

Finite Sample Bernstein – von Mises Theorem for Semiparametric Problems

Maxim Panov*

Moscow Institute of Physics and Technology,
Institute for Information Transmission Problems of RAS,
Datadvance Company,
Pokrovsky blvd. 3 building 1B,
109028 Moscow, Russia
`maxim.panov@datadvance.net`

Vladimir Spokoiny[†]

Weierstrass Institute and
Humboldt University Berlin,
Moscow Institute of Physics and Technology
Mohrenstr. 39,
10117 Berlin, Germany
`spokoiny@wias-berlin.de`

June 15, 2014

*The author is partially supported by Laboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF government grant, ag. 11.G34.31.0073.

[†]The author is partially supported by Laboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF government grant, ag. 11.G34.31.0073. Financial support by the German Research Foundation (DFG) through the Collaborative Research Center 649 "Economic Risk" is grate-

Abstract

The classical parametric and semiparametric Bernstein – von Mises (BvM) results are reconsidered in a non-classical setup allowing finite samples and model misspecification. In the case of a finite dimensional nuisance parameter we obtain an upper bound on the error of Gaussian approximation of the posterior distribution for the target parameter which is explicit in the dimension of the nuisance and target parameters. This helps to identify the so called *critical dimension* p of the full parameter for which the BvM result is applicable. In the important i.i.d. case, we show that the condition “ p^3/n is small” is sufficient for BvM result to be valid under general assumptions on the model. We also provide an example of a model with the phase transition effect: the statement of the BvM theorem fails when the dimension p approaches $n^{1/3}$. The results are extended to the case of infinite dimensional parameters with the nuisance parameter from a Sobolev class. In particular we show near normality of the posterior if the smoothness parameter s exceeds $3/2$.

AMS 2000 Subject Classification: Primary 62F15. Secondary 62F25, 62H12

Keywords: prior, posterior, Bayesian inference, semiparametric, critical dimension

1 Introduction

The prominent Bernstein – von Mises (BvM) theorem claims that the posterior measure is asymptotically normal with the mean close to the maximum likelihood estimator (MLE) and the posterior variance is nearly the inverse of the total Fisher information matrix. The BvM result provides a theoretical background for Bayesian computations of the MLE and its variance. Also it justifies usage of elliptic credible sets based on the first two moments of the posterior. The classical version of the BvM Theorem is stated for the standard parametric setup with a fixed parametric model and large samples; see Le Cam and Yang (1990); van der Vaart (2000) for a detailed overview. However, in modern statistic applications one often faces very complicated models involving a lot of parameters and with a limited sample size. This requires an extension of the classical results to such non-classical situation. We mention Cox (1993); Freedman (1999); Ghosal (1999); Johnstone (2010) and references therein for some special phenomena arising in the Bayesian analysis when the parameter dimension increases. Already consistency of

the posterior distribution in nonparametric and semiparametric models is a nontrivial problem; cf. Schwartz (1965) and Barron et al. (1996). Asymptotic normality of the posterior measure for these classes of models is even more challenging; see e.g. Shen (2002). Some results for particular semi and nonparametric problems are available from Kim and Lee (2004); Kim (2006); Leahu (2011); Castillo and Nickl (2013). Cheng and Kosorok (2008) obtained a version of the BvM statement based on a high order expansion of the profile sampler. The recent paper Bickel and Kleijn (2012) extends the BvM statement from the classical parametric case to a rather general i.i.d. framework. Castillo (2012) studies the semiparametric BvM result for Gaussian process functional priors. In Rivoirard and Rousseau (2012) semiparametric BvM theorem is derived for linear functionals of density and in forthcoming work Castillo and Rousseau (2013) the result is generalized to a broad class of models and functionals. However, all these results are limited to the asymptotic setup and to some special classes of models like i.i.d. or Gaussian.

In this paper we reconsider the BvM result for the parametric component of a general semiparametric model. An important feature of the study is that the sample size is fixed, we proceed with just one sample. A finite sample theory is especially challenging because the most of notions, methods and tools in the classical theory are formulated in the asymptotic setup with the growing sample size. Only few finite sample general results are available; see e.g. the recent paper Boucheron and Massart (2011). This paper focuses on the semiparametric problem when the full parameter is large or infinite dimensional but the target is low dimensional. In the Bayesian framework, the aim is the marginal of the posterior corresponding to the target parameter; cf. Castillo (2012). Typical examples are provided by functional estimation, estimation of a function at a point, or simply by estimating a given subvector of the parameter vector. An interesting feature of the semiparametric BvM result is that the nuisance parameter appears only via the effective score and the related efficient Fisher information; cf. Bickel and Kleijn (2012). The methods of study heavily rely on the notion of the hardest parametric sub-model. In addition, one assumes that an estimate of the nuisance parameter is available which ensures a certain accuracy of estimation; see Cheng and Kosorok (2008) or Bickel and Kleijn (2012). This essentially simplifies the study but does not allow to derive a qualitative relation between the full dimension of the parameter space and the total available information in the data.

Some recent results study the impact of a growing parameter dimension p_n on the quality of Gaussian approximation of the posterior. We mention Ghosal (1999, 2000), Boucheron and Gassiat (2009), Johnstone (2010) and Bontemps (2011) for specific examples. See the discussion after Theorem 4.1 below for more details.

In this paper we show that the *bracketing* approach of Spokoiny (2012) can be used for obtaining a finite sample semiparametric version of Bernstein – von Mises theorem even if the full parameter dimension grows with the sample size. The ultimate goal of this paper is to quantify the so called critical parameter dimension for which the BvM result can be applied. Our approach neither relies on a pilot estimate of the nuisance and target parameter nor it involves the notion of the hardest parametric submodel. In the case of finite dimensional nuisance the obtained results only require some smoothness of the log-likelihood function, its finite exponential moments, and some identifiability conditions. Further we specify this result to the i.i.d. setup and show that the imposed conditions are satisfied if p_n^3/n is small. We present an example showing that the dimension $p_n = O(n^{1/3})$ is indeed critical and the BvM result starts to fail if p_n grows over $n^{1/3}$. If the nuisance is infinite dimensional then additionally some smoothness of nonparametric part is required. We state the general BvM results and show that the smoothness $s > 3/2$ is sufficient for the validity of the BvM result for linear and generalized linear models with the nuisance parameter from Sobolev class.

Now we describe our setup. Let \mathbf{Y} denote the observed random data, and \mathbb{P} denote the data distribution. The parametric statistical model assumes that the unknown data distribution \mathbb{P} belongs to a given parametric family $(\mathbb{P}_{\mathbf{v}})$:

$$\mathbf{Y} \sim \mathbb{P} = \mathbb{P}_{\mathbf{v}^*} \in (\mathbb{P}_{\mathbf{v}}, \mathbf{v} \in \mathcal{T}),$$

where \mathcal{T} is some parameter space and $\mathbf{v}^* \in \mathcal{T}$ is the true value of parameter. In the semiparametric framework, one attempts to recover only a low dimensional component $\boldsymbol{\theta}$ of the whole parameter \mathbf{v} . This means that the target of estimation is

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \Pi_0 \mathbf{v}^*,$$

for some mapping $\Pi_0 : \mathcal{T} \rightarrow \mathbb{R}^q$, and $q \in \mathbb{N}$ stands for the dimension of the target. Usually in the classical semiparametric setup, the vector \mathbf{v} is represented as $\mathbf{v} = (\boldsymbol{\theta}, \boldsymbol{\eta})$, where $\boldsymbol{\theta}$ is the target of analysis while $\boldsymbol{\eta}$ is the *nuisance parameter*. We refer to this situation as $(\boldsymbol{\theta}, \boldsymbol{\eta})$ -setup and our presentation follows this setting. An extension to the \mathbf{v} -setup with $\boldsymbol{\theta} = \Pi_0 \mathbf{v}$ is straightforward. Also for simplicity we first develop our results for the case when the total parameter space \mathcal{T} is a subset of the Euclidean space of dimensionality p .

Another issue addressed in this paper is the model misspecification. In the most of practical problems, it is unrealistic to expect that the model assumptions are exactly fulfilled, even if some rich nonparametric models are used. This means that the true data distribution \mathbb{P} does not belong to the considered family $(\mathbb{P}_{\mathbf{v}}, \mathbf{v} \in \mathcal{T})$. The “true”

value \mathbf{v}^* of the parameter \mathbf{v} can be defined by

$$\mathbf{v}^* = \operatorname{argmax}_{\mathbf{v} \in \mathcal{Y}} \mathbb{E} \mathcal{L}(\mathbf{v}), \quad (1.1)$$

where $\mathcal{L}(\mathbf{v}) = \log \frac{d\mathbb{P}_{\mathbf{v}}}{d\mu_0}(\mathbf{Y})$ is the log-likelihood function of the family $(\mathbb{P}_{\mathbf{v}})$ for some dominating measure μ_0 . Under model misspecification, \mathbf{v}^* defines the best parametric fit to \mathbb{P} by the considered family; cf. Chernozhukov and Hong (2003), Kleijn and van der Vaart (2006, 2012) and references therein. The target $\boldsymbol{\theta}^*$ is defined by the mapping Π_0 :

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \Pi_0 \mathbf{v}^*.$$

Now we switch to the Bayesian set-up. Let π be a prior measure on the parameter set \mathcal{Y} . Below we study the properties of the posterior measure which is the random measure on \mathcal{Y} describing the conditional distribution of \mathbf{v} given \mathbf{Y} and obtained by normalization of the product $\exp\{\mathcal{L}(\mathbf{v})\}\pi(d\mathbf{v})$. This relation is usually written as

$$\mathbf{v} \mid \mathbf{Y} \propto \exp\{\mathcal{L}(\mathbf{v})\} \pi(d\mathbf{v}). \quad (1.2)$$

An important feature of our analysis is that $\mathcal{L}(\mathbf{v})$ is not assumed to be the true log-likelihood. This means that a model misspecification is possible and the underlying data distribution can be beyond the considered parametric family. In this sense, the Bayes formula (1.2) describes a *quasi posterior*; Chernozhukov and Hong (2003). Below we show that smoothness of the log-likelihood function $\mathcal{L}(\mathbf{v})$ ensures a kind of a Gaussian approximation of the posterior measure. Our focus is to describe the accuracy of such approximation as a function of the parameter dimension p and the other important characteristics of the model.

We suppose that the prior measure π has a positive density $\pi(\mathbf{v})$ w.r.t. to the Lebesgue measure on \mathcal{Y} : $\pi(d\mathbf{v}) = \pi(\mathbf{v})d\mathbf{v}$. Then (1.2) can be written as

$$\mathbf{v} \mid \mathbf{Y} \propto \exp\{\mathcal{L}(\mathbf{v})\} \pi(\mathbf{v}). \quad (1.3)$$

The famous Bernstein – von Mises (BvM) theorem claims that the posterior centered by any efficient estimator $\tilde{\mathbf{v}}$ of the parameter \mathbf{v}^* (for example MLE) and scaled by the total Fisher information matrix is nearly standard normal:

$$\mathcal{D}_0(\mathbf{v} - \tilde{\mathbf{v}}) \mid \mathbf{Y} \xrightarrow{w} \mathcal{N}(0, \mathbf{I}_p),$$

where \mathbf{I}_p is an identity matrix of dimension p .

An important feature of the posterior distribution is that it is entirely known and can be numerically assessed. If we know in addition that the posterior is nearly normal, it

suffices to compute its mean and variance for building the concentration and credible sets. The BvM result does not require prior distribution to be proper and the phenomenon can be observed in the case of improper priors as well (for examples, see Bochkina and Green (2012)).

In this work we investigate the properties of the posterior distribution for the target parameter $\boldsymbol{\vartheta} = \Pi_0 \boldsymbol{v}$. In this case (1.3) can be written as

$$\boldsymbol{\vartheta} \mid \boldsymbol{Y} \propto \int \exp\{\mathcal{L}(\boldsymbol{v})\} \pi(\boldsymbol{v}) d\boldsymbol{\eta}. \quad (1.4)$$

The BvM result in this case transforms into

$$\check{D}_0(\boldsymbol{\vartheta} - \tilde{\boldsymbol{\theta}}) \mid \boldsymbol{Y} \xrightarrow{w} \mathcal{N}(0, I_q),$$

where I_q is an identity matrix of dimension q , $\tilde{\boldsymbol{\theta}} = \Pi_0 \tilde{\boldsymbol{v}}$, and \check{D}_0^2 is given in (2.2).

We consider two important classes of priors, namely non-informative and flat priors. Our goal is to show under mild conditions that the posterior distribution of the target parameter (1.4) is close to a prescribed Gaussian law even for finite samples. The other important issue is to specify the conditions on the sample size and the dimension of the parameter space for which the BvM result is still applicable.

2 BvM Theorem with a finite dimensional nuisance

This section presents our main results for the case of a finite dimensional parameter \boldsymbol{v} , i.e. $\dim(\mathcal{T}) = p < \infty$. One of the main elements of our construction is a $p \times p$ matrix \mathcal{D}_0^2 which is defined similarly to the Fisher information matrix:

$$\mathcal{D}_0^2 \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} \mathcal{L}(\boldsymbol{v}^*). \quad (2.1)$$

Here and in what follows we implicitly assume that the log-likelihood function $\mathcal{L}(\boldsymbol{v})$ is sufficiently smooth in \boldsymbol{v} , $\nabla \mathcal{L}(\boldsymbol{v})$ stands for its gradient and $\nabla^2 \mathbb{E} \mathcal{L}(\boldsymbol{v})$ for the Hessian of the expectation $\mathbb{E} \mathcal{L}(\boldsymbol{v})$, and the true value \boldsymbol{v}^* is due to (1.1). Also define the score

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} \mathcal{D}_0^{-1} \nabla \mathcal{L}(\boldsymbol{v}^*).$$

The definition of \boldsymbol{v}^* implies $\nabla \mathbb{E} \mathcal{L}(\boldsymbol{v}^*) = 0$ and hence, $\mathbb{E} \boldsymbol{\xi} = 0$.

For the $(\boldsymbol{\theta}, \boldsymbol{\eta})$ -setup, we consider the block representation of the vector $\nabla \mathcal{L}(\boldsymbol{v}^*)$ and of the matrix and \mathcal{D}_0^2 from (2.1):

$$\nabla \mathcal{L}(\boldsymbol{v}^*) = \begin{pmatrix} \nabla_{\boldsymbol{\theta}} \\ \nabla_{\boldsymbol{\eta}} \end{pmatrix}, \quad \mathcal{D}_0^2 = \begin{pmatrix} D_0^2 & A_0 \\ A_0^\top & H_0^2 \end{pmatrix}.$$

Define also the $q \times q$ matrix \check{D}_0^2 and random vectors $\check{\nabla}_\theta, \check{\xi} \in \mathbb{R}^q$ as

$$\begin{aligned}\check{D}_0^2 &\stackrel{\text{def}}{=} D_0^2 - A_0 H_0^{-2} A_0^\top, \\ \check{\nabla}_\theta &\stackrel{\text{def}}{=} \nabla_\theta - A_0 H_0^{-2} \nabla_\eta, \\ \check{\xi} &\stackrel{\text{def}}{=} \check{D}_0^{-1} \check{\nabla}_\theta.\end{aligned}\tag{2.2}$$

The $q \times q$ matrix \check{D}_0^2 is usually called the efficient Fisher information matrix, while the random vector $\check{\xi} \in \mathbb{R}^q$ is the efficient score. Everywhere in the text for a vector \mathbf{a} we denote by $\|\mathbf{a}\|$ its Euclidean norm and for a matrix A we denote by $\|A\|$ its operator norm.

2.1 Conditions

Our results assume a number of conditions to be satisfied. The list is essentially as in Spokoiny (2012), one can find there some discussion and examples showing that the conditions are not restrictive and are fulfilled in most of classical models used in statistical studies like i.i.d., regression or Generalized Linear models. The conditions are split into local and global. The local conditions only describe the properties of the process $\mathcal{L}(\mathbf{v})$ for $\mathbf{v} \in \mathcal{V}_0(\mathbf{r}_0)$ with some fixed value \mathbf{r}_0 :

$$\mathcal{V}_0(\mathbf{r}_0) \stackrel{\text{def}}{=} \{\mathbf{v} \in \mathcal{V} : \|\mathcal{D}_0(\mathbf{v} - \mathbf{v}^*)\| \leq \mathbf{r}_0\}.$$

The global conditions have to be fulfilled on the whole \mathcal{V} . Define the stochastic component $\zeta(\mathbf{v})$ of $\mathcal{L}(\mathbf{v})$:

$$\zeta(\mathbf{v}) \stackrel{\text{def}}{=} \mathcal{L}(\mathbf{v}) - \mathbb{E}\mathcal{L}(\mathbf{v}).$$

We start with some exponential moments conditions.

(ED₀) There exists a constant $\nu_0 > 0$, a positive symmetric $p \times p$ matrix \mathcal{V}_0^2 satisfying $\text{Var}\{\nabla\zeta(\mathbf{v}^*)\} \leq \mathcal{V}_0^2$, and a constant $\mathbf{g} > 0$ such that

$$\sup_{\gamma \in \mathbb{R}^p} \log \mathbb{E} \exp \left\{ \mu \frac{\langle \nabla\zeta(\mathbf{v}^*), \gamma \rangle}{\|\mathcal{V}_0 \gamma\|} \right\} \leq \frac{\nu_0^2 \mu^2}{2}, \quad |\mu| \leq \mathbf{g}.$$

(ED₂) There exists a constant $\omega > 0$ and for each $\mathbf{r} > 0$ a constant $\mathbf{g}(\mathbf{r}) > 0$ such that for all $\mathbf{v} \in \mathcal{V}_0(\mathbf{r})$:

$$\sup_{\gamma_1, \gamma_2 \in \mathbb{R}^p} \log \mathbb{E} \exp \left\{ \frac{\mu}{\omega} \frac{\gamma_1^\top \nabla^2 \zeta(\mathbf{v}) \gamma_2}{\|\mathcal{D}_0 \gamma_1\| \cdot \|\mathcal{D}_0 \gamma_2\|} \right\} \leq \frac{\nu_0^2 \mu^2}{2}, \quad |\mu| \leq \mathbf{g}(\mathbf{r}).$$

The next condition is needed to ensure some smoothness properties of expected log-likelihood $\mathbb{E}\mathcal{L}(\mathbf{v})$ in the local zone $\mathbf{v} \in \mathcal{V}_0(\mathbf{r}_0)$. Define

$$\mathcal{D}_0^2(\mathbf{v}) \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E}\mathcal{L}(\mathbf{v}).$$

Then $\mathcal{D}_0^2 = \mathcal{D}_0^2(\mathbf{v}^*)$.

(\mathcal{L}_0) There exists a constant $\delta(\mathbf{r})$ such that it holds on the set $\mathcal{V}_0(\mathbf{r})$ for all $\mathbf{r} \leq \mathbf{r}_0$

$$|\mathcal{D}_0^{-1} \mathcal{D}_0^2(\mathbf{v}) \mathcal{D}_0^{-1} - I_p| \leq \delta(\mathbf{r}).$$

The global identification condition is:

(\mathcal{L}_r) For any \mathbf{r} there exists a value $\mathbf{b}(\mathbf{r}) > 0$, such that $\mathbf{r}\mathbf{b}(\mathbf{r}) \rightarrow \infty$, $\mathbf{r} \rightarrow \infty$ and

$$-\mathbb{E}\mathcal{L}(\mathbf{v}, \mathbf{v}^*) \geq \mathbf{r}^2 \mathbf{b}(\mathbf{r}) \quad \text{for all } \mathbf{v} \text{ with } \mathbf{r} = \|\mathcal{D}_0(\mathbf{v} - \mathbf{v}^*)\|.$$

Finally we specify the identifiability conditions. We begin by representing the information and the covariance matrices in block form:

$$\mathcal{D}_0^2 = \begin{pmatrix} D_0^2 & A_0 \\ A_0^\top & H_0^2 \end{pmatrix}, \quad \mathcal{V}_0^2 = \begin{pmatrix} V_0^2 & B_0 \\ B_0^\top & Q_0^2 \end{pmatrix}.$$

The *identifiability conditions* in Spokoiny (2012) ensure that the matrix \mathcal{D}_0^2 is positive and satisfied $\mathfrak{a}^2 \mathcal{D}_0^2 \geq \mathcal{V}_0^2$ for some $\mathfrak{a} > 0$. Here we restate these conditions in the special block form which is specific for the $(\boldsymbol{\theta}, \boldsymbol{\eta})$ -setup.

(\mathcal{I}) There are constants $\mathfrak{a} > 0$ and $\nu < 1$ such that

$$\mathfrak{a}^2 D_0^2 \geq V_0^2, \quad \mathfrak{a}^2 H_0^2 \geq Q_0^2, \quad \mathfrak{a}^2 \mathcal{D}_0^2 \geq \mathcal{V}_0^2. \quad (2.3)$$

and

$$\|D_0^{-1} A_0 H_0^{-2} A_0^\top D_0^{-1}\| \leq \nu. \quad (2.4)$$

The quantity ν bounds the angle between the target and nuisance subspaces in the tangent space. The regularity condition (\mathcal{I}) ensures that this angle is not too small and hence, the target and nuisance parameters are identifiable. In particular, the matrix \check{D}_0^2 from (2.2) is well posed under (\mathcal{I}). The bounds in (2.3) are given with the same constant \mathfrak{a} only for simplifying the notation. One can show that the last bound on \mathcal{D}_0^2 follows from the first two and (2.4) with another constant \mathfrak{a}' depending on \mathfrak{a} and ν only.

2.2 The main results

First we state the BvM result about the properties of the $\boldsymbol{\vartheta}$ -posterior given by (1.4) in case of uniform prior that is, $\pi(\boldsymbol{v}) \equiv 1$ on \mathcal{Y} . Define

$$\bar{\boldsymbol{\vartheta}} \stackrel{\text{def}}{=} \mathbb{E}(\boldsymbol{\vartheta} \mid \boldsymbol{Y}), \quad \mathfrak{S}^2 \stackrel{\text{def}}{=} \text{Cov}(\boldsymbol{\vartheta} \mid \boldsymbol{Y}) \stackrel{\text{def}}{=} \mathbb{E}\{(\boldsymbol{\vartheta} - \bar{\boldsymbol{\vartheta}})(\boldsymbol{\vartheta} - \bar{\boldsymbol{\vartheta}})^\top \mid \boldsymbol{Y}\}. \quad (2.5)$$

Also define

$$\boldsymbol{\theta}^\circ \stackrel{\text{def}}{=} \boldsymbol{\theta}^* + \check{D}_0^{-1} \check{\boldsymbol{\xi}}.$$

This random point can be viewed as first order approximation of the profile MLE $\tilde{\boldsymbol{\theta}}$. Below we present a version of the BvM result in the considered nonasymptotic setup which claims that $\bar{\boldsymbol{\vartheta}}$ is close to $\boldsymbol{\theta}^\circ$, \mathfrak{S}^2 is nearly equal to \check{D}_0^{-2} , and $\check{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ)$ is nearly standard normal conditionally on \boldsymbol{Y} .

We suppose that a large constant \mathbf{x} is fixed which specifies random events $\Omega(\mathbf{x})$ of *dominating probability*. We say that a generic random set $\Omega(\mathbf{x})$ is of dominating probability if

$$\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - \mathbb{C}e^{-\mathbf{x}}.$$

The notation \mathbb{C} for a generic absolute constant and \mathbf{x} for a positive value ensuring that $e^{-\mathbf{x}}$ is negligible. By $\Omega(\mathbf{x})$ we denote a random event of dominating probability with $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - \mathbb{C}e^{-\mathbf{x}}$. The exact values of \mathbb{C} will be specified in each particular case. The formulation of the results also involve the radius \mathbf{r}_0 and the spread $\Delta(\mathbf{r}_0, \mathbf{x})$. The radius \mathbf{r}_0 separates the local zone $\mathcal{Y}_\circ(\mathbf{r}_0)$ which is a vicinity of the central point \boldsymbol{v}_0 , and its complement $\mathcal{Y} \setminus \mathcal{Y}_\circ(\mathbf{r}_0)$ for which we establish a large deviation result. The *spread* value $\Delta(\mathbf{r}_0, \mathbf{x})$ measures the quality of local approximation of the log-likelihood $L(\boldsymbol{v}, \boldsymbol{v}_0)$ by a quadratic process $\mathbb{L}(\boldsymbol{v}, \boldsymbol{v}_0)$:

$$\Delta(\mathbf{r}_0, \mathbf{x}) \stackrel{\text{def}}{=} \{\delta(\mathbf{r}_0) + 6\nu_0 z_{\mathbb{H}}(\mathbf{x})\omega\} \mathbf{r}_0^2.$$

Here the term $\delta(\mathbf{r}_0)\mathbf{r}_0^2$ measures the error of a quadratic approximation of the expected log-likelihood $L(\boldsymbol{v})$ due to (\mathcal{L}_0) , while the second term $6\nu_0 z_{\mathbb{H}}(\mathbf{x})\omega \mathbf{r}_0^2$ controls the stochastic term and involves the entropy of the parameter space which is involved in the definition of $z_{\mathbb{H}}(\mathbf{x})$. A precise formulation is given in Theorem 6.1 below.

Theorem 2.1. *Suppose the conditions of Section 2.1. Let the prior be uniform on \mathcal{Y} . Then there exists a random event $\Omega(\mathbf{x})$ of probability at least $1 - 4e^{-\mathbf{x}}$ such that it holds*

on $\Omega(\mathbf{x})$

$$\begin{aligned}\|\check{D}_0(\bar{\boldsymbol{\vartheta}} - \boldsymbol{\theta}^\circ)\|^2 &\leq 4\Delta(\mathbf{r}_0, \mathbf{x}) + 16e^{-\mathbf{x}}, \\ \|I_q - \check{D}_0\mathfrak{S}^2\check{D}_0\| &\leq 4\Delta(\mathbf{r}_0, \mathbf{x}) + 16e^{-\mathbf{x}},\end{aligned}$$

where $\bar{\boldsymbol{\vartheta}}$ and \mathfrak{S}^2 are from (2.5).

Moreover, on $\Omega(\mathbf{x})$ for any measurable set $A \subset \mathbb{R}^q$

$$\begin{aligned}\exp(-2\Delta(\mathbf{r}_0, \mathbf{x}) - 8e^{-\mathbf{x}})\mathbb{P}(\boldsymbol{\gamma} \in A) - e^{-\mathbf{x}} \\ \leq \mathbb{P}(\check{D}_0(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}}) \in A \mid \mathbf{Y}) \\ \leq \exp(2\Delta(\mathbf{r}_0, \mathbf{x}) + 5e^{-\mathbf{x}})\mathbb{P}(\boldsymbol{\gamma} \in A),\end{aligned}$$

where $\boldsymbol{\gamma}$ is a standard Gaussian vector in \mathbb{R}^q .

The condition “ $\Delta(\mathbf{r}_0, \mathbf{x})$ is small” yields the desirable BvM result, that is, the posterior measure after centering and standardization is close in total variation to the standard normal law. The classical asymptotic results immediately follow for many classical models (see discussion in Section 4). The next corollary extends the previous result by using empirically computable objects.

Corollary 2.2. *Under the conditions of Theorem 2.1 for any measurable set $A \subset \mathbb{R}^q$ a random event $\Omega(\mathbf{x})$ of a dominating probability at least $1 - 4e^{-\mathbf{x}}$*

$$\begin{aligned}\exp(-2\Delta(\mathbf{r}_0, \mathbf{x}) - 8e^{-\mathbf{x}})\{\mathbb{P}(\boldsymbol{\gamma} \in A) - \tau\} - e^{-\mathbf{x}} \\ \leq \mathbb{P}(\mathfrak{S}^{-1}(\boldsymbol{\vartheta} - \bar{\boldsymbol{\vartheta}}) \in A \mid \mathbf{Y}) \\ \leq \exp(2\Delta(\mathbf{r}_0, \mathbf{x}) + 5e^{-\mathbf{x}})\{\mathbb{P}(\boldsymbol{\gamma} \in A) + \tau\},\end{aligned}$$

where $\boldsymbol{\gamma}$ is a standard Gaussian vector in \mathbb{R}^q and

$$\tau \stackrel{\text{def}}{=} \frac{1}{2} \left(q\Delta^2(\mathbf{r}_0, \mathbf{x}) + \{1 + \Delta(\mathbf{r}_0, \mathbf{x})\}^2 \Delta^2(\mathbf{r}_0, \mathbf{x}) \right).$$

This corollary is important as in practical applications we do not know matrix \check{D}_0 and vector $\boldsymbol{\theta}^\circ$, but matrix \mathfrak{S}^{-1} and vector $\bar{\boldsymbol{\vartheta}}$ can be computed by numerical computations. If dimension q is fixed the result becomes informative under the condition “ $\Delta(\mathbf{r}_0, \mathbf{x})$ is small”. Moreover, the statement can be extended to situations when the target dimension q grows but $\Delta(\mathbf{r}_0, \mathbf{x})q^{1/2}$ is still small.

2.3 Extension of Theorem 2.1 to a flat prior

The results of Theorem 2.1 for a non-informative prior can be extended to the case of a general prior $\pi(d\mathbf{v})$ with a density $\pi(\mathbf{v})$ which is uniformly continuous and sufficiently flat on the local set $\mathcal{Y}_0(\mathbf{r}_0)$. More precisely, let $\pi(\mathbf{v})$ satisfy

$$\sup_{\mathbf{v} \in \mathcal{Y}_0(\mathbf{r}_0)} \left| \frac{\pi(\mathbf{v})}{\pi(\mathbf{v}^*)} - 1 \right| \leq \alpha(\mathbf{r}_0), \quad \sup_{\mathbf{v} \in \mathcal{Y}} \frac{\pi(\mathbf{v})}{\pi(\mathbf{v}^*)} \leq \mathbf{C}(\mathbf{r}_0),$$

where $\alpha(\mathbf{r}_0)$ is a small constant while $\mathbf{C}(\mathbf{r}_0)$ is any fixed constant. Then the results of Theorem 2.1 continue to apply with an obvious correction of the approximation error. Indeed, for any local set $A \subseteq \mathcal{Y}_0(\mathbf{r}_0)$ one can apply the bounds

$$\begin{aligned} \int_A \exp\{L(\mathbf{v})\} \pi(\mathbf{v}) d\mathbf{v} &\leq e^{\alpha(\mathbf{r}_0)} \pi(\mathbf{v}^*) \int_A \exp\{L(\mathbf{v})\} d\mathbf{v}, \\ \int_A \exp\{L(\mathbf{v})\} \pi(\mathbf{v}) d\mathbf{v} &\geq e^{-\alpha(\mathbf{r}_0)} \pi(\mathbf{v}^*) \int_A \exp\{L(\mathbf{v})\} d\mathbf{v}, \end{aligned}$$

This particularly implies for each $A \subset \mathcal{Y}_0(\mathbf{r}_0)$,

$$\mathbb{P}_\pi(A \mid \mathbf{Y}) \leq \exp\{2\alpha(\mathbf{r}_0)\} \mathbb{P}(A \mid \mathbf{Y}). \quad (2.6)$$

The tail probability of the complement $\mathcal{Y}_0^c(\mathbf{r}_0) \stackrel{\text{def}}{=} \mathcal{Y} \setminus \mathcal{Y}_0(\mathbf{r}_0)$ of $\mathcal{Y}_0(\mathbf{r}_0)$ can be enlarged by $\mathbf{C}(\mathbf{r}_0)$ relative to the uniform prior:

$$\int_{\mathcal{Y}_0^c(\mathbf{r}_0)} \exp\{L(\mathbf{v})\} \pi(\mathbf{v}) d\mathbf{v} \leq \mathbf{C}(\mathbf{r}_0) \pi(\mathbf{v}^*) \int_{\mathcal{Y}_0^c(\mathbf{r}_0)} \exp\{L(\mathbf{v})\} d\mathbf{v},$$

hence

$$\mathbb{P}_\pi(\mathcal{Y}_0^c(\mathbf{r}_0) \mid \mathbf{Y}) \leq \mathbf{C}(\mathbf{r}_0) \mathbb{P}(\mathcal{Y}_0^c(\mathbf{r}_0) \mid \mathbf{Y}). \quad (2.7)$$

In particular, if the tail of the non-informative posterior satisfies $\mathbb{P}(\mathcal{Y}_0^c(\mathbf{r}_0) \mid \mathbf{Y}) \leq e^{-x}$, then $\mathbb{P}_\pi(\mathcal{Y}_0^c(\mathbf{r}_0) \mid \mathbf{Y}) \leq \mathbf{C}(\mathbf{r}_0) e^{-x}$.

Theorem 2.3. *Suppose the conditions of Theorem 2.1. Let also $\Pi = \mathcal{N}(0, G^{-2})$ be a Gaussian prior measure on \mathbb{R}^q such that*

$$\|\mathcal{D}_0^{-1} G^2 \mathcal{D}_0^{-1}\| \leq \varepsilon^2, \quad (2.8)$$

where ε is a given constant. Then (2.6) and (2.7) hold with $\mathbf{C}(\mathbf{r}_0) \leq \exp(\|G\mathbf{v}^*\|^2/2)$ and $\alpha(\mathbf{r}_0) = \max\{\varepsilon \mathbf{r}_0 \|G\mathbf{v}^*\|, \varepsilon^2 \mathbf{r}_0^2/2\}$.

The result of Theorem 2.3 tells us that the BvM result holds if the prior distribution is sufficiently flat.

3 Infinite dimensional nuisance

This section describes how previous results can be extended to the case where the nuisance is infinite dimensional. More specifically, we consider $(\boldsymbol{\theta}, \mathbf{f})$ -setup, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^q$ and $\mathbf{f} \in \mathcal{H}$ for some Hilbert space \mathcal{H} . Suppose that in \mathcal{H} exists a countable basis $\mathbf{e}_1, \mathbf{e}_2, \dots$. Then

$$\mathbf{f} = \mathbf{f}(\boldsymbol{\eta}) = \sum_{j=1}^{\infty} \eta_j \mathbf{e}_j \in \mathcal{H},$$

where a vector $\boldsymbol{\eta} = \{\eta_j\}_{j=1}^{\infty} \in \ell_2$ and $\eta_j = \langle \mathbf{f}, \mathbf{e}_j \rangle$.

Let the likelihood for the full model be $\mathcal{L}(\boldsymbol{\theta}, \mathbf{f})$. Denote with $\mathbf{v} = (\boldsymbol{\theta}, \boldsymbol{\eta})$

$$\mathcal{L}(\mathbf{v}) = \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \mathcal{L}(\boldsymbol{\theta}, \mathbf{f}(\boldsymbol{\eta})).$$

The underlying full and target parameters can be defined by maximizing the expected log-likelihood:

$$\mathbf{v}^* = \underset{\mathbf{v}=(\boldsymbol{\theta}, \boldsymbol{\eta})}{\operatorname{argmax}} \mathbb{E} \mathcal{L}(\mathbf{v}), \quad \boldsymbol{\theta}^* = \Pi_0 \mathbf{v}^*.$$

Also define the information matrix \mathcal{D}_0^2 for the full parameter \mathbf{v} and the efficient information matrix $\check{\mathcal{D}}_0^2$ for the target $\boldsymbol{\theta}$:

$$\begin{aligned} \mathcal{D}_0^2 &\stackrel{\text{def}}{=} \nabla^2 \mathbb{E}[\mathcal{L}(\mathbf{v}^*)] \in \operatorname{Lin}(\ell_2, \ell_2), \\ \check{\mathcal{D}}_0^2 &\stackrel{\text{def}}{=} \left(\Pi_0 \mathcal{D}_0^{-2} \Pi_0^\top \right)^{-1} \in \mathbb{R}^{q \times q}, \end{aligned}$$

where $\operatorname{Lin}(\ell_2, \ell_2)$ is the space of linear operators from ℓ_2 to ℓ_2 .

We apply the sieve approach and use an uninformative finite dimensional prior for the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}_m$. Consider a finite dimensional approximations $\boldsymbol{\eta}_m = \{\eta_j\}_{j=1}^m$ of the full parameter $\boldsymbol{\eta}$ and the corresponding finite dimensional approximation of the log-likelihood:

$$\mathcal{L}(\mathbf{v}_m) = \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}_m), \quad \mathbf{v}_m = (\boldsymbol{\theta}, \boldsymbol{\eta}_m).$$

Similarly to the finite dimensional case, introduce the quantities

$$\begin{aligned} \mathbf{v}_m^* &= \underset{\mathbf{v}_m \in \Upsilon_m}{\operatorname{argmax}} \mathbb{E} \mathcal{L}(\mathbf{v}_m), \\ \mathcal{D}_m^2 &\stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} \mathcal{L}(\mathbf{v}_m^*). \end{aligned} \tag{3.1}$$

Here $\nabla \mathcal{L}(\mathbf{v}_m)$ stands for its gradient and $\nabla^2 \mathbb{E} \mathcal{L}(\mathbf{v}_m)$ for the Hessian of the expectation $\mathbb{E} \mathcal{L}(\mathbf{v}_m)$. Also define the score vector

$$\boldsymbol{\xi}_m \stackrel{\text{def}}{=} \mathcal{D}_m^{-1} \nabla \mathcal{L}(\mathbf{v}_m^*).$$

The definition of \mathbf{v}_m^* implies $\nabla \mathbb{E}\mathcal{L}(\mathbf{v}_m^*) = 0$ and hence, $\mathbb{E}\boldsymbol{\xi}_m = 0$.

Again, we consider the block representation of the vector $\nabla \mathcal{L}(\mathbf{v}_m^*)$ and of the matrix and \mathcal{D}_m^2 from (3.1):

$$\nabla \mathcal{L}(\mathbf{v}_m^*) = \begin{pmatrix} \nabla_{\boldsymbol{\theta}} \\ \nabla_{\boldsymbol{\eta}_m} \end{pmatrix}, \quad \mathcal{D}_m^2 = \begin{pmatrix} D_m^2 & A_m \\ A_m^\top & H_m^2 \end{pmatrix}.$$

Define also the $q \times q$ matrix \check{D}_m^2 and random vectors $\check{\nabla}_{\boldsymbol{\theta},m}, \check{\boldsymbol{\xi}}_m \in \mathbb{R}^q$ as

$$\begin{aligned} \check{D}_m^2 &= D_m^2 - A_m H_m^{-2} A_m^\top, \\ \check{\nabla}_{\boldsymbol{\theta},m} &= \nabla_{\boldsymbol{\theta}} - A_m H_m^{-2} \nabla_{\boldsymbol{\eta}_m}, \\ \check{\boldsymbol{\xi}}_m &= \check{D}_m^{-1} \check{\nabla}_{\boldsymbol{\theta},m}. \end{aligned}$$

Approximation of the functional parameter \mathbf{v} by \mathbf{v}_m leads to two sources of bias. The first one is connected with approximation of the target parameter $\boldsymbol{\theta}^* - \boldsymbol{\theta}_m^*$. The second one is due to the difference between the efficient Fisher information $\check{D}_0^2 \in \mathbb{R}^{q \times q}$ and its approximation $\check{D}_m^2 \in \mathbb{R}^{q \times q}$. Both errors of approximation are due to projection of the functional parameter onto the finite dimensional space spanned by the first m basis functions. The bias terms can be bounded under smoothness assumptions on the model and on the functional nuisance parameter \mathbf{f} using the standard methods of approximation theory. To avoid tedious calculus we simply assume a kind of consistency of the sieve approximation.

(B) For any $m \in \mathbb{N}$ there exist constants α_m and β_m such that

$$\begin{aligned} \|\check{D}_0(\boldsymbol{\theta}^* - \boldsymbol{\theta}_m^*)\|^2 &\leq \alpha_m, \\ \|I_q - \check{D}_0 \check{D}_m^{-2} \check{D}_0\| &\leq \beta_m. \end{aligned}$$

For validity of our results we will need that the value of m is fixed in a proper way ensuring that the values of α_m and β_m are sufficiently small. These values can be upper bounded under usual smoothness conditions on \mathbf{f} , e.g. if \mathbf{f} belongs to a Sobolev ball with a certain regularity; cf. Bontemps (2011); Bickel and Kleijn (2012); Castillo (2012). See also an example of computing the quantities α_m and β_m in Section 5.2 below.

Consider a non-informative sieve prior on $(\boldsymbol{\theta}, \boldsymbol{\eta}_m)$ and define

$$\begin{aligned} \bar{\boldsymbol{\vartheta}}_m &\stackrel{\text{def}}{=} \mathbb{E}(\boldsymbol{\vartheta} \mid \mathbf{Y}), \\ \mathfrak{S}_m^2 &\stackrel{\text{def}}{=} \text{Cov}(\boldsymbol{\vartheta} \mid \mathbf{Y}) \stackrel{\text{def}}{=} \mathbb{E}\{(\boldsymbol{\vartheta} - \bar{\boldsymbol{\vartheta}}_m)(\boldsymbol{\vartheta} - \bar{\boldsymbol{\vartheta}}_m)^\top \mid \mathbf{Y}\}. \end{aligned}$$

Also define

$$\boldsymbol{\theta}_m^\circ \stackrel{\text{def}}{=} \boldsymbol{\theta}_m^* + \check{D}_m^{-1} \check{\boldsymbol{\xi}}_m.$$

Now we are ready to state semiparametric version of Theorem 2.1.

Theorem 3.1. *Suppose the conditions of Section 2.1 and condition (B). Consider a non-informative prior on $(\boldsymbol{\theta}, \boldsymbol{\eta}_m)$. Then there exists a random event $\Omega(\mathbf{x})$ of a dominating probability at least $1 - 4e^{-\mathbf{x}}$ such that it holds on $\Omega(\mathbf{x})$*

$$\|\check{D}_0(\bar{\boldsymbol{\vartheta}}_m - \boldsymbol{\theta}_m^\circ)\|^2 \leq (1 + \beta_m)\Delta^*(\mathbf{r}_0, \mathbf{x}) + \alpha_m.$$

$$\|I_q - \check{D}_0 \mathfrak{S}_m^2 \check{D}_0\| \leq \beta_m + (1 + \beta_m)\Delta^*(\mathbf{r}_0, \mathbf{x}),$$

where $\Delta^*(\mathbf{r}_0, \mathbf{x}) = 4\Delta(\mathbf{r}_0, \mathbf{x}) + 16e^{-\mathbf{x}}$. Moreover, on $\Omega(\mathbf{x})$ for any $A \subset \mathbb{R}^q$

$$\begin{aligned} & \exp(-2\Delta(\mathbf{r}_0, \mathbf{x}) - 8e^{-\mathbf{x}}) \{P(\gamma \in A) - \tau\} - e^{-\mathbf{x}} \\ & \leq P(\mathfrak{S}_m^{-1}(\boldsymbol{\vartheta} - \bar{\boldsymbol{\vartheta}}_m) \in A \mid \mathbf{Y}) \\ & \leq \exp(2\Delta(\mathbf{r}_0, \mathbf{x}) + 5e^{-\mathbf{x}}) \{P(\gamma \in A) + \tau\}, \end{aligned}$$

where

$$\tau \stackrel{\text{def}}{=} \frac{1}{2} \sqrt{q \Delta^2(\mathbf{r}_0, \mathbf{x}) + \{1 + \Delta(\mathbf{r}_0, \mathbf{x})\}^2 \Delta^2(\mathbf{r}_0, \mathbf{x})}.$$

4 The i.i.d. case and critical dimension

This section comments how the previously obtained general results can be linked to the classical asymptotic results in the statistical literature. The nice feature of the whole approach based on the local bracketing is that all the results are stated under the same list of conditions: once checked one can directly apply any of the mentioned results. Typical examples include i.i.d., GLM, and median regression models. Here we briefly discuss how the BvM result can be applied to one typical case, namely, to an i.i.d. experiment.

Let $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ be an i.i.d. sample from a measure P . Here we suppose the conditions of Section 5.1 in Spokoiny (2012) on P and $(P_{\mathbf{v}})$ to be fulfilled. We admit that the parametric assumption $P \in (P_{\mathbf{v}}, \mathbf{v} \in \mathcal{T})$ can be misspecified and consider the asymptotic setup with the full dimension $p = p_n$ which depends on n and grows to infinity as $p_n \rightarrow \infty$.

Theorem 4.1. *Suppose the conditions of Section 5.1 in Spokoiny (2012). Let also $p_n \rightarrow \infty$ and $p_n^3/n \rightarrow 0$. Then the result of Theorem 2.1 holds with $\Delta(\mathbf{r}_0, \mathbf{x}) = \mathbf{c}\sqrt{p_n^3/n}$, $\mathcal{D}_0^2 = n\mathbb{F}_{\mathbf{v}^*}$, where $\mathbb{F}_{\mathbf{v}^*}$ is the Fisher information of $(P_{\mathbf{v}})$ at \mathbf{v}^* .*

A similar result about asymptotic normality of the posterior in a linear regression model can be found in Ghosal (1999). However, the convergence is proved under the condition $p_n^4 \log(p_n)/n \rightarrow 0$ which appears to be too strong. Ghosal (2000) showed that the dimensionality constraint can be relaxed to $p_n^3/n \rightarrow 0$ for exponential models with a product structure. Boucheron and Gassiat (2009) proved the BvM result in a specific class of i.i.d. model with discrete probability distribution under the condition $p_n^3/n \rightarrow 0$. Further examples and the related conditions for Gaussian models are presented in Johnstone (2010).

4.1 Critical dimension

This section discusses the issue of a *critical dimension*. Namely we show that the condition $p_n = o(n^{1/3})$ in Theorem 4.1 for the validity of the BvM result cannot be dropped or relaxed in a general situation. Namely, we present an example for which $p_n^3/n \geq \beta^2 > 0$ and the posterior distribution does not concentrate around MLE.

Let n and p_n be such that $M_n = n/p_n$ is an integer. We consider a simple Poissonian model with $Y_i \sim \text{Poisson}(v_j)$ for $i \in \mathcal{I}_j$, where $\mathcal{I}_j \stackrel{\text{def}}{=} \{i : \lceil i/M_n \rceil = j\}$ for $j = 1, \dots, p_n$ and $\lceil x \rceil$ is the nearest integer greater or equal to x . Let also $u_j = \log v_j$ be the canonical parameter. The log-likelihood $\mathcal{L}(\mathbf{u})$ with $\mathbf{u} = (u_1, \dots, u_{p_n})$ reads as

$$\mathcal{L}(\mathbf{u}) = \sum_{j=1}^{p_n} (Z_j u_j - M_n e^{u_j}),$$

where

$$Z_j \stackrel{\text{def}}{=} \sum_{i \in \mathcal{I}_j} Y_i.$$

We consider the problem of estimating the mean of the u_j 's:

$$\theta = \frac{1}{p_n} (u_1 + \dots + u_{p_n}).$$

Below we study this problem in the asymptotic setup with $p_n \rightarrow \infty$ as $n \rightarrow \infty$ when the underlying measure \mathbb{P} corresponds to $u_1^* = \dots = u_{p_n}^* = u^*$ for some u^* yielding $\theta^* = u^*$. The value u^* will be specified later. We consider an i.i.d. exponential prior on the parameters v_j of Poisson distribution:

$$v_j \sim \text{Exp}(\mu).$$

Below we allow that μ may depend on n . Our results are valid for $\mu \leq \mathbb{C} \sqrt{\frac{n}{\log n}}$. The posterior is Gamma distributed:

$$v_j \mid \mathbf{Y} \sim \text{Gamma}(\alpha_j, \mu_j),$$

where $\alpha_j = 1 + \sum_{i \in \mathcal{I}_j} Y_i$, $\mu_j = \frac{\mu}{M_n \mu + 1}$.

First we describe the profile maximum likelihood estimator $\tilde{\theta}_n$ of the target parameter θ . The MLE for the full parameter \mathbf{v} reads as $\tilde{\mathbf{v}} = (\tilde{v}_1, \dots, \tilde{v}_{p_n})^\top$ with

$$\tilde{v}_j = Z_j / M_n.$$

Thus, the profile MLE $\tilde{\theta}_n$ reads as

$$\tilde{\theta}_n = \frac{1}{p_n} \sum_{j=1}^{p_n} \log(\tilde{v}_j).$$

Furthermore, the efficient Fisher information \check{D}_0^2 is equal to $p_n^{-1}n$; see Lemma 7.6 below. As $\tilde{\theta}_n$ is the profile MLE it is an efficient with the asymptotic variance equal to \check{D}_0^{-2} .

Theorem 4.2. *Let $Y_i \sim \text{Poisson}(v^*)$ for all $i = 1, \dots, n$, $v^* = 1/p_n$. Then*

1. *If $p_n^3/n \rightarrow 0$ as $p_n \rightarrow \infty$, then*

$$p_n^{1/2} n^{1/2} (\theta - \tilde{\theta}_n) \mid \mathbf{Y} \xrightarrow{w} \mathcal{N}(0, 1).$$

2. *Let $p_n^3/n \equiv \beta > 0$. Then*

$$p_n^{1/2} n^{1/2} (\theta - \tilde{\theta}_n) \mid \mathbf{Y} \xrightarrow{w} \mathcal{N}(\beta/2, 1).$$

3. *If $p_n^3/n \rightarrow \infty$, but $p_n^4/n^{3/2} \rightarrow 0$, then*

$$p_n^{1/2} n^{1/2} (\theta - \tilde{\theta}_n) \mid \mathbf{Y} \xrightarrow{w} \infty.$$

We carried out a series of experiments to numerically demonstrate the results of Theorem 4.2. The dimension of parameter space was fixed $p_n = 10000$. Three cases were considered:

1. $p_n^{3/2}/n^{1/2} = \frac{1}{\log p_n}$, which corresponds to $p_n^3/n \rightarrow 0, n \rightarrow \infty$.
2. $p_n^{3/2}/n^{1/2} \equiv 1$.
3. $p_n^{3/2}/n^{1/2} = \log p_n$, which corresponds to $p_n^3/n \rightarrow \infty, n \rightarrow \infty$.

For each sample 10000 realizations of \mathbf{Y} were generated from the exponential distribution $\text{Exp}(v_*)$ and so were corresponding posterior values $\theta \mid \mathbf{Y}$. The resulting posterior distribution for three cases is demonstrated on Figure 4.1. It can be easily seen that results of Theorem 4.2 are numerically confirmed.

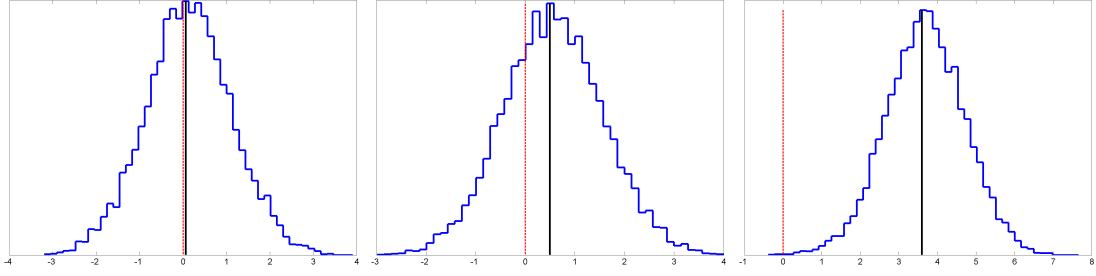


Figure 4.1: Posterior distribution of $\beta_n^{-1} p_n(\theta - \tilde{\theta}_n)$ for $\beta_n = 1/\log(p_n)$, $\beta_n = 1$, and $\beta_n = \log(p_n)$. Solid line is for posterior mean and dashed line is for true mean.

5 Examples

This section presents a number of examples illustrating the general results of Section 2.

5.1 Linear regression

Let $\mathbf{Y} = (Y_1, \dots, Y_n)$ be a random vector

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

where the errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$ are independent and the mean vector $\mathbf{f} \in \mathbb{R}^n$ is fixed. Let $\mathbf{f} = \Psi \mathbf{v}^*$, where $\Psi = \{\Psi_1, \dots, \Psi_p\}$ is a set of independent regressors and $\mathbf{v}^* \in \mathbb{R}^p$ is a vector of unknown parameters of the model.

First, we consider Gaussian case, i.e. $\varepsilon_i \in \mathcal{N}(0, \sigma_n^2 \mathbf{I}_n), i = 1, \dots, n$, with \mathbf{I}_n the $n \times n$ identity matrix. Variance of observations σ_n^2 is known but may depend on sample size n . The number of more specific settings are included in this model such as Gaussian sequence model and regression of a function in a Sobolev class. In this case approximation is valid globally and ordinary requirements for concentration of Gaussian distribution give us a condition $p = o(n)$ BvM result to be valid, see Bontemps (2011). Also it is worth noting that our condition in case of a Gaussian prior (see Section 2.3) coincide with one of Johnstone (2010).

Now consider more general situation, when errors come from a general distribution with a density $f(\cdot)$: $\varepsilon_i \sim P_f$. Denote $h(x) = \log f(x)$. The log-likelihood function for this problem reads as

$$\mathcal{L}(\mathbf{v}) = \sum_{i=1}^n \log f(Y_i - \Psi_i^\top \mathbf{v}) = \sum_{i=1}^n h(Y_i - \Psi_i^\top \mathbf{v}).$$

Suppose that $h(z)$ is twice continuously differentiable and $\mathbf{h} = \int h''(z) f(z) dz < \infty$. If

the model is correctly specified, then

$$\mathcal{D}_0^2 = -\nabla^2 \mathbb{E} \sum_{i=1}^n h(Y_i - \Psi_i^\top \mathbf{v}^*) = \int h''(z) f(z) dz \cdot \sum_{i=1}^n \Psi_i \Psi_i^\top = h \sum_{i=1}^n \Psi_i \Psi_i^\top.$$

Similarly,

$$\mathcal{D}^2(\mathbf{v}) = -\nabla^2 \mathbb{E} \sum_{i=1}^n h(Y_i - \Psi_i^\top \mathbf{v}) = \sum_{i=1}^n \Psi_i \Psi_i^\top \int h''(z - \Psi_i^\top (\mathbf{v} - \mathbf{v}^*)) f(z) dz.$$

If distribution P_f is subexponential then (ED_0) is clearly satisfied. Then (\mathcal{L}_0) and (\mathcal{L}_r) are also valid by arguments similar to the GLM case under the Lipschitz continuity assumption on $h''(\cdot)$.

Lemma 5.1. *Let*

$$|h''(z) - h''(z_0)| \leq L|z - z_0|, \quad z, z_0 \in \mathbb{R}.$$

Then

$$\delta(\mathbf{r}) \leq \frac{L \mathbf{r}}{h N_1^{1/2}},$$

where

$$N_1^{-1/2} \stackrel{\text{def}}{=} \max_i \sup_{\gamma \in \mathbb{R}^p} \frac{|\Psi_i^\top \gamma|}{\|\mathcal{D}_0 \gamma\|}.$$

The next interesting question is checking the condition (ED_2) . The stochastic part of the likelihood reads as follows:

$$\zeta(\mathbf{v}) = \sum_{i=1}^n h(Y_i - \Psi_i^\top \mathbf{v}) - \mathbb{E} h(Y_i - \Psi_i^\top \mathbf{v}).$$

To check the condition (ED_2) we need to compute the Hessian of the $\zeta(\mathbf{v})$:

$$\nabla^2 \zeta(\mathbf{v}) = \sum_{i=1}^n \{h''(Y_i - \Psi_i^\top \mathbf{v}) - \mathbb{E} h''(Y_i - \Psi_i^\top \mathbf{v})\} \Psi_i \Psi_i^\top.$$

Finally, we need to impose the mild condition on marginal likelihood:

(e₂) *There exist some constant ν_0 and for every $\mathbf{r} > 0$ exists $\mathbf{g}(\mathbf{r}) > 0$, such that for all numbers δ with $|\delta| \leq N_2^{-1/2} \mathbf{r}$*

$$\log \mathbb{E} \exp \left(\frac{\mu}{\mathfrak{s}_i} \{h''(Y_i + \delta) - \mathbb{E} h''(Y_i + \delta)\} \right) \leq \frac{\nu_0^2 \mu^2}{2}, \quad |\mu| \leq \mathbf{g}(\mathbf{r}),$$

where \mathfrak{s}_i are some known values and

$$N_2^{-1/2} \stackrel{\text{def}}{=} \max_i \sup_{\gamma \in \mathbb{R}^p} \frac{\mathfrak{s}_i |\Psi_i^\top \gamma|}{\|\mathcal{D}_0 \gamma\|}.$$

Then, we can check (ED_2) :

$$\begin{aligned}
& \sup_{\gamma_1, \gamma_2 \in \mathbb{R}^p} \log \mathbb{E} \exp \left\{ \frac{\mu}{\omega} \frac{\gamma_1^\top \nabla^2 \zeta(\mathbf{v}) \gamma_2}{\|\mathcal{D}_0 \gamma_1\| \cdot \|\mathcal{D}_0 \gamma_2\|} \right\} \\
&= \sup_{\gamma_1, \gamma_2 \in \mathbb{R}^p} \log \mathbb{E} \exp \left\{ \frac{\mu}{\omega} \frac{\sum_{i=1}^n (h''(Y_i - \Psi_i^\top \mathbf{v}) - \mathbb{E} h''(Y_i - \Psi_i^\top \mathbf{v})) \gamma_1^\top \Psi_i \Psi_i^\top \gamma_2}{\|\mathcal{D}_0 \gamma_1\| \cdot \|\mathcal{D}_0 \gamma_2\|} \right\} \\
&\leq \sup_{\gamma_1, \gamma_2 \in \mathbb{R}^p} \frac{\nu_0^2 \mu^2 \sum_{i=1}^n \mathbf{s}_i^2 |\Psi_i^\top \gamma_1|^2 \cdot |\Psi_i^\top \gamma_2|^2}{2\omega^2 \|\mathcal{D}_0 \gamma_1\|^2 \cdot \|\mathcal{D}_0 \gamma_2\|^2} \\
&\leq \frac{\nu_0^2 \mu^2}{2\omega^2} \frac{n}{N_2^2}.
\end{aligned}$$

It easily follows that $\omega = \frac{\sqrt{n}}{N_2}$ does the job. In regular situations N_2 is of order of sample size n and then $\omega \sim n^{-1/2}$. Finally,

Theorem 5.2. *Let (e_2) and conditions of Lemma 5.1 hold. Then the results of Theorem 2.1 hold for linear model with*

$$\Delta(\mathbf{r}_0, \mathbf{x}) = \{\delta(\mathbf{r}_0) + 6\nu_0 z_{\mathbb{H}}(\mathbf{x}) \omega\} \mathbf{r}_0^2 \leq \left\{ \frac{L}{\gamma} \frac{\mathbf{r}_0}{N_1^{1/2}} + 6\nu_0 z_{\mathbb{H}}(\mathbf{x}) \frac{\sqrt{n}}{N_2} \right\} \mathbf{r}_0^2.$$

5.2 Semiparametric linear regression with nuisance from Sobolev class

Now suppose that

$$\mathbf{f} = \Psi \boldsymbol{\theta}^* + \boldsymbol{\eta}^*, \quad (5.1)$$

where $\Psi = (\Psi_1, \dots, \Psi_q)$, $\Psi_i = \{\psi_1(X_i), \dots, \psi_q(X_i)\}$ for some basis functions $\{\psi_j(\cdot), j = 1, \dots, q\}$, $\boldsymbol{\eta}^* = \{\boldsymbol{\eta}(X_1), \dots, \boldsymbol{\eta}(X_N)\}$ and $\{X_i, i = 1, \dots, N\}$ is a set of regressors. Here $\boldsymbol{\eta}^*$ is an element of a functional space, e.g. a Sobolev ball \mathcal{F}_s :

$$\boldsymbol{\eta}^* \in \mathcal{F}_s \stackrel{\text{def}}{=} \left\{ \boldsymbol{\eta}(x) = \sum_{k=1}^{\infty} \eta_k \phi_k(x) : \sum_{k=1}^{\infty} \eta_k^2 k^{2s} \leq C \right\},$$

for a given functional basis $\{\phi_k\}_{k=1}^{\infty}$ (e.g. Fourier, wavelet, etc.). Define also $\Phi_i = \{\phi_1(X_i), \phi_2(X_i), \dots, \phi_m(X_i), \dots\}^\top$. Then the decomposition (5.1) can be rewritten as

$$\mathbf{f} = \Xi \mathbf{v}^*,$$

where $\Xi_i = (\Psi_i^\top, \Phi_i^\top)^\top$, $i = 1, \dots, n$ and $\Xi = (\Xi_1, \dots, \Xi_n)^\top$. Then we have the following operator as infinite dimensional counterpart of the Fisher information matrix:

$$\mathcal{D}_0^2 = \begin{pmatrix} D_0^2 & A_0 \\ A_0^\top & H_0^2 \end{pmatrix} = \mathbb{h} \sum_{i=1}^n \begin{pmatrix} \Psi_i \Psi_i^\top & \Psi_i \Phi_i^\top \\ \Phi_i \Psi_i^\top & \Phi_i \Phi_i^\top \end{pmatrix} = \mathbb{h} \sum_{i=1}^n \begin{pmatrix} \Psi_i \Psi_i^\top & \Psi_i \Phi_i^\top \\ \Phi_i \Psi_i^\top & \Phi_i \Phi_i^\top \end{pmatrix}.$$

We can equivalently write that $\mathcal{D}_0^2 = \{h \sum_{i=1}^n \xi_j(X_i) \xi_k(X_i)\}_{j,k=1}^\infty$, where ξ_j is a properly numerated joint basis of ψ_j and ϕ_k . We introduce the following matrix:

$$\mathcal{D}^2 = \left\{ h \int \xi_j(x) \xi_k(x) dx \right\}_{j,k=1}^\infty.$$

Further assume that the basis $\{\xi_i\}$ is orthonormal, i.e.

$$\mathcal{D}^2 = \{h \delta_{jk}\}_{j,k=1}^\infty.$$

If the design is regular, then we have elementwise convergence of sums to integrals. Define

$$N_3^{1/2} = \max(\|\mathcal{D}_0\|, \|A_0\|^2).$$

and also:

$$N_4^{-1/2} = \max(\|H_0^{-1}\|, \|H_m^{-1}\|, \|D_0^{-1}\|).$$

Now we check the bias. It holds

$$-EL(\mathbf{v}_m^*, \mathbf{v}^*) \leq -EL((\mathbf{v}^*)_m, \mathbf{v}^*),$$

where $(\mathbf{v}^*)_m$ is a projection of \mathbf{v}^* on a sieve space. Moreover,

$$-EL(\mathbf{v}, \mathbf{v}^*) = \|\mathcal{D}(\mathbf{v}^\circ)(\mathbf{v} - \mathbf{v}^*)\|^2/2,$$

where $\mathbf{v}^\circ \in [\mathbf{v}^*, \mathbf{v}]$. Then,

$$\begin{aligned} \|\mathcal{D}_0(\mathbf{v}_m^* - \mathbf{v}^*)\|^2 &\leq \frac{1 + \delta(\mathbf{r}_0)}{1 - \delta(\mathbf{r}_0)} \|\mathcal{D}_0((\mathbf{v}^*)_m - \mathbf{v}^*)\|^2 \\ &\leq \frac{1 + \delta(\mathbf{r}_0)}{1 - \delta(\mathbf{r}_0)} \|\mathcal{D}_0\|^2 \frac{C}{m^{2c}} = \frac{1 + \delta(\mathbf{r}_0)}{1 - \delta(\mathbf{r}_0)} \frac{CN_3}{m^{2c}}. \end{aligned}$$

The further look gives us:

$$\|\check{D}_0(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*)\|^2 \leq \|D_0(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*)\|^2$$

and

$$\begin{aligned} &\|\mathcal{D}_0(\mathbf{v}_m^* - \mathbf{v}^*)\|^2 \\ &\geq \|D_0(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*)\|^2 - 2\nu \|D_0(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*)\| \|H_0(\boldsymbol{\eta}_m^* - \boldsymbol{\eta}^*)\|^2 + \|H_0(\boldsymbol{\eta}_m^* - \boldsymbol{\eta}^*)\|^2. \end{aligned}$$

Thus,

$$\begin{aligned}
\|\check{D}_0(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*)\|^2 &\leq \|D_0(\boldsymbol{\theta}_m^* - \boldsymbol{\theta}^*)\|^2 \\
&\leq \left(\|\mathcal{D}_0(\mathbf{v}_m^* - \mathbf{v}^*)\| + \sqrt{\nu} \|H_0(\boldsymbol{\eta}_m^* - \boldsymbol{\eta}^*)\| \right)^2 \\
&\leq \left(\left[\frac{1 + \delta(\mathbf{r}_0)}{1 - \delta(\mathbf{r}_0)} \frac{CN_3}{m^{2s}} \right]^{1/2} + \left[\nu \frac{CN_3}{m^{2s}} \right]^{1/2} \right)^2 \\
&\leq \left(\left[\frac{1 + \delta(\mathbf{r}_0)}{1 - \delta(\mathbf{r}_0)} \right]^{1/2} + \nu^{1/2} \right)^2 \frac{CN_3}{m^{2s}}.
\end{aligned}$$

Now we need check the second part of the condition (B):

$$\begin{aligned}
\|\check{D}_0^{-1}(\check{D}_0^2 - \check{D}_m^2)\check{D}_0^{-1}\| &= \|\check{D}_0^{-1}(A_0 H_0^{-2} A_0^\top - A_m H_m^{-2} A_m^\top)\check{D}_0^{-1}\| \\
&\leq \|\check{D}_0^{-2}\|(\|A_0 H_0^{-2} A_0^\top\| + \|A_m H_m^{-2} A_m^\top\|) \\
&\leq \|\check{D}_0^{-2}\|(\|A_0\|^2 \|H_0^{-2}\| + \|A_m\|^2 \|H_m^{-2}\|) \leq 2 \frac{N_3}{N_4^2}.
\end{aligned}$$

Thus, we can state a bound:

$$\beta_m \leq 2 \frac{N_3}{N_4^2}.$$

In case of regular design we get $N_3, N_4 \sim n$. Then,

$$\alpha_m \leq \mathbf{C} \frac{n}{m^{2s}}, \quad \beta_m \leq \mathbf{C} \frac{1}{n},$$

where \mathbf{C} is some constant. For validity of BvM theorem with sieve prior we need $q + m \ll n^{1/3}$. Then for $s > 3/2$ under the proper choice of m the values of α_m and β_m are asymptotically negligible.

5.3 Generalized linear modeling

Now we consider a generalized linear modeling (GLM) which is often used for describing some categorical data. Let $\mathcal{P} = (P_w, w \in \mathcal{Y})$ be an exponential family with a canonical parametrization; see e.g. McCullagh and Nelder (1989). The corresponding log-density can be represented as $\ell(y, w) = yw - d(w)$ for a convex function $d(w)$. The popular examples are given by the binomial (binary response, logistic) model with $d(w) = \log(e^w + 1)$, the Poisson model with $d(w) = e^w$, the exponential model with $d(w) = -\log(w)$. Note that linear Gaussian regression is a special case with $d(w) = w^2/2$.

A GLM specification means that every observation Y_i has a distribution from the family P with the parameter w_i which linearly depends on the regressor $\Psi_i \in \mathbb{R}^q$:

$$Y_i \sim P_{\Psi_i^\top \mathbf{v}^*}. \quad (5.2)$$

The corresponding log-density of a GLM reads as

$$\mathcal{L}(\boldsymbol{\theta}) = \sum \{Y_i \Psi_i^\top \mathbf{v} - d(\Psi_i^\top \mathbf{v})\}.$$

Under $P_{\boldsymbol{\theta}^*}$ each observation Y_i follows (5.2), in particular, $\mathbb{E}Y_i = d'(\Psi_i^\top \mathbf{v}^*)$. However, similarly to the previous sections, it is accepted that the parametric model (5.2) is misspecified. Response misspecification means that the vector $\mathbf{f} \stackrel{\text{def}}{=} \mathbb{E}\mathbf{Y}$ cannot be represented in the form $d'(\Psi^\top \mathbf{v})$ whatever \mathbf{v} is. The other sort of misspecification concerns the data distribution. The model (5.2) assumes that the Y_i 's are independent and the marginal distribution belongs to the given parametric family P . In what follows, we only assume independent data having certain exponential moments. The target of estimation \mathbf{v}^* is defined by

$$\mathbf{v}^* \stackrel{\text{def}}{=} \underset{\mathbf{v}}{\operatorname{argmax}} \mathbb{E}\mathcal{L}(\mathbf{v}).$$

The quasi MLE $\tilde{\mathbf{v}}$ is defined by maximization of $\mathcal{L}(\mathbf{v})$:

$$\tilde{\mathbf{v}} = \underset{\mathbf{v}}{\operatorname{argmax}} \mathcal{L}(\mathbf{v}) = \underset{\mathbf{v}}{\operatorname{argmax}} \sum \{Y_i \Psi_i^\top \mathbf{v} - d(\Psi_i^\top \mathbf{v})\}.$$

Convexity of $d(\cdot)$ implies that $\mathcal{L}(\mathbf{v})$ is a concave function of \mathbf{v} , so that the optimization problem has a unique solution and can be effectively solved. However, a closed form solution is only available for the constant regression or for the linear Gaussian regression. The corresponding target \mathbf{v}^* is the maximizer of the expected log-likelihood:

$$\mathbf{v}^* = \underset{\mathbf{v}}{\operatorname{argmax}} \mathbb{E}\mathcal{L}(\mathbf{v}) = \underset{\mathbf{v}}{\operatorname{argmax}} \sum \{f_i \Psi_i^\top \mathbf{v} - d(\Psi_i^\top \mathbf{v})\}$$

with $f_i = \mathbb{E}Y_i$. The function $\mathbb{E}\mathcal{L}(\mathbf{v})$ is concave as well and the vector \mathbf{v}^* is also well defined.

Define the individual errors (residuals) $\varepsilon_i = Y_i - \mathbb{E}Y_i$. Below we assume that these errors fulfill some exponential moment conditions.

(e₁) *There exist some constants ν_0 and $\mathbf{g}_1 > 0$, and for every i a constant \mathbf{s}_i such that $\mathbb{E}(\varepsilon_i/\mathbf{s}_i)^2 \leq 1$ and*

$$\log \mathbb{E} \exp(\mu \varepsilon_i / \mathbf{s}_i) \leq \nu_0^2 \mu^2 / 2, \quad |\mu| \leq \mathbf{g}_1. \quad (5.3)$$

A natural candidate for \mathfrak{s}_i is σ_i where $\sigma_i^2 = \mathbb{E}\varepsilon_i^2$ is the variance of ε_i . Under (5.3), introduce a $q \times q$ matrix \mathcal{V}_0 defined by

$$\mathcal{V}_0^2 \stackrel{\text{def}}{=} \sum \mathfrak{s}_i^2 \Psi_i \Psi_i^\top. \quad (5.4)$$

Condition (e_1) effectively means that each error term $\varepsilon_i = Y_i - \mathbb{E}Y_i$ has some bounded exponential moments: for $|\lambda| \leq \mathfrak{g}_1$, it holds $f(\lambda) \stackrel{\text{def}}{=} \log \mathbb{E} \exp(\lambda \varepsilon_i / \mathfrak{s}_i) < \infty$. In words, condition (e_1) requires light (exponentially decreasing) tail for the marginal distribution of each ε_i .

Define also

$$N_1^{-1/2} \stackrel{\text{def}}{=} \max_i \sup_{\gamma \in \mathbb{R}^p} \frac{\mathfrak{s}_i |\Psi_i^\top \gamma|}{\|\mathcal{V}_0 \gamma\|}. \quad (5.5)$$

Now conditions are satisfied due to following lemma, see Spokoiny (2012) for proof.

Lemma 5.3. *Assume (e_1) and let \mathcal{V}_0^2 be defined by (5.4) and N_1 by (5.5). Then condition (ED_0) follows from (e_1) with this \mathcal{V}_0^2 and $\mathfrak{g} = \mathfrak{g}_1 N_1^{1/2}$. Moreover, the stochastic component $\zeta(\mathbf{v})$ is linear in \mathbf{v} and the condition (ED_2) is fulfilled with $\omega(\mathbf{r}) \equiv 0$.*

It only remains to bound the quality of quadratic approximation for the mean of the process $L(\mathbf{v}, \mathbf{v}^*)$ in a vicinity of \mathbf{v}^* . An interesting feature of the GLM is that the effect of model misspecification disappears in the expectation of $L(\mathbf{v}, \mathbf{v}^*)$.

Lemma 5.4. *It holds*

$$-\mathbb{E}L(\mathbf{v}, \mathbf{v}^*) = \sum \{d(\Psi_i^\top \mathbf{v}) - d(\Psi_i^\top \mathbf{v}^*) - d'(\Psi_i^\top \mathbf{v}^*) \Psi_i^\top (\mathbf{v} - \mathbf{v}^*)\} = \mathcal{K}(\mathbb{P}_{\mathbf{v}^*}, \mathbb{P}_{\mathbf{v}}),$$

where $\mathcal{K}(\mathbb{P}_{\mathbf{v}^*}, \mathbb{P}_{\mathbf{v}})$ is the Kullback-Leibler divergence between measures $\mathbb{P}_{\mathbf{v}^*}$ and $\mathbb{P}_{\mathbf{v}}$. Moreover,

$$-\mathbb{E}L(\mathbf{v}, \mathbf{v}^*) = \|\mathcal{D}(\mathbf{v}^\circ)(\mathbf{v} - \mathbf{v}^*)\|^2/2,$$

where $\mathbf{v}^\circ \in [\mathbf{v}^*, \mathbf{v}]$ and

$$\mathcal{D}^2(\mathbf{v}^\circ) = \sum d''(\Psi_i^\top \mathbf{v}^\circ) \Psi_i \Psi_i^\top.$$

The proof of this lemma can also be found in Spokoiny (2012). Define now the matrix \mathcal{D}_0^2 by

$$\mathcal{D}_0^2 \stackrel{\text{def}}{=} \mathcal{D}^2(\mathbf{v}^*) = \sum d''(\Psi_i^\top \mathbf{v}^*) \Psi_i \Psi_i^\top.$$

Let also \mathcal{V}_0^2 be defined by (5.4). Note that the matrices \mathcal{D}_0^2 and \mathcal{V}_0^2 coincide if the model $Y_i \sim P_{\Psi_i^\top \mathbf{v}^*}$ is correctly specified and $\mathfrak{s}_i^2 = d''(\Psi_i^\top \mathbf{v}^*)$. The matrix \mathcal{D}_0^2 describes a local

elliptic neighborhood of the central point \mathbf{v}^* in the form $\mathcal{V}_0(\mathbf{r}) = \{\mathbf{v} : \|\mathcal{D}_0(\mathbf{v} - \mathbf{v}^*)\| \leq \mathbf{r}\}$. If the matrix function $\mathcal{D}^2(\mathbf{v})$ is continuous in this vicinity $\mathcal{V}_0(\mathbf{r})$ then the value $\delta(\mathbf{r})$ measuring the approximation quality of $-\mathbb{E}L(\mathbf{v}, \mathbf{v}^*)$ by the quadratic function $\|\mathcal{D}_0(\mathbf{v} - \mathbf{v}^*)\|^2/2$ is small and the identifiability condition (\mathcal{L}_0) is fulfilled on $\mathcal{V}_0(\mathbf{r})$. The following lemma gives bounds for $\delta(\mathbf{r})$.

Lemma 5.5. *Let $d''(z)$ be Lipschitz continuous:*

$$|d''(z) - d''(z_0)| \leq L|z - z_0|, \quad z, z_0 \in \mathbb{R}$$

Then

$$\delta(\mathbf{r}) \leq L \frac{\mathbf{r}}{N_2^{1/2}},$$

where

$$N_2^{-1/2} \stackrel{\text{def}}{=} \max_i \sup_{\gamma \in \mathbb{R}^p} \frac{|\Psi_i^\top \gamma|}{d''(\Psi_i^\top \mathbf{v}^*) \cdot \|\mathcal{D}_0 \gamma\|}.$$

Now we are prepared to state the local results for the GLM estimation.

Theorem 5.6. *Let (e_1) and conditions of Lemma 5.5 hold. Then the results of Theorem 2.1 hold for GLM with*

$$\Delta(\mathbf{r}_0, \mathbf{x}) \leq L \frac{\mathbf{r}_0^3}{N_2^{1/2}}.$$

If the function $d(w)$ is quadratic then the approximation error δ vanishes as well and then quadratic approximation is valid globally, a localization step is not required. However, if $d(w)$ is not quadratic, the result applies only locally and it has to be accomplished with a large deviation bound. The GLM structure is helpful in the large deviation zone as well. Indeed, the gradient $\nabla \zeta(\mathbf{v})$ does not depend on \mathbf{v} and hence, the most delicate condition $(E\mathbf{r})$ is fulfilled automatically with $\mathbf{g} = \mathbf{g}_1 N^{1/2}$ for all local sets $\mathcal{V}_0(\mathbf{r})$. Further, the identifiability condition $(\mathcal{L}\mathbf{r})$ easily follows from Lemma 5.4: it suffices to bound from below the matrix $\mathcal{D}(\mathbf{v})$ for $\mathbf{v} \in \mathcal{V}_0(\mathbf{r})$:

$$\mathcal{D}(\mathbf{v}) \geq \mathbf{b}(\mathbf{r})\mathcal{D}_0, \quad \mathbf{v} \in \mathcal{V}_0(\mathbf{r}).$$

An interesting question, similarly to the i.i.d. case, is the minimal radius \mathbf{r}_0 of the local vicinity $\mathcal{V}_0(\mathbf{r}_0)$ ensuring the desirable concentration property. The required value conditions are fulfilled for $\mathbf{r}^2 \geq \mathbf{r}_0^2 = \mathbf{C}(\mathbf{x} + p)$, where \mathbf{C} only depends on ν_0, \mathbf{b} , and \mathbf{g} . Thus, the results are valid if

$$\delta(\mathbf{r}_0)\mathbf{r}_0^2 = \mathbf{C} \frac{\mathbf{r}_0^3}{N_1^{1/2}} = \mathbf{C} \frac{(\mathbf{x} + p)^{3/2}}{N_1^{1/2}}$$

is small.

The GLM model also allows a semiparametric extension, i.e.

$$w_i = \Psi_i^\top \boldsymbol{\theta}^* + \boldsymbol{\eta}^*(X_i),$$

where $\boldsymbol{\eta}^*(\cdot)$ is from a Sobolev class. This setup differs from Section 5.2 only by few technicalities and leads to similar theoretical results.

6 Supplementary

This section contains the imposed conditions and some supplementary statements which are of some interest by itself.

6.1 Bracketing and upper function devices

This section briefly overviews the main constructions of Spokoiny (2012) including the bracketing bound and the upper function results. The bracketing bound describes the quality of quadratic approximation of the log-likelihood process $\mathcal{L}(\mathbf{v})$ in a local vicinity of the point \mathbf{v}^* , while the upper function method is used to show that the full MLE $\tilde{\mathbf{v}}$ belongs to this vicinity with a dominating probability. Introduce the notation $L(\mathbf{v}, \mathbf{v}^*) = \mathcal{L}(\mathbf{v}) - \mathcal{L}(\mathbf{v}^*)$ for the (quasi) log-likelihood ratio. Given $\mathbf{r} > 0$, define the local set

$$\mathcal{V}_0(\mathbf{r}) \stackrel{\text{def}}{=} \{\mathbf{v} : (\mathbf{v} - \mathbf{v}^*)^\top \mathcal{D}_0^2(\mathbf{v} - \mathbf{v}^*) \leq \mathbf{r}^2\}. \quad (6.1)$$

Define the quadratic processes $\mathbb{L}(\mathbf{v}, \mathbf{v}^*)$:

$$\mathbb{L}(\mathbf{v}, \mathbf{v}^*) \stackrel{\text{def}}{=} (\mathbf{v} - \mathbf{v}^*)^\top \nabla \mathcal{L}(\mathbf{v}^*) - \|\mathcal{D}_0(\mathbf{v} - \mathbf{v}^*)\|^2/2.$$

The next result states the local bracketing bound. The formulation assumes that some value \mathbf{x} is fixed such that $e^{-\mathbf{x}}$ is sufficiently small. If the dimension p is large, one can select $\mathbf{x} = \mathbf{C} \log(p)$. We assume that a value $\mathbf{r} = \mathbf{r}_0$ is fixed which separates the local and global zones.

Theorem 6.1. *Suppose the conditions (ED_0) , (ED_2) , (\mathcal{L}_0) , and (\mathcal{I}) from Section 2.1 hold for some $\mathbf{r}_0 > 0$. Then on a random set $\Omega_{\mathbf{r}_0}(\mathbf{x})$ of dominating probability at least $1 - e^{-\mathbf{x}}$*

$$|L(\mathbf{v}, \mathbf{v}^*) - \mathbb{L}(\mathbf{v}, \mathbf{v}^*)| \leq \Delta(\mathbf{r}_0, \mathbf{x}), \quad \mathbf{v} \in \mathcal{V}_0(\mathbf{r}_0), \quad (6.2)$$

where

$$\begin{aligned} \Delta(\mathbf{r}_0, \mathbf{x}) &\stackrel{\text{def}}{=} \{\delta(\mathbf{r}_0) + 6\nu_0 z_{\mathbb{H}}(\mathbf{x}) \omega\} \mathbf{r}_0^2, \\ z_{\mathbb{H}}(\mathbf{x}) &\stackrel{\text{def}}{=} 2p^{1/2} + \sqrt{2\mathbf{x}} + \mathbf{g}^{-1}(\mathbf{g}^{-2}\mathbf{x} + 1)4p, \end{aligned} \quad (6.3)$$

and $\mathcal{I}_0(\mathbf{r}_0)$ is defined in (6.1). Moreover, the random vector $\boldsymbol{\xi} = \mathcal{D}_0^{-1} \nabla \mathcal{L}(\mathbf{v}^*)$ fulfills on a random set $\Omega_B(\mathbf{x})$ of dominating probability at least $1 - 2e^{-x}$

$$\|\boldsymbol{\xi}\|^2 \leq z^2(B, \mathbf{x}), \quad (6.4)$$

where $z^2(B, \mathbf{x}) \stackrel{\text{def}}{=} \mathbf{p}_B + 6\lambda_B \mathbf{x}$,

$$B \stackrel{\text{def}}{=} \mathcal{D}_0^{-1} \mathcal{V}_0^2 \mathcal{D}_0^{-1}, \quad \mathbf{p}_B \stackrel{\text{def}}{=} \text{tr}(B), \quad \lambda_B \stackrel{\text{def}}{=} \lambda_{\max}(B).$$

Furthermore, assume $(\mathcal{L}\mathbf{r})$ with $\mathbf{b}(\mathbf{r}) \equiv \mathbf{b}$ yielding

$$-EL(\mathbf{v}, \mathbf{v}^*) \geq \mathbf{b} \|\mathcal{D}_0(\mathbf{v} - \mathbf{v}^*)\|^2$$

for each $\mathbf{v} \in \mathcal{I} \setminus \mathcal{I}_0(\mathbf{r}_0)$. Let also

$$\mathbf{r} \geq \frac{2}{\mathbf{b}} \left\{ z(B, \mathbf{x}) + 6\nu_0 z_{\mathbb{H}}(\mathbf{x} + \log(2\mathbf{r}/\mathbf{r}_0)) \omega \right\}, \quad \mathbf{r} \geq \mathbf{r}_0$$

with $z_{\mathbb{H}}(\mathbf{x})$ from (6.3). Then,

$$L(\mathbf{v}, \mathbf{v}^*) \leq -\mathbf{b} \|\mathcal{D}_0(\mathbf{v} - \mathbf{v}^*)\|^2/2, \quad \mathbf{v} \in \mathcal{I} \setminus \mathcal{I}_0(\mathbf{r}_0). \quad (6.5)$$

holds on a random set $\Omega(\mathbf{x})$ of probability at least $1 - 4e^{-x}$.

The result (6.2) is an improved version of approximation bound obtained in Spokoiny (2012), Theorem 3.1. The result (6.4) can be found in the supplement to Spokoiny (2012). The result (6.5) is very similar to Theorem 4.2 from Spokoiny (2012).

6.2 Tail posterior probability for full parameter space

The next step in our analysis is to check that \mathbf{v} concentrates in a small vicinity $\mathcal{I}_0(\mathbf{r}_0)$ of the central point \mathbf{v}^* with a properly selected \mathbf{r}_0 . The concentration properties of the posterior will be described by using the random quantity

$$\rho^*(\mathbf{r}_0) = \frac{\int_{\mathcal{I} \setminus \mathcal{I}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v}}{\int_{\mathcal{I}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v}}.$$

Theorem 6.2. Suppose the conditions of Theorem 6.1. Then it holds on $\Omega_{\mathbf{r}_0}(\mathbf{x})$

$$\rho^*(\mathbf{r}_0) \leq \exp\{2\Delta(\mathbf{r}_0, \mathbf{x}) + \nu(\mathbf{r}_0)\} \mathbf{b}^{-p/2} \mathbb{P}(\|\boldsymbol{\gamma}\|^2 \geq \mathbf{b}\mathbf{r}_0^2), \quad (6.6)$$

with

$$\nu(\mathbf{r}_0) \stackrel{\text{def}}{=} -\log \mathbb{P}(\|\boldsymbol{\gamma} + \boldsymbol{\xi}\| \leq \mathbf{r}_0 \mid \mathbf{Y}).$$

If $\mathbf{r}_0 \geq z(B, \mathbf{x}) + z(p, \mathbf{x})$, then on $\Omega(\mathbf{x})$

$$\nu(\mathbf{r}_0) \leq 2e^{-\mathbf{x}}. \quad (6.7)$$

This result yields simple sufficient conditions on the value \mathbf{r}_0 which ensures the concentration of the posterior on $\Upsilon_0(\mathbf{r}_0)$.

Corollary 6.3. *Assume the conditions of Theorem 6.2. Then additional inequality $\mathbf{br}_0^2 \geq z^2(p, \mathbf{x} + \frac{p}{2} \log \frac{e}{b})$ ensures on a random set $\Omega(\mathbf{x})$ of probability at least $1 - 4e^{-\mathbf{x}}$*

$$\rho^*(\mathbf{r}_0) \leq \exp\{2\Delta(\mathbf{r}_0, \mathbf{x}) + 2e^{-\mathbf{x}} - \mathbf{x}\}.$$

The result follows from Theorem 6.2 with use of Lemma 7.2.

6.3 Tail posterior probability for target parameter

The next major step in our analysis is to check that $\boldsymbol{\theta}$ concentrates in a small vicinity $\Theta_0(\mathbf{r}_0) = \{\boldsymbol{\theta}: \|\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathbf{r}_0\}$ of the central point $\boldsymbol{\theta}^* = \Pi_0 \mathbf{v}^*$ with a properly selected \mathbf{r}_0 . The concentration properties of the posterior will be described by using the random quantity

$$\rho(\mathbf{r}_0) \stackrel{\text{def}}{=} \frac{\int_{\mathcal{Y}} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} \pi(\mathbf{v}) \mathbb{I}\{\boldsymbol{\theta} \notin \Theta_0(\mathbf{r}_0)\} d\mathbf{v}}{\int_{\mathcal{Y}} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} \pi(\mathbf{v}) \mathbb{I}\{\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)\} d\mathbf{v}}.$$

In what follows we suppose that prior is uniform, i.e. $\pi(\mathbf{v}) \equiv 1$, $\mathbf{v} \in \mathcal{Y}$. This results in the following representation for $\rho(\mathbf{r}_0)$:

$$\rho(\mathbf{r}_0) = \frac{\int_{\mathcal{Y}} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} \mathbb{I}\{\boldsymbol{\theta} \notin \Theta_0(\mathbf{r}_0)\} d\mathbf{v}}{\int_{\mathcal{Y}} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} \mathbb{I}\{\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)\} d\mathbf{v}}. \quad (6.8)$$

Obviously $P(\boldsymbol{\theta} \notin \Theta_0(\mathbf{r}_0) \mid \mathbf{Y}) \leq \rho(\mathbf{r}_0)$. Therefore, small values of $\rho(\mathbf{r}_0)$ indicate a small posterior probability of the large deviation set $\{\boldsymbol{\theta} \notin \Theta_0(\mathbf{r}_0)\}$.

Theorem 6.4. *Suppose (6.2). Then for $\mathbf{br}_0^2 \geq z^2(p, \mathbf{x} + \frac{p}{2} \log \frac{e}{b})$ on $\Omega(\mathbf{x})$ of probability at least $1 - 4e^{-\mathbf{x}}$*

$$\rho(\mathbf{r}_0) \leq \rho^*(\mathbf{r}_0) \leq \exp\{2\Delta(\mathbf{r}_0, \mathbf{x}) + 2e^{-\mathbf{x}} - \mathbf{x}\}.$$

6.4 Local Gaussian approximation of the posterior. Upper bound

It is convenient to introduce local conditional expectation: for a random variable η , define

$$\mathbb{E}^\circ \eta \stackrel{\text{def}}{=} \mathbb{E} \left[\eta \mathbb{I}\{\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)\} \mid \mathbf{Y} \right].$$

The following theorem gives exact statement about upper bound of this posterior expectation. Let

$$\boldsymbol{\theta}^\circ \stackrel{\text{def}}{=} \boldsymbol{\theta}^* + \check{D}_0^{-1} \check{\xi}.$$

Theorem 6.5. *Suppose (6.2). Then for any $f: \mathbb{R}^q \rightarrow \mathbb{R}_+$ it holds on $\Omega_{\mathbf{r}_0}(\mathbf{x})$*

$$\mathbb{E}^\circ f(\check{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ)) \leq \exp\{\Delta^+(\mathbf{r}_0, \mathbf{x})\} \mathbb{E} f(\boldsymbol{\gamma}), \quad (6.9)$$

where $\boldsymbol{\gamma} \sim \mathcal{N}(0, \mathbf{I}_q)$ and

$$\begin{aligned} \Delta^+(\mathbf{r}_0, \mathbf{x}) &\stackrel{\text{def}}{=} 2\Delta(\mathbf{r}_0, \mathbf{x}) + \nu(\mathbf{r}_0) + \rho_f(\mathbf{r}_0), \\ \rho_f(\mathbf{r}_0) &\stackrel{\text{def}}{=} \frac{\int_{\mathcal{Y} \setminus \mathcal{Y}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v}}{\int_{\mathcal{Y}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v}}. \end{aligned}$$

Define for random event $\eta \in A \subseteq \mathbb{R}^q$:

$$\mathbb{P}^\circ(\eta \in A) = \mathbb{E}^\circ \mathbb{I}\{\eta \in A\}.$$

The next result considers a special case with $f(\mathbf{u}) = |\boldsymbol{\lambda}^\top \mathbf{u}|^2$ and $f(\mathbf{u}) = \mathbb{I}(\mathbf{u} \in A)$ for any measurable set A .

Corollary 6.6. *For any $\boldsymbol{\lambda} \in \mathbb{R}^q$, it holds on $\Omega_{\mathbf{r}_0}(\mathbf{x})$*

$$\mathbb{E}^\circ |\boldsymbol{\lambda}^\top \check{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ)|^2 \leq \exp\{\Delta^+(\mathbf{r}_0, \mathbf{x})\} \|\boldsymbol{\lambda}\|^2.$$

For any measurable set $A \subseteq \mathbb{R}^q$, it holds on $\Omega_{\mathbf{r}_0}(\mathbf{x})$

$$\mathbb{P}^\circ(\check{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ) \in A) \leq \exp\{\Delta^+(\mathbf{r}_0, \mathbf{x})\} \mathbb{P}(\boldsymbol{\gamma} \in A). \quad (6.10)$$

On $\Omega(\mathbf{x})$ one obtains

$$\Delta^+(\mathbf{r}_0, \mathbf{x}) \leq 2\Delta(\mathbf{r}_0, \mathbf{x}) + 2e^{-\mathbf{x}} + 2\exp\{\Delta(\mathbf{r}_0, \mathbf{x}) + 4e^{-\mathbf{x}} - \mathbf{x}\}.$$

The next corollary describes an upper bound for the posterior probability in case of changing of scaling.

Corollary 6.7. *Let D_1 be symmetric $q \times q$ matrix such that $\|I - D_1^{-1} \check{D}_0^2 D_1^{-1}\| \leq \alpha$. Let also $\hat{\boldsymbol{\theta}} \in \mathbb{R}^q$ be such that $\|\check{D}_0(\boldsymbol{\theta}^\circ - \hat{\boldsymbol{\theta}})\| \leq \beta$. Then for any measurable set $A \subset \mathbb{R}^q$, it holds on $\Omega(\mathbf{x})$ with $\boldsymbol{\delta}_0 \stackrel{\text{def}}{=} D_1(\boldsymbol{\theta}^\circ - \hat{\boldsymbol{\theta}})$*

$$\begin{aligned} \mathbb{P}^\circ(D_1(\boldsymbol{\vartheta} - \hat{\boldsymbol{\theta}}) \in A) &\leq \exp\{\Delta^+(\mathbf{r}_0, \mathbf{x})\} \mathbb{P}(D_1 \check{D}_0^{-1} \boldsymbol{\gamma} + \boldsymbol{\delta}_0 \in A) \\ &\leq \exp\{\Delta^+(\mathbf{r}_0, \mathbf{x})\} \left(\mathbb{P}(\boldsymbol{\gamma} \in A) + \frac{1}{2} \sqrt{\alpha^2 q + (1 + \alpha)^2 \beta^2} \right). \end{aligned} \quad (6.11)$$

6.5 Local Gaussian approximation of the posterior. Lower bound

Now we present a local lower bound for the posterior measure.

Theorem 6.8. *Suppose (6.2). Then for any $f: \mathbb{R}^q \rightarrow \mathbb{R}_+$ it holds on $\Omega_{\mathbf{r}_0}(\mathbf{x})$*

$$\mathbb{E}^\circ f(\check{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ)) \geq \exp\{-\Delta^-(\mathbf{r}_0, \mathbf{x})\} \mathbb{E}\{f(\boldsymbol{\gamma}) \mathbb{I}(\|\boldsymbol{\gamma} + \check{\boldsymbol{\xi}}\| \leq \mathbf{r}_0)\}, \quad (6.12)$$

where

$$\begin{aligned} \Delta^-(\mathbf{r}_0, \mathbf{x}) &\stackrel{\text{def}}{=} 2\Delta(\mathbf{r}_0, \mathbf{x}) + \nu(\mathbf{r}_0) + \rho^*(\mathbf{r}_0) + 2\tilde{\rho}_f(\mathbf{r}_0), \\ \tilde{\rho}_f(\mathbf{r}_0) &\stackrel{\text{def}}{=} \frac{\int_{\mathbb{R}^p \setminus \mathcal{r}_0(\mathbf{r}_0)} \exp\{\mathbb{L}(\mathbf{v}, \mathbf{v}^*)\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v}}{\int_{\mathcal{r}_0(\mathbf{r}_0)} \exp\{\mathbb{L}(\mathbf{v}, \mathbf{v}^*)\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v}}. \end{aligned}$$

This result means that posterior measure can be bounded from below by the standard normal law up to (small) multiplicative and additive constants. As a corollary, we state the result for quadratic and indicator functions $f(\mathbf{u})$.

Corollary 6.9. *For any $\boldsymbol{\lambda} \in \mathbb{R}^q$, it holds on $\Omega_{\mathbf{r}_0}(\mathbf{x})$*

$$\mathbb{E}^\circ |\boldsymbol{\lambda}^\top \check{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ)|^2 \geq \exp\{-\Delta^\ominus(\mathbf{r}_0, \mathbf{x}) + \mathbf{e}^{-\mathbf{x}}\} \|\boldsymbol{\lambda}\|^2.$$

For any measurable set $A \subseteq \mathbb{R}^q$, it holds on $\Omega_{\mathbf{r}_0}(\mathbf{x})$

$$\mathbb{P}^\circ(\check{D}_0(\boldsymbol{\vartheta} - \check{\boldsymbol{\theta}}) \in A) \geq \exp\{\Delta^-(\mathbf{r}_0, \mathbf{x})\} \mathbb{P}(\boldsymbol{\gamma} \in A) - \mathbf{e}^{-\mathbf{x}}. \quad (6.13)$$

Let D_1^2 be a symmetric $q \times q$ matrix such that $\|I - D_1^{-1} \check{D}_0^2 D_1^{-1}\| \leq \alpha$ and let $\hat{\boldsymbol{\theta}} \in \mathbb{R}^q$ be such that $\|\check{D}_0(\boldsymbol{\theta}^\circ - \hat{\boldsymbol{\theta}})\| \leq \beta$. Define $\boldsymbol{\delta}_0 \stackrel{\text{def}}{=} D_1(\boldsymbol{\theta}^\circ - \hat{\boldsymbol{\theta}})$. Then for any measurable subset A in \mathbb{R}^q , it holds on $\Omega(\mathbf{x})$

$$\begin{aligned} \mathbb{P}^\circ(D_1(\boldsymbol{\vartheta} - \hat{\boldsymbol{\theta}}) \in A) &\geq \exp\{\Delta^-(\mathbf{r}_0, \mathbf{x})\} \mathbb{P}(D_1 \check{D}_0^{-1} \boldsymbol{\gamma} + \boldsymbol{\delta}_0 \in A) - \mathbf{e}^{-\mathbf{x}} \\ &\geq \exp\{\Delta^-(\mathbf{r}_0, \mathbf{x})\} \left\{ \mathbb{P}(\boldsymbol{\gamma} \in A) - \frac{1}{2} \sqrt{\alpha^2 q + (1 + \alpha)^2 \beta^2} \right\} - \mathbf{e}^{-\mathbf{x}}. \end{aligned} \quad (6.14)$$

On $\Omega(\mathbf{x})$ one obtains $\Delta^-(\mathbf{r}_0, \mathbf{x}) \leq 2\Delta(\mathbf{r}_0, \mathbf{x}) + 3\mathbf{e}^{-\mathbf{x}} + 4\exp\{\Delta(\mathbf{r}_0, \mathbf{x}) + 4\mathbf{e}^{-\mathbf{x}} - \mathbf{x}\}$.

The proof of this corollary is similar to Corollary 6.7 and Corollary 6.6.

7 Proofs

This appendix collects the proofs of the results.

7.1 Some inequalities for the normal law

This section collects 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 section γ means a standard normal vector in \mathbb{R}^q .

Lemma 7.1. *For any $\mathbf{u} \in \mathbb{R}^q$, any unit vector $\mathbf{a} \in \mathbb{R}^q$, and any $z > 0$, it holds*

$$\mathbb{P}(\|\gamma - \mathbf{u}\| \geq z) \leq \exp\{-z^2/4 + q/2 + \|\mathbf{u}\|^2/2\}, \quad (7.1)$$

$$\mathbb{E}\{|\gamma^\top \mathbf{a}|^2 \mathbb{I}(\|\gamma - \mathbf{u}\| \geq z)\} \leq (2 + |\mathbf{u}^\top \mathbf{a}|^2) \exp\{-z^2/4 + q/2 + \|\mathbf{u}\|^2/2\}. \quad (7.2)$$

Proof. By the exponential Chebyshev inequality, for any $\lambda < 1$

$$\begin{aligned} \mathbb{P}(\|\gamma - \mathbf{u}\| \geq z) &\leq \exp(-\lambda z^2/2) \mathbb{E} \exp(\lambda \|\gamma - \mathbf{u}\|^2/2) \\ &= \exp\left\{-\frac{\lambda z^2}{2} - \frac{q}{2} \log(1 - \lambda) + \frac{\lambda}{2(1 - \lambda)} \|\mathbf{u}\|^2\right\}. \end{aligned}$$

In particular, with $\lambda = 1/2$, this implies (7.1). Further, for $\|\mathbf{a}\| = 1$

$$\begin{aligned} \mathbb{E}\{|\gamma^\top \mathbf{a}|^2 \mathbb{I}(\|\gamma - \mathbf{u}\| \geq z)\} &\leq \exp(-z^2/4) \mathbb{E}\{|\gamma^\top \mathbf{a}|^2 \exp(\|\gamma - \mathbf{u}\|^2/4)\} \\ &\leq (2 + |\mathbf{u}^\top \mathbf{a}|^2) \exp(-z^2/4 + q/2 + \|\mathbf{u}\|^2/2) \end{aligned}$$

and (7.2) follows. \square

The next result explains the concentration effect for the norm $\|\xi\|^2$ of a Gaussian vector. We use a version from Spokoiny (2012).

Lemma 7.2. *For each \mathbf{x} ,*

$$\mathbb{P}(\|\gamma\| \geq z(q, \mathbf{x})) \leq \exp(-\mathbf{x}), \quad \mathbb{P}(\|\gamma\| \leq z_1(q, \mathbf{x})) \leq \exp(-\mathbf{x}),$$

where

$$z^2(q, \mathbf{x}) \stackrel{\text{def}}{=} q + \sqrt{6.6q\mathbf{x}} \vee (6.6\mathbf{x}), \quad z_1^2(q, \mathbf{x}) \stackrel{\text{def}}{=} q - 2\sqrt{q\mathbf{x}}.$$

The next lemma bounds from above the Kullback-Leibler divergence between two normal distributions.

Lemma 7.3. *Let $\mathbb{P}_0 = \mathcal{N}(0, I_q)$ and $\mathbb{P}_1 = \mathcal{N}(\beta, (U^\top U)^{-1})$ for some non-degenerated matrix U . If*

$$\|U^\top U - I_q\| \leq \epsilon \leq 1/2,$$

then

$$\begin{aligned} 2\mathcal{K}(\mathbb{P}_0, \mathbb{P}_1) &= -2\mathbb{E}_0 \log \frac{d\mathbb{P}_1}{d\mathbb{P}_0} \\ &\leq \text{tr}(U^\top U - I_q)^2 + (1 + \epsilon)\|\beta\|^2 \leq \epsilon^2 q + (1 + \epsilon)\|\beta\|^2. \end{aligned}$$

For any measurable set $A \subset \mathbb{R}^q$, it holds with $\gamma \sim \mathcal{N}(0, I_q)$

$$|\mathbb{P}_0(A) - \mathbb{P}_1(A)| = |\mathbb{P}(\gamma \in A) - \mathbb{P}(U(\gamma - \beta) \in A)| \leq \sqrt{\mathcal{K}(\mathbb{P}_0, \mathbb{P}_1)/2}.$$

Proof. It holds

$$2 \log \frac{d\mathbb{P}_1}{d\mathbb{P}_0}(\gamma) = \log \det(U^\top U) - (\gamma - \beta)^\top U^\top U (\gamma - \beta) + \|\gamma\|^2$$

with γ standard normal and

$$2\mathcal{K}(\mathbb{P}_0, \mathbb{P}_1) = -2\mathbb{E}_0 \log \frac{d\mathbb{P}_1}{d\mathbb{P}_0} = -\log \det(U^\top U) + \text{tr}(U^\top U - I_q) + \beta^\top U^\top U \beta.$$

Let a_j be the j th eigenvalue of $U^\top U - I_q$. $\|U^\top U - I_q\| \leq \epsilon \leq 1/2$ yields $|a_j| \leq 1/2$ and

$$\begin{aligned} 2\mathcal{K}(\mathbb{P}_0, \mathbb{P}_1) &= \beta^\top U^\top U \beta + \sum_{j=1}^q \{a_j - \log(1 + a_j)\} \leq (1 + \epsilon)\|\beta\|^2 + \sum_{j=1}^q a_j^2 \\ &\leq (1 + \epsilon)\|\beta\|^2 + \text{tr}(U^\top U - I_q)^2 \leq (1 + \epsilon)\|\beta\|^2 + \epsilon^2 q. \end{aligned}$$

This implies by Pinsker's inequality

$$\sup_A |\mathbb{P}_0(A) - \mathbb{P}_1(A)| \leq \sqrt{\frac{1}{2} \mathcal{K}(\mathbb{P}_0, \mathbb{P}_1)}$$

as required. \square

7.2 Proof of Theorem 6.2

Define $\mathbf{u}(\mathbf{v}) = \mathbf{b} \|\mathcal{D}_0(\mathbf{v} - \mathbf{v}^*)\|^2/2$. Now, by a change of variables, one obtains

$$\begin{aligned} &\frac{\mathbf{b}^{p/2} \det(\mathcal{D}_0)}{(2\pi)^{p/2}} \int_{\mathcal{R} \setminus \mathcal{R}_0(\mathbf{r}_0)} \exp\{-\mathbf{u}(\mathbf{v})\} d\mathbf{v} \\ &\leq \frac{\mathbf{b}^{p/2} \det(\mathcal{D}_0)}{(2\pi)^{p/2}} \int_{\mathcal{R} \setminus \mathcal{R}_0(\mathbf{r}_0)} \exp\{-\mathbf{b} \|\mathcal{D}_0(\mathbf{v} - \mathbf{v}^*)\|^2/2\} d\mathbf{v} = \mathbb{P}(\|\gamma\|^2 \geq \mathbf{b} \mathbf{r}_0^2). \end{aligned}$$

For the integral in the nominator of (6.8), it holds on $\Omega(\mathbf{x})$ by (6.5)

$$\int_{\mathcal{R} \setminus \mathcal{R}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v} \leq \int_{\mathcal{R} \setminus \mathcal{R}_0(\mathbf{r}_0)} \exp\{-\mathbf{u}(\mathbf{v})\} d\mathbf{v}.$$

For integral in the denominator it holds

$$\begin{aligned} & \int_{\mathcal{R}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v} \\ & \geq \exp\{-\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi})\} \int_{\mathcal{R}_0(\mathbf{r}_0)} \exp\{\mathbb{L}(\mathbf{v}, \mathbf{v}^*) + m(\boldsymbol{\xi})\} d\mathbf{v}. \end{aligned} \quad (7.3)$$

Inequality (7.3) implies by definition of $\nu(\mathbf{r}_0)$:

$$\int_{\mathcal{R}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v} \geq \exp\{-\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi}) - \nu(\mathbf{r}_0)\}. \quad (7.4)$$

The bound (7.4) for the local integral $\int_{\mathcal{R}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v}$ implies that

$$\rho^*(\mathbf{r}_0) \leq \exp\{\Delta(\mathbf{r}_0, \mathbf{x}) + \nu(\mathbf{r}_0) + m(\boldsymbol{\xi})\} \int_{\mathcal{R} \setminus \mathcal{R}_0(\mathbf{r}_0)} \exp\{-\mathbf{u}(\mathbf{v})\} d\mathbf{v}.$$

Finally

$$\exp\{m(\boldsymbol{\xi})\} = \exp\{-\|\boldsymbol{\xi}\|^2/2\} (2\pi)^{-p/2} \det(\mathcal{D}_0) \leq (2\pi)^{-p/2} \det(\mathcal{D}_0)$$

and the assertion (6.6) follows. The bound (6.7) is also straightforward

$$\begin{aligned} \nu(\mathbf{r}_0) &= -\log \mathbb{P}(\|\boldsymbol{\gamma} + \boldsymbol{\xi}\| \leq \mathbf{r}_0 \mid \mathbf{Y}) \leq -\log \mathbb{P}(\|\boldsymbol{\gamma}\| + \|\boldsymbol{\xi}\| \leq \mathbf{r}_0 \mid \mathbf{Y}) \\ &\leq -\log \mathbb{P}(\|\boldsymbol{\gamma}\| \leq z(p, \mathbf{x}) \mid \mathbf{Y}) \leq 2e^{-\mathbf{x}}. \end{aligned}$$

7.3 Proof of Theorem 6.4

Obviously $\{\boldsymbol{\theta} \notin \Theta_0(\mathbf{r}_0), \mathbf{v} \in \mathcal{R}\} \subset \{\mathcal{R} \setminus \mathcal{R}_0(\mathbf{r}_0)\}$. Therefore, it holds for the integral in the nominator of (6.8) in a view of (6.5)

$$\int_{\mathcal{R}} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} \mathbb{I}\{\boldsymbol{\theta} \notin \Theta_0(\mathbf{r}_0)\} d\mathbf{v} \leq \int_{\mathcal{R} \setminus \mathcal{R}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v}.$$

For the denominator, the inclusion $\mathcal{R}_0(\mathbf{r}_0) \subset \{\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0), \mathbf{v} \in \mathcal{R}\}$ and (6.5) imply

$$\int_{\mathcal{R}} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} \mathbb{I}\{\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)\} d\mathbf{v} \geq \int_{\mathcal{R}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v}.$$

Finally

$$\rho(\mathbf{r}_0) = \frac{\int_{\mathcal{R}} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} \mathbb{I}\{\boldsymbol{\theta} \notin \Theta_0(\mathbf{r}_0)\} d\mathbf{v}}{\int_{\mathcal{R}} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} \mathbb{I}\{\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)\} d\mathbf{v}} \leq \frac{\int_{\mathcal{R} \setminus \mathcal{R}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v}}{\int_{\mathcal{R}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v}} = \rho^*(\mathbf{r}_0),$$

and the assertion follows from Theorem 6.2.

7.4 Proof of Theorem 6.5

We use that $\mathbb{L}(\mathbf{v}, \mathbf{v}^*) = \boldsymbol{\xi}^\top \mathcal{D}_0(\mathbf{v} - \mathbf{v}^*) - \|\mathcal{D}_0(\mathbf{v} - \mathbf{v}^*)\|^2/2$ is proportional to the density of a Gaussian distribution. More precisely, define

$$m(\boldsymbol{\xi}) \stackrel{\text{def}}{=} -\|\boldsymbol{\xi}\|^2/2 + \log(\det \mathcal{D}_0) - p \log(\sqrt{2\pi}).$$

Then

$$m(\boldsymbol{\xi}) + \mathbb{L}(\mathbf{v}, \mathbf{v}^*) = -\|\mathcal{D}_0(\mathbf{v} - \mathbf{v}^*) - \boldsymbol{\xi}\|^2/2 + \log(\det \mathcal{D}_0) - p \log(\sqrt{2\pi})$$

is (conditionally on \mathbf{Y}) the log-density of the normal law with the mean $\mathbf{v}_0 = \mathbf{v}^* + \mathcal{D}_0^{-1}\boldsymbol{\xi}$ and the covariance matrix \mathcal{D}_0^{-2} . If we perform integration and leave only $\boldsymbol{\theta}$ part of \mathbf{v} then $m(\boldsymbol{\xi}) + \mathbb{L}(\mathbf{v}, \mathbf{v}^*)$ is (conditionally on \mathbf{Y}) the log-density of the normal law with the mean $\boldsymbol{\theta}^\circ = \check{D}_0^{-1}\check{\boldsymbol{\xi}} + \boldsymbol{\theta}^*$ and the covariance matrix \check{D}_0^{-2} . So, for any nonnegative function $f : \mathbb{R}^q \rightarrow \mathbb{R}_+$ we get

$$\begin{aligned} & \int_{\mathcal{R}} \exp\{L(\mathbf{v}, \mathbf{v}^*) + m(\boldsymbol{\xi})\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v} \\ &= \int_{\mathcal{R}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*) + m(\boldsymbol{\xi})\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v} \\ &+ \int_{\mathcal{R} \setminus \mathcal{R}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*) + m(\boldsymbol{\xi})\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v} \\ &= (1 + \rho_f(\mathbf{r}_0)) \int_{\mathcal{R}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*) + m(\boldsymbol{\xi})\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v} \\ &\leq e^{\Delta(\mathbf{r}_0, \mathbf{x}) + \rho_f(\mathbf{r}_0)} \int_{\mathcal{R}_0(\mathbf{r}_0)} \exp\{\mathbb{L}(\mathbf{v}, \mathbf{v}^*) + m(\boldsymbol{\xi})\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v} \\ &\leq e^{\Delta(\mathbf{r}_0, \mathbf{x}) + \rho_f(\mathbf{r}_0)} \int_{\mathbb{R}^p} \exp\{\mathbb{L}(\mathbf{v}, \mathbf{v}^*) + m(\boldsymbol{\xi})\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v} \\ &= e^{\Delta(\mathbf{r}_0, \mathbf{x}) + \rho_f(\mathbf{r}_0)} \mathbb{E}f(\boldsymbol{\gamma}). \end{aligned}$$

Thus,

$$\int_{\mathcal{R}} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v} \leq \exp\{\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi}) + \rho_f(\mathbf{r}_0)\} \mathbb{E}f(\boldsymbol{\gamma}). \quad (7.5)$$

Now (7.5) and (7.4) imply

$$\frac{\int_{\mathcal{R}} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} f(\check{D}_\epsilon(\boldsymbol{\theta} - \boldsymbol{\theta}_\epsilon)) d\mathbf{v}}{\int_{\mathcal{R}} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v}} \leq \exp\{2\Delta(\mathbf{r}_0, \mathbf{x}) + \nu(\mathbf{r}_0) + \rho_f(\mathbf{r}_0)\} \mathbb{E}f(\boldsymbol{\gamma})$$

and (6.9) follows by definition of $\Delta^+(\mathbf{r}_0, \mathbf{x})$.

7.5 Proof of Corollary 6.6

As a direct implication of (6.9) one easily gets

$$\mathbb{E}^\circ |\boldsymbol{\lambda}^\top \check{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ)|^2 \leq \exp(\Delta^+(\mathbf{r}_0, \mathbf{x})) \|\boldsymbol{\lambda}\|^2.$$

The only important step is to show that $\rho_{x^2}(\mathbf{r}_0)$ is small. Denote

$$\boldsymbol{\lambda}_0 = \mathcal{D}_0^{-1} \begin{pmatrix} \check{D}_0 \boldsymbol{\lambda} \\ \mathbf{0} \end{pmatrix}$$

and $\mathbf{0}$ is a zero vector of dimension $(p - q)$. We proceed separately with the nominator and denominator. For the nominator by (7.1) and (7.2) on $\Omega(\mathbf{x})$

$$\begin{aligned} & \int_{\mathcal{R} \setminus \mathcal{R}_0(\mathbf{r}_0)} |\boldsymbol{\lambda}^\top \check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)|^2 \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v} \\ & \leq \int_{\mathcal{R} \setminus \mathcal{R}_0(\mathbf{r}_0)} |\boldsymbol{\lambda}^\top \check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)|^2 \exp\{-\mathbf{b} \|\mathcal{D}_0(\mathbf{v} - \mathbf{v}^*)\|^2/2\} d\mathbf{v} \\ & = \int_{\mathcal{R} \setminus \mathcal{R}_0(\mathbf{r}_0)} |\boldsymbol{\lambda}_0^\top \mathcal{D}_0(\mathbf{v} - \mathbf{v}_0)|^2 \exp\{-\mathbf{b} \|\mathcal{D}_0(\mathbf{v} - \mathbf{v}^*)\|^2/2\} d\mathbf{v} \\ & = \exp\{-(p/2 + 2) \log \mathbf{b} - \log(\det \mathcal{D}_0) + p \log(\sqrt{2\pi})\} \mathbb{E} |\boldsymbol{\lambda}_0^\top (\boldsymbol{\gamma} + \boldsymbol{\xi})|^2 \mathbb{I}(\|\boldsymbol{\gamma}\|^2 \geq \mathbf{b} \mathbf{r}_0^2) \\ & \leq (4 + 2\|\boldsymbol{\xi}\|^2) \exp\{-(p/2 + 2) \log \mathbf{b} - \log(\det \mathcal{D}_0) + p \log(\sqrt{2\pi}) - \mathbf{b} \mathbf{r}_0^2/4 + p/2\} \|\boldsymbol{\lambda}_0\|^2 \\ & = \exp\{2 \log 2 + \log(1 + \|\boldsymbol{\xi}\|^2/2) - (p/2 + 2) \log \mathbf{b} - \log(\det \mathcal{D}_0) + p \log(\sqrt{2\pi}) - \mathbf{b} \mathbf{r}_0^2/4 + p/2\} \|\boldsymbol{\lambda}_0\|^2 \\ & \leq \exp\{\|\boldsymbol{\xi}\|^2/2 + (p/2 + 2) \log(\mathbf{e}/\mathbf{b}) - \log(\det \mathcal{D}_0) + p \log(\sqrt{2\pi}) - \mathbf{b} \mathbf{r}_0^2/4\} \|\boldsymbol{\lambda}_0\|^2 \\ & \leq \exp\{-\log(\det \mathcal{D}_0) + p \log(\sqrt{2\pi}) - \mathbf{x}\} \|\boldsymbol{\lambda}_0\|^2 \end{aligned}$$

for $\mathbf{b} \mathbf{r}_0^2 \geq (2p + 4) \log(\mathbf{e}/\mathbf{b}) + 2z(B, \mathbf{x}) + 4\mathbf{x}$. For the denominator, it holds on $\Omega(\mathbf{x})$

$$\begin{aligned} & \int_{\mathcal{R}_0(\mathbf{r}_0)} |\boldsymbol{\lambda}^\top \check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)|^2 \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v} \\ & \geq e^{-\Delta(\mathbf{r}_0, \mathbf{x})} \int_{\mathcal{R}_0(\mathbf{r}_0)} |\boldsymbol{\lambda}^\top \check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)|^2 \exp\{\mathbb{L}(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v} \\ & = e^{-\Delta(\mathbf{r}_0, \mathbf{x})} \int_{\mathcal{R}_0(\mathbf{r}_0)} |\boldsymbol{\lambda}_0^\top \mathcal{D}_0(\mathbf{v} - \mathbf{v}_0)|^2 \exp\{\mathbb{L}(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v} \\ & = e^{-\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi})} \mathbb{E} |\boldsymbol{\lambda}_0^\top (\boldsymbol{\gamma} + \boldsymbol{\xi})|^2 \mathbb{I}(\|\boldsymbol{\gamma} + \boldsymbol{\xi}\|^2 \leq \mathbf{r}_0^2) \\ & = e^{-\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi})} \left\{ \mathbb{E} |\boldsymbol{\lambda}_0^\top (\boldsymbol{\gamma} + \boldsymbol{\xi})|^2 - \mathbb{E} |\boldsymbol{\lambda}_0^\top (\boldsymbol{\gamma} + \boldsymbol{\xi})|^2 \mathbb{I}(\|\boldsymbol{\gamma} + \boldsymbol{\xi}\|^2 \geq \mathbf{r}_0^2) \right\} \\ & \geq e^{-\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi})} \left\{ \|\boldsymbol{\lambda}_0\|^2 + |\boldsymbol{\lambda}_0^\top \boldsymbol{\xi}|^2 - 2\|\boldsymbol{\lambda}_0\|^2 \exp\{-\mathbf{r}_0^2/4 + p/2 + \|\boldsymbol{\xi}\|^2/2\} \right\} \\ & \geq e^{-\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi})} \|\boldsymbol{\lambda}_0\|^2 \{1 - 2e^{-\mathbf{x}}\} \geq \exp\{-\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi}) - 4e^{-\mathbf{x}}\} \|\boldsymbol{\lambda}_0\|^2 \end{aligned}$$

for $\mathbf{r}_0^2 \geq 2p + 2z_B(\mathbf{x}) + 4\mathbf{x}$ on $\Omega(\mathbf{x})$. This yields on $\Omega(\mathbf{x})$

$$\begin{aligned} \rho_{x^2}(\mathbf{r}_0) &= \frac{\int_{\mathcal{R} \setminus \mathcal{R}_0(\mathbf{r}_0)} |\boldsymbol{\lambda}^\top \check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)|^2 \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v}}{\int_{\mathcal{R}_0(\mathbf{r}_0)} |\boldsymbol{\lambda}^\top \check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)|^2 \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v}} \\ &\leq \frac{2 \exp\{-\log(\det \mathcal{D}_0) + p \log(\sqrt{2\pi})\} - \mathbf{x}}{\exp\{-\Delta(\mathbf{r}_0, \mathbf{x}) + m(\boldsymbol{\xi}) - 4e^{-\mathbf{x}}\}} \|\boldsymbol{\lambda}_0\|^2 \\ &= 2 \exp\{\Delta(\mathbf{r}_0, \mathbf{x}) - \|\boldsymbol{\xi}\|^2/2 + 4e^{-\mathbf{x}} - \mathbf{x}\} \leq 2 \exp\{\Delta(\mathbf{r}_0, \mathbf{x}) + 4e^{-\mathbf{x}} - \mathbf{x}\}. \end{aligned}$$

7.6 Proof of Corollary 6.7

The first statement (6.14) follows from Theorem 6.5 with $f(\mathbf{u}) = \mathbb{I}(D_1 \check{D}_0^{-1} \mathbf{u} + \boldsymbol{\delta}_0 \in A)$. Further, it holds on $\Omega(\mathbf{x})$ for $\boldsymbol{\delta}_0 \stackrel{\text{def}}{=} D_1(\boldsymbol{\theta}^\circ - \widehat{\boldsymbol{\theta}})$

$$\|\boldsymbol{\delta}_0\|^2 = \|D_1(\boldsymbol{\theta}^\circ - \widehat{\boldsymbol{\theta}})\|^2 \leq (1 + \alpha) \|\check{D}_0(\boldsymbol{\theta}^\circ - \widehat{\boldsymbol{\theta}})\|^2 \leq (1 + \alpha) \beta^2.$$

For proving (6.11), we compute the Kullback–Leibler divergence between two multivariate normal distributions and apply Pinsker’s inequality. Let $\boldsymbol{\gamma}$ be standard normal in \mathbb{R}^q . The random variable $D_1 \check{D}_0^{-1} \boldsymbol{\gamma} + \boldsymbol{\delta}_0$ is normal with mean $\boldsymbol{\delta}_0$ and variance $B_1^{-1} \stackrel{\text{def}}{=} D_1 \check{D}_0^{-2} D_1$. Obviously

$$\|I_q - B_1\| = \|I_q - D_1^{-1} \check{D}_0^2 D_1^{-1}\| \leq \alpha.$$

Thus, by Lemma 7.3 for any measurable set A , it holds

$$\mathbb{P}(D_1 \check{D}_0^{-1} \boldsymbol{\gamma} + \boldsymbol{\delta}_0 \in A \mid \mathbf{Y}) \leq \mathbb{P}(\boldsymbol{\gamma} \in A) + \frac{1}{2} \sqrt{\alpha^2 q + (1 + \alpha)^2 \beta^2}.$$

7.7 Proof of Theorem 6.8

As in proof of Theorem 6.5, for any nonnegative function $f : \mathbb{R}^q \rightarrow \mathbb{R}_+$, it holds

$$\begin{aligned}
& \int_{\mathcal{R}} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) \mathbb{I}\{\boldsymbol{\theta} \in \Theta_0(\mathbf{r}_0)\} d\mathbf{v} \\
& \geq \int_{\mathcal{R}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v} \\
& \geq \exp\{-\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi})\} \int_{\mathcal{R}_0(\mathbf{r}_0)} \exp\{\mathbb{L}(\mathbf{v}, \mathbf{v}^*)\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v} \\
& \geq \exp\{-\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi})\} \int_{\mathbb{R}^p} \exp\{\mathbb{L}(\mathbf{v}, \mathbf{v}^*)\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v} \\
& \quad - \exp\{-\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi})\} \int_{\mathbb{R}^p \setminus \mathcal{R}_0(\mathbf{r}_0)} \exp\{\mathbb{L}(\mathbf{v}, \mathbf{v}^*)\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v} \\
& = \exp\{-\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi})\} (1 - \tilde{\rho}_f(\mathbf{r}_0)) \int_{\mathbb{R}^p} \exp\{\mathbb{L}(\mathbf{v}, \mathbf{v}^*)\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v} \\
& \geq \exp\{-\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi})\} (1 - \tilde{\rho}_f(\mathbf{r}_0)) \int_{\Theta_0(\mathbf{r}_0) \times \mathbb{R}^{(p-q)}} \exp\{\mathbb{L}(\mathbf{v}, \mathbf{v}^*)\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v} \\
& \geq \exp\{-\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi}) - 2\tilde{\rho}_f(\mathbf{r}_0)\} \mathbb{E}f(\boldsymbol{\gamma}) \mathbb{I}\{\|\boldsymbol{\gamma} + \check{\boldsymbol{\xi}}\| \leq \mathbf{r}_0\}, \tag{7.6}
\end{aligned}$$

Here we used that $1 - \alpha \geq e^{-2\alpha}$ for $0 \leq \alpha \leq \frac{1}{2}$. Similarly,

$$\begin{aligned}
& \int_{\mathcal{R}} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v} = \int_{\mathcal{R}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v} + \int_{\mathcal{R} \setminus \mathcal{R}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v} \\
& = \{1 + \rho^*(\mathbf{r}_0)\} \int_{\mathcal{R}_0(\mathbf{r}_0)} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v} \\
& \leq \{1 + \rho^*(\mathbf{r}_0)\} \exp\{\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi})\} \mathbb{P}(\|\boldsymbol{\gamma} + \boldsymbol{\xi}\| \leq \mathbf{r}_0 \mid \mathbf{Y}),
\end{aligned}$$

and finally

$$\int_{\mathcal{R}} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v} \leq \exp\{\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi}) + \nu(\mathbf{r}_0) + \rho^*(\mathbf{r}_0)\}. \tag{7.7}$$

The bounds (7.6) and (7.7) imply

$$\begin{aligned}
& \frac{\int_{\mathcal{R}} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} f(\check{D}_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)) d\mathbf{v}}{\int_{\mathcal{R}} \exp\{L(\mathbf{v}, \mathbf{v}^*)\} d\mathbf{v}} \\
& \geq \frac{\exp\{-\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi}) - 2\tilde{\rho}_f(\mathbf{r}_0)\} \mathbb{E}f(\boldsymbol{\gamma}) \mathbb{I}\{\|\boldsymbol{\gamma} + \check{\boldsymbol{\xi}}\| \leq \mathbf{r}_0\}}{\exp\{\Delta(\mathbf{r}_0, \mathbf{x}) - m(\boldsymbol{\xi}) + \nu(\mathbf{r}_0) + \rho^*(\mathbf{r}_0)\}} \\
& \geq \exp\{-2\Delta(\mathbf{r}_0, \mathbf{x}) - 2\tilde{\rho}_f(\mathbf{r}_0) - \nu(\mathbf{r}_0) - \rho^*(\mathbf{r}_0)\} \mathbb{E}f(\boldsymbol{\gamma}) \mathbb{I}\{\|\boldsymbol{\gamma} + \check{\boldsymbol{\xi}}\| \leq \mathbf{r}_0\}.
\end{aligned}$$

This yields (6.12).

7.8 Proof of Theorem 2.1

Due to our previous results, it is convenient to decompose the r.v. $\boldsymbol{\vartheta}$ in the form

$$\boldsymbol{\vartheta} = \boldsymbol{\vartheta} \mathbb{I}\{\boldsymbol{\vartheta} \in \Theta_0(\mathbf{r}_0)\} + \boldsymbol{\vartheta} \mathbb{I}\{\boldsymbol{\vartheta} \notin \Theta_0(\mathbf{r}_0)\} = \boldsymbol{\vartheta}^\circ + \boldsymbol{\vartheta}^c.$$

The large deviation results yields that the posterior distribution of the part $\boldsymbol{\vartheta}^c$ is negligible provided a proper choice of \mathbf{r}_0 . Below we show that $\boldsymbol{\vartheta}^\circ$ is nearly normal which yields the BvM result. Define

$$\boldsymbol{\vartheta}^\circ \stackrel{\text{def}}{=} \mathbb{E}^\circ \boldsymbol{\vartheta}, \quad \mathfrak{S}_\circ^2 \stackrel{\text{def}}{=} \text{Cov}(\boldsymbol{\vartheta}^\circ) \stackrel{\text{def}}{=} \mathbb{E}^\circ \{(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\circ)(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\circ)^\top\}.$$

It suffices to show that holds on $\Omega(\mathbf{x})$

$$\begin{aligned} \|\check{D}_0(\boldsymbol{\vartheta}^\circ - \boldsymbol{\theta}^\circ)\|^2 &\leq 2\Delta^* \\ \|I_q - \check{D}_0 \mathfrak{S}_\circ^2 \check{D}_0\| &\leq 2\Delta^*, \end{aligned}$$

where $\Delta^* = \max\{\Delta^\oplus, \Delta^\ominus\}$.

Consider $\boldsymbol{\eta} \stackrel{\text{def}}{=} \check{D}_0(\boldsymbol{\vartheta} - \boldsymbol{\theta}^\circ)$. Corollaries 6.6 and 6.9 yield for any $\boldsymbol{\lambda} \in \mathbb{R}^q$ that

$$\|\boldsymbol{\lambda}\|^2 \exp(-\Delta^-) \leq \mathbb{E}^\circ |\boldsymbol{\lambda}^\top \boldsymbol{\eta}|^2 \leq \|\boldsymbol{\lambda}\|^2 \exp(\Delta^+) \quad (7.8)$$

with $\Delta^- = \Delta^\ominus$ and $\Delta^+ = \Delta^\oplus$. Define the first two moments of $\boldsymbol{\eta}$:

$$\bar{\boldsymbol{\eta}} \stackrel{\text{def}}{=} \mathbb{E}^\circ \boldsymbol{\eta}, \quad S_\circ^2 \stackrel{\text{def}}{=} \mathbb{E}^\circ \{(\boldsymbol{\eta} - \bar{\boldsymbol{\eta}})(\boldsymbol{\eta} - \bar{\boldsymbol{\eta}})^\top\} = \check{D}_0 \mathfrak{S}_\circ^2 \check{D}_0.$$

Use the following technical statement.

Lemma 7.4. *Assume (7.8). Then with $\Delta^* = \max\{\Delta^+, \Delta^-\} \leq 1/2$*

$$\|\bar{\boldsymbol{\eta}}\|^2 \leq 2\Delta^*, \quad \|S_\circ^2 - I_q\| \leq 2\Delta^*. \quad (7.9)$$

Proof. Let \mathbf{u} be any unit vector in \mathbb{R}^q . We obtain from (7.8)

$$\exp(-\Delta^-) \leq \mathbb{E}^\circ |\mathbf{u}^\top \boldsymbol{\eta}|^2 \leq \exp(\Delta^+).$$

Note now that

$$\mathbb{E}^\circ |\mathbf{u}^\top \boldsymbol{\eta}|^2 = \mathbf{u}^\top S_\circ^2 \mathbf{u} + |\mathbf{u}^\top \bar{\boldsymbol{\eta}}|^2.$$

Hence

$$\exp(-\Delta^-) \leq \mathbf{u}^\top S_\circ^2 \mathbf{u} + |\mathbf{u}^\top \bar{\boldsymbol{\eta}}|^2 \leq \exp(\Delta^+). \quad (7.10)$$

In a similar way with $\mathbf{u} = \bar{\boldsymbol{\eta}}/\|\bar{\boldsymbol{\eta}}\|$ and $\boldsymbol{\gamma} \sim \mathcal{N}(0, \mathbf{I}_q)$

$$\mathbb{E}^\circ |\mathbf{u}^\top (\boldsymbol{\eta} - \bar{\boldsymbol{\eta}})|^2 \geq e^{-\Delta^-} \mathbb{E} |\mathbf{u}^\top (\boldsymbol{\gamma} - \bar{\boldsymbol{\eta}})|^2 = e^{-\Delta^-} (1 + \|\bar{\boldsymbol{\eta}}\|^2)$$

yielding

$$\mathbf{u}^\top \mathcal{S}_\circ^2 \mathbf{u} \geq (1 + \|\bar{\boldsymbol{\eta}}\|^2) \exp(-\Delta^-).$$

This inequality contradicts (7.10) if $\|\bar{\boldsymbol{\eta}}\|^2 > 2\Delta^* > 1$, and (7.9) follows. \square

The bound for the first moment implies with $\boldsymbol{\vartheta}^\circ = \mathbb{E}^\circ \boldsymbol{\vartheta}$

$$\|\check{D}_0(\boldsymbol{\vartheta}^\circ - \boldsymbol{\theta}^\circ)\|^2 \leq 2\Delta^*$$

while the second bound yields with

$$\|\check{D}_0 \mathfrak{S}_\circ^2 \check{D}_0 - \mathbf{I}_q\| \leq 2\Delta^*.$$

The last result follows from (6.10) and (6.13) with an additional assumption that \mathbf{x} is large enough to ensure $\Delta^+(\mathbf{r}_0, \mathbf{x}) \leq 2\Delta(\mathbf{r}_0, \mathbf{x}) + 5e^{-\mathbf{x}}$ and $\Delta^-(\mathbf{r}_0, \mathbf{x}) \geq 2\Delta(\mathbf{r}_0, \mathbf{x}) - 8e^{-\mathbf{x}}$.

7.9 Proof of Theorem 2.3

It suffices to check (2.6). First evaluate the ratio $\pi(\mathbf{v})/\pi(\mathbf{v}^*)$ for any $\mathbf{v} \in \mathcal{V}_0(\mathbf{r}_0)$. It holds

$$\log \frac{\pi(\mathbf{v})}{\pi(\mathbf{v}^*)} = -\|G\mathbf{v}\|^2/2 + \|G\mathbf{v}^*\|^2/2 = -(\mathbf{v} - \mathbf{v}^*)^\top G^2 \mathbf{v}^* - \|G(\mathbf{v} - \mathbf{v}^*)\|^2/2.$$

It follows from the definition of $\mathcal{V}_0(\mathbf{r}_0)$ and (2.8) that for $\mathbf{v} \in \mathcal{V}_0(\mathbf{r}_0)$

$$\|G(\mathbf{v} - \mathbf{v}^*)\|^2 = \|GD_0^{-1} \mathcal{D}_0(\mathbf{v} - \mathbf{v}^*)\|^2 \leq \|\mathcal{D}_0^{-1} G^2 \mathcal{D}_0^{-1}\| \mathbf{r}_0^2 \leq \varepsilon^2 \mathbf{r}_0^2.$$

Similarly

$$|(\mathbf{v} - \mathbf{v}^*)^\top G^2 \mathbf{v}^*| \leq \|G\mathbf{v}^*\| \cdot \|G(\mathbf{v} - \mathbf{v}^*)\| \leq \|G\mathbf{v}^*\| \cdot \|GD_0^{-1}\| \mathbf{r}_0 \leq \varepsilon \mathbf{r}_0 \|G\mathbf{v}^*\|.$$

This obviously implies

$$-\frac{1}{2} \varepsilon^2 \mathbf{r}_0^2 \leq \log \frac{\pi(\mathbf{v})}{\pi(\mathbf{v}^*)} \leq \varepsilon \mathbf{r}_0 \|G\mathbf{v}^*\|$$

and (2.6) follows with $\alpha(\mathbf{r}_0) = \max\{\varepsilon \mathbf{r}_0 \|G\mathbf{v}^*\|, \varepsilon^2 \mathbf{r}_0^2/2\}$.

7.10 Proof of Theorem 3.1

From finite dimensional theorem we have that

$$\begin{aligned}\|\check{D}_m(\bar{\boldsymbol{\vartheta}} - \boldsymbol{\theta}_m^\circ)\|^2 &\leq \Delta^*(\mathbf{r}_0, \mathbf{x}), \\ \|I_q - \check{D}_m \mathfrak{S}^2 \check{D}_m\| &\leq \Delta^*(\mathbf{r}_0, \mathbf{x}).\end{aligned}$$

If (B) is true, then

$$\begin{aligned}\|I_q - \check{D}_0 \mathfrak{S}^2 \check{D}_0\| &= \sup_{\|\gamma\|=1} |\gamma^\top (I_q - \check{D}_0 \mathfrak{S}^2 \check{D}_0) \gamma| \\ &= \sup_{\|\gamma\|=1} |(\check{D}_0 \gamma)^\top (\check{D}_0^{-2} - \mathfrak{S}^2) (\check{D}_0 \gamma)| \\ &\leq \sup_{\|\gamma\|=1} |(\check{D}_0 \gamma)^\top (\check{D}_0^{-2} - \check{D}_m^{-2}) (\check{D}_0 \gamma)| + \sup_{\|\gamma\|=1} |(\check{D}_0 \gamma)^\top (\check{D}_m^{-2} - \mathfrak{S}^2) (\check{D}_0 \gamma)| \\ &= \|I_q - \check{D}_0 \check{D}_m^{-2} \check{D}_0\| + \sup_{\|\gamma\|=1} |(\check{D}_m^{-1} \check{D}_0 \gamma)^\top (I_q - \check{D}_m \mathfrak{S}^2 \check{D}_m) (\check{D}_m^{-1} \check{D}_0 \gamma)| \\ &\leq \|I_q - \check{D}_0 \check{D}_m^{-2} \check{D}_0\| + \|I_q - \check{D}_m \mathfrak{S}^2 \check{D}_m\| \sup_{\|\gamma\|=1} |\gamma^\top \check{D}_0 \check{D}_m^{-2} \check{D}_0 \gamma| \\ &\leq \|I_q - \check{D}_0 \check{D}_m^{-2} \check{D}_0\| + \|I_q - \check{D}_m \mathfrak{S}^2 \check{D}_m\| \|\check{D}_0 \check{D}_m^{-2} \check{D}_0\| \\ &\leq \beta_m + (1 + \beta_m) \Delta^*(\mathbf{r}_0, \mathbf{x}).\end{aligned}$$

For the mean it follows

$$\begin{aligned}\|\check{D}_0(\bar{\boldsymbol{\vartheta}} - \boldsymbol{\theta}^\circ)\|^2 &\leq \|\check{D}_0(\bar{\boldsymbol{\vartheta}} - \boldsymbol{\theta}_m^\circ)\|^2 + \|\check{D}_0(\boldsymbol{\theta}_m^\circ - \boldsymbol{\theta}^\circ)\|^2 \\ &\leq \|(\check{D}_0 \check{D}_m^{-1}) \check{D}_m(\bar{\boldsymbol{\vartheta}} - \boldsymbol{\theta}_m^\circ)\|^2 + \|\check{D}_0(\boldsymbol{\theta}^* - \boldsymbol{\theta}_m^*)\|^2 \\ &\leq \|\check{D}_m^{-1} \check{D}_0^2 \check{D}_m^{-1}\| \|\check{D}_m(\bar{\boldsymbol{\vartheta}} - \boldsymbol{\theta}_m^\circ)\|^2 + \|\check{D}_0(\boldsymbol{\theta}^* - \boldsymbol{\theta}_m^*)\|^2 \\ &\leq (1 + \beta_m) \|\check{D}_m(\bar{\boldsymbol{\vartheta}} - \boldsymbol{\theta}_m^\circ)\|^2 + \|\check{D}_0(\boldsymbol{\theta}^* - \boldsymbol{\theta}_m^*)\|^2 \\ &\leq (1 + \beta_m) \Delta^*(\mathbf{r}_0, \mathbf{x}) + \alpha_m.\end{aligned}$$

Now we can conclude that for validity of our results we need α_m and β_m to be of order $\Delta^*(\mathbf{r}_0, \mathbf{x})$.

7.11 Proof of Theorem 4.1

The bracketing bound and the large deviation result of Theorem 6.1 apply if the sample size n fulfills $n \geq \mathbf{C}(p_n + \mathbf{x})$ for a fixed constant \mathbf{C} . It appears that the BvM result

requires a stronger condition. Indeed, in the regular i.i.d. case it holds

$$\delta(\mathbf{r}_0) \asymp \mathbf{r}_0/\sqrt{n}, \quad z_{\mathbb{H}}^2(\mathbf{x}_n) \asymp p_n + \mathbf{x}_n, \quad \omega \asymp 1/\sqrt{n}.$$

The radius \mathbf{r}_0 should fulfill $\mathbf{r}_0^2 \geq \mathbf{C}(p_n + \mathbf{x})$ to ensure the large deviation result. This yields

$$\Delta(\mathbf{r}_0, \mathbf{x}) = (\delta(\mathbf{r}_0) + 3\nu_0 z_{\mathbb{H}}^2(\mathbf{x}_n)\omega)\mathbf{r}_0^2 \geq \mathbf{C}\sqrt{(p_n^3 + \mathbf{x})/n}.$$

If we fix $\mathbf{x} = \mathbf{C}p_n$, our BvM result effectively requires the condition $p_n^3/n \rightarrow 0$ as $n \rightarrow \infty$.

7.12 Proof of Theorem 4.2

First we check that the required conditions of Section 2.1 are fulfilled in the considered example. This can be easily done if we slightly change the definition of the local set $\Upsilon_0(\mathbf{r}_0)$. Namely, for $\mathbf{u}^* = (u_1^*, \dots, u_{p_n}^*)^\top$, define $\Upsilon_0(\sqrt{\mathfrak{z}})$ as a rectangle

$$\Upsilon_0(\sqrt{\mathfrak{z}}) \stackrel{\text{def}}{=} \{\mathbf{u} : M_n \mathcal{K}(u_j, u_j^*) \leq \mathfrak{z}, j = 1, \dots, p_n\}.$$

Here $\mathcal{K}(u, u^*)$ is the Kullback-Leibler divergence for the Poisson family:

$$\mathcal{K}(u, u^*) = e^u(u - u^*) - e^u + e^{u^*}.$$

Lemma 7.5. *Let \mathfrak{z}_n be such that $2p_n e^{-\mathfrak{z}_n} \leq 1/2$. Then it holds*

$$\mathbb{P}(\tilde{\mathbf{u}} \in \Upsilon_0(\sqrt{\mathfrak{z}_n})) \geq 1 - 4p_n e^{-\mathfrak{z}_n}. \quad (7.11)$$

In particular, the choice $\mathfrak{z}_n = \mathbf{x}_n + \log(p_n)$ with $\mathbf{x}_n = \mathbf{C} \log n$ provides

$$\mathbb{P}(\tilde{\mathbf{u}} \in \Upsilon_0(\sqrt{\mathfrak{z}_n})) \geq 1 - 4e^{-\mathbf{x}_n}. \quad (7.12)$$

Proof. We use the bound from Polzehl and Spokoiny (2006)

$$\mathbb{P}(M_n \mathcal{K}(\tilde{u}_j, u_j^*) > \mathfrak{z}_n) \leq 2e^{-\mathfrak{z}_n}.$$

This yields

$$\mathbb{P}(\tilde{\mathbf{u}} \in \Upsilon_0(\sqrt{\mathfrak{z}_n})) \geq (1 - 2e^{-\mathfrak{z}_n})^{p_n}.$$

Now the elementary inequalities $\log(1 - \alpha) \geq -2\alpha$ for $0 \leq \alpha \leq 1/2$ and $e^{-\delta} \geq 1 - \delta$ for $\delta \geq 0$ applied with $\alpha_n = 2e^{-\mathfrak{z}_n}$ and $\delta_n = 2\alpha_n p_n$ imply

$$(1 - \alpha_n)^{p_n} = e^{\log(1 - \alpha_n)p_n} \geq e^{-2\alpha_n p_n} \geq 1 - 2\alpha_n p_n$$

and (7.11) follows. □

In the special case $u_1^* = \dots = u_{p_n}^* = u^*$, the set $\mathcal{Y}_0(\sqrt{\mathfrak{z}})$ is a cube which can be also viewed as a ball in the sup-norm. Moreover, if $\mathfrak{z}_n/(M_n e^{u^*}) \leq 1/2$, this cube is contained in the cube $\{\mathbf{u} : \|\mathbf{u} - \mathbf{u}^*\| \leq \sqrt{\mathfrak{z}_n/(M_n e^{u^*})}\}$ in view of $e^x - 1 - x \leq a^2 \leq 1/2$ for $|x| \leq a \leq 1$. The concentration bound (7.12) enables us to check the local conditions only on the cube $\mathcal{Y}_0(\sqrt{\mathfrak{z}_n})$. Especially the condition (\mathcal{ED}_1) is trivially fulfilled because $\zeta(\mathbf{u}) = \mathcal{L}(\mathbf{u}) - \mathbb{E}\mathcal{L}(\mathbf{u})$ is linear in \mathbf{u} and θ is a linear functional of \mathbf{u} . Condition (\mathcal{L}_0) can be checked on $\mathcal{Y}_0(\sqrt{\mathfrak{z}_n})$ with $\delta(\mathfrak{z}_n) = \sqrt{\mathfrak{z}_n/(M_n e^{u^*})}$.

It remains to compute the value \check{D}_0^2 . Define $\beta_n = p_n/M_n^{1/2} = p_n^{3/2}/n^{1/2}$. If $n = p_n^3$, then $\beta_n = 1$.

Lemma 7.6. *Let $v^* = 1/p_n$. Then it holds*

$$\check{D}_0^2 = p_n^2 \beta_n^{-2}.$$

Now we are ready to finalize the proof Theorem 4.2.

Proof. Let β_n be bounded. The definition implies

$$p_n(\theta - \tilde{\theta}_n) = \sum_{j=1}^{p_n} \log\left(\frac{v_j}{Z_j/M_n}\right)$$

The posterior distribution $v_j \mid \mathbf{Y}$ is $\text{Gamma}(\alpha_j, \mu_j)$ with $\alpha_j = 1 + Z_j$ and $\mu_j = \frac{\mu}{M_n \mu + 1}$. We use following decomposition

$$\frac{v_j}{Z_j/M_n} = \frac{M_n \mu_j \alpha_j}{\alpha_j - 1} (1 + \alpha_j^{-1/2} \gamma_j),$$

where

$$\gamma_j \stackrel{\text{def}}{=} (\alpha_j \mu_j^2)^{-1/2} (v_j - \alpha_j \mu_j)$$

has zero mean and unit variance. We can use the Taylor expansion

$$p_n(\theta - \tilde{\theta}_n) = \sum_{j=1}^{p_n} \log\left(1 - \frac{1}{M_n \mu + 1}\right) + \sum_{j=1}^{p_n} \log\left(1 + \frac{1}{\alpha_j - 1}\right) + \sum_{j=1}^{p_n} \log\left(1 + \alpha_j^{-1/2} \gamma_j\right).$$

Now take into account properties of real data distribution.

$$\alpha_j = \frac{M_n}{p_n} \left(1 + \sqrt{\frac{p_n}{M_n}} \delta_j\right),$$

where δ_j is asymptotically standard normal.

Suppose now that $\beta_n^3/\sqrt{p_n} \rightarrow 0$ as $p_n \rightarrow \infty$. Then $M_n/p_n = (\sqrt{p_n}/\beta_n^3)^{2/3} p_n^{2/3} \rightarrow \infty$ as $p_n \rightarrow \infty$. Thus for p_n sufficient large, $\alpha_j \approx M_n/p_n$. Moreover, it holds for p_n sufficiently large that $\max_{j=1, \dots, p_n} \alpha_j^{-1/2} |\gamma_j| \leq 1/2$ with a high probability. Below we

can restrict ourselves to the case when $\alpha_j^{-1/2}|\gamma_j| \leq 1/2$. This allows to use the Taylor expansion

$$\begin{aligned} p_n(\theta - \tilde{\theta}_n) &= \sum_{j=1}^{p_n} \log\left(1 - \frac{1}{M_n\mu + 1}\right) + \sum_{j=1}^{p_n} \log\left(1 + \frac{1}{\alpha_j - 1}\right) + \sum_{j=1}^{p_n} \log\left(1 + \frac{\gamma_j}{\sqrt{\alpha_j}}\right) \\ &= \sum_{j=1}^{p_n} \frac{1}{\alpha_j - 1} + \sum_{j=1}^{p_n} \frac{1}{\sqrt{\alpha_j}} \gamma_j - \sum_{j=1}^{p_n} \frac{1}{2\alpha_j} \gamma_j^2 + R. \end{aligned}$$

One can easily check that the remainder R is of order $\beta_n^3/\sqrt{p_n} \rightarrow 0$. Moreover, $p_n^{-1/2} \sum_{j=1}^{p_n} \gamma_j$ is asymptotically standard normal, while $p_n^{-1} \sum_{j=1}^{p_n} \gamma_j^2 \xrightarrow{P} 1$. CLT here can be easily checked because of the Lyapunov condition being valid. Also $\sum_{j=1}^{p_n} (\alpha_j - 1)^{-1} = \frac{p_n^2}{M_n} + o_n(\beta_n^2)$. Now check what happens if $\beta_n \rightarrow 0$

$$\beta_n^{-1} p_n(\theta - \tilde{\theta}_n) = \beta_n + \frac{1}{\sqrt{p_n}} \sum_{j=1}^{p_n} \gamma_j - \frac{\beta_n}{2p_n} \sum_{j=1}^{p_n} \gamma_j^2 + o_n(1) \xrightarrow{w} \mathcal{N}(0, 1).$$

Similarly, with $\beta_n \equiv \beta$,

$$\beta^{-1} p_n(\theta - \tilde{\theta}_n) = \beta + \frac{1}{\sqrt{p_n}} \sum_{j=1}^{p_n} \gamma_j - \frac{\beta}{2p_n} \sum_{j=1}^{p_n} \gamma_j^2 + o_n(1) \xrightarrow{w} \mathcal{N}(\beta/2, 1).$$

This proves the result for $\beta_n \equiv \beta$. Finally in the case when β_n grows to infinity, but $\beta_n^3/\sqrt{p_n} \rightarrow 0$, then $\beta_n^{-1}(\theta - \tilde{\theta}_n) \xrightarrow{P} \infty$. \square

7.13 Proof of Lemma 7.6

Proof. Let $\bar{u}_j = u_j - u_j^*$. Then

$$L(\mathbf{u}, \mathbf{u}^*) = \mathcal{L}(\mathbf{u}) - \mathcal{L}(\mathbf{u}^*) = \sum_{j=1}^{p_n} \{Z_j \bar{u}_j - M_n p_n^{-1} (e^{\bar{u}_j} - 1)\}.$$

The expected value of Z_j is M_n/p_n which leads to following expectation of likelihood:

$$\mathbb{E}L(\mathbf{u}, \mathbf{u}^*) = \frac{M_n}{p_n} \sum_{j=1}^{p_n} (\bar{u}_j - (e^{\bar{u}_j} - 1)) = -\frac{M_n}{p_n} \sum_{j=1}^{p_n} \frac{\bar{u}_j^2}{2} + O(\|\bar{\mathbf{u}}\|^3).$$

Then we substitute $\bar{u}_1 = p_n \bar{\theta} - \sum_{j=2}^{p_n} \bar{u}_j$, where $\bar{\theta} = \theta - \theta^*$. Thus we get

$$\mathbb{E}L(\mathbf{u}, \mathbf{u}^*) = -\frac{M_n}{p_n} \frac{1}{2} (p_n \bar{\theta} - \sum_{j=2}^{p_n} \bar{u}_j)^2 - \frac{M_n}{p_n} \sum_{j=2}^{p_n} \frac{\bar{u}_j^2}{2} + O(\|\bar{\mathbf{u}}\|^3).$$

This Taylor expansion allows us compute components of Fisher information matrix:

$$\mathcal{D}_0^2 = -\nabla^2 \mathbb{E} \mathcal{L}(\mathbf{u}^*) = \frac{M_n}{p_n} \begin{pmatrix} p_n^2 & -p_n & \cdots & \cdots & -p_n \\ -p_n & 2 & 1 & \cdots & 1 \\ \vdots & 1 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 1 \\ -p_n & 1 & \cdots & 1 & 2 \end{pmatrix}$$

The Fisher information for the target parameter θ can be computed as follows:

$$\check{D}_0^2 = M_n p_n (1 - \mathbf{e}^\top \mathbb{H}^{-1} \mathbf{e}),$$

where $\mathbf{e} = (1, \dots, 1)^\top$ and $\mathbb{H} = I + E$ with $E = \mathbf{e}\mathbf{e}^\top$ being the matrix of ones of size $(p_n - 1) \times (p_n - 1)$.

It follows

$$\mathbf{e}^\top \mathbb{H}^{-1} \mathbf{e} = \text{tr}(\mathbf{e}^\top \mathbb{H}^{-1} \mathbf{e}) = \text{tr}(\mathbb{H}^{-1} \mathbf{e}\mathbf{e}^\top) = \text{tr}((E + I)^{-1} E).$$

Further, $(E + I)^{-1} E = I - (E + I)^{-1}$ yielding

$$\mathbf{e}^\top \mathbb{H}^{-1} \mathbf{e} = \text{tr}\{I - (E + I)^{-1}\} = (p_n - 1) - \text{tr}\{(E + I)^{-1}\} = (p_n - 1) - \sum_{j=1}^{p_n} \lambda_j,$$

where λ_j are eigenvalues of matrix $(E + I)^{-1}$. It is easy to see that $\lambda_1 = p_n^{-1}$ while $\lambda_2 = \dots = \lambda_{p_n-1} = 1$. Thus

$$\begin{aligned} \mathbf{e}^\top \mathbb{H}^{-1} \mathbf{e} &= (p_n - 1) - \{p_n^{-1} + (p_n - 2)\} = 1 - p_n^{-1}, \\ \check{D}_0^2 &= M_n p_n (1 - \mathbf{e}^\top \mathbb{H}^{-1} \mathbf{e}) = M_n p_n \{1 - (1 - p_n^{-1})\} = M_n = p_n^2 \beta_n^{-2}, \end{aligned}$$

which completes the proof. \square

7.14 Proof of Lemma 5.5

It holds

$$\begin{aligned}
\|\mathcal{D}_0^{-1}(\mathcal{D}^2(\mathbf{v}) - \mathcal{D}_0^2)\mathcal{D}_0^{-1}\| &= \sup_{\gamma \in \mathbb{R}^p: \|\gamma\|=1} |\gamma^\top \mathcal{D}_0^{-1}(\mathcal{D}^2(\mathbf{v}) - \mathcal{D}_0^2)\mathcal{D}_0^{-1}\gamma| \\
&\leq \sup_{\gamma \in \mathbb{R}^p: \|\gamma\|=1} \left| \sum_{i=1}^n (d''(\Psi_i^\top \mathbf{v}) - d''(\Psi_i^\top \mathbf{v}^*)) \gamma^\top \mathcal{D}_0^{-1} \Psi_i \Psi_i^\top \mathcal{D}_0^{-1} \gamma \right| \\
&\leq \sup_{\gamma \in \mathbb{R}^p: \|\gamma\|=1} \sum_{i=1}^n |d''(\Psi_i^\top \mathbf{v}) - d''(\Psi_i^\top \mathbf{v}^*)| \gamma^\top \mathcal{D}_0^{-1} \Psi_i \Psi_i^\top \mathcal{D}_0^{-1} \gamma \\
&\leq \sup_{\gamma \in \mathbb{R}^p: \|\gamma\|=1} \sum_{i=1}^n L |\Psi_i^\top (\mathbf{v} - \mathbf{v}^*)| \gamma^\top \mathcal{D}_0^{-1} \Psi_i \Psi_i^\top \mathcal{D}_0^{-1} \gamma \\
&\leq LN^{-1/2} \|\mathcal{D}_0(\mathbf{v} - \mathbf{v}^*)\| \sup_{\gamma \in \mathbb{R}^p: \|\gamma\|=1} \gamma^\top \mathcal{D}_0^{-1} \left(\sum_{i=1}^n d''(\Psi_i^\top \mathbf{v}^*) \Psi_i \Psi_i^\top \right) \mathcal{D}_0^{-1} \gamma \\
&\leq L \frac{\mathbf{r}}{N_2^{1/2}}.
\end{aligned}$$

References

- Barron, A., Schervish, M. J., and Wasserman, L. (1996). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27:536–561.
- Bickel, P. J. and Kleijn, B. J. K. (2012). The semiparametric Bernstein-von Mises theorem. *Ann. Statist.*, 40(1):206–237.
- Bochkina, N. and Green, P. J. (2012). The Bernstein-von Mises theorem and non-regular models. *ArXiv e-prints*.
- Bontemps, D. (2011). Bernstein - von mises theorem for gaussian regression with increasing number of regressors. *The Annals of Statistics*, 39, No. 5:2557–2584.
- Boucheron, S. and Gassiat, E. (2009). A Bernstein-von Mises theorem for discrete probability distributions. *Electron. J. Stat.*, 3:114–148.
- Boucheron, S. and Massart, P. (2011). A high-dimensional Wilks phenomenon. *Probability Theory and Related Fields*, 150:405–433. 10.1007/s00440-010-0278-7.
- Castillo, I. (2012). A semiparametric Bernstein - von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields*, 152:53–99. 10.1007/s00440-010-0316-5.

- Castillo, I. and Nickl, R. (2013). Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.*, 41(4):1999–2028.
- Castillo, I. and Rousseau, J. (2013). A general bernstein–von mises theorem in semiparametric models. Available at arXiv:1305.4482 [math.ST].
- Cheng, G. and Kosorok, M. R. (2008). General frequentist properties of the posterior profile distribution. *The Annals of Statistics*, 36(4):1819–1853.
- Chernozhukov, V. and Hong, H. (2003). An mcmc approach to classical estimation. *Journal of Econometrics*, 115(2):293–346.
- Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Stat.*, 21(2):903–923.
- Freedman, D. (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Stat.*, 27(4):1119–1140.
- Ghosal, S. (1999). Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli*, 5(2):315–331.
- Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.*, 74(1):49–68.
- Johnstone, I. M. (2010). High dimensional Bernstein–von Mises: simple examples. In *Borrowing strength: theory powering applications—a Festschrift for Lawrence D. Brown*, volume 6 of *Inst. Math. Stat. Collect.*, pages 87–98. Inst. Math. Statist., Beachwood, OH.
- Kim, Y. (2006). The bernstein - von mises theorem for the proportional hazard model. *Ann. Statist.*, 34(4):1678–1700.
- Kim, Y. and Lee, J. (2004). A bernstein - von mises theorem in the nonparametric right-censoring model. *Ann. Statist.*, 32(4):1492–1512.
- Kleijn, B. J. K. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.*, 34(2):837–877.
- Kleijn, B. J. K. and van der Vaart, A. W. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic J. Statist.*, 6:354–381.
- Le Cam, L. and Yang, G. L. (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer in Statistics.

- Leahu, H. (2011). On the bernstein-von mises phenomenon in the gaussian white noise model. *Electronic Journal of Statistics*, 5:373–404.
- McCullagh, P. and Nelder, J. (1989). *Generalized linear models. 2nd ed.* Monographs on Statistics and Applied Probability. 37. London etc.: Chapman and Hall. xix, 511 p. .
- Polzehl, J. and Spokoiny, V. (2006). Propagation-separation approach for local likelihood estimation. *Probab. Theory Relat. Fields*, 135(3):335–362.
- Rivoirard, V. and Rousseau, J. (2012). Bernsteinvon mises theorem for linear functionals of the density. *Ann. Statist.*, 40(3):1489–1523.
- Schwartz, L. (1965). On Bayes procedures. *Probability Theory and Related Fields*, 4(1):10–26.
- Shen, X. (2002). Asymptotic normality of semiparametric and nonparametric posterior distributions. *J. Am Stat. Assoc.*, 97(457):222–235.
- Spokoiny, V. (2012). Parametric estimation. Finite sample theory. *Ann. Statist.*, 40(6):2877–2909. arXiv:1111.3029.
- van der Vaart, A. W. (2000). *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press.