$.

Presmoothing bias for $\boldsymbol{f} = I\!\!E\boldsymbol{Y}$ is described by the vector

$$\boldsymbol{B} = \Sigma^{-1/2}(\boldsymbol{f} - \Pi\boldsymbol{f}). \tag{7.30}$$

We will use the sup-norm $\|\boldsymbol{B}\|_\infty = \max_i |b_i|$ to measure the bias after presmoothing.

Regularity of the smoothing operator Π is required in Theorem 7.4.3. Namely, we assume that the rows $\Upsilon_i^\top$ of the matrix $\Upsilon \overset{\text{def}}{=} \Sigma^{-1/2}\Pi\Sigma^{1/2}$ fulfill

$$\|\Upsilon_i^\top\| \le \delta_\Pi, \qquad i = 1, \ldots, n. \tag{7.31}$$

This condition is in fact very close to the design regularity condition (7.29). To see this, consider the case of a homogeneous noise with $\Sigma = \sigma^2 \boldsymbol{I}_n$ and $\Pi = \Psi^\top\bigl(\Psi\Psi^\top\bigr)^{-1}\Psi$. Then $\Upsilon = \Pi$ and (7.29) implies

$$\|\Upsilon_i^\top\| = \bigl\|\Psi^\top\bigl(\Psi\Psi^\top\bigr)^{-1}\Psi_i\bigr\| = \bigl\|\bigl(\Psi\Psi^\top\bigr)^{-1/2}\Psi_i\bigr\| \le \delta_\Psi\,.$$

In general one can expect that δ_Ψ and δ_Π are of the same order $\sqrt{p/n}$.

7.4.3 Bootstrap validation

This section states the main results justifying the proposed bootstrap procedure: the joint distribution $\mathbb{Q}^\flat$ of the bootstrap stochastic components $\boldsymbol{\xi}_{m,m^\circ}^\flat$ for $m > m^\circ$ from (7.18) nicely reproduces the underlying distribution $\mathbb{Q}$ of the $\boldsymbol{\xi}_{m,m^\circ}$'s, and hence, all the probabilistic results obtained in Section 7.3.1 for known noise continue to apply after bootstrap parameter tuning. The next result presents a bound on the total variation distance $\|\mathbb{Q} - \mathbb{Q}^{\flat}\|_{\mathrm{TV}}$ between $\mathbb{Q}$ and $\mathbb{Q}^{\flat}$. As $\mathbb{Q}^{\flat}$ is a random measure, the result only holds with high probability.

Theorem 7.4.2. *Let $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ be a Gaussian vector in $I\!\!R^{n}$ with independent components, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0,\Sigma)$ for $\Sigma = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_n^2)$, and let also the feature matrix Ψ be such that the $p \times p$-matrix $S = \Psi\,\Sigma\,\Psi^{\top}$ is non-degenerated and (7.29) holds. For a given presmoothing operator $\Pi\colon I\!\!R^{n} \to R^{n}$, assume (7.31) to be fulfilled with $\Upsilon \overset{\mathrm{def}}{=} \Sigma^{-1/2}\Pi\Sigma^{1/2}$. Let $\mathbb{Q} = \mathcal{L}\bigl(\boldsymbol{\xi}_{m,m^{\circ}}, m > m^{\circ} \in \mathcal{M}\bigr)$ and let $\mathbb{Q}^{\flat}$ be the joint conditional distribution of the bootstrap stochastic terms $\boldsymbol{\xi}^{\flat}_{m,m^{\circ}}$ for $m > m^{\circ} \in \mathcal{M}$ given the data $\boldsymbol{Y}$. Then it holds on a random set Ω_n with $I\!\!P\bigl(\Omega_n\bigr) \ge 1 - 6/n$:*

$$\|\mathbb{Q} - \mathbb{Q}^{\flat}\|_{\mathrm{TV}} \le \frac{1}{2}\sqrt{p}\,\Delta_n, \tag{7.32}$$

$$\Delta_n \overset{\mathrm{def}}{=} \|\boldsymbol{B}\|_{\infty}^{2} + 4\bigl(\delta_{\Pi}\,\|\boldsymbol{B}\|_{\infty} + \delta_{\Psi}\bigr)\sqrt{\log n} + 4(\delta_{\Pi} + \delta_{\Pi}^{2} + \delta_{\Psi}^{2})\log n\,, \tag{7.33}$$

where the bias $\boldsymbol{B}$ is given by (7.30).

The result (7.32) enables us to control the differences $\mathbb{Q}(A) - \mathbb{Q}^{\flat}(A)$ for fixed sets A. To justify the propagation property for the bootstrap-based set of critical values $z^{\flat}_{m,m^{\circ}}(\mathtt{x} + q^{\flat}_{m^{\circ}})$, given according to (7.19), (7.20), and (7.21) with $\check{\boldsymbol{Y}} = \boldsymbol{Y} - \Pi\boldsymbol{Y}$, we also need to take into account the $\boldsymbol{Y}$-dependence of $z^{\flat}_{m,m^{\circ}}(\mathtt{x}+q^{\flat}_{m^{\circ}})$. This is done by the following theorem.

Theorem 7.4.3. *Assume the conditions of Theorem 7.4.2. Let Δ_n be from (7.33). Then for each $m^{\circ} \in \mathcal{M}$, it holds on a random set Ω_n with $I\!\!P\bigl(\Omega_n\bigr) \ge 1 - 6/n$:*

$$\begin{aligned}\biggl| I\!\!P\biggl(\max_{m>m^{\circ}}\Bigl\{\|\boldsymbol{\xi}_{m,m^{\circ}}\| - z^{\flat}_{m,m^{\circ}}(\mathtt{x} + q^{\flat}_{m^{\circ}})\Bigr\} \ge 0\biggr) & \\ - I\!\!P^{\flat}\biggl(\max_{m>m^{\circ}}\Bigl\{\|\boldsymbol{\xi}^{\flat}_{m,m^{\circ}}\| - z^{\flat}_{m,m^{\circ}}(\mathtt{x} + q^{\flat}_{m^{\circ}})\Bigr\} \ge 0\biggr)\biggr| & \le \sqrt{p}\,\Delta_n\,.\end{aligned} \tag{7.34}$$

By construction, the values $z^{\flat}_{m,m^{\circ}}(\mathtt{x} + q^{\flat}_{m^{\circ}})$ are selected as minimal ones under the propagation constraint in the bootstrap world. The presented result shows that the use of these data-dependent critical values does not destroy the propagation condition in the real world.

Now we state a bootstrap version of Theorem 7.3.1. Note that the definition (7.9) of the oracle m^{*} involves a constant β, and exactly the same constant shows up in the definition (7.12) of the $\mathfrak{z}_{m,m^{\circ}}$'s for the case of known noise distribution. For the bootstrap procedure, the value $\beta^{\flat}$ in the definition (7.22) of the $\mathfrak{z}^{\flat}_{m,m^{\circ}}$'s has to be slightly larger than β from (7.9):

$$\beta^{\flat} \geq \bigl(1 - \Delta_{n}\bigr)^{-1/2} \beta.$$

If Δ_{n} is small then one can fix $\beta^{\flat} \approx \beta$. Our default choice is again $\beta^{\flat} = 1$.

Theorem 7.4.4. *Assume the conditions of Theorem 7.4.3. Given* $\mathtt{x}$ *and* $\beta^{\flat}$*, let the critical values* $\mathfrak{z}^{\flat}_{m,m^{\circ}}$ *be given by (7.22). If the value* β *(from the definition (7.9) of* m^{*}*) and* $\beta^{\flat}$ *satisfy* $\beta^{\flat} \geq \bigl(1 - \Delta_{n}\bigr)^{-1/2} \beta$*, then*

$$I\!\!P\bigl(m^{*} \text{ is rejected}\bigr) \leq \mathrm{e}^{-\mathtt{x}} + \sqrt{p}\, \Delta_{n} ,$$

and the bootstrap calibrated SmA estimator $\widehat{\boldsymbol{\phi}} = \widetilde{\boldsymbol{\phi}}_{\widehat{m}}$ *satisfies*

$$I\!\!P\Bigl(\bigl\| \widehat{\boldsymbol{\phi}} - \widetilde{\boldsymbol{\phi}}_{m^{*}} \bigr\| > \bar{\mathfrak{z}}^{\flat}_{m^{*}} \Bigr) \leq \mathrm{e}^{-\mathtt{x}} + 6 n^{-1} + \sqrt{p}\, \Delta_{n} , \tag{7.35}$$

where $\bar{\mathfrak{z}}^{\flat}_{m^{*}}$ *satisfies on the set* Ω_{n} *(from Theorem 7.4.2) the bound*

$$\bar{\mathfrak{z}}^{\flat}_{m^{*}} \stackrel{\text{def}}{=} \max_{m \in \mathcal{M}^{-}} \mathfrak{z}^{\flat}_{m^{*},m} \leq \sqrt{1 + \Delta_{n}} \, \Bigl\{ (1 + \beta) \sqrt{\mathtt{p}_{m^{*}}} + \sqrt{2 \lambda_{m^{*}} (\mathtt{x} + \log |\mathcal{M}|)} \Bigr\} . \tag{7.36}$$

7.5 Bootstrap validity and critical dimension

Now we discuss the sense of the required conditions for bootstrap validity. The obtained results involve the error term $\sqrt{p}\, \Delta_{n}$ describing the accuracy of the bootstrap approximation. The Gaussian framework allows to reduce the proof of bootstrap validity to the comparison of two Gaussian measures and to get explicit error bounds. If the errors $\boldsymbol{\varepsilon}$ in the original model (7.1) are not Gaussian, the proof of bootstrap validity requires additional tools like high dimensional Gaussian approximation yielding much larger error bounds; cf. Spokoiny and Zhilova (2015).

Our results are only meaningful and the bootstrap approximation is accurate if the value $\sqrt{p}\, \Delta_{n}$ in (7.32) is small. One easily gets

$$\sqrt{p}\, \Delta_{n} \leq \mathtt{C} p^{1/2} \Bigl\{ \| \boldsymbol{B} \|_{\infty}^{2} + \bigl(\delta_{\Psi} + \delta_{\Pi} \bigr) \log n \Bigr\},$$

where $\mathtt{C}$ is a generic notation for an absolute constant. So, the bootstrap approximation is valid if the values $p^{1/2} \, \delta_{\Psi} \, \log n$, $p^{1/2} \, \delta_{\Pi} \, \log n$, $\| \boldsymbol{B} \|_{\infty}^{2} \, p^{1/2}$ are sufficiently small. Now we spell this condition in the typical situation with $\delta_{\Psi} \asymp \sqrt{p/n}$ and $\delta_{\Pi} \asymp \sqrt{p/n}$. Then it suffices that the values $p \, n^{-1/2} \log(n)$ and $\| \boldsymbol{B} \|_{\infty}^{2} \, p^{1/2}$ are small. Suppose that

$$\| \boldsymbol{B} \|_{\infty} \leq \mathtt{C} p^{-s}. \tag{7.37}$$

Such bounds for $\boldsymbol{B} = \boldsymbol{f} - \Pi \boldsymbol{f}$ are often used in the approximation theory when the response vector $\boldsymbol{f}$ corresponds to a Hölder-smooth regression function with the smoothness parameter s observed with noise at design points. So, the bias component does not destroy the bootstrap validity result if p^{1-4s} is small. We summarize that the bootstrap procedure is justified for $s > 1/4$ if $p = p_n \to \infty$ but $p_n n^{-1/2} \log(n) \to 0$ as $n \to \infty$.

Corollary 7.5.1. *Assume the conditions of Theorem 7.4.3 and let (7.37) hold for* $s > 1/4$. *If* $p = p_n$ *fulfill* $p_n n^{-1/2} \log(n) \to 0$ *as* $n \to \infty$, *then the results of Theorem 7.4.2 and 7.4.3 apply with* $\Delta_n \to 0$ *as* $n \to \infty$.

7.6 Simulations

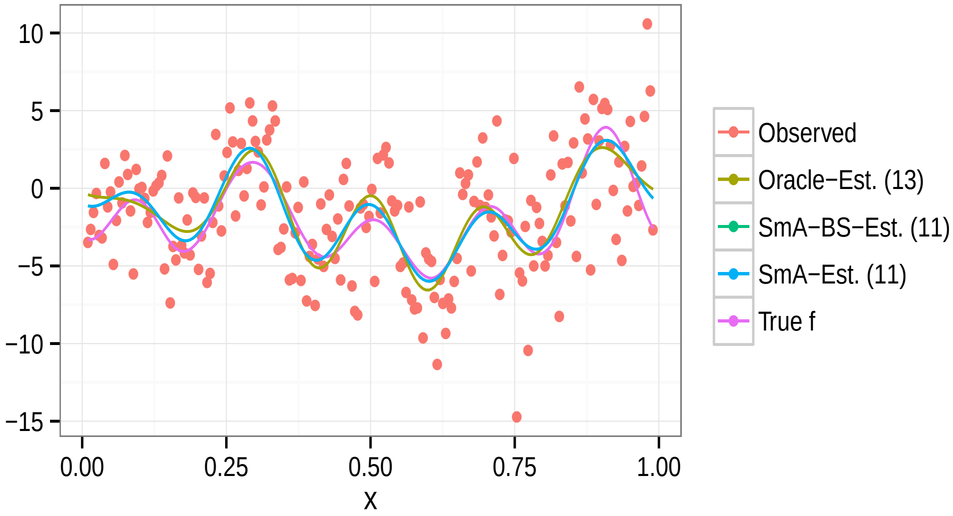

Fig. 7.1. True function and observed data with oracle estimator, the known-variance SmA-Estimator (SmA-Est.) and the Bootstrap-SmA-Estimator (SmA-BS-Est.) for inhomogeneous noise. The numbers in parentheses indicate the chosen model dimension.

This section illustrates the performance of the proposed procedure by means of simulated examples. We consider a regression model $Y_i = f(X_i) + \varepsilon_i$ for an unknown univariate function on $[0, 1]$ with unknown inhomogeneous Gaussian noise $\boldsymbol{\varepsilon}$. 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 $m^{\dagger}$.

We consider a sequence of equidistant design points $(x_i)_{1 \le i \le n}$ on $[0, 1]$ and the Fourier basis $\{\psi_j(x)\}_{j=1}^{\infty}$ to define a sequence of projection estimators where m indicates the truncation dimension of the Fourier basis. The true function is generated by

$$f(x) = c_1 \psi_1(x) + \ldots + c_n \psi_n(x),$$

where the $(c_j)_{1 \le j \le n}$ are chosen randomly: with γ_j i.i.d. standard normal

$$c_j = \begin{cases} \gamma_j, & 1 \le j \le 10, \\ \gamma_j/(j-10)^2, & 11 \le j \le n. \end{cases}$$

The noise variances are obtained in the following way: one draws a vector from a normal distribution $\mathcal{N}(2, 0.4 I_n)$, takes the square of the coefficients of this vector, puts the coefficients in ascending order and then defines the resulting vector as σ^2. The covariance matrix will then be $\Sigma = \lambda_{\text{int}} \cdot \operatorname{diag}((\sigma_i^2)_{1 \le i \le n})$, where λ_{int} governs the intensity of the noise and will be 0.2^2, 0.8^2, and 1.4^2 respectively for low, medium and high noise level. To generate the noisy observations $n = 200$ will be used. When considering smaller sample sizes, we will take equidistant subsamples of the observations. As a default the medium noise level will be used for simulations.

For each index m, we build the projection estimate using the first m basis functions. Solving the associated least squares problem gives $\boldsymbol{\theta}_m$ which we will assume to be in $I\!R^n$ by filling up the vectors of coefficients with zeros.

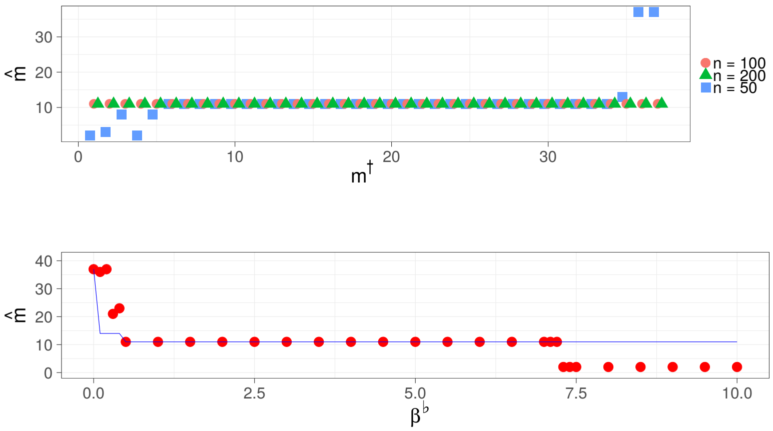

Fig. 7.2. (Top) The value $\widehat{m}$ chosen by the Bootstrap-SmA-Method as a function of the presmoothing dimension $m^\dagger$ for $n = 50, 100, 200$. (Bottom) $\widehat{m}$ as a function of β for $n = 200$. The blue line indicates the oracle m^* according to the definition (7.9).

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 $M = 37$. Our default choice for calibration is $\mathtt{x} = 2$, $\beta^\flat = 1$, and $m^\dagger = 20$.

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 7.1 three examples with different intensity of the noise term comparing the Bootstrap-method to the oracle estimator and the known-variance SmA-Method.

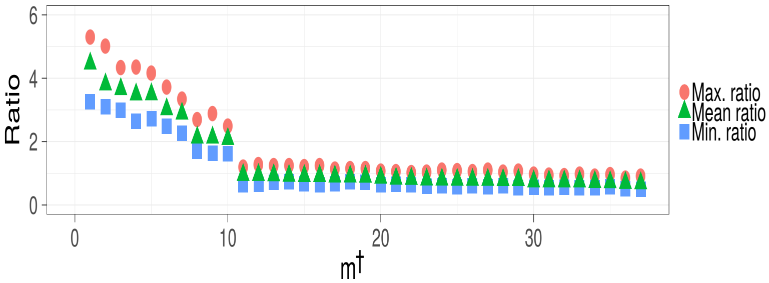

Fig. 7.3. Maximal, minimal and mean ratio of the bootstrap and theoretical critical values $|\mathfrak{z}^{\flat}_{m,m^{\circ}}/\mathfrak{z}_{m,m^{\circ}}|^{2}$ as a function of $m^{\dagger}$.

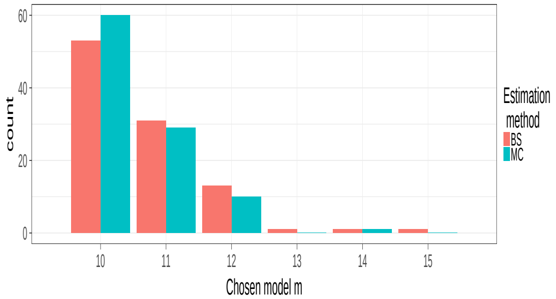

Fig. 7.4. Histograms for the selected model by the bootstrap (BS) and the known-variance method (MC), simulation size $n_{\text{hist}} = 100$.

Figure 7.2, top, illustrates the dependence of the SmA choice $\widehat{m}$ on the presmoothing dimension $m^{\dagger}$ and on the parameter $\beta^{\flat}$ for different values of the sample size n for one typical noise realization. We see that in the specific example we are considering, the impact of $m^{\dagger}$ decreases very fast with n. In particular, in the case $n = 200$, no variation in the choice of $\widehat{m}$ is observed in the whole range of $m^{\dagger}$. The oracles are respectively $m^{*} = 12$ for $n = 100, 200$ and $m^{*} = 10$ for $n = 50$. The impact of the parameter $\beta^{\flat}$ on the estimation results is illustrated on Figure 7.2, bottom, for $n = 200$. The method appears to be very stable with respect to the choice of $\beta^{\flat}$.

Figure 7.3 demonstrates the variability of the ratios $\mathfrak{z}^{\flat}_{m_1,m_2}/\mathfrak{z}_{m_1,m_2}$ w.r.t. $m^{\dagger}$. It is remarkable that it is very stable in the range $m^{\dagger} \geq 12$. Figure 7.4 shows the distribution of the selected index $\widehat{m}$ after $n_{\text{hist}} = 100$ simulations of the method with the same underlying function f observed with different realizations of the errors. Figure 7.5 shows the numerical results for the estimation of the first derivative $f'(x)$ in the same model as above. This means taking $W = (\psi_i'(x_j))_{1 \leq i,j \leq n}$. The bootstrap SmA-procedure is well competitive with the procedure based on a known noise structure and the method does a good job of mimicking the oracle in various settings.

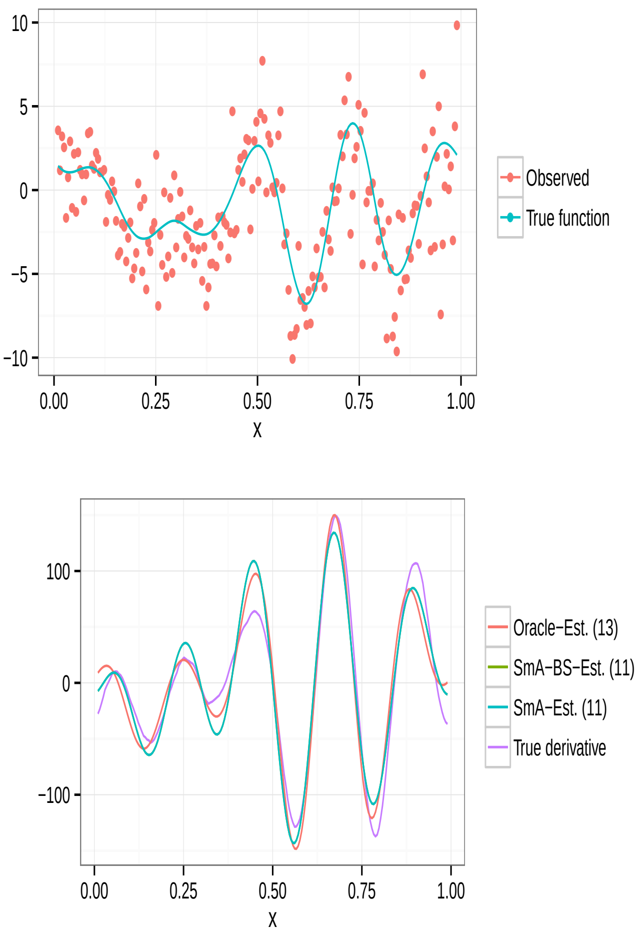

Fig. 7.5. Left: the true function and the observations. Right: the true derivative, the oracle estimator, the known-variance SmA-Estimator (SmA-Est.) and the Bootstrap-SmA-Estimator (SmA-BS-Est.).

7.7 * Proofs

The appendix collects the proofs of announced results.

7.7.1 Proof of Theorems 7.3.1 and 7.3.2

The propagation property (7.14) 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, $\widehat{m} \leq m^*$ with a probability at least $1 - \mathrm{e}^{-\mathtt{x}}$. Below we consider only this event. Let $m \in \mathcal{M}^-$. Acceptance of m requires in particular that $\mathbb{T}_{m^*,m} \leq \mathfrak{z}_{m^*,m}$. The representation $\mathbb{T}_{m^*,m} = \|\boldsymbol{b}_{m^*,m} + \boldsymbol{\xi}_{m^*,m}\|$ implies

$$I\!\!P\bigl(\mathbb{T}_{m^*,m} \leq \mathfrak{z}_{m^*,m}\bigr) \leq I\!\!P\bigl(\|\boldsymbol{\xi}_{m^*,m}\| \geq \|\boldsymbol{b}_{m^*,m}\| - \mathfrak{z}_{m^*,m}\bigr).$$

If $m \in \mathcal{M}_{\boldsymbol{b}}^-$, this yields with $\bar{\mathtt{x}} = \mathtt{x} + \log|\mathcal{M}^-|$

$$\begin{aligned} \mathbb{P}\bigl(m \text{ is accepted}\bigr) &\le \mathbb{P}\bigl(\|\boldsymbol{b}_{m^*,m} + \boldsymbol{\xi}_{m^*,m}\| \le \mathfrak{z}_{m^*,m}\bigr) \\ &\le \mathbb{P}\bigl(\|\boldsymbol{\xi}_{m^*,m}\| \ge z_{m^*,m}(\bar{\mathtt{x}})\bigr) \le \mathrm{e}^{-\bar{\mathtt{x}}}. \end{aligned}$$

This helps to bound the probability of the event $\{\widehat{m} \in \mathcal{M}_{\boldsymbol{b}}^{-}\}$:

$$\mathbb{P}\bigl(\widehat{m} \in \mathcal{M}_{\boldsymbol{b}}^{-}\bigr) \le \sum_{m \in \mathcal{M}_{\boldsymbol{b}}^{-}} \mathbb{P}\bigl(\|\boldsymbol{b}_{m^*,m} + \boldsymbol{\xi}_{m^*,m}\| \le \mathfrak{z}_{m^*,m}\bigr) \le \sum_{m \in \mathcal{M}_{\boldsymbol{b}}^{-}} \mathrm{e}^{-\bar{\mathtt{x}}} \le \mathrm{e}^{-\mathtt{x}}.$$

Therefore, the probability that the SmA-selector picks up a value $m > m^*$ or $m \in \mathcal{M}_{\boldsymbol{b}}^{-}$ is very small:

$$\mathbb{P}\Bigl(\widehat{m} \in \mathcal{M}^{+}(m^*) \cup \mathcal{M}_{\boldsymbol{b}}^{-}\Bigr) \le 2\mathrm{e}^{-\mathtt{x}}.$$

It remains to study the case when $\widehat{m} = m$ for some $m \in \mathcal{M}_{\mathrm{in}}$. We can use that this m is accepted, which implies by definition

$$\mathbb{T}_{m^*,m} = \|\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^*}\| \le \mathfrak{z}_{m^*,m} \, .$$

This yields (7.23). The bound (7.25) now follows by the triangle inequality.

7.7.2 Proof of Theorem 7.3.3

Below we use the deviation bound (C.2) for a Gaussian quadratic form from Theorem C.1 in the supplement [SW2016]. Note that similar results are available for non-Gaussian quadratic forms under exponential moment conditions; see e.g. Spokoiny (2012). The result (C.2) combined with the Bonferroni correction $q_{m^\circ} = \log|\mathcal{M}^{+}(m^\circ)| \le \log|\mathcal{M}|$ yields the following upper bound for the critical values $\mathfrak{z}_{m,m^\circ}$:

$$\begin{aligned} \mathfrak{z}_{m,m^\circ} &\le z_{m,m^\circ}(\mathtt{x} + q_{m^\circ}) + \beta \mathtt{p}_{m,m^\circ}^{1/2} \\ &\le (1+\beta)\sqrt{\mathtt{p}_{m,m^\circ}} + \sqrt{2\lambda_{m,m^\circ}\bigl(\mathtt{x} + \log|\mathcal{M}^{+}(m^\circ)|\bigr)} \\ &\le (1+\beta)\sqrt{\mathtt{p}_{m,m^\circ}} + \sqrt{2\lambda_{m,m^\circ}(\mathtt{x} + \log|\mathcal{M}|)}. \end{aligned} \tag{7.38}$$

For the payment for adaptation $\bar{\mathfrak{z}}_{m^*}$, the result (7.38) and the conditions $\mathtt{p}_{m^*,m} \le \mathtt{p}_{m^*}$ and $\lambda_{m^*,m} \le \lambda_{m^*}$ imply the required upper bound:

$$\bar{\mathfrak{z}}_{m^*} \le (1+\beta)\sqrt{\mathtt{p}_{m^*}} + \sqrt{2\lambda_{m^*}(\mathtt{x} + \log|\mathcal{M}|)}\, .$$

7.7.3 Proof of Theorem 7.4.2

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 bootstrap-based values for the original procedure, we have to account for this dependence. The statement of Theorem 7.4.2 is even more involved due to the presmoothing step and multiplicity correction (7.21). 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 $\breve{Y}_i$. Then we compare $\mathbb{Q}$ and $\mathbb{Q}^{\flat}$ 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 $\Psi_m = \Gamma_m \Psi$, where Γ_m is the projector on the subspace of the feature space $\mathbb{R}^p$ spanned by the features from the model m: $\Gamma_m = \Psi_m \Psi_m^{\top} \bigl(\Psi_m \Psi_m^{\top} \bigr)^{-}$. This allows to represent each estimator $\widetilde{\boldsymbol{\phi}}_m$ in the form

$$\begin{aligned} \widetilde{\boldsymbol{\phi}}_m &= W \widetilde{\boldsymbol{\theta}}_m = W \mathcal{S}_m \boldsymbol{Y} = W \bigl(\Psi_m \Psi_m^{\top} \bigr)^{-} \Psi_m \boldsymbol{Y} = \mathcal{T}_m \Psi \boldsymbol{Y} \\ \mathcal{T}_m &\stackrel{\mathrm{def}}{=} W \bigl(\Psi_m \Psi_m^{\top} \bigr)^{-} \Gamma_m = W \bigl(\Psi_m \Psi_m^{\top} \bigr)^{-} . \end{aligned}$$

This implies the following representation of the stochastic components $\boldsymbol{\xi}_{m,m^{\circ}}$ of the difference $\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^{\circ}}$: with $\nabla = \Psi \boldsymbol{\varepsilon}$, it holds

$$\boldsymbol{\xi}_{m,m^{\circ}} = \mathcal{T}_{m,m^{\circ}} \Psi \boldsymbol{\varepsilon} = \mathcal{T}_{m,m^{\circ}} \nabla, \qquad \mathcal{T}_{m,m^{\circ}} \stackrel{\mathrm{def}}{=} \mathcal{T}_m - \mathcal{T}_{m^{\circ}} , \tag{7.39}$$

Thus, each stochastic vector $\boldsymbol{\xi}_{m,m^{\circ}}$ is a linear function of the vector ∇. A similar representation holds true in the bootstrap world:

$$\boldsymbol{\xi}^{\flat}_{m,m^{\circ}} = \mathcal{T}_{m,m^{\circ}} \Psi \operatorname{diag}(\breve{\boldsymbol{Y}}) \boldsymbol{w}^{\flat} = \mathcal{T}_{m,m^{\circ}} \nabla^{\flat}, \qquad \nabla^{\flat} \stackrel{\mathrm{def}}{=} \Psi \operatorname{diag}(\breve{\boldsymbol{Y}}) \boldsymbol{w}^{\flat}. \tag{7.40}$$

Here the original errors $\boldsymbol{\varepsilon}$ are replaced by their bootstrap surrogates $\boldsymbol{\varepsilon}^{\flat} = \operatorname{diag}(\breve{\boldsymbol{Y}}) \boldsymbol{w}^{\flat}$. Therefore, it suffices to compare the distribution of $\nabla = \Psi \boldsymbol{\varepsilon}$ with the conditional distribution of $\nabla^{\flat} = \Psi \operatorname{diag}(\breve{\boldsymbol{Y}}) \boldsymbol{w}^{\flat}$ given $\boldsymbol{Y}$. 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 $\nabla = \Psi \boldsymbol{\varepsilon}$ is also normal zero mean:

$$\nabla \sim \mathcal{N}(0, S), \qquad S \stackrel{\mathrm{def}}{=} \Psi \Sigma \Psi^{\top}, \qquad \Sigma = \operatorname{Var}(\boldsymbol{\varepsilon}) = \operatorname{diag}\bigl(\sigma_1^2, \ldots, \sigma_n^2 \bigr).$$

Similarly, given the data $\boldsymbol{Y}$, the vector $\nabla^{\flat}$ is conditionally normal zero mean with the conditional variance

$$S^{\flat} \stackrel{\mathrm{def}}{=} \operatorname{Var}^{\flat}(\nabla^{\flat}) = \Psi \operatorname{diag}\bigl(\breve{Y}_1^2, \ldots, \breve{Y}_n^2 \bigr) \Psi^{\top} = \Psi \operatorname{diag}\bigl(\breve{\boldsymbol{Y}} \cdot \breve{\boldsymbol{Y}} \bigr) \Psi^{\top}.$$

Therefore, it remains to compare two p-dimensional Gaussian distributions with different covariance matrices. We apply Pinsker's inequality (see Lemma E.1 in the supplement [SW2016]) which only relies on the values $\|\mathcal{B}\|_{\mathrm{op}}$ and $\|\mathcal{B}\|_{\mathrm{Fr}} = \sqrt{\mathrm{tr}(\mathcal{B}^2)}$ for a random $p \times p$ matrix $\mathcal{B}$ given by

$$\mathcal{B} \stackrel{\mathrm{def}}{=} S^{-1/2}\bigl(S^{\flat} - S\bigr)S^{-1/2}. \tag{7.41}$$

Proposition 7.7.1. *There is a random set* Ω_n *with* $I\!\!P\bigl(\Omega_n\bigr) \geq 1 - 6/n$ *such that it holds on* Ω_n *with* Δ_n *given in* (7.33)*:*

$$\|\mathcal{B}\|_{\mathrm{op}} \leq \Delta_n, \qquad \|\mathcal{B}\|_{\mathrm{Fr}} \leq \sqrt{p}\,\Delta_n, \tag{7.42}$$

Proof. Define a $p \times n$ matrix $\mathcal{U} = S^{-1/2}\Psi\Sigma^{1/2}$ so that $\mathcal{U}\mathcal{U}^{\top} = \boldsymbol{I}_p$. We will use the decomposition

$$\Sigma^{-1/2}\breve{\boldsymbol{Y}} = \Sigma^{-1/2}(\boldsymbol{Y} - \Pi\boldsymbol{Y}) = \Sigma^{-1/2}(\boldsymbol{\varepsilon} - \Pi\boldsymbol{\varepsilon}) + \Sigma^{-1/2}(\boldsymbol{f} - \Pi\boldsymbol{f}) = \boldsymbol{\eta} + \boldsymbol{B}$$

with $\boldsymbol{\eta} \stackrel{\mathrm{def}}{=} \Sigma^{-1/2}(\boldsymbol{\varepsilon} - \Pi\boldsymbol{\varepsilon})$ and the result follows from Proposition D6 in the supplement [SW2016] with $\delta = \delta_{\Psi}$ and $\delta_n = \delta_{\Pi}$ and $y = \log n$ yielding $y + \log n = 2\log n$ and $y + \log p \leq 2\log n$, and thus $\Delta(y) \leq \Delta_n$.

On the set Ω_n, the claim (7.32) follows from $\|\mathcal{B}\|_{\mathrm{Fr}} \leq \sqrt{p}\,\Delta_n$ by Lemma E.1 in [SW2016] with $\boldsymbol{b} = \boldsymbol{b}^{\flat} = 0$.

7.7.4 Proof of Theorem 7.4.3

The result of Theorem 7.4.2 explains why the known bootstrap distribution can be used as a proxy for the unknown error distribution. However, it cannot be applied directly to (7.34) because the quantities $z^{\flat}_{m,m^{\circ}}(\mathtt{x})$ and $q^{\flat}_{m^{\circ}}$ are random and depend on the original data. This especially concerns the multiplicity correction $q^{\flat}_{m^{\circ}}$ which is based on the joint distribution of the vectors $\boldsymbol{\xi}^{\flat}_{m,m^{\circ}}$ from (7.19) and is defined in (7.21). The latter distribution is a Gaussian random measure in the bootstrap world. To cope with the problem of this cross-dependence, we apply the statement of Theorem F.1.1 in the Appendix. The underlying idea is to use geometric arguments to sandwich the random measure from (7.21) in two deterministic measures with high probability. The statement of Theorem 7.4.3 can be derived from Theorem F.1.1 using the bound $\|\mathcal{B}\|_{\mathrm{Fr}} \leq \sqrt{p}\,\Delta_n$ and $\|\mathcal{B}\|_{\mathrm{op}} \leq \Delta_n$ of Proposition 7.7.1 and conditioning on the set Ω_n.

7.7.5 Proof of Theorem 7.4.4

Due to Theorem 7.4.3, the bootstrap stochastic terms $\boldsymbol{\xi}^{\flat}_{m,m^{\circ}}$ nicely mimic (in distribution) their real world counterparts $\boldsymbol{\xi}_{m,m^{\circ}}$. The SmA procedure also involves the values

$\mathtt{p}_{m,m^\circ} = I\!\!E \| \boldsymbol{\xi}_{m,m^\circ} \|^2$, which are unknown and depend on the noise $\boldsymbol{\varepsilon}$. The bootstrap procedure utilizes their versions $\mathtt{p}^\flat_{m,m^\circ} = I\!\!E^\flat \| \boldsymbol{\xi}^\flat_{m,m^\circ} \|^2$. This is justified by the next lemma.

Lemma 7.7.1. *On the set* Ω_n *shown in Theorem 7.4.2, the values for* $\mathtt{p}^\flat_{m,m^\circ} \stackrel{\mathrm{def}}{=} \operatorname{tr}\bigl\{ \operatorname{Var}^\flat(\boldsymbol{\xi}^\flat_{m,m^\circ}) \bigr\}$ *and* $\lambda^\flat_{m,m^\circ} \stackrel{\mathrm{def}}{=} \lambda_{\max}\bigl\{ \operatorname{Var}^\flat(\boldsymbol{\xi}^\flat_{m,m^\circ}) \bigr\}$ *for all pairs* $m < m^\circ \in \mathcal{M}$ *fulfill*

$$
\begin{aligned}
\mathtt{p}_{m,m^\circ} \bigl(1 - \Delta_n \bigr) &\le \mathtt{p}^\flat_{m,m^\circ} \le \mathtt{p}_{m,m^\circ} \bigl(1 + \Delta_n \bigr) , \\
\lambda_{m,m^\circ} \bigl(1 - \Delta_n \bigr) &\le \lambda^\flat_{m,m^\circ} \le \lambda_{m,m^\circ} \bigl(1 + \Delta_n \bigr) .
\end{aligned}
$$

Proof. Similarly to the proof of Theorem 7.4.2, we use that $\boldsymbol{\xi}_{m,m^\circ} = \mathcal{T}_{m,m^\circ} \nabla$ and $\boldsymbol{\xi}^\flat_{m,m^\circ} = \mathcal{T}_{m,m^\circ} \nabla^\flat$ for the same deterministic linear mapping $\mathcal{T}_{m,m^\circ}$; see (7.39) and (7.40). On the set Ω_n the variances $S = \operatorname{Var}(\nabla)$ and $S^\flat = \operatorname{Var}(\nabla^\flat)$ are related by (7.42) for $\mathcal{B}$ from (7.41). This easily implies

$$
\bigl(1 - \Delta_n \bigr) S \le S^\flat \le \bigl(1 + \Delta_n \bigr) S
$$

and thus, the desired bounds follow.

The relation $\beta^\flat \ge \bigl(1 - \Delta_n \bigr)^{-1/2} \beta$ helps to bound on the set Ω_n

$$
\| \boldsymbol{b}_{m,m^\circ} \| \le \beta \sqrt{\mathtt{p}_{m,m^\circ}} \le \beta^\flat \sqrt{\mathtt{p}^\flat_{m,m^\circ}}
$$

and one can upper bound the probability of the event $\bigl\{ m^* \text{ is rejected} \bigr\}$ similarly to the case of known noise distribution. The oracle inequality (7.35) follows from the acceptance rule under the conditions that $\widehat{m} \le m^*$ and $\widehat{m}$ is accepted; cf. the proof of Theorem 7.3.2. The bound (7.36) follows from Lemma 7.7.1 and arguments of Theorem 7.3.3 applied to the bootstrap quantities $\mathfrak{z}^\flat_{m,m^\circ}$.

7.8 * 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 $\widetilde{\boldsymbol{\theta}}$ is given by $\widetilde{\boldsymbol{\theta}} = D^{-2} \Psi \boldsymbol{Y}$, 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 $\nabla^\flat = \Psi \mathcal{E}^\flat \boldsymbol{Y}$ 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 $\nabla^{\flat} = \Psi \mathcal{E}^{\flat} \boldsymbol{Y}$ given $\boldsymbol{Y}$ 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: $\nabla^{\flat}$ is normal because it is a linear combination of standard normal weights $e_i^{\flat} = w_i^{\flat} - 1$. 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 (7.29) 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.

8

* 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:

> $\varkappa^{\circ}$ 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:

> $\widehat{\varkappa}$ is the *simplest accepted* model.

Below we illustrate how this method works in two important examples: *anisotropic classes* and *subset selection* problems.

8.1 Subset selection procedure

Consider a linear model $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$. 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 $W = \boldsymbol{I}_p$ loss. Of course, these two problems coincide in the sequence space model with $n = p$ and $\Psi = \boldsymbol{I}_p$.

8.1.1 SmA procedure and multilevel synchronization

Let $\varkappa$ mean a subset of the entire index set $\{1, 2, \ldots, p\}$. We use the obvious notation $\varkappa \vee \varkappa'$ for the union, $\varkappa \wedge \varkappa^*$ for the overlap of two subsets $\varkappa$ and $\varkappa'$, $\varkappa' - \varkappa$ for the complement of $\varkappa$ within $\varkappa'$. Further we consider the usual partial ordering: $\varkappa' > \varkappa$ means that $\varkappa \subseteq \varkappa'$. The subset $\varkappa$ is good if there is no significant bias in its complement $\varkappa^c$. 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 $\mathbb{T}_{\varkappa,\varkappa^\circ}$ for $\varkappa > \varkappa^\circ$. The model-candidate $\varkappa^\circ$ is accepted if no test among $\mathbb{T}_{\varkappa,\varkappa^\circ}$ rejects the hypothesis of "no bias". Given the loss matrix W, the test statistic $\mathbb{T}_{\varkappa,\varkappa^\circ}$ reads as in the ordered case:

$$\mathbb{T}_{\varkappa,\varkappa^\circ} = \bigl\| W(\widetilde{\boldsymbol{\theta}}_{\varkappa} - \widetilde{\boldsymbol{\theta}}_{\varkappa^\circ}) \bigr\|. \tag{8.1}$$

The acceptance rule can be written as

$$\varkappa^\circ \text{ is accepted iff } \quad \mathbb{T}_{\varkappa,\varkappa^\circ} \le \mathfrak{z}_{\varkappa,\varkappa^\circ} \quad \forall \varkappa > \varkappa^\circ. \tag{8.2}$$

Now we discuss how the critical values $\mathfrak{z}_{\varkappa,\varkappa^\circ}$ can be fixed by *synchronization* (*multiplicity correction*) of the individual *tail functions*. We use the decomposition of the test statistic $\mathbb{T}_{\varkappa,\varkappa^\circ}$ from (8.1):

$$\mathbb{T}_{\varkappa,\varkappa^\circ} = \| \boldsymbol{\xi}_{\varkappa,\varkappa^\circ} + b_{\varkappa,\varkappa^\circ} \|.$$

The increase of complexity between $\varkappa^\circ$ and $\varkappa$ can be measured by via the variance of $\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}$. Namely, define $\mathtt{p}_{\varkappa,\varkappa^\circ}$ as expectation of $\|\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}\|^2$:

$$\mathtt{p}_{\varkappa,\varkappa^\circ} \stackrel{\text{def}}{=} \mathbb{E} \|\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}\|^2 = \operatorname{tr}\bigl\{ \operatorname{Var}\bigl(\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}\bigr) \bigr\}.$$

This value will be used in the bias-variance relation: the bias $b_{\varkappa,\varkappa^\circ}$ is insignificant if $\|b_{\varkappa,\varkappa^\circ}\|^2$ is smaller than the variance $\mathtt{p}_{\varkappa,\varkappa^\circ}$. More precisely, define a *good choice* $\varkappa^\circ$ as previously by "no significant bias" condition:

$$\|b_{\varkappa,\varkappa^\circ}\| \le \beta \mathtt{p}_{\varkappa,\varkappa^\circ}^{1/2} \quad \forall \varkappa > \varkappa^\circ. \tag{8.3}$$

Exercise 8.1.1. Let the true vector $\boldsymbol{\theta}^*$ contain only one positive entry θ_1^*. Describe the class of good models and the bias $b_{\varkappa,\varkappa^\circ}$ for the cases when $\varkappa^\circ$ contains the first component and when only $\varkappa$ contains the first component.

We aim at designing a procedure which accepts any such good model with a high probability. Suppose we are given for each pair $\varkappa > \varkappa^{\circ}$ a tail function $z_{\varkappa,\varkappa^{\circ}}(\mathtt{x})$ of the noise component $\|\boldsymbol{\xi}_{\varkappa,\varkappa^{\circ}}\|$ providing

$$I\!\!P\bigl(\|\boldsymbol{\xi}_{\varkappa,\varkappa^{\circ}}\| > z_{\varkappa,\varkappa^{\circ}}(\mathtt{x})\bigr) \le \mathrm{e}^{-\mathtt{x}}.$$

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 $\varkappa^{\circ}$ 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 $\mathtt{x}$ to control the overall rejection probability: define $q_{\varkappa^{\circ}}(\mathtt{x})$ by the condition

$$I\!\!P\biggl(\bigcup_{\varkappa\in\mathcal{M}(\varkappa^{\circ})}\Bigl\{\|\boldsymbol{\xi}_{\varkappa,\varkappa^{\circ}}\| > z_{\varkappa,\varkappa^{\circ}}\bigl(\mathtt{x}+q_{\varkappa^{\circ}}(\mathtt{x})\bigr)\Bigr\}\biggr) \le \mathrm{e}^{-\mathtt{x}}.$$

Denote $\mathtt{x}_{\varkappa^{\circ}} = \mathtt{x} + q_{\varkappa^{\circ}}(\mathtt{x})$ and apply the acceptance rule (8.2) with $\mathfrak{z}_{\varkappa,\varkappa^{\circ}}$ equal to such defined $z_{\varkappa,\varkappa^{\circ}}(\mathtt{x}_{\varkappa^{\circ}})$ after a small bias correction:

$$\mathfrak{z}_{\varkappa,\varkappa^{\circ}}(\beta) \stackrel{\mathrm{def}}{=} z_{\varkappa,\varkappa^{\circ}}(\mathtt{x}_{\varkappa^{\circ}}) + \beta\mathtt{p}_{\varkappa,\varkappa^{\circ}}^{1/2}.$$

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 $\varkappa^{\circ}$. For a given growing sequence $0 < \tau_1 < \tau_2 < \ldots < \tau_{\mathcal{K}}$, define

$$\mathcal{M}_m(\varkappa^{\circ}) \stackrel{\mathrm{def}}{=} \bigl\{\varkappa > \varkappa^{\circ} \colon |\varkappa^{\circ}| + \tau_{m-1} < |\varkappa| \le |\varkappa^{\circ}| + \tau_m\bigr\}$$

If $\tau_m = m$ for all m then

$$\mathcal{M}_m(\varkappa^{\circ}) \stackrel{\mathrm{def}}{=} \bigl\{\varkappa > \varkappa^{\circ} \colon |\varkappa| = |\varkappa^{\circ}| + m\bigr\}.$$

The corrections $q_{1,\varkappa^{\circ}}, q_{2,\varkappa^{\circ}}, \ldots, q_{m,\varkappa^{\circ}}$ can be defined step by step: first we fix the correction $q_{1,\varkappa^{\circ}} = q_{1,\varkappa^{\circ}}(\mathtt{x})$ for all $\varkappa \in \mathcal{M}_1(\varkappa^{\circ})$

$$I\!\!P\bigl(\mathcal{A}_1\bigr) \le \frac{1}{2}\mathrm{e}^{-\mathtt{x}}$$

with

$$\mathcal{A}_1 \stackrel{\mathrm{def}}{=} \bigcup_{\varkappa\in\mathcal{M}_1(\varkappa^{\circ})}\Bigl\{\|\boldsymbol{\xi}_{\varkappa,\varkappa^{\circ}}\| > z_{\varkappa,\varkappa^{\circ}}\bigl(\mathtt{x}+q_{1,\varkappa^{\circ}}\bigr)\Bigr\}.$$

With such defined $q_{1,\varkappa^{\circ}}$, define $q_{2,\varkappa^{\circ}} = q_{2,\varkappa^{\circ}}(\mathtt{x})$ such that

$$\begin{aligned} I\!\!P\bigl(\mathcal{A}_1 \cup \mathcal{A}_2\bigr) \; &\le \; \frac{3}{4} \mathrm{e}^{-\mathtt{x}}, \\ \mathcal{A}_2 \stackrel{\mathrm{def}}{=} & \bigcup_{\varkappa \in \mathcal{M}_2(\varkappa^\circ)} \Bigl\{ \| \boldsymbol{\xi}_{\varkappa,\varkappa^\circ} \| > z_{\varkappa,\varkappa^\circ}\bigl(\mathtt{x} + q_{1,\varkappa^\circ} + q_{2,\varkappa^\circ}\bigr) \Bigr\}. \end{aligned}$$

We continue this way and define $q_{m,\varkappa^\circ}$ by induction: if $q_{1,\varkappa^\circ}, \ldots, q_{m-1,\varkappa^\circ}$ are fixed then the correction for the set $\mathcal{M}_m(\varkappa^\circ)$ is selected as the sum $\mathtt{x} + q_{1,\varkappa^\circ} + \ldots + q_{m,\varkappa^\circ}$ to ensure

$$I\!\!P\bigl(\mathcal{A}_1 \cup \mathcal{A}_2 \cup \ldots \cup \mathcal{A}_m\bigr) \le (1 - 2^{-|m|}) \mathrm{e}^{-\mathtt{x}} \tag{8.4}$$

with

$$\mathcal{A}_m \stackrel{\mathrm{def}}{=} \bigcup_{\varkappa \in \mathcal{M}_m(\varkappa^\circ)} \Bigl\{ \| \boldsymbol{\xi}_{\varkappa,\varkappa^\circ} \| > z_{\varkappa,\varkappa^\circ}\bigl(\mathtt{x} + q_{1,\varkappa^\circ} + q_{2,\varkappa^\circ} + \ldots + q_{m,\varkappa^\circ}\bigr) \Bigr\}.$$

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

$$\mathfrak{z}_{\varkappa,\varkappa^\circ} = z_{\varkappa,\varkappa^\circ}(\mathtt{x} + q_{1,\varkappa^\circ} + q_{2,\varkappa^\circ} + \ldots + q_{m,\varkappa^\circ}) + \beta \mathtt{p}_{\varkappa,\varkappa^\circ}^{1/2}, \quad m = m(\varkappa). \tag{8.5}$$

Exercise 8.1.2. Describe the approach based on the Bonferroni correction. Explain the difference between the proposed multilevel synchronization and the Bonferroni correction.

Our selection rule chooses the smallest accepted model. It can be written as

$$\widehat{\varkappa} = \operatorname*{argmin}_{\varkappa^\circ \in \mathcal{M}} \bigl\{ |\varkappa^\circ| \colon \mathbb{T}_{\varkappa,\varkappa^\circ} \le \mathfrak{z}_{\varkappa,\varkappa^\circ}, \quad \varkappa \in \mathcal{M}(\varkappa^\circ) \bigr\}. \tag{8.6}$$

If there are many such $\varkappa^\circ$, one can select arbitrarily among them. The construction ensures that a good model in the sense (8.3) will be accepted with a high probability.

Theorem 8.1.1. *Let $\varkappa^\circ$ be a good model in the sense* (8.3)*. Then it holds for the SmA procedure with the critical values $\mathfrak{z}_{\varkappa,\varkappa^\circ}$ from* (8.4) *and* (8.5)

$$I\!\!P\bigl(\varkappa^\circ \textit{ is rejected}\bigr) \le \mathrm{e}^{-\mathtt{x}}.$$

Exercise 8.1.3. Prove the statement of Theorem 8.1.1.

Now we define the oracle choice $\varkappa^*$ as the simplest (in complexity $|\varkappa^*|$) model under the constraint (8.3):

$$\varkappa^* \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_{\varkappa^\circ \in \mathcal{M}} \bigl\{ |\varkappa^\circ| \colon \| b_{\varkappa,\varkappa^\circ} \| \le \beta \mathtt{p}_{\varkappa,\varkappa^\circ}^{1/2}, \quad \varkappa \in \mathcal{M}(\varkappa^\circ) \bigr\}. \tag{8.7}$$

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 $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}}$ for the SmA rule $\widehat{\varkappa}$.

Theorem 8.1.2. *It holds on a random set of probability at least* $1 - \mathrm{e}^{-\mathtt{x}}$

$$\| W (\widetilde{\boldsymbol{\theta}}_{\varkappa^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}}) \| \le \bar{\mathfrak{z}}_{\varkappa^*}$$

with $k^* = |\varkappa^*|$ *and* $\bar{\mathfrak{z}}_{\varkappa^*}$ *defined by*

$$\bar{\mathfrak{z}}_{\varkappa^*} \stackrel{\mathrm{def}}{=} \max_{|\varkappa| \le k^*} \bigl(\mathfrak{z}_{\varkappa \vee \varkappa^*, \varkappa^*} + \mathfrak{z}_{\varkappa \vee \varkappa^*, \varkappa} \bigr). \tag{8.8}$$

Proof. The construction and the propagation property ensures that $\varkappa^*$ is accepted with a high probability $1 - \mathrm{e}^{-\mathtt{x}}$. Below we focus on this case. Then the adaptive choice $\widehat{\varkappa}$ from (8.6) has to fulfill

$$|\widehat{\varkappa}| \le |\varkappa^*|.$$

Due to partial ordering, the models $\varkappa^*$ and $\widehat{\varkappa}$ are not directly comparable. We use the model $\breve{\varkappa} = \varkappa^* \vee \widehat{\varkappa}$ which contains both and is the smallest one with this property. As $\varkappa^*$ and $\widehat{\varkappa}$ are both accepted, it holds

$$\| W (\widetilde{\boldsymbol{\theta}}_{\breve{\varkappa}} - \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}}) \| \le \mathfrak{z}_{\breve{\varkappa}, \widehat{\varkappa}}, \quad \| W (\widetilde{\boldsymbol{\theta}}_{\breve{\varkappa}} - \widetilde{\boldsymbol{\theta}}_{\varkappa^*}) \| \le \mathfrak{z}_{\breve{\varkappa}, \varkappa^*} \, .$$

We conclude that

$$\| W (\widetilde{\boldsymbol{\theta}}_{\varkappa^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}}) \| \le \mathfrak{z}_{\breve{\varkappa}, \widehat{\varkappa}} + \mathfrak{z}_{\breve{\varkappa}, \varkappa^*} .$$

and the assertion follows.

8.1.2 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 $p \le n$. This dramatically simplifies the situation. Below we denote by $\Pi_{\varkappa}$ the projector in $\mathbb{R}^n$ onto the subspace corresponding to $\varkappa$. For a couple $\varkappa^{\circ} < \varkappa$, we also consider the projector $\Pi_{\varkappa, \varkappa^{\circ}} = \Pi_{\varkappa} - \Pi_{\varkappa^{\circ}}$ which projects to the orthogonal complement of $\Pi_{\varkappa^{\circ}}$ within the $\varkappa$-related subspace. Below we suppose a homogeneous noise $\boldsymbol{\varepsilon}$ with

$$\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n \, .$$

We also use that

$$\mathrm{p}_{\varkappa, \varkappa^{\circ}} = \mathbb{E} \| \boldsymbol{\xi}_{\varkappa, \varkappa^{\circ}} \|^2 = \sigma^2 |\varkappa - \varkappa^{\circ}|.$$

Theorem 8.1.3. *For the regression model* $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ *with homogeneous Gaussian errors and for* $W = \Psi^{\top}$, *the tail functions* $z_{\varkappa,\varkappa^{\circ}}(\mathtt{x})$ *only depends on* $|\varkappa - \varkappa^{\circ}|$. *Moreover, for* $\mathtt{x}$ *fixed, the multiplicity corrections* $q_{m,\varkappa^{\circ}} = q_{m,\varkappa^{\circ}}(\mathtt{x})$ *only depends on* $p - |\varkappa^{\circ}|$ *and on* τ_m.

Exercise 8.1.4. Prove the result. Use that $\Psi^{\top}\widetilde{\boldsymbol{\theta}}_{\varkappa} = \Pi_{\varkappa}\boldsymbol{Y}$ and

$$\sigma^{-2}\mathbb{T}^{2}_{\varkappa,\varkappa^{\circ}} = \sigma^{-2}\|\Psi^{\top}(\widetilde{\boldsymbol{\theta}}_{\varkappa} - \widetilde{\boldsymbol{\theta}}_{\varkappa^{\circ}})\|^{2} = \sigma^{-2}\|\Pi_{\varkappa,\varkappa^{\circ}}\boldsymbol{Y}\|^{2} \sim \chi^{2}_{|\varkappa-\varkappa^{\circ}|}.$$

The definition (8.7) of the oracle choice $\varkappa^{*}$ can be restated as

$$\varkappa^{*} \stackrel{\text{def}}{=} \operatorname*{argmin}_{\varkappa^{\circ}\in\mathcal{M}}\bigl\{|\varkappa^{\circ}| \colon \|b_{\varkappa,\varkappa^{\circ}}\| \le \beta|\varkappa - \varkappa^{\circ}|^{1/2}, \quad \varkappa \in \mathcal{M}(\varkappa^{\circ})\bigr\}.$$

Theorem 8.1.4. *The result of Theorem 8.1.2 holds with* $\bar{\mathfrak{z}}_{\varkappa^{*}} = \bar{\mathfrak{z}}_{k^{*}}$ *only depending on the cardinality* $k^{*} = |\varkappa^{*}|$. *Any* $\varkappa^{*}$ *with this cardinality can be used in definition (8.8).*

Exercise 8.1.5. Prove the result of Theorem 8.1.4.

Exercise 8.1.6. Compute an upper bound on $q_{m,\varkappa^{\circ}}$ using Bonferroni and Stirling formulas.

Exercise 8.1.7. Let Ψ be an identity matrix. Design an efficient algorithm for SmA subset selection based on ordered set of observations $\boldsymbol{Y}$.

8.1.3 Estimation loss

The analysis for the problem of estimation loss $W = \boldsymbol{I}_p$ is similar, but some nice features of the prediction loss do not apply here. We use

$$\begin{aligned} \widetilde{\boldsymbol{\theta}}_{\varkappa} &= \mathcal{S}_{\varkappa}\boldsymbol{Y}, \\ \mathcal{S}_{\varkappa} &= \bigl(\Psi_{\varkappa}\Psi_{\varkappa}^{\top}\bigr)^{-1}\Psi_{\varkappa}. \end{aligned}$$

Therefore,

$$\mathbb{T}_{\varkappa,\varkappa^{\circ}} = \|(\mathcal{S}_{\varkappa} - \mathcal{S}_{\varkappa^{\circ}})\boldsymbol{Y}\| = \|\boldsymbol{\xi}_{\varkappa,\varkappa^{\circ}} + b_{\varkappa,\varkappa^{\circ}}\|.$$

If the bias component vanishes, the distribution of this test statistic is completely described by the matrices $\Psi_{\varkappa}$ and $\Psi_{\varkappa^{\circ}}$ but in a more complicated way than for the prediction loss. Noise homogeneity allows us to define

$$\mathtt{p}_{\varkappa,\varkappa^{\circ}} = \mathbb{E}\|\boldsymbol{\xi}_{\varkappa,\varkappa^{\circ}}\|^{2} = \sigma^{2}\operatorname{tr}\bigl\{\bigl(\mathcal{S}_{\varkappa} - \mathcal{S}_{\varkappa^{\circ}}\bigr)\bigl(\mathcal{S}_{\varkappa} - \mathcal{S}_{\varkappa^{\circ}}\bigr)^{\top}\bigr\}.$$

The general results of Theorems 8.1.1 and 8.1.2 apply here for such defined values $\mathtt{p}_{\varkappa,\varkappa^\circ}$. 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.

8.1.4 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 $W(\widetilde{\boldsymbol{\theta}}_{\varkappa'} - \widetilde{\boldsymbol{\theta}}_{\varkappa})$ 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 $\widetilde{\phi}_{\varkappa} = W\widetilde{\boldsymbol{\theta}}_{\varkappa}$ fulfills

$$\widetilde{\phi}_{\varkappa} - \phi^* = W(\widetilde{\boldsymbol{\theta}}_{\varkappa} - \boldsymbol{\theta}^*) = W\mathcal{S}_{\varkappa}\boldsymbol{\varepsilon} + W(\mathcal{S}_{\varkappa}\boldsymbol{f}^* - \boldsymbol{\theta}^*) = \xi_{\varkappa} + b_{\varkappa},$$

where $\xi_{\varkappa} = W\mathcal{S}_{\varkappa}\boldsymbol{\varepsilon}$ is a zero mean random variable and $b_{\varkappa} = W(\mathcal{S}_{\varkappa}\boldsymbol{f}^* - \boldsymbol{\theta}^*)$ is the deterministic bias. The squared risk of $\widetilde{\phi}_{\varkappa}$ is given by the usual bias-variance decomposition:

$$\mathcal{R}_{\varkappa} = I\!\!E\bigl(\widetilde{\phi}_{\varkappa} - \phi^*\bigr)^2 = I\!\!E\bigl(\xi_{\varkappa} + b_{\varkappa}\bigr)^2 = b_{\varkappa}^2 + \operatorname{Var}(\xi_{\varkappa}) = b_{\varkappa}^2 + \mathtt{s}_{\varkappa}^2$$

with

$$\mathtt{s}_{\varkappa}^2 = \sigma^2 W\mathcal{S}_{\varkappa}\mathcal{S}_{\varkappa}^{\top}W^{\top}.$$

Monotonicity condition $\mathcal{S}_{\varkappa}\mathcal{S}_{\varkappa}^{\top} \geq \mathcal{S}_{\varkappa^\circ}\mathcal{S}_{\varkappa^\circ}^{\top}$ for $\varkappa > \varkappa^\circ$ yields monotonicity $\mathtt{s}_{\varkappa} \geq \mathtt{s}_{\varkappa^\circ}$ for the functional estimate. For each pair $\varkappa^\circ < \varkappa$

$$\widetilde{\phi}_{\varkappa} - \widetilde{\phi}_{\varkappa^\circ} = W\bigl(\widetilde{\boldsymbol{\theta}}_{\varkappa} - \widetilde{\boldsymbol{\theta}}_{\varkappa^\circ}\bigr) = W(\mathcal{S}_{\varkappa} - \mathcal{S}_{\varkappa^\circ})\boldsymbol{Y} = W\mathcal{S}_{\varkappa,\varkappa^\circ}\boldsymbol{Y}.$$

The variance of this difference reads as

$$\mathtt{s}_{\varkappa,\varkappa^\circ}^2 = \operatorname{Var}\bigl(\xi_{\varkappa,\varkappa^\circ}\bigr) = \sigma^2 W\mathcal{S}_{\varkappa,\varkappa^\circ}\,\mathcal{S}_{\varkappa,\varkappa^\circ}^{\top}W^{\top}.$$

The scaled test statistic $\mathbb{T}_{\varkappa,\varkappa^\circ}$ is given for $\varkappa > \varkappa^\circ$ by

$$\mathbb{T}_{\varkappa,\varkappa^\circ} = \mathtt{s}_{\varkappa,\varkappa^\circ}^{-1}|W\mathcal{S}_{\varkappa,\varkappa^\circ}\boldsymbol{Y}|.$$

One can use that the stochastic component $\xi_{\varkappa,\varkappa^\circ}$ of $\mathtt{s}_{\varkappa,\varkappa^\circ}^{-1}W\mathcal{S}_{\varkappa,\varkappa^\circ}\boldsymbol{Y}$ is standard normal. Thus, the multiplicity corrections are computed from the same bound (8.4) with $z_1(\cdot)$ in place of $z_{\varkappa,\varkappa^\circ}(\cdot)$. Moreover, $\mathtt{p}_{\varkappa,\varkappa^\circ} \equiv 1$, and the oracle definition reads as

$$\varkappa^* \stackrel{\text{def}}{=} \operatorname*{argmin}_{\varkappa^\circ \in \mathcal{M}} \bigl\{ |\varkappa^\circ| \colon \|b_{\varkappa,\varkappa^\circ}\| \leq \beta, \quad \varkappa \in \mathcal{M}(\varkappa^\circ) \bigr\}.$$

The general results of Theorems 8.1.1 and 8.1.2 apply here without any change. However, the involved values can be made more precise.

Exercise 8.1.8. State and prove an analog of Theorem 8.1.1 for the problem of linear functional estimation.

Exercise 8.1.9. Derive an upper bound on $q_{\varkappa,\varkappa^{\circ}}$ using Bonferroni multiplicity correction and Stirling formula.

8.1.5 Subset selection problem

The presented oracle bound of Theorem 8.1.2 claims that the risk of the adaptive estimate $\widehat{\boldsymbol{\theta}}$ is linked to the risk of the oracle $\widetilde{\boldsymbol{\theta}}_{\varkappa^*}$. However, it tells nothing about the selected set $\widehat{\varkappa}$. Now we discuss this issue of choosing the active set of important features represented by non-zero entries of $\boldsymbol{\theta}^*$.

Note that this problem has to be put in a right way: one can suppose that $\boldsymbol{\theta}^*$ is sparse and only significant entries are non-zero. Alternatively, one tries to find an approximating model with another vector $\boldsymbol{\theta}_s^*$ having a sparse representation and delivering nearly the same approximation and prediction quality. We follow our oracle result and define $\varkappa^*$ by (8.7). 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 $|\widehat{\varkappa}| \le |\varkappa^*|$. 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 $\Psi = \boldsymbol{I}_p$ 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 $\|\Pi_{\varkappa,\varkappa^*}\boldsymbol{\theta}^*\|$ does not exceed $\beta|\varkappa - \varkappa^*|^{1/2}$ for any $\varkappa > \varkappa^*$.

Let now $\varkappa$ be any subset not equal to $\varkappa^*$ with $|\varkappa| \le |\varkappa^*|$. 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 (8.5):

$$z^{+}_{\varkappa,\varkappa^{\circ}} = z_{\varkappa,\varkappa^{\circ}}(\mathtt{x} + q_{1,\varkappa^{\circ}} + q_{2,\varkappa^{\circ}} + \ldots + q_{m,\varkappa^{\circ}}), \qquad \varkappa \in \mathcal{M}_m(\varkappa^{\circ}).$$

Theorem 8.1.5. *Let* $\varkappa$ *be such that*

$$\|b_{\varkappa \vee \varkappa^*, \varkappa}\| \geq \|\Pi_{\varkappa^* \setminus \varkappa} \boldsymbol{\theta}^*\| > 2 z^{+}_{\varkappa \vee \varkappa^*, \varkappa} + \beta \, |\varkappa^* \setminus \varkappa|^{1/2}, \tag{8.9}$$

and let $\mathcal{M}^{\circ}(\varkappa^*)$ *be the collection of all such* $\varkappa$ *. Then*

$$I\!\!P\bigl(\textit{any of } \varkappa \in \mathcal{M}^{\circ}(\varkappa^*) \textit{ is accepted}\bigr) \leq \mathrm{e}^{-\mathtt{x}}.$$

Proof. We just apply the acceptance rule for $\varkappa$ requiring

$$\|\widetilde{\boldsymbol{\theta}}_{\varkappa \vee \varkappa^*} - \widetilde{\boldsymbol{\theta}}_{\varkappa}\| \leq \mathfrak{z}_{\varkappa \vee \varkappa^*, \varkappa} = z^{+}_{\varkappa \vee \varkappa^*, \varkappa} + \beta \, |\varkappa^* \setminus \varkappa|^{1/2}. \tag{8.10}$$

The usual decomposition of $\widetilde{\boldsymbol{\theta}}_{\varkappa}$ implies on a dominating set $\Omega(\mathtt{x})$ by (8.4)

$$\begin{aligned} \|\widetilde{\boldsymbol{\theta}}_{\varkappa \vee \varkappa^*} - \widetilde{\boldsymbol{\theta}}_{\varkappa}\| &\geq \|b_{\varkappa \vee \varkappa^*, \varkappa}\| - \|\boldsymbol{\xi}_{\varkappa \vee \varkappa^*, \varkappa}\| \\ &\geq \|b_{\varkappa \vee \varkappa^*, \varkappa}\| - z^{+}_{\varkappa \vee \varkappa^*, \varkappa}. \end{aligned} \tag{8.11}$$

It remains to check that the inequalities (8.10) and (8.11) are incompatible under (8.9).

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 8.1.1. *Let* $\varkappa_0^*$ *fulfill*

$$\|\Pi_{\varkappa_0^*} \boldsymbol{\theta}^*\| \geq \bar{\mathfrak{z}}_{\varkappa^*} \stackrel{\text{def}}{=} \max_{|\varkappa| \leq |\varkappa^*|} \bigl\{ 2 z_{\varkappa \vee \varkappa^*, \varkappa} + \beta \, |\varkappa^* \setminus \varkappa|^{1/2} \bigr\}.$$

Then

$$I\!\!P\bigl(\varkappa_0^* \cap \widehat{\varkappa} = \emptyset\bigr) \leq \mathrm{e}^{-\mathtt{x}}.$$

In particular, any coefficient θ_j^* *of* $\boldsymbol{\theta}^*$ *with* $|\theta_j^*| > \bar{\mathfrak{z}}_{\varkappa^*}$ *will be included in* $\widehat{\varkappa}$ *with a probability at least* $1 - \mathrm{e}^{-\mathtt{x}}$ *.*

Exercise 8.1.10. Prove the statement of Corollary 8.1.1.

8.1.6 FDR control and SmA with a block-rule

The main challenge in applying the SmA idea is that it produces a rather conservative procedure because of the multiplicity correction. Indeed, even for the first step of the procedure, we have exactly $p - |\varkappa^{\circ}|$ alternative models of cardinality $p + 1$, yielding the first step correction of order $\log(p)$. Choosing a large critical value reduces the sensitivity of the procedure. One can miss some important features, whose impact is below the threshold $\bar{\mathfrak{z}}_{\varkappa^{\circ}}$. Consider below another rule, which leads to a less conservative procedure. The idea is that any subset-candidate $\varkappa^{\circ}$ is tested against all models which are larger by $\alpha\,|\varkappa^{\circ}|$.

The FDR approach allows to select a model $\varkappa^{\circ}$ which contains $\alpha|\varkappa^{\circ}|$ "false positive" elements. We want to design a new procedure which is less conservative and still keep a prescribed probability that the true model will not be rejected.

Let $\varkappa^{\circ}$ be a subset-candidate. The FDR rule assumes that an α-fraction of the selected subset contains only noisy features. This, of course, changes the acceptance rule. A subset-candidate $\varkappa^{\circ}$ is "good" if there is its subset of a cardinality $(1-\alpha)|\varkappa^{\circ}|$ such that this subset is not rejected against any larger model $\varkappa$. Equivalently,

8.2 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 $\widetilde{\boldsymbol{\theta}}_{\varkappa}$. When only one component of $\varkappa$ is varying and the others 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 $\widetilde{\boldsymbol{\theta}}_{\varkappa}$ with a partial ordering: if $\varkappa > \varkappa^{\circ}$ then

$$\operatorname{Var}\bigl(W\widetilde{\boldsymbol{\theta}}_{\varkappa}\bigr) > \operatorname{Var}\bigl(W\widetilde{\boldsymbol{\theta}}_{\varkappa^{\circ}}\bigr).$$

This particularly implies

$$\mathtt{p}_{\varkappa} \overset{\text{def}}{=} \operatorname{tr}\bigl\{\operatorname{Var}\bigl(W\widetilde{\boldsymbol{\theta}}_{\varkappa}\bigr)\bigr\} > \mathtt{p}_{\varkappa^{\circ}} \overset{\text{def}}{=} \operatorname{tr}\bigl\{\operatorname{Var}\bigl(W\widetilde{\boldsymbol{\theta}}_{\varkappa^{\circ}}\bigr)\bigr\}.$$

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 $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ with homogeneous Gaussian error $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$ and consider a collection of pairwise test statistics

$$\mathbb{T}_{\varkappa,\varkappa^{\circ}} = \|W(\widetilde{\boldsymbol{\theta}}_{\varkappa} - \widetilde{\boldsymbol{\theta}}_{\varkappa^{\circ}})\| = \|\boldsymbol{\xi}_{\varkappa,\varkappa^{\circ}} + b_{\varkappa,\varkappa^{\circ}}\|.$$

As previously, define

$$\mathtt{p}_{\varkappa,\varkappa^{\circ}} = \operatorname{tr}\bigl\{\operatorname{Var}\bigl(\boldsymbol{\xi}_{\varkappa,\varkappa^{\circ}}\bigr)\bigr\}.$$

The acceptance rule can be written as

$$\varkappa^{\circ} \text{ is accepted iff } \quad \mathbb{T}_{\varkappa,\varkappa^{\circ}} \le \mathfrak{z}_{\varkappa,\varkappa^{\circ}} \quad \forall \varkappa \in \mathcal{M}(\varkappa^{\circ}), \tag{8.12}$$

where $\mathcal{M}(\varkappa^{\circ}) = \bigl\{\varkappa \in \mathcal{M} \colon \varkappa > \varkappa^{\circ}\bigr\}$. In words, $\varkappa^{\circ}$ is accepted if it is competitive with any larger model $\varkappa$. Now we discuss how the critical values $\mathfrak{z}_{\varkappa,\varkappa^{\circ}}$ 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 $q_{\varkappa^{\circ}}(\mathtt{x})$ by the condition

$$I\!\!P\biggl(\bigcup_{\varkappa \in \mathcal{M}(\varkappa^{\circ})} \Bigl\{ \|\boldsymbol{\xi}_{\varkappa,\varkappa^{\circ}}\| > z_{\varkappa,\varkappa^{\circ}}\bigl(\mathtt{x} + q_{\varkappa^{\circ}}(\mathtt{x})\bigr) \Bigr\} \biggr) \le \mathrm{e}^{-\mathtt{x}}. \tag{8.13}$$

Denote $\mathtt{x}_{\varkappa^{\circ}} = \mathtt{x} + q_{\varkappa^{\circ}}(\mathtt{x})$ and apply the acceptance rule (8.2) with $\mathfrak{z}_{\varkappa,\varkappa^{\circ}}$ equal to such defined $z_{\varkappa,\varkappa^{\circ}}(\mathtt{x}_{\varkappa^{\circ}})$ after a small bias correction:

$$\mathfrak{z}_{\varkappa,\varkappa^{\circ}} \stackrel{\text{def}}{=} z_{\varkappa,\varkappa^{\circ}}(\mathtt{x}_{\varkappa^{\circ}}) + \beta \mathtt{p}_{\varkappa,\varkappa^{\circ}}^{1/2}. \tag{8.14}$$

A good choice $\varkappa^{\circ}$ can be defined as previously by "no significant bias" condition:

$$\|b_{\varkappa,\varkappa^{\circ}}\| \le \beta \mathtt{p}_{\varkappa,\varkappa^{\circ}}^{1/2} \quad \varkappa \in \mathcal{M}(\varkappa^{\circ}). \tag{8.15}$$

The construction ensures that a good model will be accepted with a high probability.

Theorem 8.2.1. *Let $\varkappa^{\circ}$ be a good model in the sense (8.15). Then it holds for the acceptance rule (8.12) with the critical values $\mathfrak{z}_{\varkappa,\varkappa^{\circ}}$ from (8.13) and (8.14)*

$$I\!\!P\bigl(\varkappa^{\circ} \textit{ is rejected}\bigr) \le \mathrm{e}^{-\mathtt{x}}.$$

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 $\mathtt{p}_{\varkappa^{\circ}}$ which is proportional to $|\varkappa^{\circ}|$ for the sequence space model. Define the oracle choice $\varkappa^{*}$ as the smallest (in complexity $\mathtt{p}_{\varkappa^{\circ}}$) model under the constraint (8.15):

$$\varkappa^{*} \stackrel{\text{def}}{=} \operatorname*{argmin}_{\varkappa^{\circ} \in \mathcal{M}} \bigl\{ \mathtt{p}_{\varkappa^{\circ}} \colon \|b_{\varkappa,\varkappa^{\circ}}\| \le \beta \mathtt{p}_{\varkappa,\varkappa^{\circ}}^{1/2}, \quad \varkappa \in \mathcal{M}(\varkappa^{\circ}) \bigr\}.$$

Our selection rule chooses the smallest accepted model. It can be written as

$$\widehat{\varkappa} = \operatorname*{argmin}_{\varkappa^{\circ} \in \mathcal{M}} \bigl\{ \mathtt{p}_{\varkappa^{\circ}} \colon \mathbb{T}_{\varkappa,\varkappa^{\circ}} \le \mathfrak{z}_{\varkappa,\varkappa^{\circ}}, \quad \varkappa \in \mathcal{M}(\varkappa^{\circ}) \bigr\}. \tag{8.16}$$

The oracle bound compares the risk of the oracle estimate $\mathcal{R}_{\varkappa^{*}}$ with the risk of the adaptive estimate $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}}$ for the SmA rule $\widehat{\varkappa}$. Below for two given models $\varkappa$ and $\varkappa^{\circ}$, 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 8.2.2. *It holds on a random set of probability at least* $1 - \mathrm{e}^{-\mathtt{x}}$

$$\| W(\widetilde{\boldsymbol{\theta}}_{\varkappa^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}}) \| \le \bar{\mathfrak{z}}_{\varkappa^*}$$

where $\bar{\mathfrak{z}}_{\varkappa^*}$ *is defined by*

$$\bar{\mathfrak{z}}_{\varkappa^*} \stackrel{\text{def}}{=} \max_{\mathtt{p}_{\varkappa} \le \mathtt{p}_{\varkappa^*}} \bigl(\mathfrak{z}_{\varkappa \vee \varkappa^*, \varkappa^*} + \mathfrak{z}_{\varkappa \vee \varkappa^*, \varkappa} \bigr).$$

Proof. The construction and the propagation property ensures that $\varkappa^*$ is accepted with a high probability $1 - \mathrm{e}^{-\mathtt{x}}$. Below we focus on this case. Then the adaptive choice $\widehat{\varkappa}$ from (8.16) has to fulfill

$$\mathtt{p}_{\widehat{\varkappa}} \le \mathtt{p}_{\varkappa^*}.$$

Consider $\breve{\varkappa} = \varkappa^* \vee \widehat{\varkappa}$ which contains both $\widehat{\varkappa}$ and $\varkappa^*$. As $\varkappa^*$ and $\widehat{\varkappa}$ are both accepted, it holds

$$\| W(\widetilde{\boldsymbol{\theta}}_{\breve{\varkappa}} - \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}}) \| \le \mathfrak{z}_{\breve{\varkappa}, \widehat{\varkappa}}, \quad \| W(\widetilde{\boldsymbol{\theta}}_{\breve{\varkappa}} - \widetilde{\boldsymbol{\theta}}_{\varkappa^*}) \| \le \mathfrak{z}_{\breve{\varkappa}, \varkappa^*} .$$

Therefore,

$$\| W(\widetilde{\boldsymbol{\theta}}_{\varkappa^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}}) \| \le \mathfrak{z}_{\breve{\varkappa}, \widehat{\varkappa}} + \mathfrak{z}_{\breve{\varkappa}, \varkappa^*},$$

and the assertion follows.

The value $\bar{\mathfrak{z}}_{\varkappa^*}$ is the “payment for adaptation”, and it can be quite large relative to the oracle standard deviation $\mathtt{p}_{\varkappa^*}^{1/2}$. 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 $\mathfrak{z}_{\varkappa \vee \varkappa^*, \varkappa^*}$ 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 $\mathcal{M}^{\circ}(\varkappa^*)$.

9

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 $\boldsymbol{Y} = \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ in which the parameter dimension p can be very large. The empirical risk is just the squared norm of the difference $\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}$. The penalized procedure tries to minimize this empirical risk penalized by the complexity of the vector $\boldsymbol{\theta}$ used for prediction:

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \bigl\{ \| \boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta} \|^2 + \mathfrak{t}(\boldsymbol{\theta}) \bigr\} \tag{9.1}$$

for a penalty function $\mathfrak{t}(\boldsymbol{\theta})$.

The roughness penalty has been already discussed in Chapter 4. The resulting estimate is again linear and the general approach of linear model selection continues to apply. Here we consider two special choices of $\mathfrak{t}(\boldsymbol{\theta})$ which are essentially non-linear. The complexity penalty $\mathfrak{t}(\boldsymbol{\theta}) = \mathtt{C} \|\boldsymbol{\theta}\|_0$ just counts the number of non-zero $\boldsymbol{\theta}$-coefficients. The sparse penalty $\mathfrak{t}(\boldsymbol{\theta}) = \mathtt{C} \|\boldsymbol{\theta}\|_q^q$ use the q-norm of $\boldsymbol{\theta}$ for some $q < 2$. The most popular sparse penalty corresponds to $q = 1$.

9.1 Complexity penalization

This section considers the important special case of penalization by complexity. The famous Akaiki criteria is a special case of such penalization. We offer another viewpoint based on the SmA idea. As previously in Section **??**, for a subset $\varkappa$ of the index set $\bigl\{1, \ldots, p\bigr\}$, the estimate $\widetilde{\boldsymbol{\theta}}_{\varkappa}$ is the corresponding projection MLE:

$$\widetilde{\boldsymbol{\theta}}_{\varkappa} = \bigl(\boldsymbol{\Psi}_{\varkappa} \boldsymbol{\Psi}_{\varkappa}^{\top} \bigr)^{-1} \boldsymbol{\Psi}_{\varkappa} \boldsymbol{Y}.$$

Theorem 9.1.1. *Let* $\mathfrak{t}(\boldsymbol{\theta}) = \mathtt{C} \|\boldsymbol{\theta}\|_0$. *Then the solution* $\widehat{\boldsymbol{\theta}}$ *of the problem* (9.1) *satisfies*

$$\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}},$$

where

$$\widehat{\varkappa} = \operatorname*{argmin}_{\varkappa} \bigl\{ \|\widetilde{\varepsilon}_{\varkappa}\|^2 + \mathtt{C}\,|\varkappa| \bigr\}, \tag{9.2}$$

with

$$\widetilde{\varepsilon}_{\varkappa} = \boldsymbol{Y} - \boldsymbol{\Psi}_{\varkappa}^{\top} \widetilde{\boldsymbol{\theta}}_{\varkappa} = \bigl(I_n - \Pi_{\varkappa} \bigr) \boldsymbol{Y}$$

and $|\varkappa|$ *means the cardinality of* $\varkappa$ *or, equivalently the number of coefficients in the support set* $\varkappa$ *.*

Exercise 9.1.1. Prove the result of Theorem 9.1.1.

The unbiased risk estimation procedure corresponds to the special choice of constant $\mathtt{C} = 2\sigma^2$. This results in selecting a proper subset $\varkappa$ which provides a reasonable fit under complexity constraint:

$$\widehat{\varkappa} = \operatorname*{argmin}_{\varkappa} \bigl\{ \|\widetilde{\varepsilon}_{\varkappa}\|^2 + 2\sigma^2 |\varkappa| \bigr\}.$$

This procedure faces two essential problems when the dimension p becomes large. One of them is algorithmic: the procedure requires to compute the MLE $\widetilde{\boldsymbol{\theta}}_{\varkappa}$ and the related empirical risk for any subset $\varkappa$; 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.

9.1.1 Orthonormal case

Let $\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top} = \boldsymbol{I}_p$. Then the procedure can be reduced to thresholding of individual Fourier coefficients $\widetilde{\theta}_j = \psi_j \boldsymbol{Y}$, where ψ_j denotes the j th row of $\boldsymbol{\Psi}$.

Theorem 9.1.2. *Let* $n \ge p$ *and the matrix* $\boldsymbol{\Psi}$ *be orthonormal, that is,* $\boldsymbol{\Psi}\boldsymbol{\Psi}^{T} = \boldsymbol{I}_p$ *. Then the active set* $\widehat{\varkappa}$ *from* (9.2) *is given by hard thresholding*

$$\widehat{\varkappa} = \bigl\{ j \colon |\widetilde{\theta}_j| \ge \lambda \bigr\}$$

for $\widetilde{\boldsymbol{\theta}} = \Psi \boldsymbol{Y}$ *and a proper* $\lambda > 0$ *. The corresponding hard thresholding estimate* $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_j)$ *reads as*

$$\widehat{\theta}_j = \begin{cases} \psi_j \boldsymbol{Y}, & |\psi_j \boldsymbol{Y}| > \lambda, \\ 0, & \textit{otherwise}. \end{cases} \tag{9.3}$$

Here ψ_j *is the j'th row of* $\boldsymbol{\Psi}$ *.*

Exercise 9.1.2. Prove the result (9.3).

Let $\boldsymbol{\theta}^{*}$ be the true parameter vector and $\varkappa^{*}$ be the corresponding active set of indices for non-vanishing coefficients θ_{j}^{*}. The critical question is how the true active set $\varkappa^{*}$ and the selected active set $\widehat{\varkappa}$ relate to each other.

Theorem 9.1.3. *Let* Ψ *be orthonormal and let* $\widehat{\varkappa}$ *be selected by the rule* (9.3). *Let* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^{2} I_{n})$. *If* $\lambda > \sigma \sqrt{2 \log(p + \mathtt{x})}$, *then on a random set* $\Omega_{\mathtt{x}}$ *with* $I\!P(\Omega_{\mathtt{x}}) \geq 1 - \mathrm{e}^{-\mathtt{x}}$, *it holds*

$$\widehat{\varkappa} \subseteq \varkappa^{*}. \tag{9.4}$$

Exercise 9.1.3. Prove the inclusion (9.4). Use that

$$I\!P\left(\max_{j \leq p} |\psi_{j}^{\top} \boldsymbol{\varepsilon}| \geq \sigma \sqrt{2 \log(p + \mathtt{x})} \right) \leq \mathrm{e}^{-\mathtt{x}}.$$

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 $2\sigma^{2}|\varkappa|$ 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 (9.1) or equivalently (9.2) with a data-driven constant $\mathtt{C}$. Then we extend it to a more sophisticated non-linear choice of the penalty function $\mathtt{t}(\boldsymbol{\theta})$. Finally we discuss the saddle-point bivariate model selection.

9.1.2 Penalty tuning using propagation condition

Complexity penalization requires to fix a constant $\mathtt{C}$ in (9.2), 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 $\varkappa^{\circ}$ which is not too complex, that is, $|\varkappa^{\circ}|$ is small relative to the sample size n and the total dimension p. One says in such cases that $\varkappa^{\circ}$ is *sparse*. Further, this choice $\varkappa^{\circ}$ is *good* if the coefficients θ_{j}^{*} for $j \notin \varkappa^{\circ}$ are nearly zero. We aim at designing a data-driven procedure such that the criterium (9.2) keeps $\varkappa^{\circ}$ alive in competition with all larger models. This precisely means that

$$\|\widetilde{\varepsilon}_{\varkappa}\|^{2} + \mathtt{C}|\varkappa| \geq \|\widetilde{\varepsilon}_{\varkappa^{\circ}}\|^{2} + \mathtt{C}|\varkappa^{\circ}|, \quad \varkappa > \varkappa^{\circ}.$$

This inequality has to be verified for all $\varkappa > \varkappa^{\circ}$ with a high probability. Rearranging yields in view of $\widetilde{\boldsymbol{\varepsilon}}_{\varkappa} = (\boldsymbol{I}_n - \Pi_{\varkappa})\boldsymbol{Y}$

$$\|\widetilde{\boldsymbol{\varepsilon}}_{\varkappa^{\circ}}\|^2 - \|\widetilde{\boldsymbol{\varepsilon}}_{\varkappa}\|^2 = \|\Pi_{\varkappa,\varkappa^{\circ}}\boldsymbol{Y}\|^2 = \|\boldsymbol{\Psi}^{\top}(\widetilde{\boldsymbol{\theta}}_{\varkappa} - \widetilde{\boldsymbol{\theta}}_{\varkappa^{\circ}})\|^2 \le \mathtt{C}\bigl(|\varkappa| - |\varkappa^{\circ}|\bigr). \tag{9.5}$$

If we knew the noise distribution then we can fix the constant $\mathtt{C}$ in the "pure noise" situation. Indeed, the underlying structural assumption means that there is no significant signal θ_j^* for $j \notin \varkappa^{\circ}$, and hence, one can simply ignore such signal and consider $\theta_j^* \equiv 0$ in the complement of $\varkappa^{\circ}$. This leads to the propagation condition

$$I\!P\Bigl(\bigcup_{\varkappa > \varkappa^{\circ}} \bigl\{\|\Pi_{\varkappa,\varkappa^{\circ}}\boldsymbol{\varepsilon}\|^2 > \mathtt{C}_0\bigl(|\varkappa| - |\varkappa^{\circ}|\bigr)\bigr\}\Bigr) \le \mathrm{e}^{-\mathtt{x}} \tag{9.6}$$

for a constant $\mathtt{C}_0$. In the case of a homogeneous noise, it is obvious that this condition becomes stronger if the set $\varkappa^{\circ}$ is taken smaller. The hardest case corresponds to the empty set $\varkappa^{\circ}$, yielding the constraint

$$I\!P\Bigl(\bigcup_{\varkappa}\bigl\{\|\Pi_{\varkappa}\boldsymbol{\varepsilon}\|^2 > \mathtt{C}_0|\varkappa|\bigr\}\Bigr) \le \mathrm{e}^{-\mathtt{x}}. \tag{9.7}$$

If the dimension of $\varkappa$ only slightly higher than the dimension of $\varkappa^{\circ}$, the inequality $\|\Pi_{\varkappa,\varkappa^{\circ}}\boldsymbol{\varepsilon}\|^2 > \mathtt{C}_0\bigl(|\varkappa| - |\varkappa^{\circ}|\bigr)$ would require a very large constant $\mathtt{C}_0$. At the same time, this constant rapidly stabilizes if $|\varkappa| - |\varkappa^{\circ}|$ exceeds some prescribed value. This suggests to extend the condition (9.6) (resp. (9.7)): given $\tau \ge 1$

$$I\!P\Bigl(\bigcup_{\varkappa \in \mathcal{M}_{\tau}(\varkappa^{\circ})} \bigl\{\|\Pi_{\varkappa,\varkappa^{\circ}}\boldsymbol{\varepsilon}\|^2 > \mathtt{C}_0\bigl(|\varkappa| - |\varkappa^{\circ}|\bigr)\bigr\}\Bigr) \le \mathrm{e}^{-\mathtt{x}}. \tag{9.8}$$

Here $\mathcal{M}_{\tau}(\varkappa^{\circ}) = \bigl\{\varkappa > \varkappa^{\circ} \colon |\varkappa| \ge |\varkappa^{\circ}| + \tau\bigr\}$ 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. (8.4).

Now suppose that such a constant $\mathtt{C}_0 = \mathtt{C}_0(\tau)$ is fixed. Then the procedure can be applied with

$$\mathtt{C} \stackrel{\mathrm{def}}{=} \mathtt{C}_0(\tau) + \beta,$$

where β appears in the definition of a "good" choice: $\varkappa^{\circ}$ is good if

$$\|\Pi_{\varkappa,\varkappa^{\circ}}\boldsymbol{f}^*\|^2 \le \beta\bigl(|\varkappa| - |\varkappa^{\circ}|\bigr). \tag{9.9}$$

We now bound the difference between $\|\Pi_{\varkappa,\varkappa^{\circ}}\boldsymbol{\varepsilon}\|^2$ and its expectation using Bernstein-type inequality for quadratic forms; see Corollary C.1.2 in Section **??**: for homogeneous Gaussian noise $\boldsymbol{\varepsilon}$ and $m \stackrel{\mathrm{def}}{=} |\varkappa| - |\varkappa^{\circ}|$ with $\alpha > 0$

$$I\!P\bigl(\sigma^{-1}\|\Pi_{\varkappa,\varkappa^{\circ}}\boldsymbol{\varepsilon}\| > z(m,\mathtt{x})\bigr) \le \mathrm{e}^{-\mathtt{x}},$$

where

$$z^2(m, \mathtt{x}) \le m + 2\sqrt{m\,\mathtt{x}} + 2\mathtt{x}.$$

Given $\mathtt{C}_0$, define $\mathtt{x}_m$ by

$$z^2(m, \mathtt{x}_m) = \mathtt{C}_0\, m.$$

Then, with $N_m = \#\mathcal{M}_m(\varkappa^\circ)$ for $\mathcal{M}_m(\varkappa^\circ) \stackrel{\text{def}}{=} \bigl\{\varkappa > \varkappa^\circ \colon |\varkappa - \varkappa^\circ| = m\bigr\}$

$$\begin{aligned} & I\!\!P\Bigl(\bigcup_{\varkappa \in \mathcal{M}_\tau(\varkappa^\circ)} \{\sigma^{-2}\|\Pi_{\varkappa,\varkappa^\circ}\boldsymbol{\varepsilon}\|^2 > \mathtt{C}_0 m\}\Bigr) \\ & \qquad \le \sum_{m=\tau}^{p} \sum_{\varkappa \in \mathcal{M}_m(\varkappa^\circ)} I\!\!P\bigl(\sigma^{-2}\|\Pi_{\varkappa,\varkappa^\circ}\boldsymbol{\varepsilon}\|^2 > z^2(m,\mathtt{x}_m)\bigr) \le \sum_{m=\tau}^{p} N_m \mathrm{e}^{-\mathtt{x}_m}. \end{aligned}$$

The Stirling formula yields $N_m \le (p/m)^m$ and the sum can be bounded for $\alpha > \log(p/\tau)$.

Theorem 9.1.4. *Consider a linear model with a homogeneous Gaussian noise* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. *If* $\widehat{\varkappa}$ *is defined by* (9.2) *with* $\mathtt{C} = \mathtt{C}_0 + \beta$, *then for any good model* $\varkappa^\circ$, *it holds*

$$I\!\!P\bigl(\varkappa^\circ \text{ is rejected}\bigr) \le \sum_{m=\tau}^{p} N_m \mathrm{e}^{-\mathtt{x}_m}.$$

9.1.3 Oracle inequality for $\widehat{\varkappa}$-choice

Let $\varkappa_0 \supseteq \operatorname{supp}(\boldsymbol{\theta}^*)$ and let $\widehat{\varkappa}$ be the selected model. First we bound the probability that $\widehat{\varkappa}$ is not contained in $\varkappa_0$. Consider $\overline{\varkappa} = \widehat{\varkappa} \setminus \varkappa_0$. The definition of $\widehat{\varkappa}$ ensures by (9.5) that

$$\bigl\|\boldsymbol{\Psi}^\top\bigl(\widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}} - \widetilde{\boldsymbol{\theta}}_{\overline{\varkappa}}\bigr)\bigr\|^2 \ge \mathtt{C}(|\widehat{\varkappa}| - |\overline{\varkappa}|).$$

But

$$\|\boldsymbol{\Psi}^\top\bigl(\widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}} - \widetilde{\boldsymbol{\theta}}_{\overline{\varkappa}}\bigr)\|^2 = \|\Pi_{\widehat{\varkappa}\vee\varkappa_0,\varkappa_0}\boldsymbol{Y}\|^2 = \|\Pi_{\widehat{\varkappa}\vee\varkappa_0,\varkappa_0}\boldsymbol{\varepsilon}\|^2$$

and by (9.6), this event can only happen with a very small probability. Otherwise $\widehat{\varkappa} \subseteq \varkappa_0$ and we obtain by the acceptance rule

$$\bigl\|\boldsymbol{\Psi}^\top\bigl(\widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}} - \widetilde{\boldsymbol{\theta}}_{\varkappa_0}\bigr)\bigr\|^2 \le \mathtt{C}(|\varkappa_0| - |\widehat{\varkappa}|) \le \mathtt{C}|\varkappa_0|. \tag{9.10}$$

Theorem 9.1.5. *Consider a linear model with a homogeneous Gaussian noise* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$ *. Let* $\varkappa_0$ *be a good model due to* (9.9)*. If* $\mathtt{C}$ *is selected to ensure the propagation condition* (9.8) *then the probability of the event* $\widehat{\varkappa} \le \varkappa_0$ *is at least* $1 - \mathrm{e}^{-\mathtt{x}}$ *and on this event, it holds* (9.10)*.*

9.1.4 Bootstrap based tuning of penalty

Now we discuss the situation when the noise distribution is unknown. Then the relation (9.7) cannot be used for fixing the constant $\mathtt{C}_0$. Below we discuss the bootstrap based choice of this constant. As in Section **??** we fix some model-candidate $\varkappa^{\circ}$ and consider the family of estimates

$$\widetilde{\boldsymbol{\theta}}_{\varkappa} = \bigl(\Psi_{\varkappa} \Psi_{\varkappa}^{\top}\bigr)^{-1} \Psi_{\varkappa} \boldsymbol{Y}$$

for $\varkappa > \varkappa^{\circ}$. 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 $\breve{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \Psi^{\top} \widetilde{\boldsymbol{\theta}}$. The procedure follows the same path as in the ordered case. For each $\varkappa$ we compute and store the corresponding MLE $\widetilde{\boldsymbol{\theta}}_{\varkappa}$ and a collection of the bootstrap-based stochastic vectors $\boldsymbol{\zeta}_{\varkappa}^{\flat}$:

$$\boldsymbol{\zeta}_{\varkappa}^{\flat} = \bigl(\Psi_{\varkappa} \Psi_{\varkappa}^{\top}\bigr)^{-1} \Psi_{\varkappa} \mathcal{E}^{\flat} \breve{\boldsymbol{\varepsilon}}.$$

Further we can fix the value $\mathtt{C}_0$ using the bootstrap differences

$$\boldsymbol{\xi}_{\varkappa,\varkappa^{\circ}}^{\flat} = W\bigl(\boldsymbol{\zeta}_{\varkappa}^{\flat} - \boldsymbol{\zeta}_{\varkappa^{\circ}}^{\flat}\bigr)$$

for the weighting loss matrix W. The bootstrap critical values can be computed from these differences by the bootstrap analog of the propagation condition (9.6)

$$I\!\!P^{\flat}\biggl(\bigcup_{\varkappa \in \mathcal{M}_{\tau}(\varkappa^{\circ})} \bigl\{ \|\boldsymbol{\xi}_{\varkappa,\varkappa^{\circ}}^{\flat}\|^2 > \mathtt{C}_0\bigl(|\varkappa| - |\varkappa^{\circ}|\bigr) \bigr\}\biggr) \le \mathrm{e}^{-\mathtt{x}}.$$

Here $I\!\!P^{\flat}$ is to be understood as the empirical bootstrap measure.

To be done: Bootstrap validity

9.2 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 $\boldsymbol{\theta} \in I\!\!R^p$ we denote

$$\|\boldsymbol{\theta}\|_1 = \sum_{j=1}^{p} |\theta_j|, \qquad \|\boldsymbol{\theta}\|^2 = \sum_{j=1}^{p} \theta_j^2, \qquad \|\boldsymbol{\theta}\|_{\infty} = \max_{j \le p} |\theta_j|.$$

We consider the LASSO type procedure which is based on minimization of the empirical risk $\|\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}\|^2$ penalized by $\lambda\|\boldsymbol{\theta}\|_1$:

$$\widehat{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{\theta}} \mathcal{J}_{\lambda}(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \bigl\{ \|\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_1 \bigr\}. \tag{9.11}$$

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 (9.11) has a closed form solution only in very special situation. One of them is when the matrix $\boldsymbol{\Psi}$ is orthogonal.

Theorem 9.2.1. *Let $n \geq p$ and $\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top} = I_p$. Then $\widehat{\boldsymbol{\theta}}$ is obtained by soft-thresholding of $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\Psi}\boldsymbol{Y}$:*

$$\widehat{\theta}_j = \begin{cases} \bigl(\widetilde{\theta}_j - \lambda\bigr)_+ & \widetilde{\theta}_j \geq 0, \\ -\bigl(|\widetilde{\theta}_j| - \lambda\bigr)_+ & \widetilde{\theta}_j < 0 \end{cases} \tag{9.12}$$

Exercise 9.2.1. Prove the result (9.12).

However, compared to the complexity penalization, the problem (9.11) 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 $\boldsymbol{\theta}^*$ is sparse, that is, most of its entries vanish. By $\varkappa^*$ we denote the corresponding oracle support. We first consider the case when $\boldsymbol{\Psi}_{\varkappa^*}$ is orthogonal to the rest of $\boldsymbol{\Psi}$. Our first result shows that a proper choice of the parameter λ ensures a sparse solution: the non-zero coefficients of $\widehat{\boldsymbol{\theta}}$ are all located within $\varkappa^*$.

Theorem 9.2.2. *Let $\boldsymbol{\theta}^*$ be supported on $\varkappa_0$, $\varkappa_0^c$ be the complement of $\varkappa_0$, and*

$$\boldsymbol{\Psi}_{\varkappa_0} \boldsymbol{\Psi}_{\varkappa_0^c}^{\top} = 0. \tag{9.13}$$

If the coefficient λ fulfills

$$2\|\boldsymbol{\Psi}\boldsymbol{\varepsilon}\|_{\infty} \leq \lambda,$$

then

$$\widehat{\varkappa} \subseteq \varkappa_0 . \tag{9.14}$$

Proof. It suffices to check for each candidate $\boldsymbol{\theta}$ that the criteria in the optimization problem (9.11) only improves if we kill all its entries which do not enter in the set $\varkappa_0$. Let $\varkappa$ be the support of $\boldsymbol{\theta}$. Define $\boldsymbol{\theta}_{\varkappa_0} = \Pi_{\varkappa_0}\boldsymbol{\theta}$ as the restriction of the parameter vector $\boldsymbol{\theta}$ to $\varkappa_0$, and similarly $\boldsymbol{\theta}_{\varkappa_0^c} = \Pi_{\varkappa_0^c}\boldsymbol{\theta}$ is the projection on the complement set $\varkappa_0^c$. Obviously $\boldsymbol{\theta}_{\varkappa_0^c} = \boldsymbol{\theta} - \boldsymbol{\theta}_{\varkappa_0}$. Then the model equation $\boldsymbol{Y} = \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ implies

$$
\begin{aligned}
\|\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}\|^2 - \|\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}_{\varkappa_0}\|^2 &= \|\boldsymbol{\varepsilon} - \boldsymbol{\Psi}^{\top}(\boldsymbol{\theta}_{\varkappa_0} + \boldsymbol{\theta}_{\varkappa_0^c} - \boldsymbol{\theta}^{*})\|^2 - \|\boldsymbol{\varepsilon} - \boldsymbol{\Psi}^{\top}(\boldsymbol{\theta}_{\varkappa_0} - \boldsymbol{\theta}^{*})\|^2 \\
&= \|\boldsymbol{\Psi}^{\top}\boldsymbol{\theta}_{\varkappa_0^c}\|^2 - 2\bigl\{\boldsymbol{\varepsilon} - \boldsymbol{\Psi}^{\top}(\boldsymbol{\theta}_{\varkappa_0} - \boldsymbol{\theta}^{*})\bigr\}^{\top}\boldsymbol{\Psi}^{\top}\boldsymbol{\theta}_{\varkappa_0^c}.
\end{aligned}
$$

Exercise 9.2.2. Check that

$$
\bigl\{\boldsymbol{\Psi}^{\top}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^{*})\bigr\}^{\top}\boldsymbol{\Psi}^{\top}\boldsymbol{\theta}_{\varkappa_0^c} = 0.
$$

Hint: use that $\boldsymbol{\theta}_{\varkappa_0} - \boldsymbol{\theta}^{*}$ is supported on $\varkappa_0$, and $\boldsymbol{\theta}_{\varkappa_0^c} = \boldsymbol{\theta} - \boldsymbol{\theta}_{\varkappa_0^c}$ on $\varkappa_0^c$. Then the result follows from (9.13).

Now $\|\boldsymbol{\theta}\|_1 - \|\boldsymbol{\theta}_{\varkappa_0}\|_1 = \|\boldsymbol{\theta}_{\varkappa_0^c}\|_1$ and

$$
\begin{aligned}
&\|\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}\|^2 - \|\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}_{\varkappa_0}\|^2 + \lambda(\|\boldsymbol{\theta}\|_1 - \|\boldsymbol{\theta}_{\varkappa_0}\|_1) \\
&\quad = \|\boldsymbol{\Psi}^{\top}\boldsymbol{\theta}_{\varkappa_0^c}\|^2 + \lambda\|\boldsymbol{\theta}_{\varkappa_0^c}\|_1 - 2(\boldsymbol{\Psi}\boldsymbol{\varepsilon})^{\top}\boldsymbol{\theta}_{\varkappa_0^c}.
\end{aligned}
$$

It remains to check that the condition $2\|\boldsymbol{\Psi}\boldsymbol{\varepsilon}\|_{\infty} \le \lambda$ ensures that

$$
\lambda\|\boldsymbol{\theta}_{\varkappa_0^c}\|_1 - 2(\boldsymbol{\Psi}\boldsymbol{\varepsilon})^{\top}\boldsymbol{\theta}_{\varkappa_0^c} \ge 0.
$$

Therefore, reduction of $\boldsymbol{\theta}$ to $\boldsymbol{\theta}_{\varkappa_0}$ only improves the objective function, and the result follows.

9.2.1 Basic inequality

If we drop the orthogonality condition (9.13), the wonderful oracle result (9.14) 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 (9.11). It holds due to the model equation $\boldsymbol{Y} = \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$

$$
\begin{aligned}
\|\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}\|^2 - \|\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}^{*}\|^2 &= \|\boldsymbol{\varepsilon} - \boldsymbol{\Psi}^{\top}(\boldsymbol{\theta} - \boldsymbol{\theta}^{*})\|^2 - \|\boldsymbol{\varepsilon}\|^2 \\
&= -2(\boldsymbol{\Psi}\boldsymbol{\varepsilon})^{\top}(\boldsymbol{\theta} - \boldsymbol{\theta}^{*}) + \|\boldsymbol{\Psi}^{\top}(\boldsymbol{\theta} - \boldsymbol{\theta}^{*})\|^2.
\end{aligned}
$$

The event $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ is only possible if $\mathcal{J}_{\lambda}(\boldsymbol{\theta}) \le \mathcal{J}_{\lambda}(\boldsymbol{\theta}^{*})$. This yields

$$
\|\boldsymbol{\Psi}^{\top}(\boldsymbol{\theta} - \boldsymbol{\theta}^{*})\|^2 + \lambda\|\boldsymbol{\theta}\|_1 \le 2(\boldsymbol{\Psi}\boldsymbol{\varepsilon})^{\top}(\boldsymbol{\theta} - \boldsymbol{\theta}^{*}) + \lambda\|\boldsymbol{\theta}^{*}\|_1 \tag{9.15}
$$

Moreover, if the score $\nabla = \boldsymbol{\Psi}\boldsymbol{\varepsilon}$ fulfills $\|\nabla\|_\infty \le \mathtt{C}_0$, then

$$2(\boldsymbol{\Psi}\boldsymbol{\varepsilon})^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \le 2\mathtt{C}_0\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 \tag{9.16}$$

By the triangle inequality $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 \ge \|\boldsymbol{\theta}\|_1 - \|\boldsymbol{\theta}^*\|_1$, and (9.15) and (9.16) imply

$$\|\boldsymbol{\Psi}^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + (\lambda - 2\mathtt{C}_0)\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 \le 2\lambda\|\boldsymbol{\theta}^*\|_1$$

If $\lambda > 2\mathtt{C}_0$, 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

$$\begin{aligned} \|\boldsymbol{\Psi}^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 &\le 2\lambda\|\boldsymbol{\theta}^*\|_1, \\ \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 &\le \frac{2\lambda}{\lambda - 2\mathtt{C}_0}\|\boldsymbol{\theta}^*\|_1 . \end{aligned}$$

The last inequality can be made more precise if the true value $\boldsymbol{\theta}^*$ 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 $\boldsymbol{u} = \boldsymbol{\theta}_{\varkappa_0} - \boldsymbol{\theta}^*$. By construction, $\boldsymbol{u}$ is supported on $\varkappa_0$. It holds

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 = \|\boldsymbol{u}\|_1 + \|\boldsymbol{\theta}_{\varkappa_0^c}\|_1 .$$

Now (9.15) and (9.16) imply

$$\|\boldsymbol{\Psi}^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + \lambda\bigl(\|\boldsymbol{\theta}_{\varkappa_0}\|_1 + \|\boldsymbol{\theta}_{\varkappa_0^c}\|_1\bigr) \le 2\mathtt{C}_0\bigl(\|\boldsymbol{u}\| + \|\boldsymbol{\theta}_{\varkappa_0^c}\|_1\bigr) + \lambda\|\boldsymbol{\theta}^*\|_1$$

and therefore, the component $\boldsymbol{\theta}_{\varkappa_0^c}$ of $\boldsymbol{\theta}$ fulfills

$$\|\boldsymbol{\Psi}^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + (\lambda - 2\mathtt{C}_0)\|\boldsymbol{\theta}_{\varkappa_0^c}\|_1 \le 2\mathtt{C}_0\|\boldsymbol{u}\|_1 + \lambda\|\boldsymbol{\theta}^*\|_1 - \lambda\|\boldsymbol{\theta}_{\varkappa_0}\|_1 \le (\lambda + 2\mathtt{C}_0)\|\boldsymbol{u}\|_1 .$$

Theorem 9.2.3. *Let* $\boldsymbol{Y} = \boldsymbol{\Psi}^\top\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ *and* $\nabla = \boldsymbol{\Psi}\boldsymbol{\varepsilon}$ *fulfill* $\|\boldsymbol{\Psi}\boldsymbol{\varepsilon}\|_\infty \le \mathtt{C}_0$ *. If* $\lambda > 2\mathtt{C}_0$ *, then the estimate* $\widehat{\boldsymbol{\theta}}$ *satisfies*

$$\|\boldsymbol{\Psi}^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 + (\lambda - 2\mathtt{C}_0)\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \le 2\lambda\|\boldsymbol{\theta}^*\|_1$$

In addition, if $\boldsymbol{\theta}^*$ *is supported on* $\varkappa_0$ *, then the projections* $\widehat{\boldsymbol{\theta}}_{\varkappa_0}$ *and* $\widehat{\boldsymbol{\theta}}_{\varkappa_0^c}$ *of* $\widehat{\boldsymbol{\theta}}$ *to* $\varkappa_0$ *and* $\varkappa_0^c$ *can be related as*

$$\|\boldsymbol{\Psi}^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 + (\lambda - 2\mathtt{C}_0)\|\widehat{\boldsymbol{\theta}}_{\varkappa_0^c}\|_1 \le (\lambda + 2\mathtt{C}_0)\|\widehat{\boldsymbol{\theta}}_{\varkappa_0} - \boldsymbol{\theta}^*\|_1 .$$

To be done: Compatibility condition

To be done: Restricted isometry and oracle bound

9.2.2 Dual problem and Danzig selector

The LASSO optimization can be viewed as minimizing the fit $\|\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta}\|^{2}$ under the constraint on the ℓ_1-norm of $\boldsymbol{\theta}$. Then the objective (9.11) 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

$$\widehat{\boldsymbol{\theta}} \stackrel{\mathrm{def}}{=} \operatorname*{arginf}_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_1 \qquad \text{subject to} \quad 2\|\boldsymbol{\Psi}(\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta})\|_{\infty} \le \lambda. \tag{9.17}$$

The true value $\boldsymbol{\theta}^{*}$ is a natural candidate. Then $\|\boldsymbol{\Psi}(\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\boldsymbol{\theta})\|_{\infty} = \|\boldsymbol{\Psi}\boldsymbol{\varepsilon}\|_{\infty}$. If λ is selected properly to ensure $2\|\boldsymbol{\Psi}\boldsymbol{\varepsilon}\|_{\infty} \le \lambda$, then the constraints meet, and the objective functional is equal to $\|\boldsymbol{\theta}^{*}\|_1$. Therefore, the solution $\widehat{\boldsymbol{\theta}}$ cannot be more dense than $\boldsymbol{\theta}^{*}$ in the sense $\|\widehat{\boldsymbol{\theta}}\|_1 \le \|\boldsymbol{\theta}^{*}\|_1$.

Theorem 9.2.4. *Let $\widehat{\boldsymbol{\theta}}$ be defined by* (9.17)*. If $2\|\boldsymbol{\Psi}\boldsymbol{\varepsilon}\|_{\infty} \le \lambda$ then $\|\widehat{\boldsymbol{\theta}}\|_1 \le \|\boldsymbol{\theta}^{*}\|_1$.*

9.2.3 Data-driven choice of $\boldsymbol{\lambda}$

The necessary requirement to the choice of λ is that the score vector $\nabla = \boldsymbol{\Psi}\boldsymbol{\varepsilon}$

$$2\|\boldsymbol{\Psi}\boldsymbol{\varepsilon}\|_{\infty} \; \le \lambda.$$

This condition can be assessed by replacing the true noise by its bootstrap counterpart $\nabla^{\flat} = \boldsymbol{\Psi}\boldsymbol{\varepsilon}^{\flat}$. With a pilot $\widetilde{\boldsymbol{\theta}}$, define $\boldsymbol{\varepsilon}^{\flat} = \mathcal{W}^{\flat}(\boldsymbol{Y} - \boldsymbol{\Psi}^{\top}\widetilde{\boldsymbol{\theta}})$ and fix $\lambda^{\flat}$ by the condition

$$I\!P^{\flat}\bigl(2\|\nabla^{\flat}\|_{\infty} > \lambda^{\flat}\bigr) \; \le \mathrm{e}^{-\mathtt{x}}.$$

The original procedure has to be applied with $\lambda = \lambda^{\flat} + \beta$.

Part II

Mathematical tools

A

Univariate exponential families

This chapter reviews some basic facts about univariate exponential families (EF) distributions. This includes the Gaussian shift, Bernoulli, Poisson, exponential, volatility models.

We say that $\mathcal{P}$ is an *exponential family* if all measures $P_\theta \in \mathcal{P}$ are dominated by a σ-finite measure μ_0 on $\mathcal{Y}$ and the density functions $p(y, \theta) = dP_\theta / d\mu_0(y)$ are of the form

$$p(y, \theta) \overset{\text{def}}{=} \frac{dP_\theta}{d\mu_0}(y) = p(y) e^{yC(\theta) - B(\theta)}.$$

Here $C(\theta)$ and $B(\theta)$ are some given nondecreasing functions on Θ and $p(y)$ is a non-negative function on $\mathcal{Y}$.

Usually one assumes some regularity conditions on the family $\mathcal{P}$. One possibility was already given when we discussed the Cramér–Rao inequality; see Definition **??**. Below we assume that that condition is always fulfilled. It basically means that we can differentiate w.r.t. θ under the integral sign.

For an EF, the log-likelihood admits an especially simple representation, nearly linear in y:

$$\ell(y, \theta) \overset{\text{def}}{=} \log p(y, \theta) = yC(\theta) - B(\theta) + \log p(y)$$

so that the log-likelihood ratio for $\theta, \theta' \in \Theta$ reads as

$$\ell(y, \theta, \theta') \overset{\text{def}}{=} \ell(y, \theta) - \ell(y, \theta') = y\bigl[C(\theta) - C(\theta')\bigr] - \bigl[B(\theta) - B(\theta')\bigr].$$

A.1 Natural parametrization

Let $\mathcal{P} = \bigl(P_\theta\bigr)$ be an EF. By Y we denote one observation from the distribution $P_\theta \in \mathcal{P}$. In addition to the regularity conditions, one often assumes the *natural* parametrization for the family $\mathcal{P}$ which means the relation $E_\theta Y = \theta$. Note that this relation is fulfilled for all the examples of EF's that we considered so far in the previous section. It is obvious that the natural parametrization is only possible if the following identifiability condition is fulfilled: for any two different measures from the considered parametric family, the corresponding mean values are different. Otherwise the natural parametrization is always possible: just define θ as the expectation of Y. Below we use the abbreviation EFn for an exponential family with natural parametrization.

Some properties of an EFn

The natural parametrization implies an important property for the functions $B(\theta)$ and $C(\theta)$.

Lemma A.1.1. *Let* (P_θ) *be a naturally parameterized EF. Then*

$$B'(\theta) = \theta C'(\theta).$$

Proof. Differentiating both sides of the equation $\int p(y,\theta)\mu_0(dy) = 1$ w.r.t. θ yields

$$\begin{aligned} 0 &= \int \bigl\{yC'(\theta) - B'(\theta)\bigr\} p(y,\theta)\mu_0(dy) \\ &= \int \bigl\{yC'(\theta) - B'(\theta)\bigr\} P_\theta(dy) \\ &= \theta C'(\theta) - B'(\theta) \end{aligned}$$

and the result follows.

The next lemma computes the important characteristics of a natural EF such as the Kullback-Leibler divergence $\mathcal{K}(\theta,\theta') = E_\theta \log\bigl(p(Y,\theta)/p(Y,\theta')\bigr)$, the Fisher information $\mathbb{F}(\theta) \overset{\mathrm{def}}{=} E_\theta|\ell'(Y,\theta)|^2$, and the rate function $\mathfrak{m}(\mu,\theta,\theta') = -\log E_\theta \exp\bigl\{\mu\ell(Y,\theta,\theta')\bigr\}$.

Lemma A.1.2. *Let* (P_θ) *be an EFn. Then with* $\theta,\theta' \in \Theta$ *fixed, it holds for*

- *the Kullback-Leibler divergence* $\mathcal{K}(\theta,\theta') = E_\theta \log\bigl(p(Y,\theta)/p(Y,\theta')\bigr)$:

$$\begin{aligned} \mathcal{K}(\theta,\theta') &= \int \log \frac{p(y,\theta)}{p(y,\theta')} P_\theta(dy) \\ &= \bigl\{C(\theta) - C(\theta')\bigr\} \int y P_\theta(dy) - \bigl\{B(\theta) - B(\theta')\bigr\} \\ &= \theta\bigl\{C(\theta) - C(\theta')\bigr\} - \bigl\{B(\theta) - B(\theta')\bigr\}; \end{aligned} \tag{A.1}$$

- *the Fisher information* $\mathbb{F}(\theta) \overset{\mathrm{def}}{=} E_\theta|\ell'(Y,\theta)|^2$:

$$\mathbb{F}(\theta) = C'(\theta);$$

- *the rate function* $\mathfrak{m}(\mu,\theta,\theta') = -\log E_\theta \exp\bigl\{\mu\ell(Y,\theta,\theta')\bigr\}$:

$$\mathfrak{m}(\mu,\theta,\theta') = \mathcal{K}\bigl(\theta, \theta + \mu(\theta' - \theta)\bigr);$$

- *the variance* $\operatorname{Var}_\theta(Y)$:

$$\operatorname{Var}_\theta(Y) = 1/\mathbb{F}(\theta) = 1/C'(\theta). \tag{A.2}$$

Proof. Differentiating the equality

$$0 \equiv \int (y-\theta) P_\theta(dy) = \int (y-\theta) e^{L(y,\theta)} \mu_0(dy)$$

w.r.t. θ implies in view of Lemma A.1.1

$$1 \equiv I\!E_\theta\bigl[(Y-\theta)\bigl\{C'(\theta)Y - B'(\theta)\bigr\}\bigr] = C'(\theta) I\!E_\theta (Y-\theta)^2.$$

This yields $\operatorname{Var}_\theta(Y) = 1/C'(\theta)$. This leads to the following representation of the Fisher information:

$$\mathbb{F}(\theta) = \operatorname{Var}_\theta\bigl[\ell'(Y,\theta)\bigr] = \operatorname{Var}_\theta[C'(\theta)Y - B'(\theta)] = \bigl[C'(\theta)\bigr]^2 \operatorname{Var}_\theta(Y) = C'(\theta).$$

Exercise A.1.1. Check the equations for the Kullback-Leibler divergence and Fisher information from Lemma A.1.2.

MLE and maximum likelihood for an EFn

Now we discuss the maximum likelihood estimation for a sample from an EFn. The log-likelihood can be represented in the form

$$\begin{aligned} L(\theta) &= \sum_{i=1}^n \log p(Y_i,\theta) = C(\theta) \sum_{i=1}^n Y_i - B(\theta) \sum_{i=1}^n 1 + \sum_{i=1}^n \log p(Y_i) \\ &= SC(\theta) - nB(\theta) + R, \end{aligned} \tag{A.3}$$

where

$$S = \sum_{i=1}^n Y_i, \qquad R = \sum_{i=1}^n \log p(Y_i).$$

The remainder term R is unimportant because it does not depend on θ and thus it does not enter in the likelihood ratio. The maximum likelihood estimate $\widetilde{\theta}$ is defined by maximizing $L(\theta)$ w.r.t. θ, that is,

$$\widetilde{\theta} = \operatorname*{argmax}_{\theta\in\Theta} L(\theta) = \operatorname*{argmax}_{\theta\in\Theta} \bigl\{SC(\theta) - nB(\theta)\bigr\}.$$

In the case of an EF with the natural parametrization, this optimization problem admits a closed form solution given by the next theorem.

Theorem A.1.1. *Let (P_θ) be an EFn. Then the MLE $\widetilde{\theta}$ fulfills*

$$\widetilde{\theta} = S/n = n^{-1} \sum_{i=1}^{n} Y_i \, .$$

It holds

$$I\!\!E_\theta \widetilde{\theta} = \theta, \qquad \operatorname{Var}_\theta(\widetilde{\theta}) = \bigl[n\mathbb{F}(\theta)\bigr]^{-1} = \bigl[nC'(\theta)\bigr]^{-1}$$

so that $\widetilde{\theta}$ is R-efficient. Moreover, the fitted log-likelihood $L(\widetilde{\theta}, \theta) \stackrel{\text{def}}{=} L(\widetilde{\theta}) - L(\theta)$ satisfies for any $\theta \in \Theta$:

$$L(\widetilde{\theta}, \theta) = n\mathcal{K}(\widetilde{\theta}, \theta). \tag{A.4}$$

Proof. Maximization of $L(\theta)$ w.r.t. θ leads to the estimating equation $nB'(\theta) - SC'(\theta) = 0$. This and the identity $B'(\theta) = \theta C'(\theta)$ yield the MLE

$$\widetilde{\theta} = S/n.$$

The variance $\operatorname{Var}_\theta(\widetilde{\theta})$ is computed using (A.2) from Lemma A.1.2. The formula (A.1) for the Kullback-Leibler divergence and (A.3) yield the representation (A.4) for the fitted log-likelihood $L(\widetilde{\theta}, \theta)$ for any $\theta \in \Theta$.

One can see that the estimate $\widetilde{\theta}$ is the mean of the Y_i 's. As for the Gaussian shift model, this estimate can be motivated by the fact that the expectation of every observation Y_i under P_θ is just θ and by the law of large numbers the empirical mean converges to its expectation as the sample size n grows.

A.2 Canonical parametrization

Another useful representation of an EF is given by the so-called *canonical parametrization*. We say that υ is the *canonical* parameter for this EF if the density of each measure P_υ w.r.t. the dominating measure μ_0 is of the form:

$$p(y, \upsilon) \stackrel{\text{def}}{=} \frac{dP_\upsilon}{d\mu_0}(y) = p(y) \exp\bigl\{y\upsilon - d(\upsilon)\bigr\}.$$

Here $d(\upsilon)$ is a given *convex* function on Θ and $p(y)$ is a nonnegative function on $\mathcal{Y}$. The abbreviation EFc will indicate an EF with the canonical parametrization.

Some properties of an EFc

The next relation is an obvious corollary of the definition:

Lemma A.2.1. *An EFn (P_θ) always permits a unique canonical representation. The canonical parameter υ is related to the natural parameter θ by $\upsilon = C(\theta)$, $d(\upsilon) = B(\theta)$ and $\theta = d'(\upsilon)$.*

Proof. The first two relations follow from the definition. They imply $B'(\theta) = d'(\upsilon) \cdot d\upsilon/d\theta = d'(\upsilon) \cdot C'(\theta)$ and the last statement follows from $B'(\theta) = \theta C'(\theta)$.

The log-likelihood ratio $\ell(y, \upsilon, \upsilon_1)$ for an EFc reads as

$$\ell(Y, \upsilon, \upsilon_1) = Y(\upsilon - \upsilon_1) - d(\upsilon) + d(\upsilon_1).$$

The next lemma collects some useful facts about an EFc.

Lemma A.2.2. *Let $\mathcal{P} = \bigl(P_\upsilon, \upsilon \in \mathcal{U}\bigr)$ be an EFc and let the function $d(\cdot)$ be two times continuously differentiable. Then it holds for any $\upsilon, \upsilon_1 \in \mathcal{U}$:*

(i). The mean $E_\upsilon Y$ and the variance $\operatorname{Var}_\upsilon(Y)$ fulfill

$$E_\upsilon Y = d'(\upsilon), \qquad \operatorname{Var}_\upsilon(Y) = E_\upsilon (Y - E_\upsilon Y)^2 = d''(\upsilon).$$

(ii). The Fisher information $\mathbb{F}(\upsilon) \overset{\text{def}}{=} E_\upsilon |\ell'(Y, \upsilon)|^2$ satisfies

$$\mathbb{F}(\upsilon) = d''(\upsilon).$$

(iii). The Kullback-Leibler divergence $\mathcal{K}^c(\upsilon, \upsilon_1) = E_\upsilon \ell(Y, \upsilon, \upsilon_1)$ satisfies

$$\begin{aligned}\mathcal{K}^c(\upsilon, \upsilon_1) &= \int \log \frac{p(y, \upsilon)}{p(y, \upsilon_1)} P_\upsilon(dy) \\ &= d'(\upsilon)\bigl(\upsilon - \upsilon_1\bigr) - \bigl\{d(\upsilon) - d(\upsilon_1)\bigr\} \\ &= d''(\breve{\upsilon})\,(\upsilon_1 - \upsilon)^2/2,\end{aligned}$$

where $\breve{\upsilon}$ is a point between υ and υ_1. Moreover, for $\upsilon \le \upsilon_1 \in \mathcal{U}$

$$\mathcal{K}^c(\upsilon, \upsilon_1) = \int_\upsilon^{\upsilon_1} (\upsilon_1 - u) d''(u) du.$$

(iv). The rate function $\mathfrak{m}(\mu, \upsilon_1, \upsilon) \overset{\text{def}}{=} -\log I\!\!E_\upsilon \exp\bigl\{\mu \ell(Y, \upsilon_1, \upsilon)\bigr\}$ fulfills

$$\mathfrak{m}(\mu, \upsilon_1, \upsilon) = \mu \mathcal{K}^c\bigl(\upsilon, \upsilon_1\bigr) - \mathcal{K}^c\bigl(\upsilon, \upsilon + \mu(\upsilon_1 - \upsilon)\bigr)$$

Proof. Differentiating the equation $\int p(y,\upsilon)\mu_0(dy) = 1$ w.r.t. υ yields

$$\int \bigl\{y - d'(\upsilon)\bigr\} p(y,\upsilon)\mu_0(dy) = 0,$$

that is, $E_\upsilon Y = d'(\upsilon)$. The expression for the variance can be proved by one more differentiating of this equation. Similarly one can check (ii). The item (iii) can be checked by simple algebra and (iv) follows from (i).

Further, for any $\upsilon, \upsilon_1 \in \mathcal{U}$, it holds

$$\ell(Y,\upsilon_1,\upsilon) - E_\upsilon \ell(Y,\upsilon_1,\upsilon) = (\upsilon_1 - \upsilon)\bigl\{Y - d'(\upsilon)\bigr\}$$

and with $u = \mu(\upsilon_1 - \upsilon)$

$$\begin{aligned}
&\log E_\upsilon \exp\bigl\{u\bigl(Y - d'(\upsilon)\bigr)\bigr\} \\
&\quad = -ud'(\upsilon) + d(\upsilon + u) - d(\upsilon) + \log E_\upsilon \exp\bigl\{uY - d(\upsilon+u) + d(\upsilon)\bigr\} \\
&\quad = d(\upsilon + u) - d(\upsilon) - ud'(\upsilon) = \mathcal{K}^c(\upsilon, \upsilon + u),
\end{aligned}$$

because

$$E_\upsilon \exp\bigl\{uY - d(\upsilon+u) + d(\upsilon)\bigr\} = E_\upsilon \frac{dP_{\upsilon+u}}{dP_\upsilon} = 1$$

and (iv) follows by (iii).

Table A.1 presents the canonical parameter and the Fisher information for the examples of exponential families from Section **??**.

Table A.1. $\upsilon(\theta)$, $d(\upsilon)$, $\mathbb{F}(\upsilon) = d''(\upsilon)$ and $\theta = \theta(\upsilon)$ for the examples from Section **??**.

Model	υ	$d(\upsilon)$	$I(\upsilon)$	$\theta(\upsilon)$
Gaussian regression	θ/σ^2	$\upsilon^2\sigma^2/2$	σ^2	$\sigma^2\upsilon$
Bernoulli model	$\log\bigl(\theta/(1-\theta)\bigr)$	$\log(1+e^\upsilon)$	$e^\upsilon/(1+e^\upsilon)^2$	$e^\upsilon/(1+e^\upsilon)$
Poisson model	$\log\theta$	e^υ	e^υ	e^υ
Exponential model	$1/\theta$	$-\log\upsilon$	$1/\upsilon^2$	$1/\upsilon$
Volatility model	$-1/(2\theta)$	$-\frac{1}{2}\log(-2\upsilon)$	$1/(2\upsilon^2)$	$-1/(2\upsilon)$

Exercise A.2.1. Check (iii) and (iv) in Lemma A.2.2.

Exercise A.2.2. Check the entries of Table A.1.

Exercise A.2.3. Check that $\mathcal{K}^c(\upsilon,\upsilon') = \mathcal{K}\bigl(\theta(\upsilon), \theta(\upsilon')\bigr)$

Exercise A.2.4. Plot $\mathcal{K}^c(\upsilon^*, \upsilon)$ as a function of υ for the families from Table A.1.

Maximum likelihood estimation for an EFc

The structure of the log-likelihood in the case of the canonical parametrization is particularly simple:

$$\begin{aligned} L(\upsilon) &= \sum_{i=1}^{n} \log p(Y_i, \upsilon) = \upsilon \sum_{i=1}^{n} Y_i - d(\upsilon) \sum_{i=1}^{n} 1 + \sum_{i=1}^{n} \log p(Y_i) \\ &= S\upsilon - nd(\upsilon) + R \end{aligned}$$

where

$$S = \sum_{i=1}^{n} Y_i, \qquad R = \sum_{i=1}^{n} \log p(Y_i).$$

Again, as in the case of an EFn, we can ignore the remainder term R. The estimating equation $dL(\upsilon)/d\upsilon = 0$ for the maximum likelihood estimate $\widetilde{\upsilon}$ reads as

$$d'(\upsilon) = S/n.$$

This and the relation $\theta = d'(\upsilon)$ lead to the following result.

Theorem A.2.1. *The maximum likelihood estimates $\widetilde{\theta}$ and $\widetilde{\upsilon}$ for the natural and canonical parametrization are related by the equations*

$$\widetilde{\theta} = d'(\widetilde{\upsilon}) \qquad \widetilde{\upsilon} = C(\widetilde{\theta}).$$

The next result describes the structure of the fitted log-likelihood and basically repeats the result of Theorem A.1.1.

Theorem A.2.2. *Let (P_υ) be an EF with canonical parametrization. Then for any $\upsilon \in \mathcal{U}$ the fitted log-likelihood $L(\widetilde{\upsilon}, \upsilon) \overset{\text{def}}{=} \max_{\upsilon'} L(\upsilon', \upsilon)$ satisfies*

$$L(\widetilde{\upsilon}, \upsilon) = n\mathcal{K}^c(\widetilde{\upsilon}, \upsilon).$$

Exercise A.2.5. Check the statement of Theorem A.2.2.

A.3 Deviation probabilities for the maximum likelihood

Let $Y_1, \ldots, Y_n$ be i.i.d. observations from an EF $\mathcal{P}$. This section presents a probability bound for the fitted likelihood. To be more specific we assume that $\mathcal{P}$ is canonically parameterized, $\mathcal{P} = (P_\upsilon)$. However, the bound applies to the natural and any other parametrization because the value of maximum of the likelihood process $L(\theta)$ does not depend on the choice of parametrization. The log-likelihood ratio $L(\upsilon', \upsilon)$ is given by the expression (A.3) and its maximum over υ' leads to the fitted log-likelihood $L(\widetilde{\upsilon}, \upsilon) = n\mathcal{K}^c(\widetilde{\upsilon}, \upsilon)$.

Our first result concerns a *deviation bound* for $L(\widetilde{\upsilon}, \upsilon)$. It utilizes the representation for the fitted log-likelihood given by Theorem A.1.1. As usual, we assume that the family $\mathcal{P}$ is regular. In addition, we require the following condition.

(Pc) $\mathcal{P} = (P_\upsilon, \upsilon \in \mathcal{U} \subseteq I\!\!R)$ is a regular EF. The parameter set $\mathcal{U}$ is convex. The function $d(\upsilon)$ is two times continuously differentiable and the Fisher information $\mathbb{F}(\upsilon) = d''(\upsilon)$ satisfies $\mathbb{F}(\upsilon) > 0$ for all υ.

The condition (Pc) implies that for any compact set $\mathcal{U}_0$ there is a constant $\mathfrak{a} = \mathfrak{a}(\mathcal{U}_0) > 0$ such that

$$|\mathbb{F}(\upsilon_1)/\mathbb{F}(\upsilon_2)|^{1/2} \le \mathfrak{a}, \qquad \upsilon_1, \upsilon_2 \in \mathcal{U}_0 .$$

Theorem A.3.1. *Let Y_i be i.i.d. from a distribution P_{υ^*} which belongs to an EFc satisfying (Pc). For any $\mathfrak{z} > 0$*

$$I\!\!P_{\upsilon^*}\bigl(L(\widetilde{\upsilon}, \upsilon^*) > \mathfrak{z}\bigr) = I\!\!P_{\upsilon^*}\bigl(n\mathcal{K}^c(\widetilde{\upsilon}, \upsilon^*) > \mathfrak{z}\bigr) \le 2e^{-\mathfrak{z}}.$$

Proof. The proof is based on two properties of the log-likelihood. The first one is that the expectation of the likelihood ratio is just one: $I\!\!E_{\upsilon^*} \exp L(\upsilon, \upsilon^*) = 1$. This and the exponential Markov inequality imply for $\mathfrak{z} \ge 0$

$$I\!\!P_{\upsilon^*}\bigl(L(\upsilon, \upsilon^*) \ge \mathfrak{z}\bigr) \le \mathrm{e}^{-\mathfrak{z}}. \tag{A.5}$$

The second property is specific to the considered univariate EF and is based on geometric properties of the log-likelihood function: linearity in the observations Y_i and convexity in the parameter υ. We formulate this important fact in a separate statement.

Lemma A.3.1. *Let the EFc $\mathcal{P}$ fulfill (Pc). For given $\mathfrak{z}$ and any $\upsilon_0 \in \mathcal{U}$, there exist two values $\upsilon^+ > \upsilon_0$ and $\upsilon^- < \upsilon_0$ satisfying $\mathcal{K}^c(\upsilon^\pm, \upsilon_0) = \mathfrak{z}/n$ such that*

$$\{L(\widetilde{\upsilon}, \upsilon_0) > \mathfrak{z}\} \subseteq \{L(\upsilon^+, \upsilon_0) > \mathfrak{z}\} \cup \{L(\upsilon^-, \upsilon_0) > \mathfrak{z}\}.$$

Proof. It holds

$$\begin{aligned}\{L(\widetilde{\upsilon}, \upsilon_0) > \mathfrak{z}\} &= \bigl\{\sup_{\upsilon}\bigl[S\bigl(\upsilon - \upsilon_0\bigr) - n\bigl\{d(\upsilon) - d(\upsilon_0)\bigr\}\bigr] > \mathfrak{z}\bigr\} \\ &\subseteq \biggl\{S > \inf_{\upsilon > \upsilon_0} \frac{\mathfrak{z} + n\bigl\{d(\upsilon) - d(\upsilon_0)\bigr\}}{\upsilon - \upsilon_0}\biggr\} \cup \biggl\{-S > \inf_{\upsilon < \upsilon_0} \frac{\mathfrak{z} + n\bigl\{d(\upsilon) - d(\upsilon_0)\bigr\}}{\upsilon_0 - \upsilon}\biggr\}.\end{aligned}$$

Define for every $u > 0$

$$f(u) = \frac{\mathfrak{z} + n\bigl\{d(\upsilon_0 + u) - d(\upsilon_0)\bigr\}}{u}.$$

This function attains its minimum at a point u satisfying the equation

$$\mathfrak{z}/n + d(\upsilon_0 + u) - d(\upsilon_0) - d'(\upsilon_0 + u)u = 0$$

or, equivalently,

$$\mathcal{K}(\upsilon_0 + u, \upsilon_0) = \mathfrak{z}/n.$$

The condition (Pc) provides that there is only one solution $u \ge 0$ of this equation.

Exercise A.3.1. Check that the equation $\mathcal{K}(\upsilon_0 + u, \upsilon_0) = \mathfrak{z}/n$ has only one positive solution for any $\mathfrak{z} > 0$.
Hint: use that $\mathcal{K}(\upsilon_0 + u, \upsilon_0)$ is a convex function of u with minimum at $u = 0$.

Now, it holds with $\upsilon^{+} = \upsilon_0 + u$

$$\begin{aligned}\Bigl\{S > \inf_{\upsilon > \upsilon_0} \frac{\mathfrak{z} + n\bigl[d(\upsilon) - d(\upsilon_0)\bigr]}{\upsilon - \upsilon_0}\Bigr\} &= \Bigl\{S > \frac{\mathfrak{z} + n\bigl[d(\upsilon^{+}) - d(\upsilon_0)\bigr]}{\upsilon^{+} - \upsilon_0}\Bigr\} \\ &\subseteq \{L(\upsilon^{+}, \upsilon_0) > \mathfrak{z}\}.\end{aligned}$$

Similarly

$$\begin{aligned}\Bigl\{-S > \inf_{\upsilon < \upsilon_0} \frac{\mathfrak{z} + n\bigl\{d(\upsilon) - d(\upsilon_0)\bigr\}}{\upsilon_0 - \upsilon}\Bigr\} &= \Bigl\{-S > \frac{\mathfrak{z} + n\bigl[d(\upsilon^{-}) - d(\upsilon_0)\bigr]}{\upsilon_0 - \upsilon^{-}}\Bigr\} \\ &\subseteq \{L(\upsilon^{-}, \upsilon_0) > \mathfrak{z}\}.\end{aligned}$$

for some $\upsilon^{-} < \upsilon_0$.

The assertion of the theorem is now easy to obtain. Indeed,

$$I\!\!P_{\upsilon^*}\bigl(L(\widetilde{\upsilon}, \upsilon^*) \ge \mathfrak{z}\bigr) \le I\!\!P_{\upsilon^*}\bigl(L(\upsilon^{+}, \upsilon^*) \ge \mathfrak{z}\bigr) + I\!\!P_{\upsilon^*}\bigl(L(\upsilon^{-}, \upsilon^*) \ge \mathfrak{z}\bigr) \le 2\mathrm{e}^{-\mathfrak{z}}$$

yielding the result.

Exercise A.3.2. Let (P_υ) be a Gaussian shift experiment, that is, $P_\upsilon = \mathcal{N}(\upsilon, 1)$.

- Check that $L(\widetilde{\upsilon}, \upsilon) = n|\widetilde{\upsilon} - \upsilon|^2/2$;
- Given $\mathfrak{z} \ge 0$, find the points υ^{+} and υ^{-} such that

$$\{L(\widetilde{\upsilon}, \upsilon^*) > \mathfrak{z}\} \subseteq \{L(\upsilon^{+}, \upsilon^*) > \mathfrak{z}\} \cup \{L(\upsilon^{-}, \upsilon^*) > \mathfrak{z}\}.$$

- Plot the mentioned sets $\{\upsilon : L(\widetilde{\upsilon}, \upsilon) > \mathfrak{z}\}$, $\{\upsilon : L(\upsilon^{+}, \upsilon) > \mathfrak{z}\}$, and $\{\upsilon : L(\upsilon^{-}, \upsilon) > \mathfrak{z}\}$ as functions of υ for a fixed $S = \sum Y_i$.

Remark A.3.1. Note that the mentioned result only utilizes the geometric structure of the univariate EFc. The most important feature of the log-likelihood ratio $L(\upsilon, \upsilon^*) = S(\upsilon - \upsilon^*) - d(\upsilon) + d(\upsilon^*)$ is its linearity w.r.t. the stochastic term S. This allows us to replace the maximum over the whole set $\mathcal{U}$ by the maximum over the set consisting of two points $\upsilon^{\pm}$. Note that the proof does not rely on the distribution of the observations Y_i. In particular, Lemma A.3.1 continues to hold even within the quasi likelihood approach when $L(\upsilon)$ is not the true log-likelihood. However, the bound (A.5) relies on the nature of $L(\upsilon, \upsilon^*)$. Namely, it utilizes that $I\!\!E \exp\bigl\{L(\upsilon^{\pm}, \upsilon^*)\bigr\} = 1$, which is true under $I\!\!P = I\!\!P_{\upsilon^*}$ nut generally false in the quasi likelihood setup. Nevertheless, the exponential bound can be extended to the quasi likelihood approach under the condition of bounded exponential moments for $L(\upsilon, \upsilon^*)$: for some $\mu > 0$, it should hold $I\!\!E \exp\bigl\{\mu L(\upsilon, \upsilon^*)\bigr\} = C(\mu) < \infty$.

Theorem A.3.1 yields a simple construction of a confidence interval for the parameter υ^* and the concentration property of the MLE $\widetilde{\upsilon}$.

Theorem A.3.2. *Let Y_i be i.i.d. from $P_{\upsilon^*} \in \mathcal{P}$ with $\mathcal{P}$ satisfying (Pc).*

1. *If $\mathfrak{z}_\alpha$ satisfies $e^{-\mathfrak{z}_\alpha} \le \alpha/2$, then*

$$\mathcal{E}(\mathfrak{z}_\alpha) = \bigl\{\upsilon : n\mathcal{K}^c\bigl(\widetilde{\upsilon}, \upsilon\bigr) \le \mathfrak{z}_\alpha\bigr\}$$

is a α-confidence set for the parameter υ^*.

2. *Define for any $\mathfrak{z} > 0$ the set $\mathcal{A}(\mathfrak{z}, \upsilon^*) = \{\upsilon : \mathcal{K}^c(\upsilon, \upsilon^*) \le \mathfrak{z}/n\}$. Then*

$$I\!\!P_{\upsilon^*}\bigl(\widetilde{\upsilon} \notin \mathcal{A}(\mathfrak{z}, \upsilon^*)\bigr) \le 2\mathrm{e}^{-\mathfrak{z}}.$$

The second assertion of the theorem claims that the estimate $\widetilde{\upsilon}$ belongs with a high probability to the vicinity $\mathcal{A}(\mathfrak{z}, \upsilon^*)$ of the central point υ^* defined by the Kullback-Leibler divergence. Due to Lemma A.2.2, (iii) $\mathcal{K}^c(\upsilon, \upsilon^*) \approx \mathbb{F}(\upsilon^*)\,(\upsilon - \upsilon^*)^2/2$, where $\mathbb{F}(\upsilon^*)$ is the Fisher information at υ^*. This vicinity is an interval around υ^* of length of order $n^{-1/2}$. In other words, this result implies the root-n consistency of $\widetilde{\upsilon}$.

The deviation bound for the fitted log-likelihood from Theorem A.3.1 can be viewed as a bound for the normalized loss of the estimate $\widetilde{\upsilon}$. Indeed, define the loss function $\wp(\upsilon', \upsilon) = \mathcal{K}^{1/2}(\upsilon', \upsilon)$. Then Theorem A.3.1 yields that the loss is with high probability bounded by $\sqrt{\mathfrak{z}/n}$ provided that $\mathfrak{z}$ is sufficiently large. Similarly one can establish the bound for the risk.

Theorem A.3.3. *Let Y_i be i.i.d. from the distribution P_{υ^*} which belongs to a canonically parameterized EF satisfying (Pc). The following properties hold:*

(i). For any $r > 0$ there is a constant $\mathfrak{r}_r$ such that

$$I\!\!E_{\upsilon^*} L^r(\widetilde{\upsilon}, \upsilon^*) = n^r I\!\!E_{\upsilon^*} \mathcal{K}^r(\widetilde{\upsilon}, \upsilon^*) \le \mathfrak{r}_r\,.$$

(ii). For every $\lambda < 1$

$$I\!\!E_{\upsilon^*} \exp\bigl\{\lambda L(\widetilde{\upsilon}, \upsilon^*)\bigr\} = I\!\!E_{\upsilon^*} \exp\bigl\{\lambda n \mathcal{K}(\widetilde{\upsilon}, \upsilon^*)\bigr\} \le (1+\lambda)/(1-\lambda).$$

Proof. By Theorem A.3.1

$$\begin{aligned} I\!\!E_{\upsilon^*} L^r(\widetilde{\upsilon}, \upsilon^*) &= -\int_{\mathfrak{z} \ge 0} \mathfrak{z}^r \, dI\!\!P_{\upsilon^*}\bigl\{L(\widetilde{\upsilon}, \upsilon^*) > \mathfrak{z}\bigr\} \\ &= r \int_{\mathfrak{z} \ge 0} \mathfrak{z}^{r-1} I\!\!P_{\upsilon^*}\bigl\{L(\widetilde{\upsilon}, \upsilon^*) > \mathfrak{z}\bigr\} d\mathfrak{z} \\ &\le r \int_{\mathfrak{z} \ge 0} 2\mathfrak{z}^{r-1} e^{-\mathfrak{z}} d\mathfrak{z} \end{aligned}$$

and the first assertion is fulfilled with $\mathfrak{r}_r = 2r \int_{\mathfrak{z} \ge 0} \mathfrak{z}^{r-1} e^{-\mathfrak{z}} d\mathfrak{z}$. The assertion (ii) is proved similarly.

Deviation bound for other parameterizations

The results for the maximum likelihood and their corollaries have been stated for an EFc. An immediate question that arises in this respect is whether the use of the canonical parametrization is essential. The answer is "no": a similar result can be stated for any EF whatever the parametrization is used. This fact is based on the simple observation that the maximum likelihood is the value of the maximum of the likelihood process; this value does not depend on the parametrization.

Lemma A.3.2. *Let* (P_θ) *be an EF. Then for any* θ

$$L(\widetilde{\theta}, \theta) = n \mathcal{K}(P_{\widetilde{\theta}}, P_\theta). \tag{A.6}$$

Exercise A.3.3. Check the result of Lemma A.3.2.
Hint: use that both sides of (A.6) depend only on measures $P_{\widetilde{\theta}}, P_\theta$ and not on the parametrization.

Below we write as before $\mathcal{K}(\widetilde{\theta}, \theta)$ instead of $\mathcal{K}(P_{\widetilde{\theta}}, P_\theta)$. The property (A.6) and the exponential bound of Theorem A.3.1 imply the bound for a general EF:

Theorem A.3.4. *Let* (P_θ) *be a univariate EF. Then for any* $\mathfrak{z} > 0$

$$I\!\!P_{\theta^*}\bigl(L(\widetilde{\theta}, \theta^*) > \mathfrak{z}\bigr) = I\!\!P_{\theta^*}\bigl(n\mathcal{K}(\widetilde{\theta}, \theta^*) > \mathfrak{z}\bigr) \le 2e^{-\mathfrak{z}}.$$

This result allows us to build confidence sets for the parameter θ^* and concentration sets for the MLE $\widetilde{\theta}$ in terms of the Kullback-Leibler divergence:

$$\begin{aligned} \mathcal{A}(\mathfrak{z}, \theta^*) &= \{\theta : \mathcal{K}(\theta, \theta^*) \le \mathfrak{z}/n\}, \\ \mathcal{E}(\mathfrak{z}) &= \{\theta : \mathcal{K}(\widetilde{\theta}, \theta) \le \mathfrak{z}/n\}. \end{aligned}$$

Corollary A.3.1. *Let* (P_θ) *be an EF. If* $e^{-\mathfrak{z}_\alpha} = \alpha/2$ *then*

$$I\!\!P_{\theta^*}\bigl(\widetilde{\theta} \notin \mathcal{A}(\mathfrak{z}_\alpha, \theta^*)\bigr) \le \alpha,$$

and

$$I\!\!P_{\theta^*}\bigl(\mathcal{E}(\mathfrak{z}_\alpha) \not\ni \theta\bigr) \le \alpha.$$

Moreover, for any $r > 0$

$$I\!\!E_{\theta^*} L^r(\widetilde{\theta}, \theta^*) = n^r I\!\!E_{\theta^*} \mathcal{K}^r(\widetilde{\theta}, \theta^*) \le \mathfrak{r}_r \, .$$

Asymptotic against likelihood-based approach

The asymptotic approach recommends to apply symmetric confidence and concentration sets with width of order $[n\mathbb{F}(\theta^*)]^{-1/2}$:

$$\begin{aligned} \mathcal{A}_n(\mathfrak{z}, \theta^*) &= \{\theta : \mathbb{F}(\theta^*)\,(\theta - \theta^*)^2 \le 2\mathfrak{z}/n\}, \\ \mathcal{E}_n(\mathfrak{z}) &= \{\theta : \mathbb{F}(\theta^*)\,(\theta - \widetilde{\theta})^2 \le 2\mathfrak{z}/n\}, \\ \mathcal{E}'_n(\mathfrak{z}) &= \{\theta : \;\; I(\widetilde{\theta})\,(\theta - \widetilde{\theta})^2 \le 2\mathfrak{z}/n\}. \end{aligned}$$

Then asymptotically, i.e. for large n, these sets do approximately the same job as the non-asymptotic sets $\mathcal{A}(\mathfrak{z}, \theta^*)$ and $\mathcal{E}(\mathfrak{z})$. However, the difference for finite samples can be quite significant. In particular, for some cases, e.g. the Bernoulli of Poisson families, the sets $\mathcal{A}_n(\mathfrak{z}, \theta^*)$ and $\mathcal{E}'_n(\mathfrak{z})$ may extend beyond the parameter set Θ.

B

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 $\boldsymbol{\gamma}$ means a standard normal vector in $I\!\!R^{p}$.

B.1 Deviation bounds for a Gaussian vector

The next result describes the tails of the norm of a standard Gaussian vector.

Lemma B.1.1. *Let* $\mu \in (0,1)$. *Then for any vector* $\boldsymbol{\lambda} \in I\!\!R^{p}$ *with* $\|\boldsymbol{\lambda}\|^{2} \le p$ *and any* $\mathtt{r} > 0$

$$\log I\!\!E\bigl\{\exp(\boldsymbol{\lambda}^{\top}\boldsymbol{\gamma})\, 1\!\!\mathrm{I}\bigl(\|\boldsymbol{\gamma}\| > \mathtt{r}\bigr)\bigr\} \le -\frac{1-\mu}{2}\mathtt{r}^{2} + \frac{1}{2\mu}\|\boldsymbol{\lambda}\|^{2} + \frac{p}{2}\log(\mu^{-1}). \tag{B.1}$$

Moreover, if $\mathtt{r}^{2} \ge 4p + 4\mathtt{x}$, *then*

$$I\!\!E\Bigl\{\exp(\boldsymbol{\lambda}^{\top}\boldsymbol{\gamma})\, 1\!\!\mathrm{I}\bigl(\|\boldsymbol{\gamma}\| \le \mathtt{r}\bigr)\Bigr\} \ge \mathrm{e}^{\|\boldsymbol{\lambda}\|^{2}/2}\bigl(1 - \mathrm{e}^{-\mathtt{x}}\bigr). \tag{B.2}$$

Proof. We use that for $\mu < 1$

$$I\!\!E\bigl\{\exp(\boldsymbol{\lambda}^{\top}\boldsymbol{\gamma})\, 1\!\!\mathrm{I}\bigl(\|\boldsymbol{\gamma}\| > \mathtt{r}\bigr)\bigr\} \le \mathrm{e}^{-(1-\mu)\mathtt{r}^{2}/2} I\!\!E\exp\bigl\{\boldsymbol{\lambda}^{\top}\boldsymbol{\gamma} + (1-\mu)\|\boldsymbol{\gamma}\|^{2}/2\bigr\}.$$

It holds

$$\begin{aligned} I\!\!E\exp\bigl\{\boldsymbol{\lambda}^{\top}\boldsymbol{\gamma} + (1-\mu)\|\boldsymbol{\gamma}\|^{2}/2\bigr\} &= (2\pi)^{-p/2}\int \exp\bigl\{\boldsymbol{\lambda}^{\top}\boldsymbol{\gamma} - \mu\|\boldsymbol{\gamma}\|^{2}/2\bigr\} d\boldsymbol{\gamma} \\ &= \mu^{-p/2}\exp\bigl(\mu^{-1}\|\boldsymbol{\lambda}\|^{2}/2\bigr) \end{aligned}$$

and (B.1) follows.

Now we apply this result with $\mu = 1/2$. In view of $I\!\!E\exp(\boldsymbol{\lambda}^{\top}\boldsymbol{\gamma}) = \mathrm{e}^{\|\boldsymbol{\lambda}\|^{2}/2}$, $\mathtt{r}^{2} \ge 4p + 4\mathtt{x}$, and $1 + \log(2) < 2$, it follows for $\|\boldsymbol{\lambda}\|^{2} \le p$

$$\begin{aligned} & \mathrm{e}^{-\|\boldsymbol{\lambda}\|^2/2} I\!E\bigl\{\exp(\boldsymbol{\lambda}^\top \boldsymbol{\gamma}) \, \mathbb{I}\bigl(\|\boldsymbol{\gamma}\| \le \mathtt{r}\bigr)\bigr\} \\ & \qquad \ge 1 - \exp\Bigl(-\frac{\mathtt{r}^2}{4} + \frac{p + p\log(2)}{2}\Bigr) \ge 1 - \exp(-\mathtt{x}) \end{aligned}$$

which implies (B.2).

Lemma B.1.2. *For any* $\boldsymbol{u} \in I\!R^p$ *, any unit vector* $\boldsymbol{a} \in I\!R^p$ *, and any* $z > 0$ *, it holds*

$$I\!P\bigl(\|\boldsymbol{\gamma} - \boldsymbol{u}\| \ge z\bigr) \le \exp\bigl\{-z^2/4 + p/2 + \|\boldsymbol{u}\|^2/2\bigr\}, \tag{B.3}$$

$$I\!E\bigl\{|\boldsymbol{\gamma}^\top \boldsymbol{a}|^2 \, \mathbb{I}\bigl(\|\boldsymbol{\gamma} - \boldsymbol{u}\| \ge z\bigr)\bigr\} \le (2 + |\boldsymbol{u}^\top \boldsymbol{a}|^2) \exp\bigl\{-z^2/4 + p/2 + \|\boldsymbol{u}\|^2/2\bigr\}. \tag{B.4}$$

Proof. By the exponential Chebyshev inequality, for any $\lambda < 1$

$$\begin{aligned} I\!P\bigl(\|\boldsymbol{\gamma} - \boldsymbol{u}\| \ge z\bigr) & \le \exp\bigl(-\lambda z^2/2\bigr) I\!E \exp\bigl(\lambda \|\boldsymbol{\gamma} - \boldsymbol{u}\|^2/2\bigr) \\ & = \exp\Bigl\{-\frac{\lambda z^2}{2} - \frac{p}{2}\log(1-\lambda) + \frac{\lambda}{2(1-\lambda)}\|\boldsymbol{u}\|^2\Bigr\}. \end{aligned}$$

In particular, with $\lambda = 1/2$, this implies (B.3). Further, for $\|\boldsymbol{a}\| = 1$

$$\begin{aligned} I\!E\bigl\{|\boldsymbol{\gamma}^\top \boldsymbol{a}|^2 \, \mathbb{I}(\|\boldsymbol{\gamma} - \boldsymbol{u}\| \ge z)\bigr\} & \le \exp\bigl(-z^2/4\bigr) I\!E\bigl\{|\boldsymbol{\gamma}^\top \boldsymbol{a}|^2 \exp\bigl(\|\boldsymbol{\gamma} - \boldsymbol{u}\|^2/4\bigr)\bigr\} \\ & \le (2 + |\boldsymbol{u}^\top \boldsymbol{a}|^2) \exp\bigl(-z^2/4 + p/2 + \|\boldsymbol{u}\|^2/2\bigr) \end{aligned}$$

and (B.4) follows.

B.2 Gaussian integrals

Let $\mathcal{T}$ be a linear mapping in $I\!R^p$ with $\|\mathcal{T}\|_{\mathrm{op}} \le 1$. Given positive $\mathtt{r}_0$ and $\mathtt{C}_0$, consider the following ratio of two integrals

$$\frac{\int_{\|\mathcal{T}\boldsymbol{u}\| > \mathtt{r}_0} \exp\bigl(-\mathtt{C}_0 \|\mathcal{T}\boldsymbol{u}\| + \frac{1}{2}\mathtt{C}_0 \mathtt{r}_0^2 + \frac{1}{2}\|\mathcal{T}\boldsymbol{u}\|^2 - \frac{1}{2}\|\boldsymbol{u}\|^2\bigr) d\boldsymbol{u}}{\int_{\|\mathcal{T}\boldsymbol{u}\| \le \mathtt{r}_0} \exp\bigl(-\frac{1}{2}\|\boldsymbol{u}\|^2\bigr) d\boldsymbol{u}}.$$

Obviously, one can rewrite this value as ratio of two expectations

$$\frac{I\!E\Bigl\{\exp\bigl(-\mathtt{C}_0 \mathtt{r}_0 \|\mathcal{T}\boldsymbol{\gamma}\| + \frac{1}{2}\mathtt{C}_0 \mathtt{r}_0^2 + \frac{1}{2}\|\mathcal{T}\boldsymbol{\gamma}\|^2\bigr) \, \mathbb{I}\bigl(\|\mathcal{T}\boldsymbol{\gamma}\| > \mathtt{r}_0\bigr)\Bigr\}}{I\!P\bigl(\|\mathcal{T}\boldsymbol{\gamma}\| \le \mathtt{r}_0\bigr)}.$$

Note that without the linear term $-\mathtt{C}_0 \|\mathcal{T}\boldsymbol{\gamma}\|$ in the exponent, the expectation in the numerator can be infinite. We aim at describing $\mathtt{r}_0$ and $\mathtt{C}_0$-values which ensure that the probability in denominator is close to one while the expectation in the numerator is small.

Lemma B.2.1. *Let* $\mathcal{T}$ *be a linear operator in* $I\!\!R^{p}$ *-matrix with* $\|\mathcal{T}\|_{\mathrm{op}} \le 1$. *Define* $\mathtt{p}_{\mathcal{T}} = \operatorname{tr}(\mathcal{T}^{\top}\mathcal{T})$. *For any* $\mathtt{C}_0, \mathtt{r}_0$ *with* $1/2 < \mathtt{C}_0 \le 1$ *and* $\mathtt{C}_0\mathtt{r}_0 = 2\sqrt{\mathtt{p}_{\mathcal{T}}} + \sqrt{\mathtt{x}}$ *for* $\mathtt{x} > 0$

$$I\!\!E\Bigl\{\exp\Bigl(-\mathtt{C}_0\mathtt{r}_0\|\mathcal{T}\boldsymbol{\gamma}\| + \frac{\mathtt{C}_0\mathtt{r}_0^2}{2} + \frac{1}{2}\|\mathcal{T}\boldsymbol{\gamma}\|^2\Bigr)\,\mathrm{1\!I}\bigl(\|\mathcal{T}\boldsymbol{\gamma}\| > \mathtt{r}_0\bigr)\Bigr\} \le \mathtt{C}\mathrm{e}^{-(\mathtt{p}_{\mathcal{T}}+\mathtt{x})/2}$$

and

$$I\!\!P\bigl(\|\mathcal{T}\boldsymbol{\gamma}\| \le \mathtt{r}_0\bigr) \ge 1 - \exp\Bigl\{-\frac{1}{2}(\mathtt{r}_0 - \sqrt{\mathtt{p}_{\mathcal{T}}})^2\Bigr\} \ge 1 - \mathrm{e}^{-\mathtt{p}_{\mathcal{T}}-\mathtt{x}}.$$

Remark B.2.1. The result applies even if the full dimension p is infinite and $\boldsymbol{\gamma}$ is a Gaussian element in a Hilbert space, provided that $\mathtt{p}_{\mathcal{T}} = \operatorname{tr}(\mathcal{T}^{\top}\mathcal{T})$ is finite, that is, $\mathcal{T}^{\top}\mathcal{T}$ is a trace operator.

Proof. Define

$$\begin{aligned}
\Phi(\mathtt{r}) &\stackrel{\mathrm{def}}{=} I\!\!P\bigl(\|\mathcal{T}\boldsymbol{\gamma}\| \ge \mathtt{r}\bigr),\\
f(\mathtt{r}) &\stackrel{\mathrm{def}}{=} \exp\Bigl(-\mathtt{C}_0\mathtt{r}_0\mathtt{r} + \frac{\mathtt{C}_0\mathtt{r}_0^2}{2} + \frac{\mathtt{r}^2}{2}\Bigr)
\end{aligned}$$

Then

$$\begin{aligned}
&I\!\!E\Bigl\{\exp\Bigl(-\mathtt{C}_0\mathtt{r}_0\|\mathcal{T}\boldsymbol{\gamma}\| + \frac{\mathtt{C}_0\mathtt{r}_0^2}{2} + \frac{1}{2}\|\mathcal{T}\boldsymbol{\gamma}\|^2\Bigr)\,\mathrm{1\!I}\bigl(\|\mathcal{T}\boldsymbol{\gamma}\| > \mathtt{r}_0\bigr)\Bigr\}\\
&\quad = -\int_{\mathtt{r}_0}^{\infty} f(\mathtt{r})d\Phi(\mathtt{r}) = -f(\mathtt{r}_0)\Phi(\mathtt{r}_0) + \int_{\mathtt{r}_0}^{\infty} f'(\mathtt{r})\Phi(\mathtt{r})\,d\mathtt{r}\,.
\end{aligned}$$

Now we use that $\Phi\bigl(\sqrt{\mathtt{p}_{\mathcal{T}}} + \sqrt{2\mathtt{x}}\bigr) \le \mathrm{e}^{-\mathtt{x}}$ for any $\mathtt{x} > 0$. This can be rewritten as

$$\Phi(\mathtt{r}) \le \exp\Bigl\{-\frac{1}{2}\bigl(\mathtt{r} - \sqrt{\mathtt{p}_{\mathcal{T}}}\bigr)^2\Bigr\}$$

for $\mathtt{r} > \sqrt{\mathtt{p}_{\mathcal{T}}}$. Now we use that $f'(\mathtt{r}) = (\mathtt{r} - \mathtt{C}_0\mathtt{r}_0)f(\mathtt{r})$ and

$$\begin{aligned}
\int_{\mathtt{r}_0}^{\infty} f'(\mathtt{r})\Phi(\mathtt{r})\,d\mathtt{r} &= \int_{\mathtt{r}_0}^{\infty}(\mathtt{r} - \mathtt{C}_0\mathtt{r}_0)f(\mathtt{r})\Phi(\mathtt{r})\,d\mathtt{r}\\
&\le \int_{\mathtt{r}_0}^{\infty}(\mathtt{r} - \mathtt{C}_0\mathtt{r}_0)\exp\Bigl\{-\frac{1}{2}\bigl(\mathtt{r} - \sqrt{\mathtt{p}_{\mathcal{T}}}\bigr)^2 - \mathtt{C}_0\mathtt{r}_0\mathtt{r} + \frac{\mathtt{C}_0\mathtt{r}_0^2}{2} + \frac{\mathtt{r}^2}{2}\Bigr\}\,d\mathtt{r}\\
&= \int_{\mathtt{r}_0}^{\infty}(\mathtt{r} - \mathtt{C}_0\mathtt{r}_0)\exp\Bigl\{-\bigl(\mathtt{C}_0\mathtt{r}_0 - \sqrt{\mathtt{p}_{\mathcal{T}}}\bigr)\mathtt{r} + \frac{\mathtt{C}_0\mathtt{r}_0^2}{2} - \frac{\mathtt{p}_{\mathcal{T}}}{2}\Bigr\}\,d\mathtt{r}\\
&= \int_{0}^{\infty}(x + \mathtt{r}_0 - \mathtt{C}_0\mathtt{r}_0)\exp\Bigl\{-\bigl(\mathtt{C}_0\mathtt{r}_0 - \sqrt{\mathtt{p}_{\mathcal{T}}}\bigr)(x + \mathtt{r}_0) + \frac{\mathtt{C}_0\mathtt{r}_0^2}{2} - \frac{\mathtt{p}_{\mathcal{T}}}{2}\Bigr\}\,dx.
\end{aligned}$$

The use of $\int_0^{\infty}\mathrm{e}^{-x}dx = \int_0^{\infty}x\mathrm{e}^{-x}dx = 1$ yields

$$\int_{\mathtt{r}_0}^{\infty} f'(\mathtt{r})\Phi(\mathtt{r})\,d\mathtt{r} \le \Bigl(\frac{\mathtt{r}_0 - \mathtt{C}_0\mathtt{r}_0}{\mathtt{C}_0\mathtt{r}_0 - \sqrt{\mathtt{p}_{\mathcal{T}}}} + \frac{1}{(\mathtt{C}_0\mathtt{r}_0 - \sqrt{\mathtt{p}_{\mathcal{T}}})^2}\Bigr)\exp\Bigl\{\mathtt{r}_0\sqrt{\mathtt{p}_{\mathcal{T}}} - \frac{\mathtt{C}_0\mathtt{r}_0^2}{2} - \frac{\mathtt{p}_{\mathcal{T}}}{2}\Bigr\}.$$

It remains to check that for $\mathtt{C}_0 \in (1/2, 1)$ and $\mathtt{C}_0 \mathtt{r}_0 = 2\sqrt{\mathtt{p}_{\mathcal{T}}} + \sqrt{\mathtt{x}}$

$$-\mathtt{r}_0 \sqrt{\mathtt{p}_{\mathcal{T}}} + \frac{\mathtt{C}_0 \mathtt{r}_0^2}{2} + \frac{\mathtt{p}_{\mathcal{T}}}{2} \ge \frac{\mathtt{x} + \mathtt{p}_{\mathcal{T}}}{2} .$$

Now we consider Gaussian integrals with an additional quadratic multiplier.

Lemma B.2.2. *Let* $\mathcal{T}$ *be a linear operator in* $I\!R^p$ *with* $\|\mathcal{T}\|_{\mathrm{op}} \le 1$ *. Let* $\boldsymbol{\lambda} \in I\!R^\infty$ *be a unit norm vector:* $\|\boldsymbol{\lambda}\| = 1$ *. Define* $\mathtt{p}_{\mathcal{T}} = \operatorname{tr}(\mathcal{T}^\top \mathcal{T})$ *. For any positive* $\mathtt{C}_0, \mathtt{r}_0$ *with* $1/2 < \mathtt{C}_0 \le 1$ *and* $\mathtt{C}_0 \mathtt{r}_0 > 2\sqrt{\mathtt{p}_{\mathcal{T}} + 1} + \sqrt{\mathtt{y}}$

$$I\!E\Bigl\{ |\boldsymbol{\lambda}^\top \boldsymbol{\gamma}|^2 \exp\Bigl(-\mathtt{C}_0 \mathtt{r}_0 \|\mathcal{T}\boldsymbol{\gamma}\| + \frac{\mathtt{C}_0 \mathtt{r}_0^2}{2} + \frac{1}{2}\|\mathcal{T}\boldsymbol{\gamma}\|^2 \Bigr) \operatorname{1\!I}\bigl(\|\mathcal{T}\boldsymbol{\gamma}\| > \mathtt{r}_0 \bigr) \Bigr\} \le \mathtt{C} \mathrm{e}^{-(\mathtt{p}_{\mathcal{T}} + \mathtt{y})/2}.$$

Proof. Define $\mathcal{T}_{\boldsymbol{z}}$ by $\mathcal{T}_{\boldsymbol{z}}^\top \mathcal{T}_{\boldsymbol{z}} = \mathcal{T}^\top \mathcal{T} + \boldsymbol{\lambda}\boldsymbol{\lambda}^\top$. Obviously $\|\mathcal{T}_{\boldsymbol{z}}\boldsymbol{\gamma}\| \ge \|\mathcal{T}\boldsymbol{\gamma}\|$, $|\boldsymbol{\lambda}^\top \boldsymbol{\gamma}| \le \|\mathcal{T}\boldsymbol{\gamma}\|$. Further, $\mathtt{r}^2/2 - \mathtt{C}_0 \mathtt{r}_0 \mathtt{r}$ grows in $\mathtt{r} \ge \mathtt{r}_0$ in view of $\mathtt{C}_0 \le 1$. Therefore,

$$\begin{aligned} & |\boldsymbol{\lambda}^\top \boldsymbol{\gamma}|^2 \exp\Bigl(-\mathtt{C}_0 \mathtt{r}_0 \|\mathcal{T}\boldsymbol{\gamma}\| + \frac{\mathtt{C}_0 \mathtt{r}_0^2}{2} + \frac{1}{2}\|\mathcal{T}\boldsymbol{\gamma}\|^2 \Bigr) \operatorname{1\!I}\bigl(\|\mathcal{T}\boldsymbol{\gamma}\| > \mathtt{r}_0 \bigr) \\ & \quad \le \|\mathcal{T}_{\boldsymbol{z}}\boldsymbol{\gamma}\|^2 \exp\Bigl(-\mathtt{C}_0 \mathtt{r}_0 \|\mathcal{T}_{\boldsymbol{z}}\boldsymbol{\gamma}\| + \frac{\mathtt{C}_0 \mathtt{r}_0^2}{2} + \frac{1}{2}\|\mathcal{T}_{\boldsymbol{z}}\boldsymbol{\gamma}\|^2 \Bigr) \operatorname{1\!I}\bigl(\|\mathcal{T}_{\boldsymbol{z}}\boldsymbol{\gamma}\| > \mathtt{r}_0 \bigr) \end{aligned}$$

Now we can follow the line of the proof of Lemma B.2.1. Consider

$$\begin{aligned} \Phi_{\boldsymbol{\lambda}}(\mathtt{r}) &= I\!P\bigl(\|\mathcal{T}_{\boldsymbol{z}}\boldsymbol{\gamma}\| \ge \mathtt{r} \bigr) \le \exp\bigl\{ -\frac{1}{2}(\mathtt{r} - \sqrt{\mathtt{p}_{\boldsymbol{\lambda}}}) \bigr\}, \\ f(\mathtt{r}) &\stackrel{\mathrm{def}}{=} \mathtt{r}^2 \exp\Bigl(-\mathtt{C}_0 \mathtt{r}_0 \mathtt{r} + \frac{\mathtt{C}_0 \mathtt{r}_0^2}{2} + \frac{\mathtt{r}^2}{2} \Bigr) \end{aligned}$$

with $\mathtt{p}_{\boldsymbol{\lambda}} \stackrel{\mathrm{def}}{=} \operatorname{tr} \mathcal{T}_{\boldsymbol{z}}^\top \mathcal{T}_{\boldsymbol{z}} = \mathtt{p} + 1$. Then

$$\begin{aligned} & I\!E\Bigl\{ \bigl(\boldsymbol{\lambda}^\top \boldsymbol{\gamma} \bigr)^2 \exp\Bigl(-\mathtt{C}_0 \mathtt{r}_0 \|\mathcal{T}\boldsymbol{\gamma}\| + \frac{\mathtt{C}_0 \mathtt{r}_0^2}{2} + \frac{1}{2}\|\mathcal{T}\boldsymbol{\gamma}\|^2 \Bigr) \operatorname{1\!I}\bigl(\|\mathcal{T}\boldsymbol{\gamma}\| > \mathtt{r}_0 \bigr) \Bigr\} \\ & \quad \le -\int_{\mathtt{r}_0}^{\infty} f(\mathtt{r}) d\Phi_{\boldsymbol{\lambda}}(\mathtt{r}) = -f(\mathtt{r}_0)\Phi_{\boldsymbol{\lambda}}(\mathtt{r}_0) + \int_{\mathtt{r}_0}^{\infty} f'(\mathtt{r})\Phi_{\boldsymbol{\lambda}}(\mathtt{r})\, d\mathtt{r} . \end{aligned}$$

Now we can continue as in the proof of Lemma B.2.1.

C

Deviation bounds for quadratic forms

Here we collect some probability bounds for Gaussian quadratic forms.

C.1 Gaussian quadratic forms

The next result explains the concentration effect of $\boldsymbol{\gamma}^{\top} B \boldsymbol{\gamma}$ for a standard Gaussian vector $\boldsymbol{\gamma}$ and a symmetric matrix B. We use a version from Laurent and Massart (2000) with a complete proof.

Theorem C.1.1. *Let $\boldsymbol{\gamma}$ be a standard normal Gaussian vector and B be symmetric positively definite $p \times p$-matrix. Then with $\mathtt{p} = \operatorname{tr}(B)$, $\mathtt{v}^2 = \operatorname{tr}(B^2)$, and $\lambda = \|B\|_{\mathrm{op}}$, it holds for each $\mathtt{x} \geq 0$*

$$\mathbb{P}\Bigl(\boldsymbol{\gamma}^{\top} B \boldsymbol{\gamma} > z^2(B, \mathtt{x})\Bigr) \;\leq\; \mathrm{e}^{-\mathtt{x}}, \tag{C.1}$$

$$z(B, \mathtt{x}) \overset{\mathrm{def}}{=} \sqrt{\mathtt{p} + 2\mathtt{v}\mathtt{x}^{1/2} + 2\lambda\mathtt{x}}. \tag{C.2}$$

In particular, it implies

$$\mathbb{P}\bigl(\|B^{1/2}\boldsymbol{\gamma}\| > \mathtt{p}^{1/2} + (2\lambda\mathtt{x})^{1/2}\bigr) \leq \mathrm{e}^{-\mathtt{x}}.$$

Also

$$\mathbb{P}\bigl(\boldsymbol{\gamma}^{\top} B \boldsymbol{\gamma} < \mathtt{p} - 2\mathtt{v}\mathtt{x}^{1/2}\bigr) \leq \mathrm{e}^{-\mathtt{x}}. \tag{C.3}$$

If B is symmetric but non necessarily positive then

$$\mathbb{P}\bigl(\bigl|\boldsymbol{\gamma}^{\top} B \boldsymbol{\gamma} - \mathtt{p}\bigr| > 2\mathtt{v}\mathtt{x}^{1/2} + 2\lambda\mathtt{x}\bigr) \;\leq\; 2\mathrm{e}^{-\mathtt{x}}.$$

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 $\|\boldsymbol{\gamma}\|^2$ 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 $\mathtt{p} = \lambda_1 + \ldots + \lambda_p$, $\mathtt{v}^2 = \lambda_1^2 + \ldots + \lambda_p^2$. One can easily compute the exponential moment of $(\boldsymbol{\gamma}^{\top} B \boldsymbol{\gamma} - \mathtt{p})/2$: for each positive $\mu < 1$

$$\log I\!\!E \exp\bigl\{\mu(\boldsymbol{\gamma}^{\top} B \boldsymbol{\gamma} - \mathtt{p})/2\bigr\} = \frac{1}{2} \sum_{j=1}^{p} \bigl\{-\mu\lambda_j - \log(1-\mu\lambda_j)\bigr\}. \tag{C.4}$$

Lemma C.1.1. *Let* $\mu\lambda_j < 1$ *and* $\lambda_j \le 1$. *Then*

$$\frac{1}{2} \sum_{j=1}^{p} \bigl\{-\mu\lambda_j - \log(1-\mu\lambda_j)\bigr\} \le \frac{\mu^2 \mathtt{v}^2}{4(1-\mu)}.$$

Proof. In view of $\mu\lambda_j < 1$, it holds for every j

$$\begin{aligned} -\mu\lambda_j - \log(1-\mu\lambda_j) &= \sum_{k=2}^{\infty} \frac{(\mu\lambda_j)^k}{k} \\ &\le \frac{(\mu\lambda_j)^2}{2} \sum_{k=0}^{\infty} (\mu\lambda_j)^k \le \frac{(\mu\lambda_j)^2}{2(1-\mu\lambda_j)} \le \frac{(\mu\lambda_j)^2}{2(1-\mu)}, \end{aligned} \tag{C.5}$$

and thus

$$\frac{1}{2} \sum_{j=1}^{p} \bigl\{-\mu\lambda_j - \log(1-\mu\lambda_j)\bigr\} \le \sum_{j=1}^{p} \frac{(\mu\lambda_j)^2}{4(1-\mu)} \le \frac{\mu^2 \mathtt{v}^2}{4(1-\mu)}.$$

The next technical lemma is helpful.

Lemma C.1.2. *For each* $\mathtt{v} > 0$ *and* $\mathtt{x} > 0$, *it holds*

$$\inf_{\mu > 0} \biggl\{ -\mu\bigl(\mathtt{v}\mathtt{x}^{1/2} + \mathtt{x}\bigr) + \frac{\mu^2 \mathtt{v}^2}{4(1-\mu)} \biggr\} \le -\mathtt{x}.$$

Proof. Let pick up

$$\mu = 1 - \frac{1}{2\mathtt{x}^{1/2}/\mathtt{v} + 1} = \frac{\mathtt{x}^{1/2}}{\mathtt{x}^{1/2} + \mathtt{v}/2},$$

so that $\mu/(1-\mu) = 2\mathtt{x}^{1/2}/\mathtt{v}$. Then

$$\begin{aligned} & -\mu\bigl(\mathtt{v}\mathtt{x}^{1/2} + \mathtt{x}\bigr) + \frac{\mu^2 \mathtt{v}^2}{4(1-\mu)} \\ & \quad = -\mu\bigl(\mathtt{v}\mathtt{x}^{1/2} + \mathtt{x} + \mathtt{v}^2/4\bigr) + \frac{\mu \mathtt{v}^2}{4(1-\mu)} \\ & \quad = -\frac{\mathtt{x}^{1/2}}{\mathtt{x}^{1/2} + \mathtt{v}/2} \bigl(\mathtt{x}^{1/2} + \mathtt{v}/2\bigr)^2 + \frac{2\mathtt{x}^{1/2}\mathtt{v}}{4} = -\mathtt{x} \end{aligned} \tag{C.6}$$

and the result follows.

Now we apply the Markov inequality

$$\begin{aligned}\log I\!P\bigl(\boldsymbol{\gamma}^{\top} B \boldsymbol{\gamma} > \mathtt{p} + 2\mathtt{v}\mathtt{x}^{1/2} + 2\mathtt{x}\bigr) &= \log I\!P\bigl((\boldsymbol{\gamma}^{\top} B \boldsymbol{\gamma} - \mathtt{p})/2 > \mathtt{v}\mathtt{x}^{1/2} + \mathtt{x}\bigr) \\ &\leq \inf_{\mu>0} \Bigl\{ -\mu\bigl(\mathtt{v}\mathtt{x}^{1/2} + \mathtt{x}\bigr) + \log I\!E \exp\bigl\{\mu(\boldsymbol{\gamma}^{\top} B \boldsymbol{\gamma} - \mathtt{p})/2\bigr\} \Bigr\} \\ &\leq \inf_{\mu>0} \Bigl\{ -\mu\bigl(\mathtt{v}\mathtt{x}^{1/2} + \mathtt{x}\bigr) + \frac{\mu^2 \mathtt{v}^2}{4(1-\mu)} \Bigr\} \leq -\mathtt{x}\end{aligned}$$

and the first assertion (C.1) follows. The second statement follows from the first one by $\operatorname{tr}(B^2) \leq \|B\|_{\mathrm{op}} \operatorname{tr}(B) = \lambda\,\mathtt{p}$.

Similarly for any $\mu > 0$

$$I\!P\bigl(\boldsymbol{\gamma}^{\top} B \boldsymbol{\gamma} - \mathtt{p} < -2\mathtt{v}\sqrt{\mathtt{x}}\bigr) \leq \exp\bigl(-\mu\mathtt{v}\sqrt{\mathtt{x}}\bigr) I\!E \exp\Bigl(-\frac{\mu}{2}(\boldsymbol{\gamma}^{\top} B \boldsymbol{\gamma} - \mathtt{p})\Bigr).$$

By (C.4)

$$\log I\!E \exp\bigl\{-\mu(\boldsymbol{\gamma}^{\top} B \boldsymbol{\gamma} - \mathtt{p})/2\bigr\} = \frac{1}{2}\sum_{j=1}^{p} \bigl\{\mu\lambda_j - \log(1 + \mu\lambda_j)\bigr\}.$$

and

$$\frac{1}{2}\sum_{j=1}^{p} \bigl\{\mu\lambda_j - \log(1 + \mu\lambda_j)\bigr\} = \frac{1}{2}\sum_{j=1}^{p}\sum_{k=2}^{\infty} \frac{(-\mu\lambda_j)^k}{k} \leq \sum_{j=1}^{p} \frac{(\mu\lambda_j)^2}{4} = \frac{\mu^2 \mathtt{v}^2}{4}.$$

Here the choice $\mu = 2\sqrt{\mathtt{x}}/\mathtt{v}$ yields (C.3).

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 C.1.1. *For any unit vector* $\boldsymbol{u} = (u_i) \in I\!R^n$ *and standard normal r.v.'s* γ_i, *it holds with* $\|\boldsymbol{u}\|_{\infty} \stackrel{\mathrm{def}}{=} \max_i |u_i|$

$$I\!P\biggl(\Bigl|\sum_{i=1}^{n} u_i(\gamma_i^2 - 1)\Bigr| \geq 2\mathtt{x}^{1/2} + 2\|\boldsymbol{u}\|_{\infty}\mathtt{x}\biggr) \leq 2\mathrm{e}^{-\mathtt{x}}.$$

Proof. The statement follows directly from Theorem C.1.1. It suffices to notice $\mathtt{v}^2 = \|\boldsymbol{u}\|^2 = 1$.

As a special case, we present a bound for the chi-squared distribution corresponding to $B = \boldsymbol{I}_p$. Then $\operatorname{tr}(B) = p$, $\operatorname{tr}(B^2) = p$ and $\lambda(B) = 1$.

Corollary C.1.2. *Let $\boldsymbol{\gamma}$ be a standard normal vector in $I\!\!R^p$. Then*

$$\begin{aligned} I\!\!P\bigl(\|\boldsymbol{\gamma}\|^2 \geq p + 2\sqrt{p\mathtt{x}} + 2\mathtt{x}\bigr) &\leq \mathrm{e}^{-\mathtt{x}}, \\ I\!\!P\bigl(\|\boldsymbol{\gamma}\|^2 \leq p - 2\sqrt{p\mathtt{x}}\bigr) \quad &\leq \mathrm{e}^{-\mathtt{x}}, \\ I\!\!P\bigl(\|\boldsymbol{\gamma}\| \;\geq \sqrt{p} + \sqrt{2\mathtt{x}}\bigr) \quad &\leq \mathrm{e}^{-\mathtt{x}}. \end{aligned} \tag{C.7}$$

The previous results are mainly stated for a standard Gaussian vector $\boldsymbol{\gamma} \in I\!\!R^n$. Now we extend it to the case of a zero mean Gaussian vector $\boldsymbol{\xi}$ with the $n \times n$ covariance matrix $\mathbb{V} = (\sigma_{ij})$ with $\lambda_{\max}(\mathbb{V}) \leq \lambda^*$. Given a unit vector $\boldsymbol{u} = (u_1, \ldots, u_n)^\top \in I\!\!R^n$, consider the quadratic form

$$Q = \sum_{i=1}^{n} u_i \xi_i^2 .$$

We aim at bounding $Q - I\!\!E Q$. To apply the result of Theorem C.1.1 represent Q as $\boldsymbol{\gamma}^\top B \boldsymbol{\gamma}$ with B depending on $\boldsymbol{u}$ and $\mathbb{V}$. More precisely, let $\boldsymbol{\xi} = \mathbb{V}^{1/2}\boldsymbol{\gamma}$ for a standard Gaussian vector $\boldsymbol{\gamma} \in I\!\!R^n$. Then with $\boldsymbol{U} = \operatorname{diag}(u_1, \ldots, u_n)$, it holds

$$S = \operatorname{tr}\bigl(\boldsymbol{U}\boldsymbol{\xi}\boldsymbol{\xi}^\top\bigr) = \operatorname{tr}\bigl(\boldsymbol{U}\mathbb{V}^{1/2}\boldsymbol{\gamma}\boldsymbol{\gamma}^\top\mathbb{V}^{1/2}\bigr) = \operatorname{tr}\bigl(B\boldsymbol{\gamma}\boldsymbol{\gamma}^\top\bigr) = \boldsymbol{\gamma}^\top B \boldsymbol{\gamma}$$

with $B = \mathbb{V}^{1/2}\boldsymbol{U}\mathbb{V}^{1/2}$. Therefore, the bound $\|\mathbb{V}\|_{\mathrm{op}} \leq \lambda^*$ implies

$$\begin{aligned} \lambda &= \lambda(B) = \|\mathbb{V}^{1/2}\boldsymbol{U}\mathbb{V}^{1/2}\|_{\mathrm{op}} \leq \lambda^* \, \|\boldsymbol{u}\|_\infty \, , \\ \mathtt{v}^2 &= \operatorname{tr}(B^2) = \operatorname{tr}\bigl(\mathbb{V}^{1/2}\boldsymbol{U}\mathbb{V}\boldsymbol{U}\mathbb{V}^{1/2}\bigr) \leq \lambda^* \operatorname{tr}\bigl(\boldsymbol{U}\mathbb{V}\boldsymbol{U}\bigr) \leq \lambda^{*2}\|\boldsymbol{u}\|^2 = \lambda^{*2}. \end{aligned}$$

Now the general results of Theorem C.1.1 implies the result similar to Corollary C.1.1.

Corollary C.1.3. *For any unit vector $\boldsymbol{u} = (u_i) \in I\!\!R^n$, $\|\boldsymbol{u}\| = 1$, and normal zero mean vector $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbb{V})$ in $I\!\!R^n$ with $\|\mathbb{V}\|_{\mathrm{op}} \leq \lambda^*$, it holds*

$$I\!\!P\biggl(\biggl|\sum_{i=1}^{n} u_i(\xi_i^2 - I\!\!E\xi_i^2)\biggr| \geq 2\lambda^* \, \mathtt{x}^{1/2} + 2\lambda^* \, \|\boldsymbol{u}\|_\infty \mathtt{x}\biggr) \leq 2\mathrm{e}^{-\mathtt{x}}.$$

It is worth noting that the identity $\|\boldsymbol{u}\| = 1$ implies $\|\boldsymbol{u}\|_\infty \leq 1$. Moreover, in typical situations, $\|\boldsymbol{u}\|_\infty \asymp n^{-1/2}$, and the leading term in the bounds of Corollaries C.1.1 and C.1.3 is $2\lambda^* \, \mathtt{x}^{1/2}$.

C.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.

C.2.1 Deviation bounds for the norm of a standardized non-Gaussian vector

The bounds of Corollary C.1.2 heavily use normality of the vector $\boldsymbol{\xi}$. This section extends the upper bound (C.7) to the case when $\boldsymbol{\xi}$ has some exponential moments. More exactly, suppose for some fixed $\mathtt{g} > 0$ that

$$\log I\!\!E \exp\bigl(\boldsymbol{\gamma}^{\top}\boldsymbol{\xi}\bigr) \le \|\boldsymbol{\gamma}\|^2/2, \qquad \boldsymbol{\gamma} \in I\!\!R^p,\ \|\boldsymbol{\gamma}\| \le \mathtt{g}. \tag{C.8}$$

For ease of presentation, assume below that $\mathtt{g}$ is sufficiently large, namely, $0.3\mathtt{g} \ge \sqrt{p}$. In typical examples of an i.i.d. sample, $\mathtt{g} \asymp \sqrt{n}$. Define

$$\begin{aligned} \mathtt{x}_c &\overset{\mathrm{def}}{=} \mathtt{g}^2/4, \\ z_c^2 &\overset{\mathrm{def}}{=} p + \sqrt{p\mathtt{g}^2} + \mathtt{g}^2/2 = \mathtt{g}^2\bigl(1/2 + \sqrt{p/\mathtt{g}^2} + p/\mathtt{g}^2\bigr), \\ \mathtt{g}_c &\overset{\mathrm{def}}{=} \frac{\mathtt{g}\,\bigl(1/2 + \sqrt{p/\mathtt{g}^2} + p/\mathtt{g}^2\bigr)^{1/2}}{1 + \sqrt{p/\mathtt{g}^2}}. \end{aligned}$$

Note that with $\alpha = \sqrt{p/\mathtt{g}^2} \le 0.3$, one has

$$\begin{aligned} z_c^2 &= \mathtt{g}^2\bigl(1/2 + \alpha + \alpha^2\bigr), \\ \mathtt{g}_c &= \mathtt{g}\,\frac{\bigl(1/2 + \alpha + \alpha^2\bigr)^{1/2}}{1 + \alpha} \end{aligned}$$

so that $z_c^2/\mathtt{g}^2 \in [1/2, 1]$ and $\mathtt{g}_c^2/\mathtt{g}^2 \in [1/2, 1]$.

Theorem C.2.1. *Let (C.8) hold and $0.3\mathtt{g} \ge \sqrt{p}$. Then for each $\mathtt{x} > 0$*

$$I\!\!P\bigl(\|\boldsymbol{\xi}\| \ge z(p,\mathtt{x})\bigr) \le 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c}\, 1\!\!I(\mathtt{x} < \mathtt{x}_c), \tag{C.9}$$

where $z(p,\mathtt{x})$ is defined by

$$z(p,\mathtt{x}) \overset{\mathrm{def}}{=} \begin{cases} \bigl(p + 2\sqrt{p\mathtt{x}} + 2\mathtt{x}\bigr)^{1/2}, & \mathtt{x} \le \mathtt{x}_c, \\ z_c + 2\mathtt{g}_c^{-1}(\mathtt{x} - \mathtt{x}_c), & \mathtt{x} > \mathtt{x}_c. \end{cases}$$

Depending on the value $\mathtt{x}$, we have two types of tail behavior of the quadratic form $\|\boldsymbol{\xi}\|^2$. For $\mathtt{x} \le \mathtt{x}_c = \mathtt{g}^2/4$, 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 $\bigl(p + 2\sqrt{p\mathtt{x}} + 2\mathtt{x}\bigr)^{1/2} \le \sqrt{p} + \sqrt{2\mathtt{x}}$. For $\mathtt{x} > \mathtt{x}_c$, we switch to the special regime driven by the exponential moment condition (C.8). Usually $\mathtt{g}^2$ is a large number (of order n in the i.i.d. setup) and the second term in (C.9) can be simply ignored.

For the sub-Gaussian case with $\mathtt{g} = \infty$, the result can be simplified and it looks exactly as in the Gaussian case.

Corollary C.2.1. *Let the conditions of Theorem C.2.1 hold with* $\mathtt{g} = \infty$. *Then*

$$\mathbb{P}\bigl(\|\boldsymbol{\xi}\| \ge z(p,\mathtt{x})\bigr) \le 2\mathrm{e}^{-\mathtt{x}},$$

where $z(p,\mathtt{x})$ *is defined by*

$$z(p,\mathtt{x}) \stackrel{\mathrm{def}}{=} \bigl(p + 2\sqrt{p\mathtt{x}} + 2\mathtt{x}\bigr)^{1/2} \le \sqrt{p} + \sqrt{2\mathtt{x}}\,.$$

The main step of the proof is the following exponential bound.

Lemma C.2.1. *Suppose* (C.8). *For any* $\mu < 1$ *with* $\mathtt{g}^2 > p\mu$, *it holds*

$$\mathbb{E}\exp\Bigl(\frac{\mu\|\boldsymbol{\xi}\|^2}{2}\Bigr)\,\mathbb{1}\Bigl(\|\boldsymbol{\xi}\| \le \mathtt{g}/\mu - \sqrt{p/\mu}\Bigr) \le 2(1-\mu)^{-p/2}. \tag{C.10}$$

Proof. Let $\boldsymbol{\varepsilon}$ be a standard normal vector in $\mathbb{R}^p$ and $\boldsymbol{u} \in \mathbb{R}^p$. The bound $\mathbb{P}\bigl(\|\boldsymbol{\varepsilon}\|^2 > p\bigr) \le 1/2$ and the triangle inequality imply for any vector $\boldsymbol{u}$ and any $\mathtt{r}$ with $\mathtt{r} \ge \|\boldsymbol{u}\| + p^{1/2}$ that $\mathbb{P}\bigl(\|\boldsymbol{u} + \boldsymbol{\varepsilon}\| \le \mathtt{r}\bigr) \ge 1/2$. Let us fix some $\boldsymbol{\xi}$ with $\|\boldsymbol{\xi}\| \le \mathtt{g}/\mu - \sqrt{p/\mu}$ and denote by $\mathbb{P}_{\boldsymbol{\xi}}$ the conditional probability given $\boldsymbol{\xi}$. The previous arguments yield:

$$\mathbb{P}_{\boldsymbol{\xi}}\bigl(\|\boldsymbol{\varepsilon} + \mu^{1/2}\boldsymbol{\xi}\| \le \tau\mathtt{g}\bigr) \ge 0.5.$$

It holds with $c_p = (2\pi)^{-p/2}$

$$\begin{aligned}
c_p & \int \exp\Bigl(\boldsymbol{\gamma}^\top\boldsymbol{\xi} - \frac{\|\boldsymbol{\gamma}\|^2}{2\mu}\Bigr)\,\mathbb{1}(\|\boldsymbol{\gamma}\| \le \mathtt{g})d\boldsymbol{\gamma} \\
&= c_p \exp\bigl(\mu\|\boldsymbol{\xi}\|^2/2\bigr) \int \exp\Bigl(-\frac{1}{2}\bigl\|\tau\boldsymbol{\gamma} - \mu^{1/2}\boldsymbol{\xi}\bigr\|^2\Bigr)\,\mathbb{1}(\tau\|\boldsymbol{\gamma}\| \le \tau\mathtt{g})d\boldsymbol{\gamma} \\
&= \mu^{p/2}\exp\bigl(\mu\|\boldsymbol{\xi}\|^2/2\bigr)\mathbb{P}_{\boldsymbol{\xi}}\bigl(\|\boldsymbol{\varepsilon} + \mu^{1/2}\boldsymbol{\xi}\| \le \tau\mathtt{g}\bigr) \\
&\ge 0.5\mu^{p/2}\exp\bigl(\mu\|\boldsymbol{\xi}\|^2/2\bigr),
\end{aligned}$$

because $\|\mu^{1/2}\boldsymbol{\xi}\| + p^{1/2} \le \tau\mathtt{g}$. This implies in view of $p < \mathtt{g}^2/\mu$ that

$$\begin{aligned}
& \exp\bigl(\mu\|\boldsymbol{\xi}\|^2/2\bigr)\,\mathbb{1}\bigl(\|\boldsymbol{\xi}\|^2 \le \mathtt{g}/\mu - \sqrt{p/\mu}\bigr) \\
& \qquad \le 2\mu^{-p/2}c_p \int \exp\Bigl(\boldsymbol{\gamma}^\top\boldsymbol{\xi} - \frac{\|\boldsymbol{\gamma}\|^2}{2\mu}\Bigr)\,\mathbb{1}(\|\boldsymbol{\gamma}\| \le \mathtt{g})d\boldsymbol{\gamma}.
\end{aligned}$$

Further, by (C.8)

$$\begin{aligned}
c_p\mathbb{E} & \int \exp\Bigl(\boldsymbol{\gamma}^\top\boldsymbol{\xi} - \frac{1}{2\mu}\|\boldsymbol{\gamma}\|^2\Bigr)\,\mathbb{1}(\|\boldsymbol{\gamma}\| \le \mathtt{g})d\boldsymbol{\gamma} \\
&\le c_p \int \exp\Bigl(-\frac{\mu^{-1} - 1}{2}\|\boldsymbol{\gamma}\|^2\Bigr)\,\mathbb{1}(\|\boldsymbol{\gamma}\| \le \mathtt{g})d\boldsymbol{\gamma} \\
&\le c_p \int \exp\Bigl(-\frac{\mu^{-1} - 1}{2}\|\boldsymbol{\gamma}\|^2\Bigr)d\boldsymbol{\gamma} \\
&\le (\mu^{-1} - 1)^{-p/2}
\end{aligned}$$

and (C.10) follows.

Due to this result, the scaled squared norm $\mu\|\boldsymbol{\xi}\|^2/2$ after a proper truncation possesses the same exponential moments as in the Gaussian case. A straightforward implication is the probability bound $I\!P\bigl(\|\boldsymbol{\xi}\|^2 > p+u\bigr)$ with $u = 2\sqrt{p\mathtt{x}} + 2\mathtt{x}$. Namely, given $\mathtt{x}$, define

$$\mu = \mu(\mathtt{x}) = \frac{1}{1 + 0.5\sqrt{p/\mathtt{x}}}\,. \tag{C.11}$$

Also define for $\mathtt{x}_c = \mathtt{g}^2/4$

$$\mu_c \overset{\mathrm{def}}{=} \mu(\mathtt{x}_c) = \frac{1}{1 + \sqrt{p/\mathtt{g}^2}}\,. \tag{C.12}$$

Obviously, $\mu \le \mu_c$ for $\mathtt{x} \le \mathtt{x}_c$. Now we obtain similarly to the Gaussian case in Lemma C.1.2 for $u = 2\sqrt{p\mathtt{x}} + 2\mathtt{x}$

$$\begin{aligned} & I\!P\Bigl(\|\boldsymbol{\xi}\|^2 > p+u,\, \|\boldsymbol{\xi}\| \le \mathtt{g}/\mu - \sqrt{p/\mu}\Bigr) \\ & \quad \le \exp\Bigl\{-\frac{\mu(p+u)}{2}\Bigr\} I\!E \exp\Bigl(\frac{\mu\|\boldsymbol{\xi}\|^2}{2}\Bigr)\, \mathbb{I}\Bigl(\|\boldsymbol{\xi}\| \le \mathtt{g}/\mu - \sqrt{p/\mu}\Bigr) \\ & \quad \le 2\exp\Bigl\{-\frac{1}{2}\bigl[\mu(p+u) + p\log(1-\mu)\bigr]\Bigr\} \end{aligned} \tag{C.13}$$

and by (C.6) with $\mathsf{v}^2 = p$, it holds for μ from (C.11)

$$\mu(p + 2\sqrt{p\mathtt{x}} + 2\mathtt{x}) + p\log(1-\mu) \ge 2\mathtt{x}. \tag{C.14}$$

Now we show that the constraint $\|\boldsymbol{\xi}\| \le \mathtt{g}/\mu - \sqrt{p/\mu}$ in (C.13) can be replaced by the inequality $\|\boldsymbol{\xi}\| \le z_c$.

Lemma C.2.2. *Let* $0.3\mathtt{g} \ge \sqrt{p}$, $\mathtt{x} \le \mathtt{x}_c = \mathtt{g}^2/4$, *and* $\mu = 1/(1 + 0.5\sqrt{p/\mathtt{x}})$. *Then*

$$\begin{aligned} p + 2\sqrt{p\mathtt{x}} + 2\mathtt{x} &\le p + 2\sqrt{p\mathtt{x}_c} + 2\mathtt{x}_c, \\ \mathtt{g}/\mu - \sqrt{p/\mu} &\ge \mathtt{g}/\mu_c - \sqrt{p/\mu_c}, \\ p + 2\sqrt{p\mathtt{x}_c} + 2\mathtt{x}_c &\le \bigl(\mathtt{g}/\mu_c - \sqrt{p/\mu_c}\bigr)^2. \end{aligned} \tag{C.15}$$

Proof. The definition implies $\mu \le \mu_c$ for $\mathtt{x} \le \mathtt{x}_c$ and thus the first two inequalities of the lemma are obvious. Therefore, it remains to check (C.15). Denote $\alpha^2 = p/\mathtt{g}^2$. Then $\mu_c^{-1} = 1 + \alpha$ and

$$\mathtt{g}/\mu_c - \sqrt{p/\mu_c} = \mu_c^{-1}\mathtt{g}\bigl(1 - \sqrt{\mu_c\alpha^2}\bigr) = \mathtt{g}\,(1+\alpha)\,\bigl\{1 - \sqrt{\alpha^2/(1+\alpha)}\bigr\}.$$

For $\mathtt{x}_c = \mathtt{g}^2/4$, it holds

$$p + 2\sqrt{p\mathtt{x}_c} + 2\mathtt{x}_c = p + \sqrt{p\mathtt{g}^2} + \mathtt{g}^2/2 = \mathtt{g}^2\bigl(\alpha^2 + \alpha + 1/2\bigr).$$

Direct calculus shows that for $\alpha \le 0.3$ one can bound

$$\alpha^2 + \alpha + 1/2 \le (1+\alpha)^2\Bigl\{1 - \sqrt{\alpha^2/(1+\alpha)}\Bigr\}^2 \tag{C.16}$$

and this proves (C.15).

We conclude from this lemma, (C.13) and (C.14) that

$$I\!\!P\bigl(\|\boldsymbol{\xi}\|^2 > p + 2\sqrt{p\mathtt{x}} + 2\mathtt{x}, \|\boldsymbol{\xi}\| \le z_c\bigr) \le 2\mathrm{e}^{-\mathtt{x}}.$$

If (C.8) holds with $\mathtt{g} = \infty$, then we are back in the (sub-)Gaussian case with $z_c = \infty$. In the non-Gaussian case with a finite $\mathtt{g}$, we have to accompany the moderate deviation bound with a large deviation bound $I\!\!P\bigl(\|\boldsymbol{\xi}\| > z\bigr)$ for $z \ge z_c$. This is done by combining the bound (C.10) with the standard slicing arguments.

Lemma C.2.3. *Define* $\mathtt{g}_c = \mu_c z_c$*; see* (C.12)*. It holds for* $z \ge z_c$

$$I\!\!P\bigl(\|\boldsymbol{\xi}\| > z\bigr) \le 8.4(1 - \mathtt{g}_c/z)^{-p/2}\exp\bigl(-\mathtt{g}_c z/2\bigr) \tag{C.17}$$

$$\le 8.4\exp\bigl\{-\mathtt{x}_c - \mathtt{g}_c(z - z_c)/2\bigr\}. \tag{C.18}$$

Proof. For a fixed $z \ge z_c$, consider the growing sequence $(\mathtt{y}_k)$ with $\mathtt{y}_1 = z$ and

$$\mathtt{y}_{k+1} = z + k/\mathtt{g}_c.$$

Define also $\mu_k = \mathtt{g}_c/\mathtt{y}_k$. Then the sequence (μ_k) is decreasing, in particular, $\mu_k \le \mu_1 = \mathtt{g}_c/z \le \mu_c$. Obviously

$$I\!\!P\bigl(\|\boldsymbol{\xi}\| > z\bigr) = \sum_{k=1}^{\infty} I\!\!P\bigl(\|\boldsymbol{\xi}\| > \mathtt{y}_k, \|\boldsymbol{\xi}\| \le \mathtt{y}_{k+1}\bigr).$$

Now we try to evaluate every slicing probability in this expression. We use that

$$\mu_{k+1}\mathtt{y}_k^2 = \frac{(\mathtt{g}_c z + k - 1)^2}{\mathtt{g}_c z + k} \ge \mathtt{g}_c z + k - 2.$$

Lemma C.2.2 implies $\mathtt{g} - \sqrt{\mu_c p} \ge \mu_c z_c = \mathtt{g}_c$. This yields $\mathtt{g}/\mu_k - \sqrt{p/\mu_k} \ge \mathtt{y}_k$ because

$$\mathtt{g}/\mu_k - \sqrt{p/\mu_k} - \mathtt{y}_k = \mu_k^{-1}(\mathtt{g} - \sqrt{\mu_k p} - \mathtt{g}_c) \ge \mu_k^{-1}(\mathtt{g} - \sqrt{\mu_c p} - \mathtt{g}_c) \ge 0.$$

Hence by (C.10)

$$
\begin{aligned}
\mathbb{P}\bigl(\|\boldsymbol{\xi}\| > z\bigr) &= \sum_{k=1}^{\infty} \mathbb{P}\bigl(\|\boldsymbol{\xi}\| > \mathtt{y}_k, \|\boldsymbol{\xi}\| \le \mathtt{y}_{k+1}\bigr) \\
&\le \sum_{k=1}^{\infty} \exp\Bigl(-\frac{\mu_{k+1}\mathtt{y}_k^2}{2}\Bigr) \mathbb{E} \exp\Bigl(\frac{\mu_{k+1}\|\boldsymbol{\xi}\|^2}{2}\Bigr) \mathbb{I}\Bigl(\|\boldsymbol{\xi}\| \le \frac{\mathtt{g}}{\mu_{k+1}} - \sqrt{\frac{p}{\mu_{k+1}}}\Bigr) \\
&\le \sum_{k=1}^{\infty} 2\bigl(1 - \mu_{k+1}\bigr)^{-p/2} \exp\Bigl(-\frac{\mu_{k+1}\mathtt{y}_k^2}{2}\Bigr) \\
&\le 2\bigl(1 - \mu_1\bigr)^{-p/2} \sum_{k=1}^{\infty} \exp\Bigl(-\frac{\mathtt{g}_c z + k - 2}{2}\Bigr) \\
&= 2\mathrm{e}^{1/2}(1 - \mathrm{e}^{-1/2})^{-1}(1 - \mu_1)^{-p/2} \exp\bigl(-\mathtt{g}_c z/2\bigr) \\
&\le 8.4(1 - \mathtt{g}_c/z)^{-p/2} \exp\bigl(-\mathtt{g}_c z/2\bigr)
\end{aligned}
$$

and the assertion (C.17) follows. For $z = z_c$, it holds by (C.14)

$$
\mathtt{g}_c z_c + p\log(1 - \mu_c) = \mu_c z_c^2 + p\log(1 - \mu_c) \ge 2\mathtt{x}_c
$$

and (C.17) implies $\mathbb{P}\bigl(\|\boldsymbol{\xi}\| > z_c\bigr) \le 8.4\exp(-\mathtt{x}_c)$. Now observe that the function $f(z) = \mathtt{g}_c z/2 + (p/2)\log\bigl(1 - \mathtt{g}_c/z\bigr)$ fulfills $f(z_c) = \mathtt{x}_c$ and $f'(z) \ge \mathtt{g}_c/2$ yielding $f(z) \ge \mathtt{x}_c + \mathtt{g}_c(z - \mathtt{y}_0)/2$. This implies (C.18).

Now we can conclude that for $\mathtt{x} \ge \mathtt{x}_c$, the choice

$$
z = z(\mathtt{x}) \, = \, 2\mathtt{g}_c^{-1}(\mathtt{x} - \mathtt{x}_c) + z_c
$$

implies

$$
\mathbb{P}\bigl(\|\boldsymbol{\xi}\| > z(\mathtt{x})\bigr) \, \le \, 8.4\mathrm{e}^{-\mathtt{x}}. \tag{C.19}
$$

The statement of the theorem is obtained by a simple combination of (C.14) and (C.19).

C.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 (C.8) and B is a given symmetric positive $p \times p$ matrix. Define

$$
\mathtt{p} \overset{\mathrm{def}}{=} \operatorname{tr}\bigl(B\bigr), \qquad \mathtt{v}^2 \overset{\mathrm{def}}{=} \operatorname{tr}(B^2), \qquad \lambda \overset{\mathrm{def}}{=} \lambda_{\max}\bigl(B\bigr).
$$

For ease of presentation, suppose that $0.3\mathtt{g} \ge \sqrt{\mathtt{p}}$ so that $\alpha = \sqrt{\mathtt{p}/\mathtt{g}^2} \le 0.3$. The other case only changes the constants in the inequalities. Define also

$$\begin{aligned}
\mathtt{x}_c &\stackrel{\text{def}}{=} \mathtt{g}^2/4, \\
z_c^2 &\stackrel{\text{def}}{=} \mathtt{p} + \mathtt{v}\mathtt{g} + \lambda \mathtt{g}^2/2, \\
\mathtt{g}_c &\stackrel{\text{def}}{=} \frac{\sqrt{\mathtt{p}/\lambda + \mathtt{g}\mathtt{v}/\lambda + \mathtt{g}^2/2}}{1 + \mathtt{v}/(\lambda \mathtt{g})}.
\end{aligned}$$

Theorem C.2.2. *Let* (C.8) *hold and* $0.3\mathtt{g} \geq \sqrt{\mathtt{p}/\lambda}$ *. Then for each* $\mathtt{x} > 0$

$$I\!\!P\bigl(\|B^{1/2}\boldsymbol{\xi}\| \geq z(B,\mathtt{x})\bigr) \leq 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c}\,\mathrm{1\!I}(\mathtt{x} < \mathtt{x}_c),$$

where $z(B,\mathtt{x})$ *is defined by*

$$z(B,\mathtt{x}) \stackrel{\text{def}}{=} \begin{cases} \sqrt{\mathtt{p} + 2\mathtt{v}\mathtt{x}^{1/2} + 2\lambda\mathtt{x}}, & \mathtt{x} \leq \mathtt{x}_c, \\ z_c + 2\lambda(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c, & \mathtt{x} > \mathtt{x}_c. \end{cases}$$

Similarly to the Gaussian case, the upper quantile $z(B,\mathtt{x}) = \sqrt{\mathtt{p} + 2\mathtt{v}\mathtt{x}^{1/2} + 2\lambda\mathtt{x}}$ can be upper bounded by $\sqrt{\mathtt{p}} + \sqrt{2\lambda\mathtt{x}}$:

$$z(B,\mathtt{x}) \leq \begin{cases} \sqrt{\mathtt{p}} + \sqrt{2\lambda\mathtt{x}}, & \mathtt{x} \leq \mathtt{x}_c, \\ z_c + 2\lambda(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c, & \mathtt{x} > \mathtt{x}_c. \end{cases}$$

The result simplifies in the sub-Gaussian case with $\mathtt{g} = \infty$.

Corollary C.2.2. *Let the conditions of Theorem C.2.2 hold with* $\mathtt{g} = \infty$ *. Then for each* $\mathtt{x} > 0$

$$I\!\!P\bigl(\|B^{1/2}\boldsymbol{\xi}\| \geq z(B,\mathtt{x})\bigr) \leq 2\mathrm{e}^{-\mathtt{x}},$$

where $z(B,\mathtt{x})$ *is defined by*

$$z(B,\mathtt{x}) \stackrel{\text{def}}{=} \sqrt{\mathtt{p} + 2\mathtt{v}\mathtt{x}^{1/2} + 2\lambda\mathtt{x}} \leq \sqrt{\mathtt{p}} + \sqrt{2\lambda\mathtt{x}}.$$

The main steps of the proof are similar to the proof of Theorem C.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 $B = \operatorname{diag}\bigl(a_1, \ldots, a_p\bigr)$, where $1 = a_1 \geq a_2 \geq \ldots \geq a_p > 0$. Note that $\mathtt{p} = a_1 + \ldots + a_p$ and $\mathtt{v}^2 = a_1^2 + \ldots + a_p^2$.

Lemma C.2.4. *Suppose* (C.8) *and* $\|B\|_{\mathrm{op}} = 1$. *For any* $\mu < 1$ *with* $\mathtt{g}^2/\mu \geq \mathtt{p}$*, it holds*

$$I\!\!E \exp\bigl(\mu\|B^{1/2}\boldsymbol{\xi}\|^2/2\bigr)\,\mathrm{1\!I}\bigl(\|B\boldsymbol{\xi}\| \leq \mathtt{g}/\mu - \sqrt{\mathtt{p}/\mu}\bigr) \leq 2\det(I_p - \mu B)^{-1/2}. \qquad \text{(C.20)}$$

Proof. With $c_p(B) = \bigl(2\pi\bigr)^{-p/2}\det(B^{-1/2})$

$$\begin{aligned}
&c_p(B)\int \exp\Bigl(\boldsymbol{\gamma}^{\top}\boldsymbol{\xi} - \frac{1}{2\mu}\|B^{-1/2}\boldsymbol{\gamma}\|^2\Bigr)\,\mathrm{I\!I}(\|\boldsymbol{\gamma}\| \le \mathtt{g})d\boldsymbol{\gamma} \\
&\quad = c_p(B)\exp\Bigl(\frac{\mu\|B^{1/2}\boldsymbol{\xi}\|^2}{2}\Bigr)\int \exp\Bigl(-\frac{1}{2}\bigl\|\mu^{1/2}B^{1/2}\boldsymbol{\xi} - \tau B^{-1/2}\boldsymbol{\gamma}\bigr\|^2\Bigr)\,\mathrm{I\!I}(\|\boldsymbol{\gamma}\| \le \mathtt{g})d\boldsymbol{\gamma} \\
&\quad = \mu^{p/2}\exp\Bigl(\frac{\mu\|B^{1/2}\boldsymbol{\xi}\|^2}{2}\Bigr)I\!P_{\boldsymbol{\xi}}\bigl(\|\tau B^{1/2}\boldsymbol{\varepsilon} + B^{1/2}\boldsymbol{\xi}\| \le \mathtt{g}/\mu\bigr),
\end{aligned}$$

where $\boldsymbol{\varepsilon}$ denotes a standard normal vector in $I\!R^p$ and $I\!P_{\boldsymbol{\xi}}$ means the conditional probability given $\boldsymbol{\xi}$. Moreover, for any $\boldsymbol{u} \in I\!R^p$ and $\mathtt{r} \ge \mathtt{p}^{1/2} + \|\boldsymbol{u}\|$, it holds in view of $I\!P\bigl(\|B^{1/2}\boldsymbol{\varepsilon}\|^2 > \mathtt{p}\bigr) \le 1/2$

$$I\!P\bigl(\|B^{1/2}\boldsymbol{\varepsilon} - \boldsymbol{u}\| \le \mathtt{r}\bigr) \ge I\!P\bigl(\|B^{1/2}\boldsymbol{\varepsilon}\| \le \sqrt{\mathtt{p}}\bigr) \ge 1/2.$$

This implies

$$\begin{aligned}
&\exp\Bigl(\mu\|B^{1/2}\boldsymbol{\xi}\|^2/2\Bigr)\,\mathrm{I\!I}\bigl(\|B\boldsymbol{\xi}\| \le \mathtt{g}/\mu - \sqrt{\mathtt{p}/\mu}\bigr) \\
&\quad \le 2\mu^{-p/2}c_p(B)\int \exp\Bigl(\boldsymbol{\gamma}^{\top}\boldsymbol{\xi} - \frac{1}{2\mu}\|B^{-1/2}\boldsymbol{\gamma}\|^2\Bigr)\,\mathrm{I\!I}(\|\boldsymbol{\gamma}\| \le \mathtt{g})d\boldsymbol{\gamma}.
\end{aligned}$$

Further, by (C.8)

$$\begin{aligned}
&c_p(B)I\!E\int \exp\Bigl(\boldsymbol{\gamma}^{\top}\boldsymbol{\xi} - \frac{1}{2\mu}\|B^{-1/2}\boldsymbol{\gamma}\|^2\Bigr)\,\mathrm{I\!I}(\|\boldsymbol{\gamma}\| \le \mathtt{g})d\boldsymbol{\gamma} \\
&\quad \le c_p(B)\int \exp\Bigl(\frac{\|\boldsymbol{\gamma}\|^2}{2} - \frac{1}{2\mu}\|B^{-1/2}\boldsymbol{\gamma}\|^2\Bigr)d\boldsymbol{\gamma} \\
&\quad \le \det(B^{-1/2})\det(\mu^{-1}B^{-1} - I_p)^{-1/2} = \mu^{p/2}\det(I_p - \mu B)^{-1/2}
\end{aligned}$$

and (C.20) follows.

Now we evaluate the probability $I\!P\bigl(\|B^{1/2}\boldsymbol{\xi}\| > \mathtt{y}\bigr)$ for moderate values of $\mathtt{y}$. Given $\mathtt{x} \le \mathtt{x}_c = \mathtt{g}^2/4$, define

$$\mu \;=\; \mu(\mathtt{x}) = \frac{1}{1 + 0.5\mathtt{v}\mathtt{x}^{-1/2}}\,, \tag{C.21}$$

$$\mu_c \overset{\mathrm{def}}{=} \frac{1}{1 + 0.5\mathtt{v}\,\mathtt{x}_c^{-1/2}} = \frac{1}{1 + \mathtt{v}/\mathtt{g}}\,. \tag{C.22}$$

Obviously $\mu \le \mu_c$. Now we obtain similarly to the Gaussian case in Lemma C.1.2 for $u = 2\mathtt{v}\sqrt{\mathtt{x}} + 2\mathtt{x}$

$$
\begin{aligned}
& I\!\!P\Bigl(\|B^{1/2}\boldsymbol{\xi}\|^2 > \mathtt{p}+u,\ \|\boldsymbol{\xi}\| \le \mathtt{g}/\mu - \sqrt{\mathtt{p}/\mu}\Bigr) \\
& \quad \le \exp\Bigl\{-\frac{\mu(\mathtt{p}+u)}{2}\Bigr\} I\!\!E \exp\Bigl(\frac{\mu\|\boldsymbol{\xi}\|^2}{2}\Bigr)\, \mathbb{I}\Bigl(\|\boldsymbol{\xi}\| \le \mathtt{g}/\mu - \sqrt{\mathtt{p}/\mu}\Bigr) \\
& \quad \le 2\exp\Bigl\{-\frac{1}{2}\bigl[\mu(\mathtt{p}+u) - \log\det(I_p - \mu B)\bigr]\Bigr\}
\end{aligned}
\tag{C.23}
$$

and by (C.6), it holds for μ from (C.21)

$$
\mu(\mathtt{p} + 2\mathsf{v}\sqrt{\mathtt{x}} + 2\mathtt{x}) + \log\det(I_p - \mu B) \ge 2\mathtt{x}.
$$

Now we show that the constraint $\|\boldsymbol{\xi}\| \le \mathtt{g}/\mu - \sqrt{\mathtt{p}/\mu}$ in (C.23) can be replaced by the inequality $\|\boldsymbol{\xi}\| \le z_c$. Indeed, the definition implies $\mu \le \mu_c$ for $\mathtt{x} \le \mathtt{x}_c$ and

$$
\begin{aligned}
\mathtt{p} + 2\mathsf{v}\,\sqrt{\mathtt{x}} + 2\mathtt{x} &\le \mathtt{p} + 2\mathsf{v}\sqrt{\mathtt{x}_c} + 2\mathtt{x}_c, \\
\mathtt{g}/\mu - \sqrt{\mathtt{p}/\mu} &\ge \mathtt{g}/\mu_c - \sqrt{\mathtt{p}/\mu_c}.
\end{aligned}
$$

It remains to show that

$$
\mathtt{p} + 2\mathsf{v}\sqrt{\mathtt{x}_c} + 2\mathtt{x}_c \le \bigl(\mathtt{g}/\mu_c - \sqrt{\mathtt{p}/\mu_c}\bigr)^2. \tag{C.24}
$$

Denote $\alpha^2 = \mathtt{p}/\mathtt{g}^2$. By $\mathsf{v}^2 \le \mathtt{p}$ and $\mathtt{x}_c = \mathtt{g}^2/4$, it holds $\mu_c^{-1} = 1 + 0.5\mathsf{v}\mathtt{x}_c^{-1/2} \le 1 + \alpha$ and

$$
\mathtt{g}/\mu_c - \sqrt{\mathtt{p}/\mu_c} = \mu_c^{-1}\mathtt{g}\bigl(1 - \sqrt{\mu_c\alpha^2}\bigr) \ge \mathtt{g}\,(1+\alpha)\,\bigl\{1 - \sqrt{\alpha^2/(1+\alpha)}\bigr\}.
$$

Also in a similar way

$$
\mathtt{p} + 2\mathsf{v}\sqrt{\mathtt{x}_c} + 2\mathtt{x}_c \le \mathtt{p} + \sqrt{\mathtt{p}\mathtt{g}^2} + \mathtt{g}^2/2 = \mathtt{g}^2\bigl(\alpha^2 + \alpha + 1/2\bigr).
$$

This and (C.16) prove (C.24) yielding

$$
I\!\!P\bigl(\|B^{1/2}\boldsymbol{\xi}\|^2 > \mathtt{p} + 2\mathsf{v}\sqrt{\mathtt{x}} + 2\mathtt{x}, \|\boldsymbol{\xi}\| \le z_c\bigr) \le 2\mathrm{e}^{-\mathtt{x}}.
$$

The large deviation probability $I\!\!P\bigl(\|B^{1/2}\boldsymbol{\xi}\| > \mathtt{y}\bigr)$ for $\mathtt{y} > z_c$ can be bounded as in the case $B = I_p$.

Lemma C.2.5. *Define* $\mathtt{g}_c = \mu_c z_c$ *; see* (C.22)*. It holds for* $z \ge z_c$

$$
\begin{aligned}
I\!\!P\Bigl(\|B^{1/2}\boldsymbol{\xi}\| > z\Bigr) &\le 8.4(1 - \mathtt{g}_c/z)^{-p/2}\exp\bigl(-\mathtt{g}_c z/2\bigr) \\
&\le 8.4\exp\bigl\{-\mathtt{x}_c - \mathtt{g}_c(z - z_c)/2\bigr\}.
\end{aligned}
$$

Proof. The arguments from the case $B \equiv I_p$ apply without changes.

C.3 Deviation probability for a normalized martingale

Consider a random p-vector $\boldsymbol{M}$ and a random symmetric positive $p \times p$-matrix S^2 s.t.

$$\mathbb{E} \exp\Bigl(\boldsymbol{\gamma}^\top \boldsymbol{M} - \frac{1}{2}\boldsymbol{\gamma}^\top S^2 \boldsymbol{\gamma}\Bigr) \le 1, \qquad \|\boldsymbol{\gamma}\|_\circ \le \mathtt{g}. \tag{C.25}$$

The aim is to evaluate the deviation probability $\mathbb{P}\bigl(\|S^{-1}\boldsymbol{M}\| \ge \mathfrak{z}\bigr)$. Given a positive symmetric matrix S, define $\mathtt{r}_\circ(S)$ by

$$\mathbb{P}\Bigl(\|S^{-1}\boldsymbol{\varepsilon}\|_\circ \le \mathtt{r}_\circ(S)\Bigr) \ge 1/2, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{I}_p). \tag{C.26}$$

Here are two typical examples when $\mathtt{r}_\circ(S)$ can be easily evaluated. If $\|\cdot\|_\circ$ is the usual Euclidean norm, one can take $\mathtt{r}_\circ(S) = \operatorname{tr}(S^{-2})$. If $\|\cdot\|_\circ$ is the sup-norm, and S is a diagonal matrix with $S = \operatorname{diag}(s_1 \le \ldots \le s_p)$, define $\mathtt{r}_\circ(S) = s_p^{-1}\sqrt{2\log p}$.

Theorem C.3.1. *Suppose* (C.25) *for some* $\mathtt{g} > 0$. *For any* $\mu < 1$ *and any deterministic positive matrices* $S_- \preceq S_+$, *it holds*

$$\begin{aligned} &\mathbb{P}\Bigl(\|S^{-1}\boldsymbol{M}\| \ge \mathfrak{z},\, S_- \preceq S \preceq S_+,\, \|S^{-2}\boldsymbol{M}\|_\circ \le \mu^{-1}\mathtt{g} - \mu^{-1/2}\mathtt{r}_\circ(S)\Bigr) \\ &\qquad \le \frac{2\det(S_+)}{\det(S_-)}(1-\mu)^{-p/2}\exp(-\mu\mathfrak{z}/2). \end{aligned} \tag{C.27}$$

Optimizing w.r.t. μ *yields*

$$\begin{aligned} &\mathbb{P}\Bigl(\|S^{-1}\boldsymbol{M}\|^2 \ge p + 2\sqrt{p\,\mathtt{x}} + 2\mathtt{x},\, S_- \preceq S \preceq S_+,\, \|S^{-2}\boldsymbol{M}\|_\circ \le \mu^{-1}\mathtt{g} - \mu^{-1/2}\mathtt{r}_\circ(S)\Bigr) \\ &\quad \le \frac{2\det(S_+)}{\det(S_-)}\exp(-\mathtt{x}). \end{aligned}$$

Proof. Denote $\tau = \mu^{-1/2} > 1$. Introduce random sets $\mathcal{A}_1$ and $\mathcal{A}_2$

$$\begin{aligned} \mathcal{A}_1 &\stackrel{\text{def}}{=} \bigl\{S_- \preceq S \preceq S_+\bigr\}, \\ \mathcal{A}_2 &\stackrel{\text{def}}{=} \{\|S^{-2}\boldsymbol{M}\|_\circ \le \tau^2\mathtt{g} - \tau\mathtt{r}_\circ(S)\}. \end{aligned}$$

For $\boldsymbol{\gamma} \in \mathbb{R}^p$, define $\boldsymbol{u} = \tau S\boldsymbol{\gamma} - \tau^{-1}S^{-1}\boldsymbol{M}$, so that $\boldsymbol{\gamma} = \tau^{-1}S^{-1}\boldsymbol{u} + \tau^{-2}S^{-2}\boldsymbol{M}$. It holds on the set $\mathcal{A}_2$ by (C.26) with $\boldsymbol{\xi} = S^{-1}\boldsymbol{M}$

$$\begin{aligned} &c_p \det(\tau S) \int \exp\Bigl(\boldsymbol{\gamma}^\top\boldsymbol{M} - \frac{\tau^2}{2}\boldsymbol{\gamma}^\top S^2\boldsymbol{\gamma}\Bigr)\, \mathbb{I}(\|\boldsymbol{\gamma}\|_\circ \le \mathtt{g})d\boldsymbol{\gamma} \\ &\quad = \exp\Bigl(\frac{\|\boldsymbol{\xi}\|^2}{2\tau^2}\Bigr)c_p \det(\tau S)\int \exp\Bigl(-\frac{1}{2}\|\tau^{-1}S^{-1}\boldsymbol{M} - \tau S\boldsymbol{\gamma}\|^2\Bigr)\, \mathbb{I}(\|\boldsymbol{\gamma}\|_\circ \le \mathtt{g})d\boldsymbol{\gamma} \\ &\quad \ge \exp\Bigl(\frac{\|\boldsymbol{\xi}\|^2}{2\tau^2}\Bigr)c_p \int \exp(-\|\boldsymbol{u}\|^2/2)\, \mathbb{I}\bigl\{\|S^{-1}\boldsymbol{u}\|_\circ \le \mathtt{r}_\circ(S)\bigr\}d\boldsymbol{u} \\ &\quad \ge 0.5\exp\Bigl(\frac{\|\boldsymbol{\xi}\|^2}{2\tau^2}\Bigr). \end{aligned} \tag{C.28}$$

Further, it holds on $\mathcal{A}_1$ with $\tau_0^2 \stackrel{\mathrm{def}}{=} \tau^2 - 1$

$$\exp\Bigl(\boldsymbol{\gamma}^\top \boldsymbol{M} - \frac{\tau^2}{2} \boldsymbol{\gamma}^\top S^2 \boldsymbol{\gamma}\Bigr) \le \exp\Bigl(\boldsymbol{\gamma}^\top \boldsymbol{M} - \frac{1}{2} \boldsymbol{\gamma}^\top S^2 \boldsymbol{\gamma}\Bigr) \exp\Bigl(\frac{\tau_0^2}{2} \boldsymbol{\gamma}^\top S_-^2 \boldsymbol{\gamma}\Bigr).$$

This implies by (C.28) and (C.25)

$$\mathbb{E}\biggl[\frac{\det\bigl(\tau_0 S_-\bigr)}{\det\bigl(\tau S\bigr)} \exp\Bigl(\frac{\|\boldsymbol{\xi}\|^2}{2\tau^2}\Bigr) \mathbb{1}(\mathcal{A}_1)\biggr] \le c_p \det(\tau_0 S_-) \int \exp\Bigl(-\frac{\tau_0^2}{2} \boldsymbol{\gamma}^\top S_-^2 \boldsymbol{\gamma}\Bigr) d\boldsymbol{\gamma} = 1$$

and it follows by $\tau^2/\tau_0^2 = (1-\mu)^{-1}$ and by the Markov inequality that

$$\begin{aligned}
&\mathbb{P}\Bigl(\|\boldsymbol{\xi}\|^2 > \mathfrak{z},\, \mathcal{A}_1 \cap \mathcal{A}_2\Bigr) \\
&\quad \le \mathbb{P}\biggl(\frac{\det\bigl(\tau_0 S_-\bigr)}{\det\bigl(\tau S_+\bigr)} \exp\Bigl(-\frac{\|\boldsymbol{\xi}\|^2}{2\tau^2}\Bigr) \mathbb{1}(\mathcal{A}_1) > (\tau/\tau_0)^p \, \frac{\det\bigl(S_-\bigr)}{\det\bigl(S_+\bigr)} \, \exp\Bigl(-\frac{\mathfrak{z}}{2\tau^2}\Bigr)\biggr) \\
&\quad \le 2(1-\mu)^{-p/2} \exp\Bigl(-\frac{\mu \mathfrak{z}}{2}\Bigr) \frac{\det(S_+)}{\det(S_-)}.
\end{aligned}$$

as required.

Note that the value $\mathtt{r}_\circ(S)$ from (C.26) is monotonously decreasing in S. Therefore, the inequality $\|S^{-2}\boldsymbol{M}\|_\circ \le \mu^{-1}\mathtt{g} - \tau \mathtt{r}_\circ(S)$ follows from the stricter inequality

$$\|S_-^{-2}\boldsymbol{M}\|_\circ \le \mu^{-1}\mathtt{g} - \tau \mathtt{r}_\circ(S_-).$$

If $\lambda_{\max}(S_-^{-1}S_+) \le 1 + \delta$ 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 (C.27) 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 $\mathtt{g} = \infty$, is not sharp and requires much more involved discretization arguments.

A typical application is given by the case when $\boldsymbol{M}$ is a martingale stopped at a time instant T and $S^2 = \langle \boldsymbol{M} \rangle_T$ is its predictable quadratic characteristic at T.

Optimization of the right hand-side of (C.27) w.r.t. μ yields

$$\mathbb{P}\Bigl(\|S^{-1}\boldsymbol{M}\|^2 > p + 2\sqrt{p\,\mathtt{x}} + 2\mathtt{x}\Bigr) \le \frac{2\det(S_+)}{\det(S_-)} \mathrm{e}^{-\mathtt{x}} + \mathbb{P}(\mathcal{A}_1) + \mathbb{P}(\mathcal{A}_2);$$

cf. Spokoiny and Zhilova (2013).

D

Sums of random matrices

Here we present a number of deviation bounds for a sum of random matrices.

D.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 D.1.1 (Master bound). *Assume that* $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_n$ *are independent Hermitian matrices of the same size and* $\boldsymbol{Z} = \sum_{i=1}^n \boldsymbol{S}_i$ *. Then*

$$I\!\!E \lambda_{\max}^{+}(\boldsymbol{Z}) \leq \inf_{\theta>0} \frac{1}{\theta} \log \operatorname{tr} \exp\left(\sum_{i=1}^{n} \log I\!\!E \mathrm{e}^{\theta \boldsymbol{S}_i} \right), \tag{D.1}$$

$$I\!\!P\{\lambda_{\max}^{+}(\boldsymbol{Z}) \geq z\} \leq \inf_{\theta>0} \mathrm{e}^{-\theta z} \operatorname{tr} \exp\left(\sum_{i=1}^{n} \log I\!\!E \mathrm{e}^{\theta \boldsymbol{S}_i} \right), \tag{D.2}$$

where $\lambda_{\max}^{+}(\boldsymbol{Z})$ *denotes the algebraically largest eigenvalue of* $\boldsymbol{Z}$ *.*

Proof. By the Markov inequality

$$I\!\!P\{\lambda_{\max}^{+}(\boldsymbol{Z}) \geq z\} \leq \inf_{\theta} \mathrm{e}^{-\theta z} I\!\!E \exp(\theta \lambda_{\max}^{+}(\boldsymbol{Z})).$$

Recall the spectral mapping theorem: for any function $f \colon I\!\!R \to I\!\!R$ and Hermitian matrix A eigenvalues of $f(A)$ are equal to eigenvalues of A. Thus

$$\exp(\theta \lambda_{\max}(\boldsymbol{Z})) = \exp(\lambda_{\max}^{+}(\theta \boldsymbol{Z})) = \lambda_{\max}^{+}\bigl(\exp(\theta \boldsymbol{Z})\bigr) \leq \operatorname{tr} \mathrm{e}^{\theta \boldsymbol{Z}}.$$

Therefore,

$$I\!\!P\bigl\{\lambda_{\max}^{+}(\boldsymbol{Z}) \geq z\bigr\} \leq \inf_{\theta} \mathrm{e}^{-\theta z} I\!\!E \operatorname{tr} \exp(\theta \boldsymbol{Z}), \tag{D.3}$$

and (D.2) follows.

To prove (D.1) fix θ. Using the spectral mapping theorem one can get that

$$\mathbb{E}\lambda^{+}_{\max}(\boldsymbol{Z}) = \frac{1}{\theta}\mathbb{E}\lambda^{+}_{\max}(\theta\boldsymbol{Z}) = \frac{1}{\theta}\log\mathbb{E}\exp\bigl(\lambda^{+}_{\max}(\theta\boldsymbol{Z})\bigr) = \frac{1}{\theta}\log\mathbb{E}\lambda^{+}_{\max}\bigl(\exp(\theta\boldsymbol{Z})\bigr)\,.$$

Thus we get

$$\mathbb{E}\lambda^{+}_{\max}(\boldsymbol{Z}) \le \frac{1}{\theta}\log\operatorname{tr}\mathbb{E}\exp(\theta\boldsymbol{Z}). \tag{D.4}$$

The final step in proving the master inequalities is to bound from above $\mathbb{E}\operatorname{tr}\exp\left(\sum_{i=1}^{n}\boldsymbol{S}_i\right)$. To do this we use Jensen's inequality for the convex function $\operatorname{tr}\exp(H+\log(X))$ (in matrix X), where H is deterministic Hermitian matrix. For a random Hermitian matrix X one can write

$$\mathbb{E}\operatorname{tr}\exp(H+X) = \mathbb{E}\operatorname{tr}\exp(H+\log\mathrm{e}^{X}) \le \operatorname{tr}\exp(H+\log\mathbb{E}\mathrm{e}^{X}). \tag{D.5}$$

Denote by $\mathbb{E}_i$ the conditional expectation with respect to random matrix X_i. To bound $\mathbb{E}\operatorname{tr}\exp\left(\sum_{i=1}^{n}\boldsymbol{S}_i\right)$ we use (D.5) for the sum of independent Hermitian matrices by taking the conditional expectations with respect to i-th matrix:

$$\begin{aligned}\mathbb{E}\operatorname{tr}\exp\left(\sum_{i=1}^{n}\boldsymbol{S}_i\right) &= \mathbb{E}\mathbb{E}_n\operatorname{tr}\exp\left(\sum_{i=1}^{n-1}\boldsymbol{S}_i+\boldsymbol{S}_n\right)\\ &\le \mathbb{E}\operatorname{tr}\exp\left(\sum_{i=1}^{n-1}\boldsymbol{S}_i+\log(\mathbb{E}_n\exp(\boldsymbol{S}_n))\right)\\ &\le \operatorname{tr}\exp\left(\sum_{i=1}^{n}\log\mathbb{E}\mathrm{e}^{\theta\boldsymbol{S}_i}\right).\end{aligned} \tag{D.6}$$

To complete the prove of the Master's theorem combine (D.3) and (D.4) with (D.6).

The same result applied to $-\boldsymbol{Z}$ yields the bound for the norm $\|\boldsymbol{Z}\|_{\mathrm{op}}$:

$$\begin{aligned}\mathbb{P}\{\|\boldsymbol{Z}\|_{\mathrm{op}}\ge z\} \le \inf_{\theta>0}\mathrm{e}^{-\theta z}\operatorname{tr}\exp\left(\sum_{i=1}^{n}\log\mathbb{E}\mathrm{e}^{\theta\boldsymbol{S}_i}\right)&\\ +\inf_{\theta>0}\mathrm{e}^{-\theta z}\operatorname{tr}\exp\left(\sum_{i=1}^{n}\log\mathbb{E}\mathrm{e}^{-\theta\boldsymbol{S}_i}\right)&.\end{aligned} \tag{D.7}$$

Theorem D.1.2 (Bernstein inequality for a sum of random Hermitian matrices). *Let* $\boldsymbol{Z}=\sum_{i=1}^{n}\boldsymbol{S}_i$, *where* $\boldsymbol{S}_i$, $i=1,\ldots,n$ *are independent, random, Hermitian matrices of the dimension* $d\times d$ *and*

$$\lambda^{+}_{\max}(\boldsymbol{S}_i) \le R.$$

Denote $\mathsf{v}^2 = \mathsf{v}^2(\boldsymbol{Z}) = \|\mathbb{E}(\boldsymbol{Z}^2)\|_{\mathrm{op}}$. *Then*

$$\mathbb{E}\lambda_{\max}^{+}(\boldsymbol{Z}) \le \sqrt{2\mathsf{v}^2 \log(d)} + \frac{1}{3} R \log(d), \tag{D.8}$$

$$\mathbb{P}\bigl\{\lambda_{\max}^{+}(\boldsymbol{Z}) \ge z\bigr\} \le d \exp\left(\frac{-z^2/2}{\mathsf{v}^2 + Rz/3}\right). \tag{D.9}$$

Proof. Note that

$$\mathsf{v}^2 = \left\| \sum_{i=1}^{n} \mathbb{E}\boldsymbol{S}_i^2 \right\|_{\mathrm{op}} .$$

For the sake of simplicity let $\mathsf{v}^2 = 1$. Denote

$$g(\theta) = \frac{\theta^2/2}{1 - R\theta/3}.$$

Apart the Master inequalities, we use the following lemma:

Lemma D.1.1. *Let* $\boldsymbol{Z}$ *be a random Hermitian matrix* $\mathbb{E}\boldsymbol{Z} = 0$, $\lambda_{\max}^{+}(\boldsymbol{Z}) \le R$, *then for* $0 < \theta < 3/R$ *the following inequalities hold*

$$\mathbb{E}\mathrm{e}^{\theta \boldsymbol{Z}} \le \exp\left(\frac{\theta^2/2}{1 - R\theta/3}\mathbb{E}(\boldsymbol{Z}^2)\right),$$

$$\log \mathbb{E}\mathrm{e}^{\theta \boldsymbol{Z}} \le \frac{\theta^2/2}{1 - R\theta/3}\mathbb{E}(\boldsymbol{Z}^2).$$

Proof. Decompose the exponent in the following way

$$\mathrm{e}^{\theta \boldsymbol{Z}} = \boldsymbol{I} + \theta \boldsymbol{Z} + (\mathrm{e}^{\theta \boldsymbol{Z}} - \theta \boldsymbol{Z} - \boldsymbol{I}) = \boldsymbol{I} + \theta \boldsymbol{Z} + \boldsymbol{Z} \cdot f(\boldsymbol{Z}) \cdot \boldsymbol{Z},$$

where

$$f(x) = \frac{\mathrm{e}^{\theta x} - \theta x - 1}{x^2}, \quad \text{for} \quad x \neq 0, \quad f(0) = \frac{\theta^2}{2}.$$

One can check that the function $f(x)$ is non-decreasing, thus for $x \le R$, one has $f(x) \le f(R)$. By the matrix transfer rule $f(\boldsymbol{Z}) \le f(R)\boldsymbol{I}$ and

$$\mathbb{E}\mathrm{e}^{\theta \boldsymbol{Z}} \le \boldsymbol{I} + f(R)\mathbb{E}\boldsymbol{Z}^2.$$

In order to estimate $f(R)$ use $q! \ge 2 \cdot 3^{q-2}$ to get

$$f(R) = \frac{\mathrm{e}^{\theta R} - \theta R - \boldsymbol{I}}{R^2} = \frac{1}{R^2} \sum_{q=2}^{\infty} \frac{(\theta R)^q}{q!} \le \theta^2 \sum_{q=2}^{\infty} \frac{(R\theta)^{q-2}}{3^{q-2}} = \frac{\theta^2/2}{1 - R\theta/3}.$$

To get the result of the Lemma note that $1 + a \le \mathrm{e}^a$.

To prove (D.8) and (D.9) we apply the Master inequalities and Lemma D.1.1:

$$\begin{aligned}
I\!\!E\lambda_{\max}^{+}(\boldsymbol{Z}) &\le \inf_{\theta>0} \frac{1}{\theta} \log \operatorname{tr} \exp\left(\sum_{i=1}^{n} \log I\!\!E \exp(\theta \boldsymbol{S}_i) \right) \\
&\le \inf_{0<\theta<3/R} \frac{1}{\theta} \log \operatorname{tr} \exp\left(g(\theta) \sum_{i=1}^{n} I\!\!E \boldsymbol{S}_i^2 \right) \\
&\le \inf_{0<\theta<3/R} \frac{1}{\theta} \log \operatorname{tr} \exp\left(g(\theta) I\!\!E \boldsymbol{Z}^2 \right) \\
&\le \inf_{0<\theta<3/R} \frac{1}{\theta} \log d \exp\left(g(\theta) \| I\!\!E \boldsymbol{Z}^2 \|_{\mathrm{op}} \right) \\
&\le \inf_{0<\theta<3/R} \left\{ \frac{\log(d)}{\theta} + \frac{\theta/2}{1 - R\theta/3} \right\}.
\end{aligned}$$

Minimizing the right hand side in θ one can get (D.8).

The second inequality can be obtained in the same manner:

$$\begin{aligned}
I\!\!P\{\lambda_{\max}^{+}(\boldsymbol{Z}) \ge z\} &\le \inf_{\theta>0} \mathrm{e}^{-\theta z} \operatorname{tr} \exp\left(\sum_{i=1}^{n} \log I\!\!E \exp(\theta \boldsymbol{S}_i) \right) \\
&\le \inf_{0<\theta<3/R} \mathrm{e}^{-\theta z} \operatorname{tr} \exp\left(g(\theta) I\!\!E \boldsymbol{Z}^2 \right) \\
&\le \inf_{0<\theta<3/R} \mathrm{e}^{-\theta z} d \exp\left(g(\theta) \| I\!\!E \boldsymbol{Z}^2 \|_{\mathrm{op}} \right) \\
&\le \inf_{0<\theta<3/R} \mathrm{e}^{-\theta z} d \exp\left(g(\theta) \right).
\end{aligned}$$

Here instead of minimizing the right hand side in θ we have used $\theta = z/(1 + Rz/3)$.

Theorem D.1.3 (Bernstein inequality for a sum of random Hermitian matrices). *Let* $\boldsymbol{Z} = \sum_{i=1}^{n} \boldsymbol{S}_i$, *where* $\boldsymbol{S}_i$, $i = 1, \ldots, n$ *are independently distributed random matrices of the size* $d_1 \times d_2$ *and*

$$\|\boldsymbol{S}_i\|_{\mathrm{op}} \le R.$$

Denote $\mathsf{v}^2 = \mathsf{v}^2(\boldsymbol{Z}) = \max\left\{ \| I\!\!E(\boldsymbol{Z}^* \boldsymbol{Z}) \|_{\mathrm{op}}, \| I\!\!E(\boldsymbol{Z}\boldsymbol{Z}^*) \|_{\mathrm{op}} \right\}$. *Then*

$$\begin{aligned}
I\!\!E \|\boldsymbol{Z}\|_{\mathrm{op}} &\le \sqrt{2\mathsf{v}^2 \log(d_1 + d_2)} + \frac{1}{3} R \log(d), \\
I\!\!P\{\|\boldsymbol{Z}\|_{\mathrm{op}} \ge z\} &\le (d_1 + d_2) \exp\left(\frac{-z^2/2}{\mathsf{v}^2 + Rz/3} \right).
\end{aligned}$$

Proof. Use the following hint: define the matrix

$$H(\boldsymbol{Z}) = \begin{pmatrix} 0 & \boldsymbol{Z} \\ \boldsymbol{Z}^* & 0 \end{pmatrix}.$$

It can be easily seen that $\mathsf{v}^2 = \|H(\boldsymbol{Z})^2\|_{\mathrm{op}}$, and $\|\boldsymbol{Z}\|_{\mathrm{op}} = \lambda^{+}_{\max}\bigl(H(\boldsymbol{Z})\bigr)$, thus applying Proposition D.1.2 to $H(\boldsymbol{Z})$ the statements (D.8) and (D.9) are straightforward.

The next result provides a deviation bound for a matrix-valued quadratic forms.

Proposition D.1.1 (Deviation bound for matrix quadratic forms). *Let a $p \times n$ matrix $\mathcal{U}$ with columns $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n$ be such that*

$$\mathcal{U}\mathcal{U}^{\top} \le \boldsymbol{I}_p \,, \qquad \|\boldsymbol{\omega}_i\| \le \delta_n \tag{D.10}$$

for a fixed constant δ_n. For a random vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^{\top}$ with independent standard Gaussian components, define

$$\boldsymbol{Z} \overset{\mathrm{def}}{=} \mathcal{U}\,\mathrm{diag}\bigl\{\boldsymbol{\gamma} \cdot \boldsymbol{\gamma} - 1\bigr\}\mathcal{U}^{\top} = \sum_{i=1}^{n} (\gamma_i^2 - 1)\boldsymbol{\omega}_i\boldsymbol{\omega}_i^{\top}.$$

Then

$$I\!\!P\Bigl(\|\boldsymbol{Z}\|_{\mathrm{op}} \ge 2\delta_n\sqrt{y + \log(p)} + 2\delta_n^2(y + \log p)\Bigr) \le 2\mathrm{e}^{-y}. \tag{D.11}$$

Proof. From the Master bound (D.7)

$$\begin{aligned} I\!\!P\bigl(\|\boldsymbol{Z}\|_{\mathrm{op}} \ge z\bigr) \le {} & \inf_{\theta>0} \mathrm{e}^{-\theta z}\,\mathrm{tr}\exp\left(\sum_{i=1}^{n} \log I\!\!E \exp\bigl\{\theta(\gamma_i^2 - 1)\boldsymbol{\omega}_i\,\boldsymbol{\omega}_i^{\top}\bigr\}\right) \\ & + \inf_{\theta>0} \mathrm{e}^{-\theta z}\,\mathrm{tr}\exp\left(\sum_{i=1}^{n} \log I\!\!E \exp\bigl\{\theta(-\gamma_i^2 + 1)\boldsymbol{\omega}_i\,\boldsymbol{\omega}_i^{\top}\bigr\}\right). \end{aligned} \tag{D.12}$$

Now we use the following general fact:

Lemma D.1.2. *If χ is a random variable and Π is a projector in $I\!\!R^p$, then*

$$\log I\!\!E \exp(\chi\Pi) = \log\bigl(I\!\!E \mathrm{e}^{\chi}\bigr)\Pi. \tag{D.13}$$

Proof. The result (D.13) can be easily obtained by applying twice the spectral mapping theorem.

This result yields, in particular, for any unit vector $\boldsymbol{\omega} \in I\!\!R^p$

$$\log I\!\!E \exp\bigl(\chi\boldsymbol{\omega}\boldsymbol{\omega}^{\top}\bigr) = \log\bigl(I\!\!E \mathrm{e}^{\chi}\bigr)\boldsymbol{\omega}\boldsymbol{\omega}^{\top}.$$

Moreover, for any vector $\boldsymbol{\omega} \in I\!\!R^p$, the normalized product $\boldsymbol{\omega}\boldsymbol{\omega}^{\top}/\|\boldsymbol{\omega}\|^2$ is a rank-one projector, and hence,

$$\log I\!\!E \exp\bigl(\chi\boldsymbol{\omega}\boldsymbol{\omega}^{\top}\bigr) = \log\bigl(I\!\!E \mathrm{e}^{\chi\|\boldsymbol{\omega}\|^2}\bigr)\frac{\boldsymbol{\omega}\boldsymbol{\omega}^{\top}}{\|\boldsymbol{\omega}\|^2}.$$

With $\boldsymbol{U}_i \stackrel{\mathrm{def}}{=} \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top / \|\boldsymbol{\omega}_i\|^2$ and $\chi_i = \theta(\gamma_i^2 - 1)$, we derive

$$\begin{aligned}\log \mathbb{E} \exp\bigl\{\theta(\gamma_i^2 - 1)\boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top\bigr\} &= \log \mathbb{E} \exp\bigl\{\theta(\gamma_i^2 - 1)\|\boldsymbol{\omega}_i\|^2\bigr\}\boldsymbol{U}_i \\ &= \log\left(\frac{\exp\bigl(-\|\boldsymbol{\omega}_i\|^2\theta\bigr)}{\sqrt{1 - 2\|\boldsymbol{\omega}_i\|^2\theta}}\right)\boldsymbol{U}_i \\ &= \left\{-\|\boldsymbol{\omega}_i\|^2\theta - \frac{1}{2}\log(1 - 2\theta\|\boldsymbol{\omega}_i\|^2)\right\}\boldsymbol{U}_i\end{aligned}$$

and

$$\begin{aligned}\log \mathbb{E} \exp\bigl\{\theta(-\gamma_i^2 + 1)\boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top\bigr\} &= \log \mathbb{E} \exp\bigl(\theta(-\gamma_i^2 + 1)\|\boldsymbol{\omega}_i\|^2\bigr)\boldsymbol{U}_i \\ &\le -\|\boldsymbol{\omega}_i\|^2\theta\boldsymbol{U}_i \\ &\le \bigl\{-\|\boldsymbol{\omega}_i\|^2\theta - \frac{1}{2}\log(1 - 2\theta\|\boldsymbol{\omega}_i\|^2)\bigr\}\boldsymbol{U}_i.\end{aligned}$$

Then it follows by (D.12)

$$\begin{aligned}&\mathbb{P}\bigl(\|\boldsymbol{Z}\|_{\mathrm{op}} \ge z\bigr) \\ &\quad\le 2\inf_{\theta>0} \mathrm{e}^{-\theta z}\,\mathrm{tr}\exp\left\{\sum_{i=1}^{n}\frac{\boldsymbol{\omega}_i\boldsymbol{\omega}_i^\top}{\|\boldsymbol{\omega}_i\|^2}\left\{-\|\boldsymbol{\omega}_i\|^2\theta - \frac{1}{2}\log(1 - 2\theta\|\boldsymbol{\omega}_i\|^2)\right\}\right\}.\end{aligned} \tag{D.14}$$

Denote $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top$, where

$$\eta_i = -\theta - \frac{\log(1 - 2\|\boldsymbol{\omega}_i\|^2\theta)}{2\|\boldsymbol{\omega}_i\|^2}.$$

The use of (C.5) and (D.10) yields for $\theta < (2\delta_n^2)^{-1}$

$$\begin{aligned}\eta_i &= \frac{1}{2\|\boldsymbol{\omega}_i\|^2}\bigl\{2\theta\|\boldsymbol{\omega}_i\|^2 - \log(1 - 2\theta\|\boldsymbol{\omega}_i\|^2)\bigr\} \\ &\le \frac{\bigl(2\theta\|\boldsymbol{\omega}_i\|^2\bigr)^2}{4\|\boldsymbol{\omega}_i\|^2(1 - 2\theta\delta_n^2)} \le \frac{\theta^2\delta_n^2}{(1 - 2\theta\delta_n^2)}.\end{aligned}$$

Then by (D.14) and $\mathcal{U}\mathcal{U}^\top = \boldsymbol{I}_p$ using $\mu = 2\theta\delta_n^2$

$$\begin{aligned}\mathbb{P}\bigl(\|\boldsymbol{Z}\|_{\mathrm{op}} \ge z\bigr) &\le 2\inf_{\theta>0}\mathrm{e}^{-\theta z}\,\mathrm{tr}\exp\bigl\{\mathcal{U}\,\mathrm{diag}(\boldsymbol{\eta})\mathcal{U}^\top\bigr\} \le 2\inf_{\theta>0}\mathrm{e}^{-\theta z}\,\mathrm{tr}\exp\bigl\{\|\boldsymbol{\eta}\|_\infty\boldsymbol{I}_p\bigr\} \\ &\le 2p\inf_{\theta>0}\exp\left\{-\theta z + \frac{\theta^2\delta_n^2}{1 - 2\theta\delta_n^2}\right\} = 2p\inf_{\mu>0}\exp\left\{-\frac{\mu z}{2\delta_n^2} + \frac{\mu^2\delta_n^{-2}}{1 - \mu}\right\}.\end{aligned}$$

Lemma C.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

$$\inf_{\mu>0}\exp\left\{-\frac{\mu z}{2\delta_n^2} + \frac{\mu^2\delta_n^{-2}}{1 - \mu}\right\} = \inf_{\mu>0}\left\{-\mu\bigl(\delta_n^{-1}\sqrt{y_p} + y_p\bigr) + \frac{\mu^2\delta_n^{-2}}{4(1 - \mu)}\right\} \le -y_p\,.$$

Therefore,

$$\mathbb{P}\Bigl(\| \boldsymbol{Z} \|_{\mathrm{op}} \ge 2\delta_n \sqrt{y + \log p} + 2\delta_n^2 (y + \log p) \Bigr) \le 2p\, \mathrm{e}^{-y - \log p} = 2\mathrm{e}^{-y}$$

as required.

Proposition D.1.2 (Deviation bound for matrix Gaussian sums). *Let a $p \times n$ matrix $\mathcal{U}$ with columns $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n$ satisfy (D.10). Let γ_i be independent standard Gaussian, $i = 1, \ldots, n$. For any deterministic vector $\boldsymbol{B} = (b_1, \ldots, b_n)^\top \in \mathbb{R}^n$, consider the matrix $\boldsymbol{Z}_1$ with*

$$\boldsymbol{Z}_1 \stackrel{\mathrm{def}}{=} \sum_{i=1}^n \gamma_i \, b_i \, \boldsymbol{\omega}_i \, \boldsymbol{\omega}_i^\top .$$

It holds

$$\mathbb{P}\Bigl(\| \boldsymbol{Z}_1 \|_{\mathrm{op}} \ge \delta_n^2 \| \boldsymbol{B} \| \sqrt{2y} \Bigr) \le 2\mathrm{e}^{-y}$$

$$\mathbb{P}\Bigl(\| \boldsymbol{Z}_1 \|_{\mathrm{op}} \ge \delta_n \| \boldsymbol{B} \|_\infty \sqrt{2(y + \log p)} \Bigr) \le 2\mathrm{e}^{-y}.$$

Proof. As γ_i are i.i.d. standard normal and $\mathbb{E}\mathrm{e}^{a\gamma_i} = \mathrm{e}^{a^2/2}$ for $|a| < 1/2$, it follows from the Master inequality and Lemma D.1.2

$$\begin{aligned} \mathbb{P}\bigl(\| \boldsymbol{Z}_1 \|_{\mathrm{op}} \ge z \bigr) &\le 2 \inf_{\theta > 0} \mathrm{e}^{-\theta z} \operatorname{tr} \exp\Bigl\{ \sum_{i=1}^n \log \mathbb{E} \exp(\theta \gamma_i \, b_i \, \boldsymbol{\omega}_i \, \boldsymbol{\omega}_i^\top) \Bigr\} \\ &\le 2 \inf_{\theta > 0} \mathrm{e}^{-\theta z} \operatorname{tr} \exp\Bigl\{ \sum_{i=1}^n \frac{\theta^2 b_i^2 \| \boldsymbol{\omega}_i \|^4}{2} \, \frac{\boldsymbol{\omega}_i \boldsymbol{\omega}_i^\top}{\| \boldsymbol{\omega}_i \|^2} \Bigr\}. \end{aligned}$$

Moreover, as $\| \boldsymbol{\omega}_i \| \le \delta_n$ and $\boldsymbol{U}_i = \boldsymbol{\omega}_i \, \boldsymbol{\omega}_i^\top / \| \boldsymbol{\omega}_i \|^2$ is a rank-one projector with $\operatorname{tr} \boldsymbol{U}_i = 1$, it holds

$$\operatorname{tr} \exp\Bigl\{ \frac{\theta^2}{2} \sum_{i=1}^n b_i^2 \| \boldsymbol{\omega}_i \|^4 \boldsymbol{U}_i \Bigr\} \le \exp \operatorname{tr} \Bigl(\frac{\theta^2 \delta_n^4}{2} \sum_{i=1}^n b_i^2 \boldsymbol{U}_i \Bigr) = \exp \frac{\theta^2 \delta_n^4 \| \boldsymbol{B} \|^2}{2} .$$

This implies for $z = \delta_n^2 \| \boldsymbol{B} \| \sqrt{2y}$

$$\mathbb{P}\bigl(\| \boldsymbol{Z}_1 \|_{\mathrm{op}} \ge z \bigr) \le 2 \inf_{\theta > 0} \exp\Bigl(-\theta z + \frac{1}{2} \theta^2 \delta_n^4 \| \boldsymbol{B} \|^2 \Bigr) = 2\mathrm{e}^{-y}$$

and the assertion follows. Alternatively, the definition of $\boldsymbol{U}_i$ and (D.10) imply

$$\frac{\theta^2}{2} \sum_{i=1}^n b_i^2 \| \boldsymbol{\omega}_i \|^4 \boldsymbol{U}_i \le \frac{\theta^2 \| \boldsymbol{B} \|_\infty^2 \delta_n^2}{2} \sum_{i=1}^n \boldsymbol{\omega}_i \, \boldsymbol{\omega}_i^\top \le \frac{\theta^2 \| \boldsymbol{B} \|_\infty^2 \delta_n^2}{2} \, \boldsymbol{I}_p ,$$

so that

$$\operatorname{tr}\exp\left(\frac{\theta^2}{2}\sum_{i=1}^{n} b_i^2 \|\boldsymbol{\omega}_i\|^4 \boldsymbol{U}_i\right) \le p\exp\left(\frac{\theta^2\|\boldsymbol{B}\|_\infty^2\delta_n^2}{2}\right)$$

This implies for $z = \delta_n\|\boldsymbol{B}\|_\infty\sqrt{2(y+\log p)}$ and $\theta(z) = \bigl(\delta_n\|\boldsymbol{B}\|_\infty\bigr)^{-1}\sqrt{2(y+\log p)}$

$$\begin{aligned} I\!P\bigl(\|\boldsymbol{Z}_1\|_{\mathrm{op}} \ge z\bigr) &\le 2\inf_{\theta>0}\exp\biggl(-\theta z + \frac{\theta^2\|\boldsymbol{B}\|_\infty^2\,\delta_n^2}{2} + \log p\biggr) \\ &= 2\exp\biggl(-\theta(z)z + \frac{\theta^2(z)\|\boldsymbol{B}\|_\infty^2\,\delta_n^2}{2} + \log p\biggr) = 2\mathrm{e}^{-y} \end{aligned}$$

D.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 $I\!R^n$. Define

$$\boldsymbol{\xi} = \Sigma^{-1/2}\bigl(\boldsymbol{\varepsilon} - \Pi\boldsymbol{\varepsilon}\bigr) = \boldsymbol{\gamma} - \Upsilon\boldsymbol{\gamma} \tag{D.15}$$

where $\boldsymbol{\gamma} = \Sigma^{-1/2}\boldsymbol{\varepsilon}$ is a standard Gaussian vector, and

$$\Upsilon \overset{\mathrm{def}}{=} \Sigma^{-1/2}\Pi\Sigma^{1/2}.$$

For a deterministic bias vector $\boldsymbol{B} = (b_1,\ldots,b_n)^\top$ and for a $p\times n$ matrix $\mathcal{U}$ satisfying (D.10), we aim at bounding the Frobenius and operator norms of the matrix $\mathcal{B}$ with

$$\begin{aligned} \mathcal{B} &\overset{\mathrm{def}}{=} \mathcal{U}\Bigl[\operatorname{diag}\bigl\{(\boldsymbol{\xi}+\boldsymbol{B})\cdot(\boldsymbol{\xi}+\boldsymbol{B})\bigr\} - \boldsymbol{I}_n\Bigr]\mathcal{U}^\top \\ &= \sum_{i=1}^{n}\bigl\{(\xi_i+b_i)^2 - 1\bigr\}\boldsymbol{\omega}_i\,\boldsymbol{\omega}_i^\top . \end{aligned} \tag{D.16}$$

Proposition D.2.1. *Suppose that a vector* $\boldsymbol{\xi}$ *can be written in the form* (D.15) *for a standard Gaussian vector* $\boldsymbol{\gamma}$, *and let the rows* $\Upsilon_i^\top$ *of* $\Upsilon \overset{\mathrm{def}}{=} \Sigma^{-1/2}\Pi\Sigma^{1/2}$ *satisfy* $\|\Upsilon_i\| \le \delta$. *Further, let the matrix* $\mathcal{U}$ *fulfill* (D.10). *Then on a random set* $\Omega(y)$ *with* $I\!P\bigl(\Omega(y)\bigr) \ge 1 - 6\mathrm{e}^{-y}$, *it holds for* $\mathcal{B}$ *from* (D.16)

$$\begin{aligned} \|\mathcal{B}\|_{\mathrm{op}} &\le \Delta(y), \\ \|\mathcal{B}\|_{\mathrm{Fr}} &\le \Delta_{\mathrm{Fr}}(y) = \sqrt{p}\,\Delta(y), \end{aligned}$$

where

$$\begin{aligned} \Delta(y) \overset{\mathrm{def}}{=}\ & \|\boldsymbol{B}\|_\infty^2 + \delta\,\|\boldsymbol{B}\|_\infty\bigl(\sqrt{2y+2\log n} + \sqrt{2y+2\log p}\bigr) \\ & + 2\delta_n\sqrt{y+\log p} + 2\delta_n^2(y+\log p) + (2\delta+\delta^2)(y+\log n)\,. \end{aligned}$$

Proof. The use of $\boldsymbol{\xi} = \boldsymbol{\gamma} - \Upsilon\boldsymbol{\gamma}$ allows to decompose

$$\begin{aligned}
\mathcal{B} &= \mathcal{U}\operatorname{diag}\bigl\{\boldsymbol{B}\cdot\boldsymbol{B}\bigr\}\mathcal{U}^{\top} && \stackrel{\text{def}}{=} \mathcal{B}_1 \\
&\quad + 2\mathcal{U}\operatorname{diag}\bigl\{\boldsymbol{\gamma}\cdot\boldsymbol{B}\bigr\}\mathcal{U}^{\top} && \stackrel{\text{def}}{=} \mathcal{B}_2 \\
&\quad + 2\mathcal{U}\operatorname{diag}\bigl\{\Upsilon\boldsymbol{\gamma}\cdot\boldsymbol{B}\bigr\}\mathcal{U}^{\top} && \stackrel{\text{def}}{=} \mathcal{B}_3 \\
&\quad + \mathcal{U}\operatorname{diag}\bigl\{\boldsymbol{\xi}\cdot\boldsymbol{\xi} - \boldsymbol{\gamma}\cdot\boldsymbol{\gamma}\bigr\}\mathcal{U}^{\top} && \stackrel{\text{def}}{=} \mathcal{B}_4 \\
&\quad + \mathcal{U}\bigl\{\operatorname{diag}(\boldsymbol{\gamma}\cdot\boldsymbol{\gamma}) - \boldsymbol{I}_n\bigr\}\mathcal{U}^{\top} && \stackrel{\text{def}}{=} \mathcal{B}_5
\end{aligned}$$

Obviously

$$\|\mathcal{B}\|_{\mathrm{op}} \;\le \|\mathcal{B}_1\|_{\mathrm{op}} + \|\mathcal{B}_2\|_{\mathrm{op}} + \|\mathcal{B}_3\|_{\mathrm{op}} + \|\mathcal{B}_4\|_{\mathrm{op}} + \|\mathcal{B}_5\|_{\mathrm{op}} .$$

It is well known that the sup-norm of a standard Gaussian vector $\boldsymbol{\gamma}$ in $I\!R^n$ can be bounded in probability

$$I\!P\Bigl(\|\boldsymbol{\gamma}\|_{\infty} \ge \sqrt{2y + 2\log n}\Bigr) \le \mathrm{e}^{-y}$$

with $y_n = y + \log(n)$. Further, if each row $\Upsilon_i^{\top}$ of Υ satisfies $\|\Upsilon_i\| \le \delta$, then the scalar product $\Upsilon_i^{\top}\boldsymbol{\gamma}$ is a normal zero mean r.v. with the variance

$$\operatorname{Var}\bigl(\Upsilon_i^{\top}\boldsymbol{\gamma}\bigr) \,=\, \|\Upsilon_i\|^2 \le \delta^2$$

and

$$I\!P\bigl(|\Upsilon_i^{\top}\boldsymbol{\gamma}| > \delta z_1(y)\bigr) \;\le \mathrm{e}^{-y}$$

with $z_1(y) \le \sqrt{2y}$ yielding

$$I\!P\Bigl(\|\Upsilon\boldsymbol{\gamma}\|_{\infty} > \delta\sqrt{2y + 2\log n}\Bigr) \;\le \mathrm{e}^{-y}.$$

Due to this bounds, there is a random set $\Omega_{\infty}(y)$ with $I\!P\bigl(\Omega_{\infty}(y)\bigr) \ge 1 - 2\mathrm{e}^{-y}$ such that it holds on $\Omega_{\infty}(y)$

$$\|\Upsilon\boldsymbol{\gamma}\|_{\infty} \le \delta\sqrt{2y + 2\log n}\,, \qquad \|\boldsymbol{\gamma}\|_{\infty} \le \sqrt{2y + 2\log n}\,. \tag{D.17}$$

Bounds for $\mathcal{B}_1$

In view of $\mathcal{U}\mathcal{U}^{\top} \le \boldsymbol{I}_p$, the bias term $\mathcal{B}_1$ can be estimated as follows:

$$\bigl\|\mathcal{B}_1\bigr\|_{\mathrm{op}} = \Bigl\|\sum_{i=1}^{n} \boldsymbol{\omega}_i\,\boldsymbol{\omega}_i^{\top} b_i^2\Bigr\|_{\mathrm{op}} \le \|\boldsymbol{B}\|_{\infty}^2 \;\bigl\|\mathcal{U}\mathcal{U}^{\top}\bigr\|_{\mathrm{op}} \le \|\boldsymbol{B}\|_{\infty}^2 .$$

Bounds for $\mathcal{B}_2$

Proposition D.1.2 provides a bound for a random matrix

$$\mathcal{B}_2 = 2\mathcal{U}\,\mathrm{diag}\{\boldsymbol{\gamma}\cdot\boldsymbol{B}\}\mathcal{U}^\top = 2\sum_{i=1}^{n}\gamma_i\,b_i\,\boldsymbol{\omega}_i\,\boldsymbol{\omega}_i^\top$$

in the operator norm: on a set $\Omega_2(y)$ with $I\!P\bigl(\Omega_2(y)\bigr) \geq 1-2\mathrm{e}^{-y}$

$$\|\mathcal{B}_2\|_{\mathrm{op}} \leq \delta_n\|\boldsymbol{B}\|_\infty\sqrt{2y+2\log p}\,.$$

Bounds for $\mathcal{B}_3$

We use that $\|\Upsilon\boldsymbol{\gamma}\|_\infty \leq \delta\sqrt{2y+2\log n}$ on a set $\Omega_\infty(\mathtt{x})$. Then similarly to $\mathcal{B}_1$

$$\bigl\|\mathcal{B}_3\bigr\|_{\mathrm{op}} = \left\|\sum_{i=1}^{n}\boldsymbol{\omega}_i\,\boldsymbol{\omega}_i^\top b_i\,(\Upsilon\boldsymbol{\gamma})_i\right\|_{\mathrm{op}} \leq \|\boldsymbol{B}\|_\infty\;\|\Upsilon\boldsymbol{\gamma}\|_\infty \leq \delta\,\|\boldsymbol{B}\|_\infty\sqrt{2y+2\log n}\,.$$

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

$$\|\mathcal{B}_4\|_{\mathrm{op}} \leq \bigl\|\mathcal{U}\,\mathrm{diag}\bigl\{(\Upsilon\boldsymbol{\gamma})\cdot(\Upsilon\boldsymbol{\gamma})\bigr\}\mathcal{U}^\top\bigr\|_{\mathrm{op}} + 2\bigl\|\mathcal{U}\,\mathrm{diag}\bigl\{(\Upsilon\boldsymbol{\gamma})\cdot\boldsymbol{\gamma}\bigr\}\mathcal{U}^\top\bigr\|_{\mathrm{op}}\,.$$

The condition $\mathcal{U}\mathcal{U}^\top \leq \boldsymbol{I}_p$ helps to bound

$$\bigl\|\mathcal{U}\,\mathrm{diag}\bigl\{(\Upsilon\boldsymbol{\gamma})\cdot(\Upsilon\boldsymbol{\gamma})\bigr\}\mathcal{U}^\top\bigr\|_{\mathrm{op}} \leq \|\Upsilon\boldsymbol{\gamma}\|_\infty^2\;\|\mathcal{U}\mathcal{U}^\top\|_{\mathrm{op}} \leq \|\Upsilon\boldsymbol{\gamma}\|_\infty^2.$$

Similarly

$$\bigl\|\mathcal{U}\,\mathrm{diag}\bigl\{(\Upsilon\boldsymbol{\gamma})\cdot\boldsymbol{\gamma}\bigr\}\mathcal{U}^\top\bigr\|_{\mathrm{op}} \leq \|\Upsilon\boldsymbol{\gamma}\|_\infty\;\|\boldsymbol{\gamma}\|_\infty\,\|\mathcal{U}\mathcal{U}^\top\|_{\mathrm{op}} \leq \|\Upsilon\boldsymbol{\gamma}\|_\infty\;\|\boldsymbol{\gamma}\|_\infty.$$

By (D.17), after restricting to the set $\Omega_\infty(\mathtt{x})$, this yields the bound

$$\|\mathcal{B}_4\|_{\mathrm{op}} \leq (2\delta+\delta^2)(y+\log n)\,.$$

Bounds for $\mathcal{B}_5$

The matrix $\mathcal{B}_5$ can be bounded by a version of matrix Bernstein inequality (D.11) in Proposition D.1.1: on a set $\Omega_5(y)$ with $I\!P\bigl(\Omega_5(y)\bigr) \geq 1-2\mathrm{e}^{-y}$

$$\|\mathcal{B}_5\|_{\mathrm{op}} \leq 2\delta_n\sqrt{y+\log p} + 2\delta_n^2(y+\log p).$$

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 $\|A\|_{\mathrm{Fr}} \le \sqrt{p}\|A\|_{\mathrm{op}}$ for any $p \times p$ matrix A.

D.3 Empirical covariance matrix

Let $\boldsymbol{\varepsilon}_i$ be independent centered random vectors in $I\!\!R^p$. Consider the empirical covariance matrix

$$\widehat{S} \overset{\mathrm{def}}{=} \sum_i \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^\top .$$

The interesting question is how well this matrix stabilizes around its expectation

$$S \overset{\mathrm{def}}{=} I\!\!E\widehat{S} = \sum_{i=1}^n \mathrm{Var}(\boldsymbol{\varepsilon}_i) = \mathrm{Var}\biggl(\sum_{i=1}^n \boldsymbol{\varepsilon}_i\biggr). \tag{D.18}$$

We implicitly assume that S is large and hence each $S^{-1/2}\boldsymbol{\varepsilon}_i$ is small. Below we focus on the relative difference

$$\boldsymbol{Z} \overset{\mathrm{def}}{=} S^{-1/2}(\widehat{S} - S)S^{-1/2} = S^{-1/2}\,\widehat{S}\,S^{-1/2} - \boldsymbol{I}_p .$$

This allows to reduce the study to the case $S \equiv \boldsymbol{I}_p$ considered below. Note that each vector $\boldsymbol{\varepsilon}_i$ is replaced by $S^{-1/2}\boldsymbol{\varepsilon}_i$. For bounding $\|\boldsymbol{Z}\|_{\mathrm{op}}$, one can apply the Bernstein inequality provided that each vector $\boldsymbol{\varepsilon}_i$ is bounded by a fixed small constant δ:

$$\|\boldsymbol{\varepsilon}\|_\infty = \max_{i \le n} \|\boldsymbol{\varepsilon}_i\| \le \delta.$$

Then $\boldsymbol{S}_i = \boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\top - I\!\!E\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\top$ obviously fulfills

$$\|\boldsymbol{S}_i\|_{\mathrm{op}} = \|\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\top - I\!\!E\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\top\|_{\mathrm{op}} \le \|\boldsymbol{\varepsilon}_i\|^2 \le \delta^2.$$

Define

$$\mathtt{b}^2 = \biggl\|\sum_{i=1}^n I\!\!E\boldsymbol{S}_i^2\biggr\|_{\mathrm{op}} = \biggl\|\sum_{i=1}^n \Bigl\{ I\!\!E\bigl(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\top\bigr)^2 - \bigl(I\!\!E\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\top\bigr)^2 \Bigr\}\biggr\|_{\mathrm{op}} . \tag{D.19}$$

A rough bound on $\mathtt{b}$ can be obtained by using again $\|\boldsymbol{\varepsilon}_i\| \le \delta$ and $S = \boldsymbol{I}_p$:

$$\mathtt{b}^2 \le \delta^2 \biggl\|\sum_{i=1}^n I\!\!E\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\top\biggr\|_{\mathrm{op}} = \delta^2 \tag{D.20}$$

Now the result of Proposition D.1.2 implies for $\boldsymbol{Z} = \widehat{S} - \boldsymbol{I}_p$ with $\boldsymbol{S}_i = \boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\top - I\!\!E\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\top$

$$I\!\!P\{\|\widehat{S} - I_p\|_{\mathrm{op}} \geq \mathtt{b}\, z\} \leq 2p \exp\left(\frac{-z^2/2}{1+\delta^2 \mathtt{b}^{-1} z/3}\right).$$

The bound is particularly informative if the ratio $\delta^2/\mathtt{b}$ is small. The use of the rough upper bound (D.20) yields for each $z > 0$

$$I\!\!P\{\|\widehat{S} - I_p\|_{\mathrm{op}} \geq \delta\, z\} \leq 2p \exp\left(\frac{-z^2/2}{1+z\delta/3}\right).$$

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 D.3.1. *Let random vectors* $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ *in* $I\!\!R^p$ *be independent zero mean,* S *is given by* (D.18)*, and* $\widetilde{\boldsymbol{\varepsilon}}_i \overset{\mathrm{def}}{=} S^{-1/2}\boldsymbol{\varepsilon}_i$ *fulfill*

$$\|\widetilde{\boldsymbol{\varepsilon}}_i\| = \|S^{-1/2}\boldsymbol{\varepsilon}_i\| \leq \delta \ \textit{a.s.}\ , i = 1, \ldots, n,$$

for a constant $\delta < \infty$*. Then with* $\mathtt{b}^2$ *defined by*

$$\mathtt{b}^2 \overset{\mathrm{def}}{=} \|I\!\!E\bigl(S^{-1/2}\widehat{S}S^{-1/2}\bigr)^2\|_{\mathrm{op}} = \left\|\sum_{i=1}^{n}\Bigl\{I\!\!E\bigl(\widetilde{\boldsymbol{\varepsilon}}_i\widetilde{\boldsymbol{\varepsilon}}_i^{\top}\bigr)^2 - \bigl(I\!\!E\widetilde{\boldsymbol{\varepsilon}}_i\widetilde{\boldsymbol{\varepsilon}}_i^{\top}\bigr)^2\Bigr\}\right\|_{\mathrm{op}},$$

it holds $\mathtt{b}^2 \leq \delta^2$ *and for any* $z \geq 0$

$$I\!\!P\Bigl\{\|S^{-1/2}\,\widehat{S}\,S^{-1/2} - I_p\|_{\mathrm{op}} \geq \mathtt{b}\, z\Bigr\} \leq 2p \exp\left(\frac{-z^2/2}{1+\delta^2 \mathtt{b}^{-1} z/3}\right),$$
$$I\!\!P\Bigl\{\|S^{-1/2}\,\widehat{S}\,S^{-1/2} - I_p\|_{\mathrm{op}} \geq \delta\, z\Bigr\} \leq 2p \exp\left(\frac{-z^2/2}{1+z\delta/3}\right).$$

The result can be easily extended to the case when each $\boldsymbol{\varepsilon}_i$ is not bounded but can be bounded with a high probability.

Theorem D.3.2. *Let for some* $y > 0$ *there exists a constant* $\delta(y)$ *such that*

$$I\!\!P\bigl(\|S^{-1/2}\boldsymbol{\varepsilon}\|_{\infty} > \delta(y)\bigr) \leq \mathrm{e}^{-y}.$$

Then with $\mathtt{b}$ *from* (D.19)

$$I\!\!P\{\|S^{-1/2}\,\widehat{S}\,S^{-1/2} - I_p\|_{\mathrm{op}} \geq \mathtt{b}\, z\} \leq 2p \exp\left(\frac{-z^2/2}{1+\delta^2(y) \mathtt{b}^{-1} z/3}\right) + \mathrm{e}^{-y}.$$

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

$$I\!\!P\Bigl\{\|S^{-1/2}\,\widehat{S}\,S^{-1/2} - I_p\|_{\mathrm{op}} \geq \mathtt{b}\, z(y)\Bigr\} \leq 2\mathrm{e}^{-y}.$$

To be done: The i.i.d. case

E

Gaussian comparison via KL-divergence and Pinsker's inequality

E.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 $I\!\!R^{p}$ to $I\!\!R^{\mathbb{M}}$ and $\boldsymbol{X} = T(\boldsymbol{\xi})$ and $\boldsymbol{Y} = T(\boldsymbol{\xi}^{\flat})$. We aim to bound the distance between distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ under the conditions

$$\begin{aligned} \| S^{-1/2} S^{\flat} S^{-1/2} - \boldsymbol{I}_{p} \|_{\mathrm{op}} &\leq \boldsymbol{\epsilon} \leq 1/2, \\ \operatorname{tr} \bigl(S^{-1/2} S^{\flat} S^{-1/2} - \boldsymbol{I}_{p} \bigr)^{2} &\leq \Delta^{2} \end{aligned} \tag{E.1}$$

for some $\boldsymbol{\epsilon} \leq 1/2$ and $\Delta \geq 0$. The next lemma bounds from above the Kullback-Leibler divergence between two normal distributions.

Lemma E.1.1. *Let $I\!\!P_{0} = \mathcal{N}(\boldsymbol{b}, S)$ and $I\!\!P_{1} = \mathcal{N}(\boldsymbol{b}^{\flat}, S^{\flat})$ for some non-degenerated matrices S and $S^{\flat}$. If*

$$\begin{aligned} \| S^{-1/2} S^{\flat} S^{-1/2} - \boldsymbol{I}_{p} \|_{\mathrm{op}} &\leq 1/2, \\ \operatorname{tr} \Bigl\{ \bigl(S^{-1/2} S^{\flat} S^{-1/2} - \boldsymbol{I}_{p} \bigr)^{2} \Bigr\} &\leq \Delta^{2}, \end{aligned}$$

then

$$\mathcal{K}(I\!\!P_{0}, I\!\!P_{1}) = - I\!\!E_{0} \log \frac{dI\!\!P_{1}}{dI\!\!P_{0}} \leq \frac{\Delta^{2}}{2} + \frac{1}{2} (\boldsymbol{b} - \boldsymbol{b}^{\flat})^{\top} S^{-1} S^{\flat} S^{-1} (\boldsymbol{b} - \boldsymbol{b}^{\flat}).$$

For any measurable set $A \subset I\!\!R^{p}$, it holds

$$\bigl| I\!\!P_{0}(A) - I\!\!P_{1}(A) \bigr| \leq \sqrt{\mathcal{K}(I\!\!P_{0}, I\!\!P_{1})/2}.$$

Proof. The change of variables $\boldsymbol{u} = S^{-1/2}(\boldsymbol{x} - \boldsymbol{b})$ reduces the general case to the situation when $I\!\!P_{0}$ is standard normal in $I\!\!R^{p}$ while $I\!\!P_{1} = \mathcal{N}(\boldsymbol{\beta}, B)$ with $\boldsymbol{\beta} = S^{-1/2}(\boldsymbol{b}^{\flat} - \boldsymbol{b})$ and $B \overset{\mathrm{def}}{=} S^{-1/2} S^{\flat} S^{-1/2}$

$$2 \log \frac{dI\!\!P_{1}}{dI\!\!P_{0}}(\boldsymbol{\gamma}) = \log \det(B) - (\boldsymbol{\gamma} - \boldsymbol{\beta})^{\top} B (\boldsymbol{\gamma} - \boldsymbol{\beta}) + \|\boldsymbol{\gamma}\|^{2}$$

with $\boldsymbol{\gamma}$ standard normal and

$$2\mathcal{K}(I\!P_0, I\!P_1) = -2I\!E_0 \log \frac{dI\!P_1}{dI\!P_0} = -\log\det(B) + \operatorname{tr}(B - \boldsymbol{I}_p) + \boldsymbol{\beta}^\top B \boldsymbol{\beta}.$$

Let a_j be the j th eigenvalue of $B - \boldsymbol{I}_p$. The condition $\|B - \boldsymbol{I}_p\|_{\mathrm{op}} \le 1/2$ yields $|a_j| \le 1/2$ and

$$\begin{aligned} 2\mathcal{K}(I\!P_0, I\!P_1) &= \boldsymbol{\beta}^\top B \boldsymbol{\beta} + \sum_{j=1}^{p} \bigl\{ a_j - \log(1 + a_j) \bigr\} \\ &\le \boldsymbol{\beta}^\top B \boldsymbol{\beta} + \sum_{j=1}^{p} a_j^2 \\ &\le \boldsymbol{\beta}^\top B \boldsymbol{\beta} + \operatorname{tr}(B - \boldsymbol{I}_p)^2 \le \boldsymbol{\beta}^\top B \boldsymbol{\beta} + \Delta^2. \end{aligned}$$

This implies by Pinsker's inequality

$$\sup_A |I\!P_0(A) - I\!P_1(A)| \le \sqrt{\frac{1}{2}\mathcal{K}(I\!P_0, I\!P_1)} \le \frac{1}{2}\sqrt{\Delta^2 + \boldsymbol{\beta}^\top B \boldsymbol{\beta}} \tag{E.2}$$

as required.

Notice that the operator norm bound

$$\|S^{-1/2} S^{\flat} S^{-1/2} - \boldsymbol{I}_p\|_{\mathrm{op}} \le \boldsymbol{\epsilon} \tag{E.3}$$

implies for $B = S^{-1/2} S^{\flat} S^{-1/2}$

$$\operatorname{tr}\bigl(B - \boldsymbol{I}_p\bigr)^2 \le p\boldsymbol{\epsilon}^2, \qquad \boldsymbol{\beta}^\top B \boldsymbol{\beta} \le (1 + \boldsymbol{\epsilon})\|\boldsymbol{\beta}\|^2.$$

Corollary E.1.1. *Let $I\!P_0 = \mathcal{N}(\boldsymbol{b}, S)$ and $I\!P_1 = \mathcal{N}(\boldsymbol{b}^{\flat}, S^{\flat})$ for some non-degenerated matrices S and $S^{\flat}$ satisfying (E.3). Then with $\boldsymbol{\beta} = S^{-1/2}(\boldsymbol{b} - \boldsymbol{b}^{\flat})$*

$$\sup_A |I\!P_0(A) - I\!P_1(A)| \le \frac{1}{2}\sqrt{p\boldsymbol{\epsilon}^2 + (1 + \boldsymbol{\epsilon})\|\boldsymbol{\beta}\|^2}$$

For the special case with $\boldsymbol{\beta} \equiv 0$, we bound for any Borel set $A \subset I\!R^{\mathbb{M}}$

$$\bigl| I\!P\bigl(T(\boldsymbol{\xi}) \in A\bigr) - I\!P\bigl(T(\boldsymbol{\xi}^{\flat}) \in A\bigr) \bigr| \le \Delta/2 .$$

We state a separate corollary for the distribution of the maximum.

E.2 Gaussian comparison

Corollary E.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 (E.1) holds. Then for any mapping $T \colon I\!R^p \to I\!R^{\mathbb{M}}$ and any set of values $(q_{\boldsymbol{\eta}})$, the random vectors $\boldsymbol{X} = T(\boldsymbol{\xi})$ and $\boldsymbol{Y} = T(\boldsymbol{\xi}^{\flat})$ fulfill*

$$\bigl|I\!\!P\bigl(\max_{\boldsymbol{\eta}} X_{\boldsymbol{\eta}} - q_{\boldsymbol{\eta}} > 0\bigr) - I\!\!P\bigl(\max_{\boldsymbol{\eta}} Y_{\boldsymbol{\eta}} - q_{\boldsymbol{\eta}} > 0\bigr)\bigr| \le \Delta/2.$$

Proof. We simply apply the result of the lemma to the set $A = \{\boldsymbol{x} \in I\!\!R^p \colon T(\boldsymbol{x}) \le \boldsymbol{z}\}$.

Interestingly, this method can be used for obtaining an anti-concentration bound in the case of a homogeneous mapping $T : I\!\!R^p \to I\!\!R^{\mathbb{M}}$.

Theorem E.2.1. *Let $\boldsymbol{\xi} \sim \mathcal{N}(0, S)$ be a Gaussian vector in $I\!\!R^p$. For any homogeneous mapping $T : I\!\!R^p \to I\!\!R^{\mathbb{M}}$, and for any $z > 0$ and Δ satisfying $0 \le \Delta/z \le 1$, it holds*

$$I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \ge z\bigr) - I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \ge z + \Delta\bigr) \le \Delta\, z^{-1}\sqrt{p/2}.$$

Moreover, if $\boldsymbol{\xi}^{\flat} \sim \mathcal{N}(0, S^{\flat})$ is another Gaussian vector and (E.1) holds with $\boldsymbol{\epsilon} \le 1/2$ and some $\Delta \ge 0$, then

$$\bigl|I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \ge z\bigr) - I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}^{\flat}) \ge z + \Delta\bigr)\bigr| \le \Delta/2 + \Delta\, z^{-1}\sqrt{p/2}. \tag{E.4}$$

Proof. Given z and Δ, define $\boldsymbol{\xi}^{\flat} = z/(z+\Delta)\,\boldsymbol{\xi}$. It holds by homogeneity of T

$$I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \ge z + \Delta\bigr) = I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}^{\flat}) \ge z\bigr).$$

It is obvious that $\operatorname{Var}(\boldsymbol{\xi}^{\flat}) = (1 + \Delta/z)^{-2} S$. Now it holds for the KL-divergence between $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^{\flat}$

$$\mathcal{K}\bigl(I\!\!P_{\boldsymbol{\xi}}, I\!\!P_{\boldsymbol{\xi}^{\flat}}\bigr) = \frac{p}{2}\bigl\{2\Delta/z + (\Delta/z)^2 - 2\log(1 + \Delta/z)\bigr\} \le p(\Delta/z)^2. \tag{E.5}$$

Here we used that $\log(1+\rho) \le \rho - \rho^2/2$ for $\rho \le 1$. Now Pinsker's bound (E.2) implies

$$\begin{aligned}
&\bigl|I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \ge z\bigr) - I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}^{\flat}) \ge z + \Delta\bigr)\bigr| \\
&\quad \le I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \ge z\bigr) - I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \ge z + \Delta\bigr) \\
&\qquad + \bigl|I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \ge z + \Delta\bigr) - I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}^{\flat}) \ge z + \Delta\bigr)\bigr| \\
&\quad \le \Delta/2 + \Delta\, z^{-1}\sqrt{p/2}
\end{aligned}$$

and (E.4) follows.

We also present a simple corollary of the above result which concerns the change in the expectation $I\!\!E f(\boldsymbol{\xi})$ for a bounded function f.

Lemma E.2.1. *Let $\boldsymbol{\xi} \sim \mathcal{N}(0, S)$ and $\boldsymbol{\xi}^{\flat} \sim \mathcal{N}(0, S^{\flat})$, where $S, S^{\flat}$ satisfy (E.1). For any function f on $I\!\!R^p$ with $|f(\boldsymbol{x})| \le 1$, and any $\delta > 0$ it holds*

$$\bigl| \mathbb{E} f(\boldsymbol{\xi}) - \mathbb{E} f\bigl(\boldsymbol{\xi}^{\flat}\bigr) \bigr| \leq \Delta. \tag{E.6}$$

Also, for any $\delta \geq 0$

$$\bigl| \mathbb{E} f(\boldsymbol{\xi}) - \mathbb{E} f\bigl((1+\delta)\boldsymbol{\xi}\bigr) \bigr| \leq \delta \sqrt{2p}.$$

Proof. in view of $|f(\boldsymbol{x})| \leq 1$, it holds

$$\bigl| \mathbb{E} f(\boldsymbol{\xi}) - \mathbb{E} f(\boldsymbol{\xi}^{\flat}) \bigr| \leq \int |f(\boldsymbol{x})| \cdot \bigl| \phi_{\boldsymbol{\xi}}(\boldsymbol{x}) - \phi_{\boldsymbol{\xi}^{\flat}}(\boldsymbol{x}) \bigr| d\boldsymbol{x} \leq \int \bigl| \phi_{\boldsymbol{\xi}}(\boldsymbol{x}) - \phi_{\boldsymbol{\xi}^{\flat}}(\boldsymbol{x}) \bigr| d\boldsymbol{x} \, .$$

One more use of Pinsker's inequality yields

$$\int \bigl| \phi_{\boldsymbol{\xi}}(\boldsymbol{x}) - \phi_{\boldsymbol{\xi}^{\flat}}(\boldsymbol{x}) \bigr| d\boldsymbol{x} = 2 \| \mathbb{P}_{\boldsymbol{\xi}} - \mathbb{P}_{\boldsymbol{\xi}^{\flat}} \|_{TV} \leq \sqrt{2 \mathcal{K}(\mathbb{P}_{\boldsymbol{\xi}}, \mathbb{P}_{\boldsymbol{\xi}^{\flat}})},$$

and the assertion (E.6) follows by $2 \mathcal{K}(\mathbb{P}_{\boldsymbol{\xi}}, \mathbb{P}_{\boldsymbol{\xi}^{\flat}}) \leq \Delta^2$. It remains to note that for $S^{\flat} = (1+\delta)^2 S$, it holds $\mathcal{K}(\mathbb{P}_{\boldsymbol{\xi}}, \mathbb{P}_{\boldsymbol{\xi}^{\flat}}) \leq \delta^2 p$; see (E.5).

F

Random multiplicity correction

F.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

$$\|V^{-1/2} V^{\flat} V^{-1/2} - I_p\|_{\mathrm{op}} \le \Delta_0. \tag{F.1}$$

In what follows, $I\!\!P = \mathcal{N}(0, V)$ is the normal measure on $I\!\!R^p$ with mean zero and covariance V. Similarly $I\!\!P^{\flat}$ is a random measure on $I\!\!R^p$ which is conditionally on $V^{\flat}$ normal with $I\!\!P^{\flat} = \mathcal{N}(0, V^{\flat})$. Suppose that for each m from a given set $\mathcal{M}$, a linear mapping $T_m \colon I\!\!R^p \to I\!\!R^{p_m}$ is fixed. Given $\mathtt{x}$, define for each $m \in \mathcal{M}$ the corresponding tail function $z_m(\mathtt{x})$ by

$$I\!\!P\bigl\{\boldsymbol{u} \colon \|T_m \boldsymbol{u}\| \ge z_m(\mathtt{x})\bigr\} = \mathrm{e}^{-\mathtt{x}}. \tag{F.2}$$

Also define a set $A(\mathtt{x})$ as

$$A(\mathtt{x}) \stackrel{\mathrm{def}}{=} \bigcap_{m \in \mathcal{M}} \bigl\{\boldsymbol{u} \colon \|T_m \boldsymbol{u}\| \le z_m(\mathtt{x})\bigr\} = \bigl\{\boldsymbol{u} \colon \|T_m \boldsymbol{u}\| \le z_m(\mathtt{x}), \ \forall m \in \mathcal{M}\bigr\} .$$

Similarly, for each $m \in \mathcal{M}$ and $\mathtt{x} > 0$, define $z_m^{\flat}(\mathtt{x})$ by (F.2) with $I\!\!P^{\flat}$ in place of $I\!\!P$, and introduce the corresponding set $A^{\flat}(\mathtt{x})$. Note that all these objects are random because $I\!\!P^{\flat}$ is random. Finally, let $\mathtt{x}_{\alpha}^{\flat}$ be the random quantity providing

$$I\!\!P^{\flat}\bigl(A^{\flat}(\mathtt{x}_{\alpha}^{\flat})\bigr) = 1 - \alpha. \tag{F.3}$$

Below we try to address the question whether this random multiplicity correction based on (F.3) does a good job under $I\!\!P$. This question leads to analysis of value $I\!\!P\bigl(A^{\flat}(\mathtt{x}_{\alpha}^{\flat})\bigr)$: the goal is in evaluating the difference

$$I\!\!P\bigl(A^{\flat}(\mathtt{x}_{\alpha}^{\flat})\bigr) - (1 - \alpha).$$

Theorem F.1.1. *Let a random matrix $V^{\flat}$ satisfy (F.1) for a deterministic matrix V and $\Delta_0 < 1/2$. Then it holds*

$$\bigl| I\!\!P\bigl(A^{\flat}(\mathtt{x}_{\alpha}^{\flat})\bigr) - 1 + \alpha \bigr| \le \sqrt{p}\,\Delta_0. \tag{F.4}$$

Proof. The key property of $I\!\!P^{\flat} = \mathcal{N}(0, V^{\flat})$ is that the random matrix $V^{\flat}$ concentrates around some deterministic matrix. Below we use this property in the bracketing form:

$$\begin{gathered} V^{-} \le V^{\flat} \le V^{+} \\ V^{-} \stackrel{\text{def}}{=} (1-\Delta_0)V, \quad V^{+} \stackrel{\text{def}}{=} (1+\Delta_0)V, \quad V^{+} - V^{-} = 2\Delta_0 V. \end{gathered} \tag{F.5}$$

In other words, the random matrix $V^{\flat}$ can be sandwiched in two deterministic matrices V^{-} and V^{+}. For the proof of (F.4) we use the following well known property of the Gaussian distribution.

Lemma F.1.1. *Let $I\!\!P_1 \sim \mathcal{N}(0, V_1)$ and $I\!\!P_2 \sim \mathcal{N}(0, V_2)$ with $V_1 \le V_2$. Then for any centrally symmetric star-shaped set A, it holds*

$$I\!\!P_1(A) \ge I\!\!P_2(A).$$

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 $I\!\!P^{-} = \mathcal{N}(0, V^{-})$ and $I\!\!P^{+} = \mathcal{N}(0, V^{+})$; see (F.5). Let $z_m^{-}(\mathtt{x})$ and $z_m^{+}(\mathtt{x})$ be the corresponding tail functions, and $A^{-}(\mathtt{x})$ and $A^{+}(\mathtt{x})$ - the corresponding sets. The identities (F.5) yield

$$I\!\!P^{+}\bigl(A^{+}(\mathtt{x})\bigr) = I\!\!P^{-}\bigl(A^{-}(\mathtt{x})\bigr). \tag{F.6}$$

Lemma F.1.1 implies by (F.5) for any $\mathtt{x}$

$$I\!\!P^{+}(A(\mathtt{x})) \le I\!\!P^{\flat}(A(\mathtt{x})) \le I\!\!P^{-}(A(\mathtt{x})). \tag{F.7}$$

The key step of the proof is given by the next lemma where we sandwich the random set $A^{\flat}(\mathtt{x}^{\flat})$ in two specially constructed deterministic sets.

Lemma F.1.2. *Let the deterministic values $\mathtt{x}_{\alpha}^{-}$ and $\mathtt{x}_{\alpha}^{+}$ be define by*

$$I\!\!P^{+}\bigl(A^{-}(\mathtt{x}_{\alpha}^{+})\bigr) = 1 - \alpha, \qquad I\!\!P^{-}\bigl(A^{+}(\mathtt{x}_{\alpha}^{-})\bigr) = 1 - \alpha. \tag{F.8}$$

Then

$$\begin{gathered} \mathtt{x}_{\alpha}^{-} \quad \le \quad \mathtt{x}_{\alpha}^{\flat} \quad \le \quad \mathtt{x}_{\alpha}^{+} \\ A^{-}(\mathtt{x}_{\alpha}^{-}) \subseteq A^{\flat}(\mathtt{x}_{\alpha}^{\flat}) \subseteq A^{+}(\mathtt{x}_{\alpha}^{+}). \end{gathered} \tag{F.9}$$

Proof. By Lemma F.1.1 the following relations hold true for any $\mathtt{x}$:

$$\begin{aligned} z_m^{-}(\mathtt{x}) &\le z_m^{\flat}(\mathtt{x}) \le z_m^{+}(\mathtt{x}), \\ A^{-}(\mathtt{x}) &\subseteq A^{\flat}(\mathtt{x}) \subseteq A^{+}(\mathtt{x}). \end{aligned} \tag{F.10}$$

Now by definition (F.8) in view of (F.7) and (F.10)

$$\begin{aligned} I\!\!P^{\flat}\bigl(A^{\flat}(\mathtt{x}_{\alpha}^{+})\bigr) &\ge I\!\!P^{+}\bigl(A^{\flat}(\mathtt{x}_{\alpha}^{+})\bigr) \ge I\!\!P^{+}\bigl(A^{-}(\mathtt{x}_{\alpha}^{+})\bigr) = 1-\alpha, \\ I\!\!P^{\flat}\bigl(A^{\flat}(\mathtt{x}_{\alpha}^{-})\bigr) &\le I\!\!P^{-}\bigl(A^{\flat}(\mathtt{x}_{\alpha}^{-})\bigr) \le I\!\!P^{-}\bigl(A^{+}(\mathtt{x}_{\alpha}^{-})\bigr) = 1-\alpha. \end{aligned}$$

This yields by monotonicity of $I\!\!P^{\flat}(A^{\flat}(\mathtt{x}))$ in $\mathtt{x}$ that $\mathtt{x}_{\alpha}^{\flat}$ from (F.3) belongs to the interval $[\mathtt{x}_{\alpha}^{-}, \mathtt{x}_{\alpha}^{+}]$ and

$$A^{-}(\mathtt{x}_{\alpha}^{-}) \subseteq A^{\flat}(\mathtt{x}_{\alpha}^{-}) \subseteq A^{\flat}(\mathtt{x}_{\alpha}^{\flat}) \subseteq A^{\flat}(\mathtt{x}_{\alpha}^{+}) \subseteq A^{+}(\mathtt{x}_{\alpha}^{+}).$$

This implies the result.

Now we can finalize the proof. The relations (F.9) and (F.6) imply

$$I\!\!P^{+}\bigl(A^{\flat}(\mathtt{x}_{\alpha}^{\flat})\bigr) \le I\!\!P^{+}\bigl(A^{+}(\mathtt{x}_{\alpha}^{+})\bigr) = I\!\!P^{-}\bigl(A^{-}(\mathtt{x}_{\alpha}^{+})\bigr).$$

Furthermore, it holds by Pinsker' inequality (Corollary E.2 in the supplement [SW2016]) in view of (F.1) and (F.8)

$$I\!\!P^{-}\bigl(A^{-}(\mathtt{x}_{\alpha}^{+})\bigr) \le I\!\!P^{+}\bigl(A^{-}(\mathtt{x}_{\alpha}^{+})\bigr) + \sqrt{p}\,\Delta_0 \le 1-\alpha+\sqrt{p}\,\Delta_0.$$

Similarly

$$\begin{aligned} I\!\!P^{-}\bigl(A^{\flat}(\mathtt{x}_{\alpha}^{\flat})\bigr) &\ge I\!\!P^{-}\bigl(A^{-}(\mathtt{x}_{\alpha}^{-})\bigr) = I\!\!P^{+}\bigl(A^{+}(\mathtt{x}_{\alpha}^{-})\bigr) \\ &\ge I\!\!P^{-}\bigl(A^{+}(\mathtt{x}_{\alpha}^{-})\bigr) - \sqrt{p}\,\Delta_0 = 1-\alpha-\sqrt{p}\,\Delta_0. \end{aligned}$$

This implies (F.4) for the measure $I\!\!P$.

F.2 Max-case

Now we consider a more general situation. Let T_m be a family of test statistics in the real world, and $\mathbb{T}_m^{\flat}$

G

High-dimensional inference for a Gaussian law

This section collects some useful facts about high dimensional Gaussian measures.

G.1 Stein identity, Slepian bridge, and Gaussian comparison

Below for a $\mathbb{M} \times \mathbb{M}$ matrix A, we denote

$$\begin{aligned} \|A\|_{\mathrm{op}} &= \sup_{\|\boldsymbol{u}\|=1} \|A\boldsymbol{u}\|, & \|A\|_{\infty} &= \max_{i,j} |a_{i,j}|, \\ \|A\|_{1} &= \sum_{i,j} |a_{i,j}|, & \|A\|_{Fr}^2 &= \sum_{ij} a_{ij}^2. \end{aligned}$$

Lemma G.1.1. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two zero mean Gaussian vectors in $I\!R^{\mathbb{M}}$ with $\Sigma_{\boldsymbol{X}} = \operatorname{Var}(\boldsymbol{X})$ and $\Sigma_{\boldsymbol{Y}} = \operatorname{Var}(\boldsymbol{Y})$. Let also $f(\boldsymbol{x})$ be a smooth function on $I\!R^{\mathbb{M}}$. Then*

$$\epsilon \stackrel{\mathrm{def}}{=} \bigl| I\!E f(\boldsymbol{X}) - I\!E f(\boldsymbol{Y}) \bigr| \le \frac{1}{2} \, \|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\|_{\infty} \, \|\nabla^2 f\|_{1,\infty} \, , \tag{G.1}$$

where $\|\nabla^2 f\|_{1,\infty} \stackrel{\mathrm{def}}{=} \sup_{\boldsymbol{x}} \|\nabla^2 f(\boldsymbol{x})\|_1$.

Proof. Without loss of generality assume that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are given on the same probability space and independent. For each $t \in [0,1]$, define

$$\begin{aligned} \boldsymbol{Z}(t) &\stackrel{\mathrm{def}}{=} \sqrt{t}\,\boldsymbol{X} + \sqrt{1-t}\,\boldsymbol{Y}, \\ \Psi(t) &\stackrel{\mathrm{def}}{=} I\!E f(\boldsymbol{Z}(t)) = I\!E f\bigl(\sqrt{t}\,\boldsymbol{X} + \sqrt{1-t}\,\boldsymbol{Y}\bigr). \end{aligned}$$

Obviously

$$\epsilon = |\Psi(1) - \Psi(0)| = \left| \int_0^1 \Psi'(t) dt \right|. \tag{G.2}$$

Further,

$$\begin{aligned} \Psi'(t) &= \mathbb{E}[\nabla f(\boldsymbol{Z}(t))^{\top} \boldsymbol{Z}'(t)] \\ &= \frac{1}{2} \mathbb{E}\bigl[\bigl\{ t^{-1/2} \boldsymbol{X} - (1-t)^{-1/2} \boldsymbol{Y} \bigr\}^{\top} \nabla f(\boldsymbol{Z}(t)) \bigr]. \end{aligned} \tag{G.3}$$

To compute this expectation, we apply the *Stein identity*. Let $\boldsymbol{W}$ be a zero mean Gaussian vector in $\mathbb{R}^{\mathbb{M}}$. Then for any C^1 function $s \colon \mathbb{R}^{\mathbb{M}} \to \mathbb{R}^{\mathbb{M}}$, it holds

$$\mathbb{E}[\boldsymbol{W} \, s(\boldsymbol{W})] = \operatorname{Var}(\boldsymbol{W}) \, \mathbb{E}[\nabla s(\boldsymbol{W})]. \tag{G.4}$$

Exercise G.1.1. Prove (G.4) for standard normal $\boldsymbol{W}$ using integration by part:

$$\int_{\mathbb{R}^{\mathbb{M}}} s(\boldsymbol{w}) \, \boldsymbol{w} \, \mathrm{e}^{-\|\boldsymbol{w}\|^2/2} d\boldsymbol{w} = \int_{\mathbb{R}^{\mathbb{M}}} \nabla s(\boldsymbol{w}) \, \mathrm{e}^{-\|\boldsymbol{w}\|^2/2} d\boldsymbol{w}.$$

Reduce the case of a Gaussian zero mean $\boldsymbol{w} \sim \mathcal{N}(0, \Sigma)$ with a positive symmetric matrix Σ to the case $\Sigma = \boldsymbol{I}_{\mathbb{M}}$.

This results can be directly extended to any C^1 vector function $\boldsymbol{s} \colon \mathbb{R}^{\mathbb{M}} \to \mathbb{R}^{q}$: it holds

$$\mathbb{E}[\boldsymbol{W} \, \boldsymbol{s}(\boldsymbol{W})^{\top}] = \operatorname{Var}(\boldsymbol{W}) \, \mathbb{E}[\nabla \boldsymbol{s}(\boldsymbol{W})^{\top}]. \tag{G.5}$$

Here $\nabla \boldsymbol{s}(\boldsymbol{w})^{\top}$ means the $p \times q$ matrix with the entries $\frac{d}{d\theta_j} s_m(\boldsymbol{w})$ for $j = 1, \ldots, p$ and $m = 1, \ldots, q$.

Exercise G.1.2. Derive (G.5) by applying (G.4) columnwise.

The identity (G.5) is used with $\boldsymbol{W} = (\boldsymbol{X}^{\top}, \boldsymbol{Y}^{\top})^{\top}$ and $\boldsymbol{s}(\boldsymbol{w}) = \nabla f\bigl(\boldsymbol{z}(t)\bigr)$ for $\boldsymbol{z}(t) = \sqrt{t}\, \boldsymbol{x} + \sqrt{1-t}\, \boldsymbol{y}$. Independence of $\boldsymbol{X}$ and $\boldsymbol{Y}$ implies

$$\operatorname{Var}(\boldsymbol{W}) = \begin{pmatrix} \Sigma_{\boldsymbol{X}} & 0 \\ 0 & \Sigma_{\boldsymbol{Y}} \end{pmatrix}.$$

Also $\nabla \boldsymbol{s}(\boldsymbol{w}) = \bigl(t^{1/2} \, \nabla^2 f(\boldsymbol{z}(t)), (1-t)^{1/2} \, \nabla^2 f(\boldsymbol{z}(t)) \bigr)^{\top}$ and by (G.5)

$$\begin{aligned} \mathbb{E}\bigl[\nabla f(\boldsymbol{Z}(t)) \boldsymbol{X}^{\top}\bigr] &= t^{1/2} \Sigma_{\boldsymbol{X}} \mathbb{E}\bigl[\nabla^2 f\bigl(\boldsymbol{Z}(t)\bigr)\bigr] \\ \mathbb{E}\bigl[\nabla f(\boldsymbol{Z}(t)) \boldsymbol{Y}^{\top}\bigr] &= (1-t)^{1/2} \Sigma_{\boldsymbol{Y}} \mathbb{E}\bigl[\nabla^2 f\bigl(\boldsymbol{Z}(t)\bigr)\bigr], \end{aligned}$$

This and (G.3) imply

$$\begin{aligned} \bigl|\Psi'(t)\bigr| &\le \frac{1}{2} \Bigl| \operatorname{tr}\bigl\{ \bigl(\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\bigr) \mathbb{E}\bigl[\nabla^2 f\bigl(\boldsymbol{Z}(t)\bigr)\bigr] \bigr\} \Bigr| \\ &\le \frac{1}{2} \, \|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\|_{\infty} \, \bigl\| \mathbb{E}\bigl[\nabla^2 f\bigl(\boldsymbol{Z}(t)\bigr)\bigr] \bigr\|_1 \le \frac{1}{2} \, \|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\|_{\infty} \, \|\nabla^2 f\|_{1,\infty}. \end{aligned}$$

Now the assertion follows from (G.2).

Now we apply the obtained bound to $f(\boldsymbol{x}) \stackrel{\mathrm{def}}{=} g\bigl(\Delta^{-1} h_{\beta}(\boldsymbol{x})\bigr)$, where $g(z)$ is a smooth univariate function with bounded first and second derivatives, and the *smooth maximum* function: for some $\beta > 0$

$$h_{\beta}(\boldsymbol{x}) = \beta^{-1} \log\biggl(\sum_{j=1}^{\mathbb{M}} \mathrm{e}^{\beta x_j}\biggr). \tag{G.6}$$

Lemma G.1.2. *Let* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *be two zero mean Gaussian vectors in* $I\!R^{\mathbb{M}}$ *with* $\Sigma_{\boldsymbol{X}} = \operatorname{Var}(\boldsymbol{X})$ *and* $\Sigma_{\boldsymbol{Y}} = \operatorname{Var}(\boldsymbol{Y})$. *For a univariate function* $g(z)$ *with bounded first and second derivatives, and* $h_{\beta}(\boldsymbol{x})$ *from* (G.6)

$$\bigl| I\!E g\bigl(\Delta^{-1} h_{\beta}(\boldsymbol{X})\bigr) - I\!E g\bigl(\Delta^{-1} h_{\beta}(\boldsymbol{Y})\bigr) \bigr| \le \Bigl(\frac{\beta \|g'\|_{\infty}}{\Delta} + \frac{\|g''\|_{\infty}}{2\Delta^2} \Bigr) \|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\|_{\infty} . \tag{G.7}$$

Proof. It holds for $f(\boldsymbol{x}) \stackrel{\mathrm{def}}{=} g\bigl(\Delta^{-1} h_{\beta}(\boldsymbol{x})\bigr)$

$$\begin{aligned} \nabla f(\boldsymbol{x}) &= \Delta^{-1} g'(\Delta^{-1} h_{\beta}(\boldsymbol{x})) \nabla h_{\beta}(\boldsymbol{x}), \\ \nabla^2 f(\boldsymbol{x}) &= \Delta^{-1} g'(\Delta^{-1} h_{\beta}(\boldsymbol{x})) \nabla^2 h_{\beta}(\boldsymbol{x}) + \Delta^{-2} g''(\Delta^{-1} h_{\beta}(\boldsymbol{x})) \nabla h_{\beta}(\boldsymbol{x}) \nabla h_{\beta}(\boldsymbol{x})^{\top}. \end{aligned}$$

Also for any $\boldsymbol{x}$ by direct calculus

$$\begin{aligned} \|\nabla h_{\beta}(\boldsymbol{x})\|_1 &= 1, \\ \|\nabla^2 h_{\beta}(\boldsymbol{x})\|_1 &\le 2\beta. \end{aligned}$$

This implies

$$\begin{aligned} \|\nabla f(\boldsymbol{x})\|_1 &\le \Delta^{-1} \|g'\|_{\infty} \times \|\nabla h_{\beta}(\boldsymbol{x})\|_1 \le \Delta^{-1} \|g'\|_{\infty} , \\ \|\nabla^2 f(\boldsymbol{x})\|_1 &\le \Delta^{-1} \|g'\|_{\infty} \times \|\nabla^2 h_{\beta}(\boldsymbol{x})\|_1 + \Delta^{-2} \|g''\|_{\infty} \times \|\nabla h_{\beta}(\boldsymbol{x})\|_1^2 \\ &\le 2\Delta^{-1} \beta \|g'\|_{\infty} + \Delta^{-2} \|g''\|_{\infty} . \end{aligned}$$

Now (G.7) follows from (G.1).

A particular choice of the function g is given by

$$g(z) \stackrel{\mathrm{def}}{=} \begin{cases} 2u^2, & u \in [0, 1/2], \\ 1 - 2(1-u)^2, & u \in [1/2, 1]. \\ 0 & \text{otherwise.} \end{cases} \tag{G.8}$$

Obviously $|g'(u)| \le 2$, $|g''(u)| \le 4$ for all u. Then $\|g'_{\Delta}\|_{\infty} \le 2\Delta^{-1}$, $\|g''_{\Delta}\|_{\infty} \le 4\Delta^{-2}$. We conclude with the following bound.

Theorem G.1.1. *Let* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *be two zero mean Gaussian vectors in* $I\!R^{\mathbb{M}}$ *with* $\Sigma_{\boldsymbol{X}} = \operatorname{Var}(\boldsymbol{X})$ *and* $\Sigma_{\boldsymbol{Y}} = \operatorname{Var}(\boldsymbol{Y})$. *Then with* $g(\cdot)$ *given by* (G.8), *it holds for any* $\Delta > 0$ *and* $\beta > 0$

$$\bigl| I\!E g\bigl(\Delta^{-1} h_{\beta}(\boldsymbol{X})\bigr) - I\!E g\bigl(\Delta^{-1} h_{\beta}(\boldsymbol{Y})\bigr) \bigr| \le 2(\beta\Delta^{-1} + \Delta^{-2}) \, \|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\|_{\infty} . \tag{G.9}$$

G.2 Comparing of the maximum of Gaussians

Let $\boldsymbol{X} = (X_j)$ and $\boldsymbol{Y} = (Y_j)$ be two zero mean Gaussian vectors in $I\!R^{\mathbb{M}}$ with $\Sigma_{\boldsymbol{X}} = \operatorname{Var}(\boldsymbol{X})$ and $\Sigma_{\boldsymbol{Y}} = \operatorname{Var}(\boldsymbol{Y})$, and let $\Box = \|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\|_{\infty}$. 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

$$\max_j x_j \le h_{\beta}(\boldsymbol{x}) \le \max_j x_j + \beta^{-1} \log(\mathbb{M}).$$

As the indicator function $1\!\!I(z \ge 0)$ 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 $u \le 0$, $g(u) = 1$ for $u \ge 1$, 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 $\Delta = \beta^{-1} \log(\mathbb{M})$

$$g_{\Delta} \circ h_{\beta}(\boldsymbol{x} - \boldsymbol{\Delta}) \le 1\!\!I(\max_j x_j > 0) \le g_{\Delta} \circ h_{\beta}(\boldsymbol{x} + \boldsymbol{\Delta}).$$

Here $\boldsymbol{\Delta}$ is the vector with all entries equal to Δ. Indeed, $g_{\Delta}(z) \in [0,1]$ for any z. If $x_j \ge 0$ for some j, then $h_{\beta}(\boldsymbol{x} + \boldsymbol{\Delta}) \ge \Delta$ and hence,

$$g_{\Delta} \circ h_{\beta}(\boldsymbol{x} + \boldsymbol{\Delta}) \ge g(\Delta/\Delta) = g(1) = 1.$$

Similarly, if $\max_j x_j \le 0$, then due to $\Delta = \beta^{-1} \log(\mathbb{M})$

$$h_{\beta}(\boldsymbol{x} - \boldsymbol{\Delta}) \le \max_j (x_j - \Delta) + \beta^{-1} \log(\mathbb{M}) \le 0$$

and $g_{\Delta} \circ h_{\beta}(\boldsymbol{x} - \boldsymbol{\Delta}) = 0$. This and (G.9) yield the bound

$$\begin{aligned} I\!P\bigl(\max_j X_j > 0\bigr) &\le I\!E\bigl[g_{\Delta} \circ h_{\beta}(\boldsymbol{X} + \boldsymbol{\Delta})\bigr] \\ &\le I\!E\bigl[g_{\Delta} \circ h_{\beta}(\boldsymbol{Y} + \boldsymbol{\Delta})\bigr] + 2(\beta\Delta^{-1} + \Delta^{-2}) \, \Box \\ &\le I\!P\bigl(\max_j Y_j > -2\Delta\bigr) + 2\Delta^{-2} \bigl\{ \log(\mathbb{M}) + 1 \bigr\} \, \Box. \end{aligned}$$

Similarly one can approximate any indicator $1\!\!I(z \ge z_0)$ by shifting the function g.

Theorem G.2.1. *Let* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *be two zero mean Gaussian vectors in* $I\!R^{\mathbb{M}}$ *with* $\Sigma_{\boldsymbol{X}} = \operatorname{Var}(\boldsymbol{X})$ *and* $\Sigma_{\boldsymbol{Y}} = \operatorname{Var}(\boldsymbol{Y})$.

$$\square \overset{\text{def}}{=} \|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\|_{\infty} ,$$

it holds for any Δ *and* z

$$I\!P\bigl(\max_j X_j > z\bigr) \le I\!P\bigl(\max_j Y_j > z - 2\Delta\bigr) + 2\Delta^{-2}\bigl\{\log(\mathbb{M}) + 1\bigr\}\,\square.$$

G.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 $I\!P\bigl(\max_j X_j > 0\bigr)$ by a similar probability $I\!P\bigl(\max_j Y_j > -2\Delta\bigr)$ from above and $I\!P\bigl(\max_j Y_j > 2\Delta\bigr)$ from below up to the error term $2\Delta^{-2}\bigl\{\log(\mathbb{M}) + 1\bigr\}\,\square$. 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

$$I\!P\bigl(\max_j Y_j > -2\Delta\bigr) - I\!P\bigl(\max_j Y_j > 0\bigr).$$

The following theorem provides bounds on the Lévy concentration function of the maximum of a Gaussian random vector in $I\!R^{\mathbb{M}}$, 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

$$\mathcal{L}(\xi, \varepsilon) = \sup_{x \in I\!R} I\!P(|\xi - x| \le \varepsilon).$$

Theorem G.3.1 (Anti-concentration). *Let* $(X_1, \ldots, X_{p_n})^{\top}$ *be a centered Gaussian random vector in* $I\!R^{\mathbb{M}}$ *with* $\sigma_j^2 = I\!E[X_j^2] > 0$ *for all* $1 \le j \le \mathbb{M}$. *Moreover, let* $\underline{\sigma} = \min_{1 \le j \le \mathbb{M}} \sigma_j$, $\overline{\sigma} = \max_{1 \le j \le \mathbb{M}} \sigma_j$, *and* $a_{\mathbb{M}} = I\!E[\max_{1 \le j \le \mathbb{M}}(X_j/\sigma_j)]$.

1. *If the variances are all equal, namely* $\underline{\sigma} = \overline{\sigma} = \sigma$, *then for every* $\epsilon > 0$,

$$\mathcal{L}\Bigl(\max_{1 \le j \le \mathbb{M}} X_j, \epsilon\Bigr) \le 4\epsilon(a_{\mathbb{M}} + 1)/\sigma;$$

2. *If the variances are not equal, namely* $\underline{\sigma} < \overline{\sigma}$, *then for every* $\epsilon > 0$,

$$\mathcal{L}\Bigl(\max_{1 \le j \le \mathbb{M}} X_j, \epsilon\Bigr) \le \mathtt{C}\epsilon\{a_{\mathbb{M}} + 1 \vee \log(\underline{\sigma}/\epsilon)\}$$

where $\mathtt{C} > 0$ *depends only on* $\underline{\sigma}$ *and* $\overline{\sigma}$.

To compare the distribution of two maxima, we use the anti-concentration bound: if $\operatorname{Var}(Y_j) \equiv \sigma^2$

$$I\!P\bigl(\max_j Y_j > 0\bigr) - I\!P\bigl(\max_j Y_j > -2\Delta\bigr) \le 8\Delta(a_{\mathbb{M}} + 1)/\sigma,$$

where $a_{\mathbb{M}} \stackrel{\operatorname{def}}{=} I\!\!E \max_j |Y_j/\sigma| \le (2\log \mathbb{M})^{1/2}$. If the variances $\sigma_j^2 \stackrel{\operatorname{def}}{=} \operatorname{Var}(Y_j)$ are unequal then

$$I\!\!P\bigl(\max_j Y_j > 0\bigr) - I\!\!P\bigl(\max_j Y_j > -2\Delta\bigr) \le \mathtt{C}\Delta\sqrt{\log(\mathbb{M}/\Delta)}.$$

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 (G.9)

$$\begin{aligned}
&\bigl|I\!\!P\bigl(\max_j X_j > 0\bigr) - I\!\!P\bigl(\max_j Y_j > 0\bigr)\bigr| \\
&\quad \le (2\beta\Delta^{-1} + 2\Delta^{-2})\,\square + \mathtt{C}\Delta\sqrt{\log(\mathbb{Q})} \\
&\quad \le \mathtt{C}\,\square^{1/3}\log^{2/3}(\mathbb{Q}) + \mathtt{C}\,\square^{1/3}\log^{2/3}(\mathbb{Q}) \\
&\quad \le \mathtt{C}\,\square^{1/3}\log^{2/3}(\mathbb{Q}).
\end{aligned}$$

The definition yields $\mathbb{Q} = \mathbb{M}/\Delta \le \mathbb{M}/\square^{1/3}$. We conclude with the following result.

Theorem G.3.2. *Let* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *be two zero mean Gaussian vectors in* $I\!\!R^{\mathbb{M}}$ *with* $\Sigma_{\boldsymbol{X}} = \operatorname{Var}(\boldsymbol{X})$ *and* $\Sigma_{\boldsymbol{Y}} = \operatorname{Var}(\boldsymbol{Y})$ *. With*

$$\square \stackrel{\operatorname{def}}{=} \|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\|_\infty ,$$

it holds for any Δ

$$\bigl|I\!\!P\bigl(\max_j X_j > 0\bigr) - I\!\!P\bigl(\max_j Y_j > 0\bigr)\bigr| \le \mathtt{C}\,\square^{1/3}\log^{2/3}(\mathbb{M}/\square^{1/3}).$$

G.4 Gaussian comparison for the squared norm

This section considers a special class of functions which only depends on the Euclidean norm of a vector. Namely, given a vector $\boldsymbol{X}$ and a smooth univariate function g , we check the variability of the value $I\!\!E g(\|\boldsymbol{X}\|^2)$ with respect to the covariance of $\boldsymbol{X}$.

Lemma G.4.1. *Let* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *be two zero mean Gaussian vectors in* $I\!\!R^{\mathbb{M}}$ *with* $\Sigma_{\boldsymbol{X}} = \operatorname{Var}(\boldsymbol{X})$ *and* $\Sigma_{\boldsymbol{Y}} = \operatorname{Var}(\boldsymbol{Y})$ *. For any continuously differentiable function* g *, it holds*

$$I\!\!E g(\|\boldsymbol{X}\|^2) - I\!\!E g(\|\boldsymbol{Y}\|^2) = 2\operatorname{tr}\biggl(\bigl(\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\bigr) \int_0^1 \Sigma_s^{-1}\, I\!\!E\bigl[\boldsymbol{Z}_s \boldsymbol{Z}_s^\top g'(\|\boldsymbol{Z}_s\|^2)\bigr]\,ds \biggr), \tag{G.10}$$

where

$$\boldsymbol{Z}_s \stackrel{\operatorname{def}}{=} \sqrt{s}\,\boldsymbol{X} + \sqrt{1-s}\,\boldsymbol{Y}, \qquad \Sigma_s \stackrel{\operatorname{def}}{=} \operatorname{Var}(\boldsymbol{Z}_s) = s\,\Sigma_{\boldsymbol{X}} + (1-s)\Sigma_{\boldsymbol{Y}} .$$

Proof. It holds

$$\mathbb{E} g(\|\boldsymbol{X}\|^2) - \mathbb{E} g(\|\boldsymbol{Y}\|^2) = \mathbb{E} g(\|\boldsymbol{Z}_1\|^2) - \mathbb{E} g(\|\boldsymbol{Z}_0\|^2) = \mathbb{E} \int_0^1 \frac{d}{ds} g(\|\boldsymbol{Z}_s\|^2)\, ds \,.$$

Further,

$$\frac{d}{ds} g(\|\boldsymbol{Z}_s\|^2) = \Bigl(\frac{\boldsymbol{X}}{\sqrt{s}} - \frac{\boldsymbol{Y}}{\sqrt{1-t}}\Bigr)^\top \boldsymbol{Z}_s \, g'(\|\boldsymbol{Z}_s\|^2) \,.$$

The Stein lemma, applied to the Gaussian vector $\boldsymbol{X}$ for $\boldsymbol{Y}$ fixed, and the chain rule $\frac{d}{d\boldsymbol{X}} = \sqrt{s}\, \frac{d}{d\boldsymbol{Z}_s}$ imply

$$\begin{aligned} \mathbb{E}\bigl[\boldsymbol{X}^\top \boldsymbol{Z}_s \, g'(\|\boldsymbol{Z}_s\|^2)\bigr] &= \operatorname{tr} \mathbb{E}\bigl[\boldsymbol{X} \boldsymbol{Z}_s^\top \, g'(\|\boldsymbol{Z}_s\|^2)\bigr] = \operatorname{tr}\left(\Sigma_{\boldsymbol{X}} \, \mathbb{E}\left[\frac{d}{d\boldsymbol{X}} \{\boldsymbol{Z}_s^\top \, g'(\|\boldsymbol{Z}_s\|^2)\}\right]\right) \\ &= \sqrt{s} \, \operatorname{tr}\left(\Sigma_{\boldsymbol{X}} \, \mathbb{E}\left[\frac{d}{d\boldsymbol{Z}_s} \boldsymbol{Z}_s^\top \, g'(\|\boldsymbol{Z}_s\|^2)\right]\right). \end{aligned}$$

Similarly

$$\begin{aligned} \mathbb{E}\bigl[\boldsymbol{Y}^\top \boldsymbol{Z}_s \, g'(\|\boldsymbol{Z}_s\|^2)\bigr] &= \operatorname{tr} \mathbb{E}\bigl[\boldsymbol{Y} \boldsymbol{Z}_s^\top \, g'(\|\boldsymbol{Z}_s\|^2)\bigr] \\ &= \operatorname{tr}\left(\Sigma_{\boldsymbol{Y}} \, \mathbb{E}\left[\frac{d}{d\boldsymbol{X}} \{\boldsymbol{Z}_s^\top \, g'(\|\boldsymbol{Z}_s\|^2)\}\right]\right) \\ &= \sqrt{1-t} \, \operatorname{tr}\left(\Sigma_{\boldsymbol{Y}} \, \mathbb{E}\left[\frac{d}{d\boldsymbol{Z}_s} \{\boldsymbol{Z}_s^\top \, g'(\|\boldsymbol{Z}_s\|^2)\}\right]\right). \end{aligned}$$

Therefore,

$$\int_0^1 \mathbb{E} \frac{d}{ds} g(\|\boldsymbol{Z}_s\|^2)\, ds = \operatorname{tr}\left((\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}) \int_0^1 \mathbb{E}\left[\frac{d}{d\boldsymbol{Z}_s} \{\boldsymbol{Z}_s^\top \, g'(\|\boldsymbol{Z}_s\|^2)\}\right] ds\right).$$

Now the result follows by one more application of the Stein lemma to $\boldsymbol{Z}_s^\top \, g'(\|\boldsymbol{Z}_s\|^2)$.

If g is a monotonous function with $g'(s) \ge 0$, then we can upper bound the distance $\mathbb{E} g(\|\boldsymbol{X}\|^2) - \mathbb{E} g(\|\boldsymbol{Y}\|^2)$.

Theorem G.4.1. *Let* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *be two zero mean Gaussian vectors in* $\mathbb{R}^{\mathbb{M}}$ *with* $\Sigma_{\boldsymbol{X}} = \operatorname{Var}(\boldsymbol{X})$ *and* $\Sigma_{\boldsymbol{Y}} = \operatorname{Var}(\boldsymbol{Y})$. *For any continuously differentiable function* g, *it holds*

$$\bigl|\mathbb{E} g(\|\boldsymbol{X}\|^2) - \mathbb{E} g(\|\boldsymbol{Y}\|^2)\bigr| \le 2 \bigl\|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\bigr\|_{\mathrm{Sch},1} \|g'\|_\infty \,.$$

Proof. The result (G.10) implies

$$\begin{aligned} \bigl|\mathbb{E} g(\|\boldsymbol{X}\|^2) - \mathbb{E} g(\|\boldsymbol{Y}\|^2)\bigr| &\le 2 \bigl\|g'\bigr\|_\infty \sup_{0 \le s \le 1} \Bigl|\operatorname{tr}\Bigl((\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}) \Sigma_s^{-1} \, \mathbb{E}\bigl[\boldsymbol{Z}_s \boldsymbol{Z}_s^\top\bigr]\Bigr)\Bigr| \\ &\le 2 \bigl\|g'\bigr\|_\infty \bigl\|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\bigr\|_{\mathrm{Sch},1} \end{aligned}$$

as required.

Now we apply this result with a specific function $g(s)$ of the form

$$g(s) = \varkappa_\lambda(t-s) = \varkappa\bigl(\lambda(t-s)\bigr)$$

for a fixed t and $\varkappa(\cdot)$ from (G.11). As $0 \le \varkappa'(t) \le 1/2$, we derive

$$\bigl|\mathbb{E}\varkappa_\lambda(t - \|\boldsymbol{X}\|^2) - \mathbb{E}\varkappa_\lambda(t - \|\boldsymbol{Y}\|^2)\bigr| \le \lambda \bigl\|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\bigr\|_{\mathrm{Sch},1}.$$

G.5 Approximation of the indicator function

Let Z be a random variable with an absolutely continuous d.f. F_Z, and let its density function $p(t)$ be continuous. The next result quantifies the accuracy of approximation of the d.f. $F_Z(t) = \mathbb{P}(Z \le t)$ by expectation $\mathbb{E}\varkappa_\lambda(t-Z)$, where

$$\varkappa_\lambda(t) \overset{\mathrm{def}}{=} \varkappa\bigl(\lambda t\bigr)$$

and $\varkappa$ is a smooth approximation of the indicator function

$$\varkappa(t) \overset{\mathrm{def}}{=} \mathbb{I}(t \ge 0) - \frac{1}{2}\,\mathrm{sign}(t)\mathrm{e}^{-|t|} = \begin{cases} \frac{1}{2}\mathrm{e}^{-|t|} & t < 0, \\ 1 - \frac{1}{2}\mathrm{e}^{-t} & t \ge 0. \end{cases} \tag{G.11}$$

It obviously holds

$$\begin{aligned} \varkappa'(t) &= \frac{1}{2}\mathrm{e}^{-|t|}, \\ \varkappa''(t) &= \delta(t) - \frac{1}{2}\,\mathrm{sign}(t)\mathrm{e}^{-|t|}. \end{aligned}$$

Here $\delta(t)$ stands for the Dirac delta-function at zero.

Denote by $\tau_\lambda(t)$ the error of approximating the probability function by its smoothed counterpart:

$$\tau_\lambda(t) \overset{\mathrm{def}}{=} \mathbb{P}(Z \le t) - \mathbb{E}\varkappa_\lambda(t-Z) = \frac{1}{2}\mathbb{E}\bigl[\mathrm{sign}(t-Z)\mathrm{e}^{-\lambda|t-Z|}\bigr].$$

Lemma G.5.1. *Let the density $p_Z(t)$ be continuously differentiable. It holds for any t and any $\lambda > 0$*

$$\begin{aligned} \tau_\lambda'(t) &= \frac{1}{2}\bigl\{p_Z(t) - \lambda\mathbb{E}\mathrm{e}^{-\lambda|t-Z|}\bigr\}, \\ \tau_\lambda''(t) &= \frac{1}{2}p_Z'(t) + \lambda^2\tau_\lambda(t). \end{aligned}$$

Moreover, it holds

$$\sup_t\bigl|\tau_\lambda(t)\bigr| \le \frac{1}{2\lambda^2}\sup_t\bigl|p_Z'(t)\bigr|. \tag{G.12}$$

Proof. It holds

$$\tau_\lambda'(t) = \frac{1}{2}\mathbb{E}\bigl[\bigl\{\delta(t-Z)-\lambda\bigr\}\mathrm{e}^{-\lambda|t-Z|}\bigr] = \frac{1}{2}\bigl\{p_Z(t)-\lambda\mathbb{E}\mathrm{e}^{-\lambda|t-Z|}\bigr\}$$

$$\tau_\lambda''(t) = \frac{1}{2}\bigl\{p_Z'(t)+\lambda^2\operatorname{sign}(t-Z)\mathrm{e}^{-\lambda|t-Z|}\bigr\} = \frac{1}{2}p_Z'(t)+\lambda^2\tau_\lambda(t).$$

If $t_{\max}$ is a point of maximum of $\tau_\lambda(t)$, then $\tau_\lambda''(t_{\max}) \le 0$ and

$$\sup_t \lambda^2\tau_\lambda(t) = \lambda^2\tau_\lambda(t_{\max}) = \tau_\lambda''(t_{\max}) - \frac{1}{2}p_Z'(t_{\max}) \le -\frac{1}{2}p_Z'(t_{\max}) \le \frac{1}{2}\sup_t\bigl|p_Z'(t)\bigr|.$$

Similarly, for the point of minimum $t_{\min} = \operatorname{argmin}_t \tau_\lambda(t)$, the extremum condition $\tau_\lambda''(t_{\min}) \ge 0$ yields

$$\inf_t \lambda^2\tau_\lambda(t) = \lambda^2\tau_\lambda(t_{\min}) = \tau_\lambda''(t_{\min}) - \frac{1}{2}p_Z'(t_{\min}) \ge -\frac{1}{2}p_Z'(t_{\min}) \ge -\frac{1}{2}\sup_t\bigl|p_Z'(t)\bigr|.$$

This yields (G.12).

References

Arlot, S. (2009). Model selection by resampling penalization. *Electron. J. Statist.*, 3:557–624.

Baraud, Y., Huet, S., and Laurent, B. (2003). Adaptive tests of linear hypotheses by model selection. *Ann. Statist.*, 31(1):225–251.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413.

Beran, R. (1986). Discussion: Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14(4):1295–1298.

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. (2007). Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, 138(1-2):33–73.

Borokov, A. (1999). *Mathematical Statistics.* Taylor & Francis.

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.

Gach, F., Nickl, R., and Spokoiny, V. (2013). Spatially Adaptive Density Estimation by Localised Haar Projections. *Annales de l'Institut Henri Poincare - Probability and Statistics*, 49(3):900–914. DOI: 10.1214/12-AIHP485; arXiv:1111.2807.

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.

Green, P. J. and Silverman, B. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach.* London: Chapman & Hall. .

Ibragimov, I. and Khas'minskij, R. (1981). *Statistical estimation. Asymptotic theory. Transl. from the Russian by Samuel Kotz.* New York - Heidelberg -Berlin: Springer-Verlag .

Kneip, A. (1994). Ordered linear smoothers. *Ann. Stat.*, 22(2):835–866.

Kullback, S. (1997). *Information Theory and Statistics.* Dover Books on Mathematics. Dover Publications.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338.

Lehmann, E. and Casella, G. (1998). *Theory of point estimation. 2nd ed.* New York, NY: Springer, 2nd ed. edition.

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. (1997a). 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., Mammen, E., and Spokoiny, V. G. (1997b). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, 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.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660.

Massart, P. (2007). *Concentration inequalities and model selection.* Number 1896 in Ecole d'Eté de Probabilités de Saint-Flour. Springer.

Polzehl, J. and Spokoiny, V. (2006). Propagation-separation approach for local likelihood estimation. *Probab. Theory Relat. Fields*, 135(3):335–362.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

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. *Ann. Statist.*, 43(6):2653–2675. arXiv:1410.0347.

Spokoiny, V. G. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.*, 24(6):2477–2498.

Strasser, H. (1985). Mathematical theory of statistics. Statistical experiments and asymptotic decision theory. De Gruyter Studies in Mathematics, 7. Berlin - New York: de Gruyter.

Tropp, J. A. (2015). Found. Trends Mach. Learning. to appear.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14(4):1261–1295.