Vladimir Spokoiny

Script to VL WS2014

February 16, 2015

Springer

Berlin Heidelberg New York
Hong Kong London
Milan Paris Tokyo

Contents

1

Quasi maximum likelihood estimation in linear models

1.1 Linear Modeling

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

$$Y_i = \Psi_i^\top \boldsymbol{\theta}^* + \varepsilon_i \tag{1.1}$$

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

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

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

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

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

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

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

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

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

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

In case 1 this expression can be rewritten as

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

In case 2 the expression is similar:

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

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

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^p} L(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} (\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta})^\top \Sigma^{-1}(\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}). \tag{1.2}$$

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

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

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

$$\widetilde{\boldsymbol{\theta}} = \bigl(\Psi \Sigma^{-1} \Psi^\top\bigr)^{-1} \Psi \Sigma^{-1} \boldsymbol{Y} = \mathcal{S}\boldsymbol{Y}, \tag{1.3}$$

where

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

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

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

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

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

where

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

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

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

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

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

1.1.1 Estimation under homogeneous noise assumption

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

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

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

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

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

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

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

1.1.2 Linear basis transformation

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

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

In other words, $\boldsymbol{f}^*$ belongs to the linear subspace in $\mathbb{R}^n$ spanned by the vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p$. It is clear that this assumption still holds if we select another basis in this subspace.

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

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

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

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

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

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

yielding the solutions

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

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

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

This yields

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

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

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

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

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

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

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

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

Hint: use the formula

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

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

1.1.3 Orthogonal and orthonormal design

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.1.4 Spectral representation

Consider a linear model

$$\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{1.5}$$

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

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

The original model reads after this transformation in the form

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

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

$$\boldsymbol{Z} = \Lambda \boldsymbol{u} + \Lambda^{1/2} \boldsymbol{\xi} \tag{1.6}$$

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

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

Exercise 1.1.3. Prove the result of Theorem 1.1.2.
Hint: select any U ensuring $U^\top \Psi\Psi^\top U = \Lambda$. Then

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

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

1.2 Properties of the response estimate $\widetilde{f}$

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

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

$$\Pi = \Psi^\top \left(\Psi\Psi^\top\right)^{-1} \Psi. \tag{1.7}$$

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

1.2.1 Decomposition into a deterministic and a stochastic component

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

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

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

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

1.2.2 Properties of the operator Π

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

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

(i) Π *is symmetric (self-adjoint), that is,* $\Pi^\top = \Pi$.

(ii) Π *is a projector in* $\mathbb{R}^n$, *i.e.* $\Pi^\top\Pi = \Pi^2 = \Pi$ *and* $\Pi(\mathbf{1}_n - \Pi) = 0$, *where* $\mathbf{1}_n$ *means the unity operator in* $\mathbb{R}^n$.

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

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

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

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

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

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

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

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

Proof. It holds

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

and

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

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

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

The second property means that Π is a projector in $\mathbb{R}^n$ and the fourth one means that the dimension of its image space is equal to p. The basis vectors $\psi_1, \ldots, \psi_p$ are the rows of the matrix Ψ. It is clear that

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

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

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

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

1.2.3 Quadratic loss and risk of the response estimation

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

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

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

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

$$\mathcal{R}(\widetilde{f}) = \mathbb{E}\wp(\widetilde{f}, f^*) = \mathbb{E}\bigl[(\widetilde{f} - f^*)^\top(\widetilde{f} - f^*)\bigr]. \tag{1.9}$$

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

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

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

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

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

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

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

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

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

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

1.2.4 Misspecified "colored noise"

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

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

Theorem 1.2.2. *Suppose that* $I\!\!E\boldsymbol{\varepsilon} = 0$ *and* $\operatorname{Var}(\boldsymbol{\varepsilon}) = \Sigma_0$. *Then the loss* $\wp(\widetilde{\boldsymbol{f}}, \boldsymbol{f})$ *and the risk* $\mathcal{R}(\widetilde{\boldsymbol{f}})$ *of the LSE* $\widetilde{\boldsymbol{f}}$ *fulfill*

$$\wp(\widetilde{\boldsymbol{f}}, \boldsymbol{f}^*) = \|\boldsymbol{f}^* - \Pi \boldsymbol{f}^*\|^2 + \|\Pi \boldsymbol{\varepsilon}\|^2,$$
$$\mathcal{R}(\widetilde{\boldsymbol{f}}) = \|\boldsymbol{f}^* - \Pi \boldsymbol{f}^*\|^2 + \operatorname{tr}\big(\Pi \Sigma_0 \Pi\big).$$

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

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

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

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

as required.

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

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

$$\widetilde{\boldsymbol{\theta}} = \mathcal{S}\boldsymbol{Y} = \mathcal{S}\bigl(\boldsymbol{f}^* + \boldsymbol{\varepsilon}\bigr) = \mathcal{S}\boldsymbol{f}^* + \mathcal{S}\boldsymbol{\varepsilon}. \tag{1.10}$$

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

1.3.1 Properties of the stochastic component

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

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

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

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

$$W^2 = \sigma^2\bigl(\Psi\Psi^\top\bigr)^{-1}, \qquad \mathbb{E}\|\boldsymbol{\delta}\|^2 = \operatorname{tr}(W^2) = \sigma^2\operatorname{tr}\bigl[\bigl(\Psi\Psi^\top\bigr)^{-1}\bigr]. \tag{1.11}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

1.3.2 Properties of the deterministic component

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

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

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

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

$$\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta}} I\!\!E L(\boldsymbol{\theta}). \tag{1.12}$$

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

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

Moreover,

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

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

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

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

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

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

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

Exercise 1.3.2. State the result of Theorems 1.3.2 and 1.3.3 for the MLE $\widetilde{\boldsymbol{\theta}}$ built in the model $\boldsymbol{Y} = \Psi^\top\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with $\operatorname{Var}(\boldsymbol{\varepsilon}) = \Sigma$.
Hint: check that the statements continue to apply with $\mathcal{S} = \bigl(\Psi\Sigma^{-1}\Psi^\top\bigr)^{-1}\Psi\Sigma^{-1}$.

The last results and the decomposition (1.10) explain the behavior of the estimate $\widetilde{\boldsymbol{\theta}}$ in a very general situation. The considered model is $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$. We assume a linear parametric structure and independent homogeneous noise. The estimation procedure means in fact a kind of projection of the data $\boldsymbol{Y}$ on a p-dimensional linear subspace in $I\!R^n$ spanned by the given basis vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p$. This projection, as a linear operator,

can be decomposed into a projection of the deterministic vector $\boldsymbol{f}^*$ and a projection of the random noise $\boldsymbol{\varepsilon}$. If the linear parametric assumption $\boldsymbol{f}^* \in \langle \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p \rangle$ is correct, that is, $\boldsymbol{f}^* = \theta_1^* \boldsymbol{\psi}_1 + \ldots + \theta_p^* \boldsymbol{\psi}_p$, then this projection keeps $\boldsymbol{f}^*$ unchanged and only the random noise is reduced via this projection. If $\boldsymbol{f}^*$ cannot be exactly expanded using the basis $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p$, then the procedure recovers the projection of $\boldsymbol{f}^*$ onto this subspace. The latter projection can be written as $\Psi^\top \boldsymbol{\theta}^*$ and the vector $\boldsymbol{\theta}^*$ can be viewed as the target of estimation.

1.3.3 Risk of estimation. R-efficiency

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

as required.

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

$$I\!\!E\bigl\{(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}})\widetilde{\boldsymbol{\theta}}^\top\bigr\} = I\!\!E(A - \mathcal{S})\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\mathcal{S}^\top = \sigma^2(A - \mathcal{S})\Psi^\top(\Psi\Psi^\top)^{-1} = 0,$$

and the assertion follows.

Exercise 1.3.3. Check the details of the proof of the theorem. Show that the statement $\operatorname{Var}(\widehat{\boldsymbol{\theta}}) \geq \operatorname{Var}\bigl(\widetilde{\boldsymbol{\theta}}\bigr)$ only uses that $\widehat{\boldsymbol{\theta}}$ is unbiased and that $I\!\!E\boldsymbol{Y} = \Psi^\top\boldsymbol{\theta}^*$ and $\operatorname{Var}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}_n$.

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

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

A colored noise

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

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

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

A misspecified LPA

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

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

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

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

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

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

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

1.3.4 The case of a misspecified noise

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

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

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

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

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

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

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

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

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

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

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

1.4 Linear models and quadratic log-likelihood

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

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

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

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

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

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

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

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

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

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

$$\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = D^{-2}\boldsymbol{\zeta} \tag{1.15}$$

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

$$I\!\!E\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$$
$$\operatorname{Var}\bigl(\widetilde{\boldsymbol{\theta}}\bigr) = D^{-2}V^2D^{-2}.$$

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

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

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

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

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

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

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

Therefore, it holds for the expectation $I\!\!EL(\boldsymbol{\theta})$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.4.1 Inference based on the maximum likelihood

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Proof. By direct calculus

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

So, $\boldsymbol{\zeta}$ is a linear transformation of a Gaussian vector $\boldsymbol{Y}$ and thus it is Gaussian as well. By Theorem 1.4.1, $I\!\!E\boldsymbol{\zeta} = 0$. Moreover, $\operatorname{Var}(\boldsymbol{\varepsilon}) = \Sigma$ implies

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

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

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

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

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

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

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

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

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

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

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

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

$$\mathcal{E}(\mathfrak{z}) = \bigl\{\boldsymbol{\theta} : L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \le \mathfrak{z}\bigr\} = \Bigl\{\boldsymbol{\theta} : \sup_{\boldsymbol{\theta}'} L(\boldsymbol{\theta}') - L(\boldsymbol{\theta}) \le \mathfrak{z}\Bigr\}. \tag{1.23}$$

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

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

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

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

Theorem 1.4.6. *Assume* $\boldsymbol{Y} = \Psi^\top\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$. *Then for every* $\mathtt{x} > 0$, *it holds with* $\varkappa \ge 6.6$

$$\begin{aligned} &I\!P\bigl(2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) > p + \sqrt{\varkappa\mathtt{x}p} \vee (\varkappa\mathtt{x})\bigr) \\ &\quad = I\!P\bigl(\bigl\|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\bigr\|^2 > p + \sqrt{\varkappa\mathtt{x}p} \vee (\varkappa\mathtt{x})\bigr) \le \exp(-\mathtt{x}). \end{aligned} \tag{1.24}$$

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

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

Corollary 1.4.1. *Assume* $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$. *Then for every* $\mathtt{x} > 0$, *it holds with* $2\mathfrak{z} = p + \sqrt{\varkappa \mathtt{x} p} \vee (\varkappa \mathtt{x})$ *for* $\varkappa \ge 6.6$

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

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

1.4.2 A misspecified LPA

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

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

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

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

Exercise 1.4.5. Prove the result of Theorem 1.4.7.

1.4.3 A misspecified noise structure

This section addresses the question about the features of the maximum likelihood in the case when the likelihood is built under a wrong assumption about the noise structure. As one can expect, the chi-squared result is not valid anymore in this situation and the distribution of the maximum likelihood depends on the true noise covariance. However, the nice geometric structure of the maximum likelihood manifested by Theorems 1.4.3

and 1.4.5 does not rely on the true data distribution and it is only based on our structural assumptions on the considered model. This helps to get rigorous results about the behaviors of the maximum likelihood and particularly about its concentration properties.

Theorem 1.4.8. *Let* $\widetilde{\boldsymbol{\theta}}$ *be built for the model* $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$, *while the true noise covariance is* Σ_0: $I\!\!E\boldsymbol{\varepsilon} = 0$ *and* $\operatorname{Var}(\boldsymbol{\varepsilon}) = \Sigma_0$. *Then*

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

where

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

Further,

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

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

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

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

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

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

This implies that

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

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

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

One can see that the chi-squared result is not valid any more if the noise structure is misspecified. An interesting question is whether the CS $\mathcal{E}(\mathfrak{z})$ can be applied in the

case of a misspecified noise under some proper adjustment of the value $\mathfrak{z}$. Surprisingly, the answer is not entirely negative. The reason is that the vector $\boldsymbol{\xi}$ from (1.25) is zero mean and its norm has a similar behavior as in the case of the correct noise specification: the probability $I\!P\bigl(\|\boldsymbol{\xi}\| > z\bigr)$ starts to degenerate when z^2 exceeds $I\!E\|\boldsymbol{\xi}\|^2$. A general bound from Theorem 6.6.1 in Section 6.6 implies the following bound for the coverage probability.

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

$$I\!P\bigl(2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) > \mathtt{p} + (2v\mathtt{x}^{1/2}) \vee (6a^*\mathtt{x})\bigr) \leq \exp(-\mathtt{x}).$$

Exercise 1.4.7. Show that an overestimation of the noise in the sense $\Sigma \geq \Sigma_0$ preserves the coverage probability for the CS $\mathcal{E}(\mathfrak{z}_\alpha)$, that is, if $2\mathfrak{z}_\alpha$ is the $1-\alpha$ quantile of χ^2_p, then $I\!P\bigl(\mathcal{E}(\mathfrak{z}_\alpha) \ni \boldsymbol{\theta}^*\bigr) \leq \alpha$.

1.5 Random design regression

Consider the linear regression equation

$$\boldsymbol{Y} = \boldsymbol{\Psi}^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}, \tag{1.26}$$

where $\boldsymbol{\theta}^* \in I\!R^p$ is the target parameter, $\boldsymbol{Y}$ is the n-vector of responses, $\boldsymbol{\varepsilon}$ is the n-vector of errors, and $\boldsymbol{\Psi} = (\Psi_1, \ldots, \Psi_n)$ is a $p \times n$ design matrix with columns $\Psi_i \in I\!R^p$. We suppose that the error vector $\boldsymbol{\varepsilon}$ satisfies

$$I\!E\boldsymbol{\varepsilon} = 0, \qquad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n. \tag{1.27}$$

The corresponding Gaussian log-likelihood looks as

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

yielding the MLE $\widetilde{\boldsymbol{\theta}} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}\boldsymbol{Y}$. Below we discuss the case of a random design $\boldsymbol{\Psi}$. The problem for the analysis comes from the inversion of the random matrix $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top$. The results below present some conditions under which this random matrix can be replaced by its expectation. We distinguish between two cases. The classical random design assumes that the columns Ψ_i of $\boldsymbol{\Psi}$ are independent. The other case is when $\boldsymbol{\Psi}$ is a sum of independent random matrices.

1.5.1 Independent measurements

Let $\Psi_1, \ldots, \Psi_n$ be independent. We assume that the design is non degenerate, so that the matrix $\boldsymbol{M}^2 \stackrel{\mathrm{def}}{=} I\!E\bigl(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\bigr)$ is positive. Consider

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

where

$$A_i = \boldsymbol{M}^{-1}\bigl\{\Psi_i\Psi_i^\top - I\!E(\Psi_i\Psi_i^\top)\bigr\}\boldsymbol{M}^{-1} \tag{1.28}$$

is a symmetric $p \times p$ random matrix with $I\!EA_i = 0$. Also define the variance parameter

$$S_n^2 \stackrel{\mathrm{def}}{=} \|I\!E(A_1^2 + \ldots + A_n^2)\|. \tag{1.29}$$

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

$$\|A_i\| \le u_n \qquad a.s. \tag{1.30}$$

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

$$M_1^2 \stackrel{\text{def}}{=} I\!\!E(\Psi_1 \Psi_1^\top),$$
$$\sigma_1^2 \stackrel{\text{def}}{=} I\!\!E\big(M_1^{-1}\Psi_1\Psi_1^\top M_1^{-1} - I_p\big)^2.$$

Also suppose that with probability one

$$\|M_1^{-1}\Psi_1\Psi_1^\top M_1^{-1} - I_p\| \le u^*.$$

Then it holds

$$\begin{aligned} \boldsymbol{M}^2 &= n\, M_1^2, \\ S_n^2 &= n^{-1}\sigma_1^2, \\ u_n &\le n^{-1}u^* \end{aligned} \tag{1.31}$$

The matrix Bernstein inequality; see Theorem 1.6.1, yields:

Theorem 1.5.1. *Suppose that Ψ_i are independent and A_i from (1.28) fulfill (1.30). Then with $\boldsymbol{M}^2 = I\!\!E(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)$ and S_n^2 defined by (1.29), it holds for all $z > 0$*

$$I\!\!P\big(\|\boldsymbol{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\boldsymbol{M}^{-1} - I_p\| > z\big) \le 2p\exp\Big\{-\frac{z^2}{2S_n^2 + 2u_n z/3}\Big\}.$$

If Ψ_i are i.i.d. and (1.31) holds then

$$I\!\!P\big(n^{1/2}\|\boldsymbol{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\boldsymbol{M}^{-1} - I_p\| > z\big) \le 2p\exp\Big\{-\frac{z^2}{2\sigma_1^2 + 2n^{-1/2}u^* z/3}\Big\}.$$

Proof. (please check)

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

$$n \ge \big(2\sigma_1^2\delta^{-2} + 2u\delta^{-1}/3\big)\big\{\mathtt{x} + \log(2p)\big\} \tag{1.32}$$

to ensure

$$I\!\!P\big(\|\boldsymbol{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\boldsymbol{M}^{-1} - I_p\| > \delta\big) \le \mathrm{e}^{-\mathtt{x}}. \tag{1.33}$$

If n and $\mathtt{x}$ are fixed, then one can claim (1.33) for

$$\delta = \sqrt{\frac{2\sigma_1^2}{n}\big(\mathtt{x} + \log(2p)\big)} + \frac{2u}{3n}\big(\mathtt{x} + \log(2p)\big) \tag{1.34}$$

Corollary 1.5.1. *Suppose Ψ_i are i.i.d. and (1.31) holds. If n fulfills (1.32) for some fixed δ and $\mathtt{x}$, then (1.33) holds true. Similarly, if n and $\mathtt{x}$ are fixed, then δ from (1.34) ensures (1.33).*

Proof. (please check).

The result (1.33) guarantees that on a set $\Omega(\mathtt{x})$ with $I\!P\bigl(\Omega(\mathtt{x})\bigr) \leq \mathrm{e}^{-\mathtt{x}}$, for any $\boldsymbol{\gamma} \in I\!R^{p}$

$$(1-\delta)\boldsymbol{\gamma}^{\top}\boldsymbol{M}^{2}\boldsymbol{\gamma} \leq \boldsymbol{\gamma}^{\top}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}\boldsymbol{\gamma} \leq (1+\delta)\boldsymbol{\gamma}^{\top}\boldsymbol{M}^{2}\boldsymbol{\gamma}. \tag{1.35}$$

(please check).

Now we apply this result to the MLE in the regression model (1.26). The moments of $\boldsymbol{\varepsilon}$ in (1.27) are understood conditionally on the design $\boldsymbol{\Psi}$. In addition, we assume that the errors $\boldsymbol{\varepsilon}$ are conditionally normal given the design $\boldsymbol{\Psi}$. The log-likelihood ratio can be written in the form

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

Introduce also an approximating quadratic log-likelihood defined by

$$\begin{aligned}\mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^{*}) &= \frac{1}{\sigma^{2}}(\boldsymbol{Y}-\boldsymbol{\Psi}^{\top}\boldsymbol{\theta}^{*})^{\top}\boldsymbol{\Psi}^{\top}(\boldsymbol{\theta}-\boldsymbol{\theta}^{*}) - \frac{1}{2\sigma^{2}}(\boldsymbol{\theta}-\boldsymbol{\theta}^{*})^{\top}\boldsymbol{M}^{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^{*}) \\ &= \nabla^{\top}(\boldsymbol{\theta}-\boldsymbol{\theta}^{*}) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^{*})^{\top}D^{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^{*})\end{aligned}$$

with the symmetric $p\times p$ matrix $D^{2} = \sigma^{2}\boldsymbol{M}^{2} = \sigma^{-2}I\!E(\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top})$ and the random p vector $\nabla = \sigma^{-2}\boldsymbol{\Psi}(\boldsymbol{Y}-\boldsymbol{\Psi}^{\top}\boldsymbol{\theta}^{*})$ satisfying

$$I\!E\bigl(\nabla \,\big|\, \boldsymbol{\Psi}\bigr) = 0, \qquad \mathrm{Var}\bigl(\nabla \,\big|\, \boldsymbol{\Psi}\bigr) = \sigma^{-2}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}.$$

Note that two expressions $L(\boldsymbol{\theta},\boldsymbol{\theta}^{*})$, $\mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^{*})$ differ only in the quadratic term. We also know that $\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}$ is close to it expectation $\boldsymbol{M}^{2} = I\!E\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}$; see (1.35). Moreover, the difference between them can be evaluated using the bound (1.33):

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

This helps to study the properties of the MLE $\widetilde{\boldsymbol{\theta}}$ and of the maximum log-likelihood $L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta})$. It holds

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

Define

$$\boldsymbol{\xi} \stackrel{\mathrm{def}}{=} D^{-1}\boldsymbol{\Psi}(\boldsymbol{Y}-\boldsymbol{\Psi}^{\top}\boldsymbol{\theta}^{*}).$$

Conditioned on the design, $\boldsymbol{\xi}$ behaves as a nearly standard Gaussian vector:

$$\mathbb{E}\big(\boldsymbol{\xi}\,\big|\,\boldsymbol{\Psi}\big) = 0, \qquad \mathrm{Var}\big(\boldsymbol{\xi}\,\big|\,\boldsymbol{\Psi}\big) = D^{-1}(\sigma^{-2}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)D^{-1}$$

and $\mathbb{E}\big\{D^{-1}(\sigma^{-2}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)D^{-1}\big\} = I_p$. Moreover, restricting to $\Omega(\mathtt{x})$ helps to bound the moment generating function of $\boldsymbol{\xi}$: if $\boldsymbol{\varepsilon}$ is Gaussian conditioned on $\boldsymbol{\Psi}$, then

$$\log \mathbb{E}\Big\{\exp\Big(\lambda\frac{\boldsymbol{\gamma}^\top\boldsymbol{\xi}}{\|D\boldsymbol{\gamma}\|}\Big)\,\mathbb{1}\big(\Omega(\mathtt{x})\big)\Big\} \le (1+\delta)^2\lambda^2/2.$$

(please check. this place is nontrivial). Under the PA (1.26) this implies

$$D\big(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\big) - \boldsymbol{\xi} = \big[D\big(\sigma^{-2}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\big)^{-1}D - I_p\big]\boldsymbol{\xi}$$

yielding by (1.33) on $\Omega(\mathtt{x})$

$$\big\|D\big(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\big) - \boldsymbol{\xi}\big\| \le \frac{\delta}{1-\delta}\|\boldsymbol{\xi}\|. \tag{1.36}$$

(please check).
Similarly

$$\begin{aligned} 2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\boldsymbol{\xi}\|^2 &= \big(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\big)^\top\sigma^{-2}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\big(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\big) - \|\boldsymbol{\xi}\|^2 \\ &= \boldsymbol{\xi}^\top\big[D^{-1}\big(\sigma^{-2}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\big)D^{-1} - I_p\big]\boldsymbol{\xi} \end{aligned}$$

yielding on $\Omega(\mathtt{x})$

$$\big|2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\boldsymbol{\xi}\|^2\big| \le \delta\|\boldsymbol{\xi}\|^2. \tag{1.37}$$

These arguments yield the following statement:

Corollary 1.5.2. *Consider the model (1.26) and suppose* $\|\boldsymbol{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\boldsymbol{M}^{-1} - I_p\| \le \delta$ *for* $\boldsymbol{M}^2 = \mathbb{E}(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)$ *and some* $\delta < 1/2$ *on a dominating set* $\Omega(\mathtt{x})$. *Then the MLE* $\widetilde{\boldsymbol{\theta}}$ *fulfills (1.36) and (1.37) on this set* $\Omega(\mathtt{x})$ *with* $\boldsymbol{\xi} = D^{-1}\boldsymbol{\Psi}(\boldsymbol{Y} - \boldsymbol{\Psi}^\top\boldsymbol{\theta}^*)$.

1.5.2 Aggregated random design

Here we discuss another random design setup for the regression model (1.26). Namely, assume that the design matrix $\boldsymbol{\Psi}$ can be represented as a sum of independent matrices $\Psi_1, \ldots, \Psi_n$:

$$\boldsymbol{\Psi} = \Psi_1 + \ldots + \Psi_n. \tag{1.38}$$

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

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

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

and

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

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

$$\bigl\| \boldsymbol{M}^{-1} (\Psi_i - I\!\!E\Psi_i) \bigr\| \le u_n \qquad a.s.$$

for a small constant u_n. Now consider

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

The matrix Bernstein inequality; see Theorem 1.6.1, yields for any $z \ge 0$

$$I\!\!P\bigl(\| \boldsymbol{M}^{-1} (\boldsymbol{\Psi} - I\!\!E\boldsymbol{\Psi}) \| \ge z \bigr) \le 2(p+q) \exp\Bigl\{ - \frac{z^2}{2S_n^2 + 2u_n z/3} \Bigr\} \tag{1.39}$$

In the case with i.i.d. Ψ_i, define

$$M_1^2 \stackrel{\text{def}}{=} I\!\!E\Psi_1 I\!\!E\Psi_1^\top,$$
$$\sigma_1^2 \stackrel{\text{def}}{=} M_1^{-1} I\!\!E(\Psi_1 \Psi_1^\top) M_1^{-1} - \boldsymbol{I}_p,$$

and suppose that

$$\| M_1^{-1} (\Psi_1 - I\!\!E\Psi_1) \| \le u_1.$$

Then it holds

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

and

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

The result (1.39) implies that $\boldsymbol{\Psi}$ is close to $I\!\!E\boldsymbol{\Psi}$. The MLE $\widetilde{\boldsymbol{\theta}}$ also involves the product $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top$ which has to be close to the corresponding product $\boldsymbol{M}^2 = I\!\!E\boldsymbol{\Psi}\, I\!\!E(\boldsymbol{\Psi}^\top)$. Below we assume that $\boldsymbol{M}$ is sufficiently large and bound the difference $\boldsymbol{M}^{-1} \boldsymbol{\Psi}\boldsymbol{\Psi}^\top \boldsymbol{M}^{-1} - I_p$.

Proposition 1.5.1. *Let* $\|\boldsymbol{M}^{-1}\big(\boldsymbol{\Psi} - I\!\!E\boldsymbol{\Psi}\big)\| \le \delta$ *for some* $\delta > 0$. *Then*

$$\big\|\boldsymbol{M}^{-1}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\boldsymbol{M}^{-1} - I_p\big\| \le \delta^2 + 2\delta.$$

Proof. One can bound

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

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

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

Therefore,

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

thus

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

and the result follows.

Corollary 1.5.2 applies in this situation without any change.

1.5.3 Application to instrumental regression

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

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

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

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

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

or in vector form

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

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

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

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

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

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

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

with

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

Define

$$\begin{aligned} \boldsymbol{Z} = I\!E_n\big[\boldsymbol{Y}\,\boldsymbol{\phi}(W)\big] &= n^{-1}\textstyle\sum_i Y_i\boldsymbol{\phi}(W_i) &&\in I\!R^q, \\ \boldsymbol{T}_n = I\!E_n\big[\boldsymbol{\psi}(X)\,\boldsymbol{\phi}(W)^\top\big] &= n^{-1}\textstyle\sum_i \boldsymbol{\psi}(X_i)\,\boldsymbol{\phi}(W_i)^\top &&\in I\!R^{q\times p}, \\ \boldsymbol{\varepsilon} = I\!E_n\big[\boldsymbol{\phi}(W)\,\boldsymbol{U}\big] &= n^{-1}\textstyle\sum_i \boldsymbol{\phi}(W_i)\,U_i &&\in I\!R^q. \end{aligned} \tag{1.40}$$

The original problems reduces to

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

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

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

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

with the random design $\boldsymbol{T}_n = n^{-1}\sum_i \boldsymbol{\psi}(X_i)\,\boldsymbol{\phi}(W_i)^\top$ so that the setup (1.38) applies. The corresponding least square estimator of $\boldsymbol{\theta}^*$ reads as

$$\widetilde{\boldsymbol{\theta}} = \big(\boldsymbol{T}_n\boldsymbol{T}_n^\top\big)^{-1}\boldsymbol{T}_n\boldsymbol{Z}. \tag{1.41}$$

The results of Proposition 1.5.1 justify that the random matrix $\boldsymbol{T}_n\boldsymbol{T}_n^\top$ is very close to the product $I\!E(\boldsymbol{T}_n)I\!E(\boldsymbol{T}_n^\top) = \boldsymbol{T}\boldsymbol{T}^\top$ and the theoretical study of the properties of the estimator $\widetilde{\boldsymbol{\theta}}$ can be done with $\boldsymbol{T}\boldsymbol{T}^\top$ in place of $\boldsymbol{T}_n\boldsymbol{T}_n^\top$ in (1.41). Similarly one can justify that the product $\boldsymbol{T}_n\boldsymbol{Z}$ behaves nearly as $\boldsymbol{T}\boldsymbol{Z}$.

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

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

and suppose that it holds almost surely

$$\| \boldsymbol{M}^{-1} (T_i - \boldsymbol{T}) \| \le u. \tag{1.42}$$

Theorem 1.5.2. *Let (Y_i, X_i, W_i) be i.i.d. and T_i from (1.40) fulfill (1.42). Then for any $z > 0$*

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

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

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

1.6 Matrix Bernstein inequality

This section collects some useful facts about deviation of stochastic matrices from their mean. We mainly follows **?**.

Theorem 1.6.1 (Matrix Bernstein). *Let $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_n$ be independent random matrices with common dimension $d_1 \times d_2$. Assume that each matrix has bounded deviation from its mean:*

$$\| \boldsymbol{S}_i - I\!\!E \boldsymbol{S}_i \| \le R, \qquad i = 1, \ldots, n.$$

Form the sum $\boldsymbol{Z} = \sum_{i=1}^n \boldsymbol{S}_i$, and introduce a variance parameter

$$\sigma^2 \stackrel{\text{def}}{=} \max\Bigl\{ \bigl\| I\!\!E \bigl[(\boldsymbol{Z} - I\!\!E \boldsymbol{Z})(\boldsymbol{Z} - I\!\!E \boldsymbol{Z})^\top \bigr] \bigr\|, \bigl\| I\!\!E \bigl[(\boldsymbol{Z} - I\!\!E \boldsymbol{Z})^\top (\boldsymbol{Z} - I\!\!E \boldsymbol{Z}) \bigr] \bigr\| \Bigr\}.$$

Then

$$I\!\!P\bigl(\| \boldsymbol{Z} - I\!\!E \boldsymbol{Z} \| > z \bigr) \le (d_1 + d_2) \exp\left\{ - \frac{z^2/2}{\sigma^2 + Rz/3} \right\}.$$

Furthermore,

$$I\!\!E \| \boldsymbol{Z} - I\!\!E \boldsymbol{Z} \| \le \sqrt{2\sigma^2 \log(d_1 + d_2)} + \frac{1}{3} R \log(d_1 + d_2).$$

2

Sieve model selection in linear models

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

2.1 Projection estimation. Loss and risk

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

2.1.1 A linear model

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

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

Here Ψ is the design matrix formed by a collection of basis vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p \in I\!R^n$, and p is the parameter dimension assumed to be large or even infinity. The parametric assumption is only an idealization, for the true data distribution $I\!P$ of the vector $\boldsymbol{Y} \in I\!R^n$ we assume that the observations Y_i are independent and the errors $\varepsilon_i = Y_i - I\!E Y_i$ have some moments. We also assume that the response vector $\boldsymbol{f} = I\!E\boldsymbol{Y}$ can be well approximated by $\Psi^\top \boldsymbol{\theta}$ for a proper choice of $\boldsymbol{\theta}$, or, equivalently,

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

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

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \|\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}\|^2 = \left(\Psi\Psi^\top\right)^{-1} \Psi \boldsymbol{Y} = \mathcal{S}\boldsymbol{Y}$$

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

2.1.2 Linear decomposition

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

$$\widetilde{\boldsymbol{\theta}} = \mathcal{S}\boldsymbol{Y} = \mathcal{S}\boldsymbol{f}^* + \mathcal{S}\varepsilon.$$

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

$$I\!\!E\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^*,$$

where the target $\boldsymbol{\theta}^*$ can be defined as the vector of coefficients of the best linear fit:

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

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

$$\operatorname{Var}(\widetilde{\boldsymbol{\theta}}) = I\!\!E\bigl\{(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\bigr\} = \bigl(\Psi\Psi^\top\bigr)^{-1}\Psi I\!\!E\bigl(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\bigr)\Psi^\top\bigl(\Psi\Psi^\top\bigr)^{-1} = \sigma^2\bigl(\Psi\Psi^\top\bigr)^{-1}.$$

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

$$I\!\!E\bigl[W(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\bigr]^2 = \operatorname{tr}\bigl[W(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top W^\top\bigr] = \sigma^2\operatorname{tr}\bigl[W\bigl(\Psi\Psi^\top\bigr)^{-1}W^\top\bigr].$$

2.1.3 Inhomogeneous errors

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

$$\begin{aligned}\operatorname{Var}(\widetilde{\boldsymbol{\theta}}) &= I\!\!E I\!\!E\bigl\{(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top\bigr\} = \bigl(\Psi\Psi^\top\bigr)^{-1}\Psi I\!\!E\bigl(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\bigr)\Psi^\top\bigl(\Psi\Psi^\top\bigr)^{-1} \\ &= \bigl(\Psi\Psi^\top\bigr)^{-1}\Psi\Sigma\Psi^\top\bigl(\Psi\Psi^\top\bigr)^{-1}.\end{aligned}$$

Exercise 2.1.1. Consider the regression model

$$Y_i = \theta_1^*\psi_1(X_i) + \ldots\theta_p^*\psi_p(X_i) + \varepsilon_i \tag{2.1}$$

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

- Compute $\widetilde{\boldsymbol{\theta}}$
- Show that $I\!\!E\widetilde{\boldsymbol{\theta}} = I\!\!E\widetilde{\boldsymbol{\theta}}_{LSE} = \boldsymbol{\theta}^*$.
- Compute the variance $\operatorname{Var}\bigl(\widetilde{\boldsymbol{\theta}}\bigr)$ and the variance $\operatorname{Var}\bigl(\widetilde{\boldsymbol{\theta}}_{LSE}\bigr)$;
- show that $\operatorname{Var}\bigl(\widetilde{\boldsymbol{\theta}}_{LSE}\bigr) \geq \operatorname{Var}\bigl(\widetilde{\boldsymbol{\theta}}\bigr)$;

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

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

$$\widetilde{\boldsymbol{\theta}} = \Psi \boldsymbol{Y}, \qquad \operatorname{Var}(\widetilde{\boldsymbol{\theta}}) = \sigma^2 \boldsymbol{I}_p .$$

and for any symmetric positive matrix W, it holds

$$I\!\!E\{W(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\}^2 = \sigma^2 \operatorname{tr}(WW^\top)$$

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

2.1.4 Linear decomposition

It follows

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

Here Π is the projector in the space $I\!R^n$ on the linear subspace spanned by the basis vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p$. The model equation $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$ implies the decomposition

$$\widetilde{\boldsymbol{f}} = \Pi \boldsymbol{f} + \Pi \boldsymbol{\varepsilon}.$$

It implies

$$I\!\!E \widetilde{\boldsymbol{f}} = \Pi \boldsymbol{f}.$$

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

$$\operatorname{Var}(\widetilde{\boldsymbol{f}}) = I\!\!E(\Pi \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \Pi) = \sigma^2 \Pi.$$

2.1.5 Quadratic loss. Bias-variance decomposition

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

$$\begin{aligned}
\varrho(\widetilde{\boldsymbol{f}}, \boldsymbol{f}^*) &= \|\widetilde{\boldsymbol{f}} - \boldsymbol{f}^*\|^2 = \|\Pi \boldsymbol{f} - \boldsymbol{f}^* + \Pi \boldsymbol{\varepsilon}\|^2 \\
&= \|(\boldsymbol{I}_p - \Pi)\boldsymbol{f}^*\|^2 + \|\Pi \boldsymbol{\varepsilon}\|^2 + 2(\Pi \boldsymbol{\varepsilon})^\top (\boldsymbol{I}_p - \Pi)\boldsymbol{f}^* \\
&= \|(\boldsymbol{I}_p - \Pi)\boldsymbol{f}^*\|^2 + \|\Pi \boldsymbol{\varepsilon}\|^2. \qquad (2.2)
\end{aligned}$$

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

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

as

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

The term $\|(I_p - \Pi)\boldsymbol{f}^*\|^2$ is usually called *the squared bias* and it describes the accuracy of approximation of $\boldsymbol{f}^*$ by its projection $\Pi\boldsymbol{f}^*$ on the space generated by the basis functions $\psi_1, \ldots, \psi_p$. The term $\sigma^2 p$ called *the variance* measures the statistical error related to estimation of p unknown coefficients in the decomposition of $\Pi\boldsymbol{f}^* = \theta_1^*\boldsymbol{\psi}_1 + \ldots + \theta_p^*\boldsymbol{\psi}_p$.

Exercise 2.1.2. Consider the nonparametric model

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

and the parametric approximation (2.1). Consider the qMLE (LSE) $\widetilde{\boldsymbol{f}}_{LSE} = \Pi\boldsymbol{Y}$ for $\Pi = \Psi^\top\left(\Psi\Psi^\top\right)^{-1}\Psi$.

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

2.1.6 Projection estimation and the model choice problem

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

We associate the rows of the design matrix Ψ with basis vectors in $\mathbb{R}^n$. By Ψ_m we denote a sub matrix of Ψ composed of the first m rows $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_m$. It corresponds to the reduced regression model

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

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

$$\widetilde{\boldsymbol{\theta}}_m = \bigl(\Psi_m \Psi_m^\top\bigr)^{-1} \Psi_m \boldsymbol{Y},$$
$$\widetilde{\boldsymbol{f}}_m = \Psi_m \bigl(\Psi_m \Psi_m^\top\bigr)^{-1} \Psi_m \boldsymbol{Y} = \Pi_m \boldsymbol{Y}$$

where Π_m is a projector in $\mathbb{R}^n$ on the subspace spanned by the basis vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_m$.

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

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

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

$$\widetilde{\boldsymbol{f}}_m = z_1 \boldsymbol{\psi}_1 + \ldots + z_m \boldsymbol{\psi}_m .$$

In other words, $\widetilde{\boldsymbol{f}}_m$ is just a projection of the observed vector $\boldsymbol{Y}$ onto the subspace L_m spanned by the first m basis vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_m$: $\mathrm{L}_m = \langle \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_m \rangle$. This explains the name of the method. Clearly one can study the properties of $\widetilde{\boldsymbol{\theta}}_m$ or $\widetilde{\boldsymbol{f}}_m$ using the methods of previous sections. However, one more question for this approach is still open: a proper choice of m. The standard way of accessing this issue is based on the analysis of the quadratic risk.

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

$$\mathcal{R}(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*) = \|\boldsymbol{f}^* - \Pi_m \boldsymbol{f}^*\|^2 + \sigma^2 m = \sigma^2 m + \sum_{j=m+1}^{p} \theta_j^{*2}. \tag{2.3}$$

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

$$m^* \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_m \mathcal{R}(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*) = \operatorname*{argmin}_m \bigl\{ \|\boldsymbol{f}^* - \Pi_m \boldsymbol{f}^*\|^2 + \sigma^2 m \bigr\}. \tag{2.4}$$

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

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

$$\|\widetilde{\boldsymbol{f}}_m - \boldsymbol{f}^*\|^2 = \|\Psi^\top\widetilde{\boldsymbol{\theta}}_m - \Psi^\top\boldsymbol{\theta}^*\|^2 = \|\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\|^2$$

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

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

2.2 Unbiased risk estimation

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

$$\widehat{m} = \operatorname*{argmin}_m \|\boldsymbol{Y} - \widetilde{\boldsymbol{f}}_m\|^2 = p.$$

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

Lemma 2.2.1. *Consider the projection estimator $\widetilde{\boldsymbol{f}}_m = \Pi_m\boldsymbol{Y}$. For two different values $m' > m$, the following statements hold:*

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

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

Proof. It obviously holds

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

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

Lemma 2.2.2. *It holds under homogeneous errors* $\boldsymbol{\varepsilon}$ *:*

- *For each* m

$$\mathbb{E}\|\boldsymbol{Y} - \Pi_m \boldsymbol{Y}\|^2 = \|\boldsymbol{f}^* - \Pi_m \boldsymbol{f}^*\|^2 + \sigma^2 (n - m). \tag{2.5}$$

- *For any* $m' > m$

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

Proof. Obvious.

The first term in the statement (2.5) is exactly the squared bias which is a good news: the empirical risk contains the same term which we need in the squared risk evaluation. Unfortunately, the second term $\sigma^2 (n - m)$ behaves differently than the similar variance term $\sigma^2 m$. Another good news is that both variance terms are known to us. Therefore, one can easily make a correction of the empirical risk which delivers an unbiased risk estimate: just add $\sigma^2 (2m - n)$. Define

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

Then it holds

$$I\!\!E\widetilde{\mathcal{R}}_m = \mathcal{R}(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*) + \sigma^2 n.$$

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

$$\widehat{m} \stackrel{\text{def}}{=} \operatorname*{argmin}_m \widetilde{\mathcal{R}}_m = \operatorname*{argmin}_m \bigl(\|\boldsymbol{Y} - \Pi_m \boldsymbol{Y}\|^2 + 2\sigma^2 m \bigr) . \tag{2.6}$$

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

Exercise 2.2.1. Consider the projection estimator $\widetilde{\boldsymbol{f}}_m = \Pi_m \boldsymbol{Y}$ for the model (2.1) with $\Pi_m = \Psi_m^\top \bigl(\Psi_m \Psi_m^\top\bigr)^{-1} \Psi_m$. For two different values $m' > m$:

- Check that $\Pi_{m',m} \stackrel{\text{def}}{=} \Pi_{m'} - \Pi_m$ is a projector in $I\!R^n$. Describe its image in the orthogonal case when $\Psi\Psi^\top$ is a diagonal matrix.
- Check the identities

$$\|\boldsymbol{Y} - \Pi_m \boldsymbol{Y}\|^2 - \|\boldsymbol{Y} - \Pi_{m'} \boldsymbol{Y}\|^2 = \|\widetilde{\boldsymbol{f}}_{m'} - \widetilde{\boldsymbol{f}}_m\|^2 = \|\Pi_{m',m} \boldsymbol{f} + \Pi_{m',m} \boldsymbol{\varepsilon}\|^2,$$
$$\|\widetilde{\boldsymbol{f}}_{m'} - \boldsymbol{f}\|^2 - \|\widetilde{\boldsymbol{f}}_m - \boldsymbol{f}\|^2 = -\|\Pi_{m',m} \boldsymbol{f}\|^2 + \|\Pi_{m',m} \boldsymbol{\varepsilon}\|^2.$$

- compute $I\!\!E\|\widetilde{\boldsymbol{f}}_{m'} - \widetilde{\boldsymbol{f}}_m\|^2$ and $I\!\!E\bigl[\|\widetilde{\boldsymbol{f}}_{m'} - \boldsymbol{f}\|^2 - \|\widetilde{\boldsymbol{f}}_m - \boldsymbol{f}\|^2\bigr]$.

2.2.1 AIC and pairwise comparison

Here we try to understand whether the AIC rule does a good job in model selection. In particular, whether it mimics the oracle. Our study will be based on pairwise comparison. More precisely, we check two situations: when the data-driven choice $\widehat{m}$ is larger than the oracle and the inverse case. The most important problem is to bound the probability and the risk associated with the event $\bigl\{\widehat{m} > m^*\bigr\}$.

The definition of m^* (2.4) implies for $m > m^*$ with $\mathcal{R}_m \stackrel{\text{def}}{=} \mathcal{R}(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*)$

$$\begin{aligned} \mathcal{R}_m - \mathcal{R}_{m^*} &= \|\boldsymbol{f}^* - \Pi_m \boldsymbol{f}^*\|^2 - \|\boldsymbol{f}^* - \Pi_{m^*} \boldsymbol{f}^*\|^2 + \sigma^2 (m - m^*) \\ &= -\|\boldsymbol{b}_{m,m^*}\|^2 + \sigma^2 (m - m^*) \geq 0, \end{aligned} \tag{2.7}$$

where $\boldsymbol{b}_{m,m^*} \stackrel{\text{def}}{=} \Pi_{m,m^*} \boldsymbol{f}^*$.

Exercise 2.2.2. Consider the model $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$ with homogeneous errors $\sigma_i \equiv \sigma$. Let m^* be the oracle choice from (2.4).

- check (2.7);
- check that for $m < m^*$, it holds

$$\|\boldsymbol{b}_{m^*,m}\|^2 = \|\Pi_{m^*,m}\boldsymbol{f}\|^2 \geq \sigma^2(m^* - m). \tag{2.8}$$

- check that for $m > m^*$, it holds

$$\|\boldsymbol{b}_{m,m^*}\|^2 = \|\Pi_{m,m^*}\boldsymbol{f}\|^2 \leq \sigma^2(m - m^*) \tag{2.9}$$

In words, due to (2.8), it is reasonable to increase the model complexity towards m^*, the gain in the quality of approximation is larger than the additional complexity. However, (2.9) shows that if we increase the complexity of the model over the oracle m^*, then our additional loss due to increased complexity exceeds the gain due to bias reduction.

The next question is whether the data-driven choice $\widehat{m}$ reproduces this situation. The selected model $\widehat{m}$ is a winner in a pairwise competition with all other models, in particular, in competition with the "oracle" choice m^*. This means that $\widetilde{\mathcal{R}}_{\widehat{m}} \leq \widetilde{\mathcal{R}}_{m^*}$. If the value $\widetilde{\mathcal{R}}_m$ is close to its expectation $\mathcal{R}_m$ and if $\mathcal{R}_m$ is significantly larger than the oracle risk $\mathcal{R}_{m^*}$ then the probability of the event $\widetilde{\mathcal{R}}_m \leq \widetilde{\mathcal{R}}_{m^*}$ is very small. So, one can expect that the selected model $\widehat{m}$ is mainly located on the set where the risk $\mathcal{R}_m$ does not deviate much from $\mathcal{R}_{m^*}$. The next result quantifies this relation. We use the decomposition

$$\begin{aligned}
\widetilde{\mathcal{R}}_m - \widetilde{\mathcal{R}}_{m^*} &= \|\boldsymbol{Y} - \Pi_m\boldsymbol{Y}\|^2 + 2\sigma^2 m - \|\boldsymbol{Y} - \Pi_{m^*}\boldsymbol{Y}\|^2 - 2\sigma^2 m^* \\
&= -\|\Pi_{m,m^*}\boldsymbol{Y}\|^2 + 2\sigma^2(m - m^*) \\
&= -\|\Pi_{m,m^*}\boldsymbol{\varepsilon} + \boldsymbol{b}_{m,m^*}\|^2 + 2\sigma^2(m - m^*) \\
&= \mathcal{R}_m - \mathcal{R}_{m^*} - \bigl\{\|\Pi_{m,m^*}\boldsymbol{\varepsilon}\|^2 - \sigma^2(m - m^*)\bigr\} - 2\boldsymbol{b}_{m,m^*}^\top \Pi_{m,m^*}\boldsymbol{\varepsilon}.
\end{aligned} \tag{2.10}$$

The first stochastic term $\|\Pi_{m,m^*}\boldsymbol{\varepsilon}\|^2 - \sigma^2(m - m^*)$ of this difference is a centered quadratic form of the errors $\boldsymbol{\varepsilon}$ and the unknown regression function f does not show up there. The second one $2\boldsymbol{b}_{m,m^*}^\top \Pi_{m,m^*}\boldsymbol{\varepsilon}$ involves the bias $\boldsymbol{b}_{m,m^*}$ but it is linear in $\boldsymbol{\varepsilon}$. Both terms can be easily bounded for the Gaussian errors $\boldsymbol{\varepsilon}$.

Lemma 2.2.3. *Let $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Then it holds for $\boldsymbol{b}_{m,m^*} = \Pi_{m,m^*}\boldsymbol{f}^*$*

$$\begin{aligned}
I\!P\Bigl(\sigma^{-2}\|\Pi_{m,m^*}\boldsymbol{\varepsilon}\|^2 > \mathfrak{z}^+(m - m^*, \mathtt{x})\Bigr) &\leq \frac{1}{2}\mathrm{e}^{-\mathtt{x}}, \\
I\!P\Bigl(\sigma^{-2}\|\Pi_{m,m^*}\boldsymbol{\varepsilon}\|^2 < \mathfrak{z}^-(m - m^*, \mathtt{x})\Bigr) &\leq \frac{1}{2}\mathrm{e}^{-\mathtt{x}}, \\
I\!P\Bigl(\sigma^{-2}\bigl|\boldsymbol{b}_{m,m^*}^\top \Pi_{m,m^*}\boldsymbol{\varepsilon}\bigr| > \sigma^{-1}\|\boldsymbol{b}_{m,m^*}\| z_1(\mathtt{x})\Bigr) &\leq \mathrm{e}^{-\mathtt{x}}.
\end{aligned}$$

where $\mathfrak{z}^+(k,\mathtt{x})$ *is the upper* $1-0.5\mathrm{e}^{-\mathtt{x}}$ *quantile of* χ^2_k, $\mathfrak{z}^-(k,\mathtt{x})$ *is its lower* $0.5\mathrm{e}^{-\mathtt{x}}$ *quantile,* $z_1(\mathtt{x})$ *is the quantile of* ξ *for a standard normal r.v.* $\xi \sim \mathcal{N}(0,1)$*:* $I\!P\bigl(|\xi| > z_1(\mathtt{x})\bigr) \le \mathrm{e}^{-\mathtt{x}}$.

It holds for any $k \ge 1$ *and* $\mathtt{x} > 0$ *with* $\mathtt{x}_1 = \mathtt{x} + \log(2)$

$$\begin{aligned} \mathfrak{z}^+(k,\mathtt{x}) &\le k + \sqrt{6.6k\,\mathtt{x}_1} \vee (6.6\mathtt{x}_1), \\ \mathfrak{z}^-(k,\mathtt{x}) &\ge k - \sqrt{2k\,\mathtt{x}_1}. \end{aligned} \tag{2.11}$$

The proof only uses that $\sigma^{-1}\boldsymbol{\varepsilon}$ is a standard Gaussian vector in $I\!R^n$ and thus, $\sigma^{-1}\Pi_{m,m^*}\boldsymbol{\varepsilon}$ is standard normal in $I\!R^{m-m^*}$, while $\bigl(\sigma\|\boldsymbol{b}_{m,m^*}\|\bigr)^{-1}\boldsymbol{b}_{m,m^*}^\top\Pi_{m,m^*}\boldsymbol{\varepsilon}$ is a standard normal r.v. if the bias $\boldsymbol{b}_{m,m^*}$ does not vanish.

The presented bounds show that for moderate values of $\mathtt{x}$

$$z^{\pm}(k,\mathtt{x}) \stackrel{\mathrm{def}}{=} \mathfrak{z}^+(k,\mathtt{x}) - \mathfrak{z}^-(k,\mathtt{x}) \le \mathtt{C}\sqrt{k\,\mathtt{x}}.$$

for a fixed constant $\mathtt{C}$. Therefore, for large k, the interquartile range $z^{\pm}(k,\mathtt{x}) = \mathfrak{z}^+(k,\mathtt{x}) - \mathfrak{z}^-(k,\mathtt{x})$ is small relative to k which is the expectation of $\sigma^{-2}I\!E\|\Pi_k\boldsymbol{\varepsilon}\|^2$. This effect is called *concentration* and it explains why the AIC rule works: the difference between empirical risk and its population counterpart is small relatively to the risk itself.

2.2.2 Pairwise analysis

Now we make a more precise analysis of the term $\widetilde{\mathcal{R}}_m - \widetilde{\mathcal{R}}_{m^*}$ in (2.10). It is based on the following general property of the Gaussian distribution.

Lemma 2.2.4. *Let* $\boldsymbol{\xi}$ *be standard Gaussian vector in* $I\!R^k$ *and* $\boldsymbol{\delta}$ *be a deterministic vector in* $I\!R^k$ *with* $\|\boldsymbol{\delta}\|^2 = \Delta$. *Then*

- *the distribution of* $\|\boldsymbol{\xi}+\boldsymbol{\delta}\|^2$ *only depends on* k *and* Δ*;*
- *let, for a given* $\mathtt{x}$*, the quantiles* $\mathfrak{z}^+(k,\Delta;\mathtt{x})$ *and* $\mathfrak{z}^-(k,\Delta,\mathtt{x})$ *be defined as*

$$\begin{aligned} I\!P\bigl(\|\boldsymbol{\xi}+\boldsymbol{\delta}\|^2 \ge \mathfrak{z}^+(k,\Delta;\mathtt{x})\bigr) &= \mathrm{e}^{-\mathtt{x}}, \\ I\!P\bigl(\|\boldsymbol{\xi}+\boldsymbol{\delta}\|^2 \le \mathfrak{z}^-(k,\Delta;\mathtt{x})\bigr) &= \mathrm{e}^{-\mathtt{x}}. \end{aligned} \tag{2.12}$$

Then

$$\begin{aligned} \mathfrak{z}^+(k,\Delta;\mathtt{x}) &\le \Delta + \mathfrak{z}^+(k,\mathtt{x}) + 2\Delta^{1/2}z_1(\mathtt{x}), \\ \mathfrak{z}^-(k,\Delta;\mathtt{x}) &\ge \Delta + \mathfrak{z}^-(k,\mathtt{x}) - 2\Delta^{1/2}z_1(\mathtt{x}), \end{aligned} \tag{2.13}$$

Proof. Use the decomposition

$$\|\boldsymbol{\xi}+\boldsymbol{\delta}\|^2 = \Delta + \|\boldsymbol{\xi}\|^2 + 2\boldsymbol{\xi}^\top\boldsymbol{\delta}$$

and Lemma 2.2.3.

We apply this result to $\pm\sigma^{-2}\|\Pi_{m,m^*}\boldsymbol{Y}\|^2$ entering in the difference $\widetilde{\mathcal{R}}_m - \widetilde{\mathcal{R}}_{m^*}$. The bound $\widetilde{\mathcal{R}}_m - \widetilde{\mathcal{R}}_{m^*}$ can be rewritten for $m > m^*$ as $\sigma^{-2}\|\Pi_{m,m^*}\boldsymbol{Y}\|^2 > 2(m - m^*)$ which is directly related to the upper quantile of non-central chi-squared. Define the value of non-centrality parameter Δ to have $\mathfrak{z}^{\pm}(k, \Delta^{\pm}; \mathtt{x})$ exactly equal to $2k$:

$$\mathfrak{z}^{+}(k, \Delta^{+}(k, \mathtt{x}); \mathtt{x}) = 2k, \qquad \mathfrak{z}^{-}(k, \Delta^{-}(k, \mathtt{x}); \mathtt{x}) = 2k. \tag{2.14}$$

This definition can be rewritten as follows:

$$I\!P\bigl(\|\boldsymbol{\xi} + \boldsymbol{\delta}^{+}\|^2 > 2k\bigr) \le \mathrm{e}^{-\mathtt{x}}, \qquad \text{if } \|\boldsymbol{\delta}^{+}\|^2 \le \Delta^{+}(k, \mathtt{x}), \tag{2.15}$$

$$I\!P\bigl(\|\boldsymbol{\xi} + \boldsymbol{\delta}^{-}\|^2 < 2k\bigr) \le \mathrm{e}^{-\mathtt{x}}, \qquad \text{if } \|\boldsymbol{\delta}^{-}\|^2 \ge \Delta^{-}(k, \mathtt{x}). \tag{2.16}$$

Exercise 2.2.3. For the quantities $\Delta^{+}(k, \mathtt{x}), \Delta^{-}(k, \mathtt{x})$ from (2.14) and $\mathtt{x} \ge 1$

- show that $\Delta^{+}(k, \mathtt{x}) < k$, $\Delta^{-}(k, \mathtt{x}) > k$;
- check that Lemma 2.2.4 implies

$$\Delta^{+}(k, \mathtt{x}) \ge \mathfrak{z}^{-}(k, \mathtt{x}) - 2k^{1/2} z_1(\mathtt{x}), \tag{2.17}$$

$$\Delta^{-}(k, \mathtt{x}) \le \mathfrak{z}^{+}(k, \mathtt{x}) + 2k^{1/2} z_1(\mathtt{x}), \tag{2.18}$$

We conclude with the following statement.

Proposition 2.2.1. *Let the errors $\boldsymbol{\varepsilon}$ be normal and homogeneous: $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Then the inequalities*

$$\begin{aligned} \sigma^{-2}\|\boldsymbol{b}_{m,m^*}\|^2 &\le \Delta^{+}(m - m^*, \mathtt{x}), \qquad m > m^* \\ \sigma^{-2}\|\boldsymbol{b}_{m^*,m}\|^2 &\ge \Delta^{-}(m^* - m, \mathtt{x}), \qquad m < m^* \end{aligned} \tag{2.19}$$

ensures

$$I\!P\bigl(\widetilde{\mathcal{R}}_m \le \widetilde{\mathcal{R}}_{m^*}\bigr) \le \mathrm{e}^{-\mathtt{x}}.$$

In particular, if the bias term $\|\boldsymbol{b}_m\|$ is uniformly bounded for all $m \ge m^*$ by a fixed constant $\mathtt{C}(\mathcal{F})$, then $\|\boldsymbol{b}_{m,m^*}\| \le \mathtt{C}(\mathcal{F})$ and the inequality (2.19) is fulfilled if

$$m - m^* > \bigl(2\sigma^{-1}\mathtt{C}(\mathcal{F}) + \mathtt{Cx}\bigr)^2 \tag{2.20}$$

for $\mathtt{C} \ge 3$.

Proof. Consider the case $m > m^*$. We apply the decomposition

$$\sigma^{-2}\bigl(\widetilde{\mathcal{R}}_m - \widetilde{\mathcal{R}}_{m^*}\bigr) = -\|\sigma^{-1}\Pi_{m,m^*}\boldsymbol{\varepsilon} + \sigma^{-1}\boldsymbol{b}_{m,m^*}\|^2 + 2(m - m^*).$$

Further, $\boldsymbol{\xi} \stackrel{\text{def}}{=} \sigma^{-1}\Pi_{m,m^*}\boldsymbol{\varepsilon}$ is standard normal in $\mathbb{R}^k$ for $k = m - m^*$. The condition (2.15) with $\boldsymbol{\delta}^+ = \sigma^{-1}\boldsymbol{b}_{m,m^*}$ implies

$$\mathbb{P}\bigl(\widetilde{\mathcal{R}}_m \le \widetilde{\mathcal{R}}_{m^*}\bigr) = \mathbb{P}\bigl(\|\boldsymbol{\xi} + \boldsymbol{\delta}^+\|^2 > 2k\bigr) \le \mathrm{e}^{-\mathtt{x}}.$$

The case $m < m^*$ can be done in a similar way using (2.16) in place of (2.15).

The inequality $\|\boldsymbol{b}_m\| \le \mathtt{C}(\mathcal{F})$ implies $\|\boldsymbol{b}_{m,m^*}\| \le \mathtt{C}(\mathcal{F})$ for all $m > m^*$; see (2.22). Now it remains to check by (2.17) and (2.11) that (2.20) implies (2.19).

To be done: complete the proof

Exercise 2.2.4. Show that the value $\|\boldsymbol{b}_m\| = \|\boldsymbol{f}^* - \Pi_m\boldsymbol{f}^*\|$ monotonously decreases with m. Moreover, for any $m' > m$, the relative bias $\boldsymbol{b}_{m',m} = \Pi_{m',m}\boldsymbol{f}^*$ satisfies

$$\|\boldsymbol{b}_{m',m}\| \le \|\boldsymbol{b}_m\|, \quad m' > m. \tag{2.21}$$

Check whether also holds

$$\|\boldsymbol{b}_{m',m}\| \le \|\boldsymbol{b}_{m'}\|, \quad m' > m. \tag{2.22}$$

2.2.3 Uniform bounds and the zone of insensitivity

This section introduces the *set of insensitivity* $\mathcal{M}^\circ(\mathtt{x})$ which describes the quality of model selection. Namely, we aim at describing the set $\mathcal{M}^\circ(\mathtt{x})$ which contains all possible values of $\widehat{m}$ with a high probability. The ideal situation would be $\mathcal{M}^\circ(\mathtt{x}) = \{m^*\}$, but it is rare the case. Usually $\mathcal{M}^\circ(\mathtt{x})$ is a larger set containing m^*. Below we specify this set in terms of the difference $\mathcal{R}_m - \mathcal{R}_{m^*}$ and its decomposition (2.7).

The study of the previous section quantifies the pairwise relation $\widetilde{\mathcal{R}}_m - \widetilde{\mathcal{R}}_{m^*} \le 0$: under $\sigma^{-2}\|\boldsymbol{b}_{m,m^*}\|^2 \le \Delta^+(m - m^*, \mathtt{x})$; see (2.19), it holds

$$\mathbb{P}\bigl(\widetilde{\mathcal{R}}_m \le \widetilde{\mathcal{R}}_{m^*}\bigr) \le \mathrm{e}^{-\mathtt{x}}. \tag{2.23}$$

Now we need its uniform version over the complement of $\mathcal{M}^\circ(\mathtt{x})$:

$$\mathbb{P}\Bigl(\max_{m \notin \mathcal{M}^\circ(\mathtt{x})}\{\widetilde{\mathcal{R}}_m - \widetilde{\mathcal{R}}_{m^*}\} \ge 0\Bigr) \le \mathrm{e}^{-\mathtt{x}}.$$

One can use a uniform adjustment in each bound (2.23) by increasing the value $\mathtt{x}$ to another slightly larger level $\mathtt{x}_\mathtt{s}$. A simple way is based on the so called Bonferroni correction: $\mathtt{x}_\mathtt{s} \equiv \mathtt{x} + \log(|\mathcal{M}|)$.

Proposition 2.2.2. *Let* $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. *If* $\mathcal{M}^\circ(\mathtt{x})$ *is the set of indices* m *such that*

$$\mathcal{M}^\circ(\mathtt{x}) \stackrel{\mathrm{def}}{=} \begin{cases} \sigma^{-2}\|\boldsymbol{b}_{m,m^*}\|^2 \geq \Delta^+(m - m^*, \mathtt{x}_\mathtt{s}), & m > m^*, \\ \sigma^{-2}\|\boldsymbol{b}_{m^*,m}\|^2 \leq \Delta^-(m^* - m, \mathtt{x}_\mathtt{s}), & m < m^*, \end{cases} \tag{2.24}$$

for $\mathtt{x}_\mathtt{s} = \mathtt{x} + \log(|\mathcal{M}|)$, *then*

$$I\!P\big(\widehat{m} \notin \mathcal{M}^\circ(\mathtt{x})\big) \leq \mathrm{e}^{-\mathtt{x}}. \tag{2.25}$$

Proof. By definition of $\mathcal{M}^\circ(\mathtt{x})$ and Proposition 2.2.1

$$I\!P\big(\widehat{m} \notin \mathcal{M}^\circ(\mathtt{x})\big) \leq \sum_{m \notin \mathcal{M}^\circ(\mathtt{x})} I\!P\big(\widehat{m} = m\big) \leq \sum_{m \notin \mathcal{M}^\circ(\mathtt{x})} I\!P\big(\widetilde{\mathcal{R}}_m \leq \widetilde{\mathcal{R}}_{m^*}\big) \leq \sum_{m \in \mathcal{M}} \mathrm{e}^{-\mathtt{x}_\mathtt{s}} \leq \mathrm{e}^{-\mathtt{x}}.$$

One can conclude that if the m lies beyond the *insensitivity zone* $\mathcal{M}^\circ(\mathtt{x})$ around m^*, on which the difference $\mathcal{R}_m - \mathcal{R}_{m^*}$ is not sufficiently large, then the event $\widehat{m} = m$ is very unlikely. The result (2.25) can be stated in the form that there exists a random set $\Omega(\mathtt{x})$ such that $I\!P\big(\Omega(\mathtt{x})\big) \geq 1 - \mathrm{e}^{-\mathtt{x}}$, and on this set, it holds

$$\widetilde{\mathcal{R}}_m > \widetilde{\mathcal{R}}_{m^*}, \qquad m \notin \mathcal{M}^\circ(\mathtt{x})$$

and hence $\widehat{m} \in \mathcal{M}^\circ(\mathtt{x})$ on $\Omega(\mathtt{x})$.

2.2.4 A bound on the excess

Introduce another random set $\Omega_0(\mathtt{x})$ such that

$$\begin{aligned} \|\sigma^{-1}\Pi_{m,m^*}\varepsilon\|^2 &\leq \mathfrak{z}^+(m - m^*, \mathtt{x}_\mathtt{s}), \qquad m > m^*, \\ \|\sigma^{-1}\Pi_{m^*,m}\varepsilon\|^2 &\geq \mathfrak{z}^-(m^* - m, \mathtt{x}_\mathtt{s}), \qquad m < m^*. \end{aligned} \tag{2.26}$$

Lemma 2.2.3 implies that $I\!P(\Omega_0(\mathtt{x})) \geq 1 - \mathrm{e}^{-\mathtt{x}}$.

The next question is what happens if $\widehat{m} \in \mathcal{M}^\circ(\mathtt{x})$ and how big this set is. We will try to bound the loss difference $\varrho(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*) - \varrho(\widetilde{\boldsymbol{f}}_{m^*}, \boldsymbol{f}^*)$. It follows from (2.2) that for $m > m^*$

$$\sigma^{-2}\big\{\varrho(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*) - \varrho(\widetilde{\boldsymbol{f}}_{m^*}, \boldsymbol{f}^*)\big\} = -\|\sigma^{-1}\boldsymbol{b}_{m,m^*}\|^2 + \|\sigma^{-1}\Pi_{m,m^*}\varepsilon\|^2.$$

This implies for $m \in \mathcal{M}_+(\mathtt{x})$ by (2.24) and (2.17) on $\Omega_0(\mathtt{x})$

$$\begin{aligned} &\sigma^{-2}\big\{\varrho(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*) - \varrho(\widetilde{\boldsymbol{f}}_{m^*}, \boldsymbol{f}^*)\big\} \\ &\quad \leq -\Delta^+(m - m^*, \mathtt{x}_\mathtt{s}) + \mathfrak{z}^+(m - m^*, \mathtt{x}_\mathtt{s}) \\ &\quad \leq \mathfrak{z}^+(m - m^*, \mathtt{x}_\mathtt{s}) - \mathfrak{z}^-(m - m^*, \mathtt{x}_\mathtt{s}) + 2z_1(\mathtt{x}_\mathtt{s})\sqrt{m - m^*} \\ &\quad = z^\pm(m - m^*, \mathtt{x}_\mathtt{s}) + 2z_1(\mathtt{x}_\mathtt{s})\sqrt{m - m^*}; \end{aligned} \tag{2.27}$$

here $z^{\pm}(k, \mathtt{x}) = \mathfrak{z}^{+}(k, \mathtt{x}) - \mathfrak{z}^{-}(k, \mathtt{x})$.

Now we consider the similar difference for the parameter m from the insensitivity zone $\mathcal{M}_{-}(\mathtt{x})$ with $m < m^*$. It holds

$$\sigma^{-2}\bigl\{\varrho(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*) - \varrho(\widetilde{\boldsymbol{f}}_{m^*}, \boldsymbol{f}^*)\bigr\} = \|\sigma^{-1}\boldsymbol{b}_{m^*,m}\|^2 - \|\sigma^{-1}\Pi_{m^*,m}\boldsymbol{\varepsilon}\|^2.$$

This implies for $m \in \mathcal{M}^{\circ}(\mathtt{x})$ by (2.24) and (2.18) on $\Omega_0(\mathtt{x})$

$$\begin{aligned}
&\sigma^{-2}\bigl\{\varrho(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*) - \varrho(\widetilde{\boldsymbol{f}}_{m^*}, \boldsymbol{f}^*)\bigr\} \\
&\quad \le \|\sigma^{-1}\boldsymbol{b}_{m^*,m}\|^2 - \|\sigma^{-1}\Pi_{m^*,m}\boldsymbol{\varepsilon}\|^2 \\
&\quad \le \Delta^{-}(m^* - m, \mathtt{x}_{\mathtt{s}}) - \mathfrak{z}^{-}(m^* - m, \mathtt{x}_{\mathtt{s}}) \\
&\quad \le \mathfrak{z}^{+}(m^* - m, \mathtt{x}_{\mathtt{s}}) - \mathfrak{z}^{-}(m^* - m, \mathtt{x}_{\mathtt{s}}) + 2z_1(\mathtt{x}_{\mathtt{s}})\sqrt{m^* - m} \\
&\quad = z^{\pm}(m^* - m, \mathtt{x}_{\mathtt{s}}) + 2z_1(\mathtt{x}_{\mathtt{s}})\sqrt{m^* - m}. \qquad (2.28)
\end{aligned}$$

Now we can summarize. Define the radius $R = R(\mathcal{M}^{\circ}(\mathtt{x}))$ of the set $\mathcal{M}^{\circ}(\mathtt{x})$:

$$R \stackrel{\text{def}}{=} \max_{m \in \mathcal{M}^{\circ}(\mathtt{x})} |m - m^*|.$$

Theorem 2.2.1. *Let $\widehat{m}$ be defined by (2.6) and m^* be the oracle choice from (2.4). Suppose that $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Let $\mathtt{x}_{\mathtt{s}} = \mathtt{x} + \log(|\mathcal{M}|)$. For the insensitivity zone $\mathcal{M}^{\circ}(\mathtt{x})$ from (2.24), it holds $\widehat{m} \in \mathcal{M}^{\circ}(\mathtt{x})$ on a random set $\Omega(\mathtt{x})$ with $I\!P\bigl(\Omega(\mathtt{x})\bigr) \ge 1 - \mathrm{e}^{-\mathtt{x}}$. Moreover, on a random set $\Omega_0(\mathtt{x})$ with $I\!P\bigl(\Omega_0(\mathtt{x})\bigr) \ge 1 - 2\mathrm{e}^{-\mathtt{x}}$*

$$\sigma^{-2}\bigl\{\varrho(\widetilde{\boldsymbol{f}}_{\widehat{m}}, \boldsymbol{f}^*) - \varrho(\widetilde{\boldsymbol{f}}_{m^*}, \boldsymbol{f}^*)\bigr\} \le z^{\pm}(R, \mathtt{x}_{\mathtt{s}}) + 2z_1(\mathtt{x}_{\mathtt{s}})\sqrt{R}.$$

Proof. The result follows from (2.27) and (2.28) by monotonicity of the function $z^{\pm}(k, \mathtt{x})$ in k and $\mathtt{x}$.

In view of $z^{\pm}(R, \mathtt{x}_{\mathtt{s}}) \asymp \sqrt{R\,\mathtt{x}_{\mathtt{s}}}$, we conclude that the data-driven choice of the parameter m leads to additional loss of order $\sigma^2\sqrt{R\,\mathtt{x}_{\mathtt{s}}}$. One can say that the model selection based on unbiased risk estimation works well if the size of the zone of insensitivity $R = R(\mathcal{M}^{\circ}(\mathtt{x}))$ is not too large compared with the loss and risk of the oracle $\widetilde{\boldsymbol{f}}_{m^*}$.

Note that the loss $\varrho(\widetilde{\boldsymbol{f}}_{m^*}, \boldsymbol{f}^*)$ fulfills

$$\varrho(\widetilde{\boldsymbol{f}}_{m^*}, \boldsymbol{f}^*) = \|\boldsymbol{b}_{m^*}\|^2 + \|\Pi_{m^*}\boldsymbol{\varepsilon}\|^2 \ge \|\Pi_{m^*}\boldsymbol{\varepsilon}\|^2.$$

By Lemma 2.2.3, it is of order m^*.

Exercise 2.2.5. Let $\Omega_0(\mathtt{x})$ be a random set on which (2.26) holds. Show that $I\!P\bigl(\Omega_0(\mathtt{x})\bigr) \geq 1 - \mathrm{e}^{-\mathtt{x}}$ and for every m, the loss $\varrho(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*)$ satisfies on $\Omega_0(\mathtt{x})$

$$\sigma^{-2}\varrho(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*) \geq \mathfrak{z}^-(m, \mathtt{x}).$$

For the case when the value $R = \max_{m \in \mathcal{M}^\circ(\mathtt{x})} |m - m^*|$ is small relative to m^*, the loss of the estimate $\widehat{\boldsymbol{f}} = \widetilde{\boldsymbol{f}}_{\widehat{m}}$ corresponding to the data-driven selector $\widehat{m}$ is not significantly larger than the loss of the oracle estimate $\widetilde{\boldsymbol{f}}_{m^*}$. Unfortunately, the set $\mathcal{M}^\circ(\mathtt{x})$ can be very large if the risk $\mathcal{R}_m$ is a flat function of m. The extreme case is given by the so called "noise reproducing" model. This model is described by the equations $|\theta_j^*|^2 \equiv \sigma^2$; see (2.3). One can easily check that

$$\begin{aligned} \|\boldsymbol{b}_{m,m^*}\|^2 &\equiv \sigma^2(m - m^*), \qquad m > m^*, \\ \|\boldsymbol{b}_{m^*,m}\|^2 &\equiv \sigma^2(m^* - m), \qquad m < m^*. \end{aligned}$$

This implies that the risk function $\mathcal{R}_m$ is constant in m, and therefore, the set $\mathcal{M}^\circ(\mathtt{x})$ coincides with the whole set $\mathcal{M}$.

Exercise 2.2.6. Build an example in which the radius R is twice as large as the oracle risk $\mathcal{R}_{m^*}$.

2.3 The approach based on multiple testing. "Smallest accepted" rule

This section discusses the alternative approach to model selection based on the idea of multiple testing. Let m^* be a good choice in the sense of "bias-variance trade-off". Now we aim to develop a procedure that would treat m^* as a good choice with a high probability. We interpret the model selection procedure as pairwise comparison: the "oracle" model wins in terms of the risk against all other models:

$$\mathcal{R}_{m^*} \leq \mathcal{R}_m, \qquad m \neq m^*.$$

The selector $\widehat{m}$ suggests to apply the model which wins in term of the unbiased risk estimate $\widetilde{\mathcal{R}}_m$:

$$\widetilde{\mathcal{R}}_{\widehat{m}} \leq \widetilde{\mathcal{R}}_m, \qquad m \neq \widehat{m}.$$

Now we reconsider this approach in terms of hypothesis testing.

2.3.1 A LR test

Our null hypothesis will be that a model-candidate m° is "good" in the sense that it delivers a kind of bias-variance trade-off. This means that there is no reason for considering a larger model: the bias improvement will not be compensated by increase of model complexity. Now we interpret this check as a test of the model-candidate m° against any larger model $m > m^\circ$. The null hypothesis H_{m° means that $\theta_j^* \equiv 0$ for all $j > m^\circ$. The alternative is that there are significant coefficients θ_j^* for $m^\circ < j \le m$.

Define

$$\Theta_m \stackrel{\text{def}}{=} \bigl\{\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m, 0, \ldots, 0)^\top \in I\!R^p\bigr\}.$$

For the log-likelihood function $L(\boldsymbol{\theta}) = -(2\sigma^2)^{-1}\|\boldsymbol{Y} - \Psi^\top\boldsymbol{\theta}\|^2$, the likelihood ratio test statistic reads as

$$\begin{aligned} T_{m,m^\circ} &\stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta}\in\Theta_m} L(\boldsymbol{\theta}) - \sup_{\boldsymbol{\theta}\in\Theta_{m^\circ}} L(\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\bigl(\|\boldsymbol{Y} - \Psi^\top\widetilde{\boldsymbol{\theta}}_{m^\circ}\|^2 - \|\boldsymbol{Y} - \Psi^\top\widetilde{\boldsymbol{\theta}}_m\|^2\bigr) \\ &= \frac{1}{2\sigma^2}\bigl(\|\boldsymbol{Y} - \Pi_{m^\circ}\boldsymbol{Y}\|^2 - \|\boldsymbol{Y} - \Pi_m\boldsymbol{Y}\|^2\bigr) = \frac{1}{2\sigma^2}\|\Pi_{m,m^\circ}\boldsymbol{Y}\|^2. \end{aligned}$$

We aim to calibrate this test in a way that the null hypothesis of no significant bias $\boldsymbol{b}_{m,m^\circ}$ between m° and m is not rejected if the bias is indeed insignificant. Significance can be measured by the energy $\sigma^{-2}\|\boldsymbol{b}_{m,m^\circ}\|^2$ of this bias. Namely, we say that the bias $\boldsymbol{b}_{m,m^\circ}$ is not significant if

$$\sigma^{-2}\|\boldsymbol{b}_{m,m^\circ}\|^2 \le \beta(m - m^\circ) \tag{2.29}$$

for a fixed value β. Remind that the definition of the oracle m^* yields the inequality $\|\boldsymbol{b}_{m,m^*}\|^2 \le \sigma^2(m - m^*)$ corresponding to $\beta = 1$. We apply Lemma 2.2.4 to choose a critical value for the test statistic T_{m,m°. Define

$$\mathtt{z}_\beta(k, \mathtt{x}) \stackrel{\text{def}}{=} \mathfrak{z}^+(k, \beta k; \mathtt{x}),$$

where $\mathfrak{z}^+(k, \Delta; \mathtt{x})$ is the quantile of a non-central chi-squared from (2.12). Lemma 2.2.4 also implies

$$\mathtt{z}_\beta(k, \mathtt{x}) \le \beta k + \mathfrak{z}^+(k, \mathtt{x}) + 2\sqrt{\beta k}\, z_1(\mathtt{x}). \tag{2.30}$$

These definitions imply the following statement.

Proposition 2.3.1. *Let* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. *Then* $\|\boldsymbol{b}_{m,m^\circ}\|^2 \le \beta(m - m^\circ)$ *implies*

$$I\!P\bigl(2T_{m,m^\circ} > \mathtt{z}_\beta(m - m^\circ, \mathtt{x})\bigr) \le \mathrm{e}^{-\mathtt{x}}.$$

Exercise 2.3.1. Prove Proposition 2.3.1.

2.3.2 Multiplicity correction

The hypothesis H_{m° is not rejected if each test T_{m,m° for $m > m^\circ$ does not reject the null. This requires a multiple testing procedure and a correction for multiplicity. A simple way is the Bonferroni correction: one increases each $\mathtt{x}$ by the same value $q_{m^\circ} = \log(|\mathcal{M}(m^\circ)|) = \log(p - m^\circ)$. Here $\mathcal{M}(m^\circ) = \{m \in \mathcal{M}\colon m > m^\circ\}$. Then

$$\begin{aligned} &I\!P\Big(\bigcup_{m \in \mathcal{M}(m^\circ)} \{2T_{m,m^\circ} > \mathtt{z}_\beta(m - m^\circ, \mathtt{x} + q_{m^\circ})\} \Big) \\ &\quad \le \sum_{m \in \mathcal{M}(m^\circ)} \mathrm{e}^{-\mathtt{x} - q_{m^\circ}} = |\mathcal{M}(m^\circ)| \exp\{-\mathtt{x} - \log(|\mathcal{M}(m^\circ)|)\} = \mathrm{e}^{-\mathtt{x}}. \end{aligned}$$

However, the Bonferroni correction is known to be rather conservative especially if the test statistics T_{m,m° are correlated for different m. This is exactly the case under consideration. Another more careful way to choose the correction q_{m° is based on calibration for one special model.

First we consider the special case $\beta = 0$. Then the null hypothesis means that $\theta^*_j \equiv 0$ for all $j > m^\circ$. In this situation the relative bias $\boldsymbol{b}_{m,m^\circ} = \Pi_{m,m^\circ} \boldsymbol{f}^*$ is equal to zero, and does not depend on the first m° coefficients θ^*_j for $j \le m^\circ$. This allows to define q_{m° by the condition

$$I\!P_0\Big(\bigcup_{m \in \mathcal{M}(m^\circ)} \{2T_{m,m^\circ} > \mathfrak{z}^+(m - m^\circ, \mathtt{x} + q_{m^\circ})\} \Big) = \mathrm{e}^{-\mathtt{x}}, \tag{2.31}$$

where $I\!P_0$ is the measure corresponding to zero signal $\theta^*_j \equiv 0$, and $\mathfrak{z}^+(k, \mathtt{x})$ is the upper quantile of the χ^2_k from Lemma 2.2.3. The meaning of (2.32) is that each particular test based on the test statistic $2T_{m,m^\circ}$ is performed at a higher level $\mathrm{e}^{-\mathtt{x} - q_{m^\circ}} = A^{-1}\mathrm{e}^{-\mathtt{x}}$ with $A = \mathrm{e}^{q_{m^\circ}}$. The value q_{m° is the smallest number providing the familywise error probability $\mathrm{e}^{-\mathtt{x}}$. The Bonferroni correction uses $A = \#\{\text{set of hypotheses}\}$ but this choice is conservative especially if the test statistics are strongly dependent.

In the case of β positive, consider another special "frontier" model $\boldsymbol{f}^* = \Psi^\top \boldsymbol{\theta}^*$ with the vector of parameters $\boldsymbol{\theta}^*$ with the equalities $\sigma^{-2}\|\boldsymbol{b}_{m,m^\circ}\|^2 = \beta(m - m^\circ)$ in place of inequalities in (2.29).

Exercise 2.3.2. For the case of an orthonormal design Ψ, find a vector $\boldsymbol{\theta}^*_\beta$ for which $\sigma^{-2}\|\boldsymbol{b}_{m,m^\circ}\|^2 \equiv \beta(m - m^\circ)$.

Now we calibrate the critical values $\mathtt{z}_\beta$ for this special extreme model. Namely, we fix the smallest value $q_{m^\circ} = q_{m^\circ}(\beta)$ for which holds

$$I\!P_{\boldsymbol{\theta}^*_\beta}\Big(\bigcup_{m \in \mathcal{M}_+(m^\circ)} \{2T_{m,m^\circ} > \mathtt{z}_\beta(m - m^\circ, \mathtt{x} + q_{m^\circ})\} \Big) = \mathrm{e}^{-\mathtt{x}}. \tag{2.32}$$

Suppose that such correction $q_{m^\circ} = q_{m^\circ}(\beta)$ is defined for each $m^\circ \in \mathcal{M}$.

Exercise 2.3.3. Consider the frontier model of Exercise 2.3.2. Let $q_{m^\circ}(\beta)$ be the corresponding value for the condition (2.32). Check that for each m°,

$$q_{m^\circ}(\beta) \geq q_{m^\circ+1}(\beta)$$

yielding

$$\mathtt{z}_\beta(k, \mathtt{x} + q_{m^\circ}) \geq \mathtt{z}_\beta(k, \mathtt{x} + q_{m^\circ-1}).$$

Denote

$$\mathtt{x}_{m^\circ} = \mathtt{x} + q_{m^\circ}.$$

The suggested procedure selects the smallest m° which is accepted by the multiple test H_{m° against H_m for all $m > m^\circ$. This acceptance rule for the null hypothesis H_{m° reads as follows:

$$2T_{m,m^\circ} \leq \mathtt{z}_\beta(m - m^\circ, \mathtt{x}_{m^\circ}), \qquad \forall m > m^\circ.$$

Now the "smallest accepted" procedure can be stated in the following form:

$$\begin{aligned} \widehat{m} &\stackrel{\text{def}}{=} \min\bigl\{m^\circ \in \mathcal{M}\colon m^\circ \text{ is accepted}\bigr\} \\ &= \min\biggl\{m^\circ \in \mathcal{M}\colon \max_{m \in \mathcal{M}(m^\circ)} \bigl\{2T_{m,m^\circ} - \mathtt{z}_\beta(m - m^\circ, \mathtt{x}_{m^\circ})\bigr\} \leq 0\biggr\}. \end{aligned} \tag{2.33}$$

2.3.3 Definition of the oracle and propagation property

Further we aim at establishing the oracle inequality for this method. It is desirable to show that the data driven selector $\widehat{m}$ behaves as good as the oracle one. As the procedure does no rely on the quadratic risk we slightly extend the oracle definition. Let m^* be the oracle value which is now defined as the smallest value m^* satisfying the conditions $\sigma^{-2}\bigl\|\boldsymbol{b}_{m,m^*}\bigr\|^2 \leq \beta(m - m^*)$ for $m > m^*$:

$$m^* \stackrel{\text{def}}{=} \operatorname{argmin}\bigl\{m\colon \sigma^{-2}\bigl\|\boldsymbol{b}_{m,m^*}\bigr\|^2 \leq \beta(m - m^*),\ m > m^*\bigr\}. \tag{2.34}$$

This definition follows the structure of the null hypothesis considered in the procedure. For $\beta = 1$ it is consistent with the oracle definition based on the risk minimization. Now we aim at establishing the oracle property: the data-driven selector $\widehat{m}$ behaves essentially as good as as the oracle choice m^*. The study is done in two steps. The first step is to check that the model m^* will be accepted with a high probability. This would mean that the selected model $\widehat{m}$ satisfies $\widehat{m} \leq m^*$. The second step is in applying the test $T_{m^*,\widehat{m}}$ for this situation.

Theorem 2.3.1. *Let* m^* *be defined by* (2.34). *Let also the selector* $\widehat{m}$ *be calibrated by* (2.32). *Then*

$$\mathbb{P}\big(m^* \text{ is not accepted}\big) \le \mathrm{e}^{-\mathtt{x}}. \tag{2.35}$$

Proof. We use the following fact.

Lemma 2.3.1. *Let* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. *Let* m° *and* $\mathtt{x}$ *be fixed.*

- *The values* $\theta^*_1, \ldots, \theta^*_{m^\circ}$ *do not enter in the distribution of* T_{m,m°.
- *Consider the set* $F_\beta(m^\circ)$ *of all vectors* $\boldsymbol{f}^*$ *with* $\sigma^{-2}\big\|\boldsymbol{b}_{m,m^\circ}\big\|^2 \le \beta(m - m^\circ)$. *Within this set, the probability*

$$\mathbb{P}\big(\max_{m > m^\circ}\big\{2T_{m,m^\circ} - \mathtt{z}_\beta(m - m^\circ, \mathtt{x})\big\} > 0\big)$$

is maximized for the case $\sigma^{-2}\big\|\boldsymbol{b}_{m,m^\circ}\big\|^2 \equiv \beta(m - m^\circ)$.

To be done: Proof of Lemma 2.3.1

Now we check (2.35). The definition of m^* ensures that $\boldsymbol{f}^* \in F_\beta(m^*)$, that is, (2.34) is fulfilled. The last statement of Lemma 2.3.1 ensures that if the value q_{m^*} is fixed for the extreme model with $\sigma^{-2}\big\|\boldsymbol{b}_{m,m^*}\big\|^2 \equiv \beta(m - m^*)$, then

$$\mathbb{P}\big(m^* \text{ is not accepted}\big) = \mathbb{P}\Big(\max_{m > m^*}\big\{2T_{m,m^*} - \mathtt{z}_\beta(m - m^*, \mathtt{x}_{m^*})\big\} > 0\Big) \le \mathrm{e}^{-\mathtt{x}}$$

because the similar inequality (with $m^\circ = m^*$) is fulfilled for the extreme model.

The "propagation" property (2.35) is very important and it is usually not fulfilled for classical procedures like unbiased risk estimation (SURE). The advantage of the proposed approach is that this property is in fact intrinsic for the method and is postulated by the calibration step.

2.3.4 A bound on the loss

It remains to clarify the situation if the selected model is smaller than m^*. This probability is not small but we can control the difference of the losses similarly to the SURE procedure. First we check that the new procedure is also able to identify a significant bias $\boldsymbol{b}_{m^*,m}$ measured by $\sigma^{-2}\|\boldsymbol{b}_{m^*,m}\|^2$.

Proposition 2.3.2. *Define* $\mathtt{x}_\mathtt{s} \stackrel{\text{def}}{=} \mathtt{x} + \log(m^*)$ *and let*

$$\mathcal{M}^\circ(\mathtt{x}) = \Big\{m \le m^* \colon \mathfrak{z}^-(m^* - m, \Delta_m; \mathtt{x}_\mathtt{s}) \le \mathtt{z}_\beta(m^* - m, \mathtt{x}_\mathtt{s})\Big\}$$

with $\Delta_m \stackrel{\text{def}}{=} \sigma^{-2}\|\boldsymbol{b}_{m^*,m}\|^2$. *Then*

$$I\!P\big(\widehat{m} \notin \mathcal{M}^{\circ}(\mathtt{x})\big) \leq 2\mathrm{e}^{-\mathtt{x}}.$$

Moreover, $m \in \mathcal{M}^{\circ}(\mathtt{x})$ *is impossible if the bias* $\Delta_m \stackrel{\mathrm{def}}{=} \sigma^{-2}\|\boldsymbol{b}_{m^*,m}\|^2$ *fulfills with* $k = m^* - m$ *and with* $z^{\pm}(k,\mathtt{x}_{\mathtt{s}}) = \mathfrak{z}^{+}(k,\mathtt{x}_{\mathtt{s}}) - \mathfrak{z}^{-}(k,\mathtt{x}_{\mathtt{s}})$

$$\big\{\Delta_m^{1/2} - z_1(\mathtt{x}_{\mathtt{s}})\big\}^2 \geq z^{\pm}(k,\mathtt{x}_{\mathtt{s}}) + \big\{\sqrt{\beta k} + z_1(\mathtt{x}_{\mathtt{s}})\big\}^2 \tag{2.36}$$

or a stronger condition

$$\Delta_m^{1/2} \geq \sqrt{z^{\pm}(k,\mathtt{x}_{\mathtt{s}})} + \sqrt{\beta k} + 2z_1(\mathtt{x}_{\mathtt{s}}). \tag{2.37}$$

Proof. We already proved that $I\!P\big(\widehat{m} > m^*\big) \leq \mathrm{e}^{-\mathtt{x}}$. It remains to study the case $\widehat{m} < m^*$. Lemma 2.2.4 yields for $m < m^*$

$$2T_{m^*,m} = \sigma^{-2}\|\Pi_{m^*,m}\boldsymbol{Y}\|^2 \geq \mathfrak{z}^{-}(m^* - m, \Delta_m; \mathtt{x})$$

with the probability at least $1 - \mathrm{e}^{-\mathtt{x}}$. This implies a uniform bound

$$2T_{m^*,m} = \sigma^{-2}\|\Pi_{m^*,m}\boldsymbol{Y}\|^2 \geq \mathfrak{z}^{-}(m^* - m, \Delta_m; \mathtt{x}_{\mathtt{s}}), \qquad m < m^*$$

on a random set $\Omega_0(\mathtt{x})$ of probability at least $1 - \mathrm{e}^{-\mathtt{x}}$. For $m \notin \mathcal{M}^{\circ}(\mathtt{x})$, this implies

$$2T_{m^*,m} > \mathtt{z}_{\beta}(m^* - m, \mathtt{x}_m),$$

which makes the event "m is accepted" impossible on $\Omega_0(\mathtt{x})$ for $m \notin \mathcal{M}^{\circ}(\mathtt{x})$.

The bound $\mathfrak{z}^{-}(m^* - m, \Delta_m; \mathtt{x}_{\mathtt{s}}) > \mathtt{z}_{\beta}(m^* - m, \mathtt{x}_m)$ is implicit in Δ_m. To make it explicit, we use the lower bound (2.13) of Lemma 2.2.4 and (2.30). In view of $\mathtt{x}_{\mathtt{s}} \geq \mathtt{x}_m$ for all $m < m^*$, with $k = m^* - m$, the following inequality is sufficient for checking that $m \notin \mathcal{M}^{\circ}(\mathtt{x})$:

$$\Delta_m + \mathfrak{z}^{-}(k,\mathtt{x}_{\mathtt{s}}) - 2\Delta_m^{1/2} z_1(\mathtt{x}_{\mathtt{s}}) \geq \beta k + \mathfrak{z}^{+}(k,\mathtt{x}_{\mathtt{s}}) + 2\sqrt{\beta k}\, z_1(\mathtt{x}_{\mathtt{s}}), \tag{2.38}$$

which yields (2.36). The latter can be slightly simplified by using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$: the bound (2.38) holds if

$$\Delta_m^{1/2} - z_1(\mathtt{x}_{\mathtt{s}}) \geq \sqrt{z^{\pm}(m^* - m, \mathtt{x}_{\mathtt{s}})} + \sqrt{\beta(m^* - m)} + z_1(\mathtt{x}_{\mathtt{s}})$$

which coincides with (2.37).

We conclude that the "smallest accepted" rule leads to the choice of $\widehat{m}$ in the insensitivity zone $\mathcal{M}^{\circ}(\mathtt{x})$ with a high probability. It remains to bound the loss difference $\varrho(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*) - \varrho(\widetilde{\boldsymbol{f}}_{m^*}, \boldsymbol{f}^*)$ for $m \in \mathcal{M}^{\circ}(\mathtt{x})$. We use that for $m < m^*$

$$\sigma^{-2}\{\varrho(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*) - \varrho(\widetilde{\boldsymbol{f}}_{m^*}, \boldsymbol{f}^*)\} = \|\sigma^{-1}\boldsymbol{b}_{m^*,m}\|^2 - \|\sigma^{-1}\Pi_{m^*,m}\boldsymbol{\varepsilon}\|^2.$$

By Lemma 2.2.3, the stochastic form $\|\sigma^{-1}\Pi_{m^*,m}\boldsymbol{\varepsilon}\|^2$ can be bounded from below by $\mathfrak{z}^-(m^* - m, \mathtt{x})$ with probability $1 - \mathrm{e}^{-\mathtt{x}}$. This implies a uniform probability bound $\|\sigma^{-1}\Pi_{m^*,m}\boldsymbol{\varepsilon}\|^2 \geq \mathfrak{z}^-(m^* - m, \mathtt{x}_\mathtt{s})$ with $\mathtt{x}_\mathtt{s} = \mathtt{x} + \log(m^*)$ for $m < m^*$. For $m \in \mathcal{M}^\circ(\mathtt{x})$, this implies

$$\sigma^{-2}\{\varrho(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*) - \varrho(\widetilde{\boldsymbol{f}}_{m^*}, \boldsymbol{f}^*)\} \leq \Delta_m - \mathfrak{z}^-(m^* - m, \mathtt{x}_\mathtt{s})$$

on a random set of probability at least $1 - 3\mathrm{e}^{-\mathtt{x}}$, where Δ_m follows the bound (2.36) or (2.37).

We state the following result:

Theorem 2.3.2. *Let $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. For the selector $\widehat{m}$ from (2.33) and the oracle m^* from (2.34), it holds*

$$I\!P\big(\widehat{m} < m^*\big) \leq \mathrm{e}^{-\mathtt{x}},$$

and for the insensitivity zone $\mathcal{M}^\circ(\mathtt{x})$, it follows on a random set $\Omega(\mathtt{x})$ with $I\!P\big(\Omega(\mathtt{x})\big) \geq 1 - 3\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}\sigma^{-2}\{\varrho(\widetilde{\boldsymbol{f}}_m, \boldsymbol{f}^*) - \varrho(\widetilde{\boldsymbol{f}}_{m^*}, \boldsymbol{f}^*)\} &\leq \max_{m \in \mathcal{M}^\circ(\mathtt{x})} \{\Delta_m - \mathfrak{z}^-(m^* - m, \mathtt{x}_\mathtt{s})\} \\ &\leq \max_{m \in \mathcal{M}^\circ(\mathtt{x})} \Big\{\Big(\sqrt{z^\pm(m^* - m, \mathtt{x}_\mathtt{s})} + \sqrt{\beta(m^* - m)} + 2z_1(\mathtt{x}_\mathtt{s})\Big)^2 - \mathfrak{z}^-(m^* - m, \mathtt{x}_\mathtt{s})\Big\}.\end{aligned}$$

2.3.5 Role of β

The SURE method corresponds to $\beta = 1$. This choice is natural as long the squared risk is concerned. However, the procedure is meaningful for every $\beta > 0$. It even applies for $\beta = 0$. This is a very special case leading to a testing problem instead of estimation.

In general, consider the situation when we apply the SURE procedure with the noise level σ which is misspecified and the true noise level is $\sigma_0^2 \leq \sigma^2$. Then the true risk $\mathcal{R}_m$ is

$$\mathcal{R}_m = \sigma_0^2 m + \|\boldsymbol{b}_m\|^2 = \beta^{-1}\sigma^2 m + \|\boldsymbol{b}_m\|^2$$

for $\beta = \sigma^2/\sigma_0^2$. The use of the wrong noise level is equivalent to applying the procedure with such β.

The choice of a small β results in a strong condition on the bias, so the oracle choice with be shifted towards larger (more complex) models. With $\beta > 1$, one allows for more space for the bias, the procedure becomes more robust and tend to oversmooth the model, that is, to select $\widehat{m}$ which is smaller than the risk minimizer m^*.

3

Linear smoothers

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

$$\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon} \tag{3.1}$$

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

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

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

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

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

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

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

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

3.1 Regularization and ridge regression

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

$$\widetilde{\boldsymbol{\theta}}_R \stackrel{\text{def}}{=} \bigl(\Psi\Psi^\top + R\bigr)^{-1}\Psi\boldsymbol{Y}, \qquad \widetilde{\boldsymbol{f}}_R \stackrel{\text{def}}{=} \Psi^\top\bigl(\Psi\Psi^\top + R\bigr)^{-1}\Psi\boldsymbol{Y}. \tag{3.2}$$

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

$$\widetilde{\boldsymbol{\theta}}_\alpha \stackrel{\text{def}}{=} \bigl(\Psi\Psi^\top + \alpha I_p\bigr)^{-1}\Psi\boldsymbol{Y}, \qquad \widetilde{\boldsymbol{f}}_\alpha \stackrel{\text{def}}{=} \Psi^\top\bigl(\Psi\Psi^\top + \alpha I_p\bigr)^{-1}\Psi\boldsymbol{Y}. \tag{3.3}$$

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

3.2 Penalized likelihood. Bias and variance

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

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

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

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

The penalized MLE reads as

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

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

$$\widetilde{\boldsymbol{\theta}}_G \stackrel{\text{def}}{=} \bigl(\Psi\Psi^\top + \sigma^2 G^2\bigr)^{-1}\Psi\boldsymbol{Y}. \tag{3.5}$$

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

$$\begin{aligned} \boldsymbol{\theta}_G &= \operatorname*{argmax}_{\boldsymbol{\theta}} I\!\!E L_G(\boldsymbol{\theta}) \\ &= \operatorname*{arginf}_{\boldsymbol{\theta}} I\!\!E\bigl\{\|\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}\|^2 + \sigma^2\|G\boldsymbol{\theta}\|^2\bigr\} \\ &= \bigl(\Psi\Psi^\top + \sigma^2 G^2\bigr)^{-1}\Psi\boldsymbol{f}^* = \mathcal{S}_G\boldsymbol{f}^*. \end{aligned} \tag{3.6}$$

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

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

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

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

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

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

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

with $\Pi_G = \Psi^\top \bigl(\Psi\Psi^\top + \sigma^2 G^2\bigr)^{-1}\Psi$. Show that Π_G is a sub-projector in the sense that $\|\Pi_G \boldsymbol{u}\| \le \|\boldsymbol{u}\|$ for any $\boldsymbol{u} \in I\!R^n$.

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

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

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

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

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

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

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

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

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

Exercise 3.2.4. Let Ψ be orthonormal: $\Psi\Psi^\top = I_p$. Describe $\operatorname{Var}(\widetilde{\boldsymbol{\theta}}_G)$. Show that the variance decreases with the penalization G in the sense that $G_1 \ge G$ implies $\operatorname{Var}(\widetilde{\boldsymbol{\theta}}_{G_1}) \le \operatorname{Var}(\widetilde{\boldsymbol{\theta}}_G)$.

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

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

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

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

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

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

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

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

Theorem 3.2.2. *Assume the model* $\boldsymbol{Y} = \Psi^\top\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ *with* $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}_n$. *Then the estimate* $\widetilde{\boldsymbol{\theta}}_G$ *fulfills*

$$I\!E\|\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*\|^2 = \|\boldsymbol{\theta}_G - \boldsymbol{\theta}^*\|^2 + \operatorname{tr}\big(D_G^{-2}D^2D_G^{-2}\big).$$

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

3.3 Inference for the penalized MLE

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

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

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

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

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

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

Exercise 3.3.1. Prove (3.8).
Hint: apply the Taylor expansion to $L_G(\boldsymbol{\theta})$ at $\widetilde{\boldsymbol{\theta}}_G$. Use that $\nabla L_G(\boldsymbol{\theta}_G) = 0$ and $-\nabla^2 L_G(\boldsymbol{\theta}) \equiv \sigma^{-2}\Psi\Psi^\top + G^2$.

Exercise 3.3.2. Prove (3.9).
Hint: show that $\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G = \mathcal{S}_G\boldsymbol{\varepsilon}$ with $\mathcal{S}_G = \bigl(\Psi\Psi^\top + \sigma^2 G^2\bigr)^{-1}\Psi$.

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

$$\mathcal{E}_G(\mathfrak{z}) \stackrel{\text{def}}{=} \bigl\{\boldsymbol{\theta} : L_G(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}) \le \mathfrak{z}\bigr\}.$$

The definition implies the following result for the coverage probability:

$$I\!P\bigl(\mathcal{E}_G(\mathfrak{z}) \not\ni \boldsymbol{\theta}_G\bigr) \le I\!P\bigl(L_G(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G) > \mathfrak{z}\bigr).$$

Now the representation (3.9) for $L_G(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G)$ reduces the problem to a deviation bound for a quadratic form. We apply the general result of Section **??**.

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

$$I\!P\bigl(2L_G(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G) > \mathtt{p}_G + (2v_G\mathtt{x}^{1/2}) \vee (6\mathtt{x})\bigr) \le \exp(-\mathtt{x}).$$

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

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

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

$$I\!P\bigl(\|D_G(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G)\|^2 > \mathtt{p}_G + (2v_G\mathtt{x}^{1/2}) \vee (6\mathtt{x})\bigr) \le \exp(-\mathtt{x}).$$

In other words, $\widetilde{\boldsymbol{\theta}}_G$ concentrates on the set $\mathcal{A}(\mathfrak{z}, \boldsymbol{\theta}_G) = \bigl\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_G\|^2 \le 2\mathfrak{z}\bigr\}$ for $2\mathfrak{z} > \mathtt{p}_G$.

3.4 Projection and shrinkage estimates

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

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

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

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

$$\widetilde{\boldsymbol{f}}_m = z_1\psi_1 + \ldots + z_m\psi_m$$

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

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

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

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

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

Exercise 3.4.1. Let $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_p$. The risk $\mathcal{R}(\widetilde{\boldsymbol{f}}_{\boldsymbol{\alpha}})$ of the shrinkage estimate $\widetilde{\boldsymbol{f}}_{\boldsymbol{\alpha}}$ fulfills

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

Specify the cases of $\boldsymbol{\alpha} = \boldsymbol{\alpha}(m,\beta)$ from (3.10). Evaluate the variance term $\sum_j \alpha_j^2\sigma^2$. Hint: approximate the sum over j by the integral $\int (1 - x/m)_+^{2\beta} dx$.

The oracle choice is again defined by risk minimization:

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

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

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

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

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

$$U \widetilde{\boldsymbol{\theta}}_G = (\boldsymbol{I}_p + \sigma^2 D^2)^{-1} U \boldsymbol{Z}$$

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

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

Exercise 3.4.2. Let $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\alpha}}$ is a shrinkage estimate for a vector $\boldsymbol{\alpha} = (\alpha_j)$. Then there is a diagonal penalty matrix G such that $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\alpha}} = \widetilde{\boldsymbol{\theta}}_G$.
Hint: define the j th diagonal entry g_j by the equation $\alpha_j = (1 + \sigma^2 g_j^2)^{-1}$.

3.5 Smoothness constraints and roughness penalty approach

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

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

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

Now we consider a constrained ML problem:

$$\widetilde{\boldsymbol{\theta}}_{G,R} = \operatorname*{argmax}_{\boldsymbol{\theta}\in\mathcal{B}_G(R)} L(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta:\,\|G\boldsymbol{\theta}\|\le R} \|\boldsymbol{Y} - \Psi^\top\boldsymbol{\theta}\|^2. \tag{3.11}$$

The Lagrange multiplier method leads to an unconstrained problem

$$\widetilde{\boldsymbol{\theta}}_{G,\lambda} = \operatorname*{argmin}_{\boldsymbol{\theta}}\bigl\{\|\boldsymbol{Y} - \Psi^\top\boldsymbol{\theta}\|^2 + \lambda\|G\boldsymbol{\theta}\|^2\bigr\}.$$

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

3.6 Shrinkage in a linear inverse problem

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

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

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

3.7 Spectral cut-off and spectral penalization. Diagonal estimates

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

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

$$\widetilde{\boldsymbol{u}} = \Lambda^{-1} \boldsymbol{Z} = (\lambda_1^{-1} z_1, \ldots, \lambda_p^{-1} z_p)^\top .$$

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

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

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

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

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

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

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

The spectral cut-off method can be described as follows. Let $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$ be the intrinsic orthonormal basis of the problem composed of the standardized eigenvectors of

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

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

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

$$\mathcal{R}(\widetilde{\boldsymbol{u}}_{\boldsymbol{\alpha}}) \stackrel{\mathrm{def}}{=} I\!\!E \|\widetilde{\boldsymbol{u}}_{\boldsymbol{\alpha}} - \boldsymbol{u}^*\|^2 = \sum_{j=1}^p |\alpha_j - 1|^2 u_j^{*2} + \sum_{j=1}^p \alpha_j^2 \sigma^2 \lambda_j^{-1}.$$

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

$$I\!\!E \|\widetilde{\boldsymbol{u}}_{\boldsymbol{\alpha}} - \boldsymbol{u}^*\|^2 = \sum_{j=1}^p I\!\!E |\alpha_j \lambda_j^{-1} z_j - u_j^*|^2 = \sum_{j=1}^p \bigl\{ |\alpha_j - 1|^2 u_j^{*2} + \alpha_j^2 \sigma^2 \lambda_j^{-1} \bigr\}$$

as required.

Risk minimization leads to the oracle choice of the vector $\boldsymbol{\alpha}$ or

$$\boldsymbol{\alpha}^* = \operatorname*{argmin}_{\boldsymbol{\alpha}} \mathcal{R}(\widetilde{\boldsymbol{u}}_{\boldsymbol{\alpha}})$$

where the minimum is taken over the set of all admissible vectors $\boldsymbol{\alpha}$.

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

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

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

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

4

Ordered model selection for linear smoothers

Consider the linear model $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$. Suppose that a family of linear smoothers $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$ is given, where $\mathcal{S}_m$ is for each $m \in \mathcal{M}$ a given $p \times n$ matrix. The task is to develop a data based model selector $\widehat{m}$ which performs nearly as good as the optimal choice which depends on the model and is not available.

4.1 Model and problem

This section defines our setup.

4.1.1 Loss and risk

First we have to define the loss and risk. We focus on the quadratic loss and risk with a weighting $q \times p$ matrix W:

$$\begin{aligned} \varrho_m &\stackrel{\text{def}}{=} \|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2, \\ \mathcal{R}_m &\stackrel{\text{def}}{=} I\!\!E\|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2. \end{aligned} \tag{4.1}$$

Of course, the loss and the risk depend on the choice of W. We drop this dependence but it is important to keep in mind the role of W in the definition of ϱ_m. Typical examples of W are as follows.

Estimation of the whole vector $\boldsymbol{\theta}^*$:

Let W be the identity matrix $W = \boldsymbol{I}_p$ with $q = p$. This means that the estimation loss is measured by the usual squared Euclidean norm $\|\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\|^2$.

Prediction:

Let W be the square root of the total Fisher information matrix $\mathbb{F} = \sigma^{-2}\Psi\Psi^\top$, that is, $W^2 = \mathbb{F}$. Then minimization of the loss ϱ_m is equivalent to maximizing the log-likelihood $-(2\sigma^2)^{-1}\|\Psi^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2$. Such type of loss is usually referred to as *prediction*

loss because it measures the fit and the prediction ability of the true model by the model with the parameter $\boldsymbol{\theta}$.

Semiparametric estimation:

Let the target of estimation is not the whole vector $\boldsymbol{\theta}^*$ but its subvector $\boldsymbol{\theta}_0^*$ of dimension q. The estimate $\Pi\widetilde{\boldsymbol{\theta}}_m$ is called the *profile maximum likelihood estimate*. The matrix W can be defined as the projector Π_0 on the $\boldsymbol{\theta}_0^*$ subspace. The corresponding loss is equal to the squared Euclidean norm in this subspace:

$$\varrho_m = \|\Pi_0(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2.$$

Alternatively, one can select W^2 as the efficient Fisher information matrix defined by the relation

$$W^2 \stackrel{\text{def}}{=} \breve{\mathbb{F}} = \bigl(\Pi_0 \mathbb{F}^{-1} \Pi_0^\top\bigr)^{-1}.$$

Linear functional estimation:

The choice of the weighting matrix W can be adjusted to the problem of estimating some functionals of the whole parameter $\boldsymbol{\theta}^*$. For instance, in the regression problem $\mathbb{E}Y_i = f(X_i)$ with the Fourier expansion the target function f can be represented as

$$f(x) = \sum_{j\geq 0} \theta_j^* \psi_j(x) = \sum_{j\geq 0} \bigl\{\theta_{2j}^* \cos(2\pi j x) + \theta_{2j+1}^* \sin(2\pi j x)\bigr\}.$$

The value of this function at zero coincides with the functional

$$f(0) = \sum_j \theta_{2j}^*.$$

The first derivative of this function leads to the functional

$$f'(0) = 2\pi \sum_{j\geq 0} j\theta_{2j+1}^*.$$

Exercise 4.1.1. Describe the linear functions for the values

- $f(0.5)$;
- $f''(1)$;
- $\int_0^1 f(x)dx$;
- $\int_0^1 \sin(2\pi x) f'(x)dx$.

The most important feature of the estimate $\widetilde{\boldsymbol{\theta}}_m$ is *linearity*. It allows us to study its properties similarly to above. The next result describes the risk decomposition of $\widetilde{\boldsymbol{\theta}}_m$.

Theorem 4.1.1. *Let* $I\!\!E\boldsymbol{\varepsilon} = 0$, $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$. *Then*

$$\begin{aligned} I\!\!E\widetilde{\boldsymbol{\theta}}_m &= \boldsymbol{\theta}^*_m = \mathcal{S}_m \boldsymbol{f}^*, \\ \mathrm{Var}\bigl(\widetilde{\boldsymbol{\theta}}_m\bigr) &= V_m^2 = \sigma^2 \mathcal{S}_m \mathcal{S}_m^\top, \\ \mathcal{R}_m &= \|W(\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^*)\|^2 + \sigma^2 \operatorname{tr}\bigl(W \mathcal{S}_m \mathcal{S}_m^\top W^\top\bigr) \\ &= \|W(\mathcal{S}_m - \mathcal{S})\boldsymbol{f}^*\|^2 + \sigma^2 \operatorname{tr}\bigl(W \mathcal{S}_m \mathcal{S}_m^\top W^\top\bigr). \end{aligned} \tag{4.2}$$

The expression for the risk $\mathcal{R}_m$ in (4.2) can be called the "bias-variance" decomposition.

4.1.2 Oracle choice. Bias-variance trade-off

Let the weighting matrix W be fixed and the loss and risk of $\widetilde{\boldsymbol{\theta}}_m$ be defined by (4.1). The optimal (oracle) choice of the tuning parameter m can be defined as the risk minimizer:

$$m^* \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_{m \in \mathcal{M}} \mathcal{R}_m.$$

The *model selection* problem can be described as the choice of m by data which *mimics the oracle*, that is, we aim at constructing a selector $\widehat{m}$ leading to the adaptive estimate $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{m}}$ with the properties similar to the oracle estimate $\widetilde{\boldsymbol{\theta}}_{m^*}$.

Below we discuss the *ordered case*. The parameter m is treated as complexity of the method $\widetilde{\boldsymbol{\theta}}_m$ and this parameter may grows until the maximal value $\mathtt{M}$. In some cases the set $\mathcal{M}$ of possible m choices can be countable and/or continuous and even unbounded. For simplicity of presentation, we assume that $\mathcal{M}$ is finite. We also assume without loss of generality that m is an integer non-negative number. Complexity can be naturally expressed via the variance of the stochastic term of the estimate $\widetilde{\boldsymbol{\theta}}_m$: the larger m, the larger is the variance $\mathrm{Var}(W\widetilde{\boldsymbol{\theta}}_m)$. Due to the result of Theorem 4.1.1, this condition can be written as

$$W \mathcal{S}_m \mathcal{S}_m^\top W^\top \le W \mathcal{S}_{m'} \mathcal{S}_{m'}^\top W^\top, \qquad m' > m. \tag{4.3}$$

In the case of projection estimation with the identity matrix W, this variance is linear in m, $\mathrm{Var}(\widetilde{\boldsymbol{\theta}}_m) = \sigma^2 m$. In general the dependence of the variance term on m may be more complicated but the monotonicity constraint (4.3) has to be preserved.

Further, it is implicitly assumed that the bias term $\|W(\boldsymbol{\theta}^* - \boldsymbol{\theta}^*_m)\|^2$ becomes small when m increases. The smallest value $m = 0$ corresponds to the simplest (zero) model with probably a large bias, while m large ensures a good approximation quality $\boldsymbol{\theta}^*_m \approx \boldsymbol{\theta}^*$ and a small bias at cost of a big complexity measured by the variance term. However,

in general, in the contrary to the case of projection estimation, one cannot require that the bias term $\|W(\boldsymbol{\theta}^* - \boldsymbol{\theta}^*_m)\|$ monotonously decreases with m. One example is given by estimation-at-a-point problem.

Exercise 4.1.2. Build an example of a signal $\boldsymbol{\theta}^*$ in the sequence space model $Y_i = \theta^*_j + \varepsilon_j$ such that the bias $\|W(\boldsymbol{\theta}_m - \boldsymbol{\theta}^*)\|$ is not monotonous in m.

Hint: Consider, e.g., $\widetilde{\boldsymbol{\theta}}_m$ being the projector on the first m coordinates and $W\boldsymbol{\theta} = \sum_j \theta_j$. Take $\boldsymbol{\theta}^*$ by alternating blocks of 1's and -1's with equal length.

4.2 Unbiased risk estimation

The empirical risk is obtained by substituting the data $\boldsymbol{Y}$ in place of the true function $\boldsymbol{f}^*$. Let $\widetilde{\boldsymbol{\theta}} = \mathcal{S}\boldsymbol{Y}$ be the estimate from the largest considered model which we assume to be unbiased: $I\!\!E\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^* = \mathcal{S}\boldsymbol{f}^*$. The analog of the empirical risk is

$$\|W(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}})\|^2 = \|W(\mathcal{S}_m - \mathcal{S})\boldsymbol{Y}\|^2.$$

Lemma 4.2.1. *Let* $I\!\!E\boldsymbol{\varepsilon} = 0$, $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$. *Then*

$$I\!\!E\big\|W(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}})\big\|^2 = \|W(\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^*)\|^2 + \sigma^2 \operatorname{tr}\big\{W(\mathcal{S} - \mathcal{S}_m)(\mathcal{S} - \mathcal{S}_m)^\top W^\top\big\}.$$

Proof. The use of $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$ and $I\!\!E\boldsymbol{\varepsilon} = 0$ implies

$$\begin{aligned} I\!\!E\big\|W(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}})\big\|^2 &= I\!\!E\|W(\mathcal{S}_m - \mathcal{S})(\boldsymbol{f}^* + \boldsymbol{\varepsilon})\|^2 \\ &= \|W(\mathcal{S}_m - \mathcal{S})\boldsymbol{f}^*\|^2 + I\!\!E\|W(\mathcal{S}_m - \mathcal{S})\boldsymbol{\varepsilon}\|^2. \end{aligned}$$

For the quadratic stochastic term, we use the trace representation and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$:

$$\begin{aligned} I\!\!E\|W(\mathcal{S}_m - \mathcal{S})\boldsymbol{\varepsilon}\|^2 &= I\!\!E\big[\operatorname{tr}\big\{W(\mathcal{S}_m - \mathcal{S})\boldsymbol{\varepsilon}\big\}^\top \big\{W(\mathcal{S}_m - \mathcal{S})\boldsymbol{\varepsilon}\big\}\big] \\ &= I\!\!E\big[\operatorname{tr}\big\{W(\mathcal{S}_m - \mathcal{S})\boldsymbol{\varepsilon}\big\}\big\{W(\mathcal{S}_m - \mathcal{S})\boldsymbol{\varepsilon}\big\}^\top\big] \\ &= \operatorname{tr}\big\{W(\mathcal{S}_m - \mathcal{S})I\!\!E\big(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\big)(\mathcal{S}_m - \mathcal{S})^\top W^\top\big\} \\ &= \sigma^2 \operatorname{tr}\big\{W(\mathcal{S} - \mathcal{S}_m)(\mathcal{S} - \mathcal{S}_m)^\top W^\top\big\}. \end{aligned}$$

This proves the required decomposition.

Exercise 4.2.1. Derive a similar expansion for an inhomogeneous noise $\boldsymbol{\varepsilon}$ with $\mathrm{Var}(\boldsymbol{\varepsilon}) = \Sigma$.

Now we compare the risk $\mathcal{R}_m$ of $\widetilde{\boldsymbol{\theta}}_m$ with the expected empirical risk $I\!\!E\|W(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}})\|^2$. Similarly to the projection case, the bias term $\|W(\mathcal{S}_m - \mathcal{S})\boldsymbol{f}^*\|^2$ of the risk $\mathcal{R}_m$ and

of the expectation of the empirical risk are the same. The variance terms are different but known. So, one can make a correction and build an unbiased risk estimate $\widetilde{\mathcal{R}}_m$. Indeed, by construction

$$\mathbb{E}\|W(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}})\|^2 + \sigma^2 \operatorname{tr}\{W\big[\mathcal{S}_m\mathcal{S}_m^\top - (\mathcal{S}-\mathcal{S}_m)(\mathcal{S}-\mathcal{S}_m)^\top\big]W^\top\} = \mathcal{R}_m.$$

This suggests the following definition of the unbiased risk estimate $\widetilde{\mathcal{R}}_m$:

$$\begin{aligned}\widetilde{\mathcal{R}}_m &\stackrel{\text{def}}{=} \|W(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}})\|^2 + 2\sigma^2 \operatorname{tr}\big(W\mathcal{S}\mathcal{S}_m^\top W^\top\big) \\ &= \|W(\mathcal{S}_m - \mathcal{S})\boldsymbol{Y}\|^2 + 2\sigma^2 \operatorname{tr}\big(W\mathcal{S}\mathcal{S}_m^\top W^\top\big). \end{aligned} \tag{4.4}$$

The term $\operatorname{tr}\big(W\mathcal{S}\mathcal{S}^\top W^\top\big)$ is omitted because it does not depend on m.

Exercise 4.2.2. Let $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$ and $\mathcal{S}_m = \Pi_m$ is a projector on the subspace spanned by $\theta_1, \ldots \theta_m$.

1. Check that

$$W\mathcal{S}\mathcal{S}_m^\top W^\top = W\mathcal{S}_m\mathcal{S}_m^\top W^\top = W\mathcal{S}_m W^\top.$$

2. Compute $\widetilde{\mathcal{R}}_m$ if W projects $\boldsymbol{\theta}$ onto the first q components.

The Stein unbiased risk estimate (SURE) $\widehat{m}$ is defined by minimization of $\widetilde{\mathcal{R}}_m$ from (4.4)

$$\widehat{m} = \operatorname*{argmin}_m \widetilde{\mathcal{R}}_m. \tag{4.5}$$

4.2.1 Zone of insensitivity

Here we discuss whether the SURE method (4.5) does the job of model selection. The study is again done by the analysis of pairwise comparison. Let the oracle choice m^* is defined by risk minimization. By construction and the definition of the oracle, it implies

$$\mathbb{E}\big(\widetilde{\mathcal{R}}_m - \widetilde{\mathcal{R}}_{m^*}\big) = \mathcal{R}_m - \mathcal{R}_{m^*} \geq 0.$$

So, one can expect that it works "in mean". However, similarly to the projection case, we need a more detailed and rigorous analysis of the probability of selecting $\widehat{m} = m \neq m^*$ which is related to the probability of the inequality $\widetilde{\mathcal{R}}_m - \widetilde{\mathcal{R}}_{m^*} < 0$. Intuitively, if the systematic part (expectation) $\mathcal{R}_m - \mathcal{R}_{m^*}$ is significantly large, the probability of selecting $\widehat{m} = m$ is small. In the region where the risk difference $\mathcal{R}_m - \mathcal{R}_{m^*}$ is small, the selector $\widehat{m}$ most likely makes a random choice.

For a detailed study, we need to slightly extend the monotonicity condition. Namely, it is required that the variance of $W\widetilde{\boldsymbol{\theta}}_m$ grows with m while the variance of $W(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}})$ decreases with m. For the projection method these two conditions coincide.

We begin with the following lemma.

Lemma 4.2.2. *For each $m^\circ \neq m$, it holds*

$$\widetilde{\mathcal{R}}_m - \widetilde{\mathcal{R}}_{m^\circ} = \begin{cases} -\boldsymbol{Y}^\top \mathcal{D}^2_{m,m^\circ} \boldsymbol{Y} + \mathcal{Q}_{m,m^\circ}, & m > m^\circ, \\ \boldsymbol{Y}^\top \mathcal{D}^2_{m^\circ,m} \boldsymbol{Y} - \mathcal{Q}_{m^\circ,m}, & m < m^\circ, \end{cases} \tag{4.6}$$

where for $m > m^\circ$

$$\begin{aligned} \mathcal{D}^2_{m,m^\circ} &\stackrel{\text{def}}{=} (\mathcal{S} - \mathcal{S}_{m^\circ})^\top W^\top W (\mathcal{S} - \mathcal{S}_{m^\circ}) - (\mathcal{S} - \mathcal{S}_m)^\top W^\top W (\mathcal{S} - \mathcal{S}_m), \\ \mathcal{Q}_{m,m^\circ} &\stackrel{\text{def}}{=} \sigma^2 \operatorname{tr}\bigl(\mathcal{D}^2_{m,m^\circ}\bigr) + \sigma^2 \operatorname{tr}\bigl(D^2_{m,m^\circ}\bigr), \\ D^2_{m,m^\circ} &= \mathcal{S}_m^\top W^\top W \mathcal{S}_m - \mathcal{S}_{m^\circ}^\top W^\top W \mathcal{S}_{m^\circ}. \end{aligned} \tag{4.7}$$

Proof. The definition (4.4) yields

$$\begin{aligned} \widetilde{\mathcal{R}}_m - \widetilde{\mathcal{R}}_{m^\circ} &= \boldsymbol{Y}^\top (\mathcal{S}_m - \mathcal{S})^\top W^\top W (\mathcal{S}_m - \mathcal{S}) \boldsymbol{Y} + 2\sigma^2 \operatorname{tr}\bigl(W \mathcal{S} \mathcal{S}_m^\top W^\top\bigr) \\ &\quad - \boldsymbol{Y}^\top (\mathcal{S}_{m^\circ} - \mathcal{S})^\top W^\top W (\mathcal{S}_{m^\circ} - \mathcal{S}) \boldsymbol{Y} - 2\sigma^2 \operatorname{tr}\bigl\{W \mathcal{S} \mathcal{S}_{m^\circ}^\top W^\top\bigr\} \end{aligned}$$

and it remains to check that

$$2 \operatorname{tr}\bigl\{W \mathcal{S} \bigl(\mathcal{S}_m - \mathcal{S}_{m^\circ}\bigr)^\top W^\top\bigr\} = \operatorname{tr}\bigl(\mathcal{D}^2_{m,m^\circ}\bigr) + \operatorname{tr}\bigl(D^2_{m,m^\circ}\bigr).$$

This identity follows directly from the definition.

Exercise 4.2.3. Let $\Psi \mathcal{S}_m$ be a projector in $I\!R^n$ on a linear subspace $\mathcal{L}_m$ of dimension m for a growing system $\mathcal{L}_1 \subset \mathcal{L}_2 \subset \ldots \subset \mathcal{L}_{\mathtt{M}}$:

1. Check the monotonicity condition $\mathcal{D}^2_{m,m^\circ} \geq 0$ for all $m > m^\circ$;
2. Write explicitly the expansion (4.6);
3. Check for the case with $W = \Psi^\top$ that

$$\mathcal{D}^2_{m,m^\circ} = D^2_{m,m^\circ} = \mathcal{S}_m W W^\top \mathcal{S}_m^\top - \mathcal{S}_{m^\circ} W W^\top \mathcal{S}_{m^\circ}^\top;$$

4. Specify $\mathcal{D}^2_{m,m^\circ}$ and $\mathcal{Q}_{m,m^\circ}$ for the situation when the loss function W only depends on the first coordinate: $W\boldsymbol{\theta} = \theta_1$.

Now we apply Lemma 4.2.2 to study the probability of the event $\{\widetilde{\mathcal{R}}_m < \widetilde{\mathcal{R}}_{m^*}\}$. The decomposition (4.6) allows to represent this event as

$$(\boldsymbol{f}^* + \boldsymbol{\varepsilon})^\top \mathcal{D}^2_{m,m^*}(\boldsymbol{f}^* + \boldsymbol{\varepsilon}) \geq \sigma^2 \operatorname{tr}\bigl(\mathcal{D}^2_{m,m^*}\bigr) + \sigma^2 \operatorname{tr}\bigl(D^2_{m,m^*}\bigr), \qquad m > m^*, \tag{4.8}$$

$$(\boldsymbol{f}^* + \boldsymbol{\varepsilon})^\top \mathcal{D}^2_{m^*,m}(\boldsymbol{f}^* + \boldsymbol{\varepsilon}) \leq \sigma^2 \operatorname{tr}\bigl(\mathcal{D}^2_{m^*,m}\bigr) + \sigma^2 \operatorname{tr}\bigl(D^2_{m^*,m}\bigr), \qquad m < m^*. \tag{4.9}$$

This probability can be studied in the same way with separate consideration for $m > m^*$ and $m < m^*$. In the region $m > m^*$, in view of the definition (4.7), the condition (4.8) can be rewritten as

$$\begin{aligned} &\varepsilon^\top \mathcal{D}^2_{m,m^*}\varepsilon - I\!\!E\bigl(\varepsilon^\top \mathcal{D}^2_{m,m^*}\varepsilon\bigr) + 2\varepsilon^\top \mathcal{D}^2_{m,m^*}\boldsymbol{f}^* \\ &\quad \geq \sigma^2 \operatorname{tr}\bigl(D^2_{m,m^*}\bigr) - \boldsymbol{f}^{*\top} \mathcal{D}^2_{m,m^*}\boldsymbol{f}^* = \mathcal{R}_m - \mathcal{R}_{m^*}. \end{aligned} \tag{4.10}$$

For typical ordering schemes, one can expect that the bias component $\boldsymbol{f}^{*\top} \mathcal{D}^2_{m,m^*}\boldsymbol{f}^*$ in the risk difference $\mathcal{R}_m - \mathcal{R}_{m^*}$ remains bounded, while the variance term $\sigma^2 \operatorname{tr}\bigl(D^2_{m,m^*}\bigr)$ grows and tends to dominate the bias for $m \gg m^*$. The deviation bound of Theorem 3.3.2 can be used to check under which conditions the event in (4.8) or (4.10) can occurs only with a very small probability. In particular, with a dominating probability the centered quadratic form in (4.10) can be bounded as follows

$$\varepsilon^\top \mathcal{D}^2_{m,m^*}\varepsilon - I\!\!E\bigl(\varepsilon^\top \mathcal{D}^2_{m,m^*}\varepsilon\bigr) \leq \mathtt{C}\sigma^2 \sqrt{\mathtt{x} \operatorname{tr}\bigl(\mathcal{D}^4_{m,m^*}\bigr)}$$

for a fixed constant $\mathtt{C}$. Similarly, the linear term $2\varepsilon^\top \mathcal{D}^2_{m,m^*}\boldsymbol{f}^*$ after standardization can be bounded by a quantile of the standard normal law: with a dominating probability

$$\varepsilon^\top \mathcal{D}^2_{m,m^*}\boldsymbol{f}^* \leq \sigma \|\mathcal{D}^2_{m,m^*}\boldsymbol{f}^*\| \, z_1(\mathtt{x}).$$

One can conclude that the SURE selector does the job for values of $m > m^*$ for which

$$\operatorname{tr}(D^2_{m,m^*}) - \sigma^{-2}\boldsymbol{f}^{*\top} \mathcal{D}^2_{m,m^*}\boldsymbol{f}^* \geq \mathtt{C}\sqrt{\mathtt{x} \operatorname{tr}(\mathcal{D}^4_{m,m^*})} + \sigma^{-1}\|\mathcal{D}^2_{m,m^*}\boldsymbol{f}^*\| \, z_1(\mathtt{x})$$

for some fixed constant $\mathtt{C}$. The region of insensitivity $\mathcal{M}^\circ(\mathtt{x})$ includes those $m > m^*$, for which this condition does not meet.

Similarly one can consider the case of $m < m^*$. Here the monotonicity condition ensures that the variance term $\operatorname{tr}(\mathcal{D}^2_{m,m^*})$ is bounded from above for all $m < m^*$. However, the definition of the oracle m^* yields a significant bias for $m < m^*$. The condition (4.9) can be represented as

$$\begin{aligned} &-\varepsilon^\top \mathcal{D}^2_{m^*,m}\varepsilon + I\!\!E\bigl(\varepsilon^\top \mathcal{D}^2_{m^*,m}\varepsilon\bigr) - 2\varepsilon^\top \mathcal{D}^2_{m^*,m}\boldsymbol{f}^* \\ &\quad \geq -\sigma^2 \operatorname{tr}\bigl(D^2_{m^*,m}\bigr) + \boldsymbol{f}^{*\top} \mathcal{D}^2_{m^*,m}\boldsymbol{f}^* = \mathcal{R}_m - \mathcal{R}_{m^*}. \end{aligned}$$

Here the zone of insensitivity corresponds to values of m for which the bias $\boldsymbol{f}^{*\top} \mathcal{D}^2_{m^*,m}\boldsymbol{f}^*$ is not too large relative to the standard deviation of the quadratic form $\varepsilon^\top \mathcal{D}^2_{m^*,m}\varepsilon$. The bound of Theorem 3.3.2 leads to a sufficient condition

$$\sigma^{-2}\boldsymbol{f}^{*\top}\mathcal{D}^2_{m^*,m}\boldsymbol{f}^* - \operatorname{tr}(D^2_{m^*,m}) \geq \mathtt{C}\sqrt{\operatorname{tr}(\mathcal{D}^4_{m^*,m})} + \sigma^{-1}\|\mathcal{D}^2_{m,m^*}\boldsymbol{f}^*\|\, z_1(\mathtt{x})\,.$$

Under this condition the probability of selecting such a value m is negligible.

In general, one can conclude that the selector based on the unbiased risk estimation, works well beyond the region of insensitivity which means that the risk difference $\mathcal{R}_m - \mathcal{R}_{m^*}$ is prominent relative to the standard deviation of the quadratic form $\boldsymbol{\varepsilon}^\top\mathcal{D}^2_{m^*,m}\boldsymbol{\varepsilon}$.

4.2.2 An oracle bound

For stating an oracle bound on the loss and risk of $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{m}}$, we also have to bound the excess, that is, the difference between the loss of the adaptive and the oracle procedure in the region of insensitivity. We only present a bound for each particular m.

Lemma 4.2.3. *It holds for* $m > m^*$

$$\varrho_m - \varrho_{m^*} = \boldsymbol{\varepsilon}^\top D^2_{m,m^*}\boldsymbol{\varepsilon} - \boldsymbol{f}^{*\top}\mathcal{D}^2_{m,m^*}\boldsymbol{f}^* + 2\boldsymbol{b}^\top_{m,m^*}\boldsymbol{\varepsilon} \tag{4.11}$$

with

$$\boldsymbol{b}_{m,m^*} \stackrel{\text{def}}{=} \boldsymbol{f}^{*\top}(\mathcal{S}_m - \mathcal{S})^\top W^\top W\mathcal{S}_m - \boldsymbol{f}^{*\top}(\mathcal{S}_{m^*} - \mathcal{S})^\top W^\top W\mathcal{S}_{m^*}\,.$$

Moreover,

$$\varrho_m - \varrho_{m^*} - (\mathcal{R}_m - \mathcal{R}_{m^*}) = \boldsymbol{\varepsilon}^\top D^2_{m,m^*}\boldsymbol{\varepsilon} - I\!\!E\big(\boldsymbol{\varepsilon}^\top D^2_{m,m^*}\boldsymbol{\varepsilon}\big) + 2\boldsymbol{b}^\top_{m,m^*}\boldsymbol{\varepsilon}. \tag{4.12}$$

Proof. The use of $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m\boldsymbol{Y} = \mathcal{S}_m(\boldsymbol{f}^* + \boldsymbol{\varepsilon})$ and $\boldsymbol{\theta}^* = \mathcal{S}\boldsymbol{f}^*$ for $m > m^*$ by (4.2)

$$\begin{aligned}
\varrho_m - \varrho_{m^*} &= \|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2 - \|W(\widetilde{\boldsymbol{\theta}}_{m^*} - \boldsymbol{\theta}^*)\|^2 \\
&= \|W\mathcal{S}_m\boldsymbol{\varepsilon}\|^2 - \|W\mathcal{S}_{m^*}\boldsymbol{\varepsilon}\|^2 + \|W(\mathcal{S}_m - \mathcal{S})\boldsymbol{f}^*\|^2 - \|W(\mathcal{S}_{m^*} - \mathcal{S})\boldsymbol{f}^*\|^2 \\
&\quad + 2\boldsymbol{f}^{*\top}(\mathcal{S}_m - \mathcal{S})^\top W^\top W\mathcal{S}_m\boldsymbol{\varepsilon} - 2\boldsymbol{f}^{*\top}(\mathcal{S}_{m^*} - \mathcal{S})^\top W^\top W\mathcal{S}_{m^*}\boldsymbol{\varepsilon} \\
&= \boldsymbol{\varepsilon}^\top D^2_{m,m^*}\boldsymbol{\varepsilon} - \boldsymbol{f}^{*\top}\mathcal{D}^2_{m,m^*}\boldsymbol{f}^* + 2\boldsymbol{b}^\top_{m,m^*}\boldsymbol{\varepsilon}
\end{aligned} \tag{4.13}$$

and (4.11) follows. The assertion (4.12) holds by formula (4.2) for the risk $\mathcal{R}_m$.

Further, the deviation of the centered quadratic form $\boldsymbol{\varepsilon}^\top D^2_{m,m^*}\boldsymbol{\varepsilon} - I\!\!E\big(\boldsymbol{\varepsilon}^\top D^2_{m,m^*}\boldsymbol{\varepsilon}\big)$ can be bounded with a probability $1 - \mathrm{e}^{-\mathtt{x}}$ as

$$\big|\boldsymbol{\varepsilon}^\top D^2_{m,m^*}\boldsymbol{\varepsilon} - I\!\!E\big(\boldsymbol{\varepsilon}^\top D^2_{m,m^*}\boldsymbol{\varepsilon}\big)\|^2\big| \leq \mathtt{C}\sigma^2\sqrt{\mathtt{x}\operatorname{tr}(D^4_{m,m^*})}$$

while the linear term in (4.13) can be bounded in the same sense by

$$\big|\boldsymbol{b}^\top_{m,m^*}\boldsymbol{\varepsilon}\big| \leq \sigma\|\boldsymbol{b}_{m,m^*}\|\, z_1(\mathtt{x}).$$

An increase of $\mathtt{x}$ by $\log(\mathbb{N}(\mathtt{x})) = \log\bigl(\#\mathcal{M}^\circ(\mathtt{x})\bigr)$ yields a uniform bound over the region of insensitivity $\mathcal{M}^\circ(\mathtt{x})$. One can conclude that the risk of the adaptive estimate $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{m}}$ is close to the risk of the oracle estimate $\widetilde{\boldsymbol{\theta}}_{m^*}$ if the range of $\sqrt{\mathrm{tr}(D^4_{m,m^*})}$ over the set $\mathcal{M}^\circ(\mathtt{x})$ is small relative to the oracle risk $\mathcal{R}_{m^*}$.

To be done: An oracle risk bound for the SURE method

4.3 Smallest accepted (SmA) method

This section discusses another approach based on multiple testing. Let H_{m° vs. H_m mean a hypothesis of no significant bias between the models m° and m. Instead of the likelihood ratio test, we consider the test statistic $\mathbb{T}_{m,m^\circ}$ based on the weighted difference $W\bigl(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ}\bigr)$:

$$\mathbb{T}_{m,m^\circ} = \|W(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ})\| = \|W(\mathcal{S}_m - \mathcal{S}_{m^\circ})\boldsymbol{Y}\|.$$

4.3.1 Decomposition of the test statistic. Bias and variance

Below we use the decomposition

$$\mathbb{T}_{m,m^\circ} = \|W(\mathcal{S}_m - \mathcal{S}_{m^\circ})(\boldsymbol{f}^* + \boldsymbol{\varepsilon})\| = \|\boldsymbol{b}_{m,m^\circ} + \boldsymbol{\xi}_{m,m^\circ}\|, \tag{4.14}$$

where

$$\boldsymbol{b}_{m,m^\circ} \stackrel{\mathrm{def}}{=} W(\mathcal{S}_m - \mathcal{S}_{m^\circ})\boldsymbol{f}^*,$$
$$\boldsymbol{\xi}_{m,m^\circ} \stackrel{\mathrm{def}}{=} W(\mathcal{S}_m - \mathcal{S}_{m^\circ})\boldsymbol{\varepsilon}.$$

It obviously holds $I\!\!E\boldsymbol{\xi}_{m,m^\circ} = 0$. Introduce $q \times q$-matrix $\mathbb{V}_{m,m^\circ}$ as the variance of $W\bigl(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ}\bigr)$:

$$\mathbb{V}_{m,m^\circ} \stackrel{\mathrm{def}}{=} \mathrm{Var}\bigl\{W\bigl(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ}\bigr)\bigr\} = \mathrm{Var}\bigl\{W(\mathcal{S}_m - \mathcal{S}_{m^\circ})\boldsymbol{Y}\bigr\}.$$

If the noise $\boldsymbol{\varepsilon}$ in the decomposition $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$ is homogeneous with $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$, it holds

$$\begin{aligned}\mathbb{V}_{m,m^\circ} &= W(\mathcal{S}_m - \mathcal{S}_{m^\circ})\,\mathrm{Var}(\boldsymbol{\varepsilon})(\mathcal{S}_m - \mathcal{S}_{m^\circ})^\top W^\top \\ &= \sigma^2 W(\mathcal{S}_m - \mathcal{S}_{m^\circ})(\mathcal{S}_m - \mathcal{S}_{m^\circ})^\top W^\top.\end{aligned}$$

Note that in the case of homogeneous noise the matrix $\mathbb{V}_{m,m^\circ}$ is easily computable. The matrices W and $\mathcal{S}_m$ are given by the method and only the noise variance has to be estimated from the data.

Further,

$$\mathbb{E}\,\mathbb{T}^2_{m,m^\circ} \;=\; \|\boldsymbol{b}_{m,m^\circ}\|^2 + \mathbb{E}\|\boldsymbol{\xi}_{m,m^\circ}\|^2 = \|\boldsymbol{b}_{m,m^\circ}\|^2 + \operatorname{tr}(\mathbb{V}_{m,m^\circ}). \tag{4.15}$$

The bias term $\boldsymbol{b}_{m,m^\circ} \stackrel{\text{def}}{=} W(\mathcal{S}_m - \mathcal{S}_{m^\circ})\boldsymbol{f}^*$ is significant if its squared norm is competitive with the variance term $\mathtt{p}_{m,m^\circ} = \operatorname{tr}(\mathbb{V}_{m,m^\circ})$.

Below we proceed with the Gaussian errors $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Then $\boldsymbol{\xi}_{m,m^\circ} \sim \mathcal{N}(0, \mathbb{V}_{m,m^\circ})$ and its distribution is completely specified by the covariance matrix $\mathbb{V}_{m,m^\circ}$ and it holds for any $\mathtt{x}$

$$I\!P\Big(\|\boldsymbol{\xi}_{m,m^\circ}\| > z(\mathbb{V}_{m,m^\circ}, \mathtt{x})\Big) \;=\; \mathrm{e}^{-\mathtt{x}} \tag{4.16}$$

for the tail function $z(\mathbb{V}, \mathtt{x})$ of $\|\boldsymbol{\xi}\|$ with $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbb{V})$.

4.3.2 Multiplicity correction. A Bonferroni bound and Monte-Carlo method

Below we need a uniform in $m > m^\circ$ version of the probability bound (4.16). Let

$$\mathcal{M}^+(m^\circ) \stackrel{\text{def}}{=} \big\{m \in \mathcal{M}\colon m > m^\circ\big\}.$$

By $q_{m^\circ} = q_{m^\circ}(\mathtt{x})$ denote the corresponding multiplicity correction:

$$I\!P\Big(\bigcup_{m \in \mathcal{M}^+(m^\circ)} \big\{\|\boldsymbol{\xi}_{m,m^\circ}\| \ge z(\mathbb{V}_{m,m^\circ}, \mathtt{x} + q_{m^\circ}(\mathtt{x}))\big\}\Big) \;=\; \mathrm{e}^{-\mathtt{x}}. \tag{4.17}$$

We write $\mathtt{x}_{m^\circ} \stackrel{\text{def}}{=} \mathtt{x} + q_{m^\circ}(\mathtt{x})$. A simple way of computing the multiplicity correction q_{m° is based on the Bonferroni bound: $q_{m^\circ} = \log(\#\mathcal{M}^+(m^\circ))$. Indeed

$$\begin{aligned}
&I\!P\Big(\bigcup_{m \in \mathcal{M}^+(m^\circ)} \big\{\|\boldsymbol{\xi}_{m,m^\circ}\| \ge z\big(\mathbb{V}_{m,m^\circ}, \mathtt{x} + \log(\#\mathcal{M}^+(m^\circ))\big)\big\}\Big) \\
&\quad\le \sum_{m \in \mathcal{M}^+(m^\circ)} I\!P\Big(\|\boldsymbol{\xi}_{m,m^\circ}\| \ge z\big(\mathbb{V}_{m,m^\circ}, \mathtt{x} + \log(\#\mathcal{M}^+(m^\circ))\big)\Big) \\
&\quad\le \sum_{m \in \mathcal{M}^+(m^\circ)} \mathrm{e}^{-\mathtt{x} - \log(\#\mathcal{M}^+(m^\circ))} = \mathrm{e}^{-\mathtt{x}}.
\end{aligned}$$

However, it is well known that the Bonferroni bound is very conservative and leads to very large correction q_{m°, especially if the random vectors $\boldsymbol{\xi}_{m,m^\circ}$ are strongly correlated. This is exactly the case under consideration. Note that the joint distribution of the $\boldsymbol{\xi}_{m,m^\circ}$'s is precisely known. An analytic bound for the correction $q_{m^\circ}(\mathtt{x})$ is hardly computable but one can use Monte-Carlo experiments: draw K samples $\boldsymbol{\varepsilon}^{(k)} \sim \mathcal{N}(0, \sigma^2 I_n)$ for $k = 1, \ldots, K$, compute for each the stochastic vectors $\boldsymbol{\xi}^{(k)}_{m,m^\circ}$ for $m \in \mathcal{M}^+(m^\circ)$, and evaluate $q_{m^\circ}(\mathtt{x})$ as the smallest value for which the condition (4.16) is fulfilled under the Monte-Carlo empirical measure $I\!P^\circ$ in place of $I\!P$.

Exercise 4.3.1. Show that for each $\mathtt{x}$ there is unique value $q_{m^\circ}(\mathtt{x})$ ensuring (4.17). Check whether q_{m° depends on $\mathtt{x}$.

4.3.3 Oracle choice

We say that m° is a good choice if there is no significant bias $\boldsymbol{b}_{m,m^\circ}$ for any $m > m^\circ$. This condition can be quantified as

$$\|\boldsymbol{b}_{m,m^\circ}\|^2 \leq \beta^2 \, \mathtt{p}_{m,m^\circ}, \qquad m > m^\circ \tag{4.18}$$

for a given β and $\mathtt{p}_{m,m^\circ} = \mathrm{tr}(\mathbb{V}_{m,m^\circ}) = I\!\!E\|\boldsymbol{\xi}_{m,m^\circ}\|^2$. This condition can be viewed as the "bias-variance trade-off". The parameter β controls the bias component in the risk due to decomposition (4.15). Now define the *oracle* m^* as the minimal m° with the property (4.18):

$$m^* \stackrel{\text{def}}{=} \min\Big\{m^\circ : \max_{m > m^\circ} \big\{\|\boldsymbol{b}_{m,m^\circ}\|^2 - \beta^2 \, \mathtt{p}_{m,m^\circ}\big\} \leq 0\Big\}. \tag{4.19}$$

Exercise 4.3.2. Check that the definition (4.19) uniquely defines the value m^*.

The proposed procedure based on the family of test statistics $\mathbb{T}_{m,m^\circ}$ has to be validated by the condition that any good model in the sense (4.18) will be rejected with a very small nominal probability. In particular, this would apply to the oracle choice m^*.

4.3.4 Data-driven choice and the oracle inequality

Define the selector $\widehat{m}$ by the "smallest accepted" (SmA) rule. Namely, with $\mathtt{x}_{m^\circ} = \mathtt{x} + q_{m^\circ}(\mathtt{x})$; see (4.16), the acceptance rule reads as follows:

$$\big\{m^\circ \text{ is accepted}\big\} \Leftrightarrow \Big\{\max_{m > m^\circ} \big\{\mathbb{T}_{m,m^\circ} - \mathtt{z}_{m,m^\circ}\big\} \leq 0\Big\},$$

$$\mathtt{z}_{m,m^\circ} \stackrel{\text{def}}{=} z(\mathbb{V}_{m,m^\circ}, \mathtt{x}_{m^\circ}) + \beta\sqrt{\mathrm{tr}(\mathbb{V}_{m,m^\circ})}.$$

The SmA rule is

$$\begin{aligned} \widehat{m} &\stackrel{\text{def}}{=} \text{“smallest accepted”} \\ &= \min\Big\{m^\circ : \max_{m > m^\circ} \big\{\mathbb{T}_{m,m^\circ} - \mathtt{z}_{m,m^\circ}\big\} \leq 0\Big\}. \end{aligned} \tag{4.20}$$

Exercise 4.3.3. Check that (4.20) uniquely defines the value $\widehat{m}$.

The bound (4.16) and the definition of the oracle m^* yield the desired propagation property:

$$I\!P\bigl(m^* \text{ is rejected}\bigr) \le \mathrm{e}^{-\mathtt{x}}. \tag{4.21}$$

In some sense, this property is built in the construction of the procedure and is fulfilled automatically. It ensures that with a large probability, the procedure does not reject the point m^*, this model is "good" and has to be accepted. Therefore, the selector $\widehat{m}$ typically takes its value in the region $m \le m^*$. It remains to check the performance of the method in this region.

First we specify the zone of insensitivity for $m < m^*$. In this region we expect a bias component competitive with the variance of the oracle. If the bias $\Delta_{m^*,m} = \|\boldsymbol{b}_{m^*,m}\|$ is significantly large then such m will be accepted with only very small probability. We can use the bound

$$\begin{aligned} &I\!P\bigl(\bigl\|\boldsymbol{b}_{m^*,m} + \boldsymbol{\xi}_{m^*,m}\bigr\| < \mathtt{z}_{m^*,m}\bigr) \\ &\quad \le I\!P\bigl(\|\boldsymbol{\xi}_{m^*,m}\| > \|\boldsymbol{b}_{m^*,m}\| - \mathtt{z}_{m^*,m}\bigr). \end{aligned} \tag{4.22}$$

Acceptance of the model $m < m^*$ assumes that the check of $\mathbb{T}_{m^*,m}$ does not fail. If

$$\|\boldsymbol{b}_{m^*,m}\| \ge \mathtt{z}_{m^*,m} + z(\mathbb{V}_{m^*,m}, \mathtt{x}),$$

then (4.22) implies the upper bound for the probability of accepting this model m:

$$\begin{aligned} I\!P\bigl(m \text{ is accepted}\bigr) &\le I\!P\bigl(\bigl\|\boldsymbol{b}_{m^*,m} + \boldsymbol{\xi}_{m^*,m}\bigr\| < \mathtt{z}_{m^*,m}\bigr) \\ &\le I\!P\bigl(\bigl\|\boldsymbol{\xi}_{m^*,m}\bigr\| > z(\mathbb{V}_{m^*,m}, \mathtt{x})\bigr) \le \mathrm{e}^{-\mathtt{x}}. \end{aligned} \tag{4.23}$$

The definition of the zone of insensitivity requires a small multiplicity correction of the $\mathtt{x}$-level. Define $\overline{\mathtt{x}}_{m^*} \stackrel{\text{def}}{=} \mathtt{x} + \log(m^*)$ and

$$\mathcal{M}^\circ(\mathtt{x}) \stackrel{\text{def}}{=} \bigl\{m < m^* \colon \|\boldsymbol{b}_{m^*,m}\| \le \mathtt{z}_{m^*,m} + z(\mathbb{V}_{m^*,m}, \overline{\mathtt{x}}_{m^*})\bigr\}. \tag{4.24}$$

Theorem 4.3.1. *Let the set $\mathcal{M}^\circ(\mathtt{x})$ be defined by (4.24). Then*

$$I\!P\bigl(\widehat{m} < m^*,\ \widehat{m} \notin \mathcal{M}^\circ(\mathtt{x})\bigr) \le \mathrm{e}^{-\mathtt{x}}.$$

Proof. It suffices to note that by (4.23)

$$\begin{aligned} I\!P\bigl(\widehat{m} < m^*,\ \widehat{m} \notin \mathcal{M}^\circ(\mathtt{x})\bigr) &\le \sum_{m<m^*,\, m \notin \mathcal{M}^\circ(\mathtt{x})} I\!P\bigl(\bigl\|\boldsymbol{b}_{m^*,m} + \boldsymbol{\xi}_{m^*,m}\bigr\| < \mathtt{z}_{m^*,m}\bigr) \\ &\le \sum_{m<m^*,\, m \notin \mathcal{M}^\circ(\mathtt{x})} \mathrm{e}^{-\overline{\mathtt{x}}_{m^*}} \le \mathrm{e}^{-\mathtt{x}}. \end{aligned}$$

This implies the result because the cardinality of the set $\{m < m^*,\, m \notin \mathcal{M}^\circ(\mathtt{x})\}$ does not exceed m^*.

Note that the proof of this result is based on the crude Bonferroni upper bound and the definition of $\mathcal{M}^\circ(\mathtt{x})$ can be refined by choosing $\overline{\mathtt{x}}_{m^*}$ more carefully. However, this value only enters in the theoretical bound and is not used in the procedure, a fine tuning for this value is not required.

One can conclude that the SmA procedure typically selects some m from the set $\mathcal{M}^\circ(\mathtt{x})$. The situation is similar to the case of the SURE procedure with one essential difference: the just defined set $\mathcal{M}^\circ(\mathtt{x})$ is located from the left of the oracle, which enables us to guarantee some oracle quality of estimation.

If $\widehat{m} = m < m^*$, the acceptance of $\widehat{m}$ implies by definition

$$\mathbb{T}_{m^*,m} \leq \mathtt{z}_{m^*,m}$$

or, equivalently

$$\bigl\| W(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^*}) \bigr\| \leq \mathtt{z}_{m^*,m}\,. \tag{4.25}$$

One can see that the variability of the adaptive estimate $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{m}}$ is bounded from above by a fixed deterministic quantity. This enables us to state the final oracle inequality.

Theorem 4.3.2. *Let $\mathtt{z}_{m,m^\circ}$ ensure for each m° the condition (4.16). Then the propagation property (4.21) is fulfilled. The following oracle bound holds for the SmA estimate $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{m}}$ on a random set $\Omega(\mathtt{x})$ of dominating probability $1 - 2\mathrm{e}^{-\mathtt{x}}$:*

$$\bigl\| W(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}_{m^*}) \bigr\| \leq \overline{\mathtt{z}}(m^*) \stackrel{\mathrm{def}}{=} \max_{m \in \mathcal{M}^\circ(\mathtt{x})} \mathtt{z}_{m^*,m}\,. \tag{4.26}$$

This implies the probabilistic oracle bound: with probability at least $1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\bigl\| W(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \bigr\| \leq \bigl\| W(\widetilde{\boldsymbol{\theta}}_{m^*} - \boldsymbol{\theta}^*) \bigr\| + \overline{\mathtt{z}}(m^*). \tag{4.27}$$

Proof. First we note that

$$\mathbb{P}\bigl(\widehat{m} > m^*\bigr) \leq \mathrm{e}^{-\mathtt{x}},$$

$$\mathbb{P}\bigl(\widehat{m} \leq m^*, m \notin \mathcal{M}^\circ(\mathtt{x})\bigr) \leq \mathrm{e}^{-\mathtt{x}}.$$

This implies (4.26) by (4.25). (4.27) follows now by the triangle inequality.

The result (4.27) is called the oracle bound because it compares the loss of the data-driven selector $\widehat{m}$ and of the optimal choice m^*.

The value $\overline{\mathtt{z}}(m^*)$ in the right hand-side of (4.27) can be viewed as "payment for adaptation". An interesting feature of the presented result is that not only the oracle quality but also the payment of adaptation depend upon the unknown value $\boldsymbol{\theta}^*$ and the

corresponding oracle choice m^*. In the worst case of the model with the flat risk $\mathcal{R}_m$, the set $\mathcal{M}^\circ(\mathtt{x})$ can coincide with the whole range $m \le m^*$. Even in this case the bounds (4.25) and (4.27) are meaningful. However, the payment for adaptation $\overline{\mathtt{z}}(m^*)$ in this case can be larger then the oracle risk.

4.3.5 Analysis of the payment for adaptation $\overline{\mathtt{z}}(m^*)$

To be done: An upper bound on $\overline{\mathtt{z}}(m^*)$

4.3.6 Choice of β and $\mathtt{x}$

The value β enters in the procedure, so it has to be selected by some rule. This value shows up in the definition of the oracle choice and it controls the bias-variance relation. A standard risk minimization suggests to take $\beta = 1$. However, we prefer to keep the option of selecting β open in dependence on the type of applications. An increase of β results in the corresponding increase of the critical values $\mathtt{z}_{m,m^\circ}$, the acceptance rule becomes more conservative. Therefore, the selected model $\widehat{m}$ is smaller for β large and is bigger for β small. One can say, big values of β result in oversmoothing. In general, if the underlying model $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ is specified correctly, we recommend to take β small. Numerical results confirm this suggestion and they are best for $\beta \equiv 0$. However, if we aim at building a procedure which is robust against model misspecification then larger values of β can be recommended. This especially concerns the noise misspecification: if the real noise vector $\boldsymbol{\varepsilon}$ is not homogeneous independent Gaussian and if its covariance structure is not precisely known, one can compensate these issues by choosing a positive β, say $\beta = 1$.

Another tuning parameter entering in the procedure is the quantile level $\mathtt{x}$. As one can guess from the beginning, this value is not important for the method. A default choice $\mathtt{x} = 3$ does a good job in all conducted numerical results.

4.3.7 Power loss function

The probabilistic oracle bound of Theorem 4.3.2 provides some statement about typical behavior of the adaptive SmA estimate $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{m}}$. Unfortunately, this bound does not yield a risk bound for quadratic or polynomial losses: even if big loss occurs with a small probability, the related risk can be large. It is interesting that the SmA procedure can be easily tuned to secure an oracle risk bound.

For simplicity of notation, we only consider the quadratic risk

$$\mathcal{R}(\widehat{\boldsymbol{\theta}}) \stackrel{\text{def}}{=} I\!\!E \|W(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2.$$

We aim at comparing the risk of the SmA procedure with the risk $\mathcal{R}_{m^*}$ of the oracle estimate $\widetilde{\boldsymbol{\theta}}_{m^*}$. Remind the representation

$$W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*) = \boldsymbol{\xi}_m + \boldsymbol{b}_m$$

with $\boldsymbol{\xi}_m = W\mathcal{S}_m\boldsymbol{\varepsilon}$ and $\boldsymbol{b}_m = W(\mathcal{S}_m - \mathcal{S})\boldsymbol{f}^*$ yielding the risk decomposition

$$\mathcal{R}_m \stackrel{\text{def}}{=} I\!\!E\big\|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\big\|^2 = I\!\!E\|\boldsymbol{\xi}_m\|^2 + \|\boldsymbol{b}_m\|^2 = \mathtt{p}_m + \|\boldsymbol{b}_m\|^2$$

with $\mathtt{p}_m = \operatorname{tr}\big(\mathbb{V}_m\big)$ and $\mathbb{V}_m = \operatorname{Var}(\boldsymbol{\xi}_m)$. A similar decomposition holds for the quadratic risk of the difference $\widetilde{\boldsymbol{\theta}}_{m'} - \widetilde{\boldsymbol{\theta}}_m$ for any pair $m' > m$:

$$I\!\!E\big\|W(\widetilde{\boldsymbol{\theta}}_{m'} - \widetilde{\boldsymbol{\theta}}_m)\big\|^2 = I\!\!E\|\boldsymbol{\xi}_{m',m}\|^2 + \|\boldsymbol{b}_{m',m}\|^2 = \mathtt{p}_{m',m} + \|\boldsymbol{b}_{m',m}\|^2.$$

Usually the oracle choice is defined by minimizing the risk $\mathcal{R}_m$. For our analysis, we have to slightly modify the definition in the spirit of SmA oracle (4.19). To ensure a risk bound, it is required that not only the model m^* is "good" but also all the larger models $m > m^*$ are "good" as well.

$$m^* \stackrel{\text{def}}{=} \min\Big\{m^\circ\colon \max_{m' > m \ge m^\circ} \big\{\|\boldsymbol{b}_{m',m}\|^2 - \beta^2\,\mathtt{p}_{m',m}\big\} \le 0\Big\}. \tag{4.28}$$

Below we also suppose that the bias component $\|\boldsymbol{b}_m\|^2$ fulfills

$$\|\boldsymbol{b}_m\| \le \|\boldsymbol{b}_{m^*}\|, \qquad m > m^*. \tag{4.29}$$

Otherwise, one can define $\|\boldsymbol{b}_{m^*}\| \stackrel{\text{def}}{=} \max_{m \ge m^*} \|\boldsymbol{b}_m\|$.

Below we consider the SmA procedure with the critical values $\mathtt{z}_{m',m}$ and explain how this values can be adjusted to ensure a risk bound. For stating the result, we need a bit more detailed analysis of the SmA procedure in the propagation zone $m > m^*$. If such m is selected then $\widehat{\boldsymbol{\theta}}$ coincides with $\widetilde{\boldsymbol{\theta}}_m$ whose loss and risk can be much larger than ones for the oracle estimate $\widetilde{\boldsymbol{\theta}}_{m^*}$ because of its larger complexity. There is no essential bias component in this zone. Therefore, the SmA procedure can be tuned in the situation when there is no signal and hence no bias at all:

$$\mathbb{T}_{m',m} = \big\|W(\widetilde{\boldsymbol{\theta}}_{m'} - \widetilde{\boldsymbol{\theta}}_m)\big\| = \|\boldsymbol{\xi}_{m',m}\|\,.$$

The analysis is based on a simple but important observation that if $\widehat{m} = m > m^*$, then the good model $m^\circ = m - 1$ is rejected. The latter means that at least one check based on $\mathbb{T}_{m',m-1}$ fails. The same can be expressed as follows: the maximum of the r.v.'s $\mathbb{T}_{m',m-1}\,\mathbb{1}\!\!\mathrm{I}(\mathbb{T}_{m',m-1} > z_{m',m-1})$ is positive. Let $z(\mathbb{V}_{m',m-1}, \mathtt{x})$ be the tail function of the $\|\boldsymbol{\xi}_{m',m-1}\|$ for $\boldsymbol{\xi}_{m',m-1} \sim \mathcal{N}(0, \mathbb{V}_{m',m-1})$. For each m, we consider the expectation of

$\mathtt{p}_m^{-1}\|\boldsymbol{\xi}_m\|^2$ on the random set $\Omega_m(\mathtt{x})$ on which at least one of test statistics $\mathbb{T}_{m',m-1} = \|\boldsymbol{\xi}_{m',m-1}\|$ exceeds the critical value $z(\mathbb{V}_{m',m-1},\mathtt{x})$:

$$\mathcal{R}_m^+(\mathtt{x}) \stackrel{\text{def}}{=} I\!\!E\Big[(\mathtt{p}_m^{-1}\|\boldsymbol{\xi}_m\|^2 \vee 1)\, \mathbb{1}\Big(\max_{m'\ge m}\big\{\|\boldsymbol{\xi}_{m',m-1}\| - z(\mathbb{V}_{m',m-1},\mathtt{x})\big\} > 0\Big)\Big].$$

Similarly one can consider any other power loss function by replacing $(\mathtt{p}_m^{-1/2}\|\boldsymbol{\xi}_m\|)^2$ with $(\mathtt{p}_m^{-1/2}\|\boldsymbol{\xi}_m\|)^q$. In particular, $q=0$ yields the probability loss considered before.

Now we define the value $\mathtt{x}_{m-1}$ in a way to control the related deviation risk $\mathcal{R}_m^+(\mathtt{x})$. Let α_m be a given decreasing sequence. Its choice will be discussed below. We fix each value $\mathtt{x}_{m-1}$ such that

$$\mathcal{R}_m^+(\mathtt{x}_{m-1}) = \alpha_m\,. \tag{4.30}$$

It implies

$$\begin{aligned} I\!\!E\Big[\|\boldsymbol{\xi}_m\|^2\, \mathbb{1}\Big(\max_{m'\ge m}\big\{\|\boldsymbol{\xi}_{m',m-1}\| - z(\mathbb{V}_{m',m-1},\mathtt{x}_{m-1})\big\} > 0\Big)\Big] &\le \alpha_m \mathtt{p}_m\,, \\ I\!\!P\Big(\max_{m'\ge m}\big\{\|\boldsymbol{\xi}_{m',m-1}\| - z(\mathbb{V}_{m',m-1},\mathtt{x}_{m-1})\big\} > 0\Big) &\le \alpha_m\,. \end{aligned} \tag{4.31}$$

Now define the critical values $\mathtt{z}_{m,m^\circ}$ of the SmA procedure as

$$\mathtt{z}_{m,m^\circ} = z(\mathbb{V}_{m,m^\circ},\mathtt{x}_m) + \beta\mathtt{p}_{m,m^\circ}^{1/2}, \tag{4.32}$$

that is,

$$\widehat{m} = \min\Big\{m^\circ\colon \max_{m>m^\circ}\big\{\mathbb{T}_{m,m^\circ} - \mathtt{z}_{m,m^\circ}\big\} \le 0\Big\}. \tag{4.33}$$

It is worth mentioning that the procedure is the same, and even the critical values $\mathtt{z}_{m,m^\circ}$ are given by the same formula, as in the case of probabilistic loss. The only difference is in the propagation condition (4.30) which is a bit stronger than a similar condition for indicator loss. This implies that the values $\mathtt{x}_m$ and $\mathtt{z}_{m,m^\circ}$ are a bit larger.

Theorem 4.3.3. *Let the SmA procedure (4.33) be applied with the critical values z_{m,m° from (4.32) and the values $\mathtt{x}_m$ defined by (4.30) with the coefficients α_m satisfying*

$$\sum_{m>m^*} \alpha_m \mathtt{p}_m \le \overline{\alpha}_{m^*}\mathtt{p}_{m^*} \tag{4.34}$$

for some $\overline{\alpha}_{m^*}$. Then

$$I\!\!E\big\|W(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\big\|^2 \le 2\overline{\alpha}_{m^*}\mathcal{R}_{m^*} + \big(\mathcal{R}_{m^*}^{1/2} + \overline{\mathtt{z}}_{m^*}\big)^2$$

with

$$\overline{\mathtt{z}}_{m^*} \stackrel{\text{def}}{=} \max_{m<m^*} \mathtt{z}_{m^*,m}\,.$$

Proof. Let us fix $m > m^*$ and $m' \geq m$. The definition (4.28) of the oracle m^* and the formula (4.32) for the critical value $\mathtt{z}_{m',m-1}$ implies for the test statistic $\mathbb{T}_{m',m-1} = \|\boldsymbol{\xi}_{m',m-1} + \boldsymbol{b}_{m',m-1}\|$

$$\bigl\{\mathbb{T}_{m',m-1} > \mathtt{z}_{m',m-1}\bigr\} \subseteq \bigl\{\|\boldsymbol{\xi}_{m',m-1}\| > z(\mathbb{V}_{m',m-1}, \mathtt{x}_{m-1})\bigr\}.$$

Now we can bound the risk of $\widehat{\boldsymbol{\theta}}$ on the set $\widehat{m} > m^*$. We use that for $\widehat{m} = m > m^*$ in view of (4.29)

$$\begin{aligned}\|W(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 &= \|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2 = \|\boldsymbol{\xi}_m + \boldsymbol{b}_m\|^2 \\ &\leq 2\|\boldsymbol{\xi}_m\|^2 + 2\|\boldsymbol{b}_m\|^2 \leq 2\|\boldsymbol{\xi}_m\|^2 + 2\|\boldsymbol{b}_{m^*}\|^2\end{aligned}$$

and it holds by (4.31) and monotonicity $\mathtt{p}_m > \mathtt{p}_{m^*}$

$$\begin{aligned}&\mathbb{E}\bigl\{\|W(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 \mathbb{1}\bigl(\widehat{m} > m^*\bigr)\bigr\} \\ &\quad\leq 2 \sum_{m>m^*} \mathbb{E}\bigl\{\bigl(\|\boldsymbol{\xi}_m\|^2 + \|\boldsymbol{b}_{m^*}\|^2\bigr)\, \mathbb{1}(\widehat{m} = m)\bigr\} \\ &\quad\leq 2 \sum_{m>m^*} \mathbb{E}\bigl\{\bigl(\|\boldsymbol{\xi}_m\|^2 + \|\boldsymbol{b}_{m^*}\|^2\bigr)\, \mathbb{1}(m-1 \text{ is rejected})\bigr\} \\ &\quad= 2 \sum_{m>m^*} \mathbb{E}\Bigl[\bigl(\|\boldsymbol{\xi}_m\|^2 + \|\boldsymbol{b}_{m^*}\|^2\bigr)\, \mathbb{1}\Bigl(\max_{m' \geq m}\Bigl\{\|\boldsymbol{\xi}_{m',m-1}\| - z(\mathbb{V}_{m',m-1}, \mathtt{x}_m)\Bigr\} > 0\Bigr)\Bigr] \\ &\quad\leq 2 \sum_{m>m^*} \alpha_m(\mathtt{p}_m + \|\boldsymbol{b}_{m^*}\|^2) \leq 2\overline{\alpha}_{m^*}\bigl(\mathtt{p}_{m^*} + \|\boldsymbol{b}_{m^*}\|^2\bigr) = 2\overline{\alpha}_{m^*}\mathcal{R}_{m^*}.\end{aligned}$$

Here we have used that (4.34) and $\mathtt{p}_m > \mathtt{p}_{m^*}$ imply $\sum_{m>m^*} \alpha_m \leq \overline{\alpha}_{m^*}$. In the zone $\widehat{m} = m < m^*$, we can still use the stability property:

$$\|W(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^*})\| \leq \mathtt{z}_{m^*,m}\,;$$

cf. (4.25). We conclude

$$\begin{aligned}\mathbb{E}\bigl\|W(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\bigr\|^2 &\leq 2\overline{\alpha}_{m^*}\mathcal{R}_{m^*} + \mathbb{E}\bigl\{\bigl\|W(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\bigr\|^2 \mathbb{1}(\widehat{m} < m^*)\bigr\} \\ &\leq 2\overline{\alpha}_{m^*}\mathcal{R}_{m^*} + \mathbb{E}\bigl(\|W(\widetilde{\boldsymbol{\theta}}_{m^*} - \boldsymbol{\theta}^*)\| + \overline{\mathtt{z}}_{m^*}\bigr)^2 \\ &\leq 2\overline{\alpha}_{m^*}\mathcal{R}_{m^*} + \bigl(\mathcal{R}_{m^*}^{1/2} + \overline{\mathtt{z}}_{m^*}\bigr)^2\end{aligned}$$

as required.

Similarly to the probabilistic loss function, the result can be refined by considering the zone of insensitivity in the region $m < m^*$.

To be done: Define the zone of insensitivity. An oracle bound refined

The constants α_m have to fulfill (4.34). If $\mathtt{p}_m$ satisfy

$$\sum_{m>m^*} (\mathtt{p}_{m^*}/\mathtt{p}_m)^a \le \mathtt{C}$$

for some $a > 0$ and a fixed constant $\mathtt{C}$, then one can take

$$\alpha_m = \mathtt{p}_m^{-1-a}$$

yielding

$$\sum_{m>m^*} \alpha_m \mathtt{p}_m \le \sum_{m>m^*} \mathtt{p}_m^{-a} \le \mathtt{C}\mathtt{p}_{m^*}^{-a}.$$

This implies for $\mathtt{x}_m$ a bound

To be done: An upper bound on $\mathtt{x}_m$

In a similar way one can bound the payment for adaptation.

To be done: An upper bound on $z^{\flat}_{m^*}$

To be done: Linear vs geometric growth of $\mathtt{p}_m$

4.3.8 Penalized MLE

Now we discuss how the procedure and the results can be specified for the case of penalized maximum likelihood estimation. We consider the regression model $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with a homogeneous errors $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_p$. For simplicity of description, we assume "in the model" case when the true regression $\boldsymbol{f}^* = I\!\!E\boldsymbol{Y}$ indeed follows the linear parametric model $\boldsymbol{f}^* = \Psi^\top \boldsymbol{\theta}^*$ with the target parameter $\boldsymbol{\theta}^*$. Let us given a family of penalized procedures described by the roughness matrices G_m^2:

$$\begin{aligned} \widetilde{\boldsymbol{\theta}}_m &= \operatorname*{argmax}_{\boldsymbol{\theta}} \bigl\{ L(\boldsymbol{\theta}) - \|G_m \boldsymbol{\theta}\|^2/2 \bigr\} \\ &= \operatorname*{argmin}_{\boldsymbol{\theta}} \bigl\{ \|\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}\|^2 + \sigma^2 \|G_m \boldsymbol{\theta}\|^2 \bigr\}. \end{aligned}$$

The solution is given by

$$\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y} = \bigl(\Psi\Psi^\top + \sigma^2 G_m^2\bigr)^{-1} \Psi \boldsymbol{Y}.$$

The penalizing matrices G_m^2 are naturally ordered, the running example is given by $G_m^2 = \lambda_m G^2$ for a fixed matrix G^2 and a decreasing sequence λ_m. Note that small values of the coefficient λ correspond to a minor penalization and thus, to a large model.

For any two values $m^\circ < m$, the difference $\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ}$ can be represented as

$$\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ} = (\mathcal{S}_m - \mathcal{S}_{m^\circ}) \boldsymbol{Y} = \mathcal{S}_{m,m^\circ} \boldsymbol{Y}$$

with

$$\mathcal{S}_{m,m^\circ} = \bigl\{(\Psi\Psi^\top + \sigma^2 G_m^2)^{-1} - (\Psi\Psi^\top + \sigma^2 G_{m^\circ}^2)^{-1}\bigr\}\Psi.$$

We consider three options for the weighting matrix W: the prediction loss with $W = \Psi^\top$, the estimation loss with $W = I_p$, and functional estimation $W\colon \mathbb{R}^p \to \mathbb{R}$.

Prediction loss

In the case of prediction loss, it holds

$$\Psi^\top(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*) = \Pi_m \boldsymbol{Y} - \boldsymbol{f}^* = \boldsymbol{\xi}_m + \boldsymbol{b}_m\,,$$

where

$$\Pi_m \stackrel{\mathrm{def}}{=} \Psi^\top(\Psi\Psi^\top + \sigma^2 G_m^2)^{-1}\Psi$$

is a subprojector in $\mathbb{R}^n$ with $\Pi_m = \Pi_m^\top$ and $\|\Pi_m\| \le 1$. The bias $\boldsymbol{b}_m = \Pi_m \boldsymbol{f}^* - \boldsymbol{f}^*$ and the stochastic component $\boldsymbol{\xi}_m$ are vectors in $\mathbb{R}^n$. The quadratic risk is given by

$$\mathbb{E}\|\Psi^\top(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2 = \sigma^2 \operatorname{tr}(\Pi_m^2) + \|\boldsymbol{b}_m\|^2.$$

Further, for each $m^\circ < m$

$$\Psi^\top\bigl(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ}\bigr) = (\Pi_m - \Pi_{m^\circ})\boldsymbol{Y} = \Pi_{m,m^\circ}\boldsymbol{Y}.$$

The assumption $G_{m^\circ}^2 \ge G_m^2$ implies the monotonicity $\Pi_{m,m^\circ} \ge 0$.

Below we consider the scaled test statistic $\mathbb{T}_{m,m^\circ}$ which reads as

$$\mathbb{T}_{m,m^\circ} = \sigma^{-1}\|\Pi_{m,m^\circ}\boldsymbol{Y}\|.$$

The corresponding covariance matrix is

$$\mathbb{V}_{m,m^\circ} = \operatorname{Var}\bigl(\sigma^{-1}\Pi_{m,m^\circ}\boldsymbol{Y}\bigr) = \Pi_{m,m^\circ}^2.$$

The test statistic and the acceptance rule can be represented as the set of inequalities

$$\sigma^{-1}\|\Pi_{m,m^\circ}\boldsymbol{Y}\| \le z\bigl(\mathbb{V}_{m,m^\circ}, \mathtt{x}_{m^\circ}\bigr) + \beta\sqrt{\operatorname{tr}(\mathbb{V}_{m,m^\circ})}, \qquad m > m^\circ.$$

If the effective dimension $\mathtt{p}_{m,m^\circ} = \operatorname{tr}(\mathbb{V}_{m,m^\circ})$ is larger than a certain constant then

$$z\bigl(\mathbb{V}_{m,m^\circ}, \mathtt{x}_{m^\circ}\bigr) \approx \mathtt{p}_{m,m^\circ}^{1/2} + \mathtt{C}\mathtt{x}_{m^\circ}^{1/2},$$
$$\mathtt{z}_\beta\bigl(\mathbb{V}_{m,m^\circ}, \mathtt{x}_{m^\circ}\bigr) \approx (1+\beta)\mathtt{p}_{m,m^\circ}^{1/2} + \mathtt{C}\mathtt{x}_{m^\circ}^{1/2}$$

Estimation loss

Now we briefly discuss the situation with $W = I_p$. Then

$$\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^* = \mathcal{S}_m \boldsymbol{Y} - \boldsymbol{\theta}^* = \boldsymbol{b}_m + \boldsymbol{\xi}_m$$

where $\boldsymbol{\xi}_m = \mathcal{S}_m \boldsymbol{\varepsilon}$ and $\boldsymbol{b}_m = \mathcal{S}_m \boldsymbol{f}^* - \boldsymbol{\theta}^*$ are vectors in $I\!R^p$. The quadratic risk is given by

$$I\!E \|\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\|^2 = \sigma^2 \operatorname{tr}(\mathcal{S}_m \mathcal{S}_m^\top) + \|\boldsymbol{b}_m\|^2.$$

The assumption $G_{m^\circ}^2 \geq G_m^2$ implies the monotonicity $\mathcal{S}_m \mathcal{S}_m^\top \geq \mathcal{S}_{m^\circ} \mathcal{S}_{m^\circ}^\top$. Further, for each $m^\circ < m$

$$\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ} = (\mathcal{S}_m - \mathcal{S}_{m^\circ}) \boldsymbol{Y} = \mathcal{S}_{m,m^\circ} \boldsymbol{Y}.$$

The scaled test statistic $\mathbb{T}_{m,m^\circ}$ reads as

$$\mathbb{T}_{m,m^\circ} = \sigma^{-1} \|\mathcal{S}_{m,m^\circ} \boldsymbol{Y}\|.$$

The corresponding covariance matrix is

$$\mathbb{V}_{m,m^\circ} = \operatorname{Var}\bigl(\sigma^{-1} \mathcal{S}_{m,m^\circ} \boldsymbol{Y}\bigr) = \mathcal{S}_{m,m^\circ} \, \mathcal{S}_{m,m^\circ}^\top.$$

and the acceptance rule can be represented as the set of bounds

$$\sigma^{-1} \|\mathcal{S}_{m,m^\circ} \boldsymbol{Y}\| \leq z\bigl(\mathbb{V}_{m,m^\circ}, \mathtt{x}_{m^\circ}\bigr) + \beta \sqrt{\operatorname{tr}(\mathbb{V}_{m,m^\circ})}, \qquad m > m^\circ.$$

If the effective dimension $\mathtt{p}_{m,m^\circ} = \operatorname{tr}(\mathbb{V}_{m,m^\circ})$ is larger than a certain constant then

$$\begin{aligned} z\bigl(\mathbb{V}_{m,m^\circ}, \mathtt{x}_{m^\circ}\bigr) &\approx \mathtt{p}_{m,m^\circ}^{1/2} + \mathtt{C}\mathtt{x}_{m^\circ}^{1/2}, \\ \mathtt{z}_\beta\bigl(\mathbb{V}_{m,m^\circ}, \mathtt{x}_{m^\circ}\bigr) &\approx (1+\beta) \mathtt{p}_{m,m^\circ}^{1/2} + \mathtt{C}\mathtt{x}_{m^\circ}^{1/2}. \end{aligned}$$

To be done: make this more precise

Estimation of a linear functional

Now we discuss how the general SmA procedure can be applied to the problem of linear functional estimation in linear regression. We focus on the profile approach when one constructs the estimate $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$ and the functional is described via the weighting matrix W of rank one which can be treated as a vector in $I\!R^p$. The target of estimation is the value $\phi^* = W \boldsymbol{\theta}^*$. It holds for each estimate $\widetilde{\phi}_m = W \widetilde{\boldsymbol{\theta}}_m$

$$\widetilde{\phi}_m - \phi^* = W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*) = W\mathcal{S}_m\boldsymbol{\varepsilon} + W(\mathcal{S}_m\boldsymbol{f}^* - \boldsymbol{\theta}^*) = \xi_m + b_m\,,$$

where $\xi_m = W\mathcal{S}_m\boldsymbol{\varepsilon}$ is a zero mean random variable and $b_m = W(\mathcal{S}_m\boldsymbol{f}^* - \boldsymbol{\theta}^*)$ is the deterministic bias. The squared risk of $\widetilde{\phi}_m$ is given by the usual bias-variance decomposition:

$$\mathcal{R}_m = I\!\!E\big(\widetilde{\phi}_m - \phi^*\big)^2 = I\!\!E\big(\xi_m + b_m\big)^2 = b_m^2 + \mathrm{Var}(\xi_m) = b_m^2 + s_m^2$$

with

$$\mathbb{V}_m = s_m^2 = \sigma^2 W\mathcal{S}_m\mathcal{S}_m^\top W^\top.$$

Monotonicity condition $\mathcal{S}_m\mathcal{S}_m^\top \geq \mathcal{S}_{m^\circ}\mathcal{S}_{m^\circ}^\top$ yields monotonicity $\mathbb{V}_m \geq \mathbb{V}_{m^\circ}$ for the functional estimate. For each pair $m^\circ < m$

$$\widetilde{\phi}_m - \widetilde{\phi}_{m^\circ} = W\big(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ}\big) = W(\mathcal{S}_m - \mathcal{S}_{m^\circ})\boldsymbol{Y} = W\mathcal{S}_{m,m^\circ}\boldsymbol{Y}.$$

The variance of this difference reads as

$$\mathbb{V}_{m,m^\circ} = s_{m,m^\circ}^2 = \mathrm{Var}\big(\xi_{m,m^\circ}\big) = \sigma^2 W\mathcal{S}_{m,m^\circ}\,\mathcal{S}_{m,m^\circ}^\top W^\top.$$

The scaled test statistic $\mathbb{T}_{m,m^\circ}$ is given by

$$\mathbb{T}_{m,m^\circ} = s_{m,m^\circ}^{-1}|W\mathcal{S}_{m,m^\circ}\boldsymbol{Y}|.$$

This is the absolute value of a standard Gaussian random variable and its tail function is $z_1(\mathtt{x})$. The acceptance rule can be represented as the set of bounds

$$\mathbb{T}_{m,m^\circ} \leq z_1(\mathtt{x}_{m^\circ}) + \beta, \qquad m > m^\circ.$$

Here $z_1(\mathtt{x})$ is the $1 - \mathrm{e}^{-\mathtt{x}}$ quantile of the absolute value of a standard normal r.v. ξ

$$I\!\!P\big(|\xi| > z_1(\mathtt{x})\big) = \mathrm{e}^{-\mathtt{x}}$$

and the multiplicity correction $\mathtt{x}_{m^\circ} = \mathtt{x} + q_{m^\circ}(\mathtt{x})$ is defined by the condition

$$I\!\!P\Big(\bigcup_{m>m^\circ}\big\{s_{m,m^\circ}^{-1}|\xi_{m,m^\circ}| \geq z_1(\mathtt{x}_{m^\circ})\big\}\Big) \leq \mathrm{e}^{-\mathtt{x}}. \tag{4.35}$$

We define

$$\widehat{m} = \text{smallest accepted}, \qquad \widehat{\phi} = \widetilde{\phi}_{\widehat{m}} = W\mathcal{S}_{\widehat{m}}\boldsymbol{Y}.$$

The oracle choice m^* corresponds to the smallest index m° for which

$$|b_{m,m^\circ}| \le \beta s_{m,m^\circ}\,, \qquad m > m^\circ.$$

The construction ensures that $I\!P\bigl(\widehat{m} > m^*\bigr) \le \mathrm{e}^{-\mathtt{x}}$ and the probability oracle bound reads as

$$\bigl|\widehat{\phi} - \widetilde{\phi}_{m^*}\bigr| \le \overline{\mathtt{z}}(m^*) = \max_{m \in \mathcal{M}^\circ(\mathtt{x})} \bigl\{ z_1(\mathtt{x}_m) s_{m^*,m} \bigr\}.$$

This is a meaningful bound even in the worst case of a flat risk when $\mathcal{M}^\circ(\mathtt{x})$ coincides with the set of all $m \le m^*$. Indeed, monotonicity condition allows to bound $z_1(\mathtt{x}_m) \le z_1(\mathtt{x}_0)$ and $s_{m^*,m} \le s_{m^*,0} \le s_{m^*}$ with $s_{m^*} = \mathbb{V}_{m^*}^{1/2}$ where zero index corresponds to the smallest model with a trivial estimate. This yields

$$\bigl|\widehat{\phi} - \widetilde{\phi}_{m^*}\bigr| \le z_1(\mathtt{x}_0) s_{m^*}\,.$$

The value $z_1(\mathtt{x}_0)$ here can be easily upper bounded by usual Bonferroni arguments. Indeed, the choice $\mathtt{x}_0 \le \mathtt{x} + \log(\#\mathcal{M})$ yields the uniform bound (4.35) in a straightforward way. This implies $z_1(\mathtt{x}_0) \approx \sqrt{\mathtt{x}_0} \le \sqrt{\mathtt{x} + \log(\#\mathcal{M})}$. However, as we already mentioned, the Bonferroni correction is very conservative. In many situation it can be improved to the bound $\mathtt{x}_0 \le \mathtt{x} + \mathtt{C} \log\log(\#\mathcal{M})$. This upper bound is still pessimistic because it is computed for the full set $\mathcal{M}$. It cannot be improved in a general minimax setup: one can always build a model with a flat risk in which $\mathcal{M}^\circ(\mathtt{x})$ is nearly $\mathcal{M}$. However, for each particular configuration $\boldsymbol{\theta}^*$, the oracle result can be made more precise and accurate by considering the zone of insensitivity and using that $s_{m^*,m}$ is smaller than s_{m^*} for m close to m^*.

4.4 Lepski's method

This section briefly discusses one more method of model selection known as Lepski's method and its acceptance rule is a bit stronger than for the SmA method. Namely, a candidate model m° is accepted if all larger models are accepted and all checks for $\mathbb{T}_{m,m^\circ}$ are fulfilled. This rule can be written as

$$\widehat{m} \stackrel{\text{def}}{=} \min\bigl\{ m^\circ \colon \mathbb{T}_{m',m} \le \mathtt{z}_{m',m,m^\circ} \text{ for all } m' > m \ge m^\circ \bigr\} \tag{4.36}$$

with properly selected critical values $\mathtt{z}_{m',m,m^\circ}$. An important advantage of this procedure is its algorithmic description: one starts with the largest model and performs at each step a check of the next smaller model m°. If one of the checks for $\mathbb{T}_{m,m^\circ}$ fails, that is, if $\mathbb{T}_{m,m^\circ} > z_{m,m^\circ}$ then the procedure terminates and selects the latest previously accepted model. In the contrary, the SmA procedure screens all models. Even if one large model is rejected, it still may happen that a smaller one is accepted.

The critical values z_{m',m,m° can again be fixed by a procedure based on multiplicity correction. Let $z(\mathbb{V},\mathtt{x})$ be the tail function of the norm $\|\boldsymbol{\xi}\|$ with $\boldsymbol{\xi} \sim \mathcal{N}(0,\mathbb{V})$. Similarly to the SmA procedure, define for each $\mathtt{x}$ and m the correction $q_m(\mathtt{x})$ which ensures the probability bound

$$I\!P\Big(\bigcup_{m' \in \mathcal{M}^+(m)} \big\{ \|\boldsymbol{\xi}_{m',m}\| \geq z(\mathbb{V}_{m',m}, \mathtt{x} + q_m(\mathtt{x})) \big\} \Big) = \mathrm{e}^{-\mathtt{x}}. \tag{4.37}$$

Further, with $q_m(\mathtt{x})$ fixed for each m, introduce for each m° an additional correction $q^+_{m^\circ} = q^+_{m^\circ}(\mathtt{x})$ which ensures the uniformity in all $m \geq m^\circ$:

$$I\!P\Big(\bigcup_{m \geq m^\circ} \bigcup_{m' \in \mathcal{M}^+(m)} \big\{ \|\boldsymbol{\xi}_{m',m}\| \geq z\big(\mathbb{V}_{m',m}, \mathtt{x} + q_m(\mathtt{x}) + q^+_{m^\circ}(\mathtt{x})\big) \big\} \Big) = \mathrm{e}^{-\mathtt{x}}. \tag{4.38}$$

Exercise 4.4.1. Check that the conditions (4.37) and (4.38) uniquely define the values $q_m(\mathtt{x})$ and $q^+_m(x)$ for all m and $\mathtt{x}$.

Define

$$\mathtt{x}_{m,m^\circ} \stackrel{\text{def}}{=} \mathtt{x} + q_m(\mathtt{x}) + q^+_{m^\circ}(\mathtt{x})$$

and apply (4.36) with

$$\mathtt{z}_{m',m,m^\circ} \stackrel{\text{def}}{=} z\big(\mathbb{V}_{m',m}, \mathtt{x}_{m,m^\circ}\big) + \beta\sqrt{\mathrm{tr}(\mathbb{V}_{m',m})}.$$

Further, let the oracle choice m^* be defined in a similar way by checking the bias $\boldsymbol{b}_{m',m}$ for all $m' > m \geq m^\circ$:

$$m^* \stackrel{\text{def}}{=} \min\big\{ m^\circ \in \mathcal{M} \colon \|\boldsymbol{b}_{m',m}\|^2 \leq \beta^2 \, \mathrm{tr}\big(\mathbb{V}_{m',m}\big) \;\; \forall m' > m \geq m^\circ \big\}.$$

Now the construction ensures that

$$I\!P\big(m^* \text{ is accepted}\big) = I\!P\big(\widehat{m} \leq m^*\big) \geq 1 - \mathrm{e}^{-\mathtt{x}}.$$

The procedure (4.36) also guarantees that for $\widehat{m} = m < m^*$, it holds

$$\mathbb{T}_{m^*,m} = \|W(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^*})\| \leq z\big(\mathbb{V}_{m^*,m}, \mathtt{x}_{m,m}\big) + \beta\sqrt{\mathrm{tr}(\mathbb{V}_{m^*,m})}.$$

This bound can be used for establishing the oracle inequality similarly to the SmA procedure.

4.5 Intersection of confidence sets (ICI) method

The *intersection of confidence intervals* (ICI) procedure is another method based on multiple comparison. Here one constructs for each m a confidence set $\mathcal{E}_m$ for the target $\boldsymbol{\theta}^*$ based on the estimate $\widetilde{\boldsymbol{\theta}}_m$ and then selects the smallest m° for which the intersection of $\mathcal{E}_m$ for all $m \geq m^\circ$ is not empty.

The construction is again based on the decomposition

$$W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*) = \boldsymbol{\xi}_m + \boldsymbol{b}_m$$

with the stochastic component $\boldsymbol{\xi}_m$ and the bias $\boldsymbol{b}_m$. If we ignore the bias component $\boldsymbol{b}_m$, the use of $\boldsymbol{\xi}_m \sim \mathcal{N}(0, \mathbb{V}_m)$ leads to the confidence set

$$\mathcal{E}_m(z_m) = \bigl\{\boldsymbol{\theta}\colon \|\mathbb{V}_m^{-1/2} W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta})\| \leq \mathtt{z}_m\bigr\}.$$

The width $\mathtt{z}_m$ will be defined a bit later. Now the procedure reads as

$$\widehat{m} \stackrel{\text{def}}{=} \min\biggl\{m^\circ\colon \bigcap_{m \in \mathcal{M}^+(m^\circ)} \mathcal{E}_m(\mathtt{z}_m) \neq \emptyset\biggr\}.$$

A slightly weaker condition is that any two confidence sets $\mathcal{E}_m(z_m)$ and $\mathcal{E}_{m'}(\mathtt{z}_{m'})$ with $m' > m \geq m^\circ$ overlap.

Exercise 4.5.1. Consider the univariate case $q = \operatorname{rank}(W) = 1$. Show that the pairwise check $\mathcal{E}_{m'}(z_{m'}) \cap \mathcal{E}_m(\mathtt{z}_m) \neq \emptyset$ can be rewritten as

$$\bigl|W(\widetilde{\boldsymbol{\theta}}_{m'} - \widetilde{\boldsymbol{\theta}}_m)\bigr| \leq \mathtt{z}_{m'} s_{m'} + \mathtt{z}_m s_m\,, \tag{4.39}$$

where $s_m^2 = \mathbb{V}_m$.

Exercise 4.5.2. For the univariate case $q = 1$, check that the following conditions are equivalent:

- $\bigcap_{m \in \mathcal{M}^+(m^\circ)} \mathcal{E}_m(\mathtt{z}_m) \neq \emptyset$;
- $\mathcal{E}_{m'}(\mathtt{z}_{m'}) \cap \mathcal{E}_m(\mathtt{z}_m) \neq \emptyset$ for all $m' > m \geq m^\circ$.

Exercise 4.5.3. Check for the case $q = 2$ whether the conditions of Exercise 4.5.2 on bivariate confidence sets $\mathcal{E}_m(\mathtt{z}_m)$ are equivalent.

The algorithmic check of the condition on nonempty overlap is quite hard if the target dimension q is larger than one. Below we only consider the case $q = 1$.

The condition (4.39) of an nonempty overlap is similar to the check in the SmA procedure based on $\mathbb{T}_{m',m}$. The SmA check seems to be more accurate because it accounts

for the joint distribution of the two estimates $\widetilde{\boldsymbol{\theta}}_{m'}$ and $\widetilde{\boldsymbol{\theta}}_{m}$: their difference is scaled by its standard deviation. The ICI check normalizes the difference by the weighted sum of standard deviations.

Now we discuss how the ICI procedure can be tuned by fixing the values $\mathtt{z}_m$. The approach stays the same: one selects the "smallest" values which ensure the propagation property. The simplest way is based on a uniform multiplicity correction. The scaled stochastic components $s_m^{-1}\xi_m$ are standard normal and for each the tail function is given by the Laplace function:

$$I\!P\bigl(|s_m^{-1}\xi_m| > z_1(\mathtt{x})\bigr) \le \mathrm{e}^{-\mathtt{x}}.$$

Now one can make an adjustment $q^* = q^*(\mathtt{x})$ to the level $\mathtt{x}$ to ensure a uniform bound:

$$I\!P\Bigl(\bigcup_{m\in\mathcal{M}} \bigl\{|s_m^{-1}\xi_m| > z_1(\mathtt{x}+q^*)\bigr\}\Bigr) \le \mathrm{e}^{-\mathtt{x}}. \tag{4.40}$$

Further we make an additional correction for the bias and define with a fixed $\beta \ge 0$ the uniform confidence widths

$$\mathtt{z}_m \equiv \overline{\mathtt{z}}_\beta(\mathtt{x}) \stackrel{\mathrm{def}}{=} \beta + z_1(\mathtt{x}+q^*). \tag{4.41}$$

The ICI procedure reads now

$$\widehat{m} = \min\bigl\{m^\circ\colon \bigl|W\bigl(\widetilde{\boldsymbol{\theta}}_{m'} - \widetilde{\boldsymbol{\theta}}_{m}\bigr)\bigr| \le (s_{m'} + s_m)\overline{\mathtt{z}}_\beta(\mathtt{x}),\ \forall m' > m \ge m^\circ\bigr\}. \tag{4.42}$$

The oracle choice m^* can be defined as

$$m^* \stackrel{\mathrm{def}}{=} \min\bigl\{m\colon \bigl|b_m\bigr| \le \beta s_m,\ \forall m \ge m^\circ\bigr\}. \tag{4.43}$$

Again and again, the construction is designed to ensure the propagation property

$$I\!P\bigl(m^* \text{ is accepted}\bigr) = I\!P\bigl(\widehat{m} \le m^*\bigr) \ge 1 - \mathrm{e}^{-\mathtt{x}}.$$

This fact is an easy corollary of the definitions (4.42) and (4.43) and of the probability bound (4.40). Also the procedure secures the "stability after propagation" property: if $\widehat{m} = m < m^*$, then by definition of $\widehat{m}$

$$\bigl|W(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^*})\bigr| \le (s_m + s_{m^*})\overline{\mathtt{z}}_\beta(\mathtt{x}).$$

Monotonicity condition ensures that $s_m \le s_{m^*}$ for $m < m^*$, and we end up with the oracle inequality for the adaptive estimate $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{m}}$.

Theorem 4.5.1. *Consider the linear Gaussian model* $\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^{2}\boldsymbol{I}_{n})$ *. Let the weighting matrix* W *be of rank one. Given a set of estimates* $\widetilde{\boldsymbol{\theta}}_{m}$ *,* $m \in \mathcal{M}$ *, with* $\operatorname{Var}\bigl(W\widetilde{\boldsymbol{\theta}}_{m}\bigr) = s_{m}^{2}$ *, define the selector* $\widehat{m}$ *by* (4.42) *for* $\overline{\mathtt{z}}_{\beta}(\mathtt{x})$ *from* (4.41) *and* (4.40)*. The adaptive estimate* $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{m}}$ *and the oracle estimate* $\widetilde{\boldsymbol{\theta}}_{m^{*}}$ *with* m^{*} *from* (4.43) *satisfy*

$$\mathbb{P}\Bigl(|W(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}_{m^{*}})| > 2 s_{m^{*}} \overline{\mathtt{z}}_{\beta}(\mathtt{x})\Bigr) \leq 2\mathrm{e}^{-\mathtt{x}}.$$

To be done: Adaptive vs oracle risk

5

Unordered case. Anisotropic sets and subset selection

The SmA method of the previous section is quite general and can be extended to many statistical models and problem. However, it essentially requires the ordered structure of the set of considered models/methods. This section discusses how the SmA procedure can be extended to some other setups without ordered structure. To distinguish ordered and unordered cases, we denote by $\mathcal{A} = \{\varkappa\}$ the set of all considered models. The basic idea is to assume a kind of partial ordering which enables to define an acceptance rule:

$\varkappa^\circ$ is accepted if it is not rejected against any larger model.

This rule allows to fix a set of accepted models. Further we need some global measure of complexity which can be used for final selection:

$\widehat{\varkappa}$ is the *simplest accepted* model.

Below we illustrate how this method works in two important examples: anisotropic classes and *subset selection* problems.

5.1 Subset selection procedure

Consider a linear model $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$. The dimension p of the vector $\boldsymbol{\theta}$ can be very large and we implicitly assumes a kind of *sparse* structure:

most of $\boldsymbol{\theta}^*$-entries are nearly zero and can be dropped, there is a relatively small subvector of $\boldsymbol{\theta}^*$ containing the important features.

The aim is to find this subset and to estimate the whole vector $\boldsymbol{\theta}^*$. As in the ordered case, one can separate between prediction $W = \Psi^\top$ and estimation $W = \boldsymbol{I}_p$ loss. Of course, these two problems coincide in the sequence space model with $n = p$ and $\Psi = \boldsymbol{I}_p$.

5.1.1 SmA procedure and multilevel synchronization

Let $\varkappa$ mean a subset of the whole index set $\{1, 2, \ldots, p\}$. We use the obvious notation $\varkappa \vee \varkappa'$ for the union, $\varkappa \wedge \varkappa^*$ for the overlap of two subsets $\varkappa$ and $\varkappa'$, $\varkappa' - \varkappa$ for the complement of $\varkappa$ within $\varkappa'$. Further we consider the usual partial ordering: $\varkappa' > \varkappa'$ means that $\varkappa \subseteq \varkappa'$. The subset $\varkappa$ is good if there is no significant bias in its complement $\varkappa^c$. The approach is to design a procedure which rejects any such good model with a small probability. The proposed SmA rule will be again to select the smallest (in complexity) accepted model.

The acceptance rule is based on pairwise comparison with a family of tests $\mathbb{T}_{\varkappa,\varkappa^\circ}$ for $\varkappa > \varkappa^\circ$. The model-candidate $\varkappa^\circ$ is accepted if no of $\mathbb{T}_{\varkappa,\varkappa^\circ}$ rejects the hypothesis of "no bias". Given the loss matrix W, the test statistic $\mathbb{T}_{\varkappa,\varkappa^\circ}$ reads as in the ordered case:

$$\mathbb{T}_{\varkappa,\varkappa^\circ} = \sigma^{-1}\|W(\widetilde{\boldsymbol{\theta}}_\varkappa - \widetilde{\boldsymbol{\theta}}_{\varkappa^\circ})\|. \tag{5.1}$$

The acceptance rule can be written as

$$\varkappa^\circ \text{ is accepted iff} \quad \mathbb{T}_{\varkappa,\varkappa^\circ} \le \mathtt{z}_{\varkappa,\varkappa^\circ} \quad \forall \varkappa > \varkappa^\circ. \tag{5.2}$$

Now we discuss how the critical values $\mathtt{z}_{\varkappa,\varkappa^\circ}$ can be fixed by *synchronization* (*multiplicity correction*) of the individual *tail functions*. We use the decomposition of the test statistic $\mathbb{T}_{\varkappa,\varkappa^\circ}$ from (5.1)

$$\mathbb{T}_{\varkappa,\varkappa^\circ} = \|\boldsymbol{\xi}_{\varkappa,\varkappa^\circ} + \boldsymbol{b}_{\varkappa,\varkappa^\circ}\|.$$

Define

$$\mathtt{p}_{\varkappa,\varkappa^\circ} = \mathrm{tr}\bigl\{\mathrm{Var}\bigl(\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}\bigr)\bigr\}.$$

Suppose we are given for each pair $\varkappa > \varkappa^\circ$ a tail function $z_{\varkappa,\varkappa^\circ}(\mathtt{x})$ of the noise component $\|\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}\|$ providing

$$I\!P\bigl(\|\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}\| > z_{\varkappa,\varkappa^\circ}(\mathtt{x})\bigr) \le \mathrm{e}^{-\mathtt{x}}.$$

This tail function can be used for testing the hypothesis of no significant bias component in the test statistic $\mathbb{T}_{\varkappa,\varkappa^\circ}$. The model-candidate $\varkappa^\circ$ is accepted by the SmA method if all such tests for $\varkappa > \varkappa^\circ$ do. To keep the overall test level, we have to synchronize all performed $\mathbb{T}_{\varkappa,\varkappa^\circ}$-based tests by correcting for multiple check. The simplest way of multiplicity correction is by a uniform increase of the level $\mathtt{x}$ to control the overall rejecting probability: define $q_{\varkappa^\circ}(\mathtt{x})$ by the condition

$$I\!P\Big(\bigcup_{\varkappa \in \mathcal{M}(\varkappa^\circ)} \Big\{ \|\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}\| > z_{\varkappa,\varkappa^\circ}\big(\mathtt{x} + q_{\varkappa^\circ}(\mathtt{x})\big) \Big\} \Big) \le \mathrm{e}^{-\mathtt{x}}.$$

Denote $\mathtt{x}_{\varkappa^\circ} = \mathtt{x} + q_{\varkappa^\circ}(\mathtt{x})$ and apply the acceptance rule (5.2) with $\mathtt{z}_{\varkappa,\varkappa^\circ}$ equal to such defined $z_{\varkappa,\varkappa^\circ}(\mathtt{x}_{\varkappa^\circ})$ after a small bias correction:

$$\mathtt{z}_{\varkappa,\varkappa^\circ} \stackrel{\mathrm{def}}{=} z_{\varkappa,\varkappa^\circ}(\mathtt{x}_{\varkappa^\circ}) + \beta \mathtt{p}^{1/2}_{\varkappa,\varkappa^\circ}.$$

One can use a more sophisticated *multilevel synchronization* procedure which accounts for the complexity of the alternative model $\varkappa$. Let $|\varkappa^\circ|$ mean the cardinality (complexity) of $\varkappa^\circ$. For each $\tau > 0$, denote by $\mathcal{M}_\tau(\varkappa^\circ)$ the set of model $\varkappa > \varkappa^\circ$ with $|\varkappa| \le \tau + |\varkappa^\circ|$. For a given growing sequence $\tau_1 < \tau_2 < \ldots < \tau_{\mathcal{K}}$, we write $\mathcal{M}_k(\varkappa^\circ)$ instead of $\mathcal{M}_{\tau_k}(\varkappa^\circ)$. Obviously $\mathcal{M}_k(\varkappa^\circ)$ grow to the set of models $\mathcal{M}(\varkappa^\circ)$ containing $\varkappa^\circ$. Now one can define the correction step by step: first we fix the correction $q_{1,\varkappa^\circ} = q_{1,\varkappa^\circ}(\mathtt{x})$ for all $\varkappa \in \mathcal{M}_1(\varkappa^\circ)$

$$I\!P\Big(\bigcup_{\varkappa \in \mathcal{M}_1(\varkappa^\circ)} \Big\{ \|\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}\| > z_{\varkappa,\varkappa^\circ}\big(\mathtt{x} + q_{1,\varkappa^\circ}\big) \Big\} \Big) \le \mathrm{e}^{-\mathtt{x}}.$$

With such defined $q_{1,\varkappa^\circ}$, define $q_{2,\varkappa^\circ} = q_{2,\varkappa^\circ}(\mathtt{x})$ such that

$$I\!P\Big(\bigcup_{\varkappa \in \mathcal{M}_2(\varkappa^\circ)} \Big\{ \|\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}\| > z_{\varkappa,\varkappa^\circ}(\mathtt{x} + q_{1,\varkappa^\circ} + q_{2,\varkappa^\circ}) \Big\} \Big) \le \mathrm{e}^{-\mathtt{x}}.$$

We continue this way and define by induction: if $q_{1,\varkappa^\circ}, \ldots, q_{k-1,\varkappa^\circ}$ are fixed then the correction for the set $\mathcal{M}_k(\varkappa^\circ)$ is selected as the sum $\mathtt{x} + q_{1,\varkappa^\circ} + \ldots + q_{k,\varkappa^\circ}$

$$I\!P\Big(\bigcup_{\varkappa \in \mathcal{M}_k(\varkappa^\circ)} \Big\{ \|\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}\| > z_{\varkappa,\varkappa^\circ}(\mathtt{x} + q_{1,\varkappa^\circ} + q_{2,\varkappa^\circ} + \ldots + q_{k,\varkappa^\circ}) \Big\} \Big) \le \mathrm{e}^{-\mathtt{x}}. \tag{5.3}$$

For each $\varkappa > \varkappa^\circ$, there exists a unique $k = k(\varkappa)$ corresponding to the smallest set $\mathcal{M}_k(\varkappa^\circ)$ containing $\varkappa$. Finally, we define

$$\mathtt{z}_{\varkappa,\varkappa^\circ} = z_{\varkappa,\varkappa^\circ}(\mathtt{x} + q_{1,\varkappa^\circ} + q_{2,\varkappa^\circ} + \ldots + q_{k,\varkappa^\circ}) + \beta \mathtt{p}^{1/2}_{\varkappa,\varkappa^\circ}, \quad k = k(\varkappa). \tag{5.4}$$

Our selection rule chooses the smallest accepted model. It can be written as

$$\widehat{\varkappa} = \operatorname*{argmin}_{\varkappa^\circ \in \mathcal{M}} \big\{ |\varkappa^\circ| \colon \mathbb{T}_{\varkappa,\varkappa^\circ} \le \mathtt{z}_{\varkappa,\varkappa^\circ}, \quad \varkappa \in \mathcal{M}(\varkappa^\circ) \big\}. \tag{5.5}$$

If there are many such $\varkappa^\circ$, one can select arbitrarily among them. Now we define a good choice $\varkappa^\circ$ as previously by "no significant bias" condition:

$$\|\boldsymbol{b}_{\varkappa,\varkappa^\circ}\| \le \beta \mathtt{p}^{1/2}_{\varkappa,\varkappa^\circ} \quad \varkappa \in \mathcal{M}(\varkappa^\circ). \tag{5.6}$$

The construction ensures that a good model will be accepted with a high probability.

Theorem 5.1.1. *Let* $\varkappa^\circ$ *be a good model in the sense* (5.6)*. Then it holds for the SmA procedure with the critical values* $\mathtt{z}_{\varkappa,\varkappa^\circ}$ *from* (5.3) *and* (5.4)

$$I\!P\big(\varkappa^\circ \text{ is rejected}\big) \le \mathrm{e}^{-\mathtt{x}}.$$

Now we define the oracle choice $\varkappa^*$ as the smallest (in complexity $|\varkappa^*|$) model under the constraint (5.6):

$$\varkappa^* \stackrel{\text{def}}{=} \operatorname*{argmin}_{\varkappa^\circ \in \mathcal{M}} \big\{ |\varkappa^\circ| \colon \|\boldsymbol{b}_{\varkappa,\varkappa^\circ}\| \le \beta \mathtt{p}^{1/2}_{\varkappa,\varkappa^\circ}, \quad \varkappa \in \mathcal{M}(\varkappa^\circ) \big\}. \tag{5.7}$$

Again, this relation does not uniquely define the $\varkappa^*$ value, if there are many $\varkappa^*$ with this property, any of them can be taken. The oracle bound compares the risk of the oracle estimate $\mathcal{R}_{\varkappa^*}$ with the risk of the adaptive estimate $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}}$ for the SmA rule $\widehat{\varkappa}$.

Theorem 5.1.2. *It holds on a random set of probability at least* $1 - \mathrm{e}^{-\mathtt{x}}$

$$\sigma^{-1}\|W(\widetilde{\boldsymbol{\theta}}_{\varkappa^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}})\| \le \overline{\mathtt{z}}_{\varkappa^*}$$

with $k^* = |\varkappa^*|$ *and* $\overline{\mathtt{z}}_{\varkappa^*}$ *defined by*

$$\overline{\mathtt{z}}_{\varkappa^*} \stackrel{\text{def}}{=} \max_{|\varkappa| \le k^*} \big(\mathtt{z}_{\varkappa \vee \varkappa^*, \varkappa^*} + \mathtt{z}_{\varkappa \vee \varkappa^*, \varkappa} \big).$$

Proof. The construction and the propagation property ensures that $\varkappa^*$ is accepted with a high probability $1 - \mathrm{e}^{-\mathtt{x}}$. Below we focus on this case. Then the adaptive choice $\widehat{\varkappa}$ from (5.5) has to fulfill

$$|\widehat{\varkappa}| \le |\varkappa^*|.$$

Due to partial ordering, the models $\varkappa^*$ and $\widehat{\varkappa}$ are not directly comparable. We use the model $\breve{\varkappa} = \varkappa^* \vee \widehat{\varkappa}$ which contains both and is the smallest one with this property. As $\varkappa^*$ and $\widehat{\varkappa}$ are both accepted, it holds

$$\sigma^{-1}\|W(\widetilde{\boldsymbol{\theta}}_{\breve{\varkappa}} - \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}})\| \le \mathtt{z}_{\breve{\varkappa},\widehat{\varkappa}}, \quad \sigma^{-1}\|W(\widetilde{\boldsymbol{\theta}}_{\breve{\varkappa}} - \widetilde{\boldsymbol{\theta}}_{\varkappa^*})\| \le \mathtt{z}_{\breve{\varkappa},\varkappa^*}\,.$$

We conclude that

$$\sigma^{-1}\|W(\widetilde{\boldsymbol{\theta}}_{\varkappa^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}})\| \le \mathtt{z}_{\breve{\varkappa},\widehat{\varkappa}} + \mathtt{z}_{\breve{\varkappa},\varkappa^*}.$$

and the assertion follows.

5.1.2 Prediction loss

The case of prediction loss ($W = \Psi^\top$) in combination with projection estimates $\widetilde{\boldsymbol{\theta}}_\varkappa$ allows to reduce the study to the sequence space model with $\Psi = \boldsymbol{I}_p$. This dramatically simplifies the situation. Below we denote by $\Pi_\varkappa$ the projector in $\mathbb{R}^n$ onto the subspace corresponding to $\varkappa$. For a couple $\varkappa^\circ < \varkappa$, we also consider the projector $\Pi_{\varkappa,\varkappa^\circ} = \Pi_\varkappa - \Pi_{\varkappa^\circ} = \Pi_{\varkappa-\varkappa^\circ}$.

We also use that $\mathtt{p}_\varkappa = \varkappa$ and

$$\mathtt{p}_{\varkappa,\varkappa^\circ} = \mathbb{E}\|\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}\|^2 = |\varkappa - \varkappa^\circ|.$$

Theorem 5.1.3. *For the regression model $\boldsymbol{Y} = \Psi^\top\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with homogeneous errors and for $W = \Psi^\top$, the tail functions $z_{\varkappa,\varkappa^\circ}(\mathtt{x})$ only depends on $|\varkappa - \varkappa^\circ|$. Moreover, for $\mathtt{x}$ fixed, the multiplicity corrections $q_{k,\varkappa^\circ} = q_{k,\varkappa^\circ}(\mathtt{x})$ only depends on $p - |\varkappa^\circ|$ and on τ_k .*

Proof. We use that $\Psi^\top\widetilde{\boldsymbol{\theta}}_\varkappa = \Pi_\varkappa\boldsymbol{Y}$ and

$$\mathbb{T}^2_{\varkappa,\varkappa^\circ} = \sigma^{-2}\|\Psi^\top(\widetilde{\boldsymbol{\theta}}_\varkappa - \widetilde{\boldsymbol{\theta}}_{\varkappa^\circ})\|^2 = \sigma^{-2}\|\Pi_{\varkappa,\varkappa^\circ}\boldsymbol{Y}\|^2 \sim \chi^2_{|\varkappa-\varkappa^\circ|}\,.$$

To be done: complete the proof

The definition (5.7) of the oracle choice $\varkappa^*$ can be restated as

$$\varkappa^* \stackrel{\text{def}}{=} \operatorname*{argmin}_{\varkappa^\circ\in\mathcal{M}}\bigl\{|\varkappa^\circ|\colon \|\boldsymbol{b}_{\varkappa,\varkappa^\circ}\| \le \beta|\varkappa - \varkappa^\circ|^{1/2}, \quad \varkappa \in \mathcal{M}(\varkappa^\circ)\bigr\}. \tag{5.8}$$

Theorem 5.1.4. *The result of Theorem 5.1.2 holds with $\overline{\mathtt{z}}_{\varkappa^*} = \overline{\mathtt{z}}_{k^*}$ only depending on the cardinality $k^* = |\varkappa^*|$. Any $\varkappa^*$ with this cardinality can be used in definition (5.8).*

To be done: An upper bound on $q_{k,\varkappa^\circ}$

To be done: An upper bound on $\overline{\mathtt{z}}_{k^*}$

To be done: Algorithmic implementation

5.1.3 Estimation loss

The analysis for the problem of estimation loss $W = \boldsymbol{I}_p$ is similar, but some nice features of the prediction loss do not apply here. We use

$$\widetilde{\boldsymbol{\theta}}_\varkappa = \mathcal{S}_\varkappa\boldsymbol{Y},$$
$$\mathcal{S}_\varkappa = \bigl(\Psi_\varkappa\Psi_\varkappa^\top\bigr)^{-1}\Psi_\varkappa.$$

Therefore,

$$\mathbb{T}_{\varkappa,\varkappa^\circ} = \sigma^{-1}\|(\mathcal{S}_\varkappa - \mathcal{S}_{\varkappa^\circ})\boldsymbol{Y}\| = \|\boldsymbol{\xi}_{\varkappa,\varkappa^\circ} + \boldsymbol{b}_{\varkappa,\varkappa^\circ}\|.$$

If the bias component vanishes, the distribution of this test statistic is completely described by the matrices $\Psi_\varkappa$ and $\Psi_{\varkappa^\circ}$ but in a more complicated way than for the prediction loss. Also define

$$\mathtt{p}_{\varkappa,\varkappa^\circ} = I\!\!E\|\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}\|^2 = \operatorname{tr}\bigl\{\bigl(\mathcal{S}_\varkappa - \mathcal{S}_{\varkappa^\circ}\bigr)\bigl(\mathcal{S}_\varkappa - \mathcal{S}_{\varkappa^\circ}\bigr)^\top\bigr\}.$$

The general results of Theorems 5.1.1 and 5.1.2 apply here for such defined values $\mathtt{p}_{\varkappa,\varkappa^\circ}$. Unfortunately, the nice simplification of the formulation as in the prediction case is only possible if the design is orthonormal. Then the estimation and prediction problems coincide.

5.1.4 Linear functional estimation

Now we briefly discuss the case when W is a matrix of rank one which corresponds to estimation of a linear functional. An advantage of this situation is that each difference $W(\widetilde{\boldsymbol{\theta}}_{\varkappa'} - \widetilde{\boldsymbol{\theta}}_\varkappa)$ is univariate normal. This helps a lot in evaluating the distribution of each test statistic $\mathbb{T}_{\varkappa',\varkappa}$.

As in the ordered case, each estimate $\widetilde{\phi}_\varkappa = W\widetilde{\boldsymbol{\theta}}_\varkappa$ fulfills

$$\widetilde{\phi}_\varkappa - \phi^* = W(\widetilde{\boldsymbol{\theta}}_\varkappa - \boldsymbol{\theta}^*) = W\mathcal{S}_\varkappa\boldsymbol{\varepsilon} + W(\mathcal{S}_\varkappa\boldsymbol{f}^* - \boldsymbol{\theta}^*) = \xi_\varkappa + b_\varkappa\,,$$

where $\xi_\varkappa = W\mathcal{S}_\varkappa\boldsymbol{\varepsilon}$ is a zero mean random variable and $b_\varkappa = W(\mathcal{S}_\varkappa\boldsymbol{f}^* - \boldsymbol{\theta}^*)$ is the deterministic bias. The squared risk of $\widetilde{\phi}_\varkappa$ is given by the usual bias-variance decomposition:

$$\mathcal{R}_\varkappa = I\!\!E\bigl(\widetilde{\phi}_\varkappa - \phi^*\bigr)^2 = I\!\!E\bigl(\xi_\varkappa + b_\varkappa\bigr)^2 = b_\varkappa^2 + \operatorname{Var}(\xi_\varkappa) = b_\varkappa^2 + s_\varkappa^2$$

with

$$s_\varkappa^2 = \sigma^2 W\mathcal{S}_\varkappa\mathcal{S}_\varkappa^\top W^\top.$$

Monotonicity condition $\mathcal{S}_\varkappa\mathcal{S}_\varkappa^\top \geq \mathcal{S}_{\varkappa^\circ}\mathcal{S}_{\varkappa^\circ}^\top$ for $\varkappa > \varkappa^\circ$ yields monotonicity $s_\varkappa \geq s_{\varkappa^\circ}$ for the functional estimate. For each pair $\varkappa^\circ < \varkappa$

$$\widetilde{\phi}_\varkappa - \widetilde{\phi}_{\varkappa^\circ} = W\bigl(\widetilde{\boldsymbol{\theta}}_\varkappa - \widetilde{\boldsymbol{\theta}}_{\varkappa^\circ}\bigr) = W(\mathcal{S}_\varkappa - \mathcal{S}_{\varkappa^\circ})\boldsymbol{Y} = W\mathcal{S}_{\varkappa,\varkappa^\circ}\boldsymbol{Y}.$$

The variance of this difference reads as

$$s_{\varkappa,\varkappa^\circ}^2 = \operatorname{Var}\bigl(\xi_{\varkappa,\varkappa^\circ}\bigr) = \sigma^2 W\mathcal{S}_{\varkappa,\varkappa^\circ}\,\mathcal{S}_{\varkappa,\varkappa^\circ}^\top W^\top.$$

The scaled test statistic $\mathbb{T}_{\varkappa,\varkappa^\circ}$ is given for $\varkappa > \varkappa^\circ$ by

$$\mathbb{T}_{\varkappa,\varkappa^\circ} = s^{-1}_{\varkappa,\varkappa^\circ} |W \mathcal{S}_{\varkappa,\varkappa^\circ} \boldsymbol{Y}|.$$

One can use that the stochastic component $\xi_{\varkappa,\varkappa^\circ}$ of $s^{-1}_{\varkappa,\varkappa^\circ} W \mathcal{S}_{\varkappa,\varkappa^\circ} \boldsymbol{Y}$ is standard normal. Thus, the multiplicity corrections are computed from the same bound (5.3) with $z_1(\cdot)$ in place of $z_{\varkappa,\varkappa^\circ}(\cdot)$. Moreover, $\mathtt{p}_{\varkappa,\varkappa^\circ} \equiv 1$, and the oracle definition reads as

$$\varkappa^* \stackrel{\text{def}}{=} \operatorname*{argmin}_{\varkappa^\circ \in \mathcal{M}} \bigl\{ |\varkappa^\circ| \colon \|\boldsymbol{b}_{\varkappa,\varkappa^\circ}\| \le \beta, \quad \varkappa \in \mathcal{M}(\varkappa^\circ) \bigr\}.$$

The general results of Theorems 5.1.1 and 5.1.2 apply here without any change. However, the involved values can be made more precise.

To be done: An upper bound on $q_{k,\varkappa^\circ}$

To be done: An upper bound on $\overline{\mathtt{z}}_{k^*}$

To be done: Algorithmic implementation

5.1.5 Subset selection problem

The presented oracle bound of Theorem 5.1.2 claims that the risk of the adaptive estimate $\widehat{\boldsymbol{\theta}}$ is linked to the risk of the oracle $\widetilde{\boldsymbol{\theta}}_{\varkappa^*}$. However, it tells nothing about the selected set $\widehat{\varkappa}$. Now we discuss this issue of choosing the active set of important features represented by non-zero entries of $\boldsymbol{\theta}^*$.

Note that this problem has to be put in a right way: one can suppose that $\boldsymbol{\theta}^*$ is sparse and only significant entries are non-zero. Alternatively, one tries to find an approximating model with another vector $\boldsymbol{\theta}^\star$ having a sparse representation and delivering nearly the same approximation and prediction quality. We follow our oracle result and define $\varkappa^*$ by (5.7). This will be our target. We aim to find some sufficient conditions ensuring that $\widehat{\varkappa} \approx \varkappa^*$. We already know that the set $\varkappa^*$ will be accepted with a high probability. This particularly implies that $|\widehat{\varkappa}| \le |\varkappa^*|$. So, the question under study is the probability of the situation when $\widehat{\varkappa}$ selects some other features instead of those in $\varkappa^*$.

For simplicity of notation, we consider the sequence space model with $\Psi = I_p$ and also assume $\sigma^2 = 1$. Our first result describes which candidate set $\varkappa$ will be rejected with a high probability. The definition of the oracle $\varkappa^*$ implies that norm of the remaining part $\|\Pi_{\varkappa,\varkappa^*}\boldsymbol{\theta}^*\|$ does not exceed $\beta|\varkappa - \varkappa^*|^{1/2}$ for any $\varkappa > \varkappa^*$. Let $\varkappa$ be another subset. We measure its departure from $\varkappa^*$ by the norm of $\Pi_{\varkappa^* \setminus \varkappa}\boldsymbol{\theta}^*$, which is the projection of the true signal onto components which are in $\varkappa^*$ but not in $\varkappa$. Our result says that if this departure is large, such a candidate will be killed with a high probability. This result can be viewed as an extension of the zone-of-insensitivity result from the ordered case. Below we use the short notation for the tail functions of $\|\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}\|$ with the proper multiplicity correction; cf (5.4):

$$z_{\varkappa,\varkappa^\circ} = z_{\varkappa,\varkappa^\circ}(\mathtt{x} + q_{1,\varkappa^\circ} + q_{2,\varkappa^\circ} + \ldots + q_{k,\varkappa^\circ}), \qquad \varkappa \in \mathcal{M}_k(\varkappa^\circ).$$

Theorem 5.1.5. *Let* $\varkappa$ *be such that*

$$\|\boldsymbol{b}_{\varkappa\vee\varkappa^*,\varkappa}\| = \|\Pi_{\varkappa^*\setminus\varkappa}\boldsymbol{\theta}^*\| > 2z_{\varkappa\vee\varkappa^*,\varkappa} + \beta\,|\varkappa^* \setminus \varkappa|^{1/2}, \tag{5.9}$$

and let $\mathcal{M}^\circ(\varkappa^*)$ *be the collection of all such* $\varkappa$*. Then*

$$I\!P\big(\textit{any of } \varkappa \in \mathcal{M}^\circ(\varkappa^*) \textit{ is accepted}\big) \le \mathrm{e}^{-\mathtt{x}}.$$

Proof. We just apply the acceptance rule for $\varkappa$ requiring

$$\|\widetilde{\boldsymbol{\theta}}_{\varkappa\vee\varkappa^*} - \widetilde{\boldsymbol{\theta}}_{\varkappa}\| \le \mathtt{z}_{\varkappa\vee\varkappa^*,\varkappa} = z_{\varkappa\vee\varkappa^*,\varkappa} + \beta\,|\varkappa^* \setminus \varkappa|^{1/2}. \tag{5.10}$$

The usual decomposition of $\widetilde{\boldsymbol{\theta}}_{\varkappa}$ implies on a dominating set $\Omega(\mathtt{x})$ by (5.3)

$$\begin{aligned}\|\widetilde{\boldsymbol{\theta}}_{\varkappa\vee\varkappa^*} - \widetilde{\boldsymbol{\theta}}_{\varkappa}\| &\ge \|\boldsymbol{b}_{\varkappa\vee\varkappa^*,\varkappa}\| - \|\boldsymbol{\xi}_{\varkappa\vee\varkappa^*,\varkappa}\| \\ &\ge \|\boldsymbol{b}_{\varkappa\vee\varkappa^*,\varkappa}\| - z_{\varkappa\vee\varkappa^*,\varkappa}. \end{aligned} \tag{5.11}$$

It remains to check that the inequalities (5.10) and (5.11) are incompatible under (5.9).

This result can be directly applied to check for a subset $\varkappa_0^*$ of $\varkappa^*$ whether it will completely missed by $\widehat{\varkappa}$.

Corollary 5.1.1. *Let* $\varkappa_0^*$ *fulfill*

$$\|\Pi_{\varkappa_0^*}\boldsymbol{\theta}^*\| \ge \overline{\mathtt{z}}_{\varkappa^*}$$

with

$$\overline{\mathtt{z}}_{\varkappa^*} \stackrel{\mathrm{def}}{=} \max_{|\varkappa|\le|\varkappa^*|} \big\{2z_{\varkappa\vee\varkappa^*,\varkappa} + \beta\,|\varkappa^* \setminus \varkappa|^{1/2}\big\}.$$

Then

$$I\!P\big(\varkappa_0^* \cap \widehat{\varkappa} = \emptyset\big) \le \mathrm{e}^{-\mathtt{x}}.$$

In particular, any coefficient θ_j^* *of* $\boldsymbol{\theta}^*$ *with* $|\theta_j^*| > \overline{\mathtt{z}}_{\varkappa^*}$ *will be included in* $\widehat{\varkappa}$ *with a probability at least* $1 - \mathrm{e}^{-\mathtt{x}}$.

5.2 Anisotropic models

This section studies the so called anisotropic models when one has a number tuning parameters to be selected, and each of them is ordered. In other words, $\varkappa$ is a vector with two or more components, and we consider the set of the estimates $\widetilde{\boldsymbol{\theta}}_{\varkappa}$. When only one component of $\varkappa$ is varying and the other are fixed, the monotonicity assumption is assumed to be fulfilled. However, this only yields a componentwise partial ordering of the set $\mathcal{M}$ of all considered models.

To simplify the presentation, we consider below the two dimensional case and a product structure. An extension to the general case is straightforward.

Let $\varkappa = (\varkappa_1, \varkappa_2)$ with $\varkappa_j \in \mathcal{M}_j$ for $j = 1, 2$ and $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$. We write $\varkappa = (\varkappa_1, \varkappa_2) \geq \varkappa^\circ = (\varkappa_1^\circ, \varkappa_2^\circ)$ if $\varkappa_1 \geq \varkappa_1^\circ$ and $\varkappa_2 \geq \varkappa_2^\circ$. Assume we are given a collection of linear smoothers $\widetilde{\boldsymbol{\theta}}_{\varkappa}$ with a partial ordering: if $\varkappa > \varkappa^\circ$ then

$$\operatorname{Var}\bigl(W\widetilde{\boldsymbol{\theta}}_{\varkappa}\bigr) > \operatorname{Var}\bigl(W\widetilde{\boldsymbol{\theta}}_{\varkappa^\circ}\bigr).$$

This particularly implies

$$\mathtt{p}_{\varkappa} \stackrel{\text{def}}{=} \operatorname{tr}\bigl\{\operatorname{Var}\bigl(W\widetilde{\boldsymbol{\theta}}_{\varkappa}\bigr)\bigr\} > \mathtt{p}_{\varkappa^\circ} \stackrel{\text{def}}{=} \operatorname{tr}\bigl\{\operatorname{Var}\bigl(W\widetilde{\boldsymbol{\theta}}_{\varkappa^\circ}\bigr)\bigr\}.$$

We aim at applying the SmA method to this special situation. The setup and notation of the previous section are kept. We focus on a linear regression model $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with homogeneous Gaussian error $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$ and consider a collection of pairwise test statistics

$$\mathbb{T}_{\varkappa,\varkappa^\circ} = \sigma^{-1}\|W(\widetilde{\boldsymbol{\theta}}_{\varkappa} - \widetilde{\boldsymbol{\theta}}_{\varkappa^\circ})\| = \|\boldsymbol{\xi}_{\varkappa,\varkappa^\circ} + \boldsymbol{b}_{\varkappa,\varkappa^\circ}\|.$$

As previously, define

$$\mathtt{p}_{\varkappa,\varkappa^\circ} = \operatorname{tr}\bigl\{\operatorname{Var}\bigl(\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}\bigr)\bigr\}.$$

The acceptance rule can be written as

$$\varkappa^\circ \text{ is accepted iff } \quad \mathbb{T}_{\varkappa,\varkappa^\circ} \leq \mathtt{z}_{\varkappa,\varkappa^\circ} \quad \forall \varkappa \in \mathcal{M}(\varkappa^\circ), \tag{5.12}$$

where $\mathcal{M}(\varkappa^\circ) = \bigl\{\varkappa \in \mathcal{M} \colon \varkappa > \varkappa^\circ\bigr\}$. In words, $\varkappa^\circ$ is accepted if it is competitive with any larger model $\varkappa$. Now we discuss how the critical values $\mathtt{z}_{\varkappa,\varkappa^\circ}$ can be fixed by the multiplicity correction of the individual tail functions. Here we only discuss a uniform correction, however, it can be easily done in a multilevel form. Define $q_{\varkappa^\circ}(\mathtt{x})$ by the condition

$$I\!P\Big(\bigcup_{\varkappa \in \mathcal{M}(\varkappa^\circ)} \Big\{ \|\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}\| > z_{\varkappa,\varkappa^\circ}\big(\mathtt{x} + q_{\varkappa^\circ}(\mathtt{x})\big) \Big\} \Big) \leq \mathrm{e}^{-\mathtt{x}}. \tag{5.13}$$

Denote $\mathtt{x}_{\varkappa^\circ} = \mathtt{x} + q_{\varkappa^\circ}(\mathtt{x})$ and apply the acceptance rule (5.2) with $\mathtt{z}_{\varkappa,\varkappa^\circ}$ equal to such defined $z_{\varkappa,\varkappa^\circ}(\mathtt{x}_{\varkappa^\circ})$ after a small bias correction:

$$\mathtt{z}_{\varkappa,\varkappa^\circ} \stackrel{\text{def}}{=} z_{\varkappa,\varkappa^\circ}(\mathtt{x}_{\varkappa^\circ}) + \beta \mathtt{p}^{1/2}_{\varkappa,\varkappa^\circ}. \tag{5.14}$$

A good choice $\varkappa^\circ$ can be defined as previously by "no significant bias" condition:

$$\|\boldsymbol{b}_{\varkappa,\varkappa^\circ}\| \leq \beta \mathtt{p}^{1/2}_{\varkappa,\varkappa^\circ} \quad \varkappa \in \mathcal{M}(\varkappa^\circ). \tag{5.15}$$

The construction ensures that a good model will be accepted with a high probability.

Theorem 5.2.1. *Let $\varkappa^\circ$ be a good model in the sense (5.15). Then it holds for the acceptance rule (5.12) with the critical values $\mathtt{z}_{\varkappa,\varkappa^\circ}$ from (5.13) and (5.14)*

$$I\!P\big(\varkappa^\circ \text{ is rejected}\big) \leq \mathrm{e}^{-\mathtt{x}}.$$

So, the construction allows to figure out a set of good models, each of them will be kept by the procedure with a high probability. It remains to introduce a natural ordering on the set of such good models. This can be done by the value $\mathtt{p}_{\varkappa^\circ}$ which is proportional to $|\varkappa^\circ|$ for the sequence space model. Define the oracle choice $\varkappa^*$ as the smallest (in complexity $\mathtt{p}_{\varkappa^\circ}$) model under the constraint (5.15):

$$\varkappa^* \stackrel{\text{def}}{=} \operatorname*{argmin}_{\varkappa^\circ \in \mathcal{M}} \big\{ \mathtt{p}_{\varkappa^\circ} : \|\boldsymbol{b}_{\varkappa,\varkappa^\circ}\| \leq \beta \mathtt{p}^{1/2}_{\varkappa,\varkappa^\circ}, \quad \varkappa \in \mathcal{M}(\varkappa^\circ) \big\}.$$

Our selection rule chooses the smallest accepted model. It can be written as

$$\widehat{\varkappa} = \operatorname*{argmin}_{\varkappa^\circ \in \mathcal{M}} \big\{ \mathtt{p}_{\varkappa^\circ} : \mathbb{T}_{\varkappa,\varkappa^\circ} \leq \mathtt{z}_{\varkappa,\varkappa^\circ}, \quad \varkappa \in \mathcal{M}(\varkappa^\circ) \big\}. \tag{5.16}$$

The oracle bound compares the risk of the oracle estimate $\mathcal{R}_{\varkappa^*}$ with the risk of the adaptive estimate $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}}$ for the SmA rule $\widehat{\varkappa}$. Below for two given models $\varkappa$ and $\varkappa^\circ$, we denote by $\varkappa \vee \varkappa^\circ$ the smallest model which is larger than each:

$$\varkappa \vee \varkappa^\circ \stackrel{\text{def}}{=} (\varkappa_1 \vee \varkappa_1^\circ, \varkappa_2 \vee \varkappa_2^\circ).$$

Theorem 5.2.2. *It holds on a random set of probability at least $1 - \mathrm{e}^{-\mathtt{x}}$*

$$\sigma^{-1} \|W(\widetilde{\boldsymbol{\theta}}_{\varkappa^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}})\| \leq \overline{\mathtt{z}}_{\varkappa^*}$$

where $\overline{\mathtt{z}}_{\varkappa^*}$ is defined by

$$\overline{\mathtt{z}}_{\varkappa^*} \stackrel{\text{def}}{=} \max_{\mathtt{p}_\varkappa \leq \mathtt{p}_{\varkappa^*}} \big(\mathtt{z}_{\varkappa \vee \varkappa^*, \varkappa^*} + \mathtt{z}_{\varkappa \vee \varkappa^*, \varkappa} \big).$$

Proof. The construction and the propagation property ensures that $\varkappa^*$ is accepted with a high probability $1 - \mathrm{e}^{-\mathtt{x}}$. Below we focus on this case. Then the adaptive choice $\widehat{\varkappa}$ from (5.16) has to fulfill

$$\mathtt{p}_{\widehat{\varkappa}} \leq \mathtt{p}_{\varkappa^*}.$$

Consider $\breve{\varkappa} = \varkappa^* \vee \widehat{\varkappa}$ which contains both $\widehat{\varkappa}$ and $\varkappa^*$. As $\varkappa^*$ and $\widehat{\varkappa}$ are both accepted, it holds

$$\sigma^{-1}\|W(\widetilde{\boldsymbol{\theta}}_{\breve{\varkappa}} - \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}})\| \leq \mathtt{z}_{\breve{\varkappa},\widehat{\varkappa}}, \quad \sigma^{-1}\|W(\widetilde{\boldsymbol{\theta}}_{\breve{\varkappa}} - \widetilde{\boldsymbol{\theta}}_{\varkappa^*})\| \leq \mathtt{z}_{\breve{\varkappa},\varkappa^*}\,.$$

Therefore,

$$\sigma^{-1}\|W(\widetilde{\boldsymbol{\theta}}_{\varkappa^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}})\| \leq \mathtt{z}_{\breve{\varkappa},\widehat{\varkappa}} + \mathtt{z}_{\breve{\varkappa},\varkappa^*}.$$

and the assertion follows.

The value $\overline{\mathtt{z}}_{\varkappa^*}$ is the "payment for adaptation", and it can be quite large relative to the oracle standard deviation $\mathtt{p}_{\varkappa^*}^{1/2}$. The worst case is given by the anisotropic situation with the oracle of the form like $\varkappa^* = (\varkappa_1^*, \varkappa_{2,\max})$. In this case of a competitive model $\varkappa = (\varkappa_{1,\max}^*, \varkappa_2)$, the maximum of $\varkappa^*$ and $\varkappa$ is the largest possible model

$$\varkappa^* \vee \varkappa = (\varkappa_{1,\max}, \varkappa_{2,\max})$$

and the corresponding critical value $\mathtt{z}_{\varkappa\vee\varkappa^*,\varkappa^*}$ is very large.

To be done: In the isotropic case with $\varkappa_1^* \asymp \varkappa_2^*$, this problem disappears.

To be done: One can refine the result by considering the zone of insensitivity $\mathcal{M}^\circ(\varkappa^*)$.

6

Fisher and Wilks expansion

This chapter presents two prominent results of classical parametric statistics, namely the Fisher and Wilks Theorems, in a non-classical framework. The main features to be addressed here are a finite sample setup with large parameter dimension and a possible model misspecification.

First we specify our set-up. Let $\boldsymbol{Y}$ denote the observed data and $I\!P$ mean their distribution. A general parametric assumption (PA) means that $I\!P$ belongs to p-dimensional family $(I\!P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \subseteq I\!R^p)$ dominated by a measure $\boldsymbol{\mu}_0$. This family yields the log-likelihood function $L(\boldsymbol{\theta}) = L(\boldsymbol{Y}, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \log \frac{dI\!P_{\boldsymbol{\theta}}}{d\boldsymbol{\mu}_0}(\boldsymbol{Y})$. The PA can be misspecified, so, in general, $L(\boldsymbol{\theta})$ is a *quasi log-likelihood*. The classical likelihood principle suggests to estimate $\boldsymbol{\theta}$ by maximizing the function $L(\boldsymbol{\theta})$:

$$\widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}). \tag{6.1}$$

If $I\!P \notin \left(I\!P_{\boldsymbol{\theta}}\right)$, then the (quasi) MLE estimate $\widetilde{\boldsymbol{\theta}}$ from (6.1) is still meaningful and it appears to be an estimate of the value $\boldsymbol{\theta}^*$ defined by maximizing the expected value of $L(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} I\!E L(\boldsymbol{\theta}).$$

Here $\boldsymbol{\theta}^*$ is the true value in the parametric situation and can be viewed as the parameter of the best parametric fit in the general case. The study is non-asymptotic, that is, we proceed with only one sample $\boldsymbol{Y}$. One can easily extend it to an asymptotic setup in which the data, its distribution, the parameter space and the parametric family depend on the asymptotic parameter like the sample size. One example is given below in Section 6.4 for the case of an i.i.d. sample.

The Fisher expansion of the qMLE $\widetilde{\boldsymbol{\theta}}$ is given as follows:

$$D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \approx \boldsymbol{\xi} \stackrel{\text{def}}{=} D^{-1} \nabla L(\boldsymbol{\theta}^*),$$

where $\nabla L(\boldsymbol{\theta}) = \frac{dL}{d\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $D^2 \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}^*)$ is the analog of the total Fisher information matrix. In classical situations, the standardized score $\boldsymbol{\xi}$ is asymptotically standard normal yielding asymptotic root-n normality and efficiency of the MLE $\widetilde{\boldsymbol{\theta}}$. Theorem 6.3.2 carefully describes how the error of this expansion depends on the parameter dimension p and the regularity of the model. The Wilks expansion means

$$L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) \approx \|\boldsymbol{\xi}\|^2/2.$$

Again, if the vector $\boldsymbol{\xi}$ is asymptotically standard normal, the expansion yields the classical χ^2_p asymptotic distribution for the excess $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$.

The whole study is nonasymptotic and all "small" terms are carefully described. This helps to understand how the parameter dimension is involved and particularly to address the question of a *critical dimension*; see Section 6.4 which specifies the result to the i.i.d. case with n observations and links the obtained results to the classical literature.

6.1 Main results

This section presents our main results which include the Fisher and Wilks expansions for a non-classical and non-asymptotic framework. First we present the frequentist results: concentration and large deviation properties of the maximum likelihood estimator $\widetilde{\boldsymbol{\theta}}$, the Fisher expansion for the difference $\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}$ and the Wilks expansion for the excess $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta})$. The results are stated in a concise way, all the terms are given explicitly. Surprisingly, the leading terms in all bounds are sharp, in particular, the classical results on asymptotic efficiency can be easily derived from the obtained expansions.

Introduce the notation $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$ for the (quasi) log-likelihood ratio. The main step in the approach is the following *uniform local bracketing result*:

$$\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \Delta \le L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \le \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \Delta, \qquad \boldsymbol{\theta} \in \Theta_0. \tag{6.2}$$

Here $\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is a quadratic in $\boldsymbol{\theta} - \boldsymbol{\theta}^*$ expression, Δ is a small error only depending on Θ_0 which is a local vicinity of the central point $\boldsymbol{\theta}^*$. This result can be viewed as an extension of the famous Le Cam *local asymptotic normality* (LAN) condition. The LAN condition postulates an approximation of the log-likelihood $L(\boldsymbol{\theta})$ by a nearly Gaussian process; see e.g. Ibragimov and Khas'minskij (1981) or Kleijn and van der Vaart (2012) for an extension of this condition (stochastic LAN). The bracketing bound (6.2) requires only some general conditions listed in Section 6.2. A model misspecification case is included. Similarly to the LAN theory, the bracketing result has a number of remarkable corollaries like the Wilks and Fisher Theorems; see Theorems 6.3.2 and 6.3.3.

For making a precise statement, we have to specify the ingredients of the bracketing device. The most important one is a symmetric positive $p \times p$-matrix D^2. In typical situations, it can be defined as the negative Hessian of the expected log-likelihood: $D^2 = -\nabla^2 I\!\!E L(\boldsymbol{\theta}^*)$. Also one has to specify a radius $\mathtt{r}_0$ entering in the definition of the local vicinity $\Theta_0(\mathtt{r}_0)$ of the central point $\boldsymbol{\theta}^*$: $\Theta_0(\mathtt{r}_0) = \{\boldsymbol{\theta}\colon \|D(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\| \le \mathtt{r}_0\}$. The bracketing result (6.2) can be stated for $\Theta_0 = \Theta_0(\mathtt{r}_0)$ with

$$\begin{aligned}\mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*) &\stackrel{\text{def}}{=} (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top \nabla L(\boldsymbol{\theta}^*) - \|D(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2/2 \\ &= \boldsymbol{\xi}^\top D(\boldsymbol{\theta}-\boldsymbol{\theta}^*) - \|D(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2/2\end{aligned}$$

and

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} D^{-1}\nabla L(\boldsymbol{\theta}^*). \tag{6.3}$$

The construction is essentially changed relative to Spokoiny (2012) (in fact, it is simplified) by using only one matrix D^2 while Spokoiny (2012) used three matrices $D^2_{\boldsymbol{\epsilon}}$, $D^2_{\underline{\boldsymbol{\epsilon}}}$, and V^2. The bracketing bound (6.2) becomes useful if the error Δ is relatively small and can be neglected.

6.2 Conditions

This section collects the conditions which are systematically used in the text. The conditions are quite general and seem to be non-restrictive; see the discussion at the end of the section. We mainly require some regularity and smoothness of the log-likelihood process $L(\boldsymbol{\theta})$. With $D^2 = -\nabla^2 I\!\!E L(\boldsymbol{\theta}^*)$, define the local elliptic sets $\Theta_0(\mathtt{r})$ as

$$\Theta_0(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\theta} \in \Theta\colon \|D(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\| \le \mathtt{r}\}.$$

We distinguish between local and global conditions. Local ones are stated on $\Theta_0(\mathtt{r}_0)$, while the global one corresponds to $\mathtt{r} \ge \mathtt{r}_0$, where the value $\mathtt{r}_0$ will be specified later.

The first condition requires that the expected log-likelihood $I\!\!E L(\boldsymbol{\theta})$ is twice continuously differentiable.

$(\mathcal{L}_0)$ *For each $\mathtt{r} \le \mathtt{r}_0$, there is a constant $\delta(\mathtt{r}) \le 1/2$ such that it holds for any $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ and $D^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2 I\!\!E L(\boldsymbol{\theta})$:*

$$\bigl\|D^{-1}D^2(\boldsymbol{\theta})D^{-1} - I_p\bigr\|_\infty \le \delta(\mathtt{r}).$$

Under $(\mathcal{L}_0)$, it follows from the second order Taylor expansion at $\boldsymbol{\theta}^*$:

$$\bigl|-2I\!\!E L(\boldsymbol{\theta},\boldsymbol{\theta}^*) - \|D(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2\bigr| \le \delta(\mathtt{r})\|D(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2, \quad \boldsymbol{\theta} \in \Theta_0(\mathtt{r}).$$

For $\mathtt{r} > \mathtt{r}_0$, we need a global identification property which ensures that the deterministic component $I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ of the log-likelihood is competitive with the variation of the stochastic component.

($\mathcal{L}$) *There exists* $\mathtt{b}(\mathtt{r}) > 0$ *such that* $\mathtt{r}\mathtt{b}(\mathtt{r})$ *is non-decreasing and*

$$\frac{-2I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2} \geq \mathtt{b}(\mathtt{r}), \qquad \forall \mathtt{r} \geq \mathtt{r}_0,\ \boldsymbol{\theta} \in \Theta_0(\mathtt{r}).$$

Now we consider the stochastic component of the process $L(\boldsymbol{\theta})$:

$$\zeta(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - I\!\!E L(\boldsymbol{\theta}).$$

We assume that it is twice differentiable and denote by $\nabla\zeta(\boldsymbol{\theta})$ its gradient and by $\nabla^2\zeta(\theta)$ its Hessian matrix.

(ED_0) *There exist a positive symmetric matrix* V^2, *and constants* $\mathtt{g} > 0$, $\nu_0 \geq 1$ *such that* $\mathrm{Var}\bigl\{\nabla\zeta(\boldsymbol{\theta}^*)\bigr\} \leq V^2$ *and*

$$\sup_{\boldsymbol{\gamma} \in I\!R^p} \log I\!\!E \exp\biggl\{ \lambda \frac{\boldsymbol{\gamma}^\top \nabla\zeta(\boldsymbol{\theta}^*)}{\|V\boldsymbol{\gamma}\|} \biggr\} \leq \nu_0^2 \lambda^2/2, \qquad |\lambda| \leq \mathtt{g}.$$

(ED_2) *There exist a value* $\omega > 0$ *and for each* $\mathtt{r} > 0$, *a constant* $\mathtt{g}(\mathtt{r}) > 0$ *such that it holds for any* $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$*:*

$$\sup_{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in I\!R^p} \log I\!\!E \exp\biggl\{ \frac{\lambda}{\omega} \frac{\boldsymbol{\gamma}_1^\top \nabla^2\zeta(\boldsymbol{\theta})\boldsymbol{\gamma}_2}{\|D\boldsymbol{\gamma}_1\| \cdot \|D\boldsymbol{\gamma}_2\|} \biggr\} \leq \frac{\nu_0^2 \lambda^2}{2}, \qquad |\lambda| \leq \mathtt{g}(\mathtt{r}).$$

Below we only need that the constant $\mathtt{g}(\mathtt{r})$ is larger than $\mathtt{C}p$ for a fixed constant $\mathtt{C}$ and all $\mathtt{r}$.

The *identifiability condition* relates the matrices V^2 and D^2.

($\mathcal{I}$) There is a constant $\mathfrak{a} > 0$ such that

$$\mathfrak{a}^2 D^2 \geq V^2.$$

Remark 6.2.1. The conditions involve some constants. We distinguish between important constants and technical ones. The impact of the important constants is shown in our results, the list includes $\delta(\mathtt{r})$, ω, and $\mathfrak{a}$. The constant $\mathfrak{a}$ can be viewed as the largest eigenvalue of $B = D^{-1}V^2D^{-1}$ and it enters in the definition of the upper quantile function $q(B, \mathtt{x})$ for $\|\boldsymbol{\xi}\|$; see Proposition 6.5.3 below. The other constants like ν_0 or $\mathtt{g}(\mathtt{r})$ are technical. The constant ν_0 is introduced for convenience only, it can be omitted by rescaling the matrix V. In the asymptotic setup it can usually be selected very close to one.

Remark 6.2.2. We briefly comment how restrictive the imposed conditions are. Spokoiny (2012), Section 5.1, considered in details the i.i.d. case and presented some mild sufficient conditions on the parametric family which imply the above general conditions. Condition (ED_0) requires some exponential moments of the observations (errors). Usually one only assumes some finite moments of the errors; cf. Ibragimov and Khas'minskij (1981), Chapter 2. Our condition is a bit more restrictive but it allows to obtain some finite sample bounds. Condition $(\mathcal{L}_0)$ only requires some regularity of the considered parametric family and is not restrictive. Conditions (ED_2) with $\mathtt{g}(\mathtt{r}) \equiv \mathtt{g} > 0$ and $(\mathcal{L})$ with $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b} > 0$ are easy to verify if the parameter set Θ is compact and the sample size n exceeds $\mathtt{C}p$ for a fixed constant $\mathtt{C}$. It suffices to check a usual identifiability condition that the value $I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ does not vanish for $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$.

The regression and generalized regression models are included as well; cf. Ghosal (1999, 2000) or Kim (2006). Spokoiny (2012), Section 5.2, argued that (ED_2) is automatically fulfilled for a generalized linear model, while (ED_0) requires that regression errors have to fulfill some exponential moments conditions. If this condition is too restrictive and a more stable (robust) estimation procedure is desirable, one can apply the LAD-type contrast leading to median regression. Spokoiny (2012), Section 5.3, showed for the case of linear median regression that all the required conditions are fulfilled automatically if the sample size n exceeds $\mathtt{C}p$ for a fixed constant $\mathtt{C}$. Spokoiny et al. (2013) applied this approach for local polynomial quantile regression. Zaitsev et al. (2013) applied the approach to the problem of regression with Gaussian process where the unknown parameters enter in the likelihood in a rather complicated way.

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

This section collects the main results about the MLE $\widetilde{\boldsymbol{\theta}}$. We begin by a large deviation bound which ensures a small probability of the event $\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r}_0)$. Then we present the Fisher and Wilks expansions. The formulation involves two growing functions of the argument $\mathtt{x}$: $q(B, \mathtt{x})$ and $q_{\mathbb{H}}(\mathtt{x})$. The functions are given analytically and only depend on the parameters of the model. The function $q(B, \mathtt{x})$ describes the quantiles of the norm of the vector $\|\boldsymbol{\xi}\|$ from (6.3). The definition is given in (6.25). Further, the function $q_{\mathbb{H}}(\mathtt{x})$ is related to the entropy of the parameter space and it is given by (6.13). In typical situations one can use the upper bound $q^2(\mathtt{x}) \leq \mathtt{C}(p + \mathtt{x})$ for both functions. The first result explains the choice of $\mathtt{r}_0$ ensuring with a high probability that $\widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r}_0)$.

Theorem 6.3.1. *Suppose* (ED_0) *and* (ED_2), $(\mathcal{L}_0)$, $(\mathcal{L})$, *and* $(\mathcal{I})$. *Let also the function* $\mathtt{b}(\mathtt{r})$ *in* $(\mathcal{L})$ *satisfy*

$$\mathtt{b}(\mathtt{r})\,\mathtt{r} \geq 2q(B,\mathtt{x}) + 2\varrho(\mathtt{r},\mathtt{x}), \quad \mathtt{r} > \mathtt{r}_0, \tag{6.4}$$

where

$$\varrho(\mathtt{r},\mathtt{x}) \stackrel{\text{def}}{=} 6\nu_0\, q_{\mathbb{H}}\bigl(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0)\bigr)\,\omega. \tag{6.5}$$

Then

$$I\!\!P\bigl(\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r}_0)\bigr) \leq 3\mathrm{e}^{-\mathtt{x}}.$$

Remark 6.3.1. The radius $\mathtt{r}_0$ has to fulfill (6.4). This condition is easy to check in typical situations. One can use that $\mathtt{b}(\mathtt{r}_0) \geq 1 - \delta(\mathtt{r}_0) \approx 1$, that the constant ω is small, and $\mathtt{r}\mathtt{b}(\mathtt{r})$ grows with $\mathtt{r}$. A simple rule $\mathtt{r}_0 \geq (2+\delta)q(B,\mathtt{x})$ for some $\delta > 0$ works in most of cases.

Now we state the result about the Fisher expansion for the qMLE $\widetilde{\boldsymbol{\theta}}$.

Theorem 6.3.2. *Suppose the conditions of Theorem 6.3.1. On a random set $\Omega(\mathtt{x})$ of a dominating probability at least $1 - 4\mathrm{e}^{-\mathtt{x}}$, it holds*

$$\|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi}\| \leq \diamondsuit(\mathtt{r}_0,\mathtt{x}), \tag{6.6}$$

where for the function $q_{\mathbb{H}}(\mathtt{x})$ given by (6.13), the value $\diamondsuit(\mathtt{r},\mathtt{x})$ is given by

$$\diamondsuit(\mathtt{r},\mathtt{x}) \stackrel{\text{def}}{=} \bigl\{\delta(\mathtt{r}) + 6\nu_0\, q_{\mathbb{H}}(\mathtt{x})\,\omega\bigr\}\,\mathtt{r}. \tag{6.7}$$

Our version of the Wilks result can be stated in the following form.

Theorem 6.3.3. *Suppose the conditions of Theorem 6.3.1. On a random set $\Omega(\mathtt{x})$ of a dominating probability at least $1 - 5\mathrm{e}^{-\mathtt{x}}$, it holds with $\diamondsuit(\mathtt{r}_0,\mathtt{x})$ from (6.7)*

$$\begin{aligned} \bigl|2L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*) - \|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2\bigr| &\leq 2\mathtt{r}_0\,\diamondsuit(\mathtt{r}_0,\mathtt{x}) + \diamondsuit^2(\mathtt{r}_0,\mathtt{x}), \\ \bigl|2L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*) - \|\boldsymbol{\xi}\|^2\bigr| &\leq 2\mathtt{r}_0\,\diamondsuit(\mathtt{r}_0,\mathtt{x}) + \diamondsuit^2(\mathtt{r}_0,\mathtt{x}). \end{aligned} \tag{6.8}$$

Furthermore,

$$\begin{aligned} \Bigl|\sqrt{2L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*)} - \|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|\Bigr| &\leq 2\diamondsuit(\mathtt{r}_0,\mathtt{x}), \\ \Bigl|\sqrt{2L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*)} - \|\boldsymbol{\xi}\|\Bigr| &\leq 3\diamondsuit(\mathtt{r}_0,\mathtt{x}), \end{aligned} \tag{6.9}$$

and for any $\boldsymbol{\theta}^\circ \in \Theta_0(\mathtt{r}_0)$, it holds on $\Omega(\mathtt{x})$

$$\begin{aligned} \Bigl|\bigl\{2L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^\circ)\bigr\}^{1/2} - \|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ)\|\Bigr| &\leq 4\diamondsuit(\mathtt{r}_0,\mathtt{x}), \\ \Bigl|\bigl\{2L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^\circ)\bigr\}^{1/2} - \|\boldsymbol{\xi} + D(\boldsymbol{\theta}^* - \boldsymbol{\theta}^\circ)\|\Bigr| &\leq 5\diamondsuit(\mathtt{r}_0,\mathtt{x}). \end{aligned}$$

Remark 6.3.2. The classical Fisher and Wilks results describe asymptotic behavior of the MLE $\widetilde{\boldsymbol{\theta}}$ and of the excess $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$. The whole derivations are based on expansions similar to (6.6) and (6.8) and on the limiting behavior of the vector $\boldsymbol{\xi}$ and its squared norm. Under standard assumptions in the regression or i.i.d. setup the vector $\boldsymbol{\xi}$ is standard normal and $\|\boldsymbol{\xi}\|^2$ is asymptotically χ^2 with p degrees of freedom. The asymptotic distribution of the MLE $\widetilde{\boldsymbol{\theta}}$ or of the excess $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ can be used for building the confidence sets or for test critical values. However, the use of asymptotic arguments is limited and faces serious problems in practical applications.

This especially concerns the likelihood ratio statistic $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$. It is well recognized that the accuracy of χ^2-approximation of the tails of $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ is very poor and a reasonable quality requires a huge sample size. If the parameter dimension grows with n this problem becomes even more crucial. The qualitative tail behavior of $\|\boldsymbol{\xi}\|^2$ is described in Proposition 6.5.3 but the upper bounds given there appear to be too conservative for practical use.

Remark 6.3.3. Another issue is a possible model misspecification. The expansions (6.8) and (6.9) apply, even if $L(\boldsymbol{\theta})$ is a quasi-log-likelihood function. However, the covariance matrix $V^2 = \operatorname{Var}\bigl\{\nabla L(\boldsymbol{\theta}^*)\bigr\}$ of the score does not necessarily coincide with the information matrix D^2. Then the covariance matrix of the vector $\boldsymbol{\xi}$ follows the famous "sandwich" formula $\operatorname{Var}(\boldsymbol{\xi}) = D^{-1} V^2 D^{-1}$, and the distribution of the squared norm $\|\boldsymbol{\xi}\|^2$ depends on the unknown covariance matrix V^2.

The results presented above focus on the expansions of the MLE $\widetilde{\boldsymbol{\theta}}$ and on the excess $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$. Numerical results (not presented here) indicate that the accuracy of the expansions (6.6) and (6.8) is very reasonable even for moderate and small samples and it is stable w.r.t. possible model misspecifications. It seems that such expansions are of independent interest and can be used for many further statistical tasks.

6.4 Critical dimension. Examples

This section discusses the issue of the critical dimension of the parameter space Θ. This particularly allows to consider the case of a growing parameter dimension. It appears that the error of expansions is different for different types of problems under consideration. The large deviation result of Theorem 6.3.1 for the MLE $\widetilde{\boldsymbol{\theta}}$ requires (6.4). In particular, the value $\mathtt{r}_0^2$ should be at least $\mathtt{C}(p + \mathtt{x})$. This result yields consistency of the MLE if the neighborhood $\Theta_0(\mathtt{r}_0)$ is small, or, equivalently, $(p + \mathtt{x})^{1/2} \|D^{-1}\|_\infty$ is small. The Fisher expansion (6.6) of Theorem 6.3.2 requires that the error term $\diamondsuit(\mathtt{r}_0, \mathtt{x})$ from (6.5) is small. Finally, the Wilks expansion holds if the error term $\Delta(\mathtt{r}_0, \mathtt{x})$ is small. Now we specify the results for some popular statistical models.

6.4.1 Linear and generalized linear models

In the case of a linear model $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon}$ with a given design $p \times n$ matrix Ψ under the assumption of Gaussian noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$, the standard calculus leads to the log-likelihood

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

where the remainder R does not depend on $\boldsymbol{\theta}$. Moreover, $L(\boldsymbol{\theta})$ is quadratic in $\boldsymbol{\theta}$ and its Hessian is constant: $\nabla^2 I\!\!E L(\boldsymbol{\theta}) = -\Psi \Sigma^{-1} \Psi^\top$. One can summarize as follows: with $I\!\!E \boldsymbol{Y} = \boldsymbol{f}$

$$D^2 \stackrel{\text{def}}{=} -\nabla^2 I\!\!E L(\boldsymbol{\theta}) = \Psi \Sigma^{-1} \Psi^\top, \qquad \widetilde{\boldsymbol{\theta}} = D^{-2} \Psi \Sigma^{-1} \boldsymbol{Y},$$
$$\boldsymbol{\xi} \stackrel{\text{def}}{=} D^{-1} \nabla L(\boldsymbol{\theta}^*) = D^{-1} \Psi \Sigma^{-1} (\boldsymbol{Y} - \boldsymbol{f}). \qquad \boldsymbol{\theta}^* = D^{-2} \Psi \Sigma^{-1} \boldsymbol{f},$$

Moreover,

$$D\big(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\big) \equiv \boldsymbol{\xi}, \qquad L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) \equiv \|\boldsymbol{\xi}\|^2/2.$$

All these results are straightforward, the last one is obtained by the Tailor expansion of the second order around $\widetilde{\boldsymbol{\theta}}$ with the use of $\nabla L(\widetilde{\boldsymbol{\theta}}) = 0$:

$$L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) = -\frac{1}{2}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \nabla^2 L(\widetilde{\boldsymbol{\theta}})(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = -\frac{1}{2}\big\|D\big(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\big)\big\|^2 = -\frac{1}{2}\|\boldsymbol{\xi}\|^2.$$

The presented derivations mean that the Fisher and Wilks expansions are *identities*, they apply for *any sample size* without *any conditions*, and are only based on *quadraticity* of the likelihood function $L(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$. The *true distribution* of $\boldsymbol{Y}$ can be whatever and it is not involved at all. There is *no dimensional restrictions*. However, for *inference*, the parametric assumption is important. It only concerns the *distribution of* $\boldsymbol{\xi}$. Let $\mathrm{Var}\big(\boldsymbol{Y}\big) = \Sigma_0 \neq \Sigma$. Then with $D^2 = \Psi \Sigma^{-1} \Psi^\top$

$$\mathrm{Var}\big\{\nabla L(\boldsymbol{\theta}^*)\big\} = \mathrm{Var}\big\{\Psi \Sigma^{-1} \boldsymbol{Y}\big\} = \Psi \Sigma^{-1} \Sigma_0 \Sigma^{-1} \Psi^\top \stackrel{\text{def}}{=} V^2 \neq D^2$$

which leads to the famous *sandwich formula*

$$\mathrm{Var}(\boldsymbol{\xi}) = \mathrm{Var}\big\{D^{-1} \nabla L(\boldsymbol{\theta}^*)\big\} = D^{-1} V^2 D^{-1} \neq \boldsymbol{I}_p.$$

6.4.2 Generalized linear models (GLM)

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \sim I\!\!P$ be a sample of independent r.v.s. The parametric GLM model is given by $Y_i \sim P_{\Psi_i^\top \boldsymbol{\theta}} \in (P_\upsilon)$, where Ψ_i are given factors in $I\!R^p$, $\boldsymbol{\theta} \in I\!R^p$

is the unknown parameter in $\mathbb{R}^p$, and (P_υ) is an exponential family with canonical parametrization yielding the log-density $\ell(y,\upsilon) = y\upsilon - d(\upsilon)$ for a convex function $d(\upsilon)$. The MLE $\widetilde{\boldsymbol{\theta}}$ and the target $\boldsymbol{\theta}^*$ for this GLM read as

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^n \bigl\{ Y_i \Psi_i^\top \boldsymbol{\theta} - d(\Psi_i^\top \boldsymbol{\theta}) \bigr\},$$

$$\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^n \bigl\{ f_i \Psi_i^\top \boldsymbol{\theta} - d(\Psi_i^\top \boldsymbol{\theta}) \bigr\}$$

with $f_i = \mathbb{E} Y_i$. An important feature of a GLM is that the stochastic component $\zeta(\boldsymbol{\theta})$ of $L(\boldsymbol{\theta})$ is *linear in* $\boldsymbol{\theta}$: with $\varepsilon_i = Y_i - \mathbb{E} Y_i$

$$\zeta(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathbb{E} L(\boldsymbol{\theta}) = \Bigl(\sum_{i=1}^n \varepsilon_i \Psi_i \Bigr)^\top \boldsymbol{\theta}, \quad \nabla \zeta(\boldsymbol{\theta}) = \sum_{i=1}^n \varepsilon_i \Psi_i .$$

In the contrary to the linear case, the Fisher information matrix D^2 depends on the true data distribution via the target $\boldsymbol{\theta}^*$:

$$D^2 = \sum_i \Psi_i \Psi_i^\top d''(\Psi_i^\top \boldsymbol{\theta}^*).$$

The vector $\boldsymbol{\xi}$ is given by

$$\boldsymbol{\xi} = D^{-1} \nabla \zeta(\boldsymbol{\theta}^*) = D^{-1} \sum_{i=1}^n \varepsilon_i \Psi_i \, .$$

Here is a list of sufficient conditions which ensure our general conditions from Section 6.2: the functions $d''(\Psi_i^\top \boldsymbol{\theta})$ are uniformly continuous in $\boldsymbol{\theta}$ and $i \le n$, for some matrices $\mathtt{v}_i^2$ and fixed constants $\mathtt{C}_0, \lambda_0 > 0$

$$\mathbb{E} \exp\bigl\{ \lambda_0 \mathtt{v}_i^{-1} \varepsilon_i \bigr\} \le \mathtt{C}_0, \qquad i = 1, \ldots, n,$$

the sample size n is larger than $\mathtt{C} p$ for a prescribed constant $\mathtt{C}$, and the matrix $V^2 = \sum_i \mathtt{v}_i^2$ fulfills

$$V^2 \le \mathfrak{a}^2 D^2 .$$

The details can be found in Spokoiny (2012).

As already mentioned, the error of Fisher and Wilks approximations only include the term $\delta(\mathtt{r}_0)$ because the stochastic term is linear by definition. By using the higher order expansion of the function $d(\cdot)$, Portnoy (1988) showed that the error $\diamondsuit(\mathtt{r}_0, \mathtt{x})$ can be improved under the true the parametric assumption, however, the case of a model misspecification is not include. Under mild regularity conditions on the design Ψ and on

smoothness of $d(\cdot)$, the consistency result applies with $\mathtt{r}_0^2 = \mathtt{C}(p+\mathtt{x})$, and the bound $\delta(\mathtt{r}_0) \asymp \mathtt{r}_0/\sqrt{n}$ yields the errors $\diamondsuit(\mathtt{r}_0, \mathtt{x}) \asymp \mathtt{r}_0\delta(\mathtt{r}_0) \le \mathtt{C}(p+\mathtt{x})/\sqrt{n}$ and $\Delta(\mathtt{r}_0, \mathtt{x}) \asymp \mathtt{r}_0^2\delta(\mathtt{r}_0) \le \mathtt{C}(p+\mathtt{x})^{3/2}/\sqrt{n}$.

6.4.3 I.i.d. case

Now we consider the asymptotic setup in which $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ is an i.i.d. sample from a measure P. The parametric assumption means that $P \in (P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta)$ for a given family of marginal measures $(P_{\boldsymbol{\theta}})$. We admit that the parametric family depends on n and the parameter dimension $p = p_n$ grows to infinity with the sample size n. We also allow that the parametric assumption can be misspecified.

Similarly to Spokoiny (2012), our general conditions can be transformed into the conditions on the family $(P_{\boldsymbol{\theta}})$ and the marginal measure P. It suffices to check that the first and second derivatives of the log-density function $\ell(y, \boldsymbol{\theta}) = \log dP_{\boldsymbol{\theta}}/d\mu_0(y)$ have exponential moments and the expectation $I\!\!E\ell(Y_1, \boldsymbol{\theta})$ is three times continuously differentiable in $\boldsymbol{\theta}$. See Section 5.1 in Spokoiny (2012) for more details.

Also select $\mathtt{x} = \mathtt{x}_n$ depending on n and growing slowly with n, for instance, $\mathtt{x}_n = \log n$. The matrix D^2 satisfies $D^2 = n\mathbb{F}_{\boldsymbol{\theta}^*}$, where $\mathbb{F}_{\boldsymbol{\theta}^*}$ is the Fisher information of $(P_{\boldsymbol{\theta}})$ at $\boldsymbol{\theta}^*$ if the parametric assumption holds.

The bracketing bound and the large deviation result from Spokoiny (2012) and from Section 6.3 apply if the sample size n fulfills $n \ge \mathtt{C}p_n$ for a fixed constant $\mathtt{C}$. It appears that the Fisher and Wilks results require stronger conditions. Indeed, in the regular i.i.d. case it holds $\delta(\mathtt{r}_0) \asymp \mathtt{r}_0/\sqrt{n}$, $q^2_{\mathbb{H}}(\mathtt{x}_n) \asymp p_n + \mathtt{x}_n$, and $\omega \asymp 1/\sqrt{n}$. The radius $\mathtt{r}_0$ should fulfill $\mathtt{r}_0^2 \ge \mathtt{C}p_n$ to ensure the large deviation result. This yields

$$\diamondsuit(\mathtt{r}_0, \mathtt{x}_n) \le \bigl\{\delta(\mathtt{r}_0) + 3\nu_0 q_{\mathbb{H}}(\mathtt{x}_n)\omega\bigr\}\mathtt{r}_0 \le \mathtt{C}p_n/\sqrt{n}.$$

Similarly

$$\Delta(\mathtt{r}_0, \mathtt{x}_n) \le \bigl\{\delta(\mathtt{r}_0) + 6\nu_0 q_{\mathbb{H}}(\mathtt{x}_n)\omega\bigr\}\mathtt{r}_0^2 \le \mathtt{C}\sqrt{p_n^3/n}.$$

One can conclude that the consistency result is valid under $p_n/n \to 0$, the Fisher expansion requires $p_n^2/n \to 0$, while the Wilks is applicable under $p_n^3/n \to 0$ as $n \to \infty$.

Theorem 6.4.1. *Suppose the conditions of Theorem 5.4 in Spokoiny (2012). If $p_n^2/n \to 0$, then the Fisher expansion (6.6) of Theorem 6.3.2 holds with the error term $\diamondsuit(\mathtt{r}_0, \mathtt{x}_n) \to 0$. Let also $p_n^3/n \to 0$. Then the error term $\Delta(\mathtt{r}_0, \mathtt{x}_n)$ in the Wilks expansion (6.8) satisfies $\Delta(\mathtt{r}_0, \mathtt{x}_n) \to 0$.*

Existing statistical literature addresses the issue of a growing parameter dimension in different set-ups. The classical results by Portnoy (1984, 1985, 1986) provide some

constraints on parameter dimension for consistency and asymptotic normality of the M-estimator for regression models. Our results are consistent with the conclusion of that papers. Mammen (1993, 1996) discussed the validity of bootstrap procedures in linear models with many parameters. The obtained results are valid under $p_n^{3/2}/n \to 0$, however are limited to the testing problem for linear models.

The setup with growing parameter dimension is naturally used in sieve nonparametric estimation when a nonparametric model is approximated by a sequence of parametric ones. We mention papers by Shen and Wong (1994); Shen (1997), Birgé and Massart (1993), Van de Geer (1993); van de Geer (2002). Some minimal smoothness assumptions are normally imposed on the underlying nonparametric function which ensure that the parameter dimension of a sieve is smaller in order than the sample size.

6.5 Some auxiliary results and proofs

This section collects some auxiliary results about the behavior of the posterior measures which might be of independent interest.

6.5.1 Local linear approximation of the gradient of the log-likelihood

The principle step of the proof is a bound on the local linear approximation of the gradient $\nabla L(\boldsymbol{\theta})$. Below we study separately its stochastic and deterministic components coming from the decomposition $L(\boldsymbol{\theta}) = I\!\!E L(\boldsymbol{\theta}) + \zeta(\boldsymbol{\theta})$. With $D^2 = -\nabla^2 I\!\!E L(\boldsymbol{\theta}^*)$, this leads to the decomposition

$$\begin{aligned} \chi(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &\stackrel{\text{def}}{=} D^{-1}\bigl\{\nabla L(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta}^*) + D^2 (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr\} \\ &= D^{-1}\bigl\{\nabla \zeta(\boldsymbol{\theta}) - \nabla \zeta(\boldsymbol{\theta}^*) + \nabla I\!\!E L(\boldsymbol{\theta}) - \nabla I\!\!E L(\boldsymbol{\theta}^*) + D^2 (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr\}. \end{aligned} \tag{6.10}$$

First we check the deterministic part. For any $\boldsymbol{\theta}$ with $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}$

$$\begin{aligned} I\!\!E\bigl[\chi(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\bigr] &\stackrel{\text{def}}{=} D^{-1}\bigl\{\nabla I\!\!E L(\boldsymbol{\theta}) - \nabla I\!\!E L(\boldsymbol{\theta}^*)\bigr\} + D(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\ &= \bigl\{I_p - D^{-1} D^2(\boldsymbol{\theta}^\circ) D^{-1}\bigr\}\, D(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \end{aligned}$$

where $\boldsymbol{\theta}^\circ$ is a point on the line connecting $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$. This implies by $(\mathcal{L}_0)$

$$I\!\!E\bigl[\chi(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\bigr] \le \|I_p - D^{-1} D^2(\boldsymbol{\theta}^\circ) D^{-1}\|_\infty \, \mathtt{r} \le \delta(\mathtt{r})\mathtt{r}. \tag{6.11}$$

Now we study the stochastic part. Consider the vector process

$$\mathcal{U}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \stackrel{\text{def}}{=} D^{-1}\bigl\{\nabla \zeta(\boldsymbol{\theta}) - \nabla \zeta(\boldsymbol{\theta}^*)\bigr\}.$$

It is convenient to change the variable by $\upsilon = D(\theta - \theta^*)$ and consider the vector process $\mathcal{Y}(\upsilon) = \mathcal{U}(\theta, \theta^*)$. It obviously holds $\nabla \mathcal{Y}(\upsilon) = D^{-1} \nabla^2 \zeta(\theta) D^{-1}$. Moreover, for any $\gamma_1, \gamma_2 \in \mathbb{R}^p$ with $\|\gamma_1\| = \|\gamma_2\| = 1$, condition (ED_2) implies

$$\log \mathbb{E} \exp\left\{ \frac{\lambda}{\omega} \gamma_1^\top \nabla \mathcal{Y}(\upsilon) \gamma_2 \right\} = \log \mathbb{E} \exp\left\{ \frac{\lambda}{\omega} \gamma_1^\top D^{-1} \nabla^2 \zeta(\theta) D^{-1} \gamma_2 \right\} \le \frac{\nu_0^2 \lambda^2}{2}.$$

Define $\Upsilon_\circ(\mathtt{r}) \stackrel{\text{def}}{=} \{\upsilon \colon \|\upsilon\| \le \mathtt{r}\}$. Then

$$\sup_{\theta \in \Theta_0(\mathtt{r})} \|\mathcal{U}(\theta, \theta^*)\| = \sup_{\upsilon \in \Upsilon_\circ(\mathtt{r})} \|\mathcal{Y}(\upsilon)\|. \tag{6.12}$$

Theorem 6.5.3 yields

$$\sup_{\upsilon \in \Upsilon_\circ(\mathtt{r})} \|\mathcal{Y}(\upsilon)\| \le 6\nu_0 \, q_{\mathbb{H}}(\mathtt{x}) \, \omega \, \mathtt{r}$$

on a set of a dominating probability at least $1 - \mathrm{e}^{-\mathtt{x}}$, where the function $q_{\mathbb{H}}(\mathtt{x})$ is given by the following rules:

$$q_{\mathbb{H}}(\mathtt{x}) = \begin{cases} \sqrt{\mathbb{H}_2 + 2\mathtt{x}}, & \text{if } \mathbb{H}_2 + 2\mathtt{x} \le \mathtt{g}^2, \\ \mathtt{g}^{-1}\mathtt{x} + \frac{1}{2}\bigl(\mathtt{g}^{-1}\mathbb{H}_2 + \mathtt{g}\bigr), & \text{if } \mathbb{H}_2 + 2\mathtt{x} > \mathtt{g}^2. \end{cases} \tag{6.13}$$

Here $\mathbb{H}_2 = 4p$ and $\mathbb{H}_1 = 2p^{1/2}$; see Theorem 6.5.1 in the Appendix.

Putting together the bounds (6.11) and (6.12) imply the following result.

Proposition 6.5.1. *Suppose that the matrix* $D^2(\theta) \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} L(\theta)$ *fulfills the condition* $(\mathcal{L}_0)$ *and let also* (ED_2) *be fulfilled on* $\Theta_0(\mathtt{r})$ *for any fixed* $\mathtt{r}$. *Then*

$$\mathbb{P}\left\{ \sup_{\theta \in \Theta_0(\mathtt{r})} \bigl\| D^{-1}\bigl\{ \nabla L(\theta) - \nabla L(\theta^*) \bigr\} + D(\theta - \theta^*) \bigr\| \ge \diamondsuit(\mathtt{r}, \mathtt{x}) \right\} \le \mathrm{e}^{-\mathtt{x}},$$

where

$$\diamondsuit(\mathtt{r}, \mathtt{x}) \stackrel{\text{def}}{=} \bigl\{ \delta(\mathtt{r}) + 6\nu_0 \, q_{\mathbb{H}}(\mathtt{x}) \, \omega \bigr\} \mathtt{r}. \tag{6.14}$$

The result of Proposition 6.5.1 can be extended to the differences $\mathcal{U}(\theta, \theta^\circ) = D^{-1}\bigl\{ \nabla \zeta(\theta) - \nabla \zeta(\theta^\circ) \bigr\}$: on a set of probability at least $1 - \mathrm{e}^{-\mathtt{x}}$, it holds for any $\theta, \theta^\circ \in \Theta_0(\mathtt{r})$ and $\chi(\theta, \theta^\circ) = D^{-1}\bigl\{ \nabla L(\theta) - \nabla L(\theta^\circ) \bigr\} + D\,(\theta - \theta^\circ)$

$$\begin{aligned} \mathbb{E}\bigl[\chi(\theta, \theta^\circ) \bigr] &\le \delta(\mathtt{r}) \, \|D(\theta - \theta^\circ)\| \le 2\mathtt{r} \, \delta(\mathtt{r}), \\ \|\mathcal{U}(\theta, \theta^\circ)\| &\le \|\mathcal{U}(\theta, \theta^*)\| + \|\mathcal{U}(\theta^\circ, \theta^*)\| \le 2\mathtt{r} \, \varrho(\mathtt{x}), \\ \bigl\| \chi(\theta, \theta^\circ) \bigr\| &\le 2 \diamondsuit(\mathtt{r}, \mathtt{x}). \end{aligned} \tag{6.15}$$

6.5.2 Local quadratic approximation of the log-likelihood

As the next step, we derive a uniform deviation bound on the error of a quadratic approximation $\mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^\circ) = (\boldsymbol{\theta}-\boldsymbol{\theta}^\circ)^\top \nabla L(\boldsymbol{\theta}^\circ) - \|D(\boldsymbol{\theta}-\boldsymbol{\theta}^\circ)\|^2/2$ of $L(\boldsymbol{\theta},\boldsymbol{\theta}^\circ)$:

$$\begin{aligned}\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^\circ) &\stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^\circ) - (\boldsymbol{\theta}-\boldsymbol{\theta}^\circ)^\top \nabla L(\boldsymbol{\theta}^\circ) + \frac{1}{2}\|D(\boldsymbol{\theta}-\boldsymbol{\theta}^\circ)\|^2 \\ &= L(\boldsymbol{\theta},\boldsymbol{\theta}^\circ) - \mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^\circ)\end{aligned}$$

in all $\boldsymbol{\theta},\boldsymbol{\theta}^\circ \in \Theta_0$, where Θ_0 is some vicinity of a fixed point $\boldsymbol{\theta}^*$. With $\boldsymbol{\theta}^\circ$ fixed, the gradient $\nabla\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^\circ) \stackrel{\text{def}}{=} \frac{d}{d\boldsymbol{\theta}}\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^\circ)$ fulfills

$$\nabla\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^\circ) = \nabla L(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta}^\circ) + D^2(\boldsymbol{\theta}-\boldsymbol{\theta}^\circ) = D\,\chi(\boldsymbol{\theta},\boldsymbol{\theta}^\circ);$$

cf. (6.10). This implies

$$\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^\circ) = (\boldsymbol{\theta}-\boldsymbol{\theta}^\circ)^\top \nabla\alpha(\boldsymbol{\theta}',\boldsymbol{\theta}^\circ),$$

where $\boldsymbol{\theta}'$ is a point on the line connecting $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\circ$. Further,

$$\bigl|\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^\circ)\bigr| = \bigl|(\boldsymbol{\theta}-\boldsymbol{\theta}^\circ)^\top DD^{-1}\nabla\alpha(\boldsymbol{\theta}',\boldsymbol{\theta}^\circ)\bigr| \le \|D(\boldsymbol{\theta}-\boldsymbol{\theta}^\circ)\| \sup_{\boldsymbol{\theta}'\in\Theta_0(\mathtt{r})} \bigl|\chi(\boldsymbol{\theta}',\boldsymbol{\theta}^\circ)\bigr|.$$

and one can apply (6.15). This yields the following result.

Proposition 6.5.2. *Suppose* $(\mathcal{L}_0)$, (ED_0), *and* (ED_2). *For each* $\mathtt{r}$, *it holds on a random set* $\Omega(\mathtt{r},\mathtt{x})$ *of a dominating probability at least* $1-\mathrm{e}^{-\mathtt{x}}$, *it holds with any* $\boldsymbol{\theta},\boldsymbol{\theta}^\circ \in \Theta_0(\mathtt{r})$

$$\begin{aligned}\frac{\bigl|\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^*)\bigr|}{\|D(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|} &\le \diamondsuit(\mathtt{r},\mathtt{x}), \qquad \bigl|\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^*)\bigr| \le \mathtt{r}\,\diamondsuit(\mathtt{r},\mathtt{x}), \\ \frac{\bigl|\alpha(\boldsymbol{\theta}^*,\boldsymbol{\theta})\bigr|}{\|D(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|} &\le 2\diamondsuit(\mathtt{r},\mathtt{x}), \qquad \bigl|\alpha(\boldsymbol{\theta}^*,\boldsymbol{\theta})\bigr| \le 2\mathtt{r}\,\diamondsuit(\mathtt{r},\mathtt{x}), \\ \frac{\bigl|\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^\circ)\bigr|}{\|D(\boldsymbol{\theta}-\boldsymbol{\theta}^\circ)\|} &\le 2\diamondsuit(\mathtt{r},\mathtt{x}), \qquad \bigl|\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^\circ)\bigr| \le 4\mathtt{r}\,\diamondsuit(\mathtt{r},\mathtt{x}),\end{aligned}$$

where $\diamondsuit(\mathtt{r},\mathtt{x})$ *is from* (6.14).

6.5.3 Proof of Theorem 6.3.1

By definition $\sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r}_0)} L(\boldsymbol{\theta},\boldsymbol{\theta}^*) \ge 0$. So, it suffices to check that $L(\boldsymbol{\theta},\boldsymbol{\theta}^*) < 0$ for all $\boldsymbol{\theta} \in \Theta \setminus \Theta_0(\mathtt{r}_0)$. The proof is based on the following bound: for each $\mathtt{r}$

$$I\!P\Bigl(\sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})} \bigl|\zeta(\boldsymbol{\theta},\boldsymbol{\theta}^*) - (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top\nabla\zeta(\boldsymbol{\theta}^*)\bigr| \ge 3\nu_0\, q_{\mathbb{H}}(\mathtt{x})\,\omega\,\mathtt{r}\Bigr) \le \mathrm{e}^{-\mathtt{x}}.$$

This bound is a special case of the general result from Theorem 6.5.1 below. It implies by Theorem 6.5.2 with $\rho = 1/2$ on a set $\Omega(\mathtt{x})$ of probability at least $1 - \mathrm{e}^{-\mathtt{x}}$ that for all $\mathtt{r} \geq \mathtt{r}_0$ and all $\boldsymbol{\theta}$ with $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathtt{r}$

$$\bigl|\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla \zeta(\boldsymbol{\theta}^*)\bigr| \leq \varrho(\mathtt{r}, \mathtt{x})\, \mathtt{r},$$

where

$$\varrho(\mathtt{r}, \mathtt{x}) = 6\nu_0\, q_{\mathbb{H}}\bigl(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0)\bigr)\, \omega\,.$$

The use of $\nabla I\!\!E L(\boldsymbol{\theta}^*) = 0$ yields

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \bigl|L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla L(\boldsymbol{\theta}^*)\bigr| \leq \varrho(\mathtt{r}, \mathtt{x})\, \mathtt{r}.$$

By Proposition 6.5.3, the vector $\boldsymbol{\xi} = D^{-1} \nabla \zeta(\boldsymbol{\theta}^*)$ fulfills $I\!\!P\bigl(\|\boldsymbol{\xi}\| \geq q(B, \mathtt{x})\bigr) \leq 2\mathrm{e}^{-\mathtt{x}}$. We ignore here the negligible term of order $\mathrm{e}^{-\mathtt{x}_c}$. The condition $\|\boldsymbol{\xi}\| \leq q(B, \mathtt{x})$ implies for $\mathtt{r} \geq \mathtt{r}_0$

$$\begin{aligned} &\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \bigl|(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla L(\boldsymbol{\theta}^*)\bigr| \\ &\quad \leq \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \times \|D^{-1} \nabla L(\boldsymbol{\theta}^*)\| = \mathtt{r} \|\boldsymbol{\xi}\| \leq q(B, \mathtt{x})\, \mathtt{r}. \end{aligned}$$

Condition $(\mathcal{L})$ implies $-2I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \geq \mathtt{r}^2 \mathtt{b}(\mathtt{r})$ for each $\boldsymbol{\theta}$ with $\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathtt{r}$. We conclude that the condition

$$\mathtt{r}\mathtt{b}(\mathtt{r}) \geq 2q(B, \mathtt{x}) + 2\varrho(\mathtt{r}, \mathtt{x}), \quad \mathtt{r} > \mathtt{r}_0,$$

ensure $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) < 0$ for all $\boldsymbol{\theta} \notin \Theta_0(\mathtt{r}_0)$ with a dominating probability.

6.5.4 Proof of Theorem 6.3.2

Let $\mathtt{r}_0$ be selected to ensure that $I\!\!P\bigl\{\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r}_0)\bigr\} \leq \mathrm{e}^{-\mathtt{x}}$. Furthermore, the definition of $\widetilde{\boldsymbol{\theta}}$ yields $\nabla L(\widetilde{\boldsymbol{\theta}}) = 0$ and

$$\chi(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = -D^{-1} \nabla L(\boldsymbol{\theta}^*) + D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*).$$

Now Proposition 6.5.1 implies on a set of a dominating probability

$$\|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi}\| \leq \diamondsuit(\mathtt{r}, \mathtt{x}) \tag{6.16}$$

and the assertion follows.

6.5.5 Proof of Theorem 6.3.3

We apply the result of Proposition 6.5.2 on a random set of dominating probability $1 - 2\mathrm{e}^{-\mathtt{x}}$ on which $\widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r}_0)$ and the inequalities from that proposition are fulfilled. For the special case with $\boldsymbol{\theta}^\circ = \widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ we obtain in view of $\nabla L(\widetilde{\boldsymbol{\theta}}) = 0$ that

$$\bigl| L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2/2 \bigr| \le \bigl| \alpha(\boldsymbol{\theta}^*, \widetilde{\boldsymbol{\theta}}) \bigr| \le 2\mathtt{r}_0 \diamondsuit(\mathtt{r}_0, \mathtt{x}). \tag{6.17}$$

Furthermore, with $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^\circ = \boldsymbol{\theta}^*$

$$\bigl| L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \boldsymbol{\xi}^\top D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + \|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2/2 \bigr| = \bigl| \alpha(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \bigr| \le \mathtt{r}_0 \diamondsuit(\mathtt{r}_0, \mathtt{x})$$

which implies

$$\bigl| L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\boldsymbol{\xi}\|^2/2 + \|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi}\|^2/2 \bigr| \le \mathtt{r}_0 \diamondsuit(\mathtt{r}_0, \mathtt{x}).$$

Now it follows by (6.16) that

$$\bigl| L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\boldsymbol{\xi}\|^2/2 \bigr| \le \mathtt{r}_0 \diamondsuit(\mathtt{r}_0, \mathtt{x}) + \diamondsuit^2(\mathtt{r}_0, \mathtt{x})/2.$$

For the squared root of the excess, (6.17) implies

$$\begin{aligned} \Bigl| \bigl\{ 2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \bigr\}^{1/2} - \|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \Bigr| &\le \frac{\bigl| 2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 \bigr|}{\|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|} \\ &\le \frac{2\bigl| \alpha(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \bigr|}{\|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|} \le \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0)} \frac{2\bigl| \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \bigr|}{\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|} \le 2 \diamondsuit(\mathtt{r}_0, \mathtt{x}). \end{aligned} \tag{6.18}$$

Similarly, for any $\boldsymbol{\theta}^\circ \in \Theta_0(\mathtt{r})$, it holds

$$\Bigl| \bigl\{ 2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^\circ) \bigr\}^{1/2} - \|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^\circ)\| \Bigr| \le \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0)} \frac{2\bigl| \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) \bigr|}{\|D(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\|} \le 4 \diamondsuit(\mathtt{r}_0, \mathtt{x}).$$

The Fisher expansion (6.16) allows to replace in (6.18) the norm of the standardized error $D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ with the norm of the normalized score $\boldsymbol{\xi}$. This completes the proof of Theorem 6.3.3.

6.5.6 An entropy bound for the maximum of a random process

We use one general result on the upper bound for the maximum of a centered random process in the form of Spokoiny (2012); see Corollary 7.2 in the supplement of that paper.

Here we discuss the special case when Υ is an open subset in $\mathbb{R}^p$, the stochastic process $\mathcal{U}(\boldsymbol{\upsilon})$ is absolutely continuous and its gradient $\nabla \mathcal{U}(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} d\mathcal{U}(\boldsymbol{\upsilon})/d\boldsymbol{\upsilon}$ has bounded exponential moments.

($\mathcal{ED}$) *There exist* $\mathtt{g} > 0$, $\nu_0 \geq 1$, *and a symmetric non-negative matrix* H_0 *such that for any* $\lambda \leq \mathtt{g}$ *and any unit vector* $\gamma \in I\!R^p$, *it holds*

$$\log I\!E \exp\Big\{ \lambda \frac{\gamma^\top \nabla \mathcal{U}(\upsilon)}{\|H_0 \gamma\|} \Big\} \leq \frac{\nu_0^2 \lambda^2}{2}.$$

We consider the local sets of the elliptic form $\Upsilon_\circ(\mathtt{r}) \stackrel{\text{def}}{=} \{\upsilon : \|H_0(\upsilon - \upsilon^*)\| \leq \mathtt{r}\}$.

Theorem 6.5.1 (Spokoiny (2012)). *Let* ($\mathcal{ED}$) *hold with some* $\mathtt{g} > 0$, *and a matrix* H_0. *For any* $\mathtt{x} \geq 0$ *and any* $\mathtt{r} > 0$

$$I\!P\Big\{ \sup_{\upsilon \in \Upsilon_\circ(\mathtt{r})} \big| \mathcal{U}(\upsilon, \upsilon^*) \big| \geq 3\nu_0 \mathtt{r}\, q_{\mathbb{H}}(\mathtt{x}) \Big\} \leq \mathrm{e}^{-\mathtt{x}},$$

where $q_{\mathbb{H}}(\mathtt{x})$ *is given by the following rule:*

$$q_{\mathbb{H}}(\mathtt{x}) = \begin{cases} \sqrt{\mathbb{H} + 2\mathtt{x}} & \text{if } \mathbb{H} + 2\mathtt{x} \leq \mathtt{g}^2, \\ \mathtt{g}^{-1}\mathtt{x} + \frac{1}{2}\big(\mathtt{g}^{-1}\mathbb{H} + \mathtt{g}\big) & \text{if } \mathbb{H} + 2\mathtt{x} > \mathtt{g}^2, \end{cases} \tag{6.19}$$

with $\mathbb{H} = 4p$.

Due to the result of Theorem 6.5.1, the bound for the maximum of $\mathcal{U}(\upsilon, \upsilon^*)$ over $\upsilon \in \mathcal{B}_{\mathtt{r}}(\upsilon^*)$ grows linearly in $\mathtt{r}$. So, its applications to situations with $\mathtt{r} \gg \mathbb{Q}_1(\Upsilon^\circ)$ are limited. The next result shows that introducing a negative drift helps to state a uniform in $\mathtt{r}$ local probability bound. Namely, the bound for the process $\mathcal{U}(\upsilon, \upsilon^*) - f(d(\upsilon, \upsilon^*))$ for some function $f(\mathtt{r})$ over a ball $\mathcal{B}_{\mathtt{r}}(\upsilon^*)$ around the point υ^* does not depend on $\mathtt{r}$. Here the generic chaining arguments are accomplished with the slicing technique. The idea is for a given $\mathtt{r}^* > 1$ to split the ball $\mathcal{B}_{\mathtt{r}^*}(\upsilon^*)$ into the slices $\mathcal{B}_{\mathtt{r}_k}(\upsilon^*) \setminus \mathcal{B}_{\mathtt{r}_{k-1}}(\upsilon^*)$ and to apply Theorem 6.5.1 to each slice separately.

Theorem 6.5.2. *Let* $\mathtt{r}^*$ *be such that* ($\mathcal{ED}$) *holds on* $\mathcal{B}_{\mathtt{r}^*}(\upsilon^*)$. *Given* $\mathtt{r}_0 < \mathtt{r}^*$, *let a monotonous function* $f(\mathtt{r}, \mathtt{r}_0)$ *fulfill*

$$f(\mathtt{r}, \mathtt{r}_0) \geq 3\nu_0 \mathtt{r}\, q_{\mathbb{H}}\big(\mathtt{x} + \log(\mathtt{r}/\mathtt{r}_0)\big), \quad \mathtt{r}_0 \leq \mathtt{r} \leq \mathtt{r}^*, \tag{6.20}$$

where the function $q_{\mathbb{H}}(\cdot)$ *is given by* (6.19). *Then it holds for any* $\rho < 1$

$$I\!P\Big(\sup_{\mathtt{r}_0 \leq \mathtt{r} \leq \mathtt{r}^*} \sup_{\upsilon \in \Upsilon_\circ(\mathtt{r})} \big\{ \mathcal{U}(\upsilon, \upsilon^*) - f\big(\rho^{-1}\mathtt{r}, \mathtt{r}_0\big) \big\} \geq 0 \Big) \leq \frac{\rho}{1-\rho} \mathrm{e}^{-\mathtt{x}}.$$

Remark 6.5.1. Formally the bound applies even with $\mathtt{r}^* = \infty$ provided that ($\mathcal{ED}$) is fulfilled on the whole set Υ°.

Remark 6.5.2. If $\mathtt{g} = \infty$, then $q_{\mathbb{H}}(\mathtt{x}) = \sqrt{2\mathtt{x} + 4p}$ and the condition (6.20) on the drift simplifies to $(3\nu_0 \mathtt{r})^{-1} f(\mathtt{r}, \mathtt{r}_0) \geq \sqrt{2\mathtt{x} + 4p + 2\log(\mathtt{r}/\mathtt{r}_0)}$.

Proof. By (6.20) and Theorem 6.5.1 for any $\mathtt{r} > \mathtt{r}_0$

$$\begin{aligned} &I\!P\Big(\sup_{\upsilon \in \mathcal{B}_{\mathtt{r}}(\upsilon^*) \setminus \mathcal{B}_{\rho \mathtt{r}}(\upsilon^*)} \big\{ \mathcal{U}(\upsilon, \upsilon^*) - f\big(\mathtt{r}, \mathtt{r}_0\big) \big\} \geq 0 \Big) \\ &\quad \leq I\!P\Big(\frac{1}{3\nu_0 \mathtt{r}} \sup_{\upsilon \in \mathcal{B}_{\mathtt{r}}(\upsilon^*)} \mathcal{U}(\upsilon, \upsilon^*) \geq q_{I\!\mathbb{H}}\big(\mathtt{x} + \log(\mathtt{r}/\mathtt{r}_0)\big) \Big) \leq \frac{\mathtt{r}_0}{\mathtt{r}} \mathrm{e}^{-\mathtt{x}}. \end{aligned} \tag{6.21}$$

Now define $\mathtt{r}_k = \mathtt{r}_0 \rho^{-k}$ for $k = 0, 1, 2, \ldots$. Define also $k^* \stackrel{\text{def}}{=} \log(\mathtt{r}^*/\mathtt{r}_0) + 1$. It follows from (6.21) that

$$\begin{aligned} &I\!P\Big(\sup_{\upsilon \in \mathcal{B}_{\mathtt{r}^*}(\upsilon^*) \setminus \mathcal{B}_{\mathtt{r}_0}(\upsilon^*)} \Big\{ \mathcal{U}(\upsilon, \upsilon^*) - f\big(\rho^{-1} d(\upsilon, \upsilon^*), \mathtt{r}_0\big) \Big\} \geq 0 \Big) \\ &\quad \leq \sum_{k=1}^{k^*} I\!P\Big(\frac{1}{\mathtt{r}_k} \sup_{\upsilon \in \mathcal{B}_{\mathtt{r}_k}(\upsilon^*) \setminus \mathcal{B}_{\mathtt{r}_{k-1}}(\upsilon^*)} \Big\{ \mathcal{U}(\upsilon, \upsilon^*) - f\big(\mathtt{r}_k, \mathtt{r}_0\big) \Big\} \geq 0 \Big) \\ &\quad \leq \mathrm{e}^{-\mathtt{x}} \sum_{k=1}^{k^*} \rho^k \leq \frac{\rho}{1-\rho} \mathrm{e}^{-\mathtt{x}} \end{aligned}$$

as required.

6.5.7 A bound for the norm of a vector random process

Let $\mathcal{Y}(\upsilon)$, $\upsilon \in \Upsilon$, be a smooth centered random vector process with values in $I\!R^q$, where $\Upsilon \subseteq I\!R^p$. Let also $\mathcal{Y}(\upsilon^*) = 0$ for a fixed point $\upsilon^* \in \Upsilon$. Without loss of generality assume $\upsilon^* = 0$. We aim to bound the maximum of the norm $\|\mathcal{Y}(\upsilon)\|$ over a vicinity $\Upsilon_\circ$ of υ^*. Suppose that $\mathcal{Y}(\upsilon)$ satisfies for each $\gamma \in I\!R^p$ and $\alpha \in I\!R^q$ with $\|\gamma\| = \|\alpha\| = 1$

$$\sup_{\upsilon \in \Upsilon} \log I\!E \exp\Big\{ \lambda \gamma^\top \nabla \mathcal{Y}(\upsilon) \alpha \Big\} \leq \frac{\nu_0^2 \lambda^2}{2}, \qquad \lambda^2 \leq 2\mathtt{g}^2. \tag{6.22}$$

Condition (6.22) implies for any $\upsilon \in \Upsilon_\circ$ with $\|\upsilon\| \leq \mathtt{r}$ and $\|\gamma\| = 1$ in view of $\mathcal{Y}(\upsilon^*) = 0$

$$\log I\!E \exp\Big\{ \frac{\lambda}{\mathtt{r}} \gamma^\top \mathcal{Y}(\upsilon) \Big\} \leq \frac{\nu_0^2 \lambda^2 \|\upsilon\|^2}{2\mathtt{r}^2}, \qquad \lambda^2 \leq 2\mathtt{g}^2; \tag{6.23}$$

In what follows, we use the representation

$$\|\mathcal{Y}(\upsilon)\| = \sup_{\|u\| \leq \mathtt{r}} \frac{1}{\mathtt{r}} u^\top \mathcal{Y}(\upsilon).$$

This implies for $\Upsilon_\circ(\mathtt{r}) = \big\{ \upsilon \in \Upsilon \colon \|\upsilon - \upsilon^*\| \leq \mathtt{r} \big\}$

$$\sup_{\upsilon \in \Upsilon_\circ(\mathtt{r})} \|\mathcal{Y}(\upsilon)\| = \sup_{\upsilon \in \Upsilon_\circ(\mathtt{r})} \sup_{\|u\| \leq \mathtt{r}} \frac{1}{\mathtt{r}} u^\top \mathcal{Y}(\upsilon).$$

Consider a bivariate process $\boldsymbol{u}^\top \mathcal{Y}(\boldsymbol{v})$ of $\boldsymbol{u} \in I\!R^q$ and $\boldsymbol{v} \in \Upsilon \subset I\!R^p$. By definition $I\!E \boldsymbol{u}^\top \mathcal{Y}(\boldsymbol{v}) = 0$. Further, $\nabla_{\boldsymbol{u}} \bigl[\boldsymbol{u}^\top \mathcal{Y}(\boldsymbol{v})\bigr] = \mathcal{Y}(\boldsymbol{v})$ while $\nabla_{\boldsymbol{v}} \bigl[\boldsymbol{u}^\top \mathcal{Y}(\boldsymbol{v})\bigr] = \boldsymbol{u}^\top \nabla \mathcal{Y}(\boldsymbol{v}) = \|\boldsymbol{u}\| \boldsymbol{\gamma}^\top \nabla \mathcal{Y}(\boldsymbol{v})$ for $\boldsymbol{\gamma} = \boldsymbol{u}/\|\boldsymbol{u}\|$. Suppose that $\boldsymbol{u} \in I\!R^q$ and $\boldsymbol{v} \in \Upsilon$ are such that $\|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 \leq 2\mathtt{r}^2$. By the Hölder inequality, (6.23), and (6.22), it holds for $\|\boldsymbol{\gamma}\| = \|\boldsymbol{\alpha}\| = 1$ and $\boldsymbol{v} \in \Upsilon_\circ(\mathtt{r})$

$$\begin{aligned}
&\log I\!E \exp\biggl\{ \frac{\lambda}{2\mathtt{r}} (\boldsymbol{\gamma}, \boldsymbol{\alpha})^\top \nabla \bigl[\boldsymbol{u}^\top \mathcal{Y}(\boldsymbol{v})\bigr] \biggr\} \\
&\quad \leq \frac{1}{2} \log I\!E \exp\biggl\{ \frac{\lambda}{\mathtt{r}} \boldsymbol{\gamma}^\top \mathcal{Y}(\boldsymbol{v}) \biggr\} + \frac{1}{2} \log I\!E \exp\biggl\{ \frac{\lambda}{\mathtt{r}} \boldsymbol{u}^\top \nabla \mathcal{Y}(\boldsymbol{v}) \boldsymbol{\alpha} \biggr\} \\
&\quad \leq \frac{1}{2} \log I\!E \exp\biggl\{ \frac{\lambda}{\mathtt{r}} \boldsymbol{\gamma}^\top \mathcal{Y}(\boldsymbol{v}) \biggr\} + \frac{1}{2} \log I\!E \exp\biggl\{ \frac{\lambda}{\mathtt{r}} \|\boldsymbol{u}\| \boldsymbol{\gamma}^\top \nabla \mathcal{Y}(\boldsymbol{v}) \boldsymbol{\alpha} \biggr\} \\
&\quad \leq \frac{\nu_0^2 \lambda^2}{4\mathtt{r}^2} \bigl(\|\boldsymbol{v}\|^2 + \|\boldsymbol{u}\|^2 \bigr) \leq \frac{\nu_0^2 \lambda^2}{2}, \qquad |\lambda| \leq \mathtt{g}.
\end{aligned}$$

We summarize our findings in the following theorem.

Theorem 6.5.3. *Let a random p-vector process $\mathcal{Y}(\boldsymbol{v})$ for $\boldsymbol{v} \in \Upsilon \subseteq I\!R^p$ fulfill $\mathcal{Y}(\boldsymbol{v}^*) = 0$, $I\!E \mathcal{Y}(\boldsymbol{v}) \equiv 0$, and the condition (6.22) be satisfied. Then for each $\mathtt{r}$ and any $\mathtt{x} \geq 1/2$, it holds*

$$I\!P\Bigl\{ \sup_{\boldsymbol{v} \in \Upsilon_\circ(\mathtt{r})} \bigl\| \mathcal{Y}(\boldsymbol{v}) \bigr\| > 6\nu_0 \mathtt{r}\, q_{\mathbb{H}}(\mathtt{x}) \Bigr\} \leq \mathrm{e}^{-\mathtt{x}},$$

where $q_{\mathbb{H}}(\mathtt{x})$ is given by (6.19).

Proof. The results follow from Theorem 6.5.1 applied to the process $\boldsymbol{u}^\top \mathcal{Y}(\boldsymbol{v})/2$ of the variable $(\boldsymbol{u}, \boldsymbol{v}) \in I\!R^{p+q}$.

6.5.8 A deviation bound for the quadratic form $\|\boldsymbol{\xi}\|^2$

This section presents a bound for a quadratic form $\|\boldsymbol{\xi}\|^2$ where $\boldsymbol{\xi} = D^{-1} \nabla \zeta(\boldsymbol{\theta}^*)$. The result only uses the condition (ED_0) which we restate in a slightly different form. For $B = D^{-1} V^2 D^{-1}$, define

$$\mathtt{p}_B \stackrel{\text{def}}{=} \operatorname{tr}\bigl(B\bigr), \qquad \mathtt{v}_B^2 \stackrel{\text{def}}{=} 2 \operatorname{tr}(B^2), \qquad \lambda_B \stackrel{\text{def}}{=} \lambda_{\max}\bigl(B\bigr).$$

Note that $\mathtt{p}_B = I\!E \|\boldsymbol{\xi}\|^2$. Moreover, if $\boldsymbol{\xi}$ is a Gaussian vector then $\mathtt{v}_B^2 = \operatorname{Var}\bigl(\|\boldsymbol{\xi}\|^2\bigr)$. If $V^2 = D^2$, then $\lambda_B = 1$. The condition (ED_0) means that the vector $V^{-1} \nabla \zeta(\boldsymbol{\theta}^*)$ fulfills the following exponential moment condition:

$$\log I\!E \exp\bigl(\boldsymbol{\gamma}^\top V^{-1} \nabla \zeta(\boldsymbol{\theta}^*) \bigr) \leq \|\boldsymbol{\gamma}\|^2/2, \qquad \boldsymbol{\gamma} \in I\!R^p,\ \|\boldsymbol{\gamma}\| \leq \mathtt{g}.$$

Here ν_0 is set to one. Spokoiny (2012) argued how the case of any $\nu_0 \geq 1$ can be reduced to $\nu_0 \approx 1$ by a slight change of scale and reducing the value $\mathtt{g}$ which is typically large. For ease of presentation, suppose that $\mathtt{g}^2 \geq 2\mathtt{p}_B$. The other case only changes the constants in the inequalities. Note that $\|\boldsymbol{\xi}\|^2 = \boldsymbol{\eta}^\top B\,\boldsymbol{\eta}$. Define $\mu_c = 2/3$ and

$$\begin{aligned} \mathtt{g}_c &\stackrel{\text{def}}{=} \sqrt{\mathtt{g}^2 - \mu_c \mathtt{p}_B}, \\ 2\mathtt{x}_c &\stackrel{\text{def}}{=} (\mathtt{g}^2/\mu_c - \mathtt{p}_B)/\lambda_B + \log\det\bigl(I_p - \mu_c B/\lambda_B\bigr). \end{aligned} \tag{6.24}$$

Proposition 6.5.3 (Spokoiny (2012)). *Let* (ED_0) *hold with* $\nu_0 = 1$ *and* $\mathtt{g}^2 \geq 2\mathtt{p}_B$. *Then for each* $\mathtt{x} > 0$

$$I\!P\bigl(\|\boldsymbol{\xi}\| \geq q(B,\mathtt{x})\bigr) \leq 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c},$$

where $q(B,\mathtt{x})$ *is defined by*

$$q^2(B,\mathtt{x}) \stackrel{\text{def}}{=} \begin{cases} \mathtt{p}_B + 2\mathtt{v}_B\mathtt{x}^{1/2}, & \mathtt{x} \leq \mathtt{v}_B/(18\lambda_B), \\ \mathtt{p}_B + 6\lambda_B\mathtt{x}, & \mathtt{v}_B/(18\lambda_B) < \mathtt{x} \leq \mathtt{x}_c, \\ \bigl|\mathtt{y}_c + 2\lambda_B(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c\bigr|^2, & \mathtt{x} > \mathtt{x}_c. \end{cases} \tag{6.25}$$

with $\mathtt{y}_c^2 \leq \mathtt{p}_B + 6\lambda_B\mathtt{x}_c$.

Depending on the value $\mathtt{x}$, we observe three types of tail behavior of the quadratic form $\|\boldsymbol{\xi}\|^2$. The sub-Gaussian regime for $\mathtt{x} \leq \mathtt{v}_B/(18\lambda_B)$ and the Poissonian regime for $\mathtt{x} \leq \mathtt{x}_c$ are similar to the case of a Gaussian quadratic form. The value $\mathtt{x}_c$ from (6.24) is of order $\mathtt{g}^2$. In all our results we suppose that $\mathtt{g}^2$ and hence, $\mathtt{x}_c$ is sufficiently large and the quadratic form $\|\boldsymbol{\xi}\|^2$ can be bounded with a dominating probability by $\mathtt{p}_B + 6\lambda_B\mathtt{x}$ for a proper $\mathtt{x}$. We refer to Spokoiny (2012) for the proof of this and related results, further discussion and references.

6.6 Some results 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.

6.6.1 Deviation bounds

This section collects some deviation bounds on the norm or quadratic form of a standard normal vector. Everywhere in this section $\boldsymbol{\gamma}$ means a standard normal vector in $I\!R^p$.

Lemma 6.6.1. *Let* $\mu \in (0,1)$. *Then for any vector* $\boldsymbol{\lambda} \in I\!R^p$ *with* $\|\boldsymbol{\lambda}\|^2 \le p$ *and any* $\mathtt{r} > 0$

$$\log I\!E\bigl\{\exp(\boldsymbol{\lambda}^\top \boldsymbol{\gamma})\, 1\!\!1\bigl(\|\boldsymbol{\gamma}\| > \mathtt{r}\bigr)\bigr\} \le -\frac{1-\mu}{2}\mathtt{r}^2 + \frac{1}{2\mu}\|\boldsymbol{\lambda}\|^2 + \frac{p}{2}\log(\mu^{-1}). \tag{6.26}$$

Moreover, if $\mathtt{r}^2 \ge 6p + 4\mathtt{x}$, *then*

$$I\!E\Bigl\{\exp(\boldsymbol{\lambda}^\top \boldsymbol{\gamma})\, 1\!\!1\bigl(\|\boldsymbol{\gamma}\| \le \mathtt{r}\bigr)\Bigr\} \ge \mathrm{e}^{\|\boldsymbol{\lambda}\|^2/2}\bigl(1 - \mathrm{e}^{-\mathtt{x}}\bigr). \tag{6.27}$$

Proof. We use that for $\mu < 1$

$$I\!E\bigl\{\exp(\boldsymbol{\lambda}^\top \boldsymbol{\gamma})\, 1\!\!1\bigl(\|\boldsymbol{\gamma}\| > \mathtt{r}\bigr)\bigr\} \le \mathrm{e}^{-(1-\mu)\mathtt{r}^2/2} I\!E \exp\bigl\{\boldsymbol{\lambda}^\top \boldsymbol{\gamma} + (1-\mu)\|\boldsymbol{\gamma}\|^2/2\bigr\}.$$

It holds

$$\begin{aligned} I\!E \exp\bigl\{\boldsymbol{\lambda}^\top \boldsymbol{\gamma} + (1-\mu)\|\boldsymbol{\gamma}\|^2/2\bigr\} &= (2\pi)^{-p/2} \int \exp\bigl\{\boldsymbol{\lambda}^\top \boldsymbol{\gamma} - \mu\|\boldsymbol{\gamma}\|^2/2\bigr\} d\boldsymbol{\gamma} \\ &= \mu^{-p/2} \exp\bigl(\mu^{-1}\|\boldsymbol{\lambda}\|^2/2\bigr) \end{aligned}$$

and (6.26) follows.

Now we apply this result with $\mu = 1/2$. In view of $I\!E \exp(\boldsymbol{\lambda}^\top \boldsymbol{\gamma}) = \mathrm{e}^{\|\boldsymbol{\lambda}\|^2/2}$, $\mathtt{r}^2 \ge 6p + 4\mathtt{x}$, and $2 + \log(2) < 3$, it follows for $\|\boldsymbol{\lambda}\|^2 \le p$

$$\begin{aligned} &\mathrm{e}^{-\|\boldsymbol{\lambda}\|^2/2} I\!E\bigl\{\exp(\boldsymbol{\lambda}^\top \boldsymbol{\gamma})\, 1\!\!1\bigl(\|\boldsymbol{\gamma}\| \le \mathtt{r}\bigr)\bigr\} \\ &\quad \ge 1 - \exp\bigl(-\mathtt{r}^2/4 + p + (p/2)\log(2)\bigr) \ge 1 - \exp(-\mathtt{x}) \end{aligned}$$

which implies (6.27).

Lemma 6.6.2. *For any* $\boldsymbol{u} \in I\!R^p$, *any unit vector* $\boldsymbol{a} \in I\!R^p$, *and any* $q > 0$, *it holds*

$$I\!P\bigl(\|\boldsymbol{\gamma} - \boldsymbol{u}\| \ge q\bigr) \le \exp\bigl\{-q^2/4 + p/2 + \|\boldsymbol{u}\|^2/2\bigr\}, \tag{6.28}$$

$$I\!E\bigl\{|\boldsymbol{\gamma}^\top \boldsymbol{a}|^2\, 1\!\!1\bigl(\|\boldsymbol{\gamma} - \boldsymbol{u}\| \ge q\bigr)\bigr\} \le (2 + |\boldsymbol{u}^\top \boldsymbol{a}|^2) \exp\bigl\{-q^2/4 + p/2 + \|\boldsymbol{u}\|^2/2\bigr\}. \tag{6.29}$$

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

$$\begin{aligned} I\!P\bigl(\|\boldsymbol{\gamma} - \boldsymbol{u}\| \ge q\bigr) &\le \exp\bigl(-\lambda q^2/2\bigr) I\!E \exp\bigl(\lambda\|\boldsymbol{\gamma} - \boldsymbol{u}\|^2/2\bigr) \\ &= \exp\Bigl\{-\frac{\lambda q^2}{2} - \frac{p}{2}\log(1-\lambda) + \frac{\lambda}{2(1-\lambda)}\|\boldsymbol{u}\|^2\Bigr\}. \end{aligned}$$

In particular, with $\lambda = 1/2$, this implies (6.28). Further, for $\|\boldsymbol{a}\| = 1$

$$\begin{aligned} I\!E\bigl\{|\boldsymbol{\gamma}^\top \boldsymbol{a}|^2\, 1\!\!1(\|\boldsymbol{\gamma} - \boldsymbol{u}\| \ge q)\bigr\} &\le \exp\bigl(-q^2/4\bigr) I\!E\bigl\{|\boldsymbol{\gamma}^\top \boldsymbol{a}|^2 \exp\bigl(\|\boldsymbol{\gamma} - \boldsymbol{u}\|^2/4\bigr)\bigr\} \\ &\le (2 + |\boldsymbol{u}^\top \boldsymbol{a}|^2) \exp\bigl(-q^2/4 + p/2 + \|\boldsymbol{u}\|^2/2\bigr) \end{aligned}$$

and (6.29) follows.

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

Lemma 6.6.3. *For each* $\mathtt{x}$,

$$\mathbb{P}\bigl(\|\boldsymbol{\gamma}\| \geq q(p,\mathtt{x})\bigr) \leq \exp\bigl(-\mathtt{x}\bigr), \quad \mathbb{P}\bigl(\|\boldsymbol{\gamma}\| \leq q_1(p,\mathtt{x})\bigr) \leq \exp\bigl(-\mathtt{x}\bigr), \tag{6.30}$$

where

$$q^2(p,\mathtt{x}) \stackrel{\text{def}}{=} p + \sqrt{6.6p\mathtt{x}} \vee (6.6\mathtt{x}), \qquad q_1^2(p,\mathtt{x}) \stackrel{\text{def}}{=} p - 2\sqrt{p\mathtt{x}}.$$

Proof. The exponential Chebyshev inequality implies for any positive $\lambda \leq 1/2$ in view of $-\log(1-\lambda) \leq \lambda + \lambda^2$

$$\begin{aligned}
\mathbb{P}\bigl(\|\boldsymbol{\gamma}\|^2 > p + \sqrt{2p\mathtt{x}}\bigr) &\leq \exp\Bigl\{-\frac{\lambda}{2}\bigl(p + \sqrt{2p\mathtt{x}}\bigr)\Bigr\}\mathbb{E}\exp\Bigl(\frac{\lambda}{2}\|\boldsymbol{\gamma}\|^2\Bigr) \\
&\leq \exp\Bigl\{-\frac{\lambda}{2}\bigl(p + \sqrt{2p\mathtt{x}}\bigr)\Bigr\}\mathbb{E}\exp\Bigl(\frac{\lambda}{2}\|\boldsymbol{\gamma}\|^2\Bigr) \\
&= \exp\Bigl\{-\frac{\lambda}{2}\bigl(p + \sqrt{2p\mathtt{x}}\bigr) - \frac{p}{2}\log(1-\lambda)\Bigr\} \\
&\leq \exp\bigl\{-\lambda\sqrt{p\mathtt{x}/2} + p\lambda^2/2\bigr\}.
\end{aligned}$$

The choice $\lambda = \sqrt{\mathtt{x}/(2p)}$ yields the result (6.30). Similarly

$$\begin{aligned}
\mathbb{P}\bigl(\|\boldsymbol{\gamma}\|^2 - p < -\sqrt{2p\mathtt{x}}\bigr) &\leq \exp\Bigl\{-\frac{\lambda}{2}\bigl(\sqrt{2p\mathtt{x}} - p\bigr)\Bigr\}\mathbb{E}\exp\Bigl(-\frac{\lambda}{2}\|\boldsymbol{\gamma}\|^2\Bigr) \\
&\leq \exp\Bigl\{-\frac{\lambda}{2}\bigl(\sqrt{2p\mathtt{x}} - p\bigr)\Bigr\}\mathbb{E}\exp\Bigl(-\frac{\lambda}{2}\|\boldsymbol{\gamma}\|^2\Bigr) \\
&= \exp\Bigl\{-\frac{\lambda}{2}\bigl(\sqrt{2p\mathtt{x}} - p\bigr) - \frac{p}{2}\log(1+\lambda)\Bigr\} \\
&\leq \exp\bigl\{-\lambda\sqrt{p\mathtt{x}/2} + p\lambda^2/4\bigr\}.
\end{aligned}$$

Here the choice $\lambda = \sqrt{2\mathtt{x}/p}$ does the job.

A similar bound can be obtained for a norm of the vector $B\boldsymbol{\xi}$ where B is some given deterministic matrix. For notational simplicity we assume that B is symmetric. Otherwise one should replace it with $(B^\top B)^{1/2}$.

Theorem 6.6.1. *Let* $\boldsymbol{\xi}$ *be standard normal in* $\mathbb{R}^p$. *Then for every* $\mathtt{x} > 0$ *and any symmetric matrix* B, *it holds with* $\mathtt{p} = \operatorname{tr}(B^2)$, $\mathtt{v}^2 = 2\operatorname{tr}(B^4)$, *and* $a^* = \|B^2\|_\infty$

$$\mathbb{P}\bigl(\|B\boldsymbol{\xi}\|^2 > \mathtt{p} + (2\mathtt{v}\mathtt{x}^{1/2}) \vee (6a^*\mathtt{x})\bigr) \leq \exp(-\mathtt{x}).$$

Proof. The matrix B^2 can be represented as $U^\top \operatorname{diag}(a_1, \ldots, a_p) U$ for an orthogonal matrix U. The vector $\widetilde{\boldsymbol{\xi}} = U\boldsymbol{\xi}$ is also standard normal and $\|B\boldsymbol{\xi}\|^2 = \widetilde{\boldsymbol{\xi}}^\top U B^2 U^\top \widetilde{\boldsymbol{\xi}}$. This means that one can reduce the situation to the case of a diagonal matrix $B^2 = \operatorname{diag}(a_1, \ldots, a_p)$. We can also assume without loss of generality that $a_1 \geq a_2 \geq \ldots \geq a_p$. The expressions for the quantities $\mathtt{p}$ and $\mathtt{v}^2$ simplifies to

$$\begin{aligned} \mathtt{p} &= \operatorname{tr}(B^2) = a_1 + \ldots + a_p, \\ \mathtt{v}^2 &= 2\operatorname{tr}(B^4) = 2(a_1^2 + \ldots + a_p^2). \end{aligned}$$

Moreover, rescaling the matrix B^2 by a_1 reduces the situation to the case with $a_1 = 1$.

Lemma 6.6.4. *It holds*

$$I\!\!E \|B\boldsymbol{\xi}\|^2 = \operatorname{tr}(B^2), \qquad \operatorname{Var}\bigl(\|B\boldsymbol{\xi}\|^2\bigr) = 2\operatorname{tr}(B^4).$$

Moreover, for $\mu < 1$

$$I\!\!E \exp\bigl\{\mu \|B\boldsymbol{\xi}\|^2/2\bigr\} = \det(1-\mu B^2)^{-1/2} = \prod_{i=1}^{p} (1-\mu a_i)^{-1/2}. \tag{6.31}$$

Proof. If B^2 is diagonal, then $\|B\boldsymbol{\xi}\|^2 = \sum_i a_i \xi_i^2$ and the summands $a_i\xi_i^2$ are independent. It remains to note that $I\!\!E(a_i\xi_i^2) = a_i$, $\operatorname{Var}(a_i\xi_i^2) = 2a_i^2$, and for $\mu a_i < 1$,

$$I\!\!E \exp\bigl\{\mu a_i \xi_i^2/2\bigr\} = (1-\mu a_i)^{-1/2}$$

yielding (6.31).

Given u, fix $\mu < 1$. The exponential Markov inequality yields

$$\begin{aligned} I\!\!P\bigl(\|B\boldsymbol{\xi}\|^2 > \mathtt{p} + u\bigr) &\leq \exp\Bigl\{-\frac{\mu(\mathtt{p}+u)}{2}\Bigr\} I\!\!E \exp\Bigl(\frac{\mu\|B\boldsymbol{\xi}\|^2}{2}\Bigr) \\ &\leq \exp\Bigl\{-\frac{\mu u}{2} - \frac{1}{2}\sum_{i=1}^{p}\bigl[\mu a_i + \log(1-\mu a_i)\bigr]\Bigr\}. \end{aligned}$$

We start with the case when $\mathtt{x}^{1/2} \leq \mathtt{v}/3$. Then $u = 2\mathtt{x}^{1/2}\mathtt{v}$ fulfills $u \leq 2\mathtt{v}^2/3$. Define $\mu = u/\mathtt{v}^2 \leq 2/3$ and use that $t + \log(1-t) \geq -t^2$ for $t \leq 2/3$. This implies

$$\begin{aligned} &I\!\!P\bigl(\|B\boldsymbol{\xi}\|^2 > \mathtt{p} + u\bigr) \\ &\quad\leq \exp\Bigl\{-\frac{\mu u}{2} + \frac{1}{2}\sum_{i=1}^{p}\mu^2 a_i^2\Bigr\} = \exp\bigl(-u^2/(4\mathtt{v}^2)\bigr) = e^{-\mathtt{x}}. \end{aligned} \tag{6.32}$$

Next, let $\mathtt{x}^{1/2} > \mathtt{v}/3$. Set $\mu = 2/3$. It holds similarly to the above

$$\sum_{i=1}^{p}\bigl[\mu a_i + \log\bigl(1-\mu a_i\bigr)\bigr] \geq -\sum_{i=1}^{p}\mu^2 a_i^2 \geq -2\mathtt{v}^2/9 \geq -2\mathtt{x}.$$

Now, for $u = 6\mathtt{x}$ and $\mu u/2 = 2\mathtt{x}$, (6.32) implies

$$I\!P\bigl(\|B\boldsymbol{\xi}\|^2 > \mathtt{p} + u\bigr) \leq \exp\bigl\{-\bigl(2\mathtt{x} - \mathtt{x}\bigr)\bigr\} = \exp(-\mathtt{x})$$

as required.

6.6.2 Gaussian comparison via KL-divergence and Pinsker's inequality

Suppose that two p-dimensional zero mean Gaussian vectors $\boldsymbol{\xi} \sim \mathcal{N}(0, \Sigma)$ and $\boldsymbol{\xi}^\flat \sim \mathcal{N}(0, \Sigma^\flat)$ are given. Let also T map $I\!R^p$ to $I\!R^{\mathbb{M}}$ and $\boldsymbol{X} = T(\boldsymbol{\xi})$ and $\boldsymbol{Y} = T(\boldsymbol{\xi}^\flat)$. We aim to bound the distance between distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ under the conditions

$$\|\Sigma^{-1/2}\Sigma^\flat\Sigma^{-1/2} - \boldsymbol{I}_p\| \leq \boldsymbol{\epsilon} \leq 1/2, \quad \operatorname{tr}\bigl(\Sigma^{-1/2}\Sigma^\flat\Sigma^{-1/2} - \boldsymbol{I}_p\bigr)^2 \leq \square^2 \tag{6.33}$$

for some $\boldsymbol{\epsilon} \leq 1/2$ and $\square \geq 0$. The next lemma bounds from above the Kullback-Leibler divergence between two normal distributions.

Lemma 6.6.5. *Let $I\!P_0 = \mathcal{N}(\boldsymbol{b}, \Sigma)$ and $I\!P_1 = \mathcal{N}(\boldsymbol{b}^\flat, \Sigma^\flat)$ some non-degenerated matrices Σ and $\Sigma^\flat$. If*

$$\|\Sigma^{-1/2}\Sigma^\flat\Sigma^{-1/2} - \boldsymbol{I}_p\| \leq 1/2, \qquad \operatorname{tr}\Bigl\{\bigl(\Sigma^{-1/2}\Sigma^\flat\Sigma^{-1/2} - \boldsymbol{I}_p\bigr)^2\Bigr\} \leq \square^2,$$

then

$$\mathcal{K}(I\!P_0, I\!P_1) = -I\!E_0 \log\frac{dI\!P_1}{dI\!P_0} \leq \frac{\square^2}{2} + \frac{1}{2}(\boldsymbol{b} - \boldsymbol{b}^\flat)^\top\Sigma^\flat(\boldsymbol{b} - \boldsymbol{b}^\flat).$$

For any measurable set $A \subset I\!R^p$, it holds

$$\bigl|I\!P_0(A) - I\!P_1(A)\bigr| \leq \sqrt{\mathcal{K}(I\!P_0, I\!P_1)/2}.$$

Proof. The change of variables $\boldsymbol{u} = \Sigma^{-1/2}(\boldsymbol{x} - \boldsymbol{b})$ reduces the general case to the situation when $I\!P_0$ is standard normal in $I\!R^p$ while $P_1 = \mathcal{N}(\boldsymbol{\beta}, B)$ with $\boldsymbol{\beta} = \Sigma^{1/2}(\boldsymbol{b}^\flat - \boldsymbol{b})$ and $B \stackrel{\text{def}}{=} \Sigma^{-1/2}\Sigma^\flat\Sigma^{-1/2}$

$$2\log\frac{dI\!P_1}{dI\!P_0}(\boldsymbol{\gamma}) = \log\det(B) - (\boldsymbol{\gamma} - \boldsymbol{\beta})^\top B(\boldsymbol{\gamma} - \boldsymbol{\beta}) + \|\boldsymbol{\gamma}\|^2$$

with $\boldsymbol{\gamma}$ standard normal and

$$2\mathcal{K}(I\!P_0, I\!P_1) = -2I\!E_0 \log\frac{dI\!P_1}{dI\!P_0} = -\log\det(B) + \operatorname{tr}(B - \boldsymbol{I}_p) + \boldsymbol{\beta}^\top B\boldsymbol{\beta}. \tag{6.34}$$

Let a_j be the j th eigenvalue of $B - I_p$. The condition $\|B - I_p\| \le 1/2$ yields $|a_j| \le 1/2$ and

$$\begin{aligned} 2\mathcal{K}(I\!P_0, I\!P_1) &= \boldsymbol{\beta}^\top B\boldsymbol{\beta} + \sum_{j=1}^{p} \bigl\{ a_j - \log(1+a_j) \bigr\} \\ &\le \boldsymbol{\beta}^\top B\boldsymbol{\beta} + \sum_{j=1}^{p} a_j^2 \\ &\le \boldsymbol{\beta}^\top B\boldsymbol{\beta} + \operatorname{tr}(B - I_p)^2 \le \boldsymbol{\beta}^\top B\boldsymbol{\beta} + \square^2. \end{aligned}$$

This implies by Pinsker's inequality

$$\sup_A |I\!P_0(A) - I\!P_1(A)| \le \sqrt{\frac{1}{2}\mathcal{K}(I\!P_0, I\!P_1)} \le \frac{1}{2}\sqrt{\square^2 + \boldsymbol{\beta}^\top B\boldsymbol{\beta}} \tag{6.35}$$

as required.

Theorem 6.6.2. *Let two p-dimensional zero mean Gaussian vectors $\boldsymbol{\xi} \sim \mathcal{N}(0, \Sigma)$ and $\boldsymbol{\xi}^\flat \sim \mathcal{N}(0, \Sigma^\flat)$ be given, and (6.33) holds. Then for any mapping $T \colon I\!R^p \to I\!R^{\mathbb{M}}$ and any set of values $(q_{\boldsymbol{\eta}})$, the random vectors $\boldsymbol{X} = T(\boldsymbol{\xi})$ and $\boldsymbol{Y} = T(\boldsymbol{\xi}^\flat)$ fulfill*

$$\bigl| I\!P\bigl(\max_{\boldsymbol{\eta}} X_{\boldsymbol{\eta}} - q_{\boldsymbol{\eta}} > 0\bigr) - I\!P\bigl(\max_{\boldsymbol{\eta}} Y_{\boldsymbol{\eta}} - q_{\boldsymbol{\eta}} > 0\bigr) \bigr| \le \square/2.$$

Proof. We simply apply the result of the lemma to the set $A = \{\boldsymbol{x} \in I\!R^p \colon T(\boldsymbol{x}) \le \boldsymbol{z}\}$.

As a corollary, for the special case with $\boldsymbol{\beta} \equiv 0$, we bound for any Borel set $A \subset I\!R^{\mathbb{M}}$

$$\bigl| I\!P\bigl(T(\boldsymbol{\xi}) \in A\bigr) - I\!P\bigl(T(\boldsymbol{\xi}^\flat) \in A\bigr) \bigr| \le \square/2\,.$$

Notice that the operator norm bound

$$\|\Sigma^{-1/2}\Sigma^\flat\Sigma^{-1/2} - I_p\| \le \boldsymbol{\epsilon}$$

implies for $B = \Sigma^{-1/2}\Sigma^\flat\Sigma^{-1/2}$

$$\operatorname{tr}\bigl(B - I_p\bigr)^2 \le p\boldsymbol{\epsilon}^2, \qquad \boldsymbol{\beta}^\top B\boldsymbol{\beta} \le (1+\boldsymbol{\epsilon})\|\boldsymbol{\beta}\|^2.$$

Interestingly, this method can be used for obtaining an anti-concentration bound in the case of a homogeneous mapping $T : I\!R^p \to I\!R^{\mathbb{M}}$.

Theorem 6.6.3. *Let $\boldsymbol{\xi} \sim \mathcal{N}(0, \Sigma)$ be a Gaussian vector in $I\!R^p$. For any homogeneous mapping $T : I\!R^p \to I\!R^{\mathbb{M}}$, and for any $q > 0$ and Δ satisfying $0 \le \Delta/q \le 1$, it holds*

$$I\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \ge q\bigr) - I\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \ge q + \Delta\bigr) \le \Delta\, q^{-1}\sqrt{p/2}.$$

Moreover, if $\boldsymbol{\xi}^\flat \sim \mathcal{N}(0, \Sigma^\flat)$ *is another Gaussian vector and* (6.33) *holds with* $\boldsymbol{\epsilon} \leq 1/2$ *and some* $\square \geq 0$, *then*

$$\bigl| I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \geq q\bigr) - I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}^\flat) \geq q + \Delta\bigr) \bigr| \leq \square/2 + \Delta\, q^{-1}\sqrt{p/2}. \tag{6.36}$$

Proof. Given q and Δ, define $\boldsymbol{\xi}^\flat = q/(q+\Delta)\,\boldsymbol{\xi}$. It holds by homogeneity of T

$$I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \geq q + \Delta\bigr) = I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}^\flat) \geq q\bigr).$$

It is obvious that $\operatorname{Var}(\boldsymbol{\xi}^\flat) = (1 + \Delta/q)^{-2}\Sigma$. Now it holds for the KL-divergence between $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^\flat$

$$\mathcal{K}\bigl(I\!\!P_{\boldsymbol{\xi}}, I\!\!P_{\boldsymbol{\xi}^\flat}\bigr) = \frac{p}{2}\bigl\{2\Delta/q + (\Delta/q)^2 - 2\log(1 + \Delta/q)\bigr\} \leq p(\Delta/q)^2. \tag{6.37}$$

Here we used that $\log(1+\rho) \leq \rho - \rho^2/2$ for $\rho \leq 1$. Now Pinsker's bound (6.35) implies

$$\begin{aligned}
&\bigl| I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \geq q\bigr) - I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}^\flat) \geq q + \Delta\bigr) \bigr| \\
&\quad \leq I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \geq q\bigr) - I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \geq q + \Delta\bigr) \\
&\qquad + \bigl| I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \geq q + \Delta\bigr) - I\!\!P\bigl(\max_{\boldsymbol{\eta}} T_{\boldsymbol{\eta}}(\boldsymbol{\xi}^\flat) \geq q + \Delta\bigr) \bigr| \\
&\quad \leq \square/2 + \Delta\, q^{-1}\sqrt{p/2}
\end{aligned}$$

and (6.36) follows.

We also present a simple corollary of the above result which concerns the change in the expectation $I\!\!E f(\boldsymbol{\xi})$ for a bounded function f.

Lemma 6.6.6. *Let* $\boldsymbol{\xi} \sim \mathcal{N}(0, \Sigma)$ *and* $\boldsymbol{\xi}^\flat \sim \mathcal{N}(0, \Sigma^\flat)$, *where* $\Sigma, \Sigma^\flat$ *satisfy* (6.33). *For any function* f *on* $I\!R^p$ *with* $|f(\boldsymbol{x})| \leq 1$, *and any* $\delta > 0$ *it holds*

$$\bigl| I\!\!E f(\boldsymbol{\xi}) - I\!\!E f\bigl(\boldsymbol{\xi}^\flat\bigr) \bigr| \leq \square. \tag{6.38}$$

Also, for any $\delta \geq 0$

$$\bigl| I\!\!E f(\boldsymbol{\xi}) - I\!\!E f\bigl((1+\delta)\boldsymbol{\xi}\bigr) \bigr| \leq \delta\sqrt{2p}. \tag{6.39}$$

Proof. in view of $|f(\boldsymbol{x})| \leq 1$, it holds

$$\bigl| I\!\!E f(\boldsymbol{\xi}) - I\!\!E f(\boldsymbol{\xi}^\flat) \bigr| \leq \int |f(\boldsymbol{x})| \cdot \bigl| \phi_{\boldsymbol{\xi}}(\boldsymbol{x}) - \phi_{\boldsymbol{\xi}^\flat}(\boldsymbol{x}) \bigr| d\boldsymbol{x} \leq \int \bigl| \phi_{\boldsymbol{\xi}}(\boldsymbol{x}) - \phi_{\boldsymbol{\xi}^\flat}(\boldsymbol{x}) \bigr| d\boldsymbol{x}\,.$$

One more use of Pinsker's inequality yields

$$\int \bigl| \phi_{\boldsymbol{\xi}}(\boldsymbol{x}) - \phi_{\boldsymbol{\xi}^\flat}(\boldsymbol{x}) \bigr| d\boldsymbol{x} = 2 \| I\!\!P_{\boldsymbol{\xi}} - I\!\!P_{\boldsymbol{\xi}^\flat} \|_{TV} \leq \sqrt{2 \mathcal{K}(I\!\!P_{\boldsymbol{\xi}}, I\!\!P_{\boldsymbol{\xi}^\flat})},$$

and the assertion (6.39) follows by $2\mathcal{K}(I\!\!P_{\boldsymbol{\xi}}, I\!\!P_{\boldsymbol{\xi}^\flat}) \leq \square^2$. It remains to note that for $\Sigma^\flat = (1+\delta)^2 \Sigma$, it holds $\mathcal{K}(I\!\!P_{\boldsymbol{\xi}}, I\!\!P_{\boldsymbol{\xi}^\flat}) \leq \delta^2 p$; see (6.37).

7

Sieve Model Selection

This chapter discusses the problem of sieve model selection in the situation when no prior information about the underlying noise distribution is available. The SmA procedure from Section 2.3 requires the set of critical values $\mathtt{z}_{m,m^\circ}$ to be fixed. Here we discuss how this can be done in a data driven way with a resampling procedure.

7.1 Sieve SmA procedure

We consider a general parametric setup $\boldsymbol{Y} \sim I\!\!P \in (I\!\!P_{\boldsymbol{\theta}})$, where the parameter $\boldsymbol{\theta}$ is high or infinite dimensional. The sieve approximations assumes that there is a growing sequence of subspaces $\Theta_1 \subset \Theta_2 \subset \ldots$, one fixes a proper value m and applied the MLE $\widetilde{\boldsymbol{\theta}}_m$ obtained by maximization of the log-likelihood $L(\boldsymbol{\theta})$ over Θ_m:

$$\widetilde{\boldsymbol{\theta}}_m \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta_m} L(\boldsymbol{\theta}). \tag{7.1}$$

The main issue in applying this approach is the choice of the model parameter m. This problem was discussed in details for linear models with a quadratic log-likelihood function in Chapters 2 and 4. Now we aim at extending the SmA approach to the general sieve maximum likelihood setup.

Define the sieve target

$$\boldsymbol{\theta}^*_m = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta_m} I\!\!E L(\boldsymbol{\theta}).$$

The Fisher Theorem claims that the sieve MLS $\widetilde{\boldsymbol{\theta}}_m$ estimates $\boldsymbol{\theta}^*_m$ with the parametric accuracy corresponding to the parameter dimension p_m of the subset Θ_m. The value $\boldsymbol{\theta}^*_m$ differs in general from $\boldsymbol{\theta}^*$ yielding some sieve bias which decreases with m. At the same time, the use of large m leads to a rather complicated parametric problem (7.1) with p_m parameters. Therefore, a proper model choice has to balance the parametric complexity within the sieve model Θ_m and the bias occurring by replacing the full model by its approximation.

Similarly to the linear case we suppose to be given by a loss weighting matrix W and measure the loss of estimation by $\|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|$. The quadratic risk is defined by its expectation

$$\mathbb{E}\|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2.$$

More generally, one can consider polynomial loss function $\|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^q$ for a given $q \geq 0$. For the quadratic risk, one can use the bias-variance decomposition

$$\mathbb{E}\|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2 = \bigl\|W(\mathbb{E}\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\bigr\|^2 + \operatorname{tr}\bigl\{\operatorname{Var}\bigl(W\widetilde{\boldsymbol{\theta}}_m\bigr)\bigr\}.$$

Below we use a slightly different decomposition based on the Fisher expansion. Define

$$\begin{aligned} D_m^2 &= -\nabla_m^2 \mathbb{E}L(\boldsymbol{\theta}^*), \\ \nabla &= \nabla L(\boldsymbol{\theta}^*). \end{aligned}$$

The Fisher expansion for the largest sieve $m = \mathtt{M}$ yields a similar statement for each $m < \mathtt{M}$: on a random set of probability at least $1 - \mathrm{e}^{-\mathtt{x}}$

$$\bigl\|D_m\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*_m\bigr) - D_m^{-1}\nabla\bigr\| \leq \diamondsuit(\mathtt{x})$$

where $\diamondsuit(\mathtt{x}) = \diamondsuit(\mathtt{r}_{\mathtt{M}}, \mathtt{x})$ is the error of approximation in the $\mathtt{M}$ sieve.

Now consider the estimation loss with a weighting matrix W satisfying

$$\|WD_m^{-1}\| \leq 1.$$

Then

$$\bigl\|W\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*_m\bigr) - WD_m^{-2}\nabla\bigr\| = \bigl\|WD_m^{-1}\bigl\{D_m\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*_m\bigr) - D_m^{-1}\nabla\bigr\}\bigr\| \leq \diamondsuit(\mathtt{x})$$

For any two $m > m^\circ$,

$$\bigl\|W\bigl(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ}\bigr) - W(\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^*_{m^\circ}) - W\bigl(D_m^{-2} - D_{m^\circ}^{-2}\bigr)\nabla\bigr\| \leq 2\diamondsuit(\mathtt{x}).$$

This expansion can be rewritten as

$$\bigl\|W\bigl(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ}\bigr) - \boldsymbol{b}_{m,m^\circ} - \boldsymbol{\xi}_{m,m^\circ}\bigr\| \leq 2\diamondsuit(\mathtt{x}) \tag{7.2}$$

with

$$\begin{aligned} \boldsymbol{b}_{m,m^\circ} &\stackrel{\text{def}}{=} W(\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^*_{m^\circ}), \\ \boldsymbol{\xi}_{m,m^\circ} &\stackrel{\text{def}}{=} W\bigl(D_m^{-2} - D_{m^\circ}^{-2}\bigr)\nabla. \end{aligned} \tag{7.3}$$

In the linear case, the expansion $W\bigl(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ}\bigr) - \boldsymbol{b}_{m,m^\circ} - \boldsymbol{\xi}_{m,m^\circ} = 0$ is exact. The formula (7.2) is an extension to a general nonlinear regular case.

Now we suppose that the error term is small enough and proceed as if this expansion is identity. The SmA procedure selects the "smallest accepted" with the acceptance rule

$$m^\circ \text{ is accepted if } \quad \bigl\|W\bigl(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ}\bigr)\bigr\| \le \mathtt{z}_{m,m^\circ} \quad \forall m \in \mathcal{M}(m^\circ).$$

The critical values $\mathtt{z}_{m,m^\circ}$ have to be selected to ensure the propagation property: if there is no bias for $m > m^\circ$ then the procedure should not reject m°.

Below we discuss how these values can be selected by resampling methods.

7.2 Resampling methods for parameter tuning in generalized regression

This section explains the choice of the critical values $\mathtt{z}_{m,m^\circ}$ for the generalized regression model by a multiplier bootstrap procedure.

7.2.1 Generalized regression

We consider a sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ with independent observations Y_i.

Our parametric assumption concerns the marginal distribution of each observation Y_i and the structure of the regression function f.

We suppose that the distribution P_i of each observation Y_i belongs to a given family $\mathcal{P} = (P_{\boldsymbol{\upsilon}})$ and the value of this parameter is an unknown function f of the regressor $\boldsymbol{X}_i$. We write this relation in the form

$$Y_i \sim P_{f(\boldsymbol{X}_i)}\,. \tag{7.4}$$

Further we suppose the family $\mathcal{P}$ to be regular and dominated by a sigma-finite measure μ_0. By $\ell(y, \boldsymbol{\upsilon})$ we denote the corresponding log-density function: $\ell(y, \boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \log \frac{dP_{\boldsymbol{\upsilon}}}{d\mu_0}(y)$.

The regression function f will be modelled by a linear expansion

$$f(\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{j=1}^{p} \theta_j^* \psi_j(\boldsymbol{x}) \tag{7.5}$$

for a given basis system $\{\psi_j\}$. For simplicity this basis will be considered finite: $j \le p$ for a finite p. We write this expansion in the vector form $\boldsymbol{f}^* = \Psi^\top \boldsymbol{\theta}^*$.

Our parametric assumptions (7.4) about the marginal distribution of each Y_i and (7.5) about the structure of the regression function f lead to the log-likelihood function

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell\bigl(Y_i, f(X_i, \boldsymbol{\theta})\bigr) = \sum_{i=1}^{n} \ell\bigl(Y_i, \Psi_i^\top \boldsymbol{\theta}\bigr) \tag{7.6}$$

and the MLE $\widetilde{\boldsymbol{\theta}}$ maximizes $L(\boldsymbol{\theta})$ over the large set of all feasible $\boldsymbol{\theta}$-values. The sieve approach leads to the family of estimates $\widetilde{\boldsymbol{\theta}}_m$ each of them is defined by restricting the parameter set to a subset Θ_m in which only first m components of $\boldsymbol{\theta}$ are varying. This means that for $\boldsymbol{\theta} \in \Theta_m$, the expansion (7.5) reads as

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{j=1}^{m} \theta_j^* \psi_j(\boldsymbol{x}). \tag{7.7}$$

Exercise 7.2.1. Write the sieve MLE $\widetilde{\boldsymbol{\theta}}_m$ for the generalized linear regression (7.7) in

- Poisson regression with canonical parameter $\boldsymbol{\upsilon} = 1/\lambda$, where λ is the Poisson intensity parameter;
- logit (Bernoulli canonical) model

Describe in each case the Fisher matrices D_m, the score ∇_m, and the stochastic term $\boldsymbol{\xi}_{m,m^\circ}$ in expansion (7.3).

7.2.2 Multiplier bootstrap

We consider now the SmA procedure based on the family of estimates $\widetilde{\boldsymbol{\theta}}_m$ and discuss how the critical values $\mathtt{z}_{m,m^\circ}$ can be selected in a data-driven way. In what follows we suppose that the data sample $\boldsymbol{Y}$ is fixed as well as the corresponding feature vectors Ψ_i. This means that the critical values will be computed given data and are in general data-dependent.

Introduce the central object of analysis - the weighted log-likelihood $L^\flat(\boldsymbol{\theta})$ defined via the family of random weights $w_i^\flat$:

$$L^\flat(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \ell\big(Y_i, f(X_i, \boldsymbol{\theta})\big) w_i^\flat . \tag{7.8}$$

The weights $w_i^\flat$ are assumed i.i.d. given the data $\boldsymbol{Y}$ with $I\!\!E w_i^\flat = \operatorname{Var} w_i^\flat = 1$. Two leading examples of such weights are Gaussian weights with $w_i^\flat \sim \mathcal{N}(1,1)$ and the exponential weights $w_i^\flat \sim \operatorname{Exp}(1)$. Note the the expression (7.8) looks very similar to (7.6) but there is an essential difference in the probabilistic nature of them. The original log-likelihood is a function of the random data $\boldsymbol{Y}$, its distribution is described via the unknown data distribution. In the expression (7.8) the data $\boldsymbol{Y}$ are considered as fixed and non-random, the only random element there is a collection of weights $w_i^\flat$ with known distribution. In particular, if $w_i^\flat$ are i.i.d. normal then $L^\flat(\boldsymbol{\theta})$ is also normal as a linear combination of normal r.v.s. One can conclude that $L(\boldsymbol{\theta})$ and $L^\flat(\boldsymbol{\theta})$ are living on two different probability spaces and have very different properties. The link between these two worlds (of real and of bootstrap ones) is given by a very simple observation

$$I\!\!E^{\flat} L^{\flat}(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}).$$

Here and everywhere below $I\!\!P^{\flat}$ means the distribution of the weights $w_i^{\flat}$ given the data $\boldsymbol{Y}$. By $I\!\!E^{\flat}$ we denote the corresponding expectation. The main benefit of considering the measure $I\!\!P^{\flat}$ is that it is completely known.

Now we interpret $L^{\flat}(\boldsymbol{\theta})$ as a log-likelihood process and define

$$\widetilde{\boldsymbol{\theta}}^{\flat} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}} L^{\flat}(\boldsymbol{\theta}). \tag{7.9}$$

Similarly one can define $\widetilde{\boldsymbol{\theta}}_m^{\flat}$ by maximizing $L^{\flat}(\boldsymbol{\theta})$ over Θ_m:

$$\widetilde{\boldsymbol{\theta}}_m^{\flat} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta_m} L^{\flat}(\boldsymbol{\theta}). \tag{7.10}$$

The value $\widetilde{\boldsymbol{\theta}}_m^{\flat}$ formally depends on both the data $\boldsymbol{Y}$ and the weights $w_i^{\flat}$ but we consider its conditional distribution given the data $\boldsymbol{Y}$ with the hope that this distribution somehow reflects the original distribution of $\widetilde{\boldsymbol{\theta}}_m$. In fact, for running the SmA procedure we only need to know the distribution (tail function) of stochastic parts $\boldsymbol{\xi}_{m,m^\circ}$ of considered test statistics $W(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ})$. Define their analog in the bootstrap world:

$$\mathbb{T}_{m,m^\circ}^{\flat} \stackrel{\text{def}}{=} \bigl\| W(\widetilde{\boldsymbol{\theta}}_m^{\flat} - \widetilde{\boldsymbol{\theta}}_{m^\circ}^{\flat}) \bigr\|.$$

Now we use the great advantage of considering the bootstrap world: the distribution of the $\widetilde{\boldsymbol{\theta}}_m^{\flat}$ is known, in particular, one can compute the expectation $I\!\!E^{\flat}(\widetilde{\boldsymbol{\theta}}_m^{\flat})$ or use the knowledge of the true value in the bootstrap model which coincides with the real world estimate $\widetilde{\boldsymbol{\theta}}_m$. The corresponding stochastic component $\boldsymbol{\xi}_{m,m^\circ}^{\flat}$ is given by

$$\boldsymbol{\xi}_{m,m^\circ}^{\flat} \stackrel{\text{def}}{=} W\bigl\{ \widetilde{\boldsymbol{\theta}}_m^{\flat} - \widetilde{\boldsymbol{\theta}}_{m^\circ}^{\flat} - I\!\!E^{\flat}(\widetilde{\boldsymbol{\theta}}_m^{\flat} - \widetilde{\boldsymbol{\theta}}_{m^\circ}^{\flat}) \bigr\}.$$

Now define the tail function $z_{m,m^\circ}^{\flat}(\mathtt{x})$ by the relation

$$I\!\!P^{\flat}\bigl(\|\boldsymbol{\xi}_{m,m^\circ}^{\flat}\| \geq z_{m,m^\circ}^{\flat}(\mathtt{x}) \bigr) = \mathrm{e}^{-\mathtt{x}} \tag{7.11}$$

and further proceed as in the case of known functions z_{m,m°'s. In particular, the multiplicity correction $q_{m^\circ} = q_{m^\circ}(\mathtt{x})$ is defined as the smallest value q satisfying the relation

$$I\!\!P^{\flat}\biggl(\max_{m > m^\circ} \Bigl\{ \|\boldsymbol{\xi}_{m,m^\circ}^{\flat}\| - z_{m,m^\circ}^{\flat}(\mathtt{x}+q) \Bigr\} \geq 0 \biggr) \leq \mathrm{e}^{-\mathtt{x}}. \tag{7.12}$$

Finally the SmA procedure is applied with

$$\mathtt{z}_{m,m^\circ} \stackrel{\text{def}}{=} z_{m,m^\circ}^{\flat}(\mathtt{x} + q_{m^\circ}) + \beta \sqrt{\mathtt{p}_{m,m^\circ}^{\flat}}. \tag{7.13}$$

7.2.3 Numerical issues

It was mentioned many times that the joint distribution of the test statistics $\mathbb{T}^{\flat}_{m,m^\circ}$ under $I\!P^{\flat}$ is known. However, its analytic study is a hard task even if the weights $w_i^{\flat}$ are normal. Similarly to linear regression case, one can use a numerical Monte Carlo procedure for evaluating the tail functions of $\mathbb{T}^{\flat}_{m,m^\circ}$ and the multiplicity corrections q_{m°. The approach can be described as follows:

- generate B samples of weights $\boldsymbol{w}^{\flat} = (w_i^{\flat})$;
- for each sample, compute and store $\widetilde{\boldsymbol{\theta}}^{\flat}_m = \widetilde{\boldsymbol{\theta}}^{\flat}_m(\boldsymbol{w}^{\flat})$ for all m;
- compute the bootstrap empirical means

$$I\!E^{\flat}(\widetilde{\boldsymbol{\theta}}^{\flat}_m) \stackrel{\text{def}}{=} \frac{1}{B}\sum_{\boldsymbol{w}^{\flat}} \widetilde{\boldsymbol{\theta}}^{\flat}_m(\boldsymbol{w}^{\flat});$$

- For each $m > m^\circ$ and all feasible $\mathtt{x}$, compute the tail functions $z_{m,m^\circ}(\mathtt{x})$ by the relations (7.11) when $I\!P^{\flat}$ is replaced by its bootstrap empirical distribution:

$$\frac{1}{B}\sum_{\boldsymbol{w}^{\flat}} \mathbb{1}\bigl(\|\boldsymbol{\xi}^{\flat}_{m,m^\circ}(\boldsymbol{w}^{\flat})\| \geq z^{\flat}_{m,m^\circ}(\mathtt{x})\bigr) \leq \mathrm{e}^{-\mathtt{x}}$$

with

$$\boldsymbol{\xi}^{\flat}_{m,m^\circ}(\boldsymbol{w}^{\flat}) \stackrel{\text{def}}{=} W\Bigl\{\widetilde{\boldsymbol{\theta}}^{\flat}_m(\boldsymbol{w}^{\flat}) - \widetilde{\boldsymbol{\theta}}^{\flat}_{m^\circ}(\boldsymbol{w}^{\flat}) - I\!E^{\flat}(\widetilde{\boldsymbol{\theta}}^{\flat}_m - \widetilde{\boldsymbol{\theta}}^{\flat}_{m^\circ})\Bigr\}.$$

Alternatively,

$$\boldsymbol{\xi}^{\flat}_{m,m^\circ}(\boldsymbol{w}^{\flat}) \stackrel{\text{def}}{=} W\Bigl\{\widetilde{\boldsymbol{\theta}}^{\flat}_m(\boldsymbol{w}^{\flat}) - \widetilde{\boldsymbol{\theta}}^{\flat}_{m^\circ}(\boldsymbol{w}^{\flat}) - (\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ})\Bigr\}.$$

Also estimate

$$\mathtt{p}^{\flat}_{m,m^\circ} = \mathrm{Var}^{\flat}(\boldsymbol{\xi}^{\flat}_{m,m^\circ}) \approx \mathrm{tr}\Bigl\{\frac{1}{B}\sum_{\boldsymbol{w}^{\flat}} \boldsymbol{\xi}^{\flat}_{m,m^\circ}(\boldsymbol{w}^{\flat})\, \boldsymbol{\xi}^{\flat}_{m,m^\circ}(\boldsymbol{w}^{\flat})^{\top}\Bigr\}.$$

- Compute for each m° the multiplicity corrections q_{m° as the smallest value with

$$\frac{1}{B}\sum_{\boldsymbol{w}^{\flat}} \mathbb{1}\Bigl(\max_{m>m^\circ}\Bigl\{\|\boldsymbol{\xi}^{\flat}_{m,m^\circ}(\boldsymbol{w}^{\flat})\| - z^{\flat}_{m,m^\circ}(\mathtt{x}+q_{m^\circ})\Bigr\} \geq 0\Bigr) \leq \mathrm{e}^{-\mathtt{x}}.$$

Complexity of this procedure is mainly determined by the complexity of computing the family of estimates $\widetilde{\boldsymbol{\theta}}^{\flat}_m(\boldsymbol{w}^{\flat})$ for each bootstrap sample $\boldsymbol{w}^{\flat}$. This has to be done many many times to reduce the Monte-Carlo error. Each estimate $\widetilde{\boldsymbol{\theta}}^{\flat}_m$ is given implicitly via the optimization problem (7.10). This can be a hard task especially if the parameter dimension p is large. Note however one important feature of the bootstrap world: we know the true value, which the estimate $\widetilde{\boldsymbol{\theta}}$ computed from the original data. This true

value is actually there target of estimation from the resampled data and automatically it is a very good starting point of the estimation procedure; see further details of reducing the computational burden below in Section **??**.

In the next section we discuss how the proposed procedure can be justified. First we consider the linear Gaussian case. Then we extend the results to the case of linear models with non-Gaussian errors. Finally we consider the general case of sieve model selection and show how the bootstrap validity cab be derived from the Fisher expansions (7.2) in the asymptotic sense for a reasonably large sample size n.

7.3 Why does it work? Linear Gaussian case

This section studies the validity of the bootstrap procedure for a linear Gaussian model

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

In words, the model assumes linear systematic dependence $I\!\!E\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^*$ and standard Gaussian noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I_n)$. In reality, both assumptions can be misspecified. We aim to check to which expend the proposed procedure is robust w.r.t. possible model misspecifications. First we make the analysis for just one estimate, then we discuss how the results apply to the SmA model selection procedure.

7.3.1 Small modeling bias condition

This section discusses in the setup of linear Gaussian regression whether the bootstrap procedure does the job if the model assumptions are misspecified. Our study admits that the model is misspecified: the linear dependence can be violated and we admit a inhomogeneous Gaussian noise, that is, $\boldsymbol{\varepsilon}$ be zero mean normal with a diagonal covariance matrix $\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. The main question is whether the bootstrap procedure mimics well the noise distribution and is robust with respect to a moderate deviation from the linear modeling assumption. The final result presents some sufficient conditions on the deviation $I\!\!E\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}^*$ and on the variance coefficients σ_i^2 which ensure the bootstrap validity.

Let $\widetilde{\boldsymbol{\theta}} = \mathcal{S}\boldsymbol{Y} = \left(\Psi\Psi^\top\right)^{-1}\Psi\boldsymbol{Y}$ be the qMLE for our linear Gaussian parametric model. The corresponding target value is $\boldsymbol{\theta}^* = \left(\Psi\Psi^\top\right)^{-1}\Psi\boldsymbol{f}^*$. Below we suppose that $\boldsymbol{\theta}^* = 0$ which does not restrict generality, because this case can be obtained by reparametrization (a linear shift of the parameter $\boldsymbol{\theta}$). For the real world estimate $\widetilde{\boldsymbol{\theta}} = D^{-2}\Psi\boldsymbol{Y}$ holds

$$\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = D^{-2}\Psi(\boldsymbol{Y} - \Psi^\top\boldsymbol{\theta}^*) = D^{-2}\Psi\boldsymbol{\varepsilon} = D^{-2}\nabla$$

with $\nabla = \Psi\varepsilon$. Under the assumption of an inhomogeneous Gaussian noise with ε being zero mean normal with a diagonal covariance matrix $\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, the score $\nabla = \Psi\varepsilon$ is zero mean normal as well and

$$V^2 \stackrel{\text{def}}{=} \operatorname{Var}(\nabla) = I\!\!E(\nabla\nabla^\top) = \Psi \operatorname{Var}(\varepsilon)\Psi^\top = \Psi\Sigma\Psi^\top. \tag{7.14}$$

Now we check what happens in the bootstrap world. One can easily see that

$$L^\flat(\boldsymbol{\theta}) = -\frac{1}{2}\sum_i \bigl(Y_i - \Psi_i^\top\boldsymbol{\theta}\bigr)^2 w_i^\flat + R,$$

where the remainder R does not depend on $\boldsymbol{\theta}$. The corresponding estimates $\widetilde{\boldsymbol{\theta}}^\flat$ read as

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

Alternatively one can consider the estimate

$$\widetilde{\boldsymbol{\theta}}^\flat = \bigl(\Psi\Psi^\top\bigr)^{-1}\Psi\,\mathcal{W}^\flat\boldsymbol{Y} = D^{-2}\Psi\mathcal{W}^\flat\boldsymbol{Y} \tag{7.15}$$

with $D^2 = \Psi\Psi^\top$. These two versions are very close to each other because the matrix $\Psi\Psi^\top$ and its bootstrap counterpart $\Psi\mathcal{W}^\flat\Psi^\top$ are close to each other. The version (7.15) is preferable for the bootstrap procedure because one can use the same matrix $D^{-2}\Psi$ for all bootstrap runs, otherwise one has to recompute it for each $\boldsymbol{w}^\flat$. The theoretical study is also simpler for the version (7.15). By construction $I\!\!E^\flat(w_i^\flat) = 1$ yielding $I\!\!E^\flat\mathcal{W}^\flat = I_n$ and

$$\widetilde{\boldsymbol{\theta}}^\flat - \widetilde{\boldsymbol{\theta}} = D^{-2}\Psi(\mathcal{W}^\flat - I_p)\boldsymbol{Y} = D^{-2}\nabla^\flat,$$

where for $\mathcal{E}^\flat = \mathcal{W}^\flat - I_p$

$$\nabla^\flat \stackrel{\text{def}}{=} \Psi\mathcal{E}^\flat\boldsymbol{Y} \tag{7.16}$$

and $I\!\!E^\flat(\nabla^\flat) = 0$.

Further, the model equation $\boldsymbol{Y} = \boldsymbol{f}^* + \varepsilon$ yields the decomposition of the bootstrap score $\nabla^\flat = \Psi\mathcal{E}^\flat\boldsymbol{Y}$ from (7.16) as

$$\nabla^\flat = \Psi\mathcal{E}^\flat\varepsilon + \Psi\mathcal{E}^\flat\boldsymbol{f}^*.$$

Below we show that the first term $\Psi\mathcal{E}^\flat\varepsilon$ nicely mimics in distribution the original score ∇, while the second term is small under the so called "small modeling bias" condition. Let the weights $w_i^\flat$ be normal $\mathcal{N}(1,1)$. Then each $e_i^\flat = w_i^\flat - 1$ is standard normal and $\Psi\mathcal{E}^\flat\varepsilon$ is also zero mean normal under $I\!\!P^\flat$ with

$$\mathrm{Var}^{\flat}(\Psi\mathcal{E}^{\flat}\boldsymbol{\varepsilon}) = I\!\!E^{\flat}\{\Psi\mathcal{E}^{\flat}\boldsymbol{\varepsilon}(\Psi\mathcal{E}^{\flat}\boldsymbol{\varepsilon})^{\top}\} = \Psi\ \mathrm{Var}(\mathcal{E}^{\flat}\boldsymbol{\varepsilon})\,\Psi^{\top} = \Psi\ \mathrm{diag}(\boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon})\,\Psi^{\top}. \quad (7.17)$$

Exercise 7.3.1. Let $\mathcal{E}^{\flat}\boldsymbol{\varepsilon}$ be the vector with the entries $w_i^{\flat}\varepsilon_i$.

- Check that

$$\mathrm{Var}^{\flat}(\mathcal{E}^{\flat}\boldsymbol{\varepsilon}) = \mathrm{diag}(\boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon}),$$

  where $\mathrm{diag}(\boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon}) = \mathrm{diag}(\varepsilon_1^2,\ldots,\varepsilon_n^2)$ is the diagonal matrix with the entries ε_i^2.
- Check the formula (7.17)

For the study we need a kind of Lindeberg condition which requires for each vector Ψ_i to be small relative to the standardized sum $\Sigma^{1/2}$. More precisely, define

$$\delta_{\Psi} \stackrel{\mathrm{def}}{=} \max_{i=1,\ldots,n} \|V^{-1}\Psi_i\|\sigma_i\,. \quad (7.18)$$

Exercise 7.3.2. Consider the case of a regular design, when all the Ψ_i's belongs to a compact subset $\mathcal{X}$ of $I\!R^p$, the matrix

$$d_{\Psi}^2 \stackrel{\mathrm{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\Psi_i\Psi_i^{\top}$$

is non-degenerate, and the value

$$a_{\Psi} \stackrel{\mathrm{def}}{=} \max_{i=1,\ldots,n} \|d_{\Psi}^{-1}\Psi_i\| \quad (7.19)$$

is finite. Moreover, let the ratio $\lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$ of the maximal and minimal values of σ_i^2's be bounded by a_{Σ}^2, that is, for all $i,j \le n$

$$\sigma_i/\sigma_j \ \le a_{\Sigma}\,. \quad (7.20)$$

Check that δ_{Ψ} from (7.18) fulfills

$$\delta_{\Psi} \le a_{\Psi}\, a_{\Sigma}\, n^{-1/2}. \quad (7.21)$$

Two centered Gaussian distributions with different covariance matrices can be compared via the Pinsker inequality. The general result of Theorem 6.6.2 claims that these two distributions are close to each other if the quantity

$$\begin{aligned}\square^2 &\stackrel{\mathrm{def}}{=} \mathrm{tr}\Big[\Big\{V^{-1}\ \mathrm{Var}^{\flat}(\Psi\mathcal{E}^{\flat}\boldsymbol{\varepsilon})\,V^{-1} - \boldsymbol{I}_p\Big\}^2\Big] \\ &= \mathrm{tr}\Big[\Big\{V^{-1}\Psi\,\mathrm{diag}(\boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon})\Psi^{\top}V^{-1} - \boldsymbol{I}_p\Big\}^2\Big]\end{aligned}$$

is small. For a symmetric $p \times p$ matrix A, the trace $\operatorname{tr}(A^2)$ of A^2 coincides with its squared Frobenius norm and it can be bounded from above using the operator norm (maximal eigenvalue) $\|A\|$ of the matrix A:

$$\operatorname{tr}(A^2) \le p\|A\|^2.$$

We apply this bound for the matrix $\boldsymbol{A}$ in the form

$$\boldsymbol{A} \stackrel{\text{def}}{=} V^{-1}\Psi \operatorname{diag}(\boldsymbol{\varepsilon} \cdot \boldsymbol{\varepsilon})\Psi^\top V^{-1} - \boldsymbol{I}_p.$$

Now we use that $\widetilde{\boldsymbol{\varepsilon}} = \Sigma^{-1/2}\boldsymbol{\varepsilon}$ is a standard normal vector in $\mathbb{R}^n$. Define

$$\boldsymbol{U}^\top \stackrel{\text{def}}{=} V^{-1}\Psi\Sigma^{1/2}. \tag{7.22}$$

Exercise 7.3.3. Check whether the definition (7.14) implies that the columns $\boldsymbol{u}_j \in \mathbb{R}^n$ of the matrix $\boldsymbol{U}$ are orthonormal:

$$\boldsymbol{u}_i^\top \boldsymbol{u}_j = \mathbb{1}(i = j).$$

Suppose that

$$\max_{i=1,\ldots,n} \|V^{-1}\Psi_i\|\sigma_i \le \delta_\Psi$$

Then for any unit vector $\boldsymbol{\gamma} \in \mathbb{R}^p$, the vector $\boldsymbol{u} = \boldsymbol{U}\boldsymbol{\gamma} \in \mathbb{R}^n$ fulfills $\|\boldsymbol{u}\| = 1$ and

$$\|\boldsymbol{u}\|_\infty \le \delta_\Psi\,. \tag{7.23}$$

Hint: suppose that all $\sigma_i \equiv 1$. Then the Cauchy-Schwartz inequality, each component u_i of $\boldsymbol{u}$ satisfies

$$u_i = \boldsymbol{\gamma}^\top V^{-1}\Psi_i \le \|\boldsymbol{\gamma}\|\,\|V^{-1}\Psi_i\|.$$

With $\boldsymbol{U}$ from (7.22), one can write

$$\boldsymbol{A} = \boldsymbol{U}^\top \operatorname{diag}\bigl\{\widetilde{\boldsymbol{\varepsilon}} \cdot \widetilde{\boldsymbol{\varepsilon}} - 1\bigr\}\boldsymbol{U}. \tag{7.24}$$

The spectral norm of A can be easily bounded in the univariate case with $p = 1$. Then $\boldsymbol{U}$ coincides with a unit vector $\boldsymbol{u} \in \mathbb{R}^n$ satisfying $\|\boldsymbol{u}\|_\infty \le \delta_\Psi$ and the value

$$\boldsymbol{u}^\top \operatorname{diag}\bigl\{\widetilde{\boldsymbol{\varepsilon}} \cdot \widetilde{\boldsymbol{\varepsilon}} - 1\bigr\}\boldsymbol{u} = \sum_{i=1}^n u_i^2(\widetilde{\varepsilon}_i^2 - 1)$$

can be bounded using the lower and upper bounds for centered Gaussian quadratic forms; cf. Theorem 6.6.1: with $\mathrm{v}_{\boldsymbol{u}}^2 = 2\sum_i u_i^4$ and $a^* = \max u_i^2 = \|\boldsymbol{u}\|_\infty^2$

$$I\!P\Big(\Big|\sum_{i=1}^{n} u_i^2(\widetilde{\varepsilon}_i^2 - 1)\Big| \geq \sqrt{4\mathtt{v}_{\boldsymbol{u}}^2 \mathtt{x}} \vee (6\|\boldsymbol{u}\|_\infty^2 \mathtt{x})\Big) \leq \mathrm{e}^{-\mathtt{x}}.$$

One can also bound

$$\mathtt{v}_{\boldsymbol{u}}^2 = \sum_{i=1}^{n} u_i^4 \leq \|\boldsymbol{u}\|_\infty^2 \sum_i u_i^2 = \|\boldsymbol{u}\|_\infty^2 \leq \delta_\Psi^2.$$

Therefore, with a high probability

$$\boldsymbol{u}^\top \operatorname{diag}\bigl\{\widetilde{\boldsymbol{\varepsilon}} \cdot \widetilde{\boldsymbol{\varepsilon}} - 1\bigr\}\boldsymbol{u} \leq (2\delta_\Psi \mathtt{x}^{1/2}) \vee \bigl(6\delta_\Psi^2 \mathtt{x}\bigr). \tag{7.25}$$

Exercise 7.3.4. Complete the proof of (7.25).

In the matrix case one has to increase the tail level $\mathtt{x}$ by $\log p$. Putting together all obtained bounds yields for $3\mathtt{x}^{1/2}\delta_\Psi \leq 1$ with probability at least $1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\|A\| = \bigl\|\boldsymbol{U}^\top \operatorname{diag}\bigl\{\widetilde{\boldsymbol{\varepsilon}} \cdot \widetilde{\boldsymbol{\varepsilon}} - 1\bigr\}\boldsymbol{U}\bigr\| \leq 2\delta_\Psi(\mathtt{x} + \log p)^{1/2}. \tag{7.26}$$

To be done: complete the case $p > 1$.

The bootstrap procedure in the unbiased case is justified if $\operatorname{tr}(\boldsymbol{A}^2)$ is small. In view of the bound $\operatorname{tr}(\boldsymbol{A}^2) \leq p\|\boldsymbol{A}\|^2$ and (7.26), it follows from the condition "$p\delta_\Psi^2(\mathtt{x} + \log p)$ is small". Under regular noise and regular design, δ_Ψ is of order $n^{-1/2}$; see (7.21). Therefore, it suffices that $n^{-1}p\log(p)$ is small.

Now we consider the general case and try to evaluate the impact of the bias $\boldsymbol{f}^* - \Psi^\top\boldsymbol{\theta}^*$ which under $\boldsymbol{\theta}^* = 0$ coincides with $\boldsymbol{f}^*$. For $\nabla^\flat = \Psi\mathcal{E}^\flat\boldsymbol{Y} = \Psi\mathcal{E}^\flat(\boldsymbol{\varepsilon} + \boldsymbol{f}^*)$, it holds with $\boldsymbol{U}$ from (7.22)

$$V^{-1}\operatorname{Var}^\flat\bigl(\nabla^\flat\bigr)V^{-1} = \boldsymbol{U}^\top \operatorname{diag}\bigl\{(\widetilde{\boldsymbol{\varepsilon}} + \boldsymbol{B}) \cdot (\widetilde{\boldsymbol{\varepsilon}} + \boldsymbol{B})\bigr\}\boldsymbol{U}, \tag{7.27}$$

where

$$\boldsymbol{B} = \Sigma^{-1/2}\boldsymbol{f}^* = \Sigma^{-1/2}\bigl(\boldsymbol{f}^* - \Pi\boldsymbol{f}^*\bigr).$$

Then

$$\begin{aligned}\boldsymbol{A} &= \boldsymbol{U}^\top \operatorname{diag}\bigl\{(\widetilde{\boldsymbol{\varepsilon}} + \boldsymbol{B}) \cdot (\widetilde{\boldsymbol{\varepsilon}} + \boldsymbol{B}) - 1\bigr\}\boldsymbol{U} \\ &= \boldsymbol{U}^\top \operatorname{diag}\bigl\{\widetilde{\boldsymbol{\varepsilon}} \cdot \widetilde{\boldsymbol{\varepsilon}} - 1\bigr\}\boldsymbol{U} + \boldsymbol{U}^\top \operatorname{diag}\bigl\{\boldsymbol{B} \cdot \boldsymbol{B}\bigr\}\boldsymbol{U} + 2\boldsymbol{U}^\top \operatorname{diag}\bigl\{\widetilde{\boldsymbol{\varepsilon}} \cdot \boldsymbol{B}\bigr\}\boldsymbol{U}. \end{aligned} \tag{7.28}$$

The quadratic bias term can be easily evaluated. Indeed, for any unit vector $\boldsymbol{u} = (u_1, \ldots, u_n)^\top \in I\!R^n$ with $\|\boldsymbol{u}\|_\infty \leq \delta_\Psi$, it holds

$$\boldsymbol{u}^\top \operatorname{diag}\bigl\{\boldsymbol{B}\cdot\boldsymbol{B}\bigr\}\boldsymbol{u} = \sum_{i=1}^n u_i^2 b_i^2 \le \|\boldsymbol{u}\|_\infty^2 \|\boldsymbol{B}\|^2 \le \delta_\Psi^2 \|\boldsymbol{B}\|^2. \tag{7.29}$$

This also implies by (7.23)

$$\begin{aligned} \bigl\|\boldsymbol{U} \operatorname{diag}\bigl\{\boldsymbol{B}\cdot\boldsymbol{B}\bigr\}\boldsymbol{U}^\top\bigr\| &= \sup_{\boldsymbol{\gamma}\in I\!R^p:\ \|\boldsymbol{\gamma}\|=1} \boldsymbol{\gamma}^\top \boldsymbol{U} \operatorname{diag}\bigl\{\boldsymbol{B}\cdot\boldsymbol{B}\bigr\}\boldsymbol{U}^\top\boldsymbol{\gamma} \\ &\le \sup_{\boldsymbol{u}\in I\!R^n:\ \|\boldsymbol{u}\|=1, \|\boldsymbol{u}\|_\infty \le \delta_\Psi} \boldsymbol{u}^\top \operatorname{diag}\bigl\{\boldsymbol{B}\cdot\boldsymbol{B}\bigr\}\boldsymbol{u}^\top \le \delta_\Psi^2 \|\boldsymbol{B}\|^2. \end{aligned} \tag{7.30}$$

Exercise 7.3.5. Check (7.27) and the inequalities (7.29) and (7.30).

The cross term can be easily bounded in the univariate case with $p = 1$. Then $\boldsymbol{U}$ coincides with a unit vector $\boldsymbol{u} \in I\!R^n$ satisfying $\|\boldsymbol{u}\|_\infty \le \delta_\Psi$ and

$$\begin{aligned} \boldsymbol{u}^\top \operatorname{diag}\bigl\{\boldsymbol{B}\cdot\widetilde{\boldsymbol{\varepsilon}}\bigr\}\boldsymbol{u} &= \sum_{i=1}^n u_i^2 b_i \widetilde{\varepsilon}_i \sim \mathcal{N}(0, v_{\boldsymbol{u}}^2), \\ v_{\boldsymbol{u}}^2 &\stackrel{\text{def}}{=} \sum_{i=1}^n u_i^4 b_i^2 \le \|\boldsymbol{u}\|_\infty^4 \|\boldsymbol{B}\|^2 \le \delta_\Psi^4 \|\boldsymbol{B}\|^2. \end{aligned}$$

This implies that on a dominating set of probability $1 - \mathrm{e}^{-\mathtt{x}}$

$$\boldsymbol{u}^\top \operatorname{diag}\bigl\{\boldsymbol{B}\cdot\widetilde{\boldsymbol{\varepsilon}}\bigr\}\boldsymbol{u} \le \delta_\Psi^2 \|\boldsymbol{B}\| z_1(\mathtt{x}). \tag{7.31}$$

In the general case $p > 1$, the result continues to apply with $\mathtt{x}$ increased by $\log(p)$.

Putting together (7.28), (7.26), (7.30), and (7.31) yields on a random set of probability at least $1 - 3\mathrm{e}^{-\mathtt{x}}$

$$\|\boldsymbol{A}\| \le 2(\mathtt{x} + \log p)^{1/2} \delta_\Psi + \delta_\Psi^2 \|\boldsymbol{B}\|^2 + 2\delta_\Psi^2 \|\boldsymbol{B}\| z_1(\mathtt{x} + \log p). \tag{7.32}$$

Now we evaluate the distance $\square$ between the distribution of $\nabla = \Psi\boldsymbol{\varepsilon}$ under $I\!P$ and of $\nabla^\flat = \Psi\mathcal{E}^\flat\boldsymbol{Y}$ under $I\!P^\flat$ using that δ_Ψ is a small number. Remind that under regular noise and regular design, it is of order $n^{-1/2}$; see (7.21). In view of $\operatorname{tr}(\boldsymbol{A}^2) \le p\|\boldsymbol{A}\|^2$ and of (7.32), we ensure that $\square$ is small if $n^{-1}p(\mathtt{x} + \log p)$ is small and $p^{1/2}n^{-1}\|\boldsymbol{B}\|^2$ is small as well. Under a homogeneous noise with the variance σ^2, it holds $\|\boldsymbol{B}\|^2 = \sigma^{-2}\|\boldsymbol{f}^*\|^2$. A similar bound holds under "regular noise" condition (7.20). If this condition is fulfilled, it suffices that

$$\frac{p^{1/2}}{n\overline{\sigma}^2}\|\boldsymbol{f}^* - \Psi^\top\boldsymbol{\theta}^*\|^2 \quad \text{is small}$$

with $\overline{\sigma}^2 \stackrel{\text{def}}{=} n^{-1}\operatorname{tr}\Sigma$.

We are prepared to state the following result.

Theorem 7.3.1. *Let $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$ be a Gaussian vector in $\mathbb{R}^n$ with independent components, $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{f}^*, \Sigma)$ for a diagonal matrix $\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. Let also Ψ be a $p \times n$ feature matrix such that the $p \times p$-matrix $V^2 = \Psi \Sigma \Psi^\top$ is non-degenerated and the value*

$$\delta_\Psi = \max_{i=1,\ldots,n} \|V^{-1}\Psi_i\|\sigma_i \tag{7.33}$$

is small. Denote by $\Pi = \Psi^\top (\Psi\Psi^\top)^{-1}\Psi$ the projector on the feature space and define

$$\boldsymbol{B} = \Sigma^{-1/2}(\boldsymbol{f}^* - \Pi \boldsymbol{f}^*).$$

Further, let $\mathcal{W}^\flat = \operatorname{diag}\{w_1^\flat, \ldots, w_n^\flat\}$ be a diagonal matrix of bootstrap $\mathcal{N}(1,1)$ multipliers and $\mathcal{E}^\flat = \mathcal{W}^\flat - \mathbb{E}^\flat \mathcal{W}^\flat$. Then on a set $\Omega(\mathtt{x})$ with $\mathbb{P}\bigl(\Omega(\mathtt{x})\bigr) \geq 1 - 3\mathrm{e}^{-\mathtt{x}}$, the distribution $\mathbb{Q}$ of the score $\nabla = \Psi\boldsymbol{\varepsilon}$ and the conditional distribution $\mathbb{Q}^\flat$ of the bootstrap score $\nabla^\flat = \Psi\mathcal{E}^\flat \boldsymbol{Y}$ given $\boldsymbol{Y}$ are related by

$$\bigl\|\mathbb{Q} - \mathbb{Q}^\flat\bigr\|_{TV} \leq \sqrt{p/2}\, \Delta \tag{7.34}$$

with

$$\Delta = 2(\mathtt{x} + \log p)^{1/2}\delta_\Psi + \delta_\Psi^2 \|\boldsymbol{B}\|^2 + 2\delta_\Psi^2 \|\boldsymbol{B}\| z_1(\mathtt{x} + \log p). \tag{7.35}$$

The same bounds applies to the distance between the distribution of the qMLE $\widetilde{\boldsymbol{\theta}}$ and of the bootstrap estimate $\widetilde{\boldsymbol{\theta}}^\flat = (\Psi\Psi^\top)^{-1}\Psi\,\mathcal{W}^\flat \boldsymbol{Y}$.

Furthermore, if the design Ψ and the noise $\boldsymbol{\varepsilon}$ are regular in the sense (7.19) and (7.20) with some fixed constants a_Ψ and a_Σ, then

$$\delta_\Psi \leq a_\Psi\, a_\Sigma\, n^{-1/2}.$$

The statement (7.34) yields

$$\sup_A \bigl|\mathbb{Q}(A) - \mathbb{Q}^\flat(A)\bigr| \leq \sqrt{p/2}\, \Delta, \tag{7.36}$$

where sup is taken over all measurable subsets of $\mathbb{R}^p$. This result has a number of remarkable corollaries. First we specify the result to the considered estimation loss $\|W(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|$.

Corollary 7.3.1. *Under the conditions of Theorem 7.3.1, it holds on a set $\Omega(\mathtt{x})$ of dominating probability*

$$\sup_{z>0} \Bigl| \mathbb{P}\bigl(\|W(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \leq z\bigr) - \mathbb{P}^\flat\bigl(\|W(\widetilde{\boldsymbol{\theta}}^\flat - \widetilde{\boldsymbol{\theta}}\| \leq z\bigr) \Bigr| \leq \sqrt{p/2}\, \Delta. \tag{7.37}$$

Exercise 7.3.6. Derive (7.37) as a special case of (7.36).

Our next corollary concerns the "no-bias" case with $\boldsymbol{B} \equiv 0$.

Corollary 7.3.2. *Assume the conditions of Theorem 7.3.1 including the regularity conditions (7.19) and (7.20). If the linear parametric assumption is correct, that is, if $\boldsymbol{f}^* = \Psi^\top \boldsymbol{\theta}^*$, then the bootstrap procedure is justified with the accuracy*

$$\square \asymp \sqrt{n^{-1} p \log(p)}\,.$$

In particularly, the procedure is asymptotically valid if the parameter dimension $p = p_n$ satisfies

$$n^{-1} p_n \log(p_n) \to 0, \qquad n \to \infty.$$

One can conclude from (7.34) and (7.35) that the bootstrap procedure does the job if the value

$$p \delta_\Psi^2 \|\boldsymbol{B}\|^2 \quad \text{is small.}$$

In the regular case it can be rewritten as "$p\|\boldsymbol{B}\|^2/n$ is small". As a special case we present the following corollary.

Corollary 7.3.3. *Assume the conditions of Theorem 7.3.1 including the regularity conditions (7.19) and (7.20). If $\boldsymbol{f}^*$ is Sobolev-smooth with the parameter s, that is, if*

$$\frac{1}{n}\|\boldsymbol{B}\|^2 \le \mathtt{C} p^{-2s}$$

then the bootstrap procedure is justified for $s > 1/2$ if $p = p_n \to \infty$ as $n \to \infty$.

7.3.2 The "large bias" case

The result of Theorem 7.3.1 claims that if the modeling bias $\|\boldsymbol{B}\|^2 = \|\Sigma^{-1/2}(\boldsymbol{f}^* - \Pi \boldsymbol{f}^*)\|^2$ is small then the bootstrap mimics well the real distribution. The next question is whether we can apply the bootstrap when the bias is large. The answer is a bit surprising: yes, we can apply, the bootstrap can be validated but the corresponding bootstrap quantiles are larger the the quantiles for the original distributions. In other words, the bootstrap procedure in the "large bias" situation becomes conservative and yields higher coverage probability than the nominal one. The reason is that the bias in the original data transfers to the additional variance in the bootstrap world. More exactly, the vector $V^{-1}\nabla = V^{-1}\Psi\boldsymbol{\varepsilon}$ is standard normal in $\mathbb{R}^p$ under the underlying measure $\mathbb{P}$. At the same time, the vector $V^{-1}\nabla^\flat = V^{-1}\Psi\mathcal{E}^\flat \boldsymbol{Y}$ is Gaussian zero mean under $\mathbb{P}^\flat$ but its variance is larger than the identity if the bias $\boldsymbol{B}$ is significant.

Define

$$\mathcal{V}^2 \stackrel{\text{def}}{=} V^2 + \Psi \operatorname{diag}\bigl\{ \bigl(\boldsymbol{f}^* - \Pi \boldsymbol{f}^*\bigr) \cdot \bigl(\boldsymbol{f}^* - \Pi \boldsymbol{f}^*\bigr) \bigr\} \Psi^\top . \tag{7.38}$$

Then we can approximate the variance of $\nabla^\flat$ by $\mathcal{V}^2$, and the latter can be significantly larger than V^2 depending on the bias $\boldsymbol{f}^* - \Pi \boldsymbol{f}^*$.

Theorem 7.3.2. *Suppose the conditions of Theorem 7.3.1. Then on a set $\Omega(\mathtt{x})$ of dominating probability, it holds with $\mathcal{V}^2$ from (7.38)*

$$\bigl\| \operatorname{Var}^\flat(\mathcal{V}^{-1} \nabla^\flat) - \boldsymbol{I}_p \bigr\| \le 2\delta_\Psi \bigl(\mathtt{x} + \log p\bigr)^{1/2}. \tag{7.39}$$

Moreover, if

$$\|\mathcal{V}^{-1} V^2 \mathcal{V}^{-1}\| \le 1 - 2\delta_\Psi \bigl(\mathtt{x} + \log p\bigr)^{1/2}, \tag{7.40}$$

then on the same set $\Omega(\mathtt{x})$

$$\operatorname{Var}^\flat(V^{-1} \nabla^\flat) \ge \boldsymbol{I}_p .$$

Proof. The second statement follows directly from the first one and the definition of $\mathcal{V}^2$.

As the variance of $\nabla^\flat$ is larger than the variance of ∇, the related probability of any centrally symmetric convex set A in $\mathbb{R}^p$ is larger under $\mathbb{P}$ than under $\mathbb{P}^\flat$.

Corollary 7.3.4. *For any convex centrally symmetric set $A \subset \mathbb{R}^p$, it holds on a set $\Omega(\mathtt{x})$ of dominating probability*

$$\mathbb{Q}(A) - \mathbb{Q}^\flat(A) \ \ge \delta_\Psi \sqrt{2p\bigl(\mathtt{x} + \log p\bigr)} \ . \tag{7.41}$$

Moreover, under (7.40)

$$\mathbb{Q}(A) - \mathbb{Q}^\flat(A) \ \ge 0. \tag{7.42}$$

Exercise 7.3.7. Check that (7.39) implies (7.41).

In particular, one can bound from above the deviation probabilities.

Corollary 7.3.5. *For any $W : \mathbb{R}^p \to \mathbb{R}^q$, it holds on a set $\Omega(\mathtt{x})$ of dominating probability*

$$\sup_{z>0} \Bigl\{ \mathbb{P}\bigl(\|W(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \le z\bigr) - \mathbb{P}^\flat\bigl(\|W(\widetilde{\boldsymbol{\theta}}^\flat - \widetilde{\boldsymbol{\theta}}\| \le z\bigr) \Bigr\} \ge \delta_\Psi \sqrt{2p\bigl(\mathtt{x} + \log p\bigr)} \ . \tag{7.43}$$

Moreover, under (7.40)

$$\sup_{z>0} \Bigl\{ \mathbb{P}\bigl(\|W(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \le z\bigr) - \mathbb{P}^\flat\bigl(\|W(\widetilde{\boldsymbol{\theta}}^\flat - \widetilde{\boldsymbol{\theta}}\| \le z\bigr) \Bigr\} \ge 0. \tag{7.44}$$

Exercise 7.3.8. Derive (7.43) from (7.41) and (7.44) from (7.42).

The inequality (7.44) implies that the $1-\alpha$ quantile of $\|W(\widetilde{\boldsymbol{\theta}}^{\flat} - \widetilde{\boldsymbol{\theta}}\|$ is systematically larger than the corresponding quantile of $\|W(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|$, and thus, the bootstrap procedure becomes conservative in the "large bias" situation.

7.3.3 Bootstrap and the SmA procedure

Now we come back to the model selection setup. We again assume a linear Gaussian model with an inhomogeneous noise and $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$.

Let $(\widetilde{\boldsymbol{\theta}}_m, m \in \mathcal{M})$ be the considered ordered family of estimates $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$ with $\mathcal{S}_m = (\Psi_m \Psi_m^\top)^{-1} \Psi_m$ for $\Psi_m = \Pi_m \Psi$, $m \in \mathcal{M}$. For notational simplicity $\mathcal{M} = \{1, 2, \ldots, p\}$. In particular, p is the largest considered parameter dimension, and here it is supposed to be finite. Below we identify Ψ and Ψ_p.

Let also $m^\circ \in \mathcal{M}$ be fixed. We measure the corresponding modeling bias by the vector

$$\boldsymbol{f}^* - \Psi_{m^\circ}^\top \boldsymbol{\theta}^*_{m^\circ} = \boldsymbol{f}^* - \Pi_{m^\circ} \boldsymbol{f}^*.$$

The aim is to compare the joint distribution of scaled differences $W(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ})$ for all $m > m^\circ$ with the analogous distribution of their bootstrap counterparts. We use the decomposition (4.14)

$$W(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ}) = \boldsymbol{b}_{m,m^\circ} + \boldsymbol{\xi}_{m,m^\circ}$$

with

$$\begin{aligned} \boldsymbol{b}_{m,m^\circ} &= W(\Pi_m \boldsymbol{f}^* - \Pi_{m^\circ} \boldsymbol{f}^*), \\ \boldsymbol{\xi}_{m,m^\circ} &= W(D_m^{-2} \Pi_m - D_{m^\circ}^{-2} \Pi_{m^\circ}) \nabla, \end{aligned} \tag{7.45}$$

and $\nabla = \Psi \boldsymbol{\varepsilon}$. The bias term $\boldsymbol{b}_{m,m^\circ}$ can be well controlled if m° is "good".

Now we switch to the bootstrap counterpart. The main problem is a possible systematic component in $\boldsymbol{Y}$. We aim to design a sensitive procedure which detects the bias of the smoothing method $\mathcal{S}_m$ as precise as possible. The use of $\boldsymbol{Y}$ in the bootstrap world leads for $m^\circ < m^*$ to quantiles of $\boldsymbol{\xi}^{\flat}_{m,m^\circ}$ which are much larger than the corresponding quantiles of $\boldsymbol{\xi}_{m,m^\circ}$ due to the large bias component. This would result in a procedure which heavily oversmooth the data. A possible way out of this problem is based on using an under smoothing pilot $\widetilde{\boldsymbol{\theta}}$ which removes the bias but preserves the variance. Typically one can use the largest considered model $m = p$ and the corresponding estimate $\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_p = \mathcal{S}_p \boldsymbol{Y}$. Now we replace the data by the residuals for the pilot fit:

$$\breve{\boldsymbol{\varepsilon}} \stackrel{\text{def}}{=} \boldsymbol{Y} - \Psi^{\top}\widetilde{\boldsymbol{\theta}}_p = \boldsymbol{Y} - \Pi\boldsymbol{Y}$$

with $\Pi = \Pi_p$. Below we use that for any $m \le p$, it holds

$$\Psi_m\breve{\boldsymbol{\varepsilon}} = \Psi_m(\boldsymbol{I} - \Pi)\boldsymbol{Y} = 0. \tag{7.46}$$

Now we apply these residuals in place of the original data yielding the family of bootstrap residuals

$$\boldsymbol{\zeta}_m^{\flat} = \mathcal{S}_m\mathcal{W}^{\flat}\breve{\boldsymbol{\varepsilon}}$$

for a diagonal weighting matrix $\mathcal{W}^{\flat} = \operatorname{diag}(w_1^{\flat}, \ldots, w_n^{\flat})$ of bootstrap weights. If $\mathcal{E}^{\flat} = \mathcal{W}^{\flat} - I\!\!E^{\flat}\mathcal{W}^{\flat}$, then (7.46) and the definition imply

$$\boldsymbol{\zeta}_m^{\flat} = \bigl(\Psi_m\Psi_m^{\top}\bigr)^{-1}\Psi_m\,\mathcal{E}^{\flat}\breve{\boldsymbol{\varepsilon}} = D_m^{-2}\Pi_m\breve{\nabla}^{\flat},$$

where $D_m^2 = \Psi_m\Psi_m^{\top}$ and

$$\breve{\nabla}^{\flat} = \Psi\mathcal{E}^{\flat}\breve{\boldsymbol{\varepsilon}} = \Psi\mathcal{E}^{\flat}(\boldsymbol{I} - \Pi)\boldsymbol{Y}.$$

For the differences $\boldsymbol{\zeta}_m^{\flat} - \boldsymbol{\zeta}_{m^\circ}^{\flat}$, $m > m^\circ$, this yields:

$$\boldsymbol{\zeta}_m^{\flat} - \boldsymbol{\zeta}_{m^\circ}^{\flat} = (D_m^{-2}\Pi_m - D_{m^\circ}^{-2}\Pi_{m^\circ})\breve{\nabla}^{\flat}.$$

The bootstrap procedure involves the stochastic component obtained by multiplying with the weighting matrix W:

$$\boldsymbol{\xi}_{m,m^\circ}^{\flat} = W\bigl(\boldsymbol{\zeta}_m^{\flat} - \boldsymbol{\zeta}_{m^\circ}^{\flat}\bigr) = W\bigl(D_m^{-2}\Pi_m - D_{m^\circ}^{-2}\Pi_{m^\circ}\bigr)\breve{\nabla}^{\flat} \tag{7.47}$$

for all $m > m^\circ$. One can see that each $\boldsymbol{\xi}_{m,m^\circ}$ in (7.45) is a deterministic functions of the vector ∇ while $\boldsymbol{\xi}_{m,m^\circ}^{\flat}$ is the same function of $\breve{\nabla}^{\flat}$. Therefore, it suffices to compare the distribution of ∇ with the conditional distribution of $\breve{\nabla}^{\flat}$ given $\boldsymbol{Y}$ and show that the latter mimics well the former. This question was studied in the previous section in terms of the quantity δ_Ψ defined by (7.33) for the model p, and of the bias $\boldsymbol{B} = \Sigma^{-1/2}\bigl(\boldsymbol{f}^* - \Pi\boldsymbol{f}^*\bigr)$. The result of Theorem 7.3.1 does not directly applies here, because we use $\breve{\nabla}^{\flat}$ in place of $\nabla^{\flat}$. However, the main arguments of the approach can still be used: given $\boldsymbol{Y}$, the vector $\breve{\nabla}^{\flat}$ is Gaussian zero mean with the covariance matrix

$$\Psi\operatorname{diag}\bigl\{\breve{\boldsymbol{\varepsilon}}\cdot\breve{\boldsymbol{\varepsilon}}\bigr\}\Psi^{\top}.$$

If this random matrix is close in probability to the covariance matrix $\Psi\Sigma\Psi^{\top}$ of the score ∇, then the approach of Section 7.3.1 is still validated. A linear transformation

$\Psi \to \Psi\Sigma^{1/2}$ reduces the study to the case with $\Sigma = \mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{I}_n$. It obviously holds in this situation

$$\begin{aligned} I\!\!E(\check{\boldsymbol{\varepsilon}}\check{\boldsymbol{\varepsilon}}^\top) &= (\boldsymbol{I} - \Pi) I\!\!E\big(\boldsymbol{Y}\boldsymbol{Y}^\top\big)(\boldsymbol{I} - \Pi) \\ &= (\boldsymbol{f}^* - \Pi\boldsymbol{f}^*)(\boldsymbol{f}^* - \Pi\boldsymbol{f}^*)^\top + (\boldsymbol{I} - \Pi)\Sigma(\boldsymbol{I} - \Pi) = \boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{I} - \Pi \end{aligned}$$

with $\Pi = \Pi_p$. Now one can check that each diagonal element Π_{ii} of Π satisfies

$$|\Pi_{ii}| \le \delta_\Psi^2. \tag{7.48}$$

Exercise 7.3.9. Check (7.48) with δ_Ψ from (7.33) for the case $\sigma_i \equiv 1$.

This implies the operator-norm inequality

$$\|I_n - \mathrm{diag}\{\check{\boldsymbol{\varepsilon}} \cdot \check{\boldsymbol{\varepsilon}}\}\| \le \delta_\Psi^2.$$

Now we can summarize that under the small modeling bias condition $\boldsymbol{B} \approx 0$, the covariance of $\check{\nabla}^\flat$ is close in mean to the score covariance $\Psi\Psi^\top$ up to the error of order δ_Ψ. Moreover, the bias term $\boldsymbol{B}\boldsymbol{B}^\top$ can be treated exactly as in Section 7.3.2. These informal considerations motivate the following result.

Theorem 7.3.3. *Let $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}$ be a Gaussian vector in $I\!R^n$ with independent components, $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{f}^*, \Sigma)$ for $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. Let also Ψ be a $p \times n$ feature matrix such that the $p \times p$-matrix $V^2 = \Psi\Sigma\Psi^\top$ is non-degenerated and the value*

$$\delta_\Psi = \max_{i=1,\ldots,n} \|V^{-1}\Psi_i\|\sigma_i$$

is small. Define

$$\boldsymbol{B} = \Sigma^{-1/2}(\boldsymbol{f}^* - \Pi\boldsymbol{f}^*).$$

Given $m^\circ < p$, let the values $z^\flat_{m,m^\circ}(\mathtt{x})$ and q_{m° be fixed for the bootstrap distribution due to (7.11) and (7.12). Then, with Δ from (7.35), it holds

$$I\!\!P\left(\max_{m>m^\circ}\Big\{\|\boldsymbol{\xi}_{m,m^\circ}\| - z^\flat_{m,m^\circ}(\mathtt{x} + q_{m^\circ})\Big\} \ge 0\right) \le 4\mathrm{e}^{-\mathtt{x}} + \Delta.$$

Proof. **To be done:**

The result of this theorem justifies the SmA procedure with the bootstrap-based critical values (7.13) in the asymptotic sense.

7.4 Linear non-Gaussian case and GAR

This section briefly comment why the bootstrap procedure can be validated even if the true error distribution is not Gaussian. This means that we again consider the linear Gaussian likelihood and the corresponding qMLE $\widetilde{\boldsymbol{\theta}}$ is given by $\widetilde{\boldsymbol{\theta}} = D^{-2}\Psi\boldsymbol{Y}$, the errors $\boldsymbol{\varepsilon}$ are independent but no more Gaussian. The discussion of the previous section shows that the most challenging step of analysis is to check that two vectors $\nabla = \Psi\boldsymbol{\varepsilon}$ and $\nabla^\flat = \Psi\mathcal{E}^\flat\boldsymbol{Y}$ have a similar distribution under the corresponding measures. In the Gaussian case, both vectors are normal zero mean and it suffices to compare their covariance matrices. In the non-Gaussian case the situation is more involved. A nice feature of Gaussian bootstrap multipliers is that the distribution of $\nabla^\flat = \Psi\mathcal{E}^\flat\boldsymbol{Y}$ given $\boldsymbol{Y}$ is again Gaussian, and this fact does not rely on the true data distribution. It is entirely due to the construction of the bootstrap multipliers: $\nabla^\flat$ is normal because it is a linear combination of standard normal weights $e_i^\flat = w_i^\flat - 1$. The real score $\nabla = \Psi\boldsymbol{\varepsilon}$ is again a linear combination of errors ε_i, however these errors can be non-normal. If fact, in typical applications, there is no reason to assume that the errors are exactly normal. However, $\Psi\boldsymbol{\varepsilon}$ can be viewed as a linear combination of the errors ε_i. In combination with the condition that the value δ_Ψ from (7.33) is small, the central limit theorem applies and the zero mean standardized vector $V^{-1}\nabla$ is nearly standard normal under some further regularity and moment conditions. This allows to extend the result on bootstrap validity to the non-Gaussian case in some asymptotic sense. In the univariate case with $p = 1$ one can use the famous Berry-Esseen theorem, which can be also extended to the multivariate case in various special setups.

7.5 Sieve Generalized Linear regression

In a special case of a generalized liner model, $\mathcal{P}$ is an exponential family with canonical parametrization. Then $\ell(y, \boldsymbol{\upsilon}) = y\boldsymbol{\upsilon} - d(\boldsymbol{\upsilon})$ and

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell\big(Y_i, f(X_i)\big) = \sum_{i=1}^{n} \big\{ Y_i \Psi_i^\top \boldsymbol{\theta} - d(\Psi_i^\top \boldsymbol{\theta}) \big\} = \boldsymbol{Y}^\top \Psi^\top \boldsymbol{\theta} - A(\boldsymbol{\theta})$$

with

$$A(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} d(\Psi_i^\top \boldsymbol{\theta}).$$

Also define

$$D^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \nabla^2 A(\boldsymbol{\theta}) = \sum_i \Psi_i \Psi_i^\top d''(\Psi_i^\top \boldsymbol{\theta}).$$

The estimate $\widetilde{\boldsymbol{\theta}}$ can be computed by the iterative Newton procedure: start with any $\widetilde{\boldsymbol{\theta}}_0$, e.g. LSE, and then compute

$$\widetilde{\boldsymbol{\theta}}_{k+1} = \widetilde{\boldsymbol{\theta}}_k + D^{-2}(\widetilde{\boldsymbol{\theta}}_k)\Psi\big(\boldsymbol{Y} - \Psi^\top \widetilde{\boldsymbol{\theta}}_k\big), \qquad k = 1, 2, \ldots.$$

Under standard conditions this procedure converges very fast after just few iterations.

The target $\boldsymbol{\theta}^*$ is described by the optimizing the value $I\!\!E L(\boldsymbol{\theta})$. In view of $I\!\!E Y_i = f(\boldsymbol{X}_i)$

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}} I\!\!E L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_i \big\{ f(\boldsymbol{X}_i)\Psi_i^\top \boldsymbol{\theta} - d(\Psi_i^\top \boldsymbol{\theta}) \big\}$$

which leads to the normal equation

$$\sum_{i=1}^n \big\{ f(\boldsymbol{X}_i) - d'(\Psi_i^\top \boldsymbol{\theta}) \big\} \Psi_i = 0.$$

One also has

$$\mathbb{F} = D^2 = -\nabla^2 I\!\!E L(\boldsymbol{\theta}^*) = \sum_i \Psi_i \Psi_i^\top d''(\Psi_i^\top \boldsymbol{\theta}^*) = \Psi \, \boldsymbol{d}''(\Psi^\top \boldsymbol{\theta}^*) \, \Psi^\top,$$

where $\boldsymbol{d}''(\boldsymbol{f})$ is a $n \times n$ diagonal matrix with the diagonal entries $d''(f_i)$.

Further, for the stochastic component $\zeta(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - I\!\!E L(\boldsymbol{\theta})$, it holds with $\varepsilon_i = Y_i - I\!\!E Y_i$

$$\nabla \zeta(\boldsymbol{\theta}) = \sum_i \varepsilon_i \Psi_i = \Psi \boldsymbol{\varepsilon} \tag{7.49}$$

and its covariance matrix fulfills

$$V^2 = \operatorname{Var}\big\{ \nabla \zeta(\boldsymbol{\theta}) \big\} = \sum_i \operatorname{Var}\big(Y_i\big) \Psi_i \Psi_i^\top = \Psi \operatorname{Var}(\boldsymbol{\varepsilon}) \Psi^\top. \tag{7.50}$$

If the assumption on marginal distributions is correct, that is, if $Y_i \sim P_{f(\boldsymbol{X}_i)}$, then

$$\operatorname{Var}\big(Y_i\big) = d''\big(f(\boldsymbol{X}_i)\big). \tag{7.51}$$

By comparing the equations (7.50) and (7.51), one can conclude that D^2 and V^2 coincides if $d''(\Psi_i^\top \boldsymbol{\theta}^*) = d''\big(f(\boldsymbol{X}_i)\big)$, that is, if PA is correct. Otherwise these two matrices may differ from each other. The Fisher expansion reads as

$$\big\| D\big(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\big) - \boldsymbol{\xi} \big\| \le \diamondsuit(\mathtt{x}) \tag{7.52}$$

on a dominating set of probability at least $1 - \mathrm{e}^{-\mathtt{x}}$. Here in view of (7.49)

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

7.5.1 Sieve MLE

Now we consider the sieve MLE setup when a sequence of subsets Θ_m of growing dimension is given and for each m the MLE $\widetilde{\boldsymbol{\theta}}_m$ is computed. The corresponding quadratic risk $\mathcal{R}_m$ is

$$\mathcal{R}_m = I\!\!E\|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2.$$

The Fisher expansion (7.52) restricted to the sieve Θ_m yields

$$\bigl\|D_m\bigl(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*_m\bigr) - \boldsymbol{\xi}_m\bigr\| \le \diamondsuit(\mathtt{x}), \tag{7.53}$$

where

$$\begin{aligned} D_m^2 &\stackrel{\text{def}}{=} -\nabla_m^2 I\!\!E L(\boldsymbol{\theta}^*) = \Psi_m\, \boldsymbol{d}''(\Psi_m^\top \boldsymbol{\theta}^*)\, \Psi_m, \\ \boldsymbol{\xi}_m &\stackrel{\text{def}}{=} D_m^{-1}\nabla_m \zeta(\boldsymbol{\theta}^*) = D_m^{-1}\Psi_m \boldsymbol{\varepsilon}, \\ \boldsymbol{\theta}^*_m &\stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta_m} I\!\!E L(\boldsymbol{\theta}). \end{aligned}$$

Alternatively one can define

$$\boldsymbol{\theta}^*_m = \Pi_m \boldsymbol{\theta}^*$$

the expansion (7.53) continues to apply.

The test statistic $\mathbb{T}_{m,m^\circ}$ can be written as

$$\mathbb{T}_{m,m^\circ} = \|W(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ})\|$$

and the decomposition (7.53) implies

$$\Bigl|\mathbb{T}_{m,m^\circ} - \|\boldsymbol{b}_{m,m^\circ} + \boldsymbol{\xi}_{m,m^\circ}\|\Bigr| \le 2\diamondsuit(\mathtt{x})$$

with

$$\begin{aligned} \boldsymbol{b}_{m,m^\circ} &= W\bigl(\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^*_{m^\circ}\bigr), \\ \boldsymbol{\xi}_{m,m^\circ} &= W\bigl(D_m^{-2}\nabla_m - D_{m^\circ}^{-2}\nabla_{m^\circ}\bigr) = W\bigl(D_m^{-2}\Pi_m - D_{m^\circ}^{-2}\Pi_{m^\circ}\bigr)\nabla \end{aligned} \tag{7.54}$$

for $\nabla = \Psi\boldsymbol{\varepsilon}$.

7.5.2 Bootstrap counterpart

Now we check what happens in the bootstrap world. It holds

$$L^{\flat}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}) w_i^{\flat} = \sum_{i=1}^{n} \{ Y_i \Psi_i^{\top} \boldsymbol{\theta} - d(\Psi_i^{\top} \boldsymbol{\theta}) \} w_i^{\flat} = \boldsymbol{\theta}^{\top} \Psi \mathcal{W}^{\flat} \boldsymbol{Y} - A^{\flat}(\boldsymbol{\theta})$$

with

$$A^{\flat}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} d(\Psi_i^{\top} \boldsymbol{\theta}) w_i^{\flat}.$$

One can use that $\mathbb{E}^{\flat} A^{\flat}(\boldsymbol{\theta}) = A$ and even more, the matrix $A^{\flat}(\boldsymbol{\theta})$ is close to its expectation $A(\boldsymbol{\theta})$ for n large. This suggests to replace $A^{\flat}(\boldsymbol{\theta})$ by $A(\boldsymbol{\theta})$ in our analysis. The same applies to its Hessian $D^2(\boldsymbol{\theta})$.

This implies similarly to the linear case that the stochastic part of the bootstrap-world estimate $\widetilde{\boldsymbol{\theta}}^{\flat}$ is $D^{-2} \breve{\nabla}^{\flat}$ for $\breve{\nabla}^{\flat} = \Psi \mathcal{E}^{\flat} \breve{\boldsymbol{\varepsilon}}$ for $\breve{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \Psi^{\top} \widetilde{\boldsymbol{\theta}}$. One can see that the question of bootstrap validity is again reduced to comparing of two distributions: of ∇ w.r.t. the original measure $\mathbb{P}$ and of $\breve{\nabla}^{\flat}$ w.r.t. to the bootstrap measure $\mathbb{P}^{\flat}$ given $\boldsymbol{Y}$. If the bootstrap weights $w_i^{\flat}$ are Gaussian then the bootstrap score $\breve{\nabla}^{\flat} = \Psi \mathcal{E}^{\flat} \breve{\boldsymbol{\varepsilon}}$ is Gaussian as well, because it is a linear combination of the $e_i^{\flat} = w_i^{\flat} - 1$'s which are i.i.d. standard normal. Unfortunately, the real world score ∇ is a linear combination of the errors ε_i which in general are not Gaussian. Therefore, we cannot assume that the score ∇ is Gaussian. However, the desirable justification of the bootstrap procedure can be obtained by GAR arguments for the score vector ∇.

7.5.3 Bootstrap for the SmA procedure

Now we return to the SmA procedure for the sieve GLM estimation scheme. For each m, we consider the sieve estimate $\widetilde{\boldsymbol{\theta}}_m$ and its bootstrap counterpart $\widetilde{\boldsymbol{\theta}}_m^{\flat}$ and try to design the procedure by its desired performance in the bootstrap world where we know the true value $\widetilde{\boldsymbol{\theta}}_m$ in each sieve. We use the Fisher expansion as the main tool. Similarly to the real world case, it allows to decompose the difference $\widetilde{\boldsymbol{\theta}}_m^{\flat} - \widetilde{\boldsymbol{\theta}}_m$ into the deterministic and stochastic part. For the SmA procedure, we have to consider the pairs $\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ}$ for all $m > m^\circ$ and their bootstrap versions. The counterpart of the real world expansion (7.54) looks as

$$\boldsymbol{\xi}_{m,m^\circ}^{\flat} = W \bigl(D_m^{-2} \Pi_m - D_{m^\circ}^{-2} \Pi_{m^\circ} \bigr) \breve{\nabla}^{\flat}$$

with $\breve{\nabla}^{\flat} = \Psi \mathcal{E}^{\flat} \breve{\boldsymbol{\varepsilon}}$. Again, as in the linear case, for m° fixed, all the $\boldsymbol{\xi}_{m,m^\circ}^{\flat}$ are the deterministic linear functions of the score $\breve{\nabla}^{\flat}$, and this is exactly as for the real world stochastic vectors $\boldsymbol{\xi}_{m,m^\circ}$. Therefore, it suffices to compare the distribution of ∇ and of $\breve{\nabla}^{\flat}$. These two distributions do not depend on m° and they are close to each other in probability if the small modeling bias condition is fulfilled for the largest considered

model p. Therefore, the bootstrap-adjusted critical values $\mathtt{z}_{m,m^\circ}$ can be used in the real world, and the central propagation result continues to hold: any good model will be accepted with a high probability.

8

SmA and parameter tuning in high dimensional regression

Extending the methods and results on model selection to situation with a large or even huge parameter dimension is one of the main challenge of modern statistics. An important requirement to any such method is an automatic parameter tuning. If a proposed procedure involves some tuning parameters without explaining their automatic choice, then one problem is just replaced by another. This chapter focuses a special problem of subset selection for linear models with unknown noise structure. We aim to bring together the SmA procedure of Section 5.1 and the resampling idea of Chapter 7. In this chapter we restrict ourselves to a linear model: the observation vector $\boldsymbol{Y} \in \mathbb{R}^n$ is described by the equation

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

for a given dictionary Ψ. We allow that the dictionary is overcomplete, that is, the dimension p of the vector $\boldsymbol{\theta}^*$ can be much larger than the number of observations n. Some structural assumptions are necessary to make the problem of recovering $\boldsymbol{\theta}^*$ meaningful. As in Chapter 7 we assume that $\boldsymbol{\theta}^*$ is sparse or can be approximated by a sparse vector. For a subset $\varkappa$, by $\widetilde{\boldsymbol{\theta}}_\varkappa$ we denote the corresponding LSE $\widetilde{\boldsymbol{\theta}}_\varkappa$:

$$\widetilde{\boldsymbol{\theta}}_\varkappa = \left(\Psi_\varkappa \Psi_\varkappa^\top\right)^{-1} \Psi_\varkappa \boldsymbol{Y} = D_\varkappa^{-2} \Pi_\varkappa \Psi \boldsymbol{Y},$$

where $\Psi_\varkappa = \Pi_\varkappa \Psi$ with $\Pi_\varkappa$ being a projector on the $\varkappa$-subspace of the $\boldsymbol{\theta}$-space and $\nabla = \Psi \boldsymbol{\varepsilon}$ is the score vector in $\mathbb{R}^p$. By $D_\varkappa^{-2}$ we denote the pseudo inverse of the matrix $D_\varkappa^2 = \Psi_\varkappa \Psi_\varkappa^\top$. Given a weighing loss matrix W, the SmA procedure can be applied as soon as the family of tail functions $z_{\varkappa,\varkappa^\circ}(\mathtt{x})$ for the norm of the stochastic component

$$\boldsymbol{\xi}_{\varkappa,\varkappa^\circ} = W(D_\varkappa^{-2} - D_{\varkappa^\circ}^{-2})\nabla = \mathcal{S}_{\varkappa,\varkappa^\circ} \nabla$$

is fixed. Prior information about the noise $\boldsymbol{\varepsilon}$ makes this possible. Here we discuss how this information can be recovered from the data using a resampling method.

8.1 SmA subset selection in high dimensional regression

This section discusses the parameter choice in the problem of subset selection for a high dimensional linear regression model with unknown noise structure. The SmA procedure of Section 5.1 will be extended by the bootstrap step for choosing the critical values $\mathtt{z}_{\varkappa,\varkappa^\circ}$. We follow the approach of previous chapters and consider for a given sample $\boldsymbol{Y}$ a family of Gaussian multipliers $\boldsymbol{w}^\flat = (w_i^\flat)$. We also need a pilot estimate $\widetilde{\boldsymbol{\theta}}$ which removes most of systematic part from the data $\boldsymbol{Y}$. Further we proceed with the residuals $\breve{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \Psi^\top \widetilde{\boldsymbol{\theta}}$ for the bootstrap step. For each pair $\varkappa > \varkappa^\circ$, the bootstrap counterpart $\boldsymbol{\xi}^\flat_{\varkappa,\varkappa^\circ}$ of $\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}$ looks as

$$\boldsymbol{\xi}^\flat_{\varkappa,\varkappa^\circ} = W\bigl(D_\varkappa^{-2}\Pi_\varkappa - D_{\varkappa^\circ}^{-2}\Pi_{\varkappa^\circ}\bigr)\Psi\mathcal{E}^\flat\breve{\boldsymbol{\varepsilon}},$$

where $\mathcal{E}^\flat$ is the diagonal matrix of the centered bootstrap multipliers; cf. (7.47). The proposed approach uses the bootstrap tail function $z_{\varkappa,\varkappa^\circ}(\mathtt{x})$ as a proxy for the true one:

$$I\!\!P^\flat\bigl(\|\boldsymbol{\xi}^\flat_{\varkappa,\varkappa^\circ}\| > z_{\varkappa,\varkappa^\circ}(\mathtt{x})\bigr) \le \mathrm{e}^{-\mathtt{x}}.$$

After each tail function $z_{\varkappa,\varkappa^\circ}(\mathtt{x})$ is built, one can apply either uniform or multilevel synchronization of such tail functions for fixing the set of critical values $\mathtt{z}_{\varkappa,\varkappa^\circ}$. The main argument in validating this bootstrap procedure in the ordered case was that all real-world stochastic terms $\boldsymbol{\xi}_{\varkappa,\varkappa^\circ}$ and bootstrap-world stochastic components $\boldsymbol{\xi}^\flat_{\varkappa,\varkappa^\circ}$ are deterministic linear functions of the corresponding scores: $\breve{\nabla}^\flat = \Psi\mathcal{E}^\flat\breve{\boldsymbol{\varepsilon}}$ in the bootstrap world and $\nabla = \Psi\boldsymbol{\varepsilon}$ in the underlying model. This is still the case.

If the errors $\boldsymbol{\varepsilon}$ are Gaussian then the score vector $\nabla = \Psi\boldsymbol{\varepsilon}$ is Gaussian as well. The bootstrap score $\breve{\nabla}^\flat$ is Gaussian under the bootstrap measure $I\!\!P^\flat$ by construction. So, validation of the bootstrap procedure can now be restated as comparison of two Gaussian distributions in a rather special sense. Below we admit a possibly inhomogeneous noise. The only necessary assumptions are independence of the errors ε_i and a kind of the Lindeberg condition on the rows of the matrix Ψ. If $\psi_j = (\psi_{i,j})$ denotes the jth row of Ψ, then each component $\nabla_j = \psi_j\boldsymbol{\varepsilon}$ of the score $\nabla = \Psi\boldsymbol{\varepsilon}$ reads

$$\nabla_j = \sum_{i=1}^n \psi_{i,j}\varepsilon_i$$

Under the assumption that the errors ε_i are Gaussian, the same holds for ∇_j:

$$\nabla_j \sim \mathcal{N}(0, \mathrm{v}_j^2), \qquad \mathrm{v}_j^2 = \sum_{i=1}^n \psi_{i,j}^2\sigma_i^2.$$

The corresponding bootstrap-score component can be represented as

$$\nabla_j^\flat = \sum_{i=1}^n \psi_{i,j} \breve{\varepsilon}_i w_i^\flat ,$$

where $\breve{\varepsilon}_i$ are the components of $\breve{\boldsymbol{\varepsilon}}$. With Gaussian multipliers, it is normal as well conditioned on the data $\boldsymbol{Y}$:

$$\nabla_j^\flat \sim \mathcal{N}(0, \mathrm{v}_j^{\flat 2}), \qquad \mathrm{v}_j^{\flat 2} = \sum_{i=1}^n \psi_{i,j}^2 \breve{\varepsilon}_i^2 .$$

Even if we ignore the systematic component in the residuals $\breve{\varepsilon}_i$ and use ε_i in place of $\breve{\varepsilon}_i$, we can only hope that two covariances v_j^2 and $\mathrm{v}_j^{\flat 2}$ are close to each other with high probability under Lindeberg type conditions on ε_i. The same applies to cross-covariance of the different component of ∇ and similar components of $\breve{\nabla}^\flat$. Therefore, the problem can be reduced to comparing of two high dimensional Gaussian measures with similar covariance structure. Unfortunately the dimension p of the vectors ∇ and $\breve{\nabla}^\flat$ can be very large. The tools of previous chapters based on the Pinsker inequality hardly apply here, because the dimension p enters in the error bound. In regular situation the error term is of order $\sqrt{p/n}$, and this value is not small if p is larger than n. One needs other technique which applies in a very high dimensional space.

8.1.1 Norm comparison for a family of Gaussian vectors

The SmA procedure involves a multiple comparison of many test statistics each of them is based on the norm of a Gaussian vector. Suppose we are given a family of vectors $\{\boldsymbol{\xi}_\varkappa, \varkappa \in \mathcal{M}_m\}$ each of dimension m. For each $\boldsymbol{\xi}_\varkappa$ suppose that the corresponding critical value $\mathtt{z}_\varkappa$ is fixed in a way that

$$I\!P\Big(\bigcup_{\varkappa \in \mathcal{M}_m} \{\|\boldsymbol{\xi}_\varkappa\| > \mathtt{z}_\varkappa\} \Big) \le \mathrm{e}^{-\mathtt{x}}.$$

We now want to control a similar deviation probability for another family of vectors $\{\boldsymbol{\xi}_\varkappa^\flat\}$ whose covariance structure is close to that of $\{\boldsymbol{\xi}_\varkappa\}$. To justify the bootstrap procedure in the case of high parameter dimension, we need to reduce this problem to the problem of comparison of maxima for two large Gaussian vectors.

Let $\mathcal{S}_m$ denote a unit sphere in $I\!R^m$. We use that for any vector $\boldsymbol{\xi}$ in $I\!R^m$, it holds

$$\|\boldsymbol{\xi}\| = \sup_{\boldsymbol{\gamma} \in \mathcal{S}_m} \boldsymbol{\gamma}^\top \boldsymbol{\xi}.$$

Further we have to replace the maximum over the whole sphere by the maximum over a finite subset. Given $\delta > 0$, consider a finite δ-net $\mathcal{S}_m(\delta)$ in $\mathcal{S}_m$. It is obvious that

$$(1-\delta)\|\boldsymbol{\xi}\| \le \max_{\mathcal{S}_m(\delta)} \boldsymbol{\gamma}^\top \boldsymbol{\xi} \le \|\boldsymbol{\xi}\|. \tag{8.1}$$

Putting together for all $\varkappa$ implies

$$\begin{aligned} I\!P\Big(\bigcup_{\varkappa \in \mathcal{M}_m} \bigcup_{\gamma \in \mathcal{S}_m} \Big\{ \frac{1}{\mathtt{z}_\varkappa} \gamma^\top \boldsymbol{\xi}_\varkappa > 1 \Big\} \Big) &\le I\!P\Big(\bigcup_{\varkappa \in \mathcal{M}_m} \big\{ \|\boldsymbol{\xi}_\varkappa\| > \mathtt{z}_\varkappa \big\} \Big) \\ &\le I\!P\Big(\bigcup_{\varkappa \in \mathcal{M}_m} \bigcup_{\gamma \in \mathcal{S}_m(\delta)} \Big\{ \frac{1}{\mathtt{z}_\varkappa} \gamma^\top \boldsymbol{\xi}_\varkappa > 1 - \delta \Big\} \Big) \end{aligned}$$

Now introduce the vector $\boldsymbol{X}$ of dimension $\mathbb{M}_m = \mathbb{M}_m(\delta) = |\mathcal{S}_m(\delta)| \times |\mathcal{M}_m|$ with the entries $\mathtt{z}_\varkappa^{-1} \gamma^\top \boldsymbol{\xi}_\varkappa$ for $\gamma \in \mathcal{S}_m(\delta)$ and $\varkappa \in \mathcal{M}_m$.

Below we consider a special case when $\boldsymbol{\xi}_\varkappa = \Pi_\varkappa \boldsymbol{\xi}$ for some linear mapping $\Pi_\varkappa \colon I\!R^p \to I\!R^m$ for $\varkappa \in \mathcal{M}_m$. Suppose also that another Gaussain zero mean vector $\boldsymbol{\xi}^\flat$ is given and $\boldsymbol{\xi}_\varkappa^\flat = \Pi_\varkappa \boldsymbol{\xi}^\flat$. Build a vector $\boldsymbol{X}^\flat$ out of the $\boldsymbol{\xi}_\varkappa^\flat$'s in the same way as $\boldsymbol{X}$ was constructed out of the $\boldsymbol{\xi}_\varkappa$'s. As a next step we evaluate the distance between two covariance operators for $\boldsymbol{X}$ and $\boldsymbol{X}^\flat$. Let $\Sigma = \mathrm{Var}(\boldsymbol{\xi})$. Obviously, for each two pairs $(\varkappa, \gamma)$ and $(\varkappa_1, \gamma_1)$,

$$\begin{aligned} I\!E\big[\gamma^\top \boldsymbol{\xi}_\varkappa \gamma_1^\top \boldsymbol{\xi}_{\varkappa_1} \big] &= I\!E\big[\gamma^\top \boldsymbol{\xi}_\varkappa \boldsymbol{\xi}_{\varkappa_1}^\top \gamma_1 \big] = I\!E\big[\gamma^\top \Pi_\varkappa \boldsymbol{\xi}\, \boldsymbol{\xi}^\top \Pi_{\varkappa_1}^\top \gamma_1 \big] \\ &= \gamma^\top \Pi_\varkappa \Sigma \Pi_{\varkappa_1}^\top \gamma_1 \le \|\Pi_\varkappa \Sigma \Pi_{\varkappa_1}^\top\|. \end{aligned} \tag{8.2}$$

A similar formula holds for the covariance operator of $\boldsymbol{\xi}^\flat$. Below we denote

$$\square_m \stackrel{\mathrm{def}}{=} \max_{\varkappa, \varkappa_1 \in \mathcal{M}_m} \|\Pi_\varkappa (\Sigma - \Sigma^\flat) \Pi_{\varkappa_1}^\top\|. \tag{8.3}$$

To be done: Exp-Bernstein inequality implies $\square_m \le \mathtt{C} n^{-1/2} \log(p)$

Then (8.2) and (8.3) imply for $\mathtt{z}_\varkappa \ge 1$

$$\|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{X}^\flat}\|_\infty \le \square_m .$$

This and the result (8.15) of Theorem 8.2.2 imply

$$\begin{aligned} &I\!P\Big(\max_{\varkappa \in \mathcal{M}_m} \max_{\gamma \in \mathcal{S}_m(\delta)} \frac{1}{\mathtt{z}_\varkappa} \gamma^\top \boldsymbol{\xi}_\varkappa \ge 1 \Big) \\ &\quad \le I\!P\Big(\max_{\varkappa \in \mathcal{M}_m} \max_{\gamma \in \mathcal{S}_m(\delta)} \frac{1}{\mathtt{z}_\varkappa} \gamma^\top \boldsymbol{\xi}_\varkappa^\flat \ge 1 - 2\Delta \Big) + 2\Delta^{-2} \big\{ \log(\mathbb{M}_m) + 1 \big\} \square_m . \end{aligned}$$

Together with the norm approximation (8.1) this yields

$$\begin{aligned} &I\!P\Big(\max_{\varkappa \in \mathcal{M}_m} \frac{1}{\mathtt{z}_\varkappa} \|\boldsymbol{\xi}_\varkappa\| \ge 1 \Big) \\ &\quad \le I\!P\Big(\max_{\varkappa \in \mathcal{M}_m} \frac{1}{\mathtt{z}_\varkappa} \|\boldsymbol{\xi}_\varkappa^\flat\| \ge (1-\delta)(1-2\Delta) \Big) + 2\Delta^{-2} \big\{ \log(\mathbb{M}_m) + 1 \big\} \square_m . \end{aligned} \tag{8.4}$$

It remains to account for δ. The cardinality $|\mathcal{S}_m(\delta)|$ can be roughly upper bounded by $(1 + 2\delta^{-1})^m$, see Lemma 5.2 in www-personal.umich.edu/$\sim$romanv/papers/non-asymptotic-rmt-plain.pdf, yielding

$$\mathbb{M}_m < |\mathcal{M}_m|\,(1+2\delta^{-1})^m. \tag{8.5}$$

The above calculus with $\delta = \Delta$ imply the following result.

Theorem 8.1.1. *Let $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^\flat$ be two zero mean Gaussian vectors in $I\!R^p$, and let $\{\Pi_\varkappa, \varkappa \in \mathcal{M}_m\}$ be a collection of linear mappings $I\!R^{\mathcal{M}} \to I\!R^m$. For the quantity $\square_m$ from (8.3) and for any set of critical values $\mathtt{z}_\varkappa \geq 1$, it holds with any $\Delta < 1/3$*

$$\begin{aligned} &I\!P\Big(\max_{\varkappa\in\mathcal{M}_m} \frac{1}{\mathtt{z}_\varkappa}\|\boldsymbol{\xi}_\varkappa\| \geq 1\Big) \\ &\quad \leq I\!P\Big(\max_{\varkappa\in\mathcal{M}_m} \frac{1}{\mathtt{z}_\varkappa}\|\boldsymbol{\xi}^\flat_\varkappa\| \geq 1-3\Delta\Big) + 2\Delta^{-2}\Big\{\log(|\mathcal{M}_m|) + m\log\big(1+2/\Delta\big)\Big\}\square_m\,. \end{aligned}$$

Proof. Apply (8.4) and (8.5) and use that $(1-\Delta)(1-2\Delta) \geq 1-3\Delta$.

If $\mathcal{M}_m$ is the set of all subsets of the full index set $\{1,\ldots,p\}$, then by the Stirling formula for $m!$

$$\log|\mathcal{M}_m| \leq \log\binom{p}{m} \leq \log\big(p^m/m!\big) \leq m\log(ep/m).$$

In particular, with $\Delta < 1/3$ we obtain

$$\begin{aligned} &I\!P\Big(\max_{\varkappa\in\mathcal{M}_m} \frac{1}{\mathtt{z}_\varkappa}\|\boldsymbol{\xi}_\varkappa\| \geq 1\Big) \\ &\quad \leq I\!P^\flat\Big(\max_{\varkappa\in\mathcal{M}_m} \frac{1}{\mathtt{z}_\varkappa}\|\boldsymbol{\xi}^\flat_\varkappa\| \geq 1-3\Delta\Big) + 2\Delta^{-2}m\log\Big(\frac{2ep}{m\Delta}\Big)\square_m\,. \end{aligned}$$

To be done: put altogether and find the bound on n, m, and p

8.2 Gaussian comparison in high dimension

If the dimension p of the vector $\boldsymbol{\theta}$ is large, the approach of Section 7.3 based on the Pinsker inequality faces a crucial problem: the error term is proportional to $p^{1/2}$ and can be very large. One needs a different technique which allows comparing two Gaussian measures in a high dimensional space in terms of the corresponding covariance operators.

8.2.1 Stein identity, Slepian bridge, and Gaussian comparison

Below for a $\mathbb{M}\times\mathbb{M}$ matrix A, we denote

$$\begin{aligned} \|A\| &= \sup_{\|\boldsymbol{u}\|=1}\|A\boldsymbol{u}\|, & \|A\|_\infty &= \max_{i,j}|a_{i,j}|, \\ \|A\|_1 &= \sum_{i,j}|a_{i,j}|, & \|A\|^2_{Fr} &= \sum_{ij} a^2_{ij}. \end{aligned}$$

Lemma 8.2.1. *Let* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *be two zero mean Gaussian vectors in* $I\!R^{\mathbb{M}}$ *with* $\Sigma_{\boldsymbol{X}} = \mathrm{Var}(\boldsymbol{X})$ *and* $\Sigma_{\boldsymbol{Y}} = \mathrm{Var}(\boldsymbol{Y})$. *Let also* $f(\boldsymbol{x})$ *be a smooth function on* $I\!R^{\mathbb{M}}$. *Then*

$$\epsilon \stackrel{\mathrm{def}}{=} \bigl| I\!E f(\boldsymbol{X}) - I\!E f(\boldsymbol{Y}) \bigr| \leq \frac{1}{2} \, \| \Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}} \|_{\infty} \, \| \nabla^2 f \|_{1,\infty} \, , \tag{8.6}$$

where $\| \nabla^2 f \|_{1,\infty} \stackrel{\mathrm{def}}{=} \sup_{\boldsymbol{x}} \| \nabla^2 f(\boldsymbol{x}) \|_1$.

Proof. Without loss of generality assume that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are given on the same probability space and independent. For each $t \in [0,1]$, define

$$\begin{aligned} \boldsymbol{Z}(t) &\stackrel{\mathrm{def}}{=} \sqrt{t}\, \boldsymbol{X} + \sqrt{1-t}\, \boldsymbol{Y}, \\ \Psi(t) &\stackrel{\mathrm{def}}{=} I\!E f(\boldsymbol{Z}(t)) = I\!E f\bigl(\sqrt{t}\, \boldsymbol{X} + \sqrt{1-t}\, \boldsymbol{Y} \bigr). \end{aligned}$$

Obviously

$$\epsilon = |\Psi(1) - \Psi(0)| = \biggl| \int_0^1 \Psi'(t) dt \biggr|. \tag{8.7}$$

Further,

$$\Psi'(t) = I\!E[\nabla f(\boldsymbol{Z}(t))^{\top} \boldsymbol{Z}'(t)] = \frac{1}{2} I\!E\bigl[\bigl\{ t^{-1/2} \boldsymbol{X} - (1-t)^{-1/2} \boldsymbol{Y} \bigr\}^{\top} \nabla f(\boldsymbol{Z}(t)) \bigr].$$

To compute this expectation, we apply the *Stein identity*. Let $\boldsymbol{W}$ be a zero mean Gaussian vector in $I\!R^{\mathbb{M}}$. Then for any C^1 function $s \colon I\!R^{\mathbb{M}} \to I\!R^{\mathbb{M}}$, it holds

$$I\!E[\boldsymbol{W} \, s(\boldsymbol{W})] = \mathrm{Var}(\boldsymbol{W}) \, I\!E[\nabla s(\boldsymbol{W})]. \tag{8.8}$$

Exercise 8.2.1. Prove (8.9) for standard normal $\boldsymbol{W}$ using integration by part:

$$\int_{I\!R^{\mathbb{M}}} s(\boldsymbol{w}) \, \boldsymbol{w} \, \mathrm{e}^{-\|\boldsymbol{w}\|^2/2} d\boldsymbol{w} = \int_{I\!R^{\mathbb{M}}} \nabla s(\boldsymbol{w}) \, \mathrm{e}^{-\|\boldsymbol{w}\|^2/2} d\boldsymbol{w}.$$

Reduce the case of a Gaussian zero mean $\boldsymbol{w} \sim \mathcal{N}(0, \Sigma)$ with a positive symmetric matrix Σ to the case $\Sigma = I_{\mathbb{M}}$.

This results can be directly extended to any C^1 vector function $\boldsymbol{s} \colon I\!R^{\mathbb{M}} \to I\!R^q$: it holds

$$I\!E[\boldsymbol{W} \, \boldsymbol{s}(\boldsymbol{W})^{\top}] = \mathrm{Var}(\boldsymbol{W}) \, I\!E[\nabla \boldsymbol{s}(\boldsymbol{W})^{\top}]. \tag{8.9}$$

Here $\nabla \boldsymbol{s}(\boldsymbol{w})^{\top}$ means the $p \times q$ matrix with the entries $\frac{d}{d\theta_j} s_m(\boldsymbol{w})$ for $j = 1, \ldots, p$ and $m = 1, \ldots, q$.

Exercise 8.2.2. Derive (8.9) by applying (8.8) columnwise.

The identity (8.9) is used with $\boldsymbol{W} = (\boldsymbol{X}^\top, \boldsymbol{Y}^\top)^\top$ and $\boldsymbol{s}(\boldsymbol{w}) = \nabla f\bigl(\boldsymbol{z}(t)\bigr)$ for $\boldsymbol{z}(t) = \sqrt{t}\,\boldsymbol{x} + \sqrt{1-t}\,\boldsymbol{y}$. Independence of $\boldsymbol{X}$ and $\boldsymbol{Y}$ implies

$$\operatorname{Var}(\boldsymbol{W}) = \begin{pmatrix} \Sigma_{\boldsymbol{X}} & 0 \\ 0 & \Sigma_{\boldsymbol{Y}} \end{pmatrix}.$$

Also $\nabla \boldsymbol{s}(\boldsymbol{w}) = \bigl(t^{1/2}\,\nabla^2 f(\boldsymbol{z}(t)), (1-t)^{1/2}\,\nabla^2 f(\boldsymbol{z}(t))\bigr)^\top$ and by (8.9)

$$\begin{aligned} I\!\!E\bigl[\nabla f(\boldsymbol{Z}(t))\boldsymbol{X}^\top\bigr] &= t^{1/2}\Sigma_{\boldsymbol{X}}\, I\!\!E\bigl[\nabla^2 f\bigl(\boldsymbol{Z}(t)\bigr)\bigr] \\ I\!\!E\bigl[\nabla f(\boldsymbol{Z}(t))\boldsymbol{Y}^\top\bigr] &= (1-t)^{1/2}\Sigma_{\boldsymbol{Y}}\, I\!\!E\bigl[\nabla^2 f\bigl(\boldsymbol{Z}(t)\bigr)\bigr], \end{aligned}$$

This and (8.12) imply

$$\begin{aligned} \bigl|\Psi'(t)\bigr| &\le \frac{1}{2}\Bigl|\operatorname{tr}\bigl\{\bigl(\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\bigr) I\!\!E\bigl[\nabla^2 f\bigl(\boldsymbol{Z}(t)\bigr)\bigr]\bigr\}\Bigr| \\ &\le \frac{1}{2}\,\|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\|_\infty \bigl\|I\!\!E\bigl[\nabla^2 f\bigl(\boldsymbol{Z}(t)\bigr)\bigr]\bigr\|_1 \le \frac{1}{2}\,\|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\|_\infty \,\|\nabla^2 f\|_{1,\infty}\,. \end{aligned}$$

Now the assertion follows from (8.7).

Now we apply the obtained bound to $f(\boldsymbol{x}) \stackrel{\text{def}}{=} g\bigl(\Delta^{-1} h_\beta(\boldsymbol{x})\bigr)$, where $g(z)$ is a smooth univariate function with bounded first and second derivatives, and the *smooth maximum* function: for some $\beta > 0$

$$h_\beta(\boldsymbol{x}) = \beta^{-1} \log\Bigl(\sum_j \mathrm{e}^{\beta x_j}\Bigr). \tag{8.10}$$

Lemma 8.2.2. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two zero mean Gaussian vectors in $I\!R^{\mathbb{M}}$ with $\Sigma_{\boldsymbol{X}} = \operatorname{Var}(\boldsymbol{X})$ and $\Sigma_{\boldsymbol{Y}} = \operatorname{Var}(\boldsymbol{Y})$. For a univariate function $g(z)$ with bounded first and second derivatives, and $h_\beta(\boldsymbol{x})$ from (8.10)*

$$\bigl|I\!\!E g\bigl(\Delta^{-1} h_\beta(\boldsymbol{X})\bigr) - I\!\!E g\bigl(\Delta^{-1} h_\beta(\boldsymbol{Y})\bigr)\bigr| \le \Bigl(\frac{\beta\|g'\|_\infty}{\Delta} + \frac{\|g''\|_\infty}{2\Delta^2}\Bigr)\,\|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\|_\infty\,. \tag{8.11}$$

Proof. It holds for $f(\boldsymbol{x}) \stackrel{\text{def}}{=} g\bigl(\Delta^{-1} h_\beta(\boldsymbol{x})\bigr)$

$$\begin{aligned} \nabla f(\boldsymbol{x}) &= \Delta^{-1} g'(\Delta^{-1} h_\beta(\boldsymbol{x}))\nabla h_\beta(\boldsymbol{x}), \\ \nabla^2 f(\boldsymbol{x}) &= \Delta^{-1} g'(\Delta^{-1} h_\beta(\boldsymbol{x}))\nabla^2 h_\beta(\boldsymbol{x}) + \Delta^{-2} g''(\Delta^{-1} h_\beta(\boldsymbol{x}))\nabla h_\beta(\boldsymbol{x})\nabla h_\beta(\boldsymbol{x})^\top. \end{aligned}$$

Also for any $\boldsymbol{x}$ by direct calculus

$$\begin{aligned} \|\nabla h_\beta(\boldsymbol{x})\|_1 &= 1, \\ \|\nabla^2 h_\beta(\boldsymbol{x})\|_1 &\le 2\beta. \end{aligned} \tag{8.12}$$

This implies

$$\begin{aligned}\|\nabla f(\boldsymbol{x})\|_1 &\le \Delta^{-1}\|g'\|_\infty \times \|\nabla h_\beta(\boldsymbol{x})\|_1 \le \Delta^{-1}\|g'\|_\infty\,,\\ \|\nabla^2 f(\boldsymbol{x})\|_1 &\le \Delta^{-1}\|g'\|_\infty \times \|\nabla^2 h_\beta(\boldsymbol{x})\|_1 + \Delta^{-2}\|g''\|_\infty \times \|\nabla h_\beta(\boldsymbol{x})\|_1^2\\ &\le 2\Delta^{-1}\beta\|g'\|_\infty + \Delta^{-2}\|g''\|_\infty\,.\end{aligned}$$

Now (8.11) follows from (8.6).

A particular choice of the function g is given by

$$g(z) \stackrel{\mathrm{def}}{=} \begin{cases} 2u^2, & u \in [0,1/2],\\ 1-2(1-u)^2, & u \in [1/2,1].\\ 0 & \text{otherwise.}\end{cases} \tag{8.13}$$

Obviously $|g'(u)| \le 2$, $|g''(u)| \le 4$ for all u. Then $\|g'_\Delta\|_\infty \le 2\Delta^{-1}$, $\|g''_\Delta\|_\infty \le 4\Delta^{-2}$. We conclude with the following bound.

Theorem 8.2.1. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two zero mean Gaussian vectors in $\mathbb{R}^{\mathbb{M}}$ with $\Sigma_{\boldsymbol{X}} = \mathrm{Var}(\boldsymbol{X})$ and $\Sigma_{\boldsymbol{Y}} = \mathrm{Var}(\boldsymbol{Y})$. Then with $g(\cdot)$ given by (8.13), it holds for any $\Delta > 0$ and $\beta > 0$*

$$\bigl|\mathbb{E}g\bigl(\Delta^{-1}h_\beta(\boldsymbol{X})\bigr) - \mathbb{E}g\bigl(\Delta^{-1}h_\beta(\boldsymbol{Y})\bigr)\bigr| \le 2(\beta\Delta^{-1} + \Delta^{-2})\,\|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\|_\infty\,. \tag{8.14}$$

8.2.2 Comparing of the maximum of Gaussians

Let $\boldsymbol{X} = (X_j)$ and $\boldsymbol{Y} = (Y_j)$ be two zero mean Gaussian vectors in $\mathbb{R}^{\mathbb{M}}$ with $\Sigma_{\boldsymbol{X}} = \mathrm{Var}(\boldsymbol{X})$ and $\Sigma_{\boldsymbol{Y}} = \mathrm{Var}(\boldsymbol{Y})$, and let $\square = \|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\|_\infty$. Now we aim at comparing the distributions of $\max_j X_j$ and $\max_j Y_j$. We use that the smooth maximum h_β fulfills

$$\max_j x_j \le h_\beta(\boldsymbol{x}) \le \max_j x_j + \beta^{-1}\log(\mathbb{M}).$$

As the indicator function $\mathbb{1}(z \ge 0)$ is not differentiable, we approximate it by a smooth function g_Δ. Namely, select a two times differentiable function g with $g(u) = 0$ for $u \le 0$, $g(u) = 1$ for $u \ge 1$, and $g(u)$ monotonously grows from zero to one when u grows from zero to one. Define also $g_\Delta(u) = g(\Delta^{-1}u)$ for $\Delta > 0$. With $\Delta = \beta^{-1}\log(\mathbb{M})$

$$g_\Delta \circ h_\beta(\boldsymbol{x} - \boldsymbol{\Delta}) \le \mathbb{1}(\max_j x_j > 0) \le g_\Delta \circ h_\beta(\boldsymbol{x} + \boldsymbol{\Delta}).$$

Here $\boldsymbol{\Delta}$ is the vector with all entries equal to Δ. Indeed, $g_\Delta(z) \in [0,1]$ for any z. If $x_j \ge 0$ for some j, then $h_\beta(\boldsymbol{x} + \boldsymbol{\Delta}) \ge \Delta$ and hence,

$$g_{\Delta} \circ h_{\beta}(\boldsymbol{x} + \boldsymbol{\Delta}) \geq g(\Delta/\Delta) = g(1) = 1.$$

Similarly, if $\max_j x_j \leq 0$, then due to $\Delta = \beta^{-1} \log(\mathbb{M})$

$$h_{\beta}(\boldsymbol{x} - \boldsymbol{\Delta}) \leq \max_j (x_j - \Delta) + \beta^{-1} \log(\mathbb{M}) \leq 0$$

and $g_{\Delta} \circ h_{\beta}(\boldsymbol{x} - \boldsymbol{\Delta}) = 0$. This and (8.14) yield the bound

$$\begin{aligned} I\!P\bigl(\max_j X_j > 0\bigr) &\leq I\!E\bigl[g_{\Delta} \circ h_{\beta}(\boldsymbol{X} + \boldsymbol{\Delta})\bigr] \\ &\leq I\!E\bigl[g_{\Delta} \circ h_{\beta}(\boldsymbol{Y} + \boldsymbol{\Delta})\bigr] + 2(\beta \Delta^{-1} + \Delta^{-2})\, \square \\ &\leq I\!P\bigl(\max_j Y_j > -2\Delta\bigr) + 2\Delta^{-2}\bigl\{\log(\mathbb{M}) + 1\bigr\}\, \square. \end{aligned}$$

Similarly one can approximate any indicator $\mathbb{1}(z \geq z_0)$ by shifting the function g.

Theorem 8.2.2. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two zero mean Gaussian vectors in $I\!R^{\mathbb{M}}$ with $\Sigma_{\boldsymbol{X}} = \mathrm{Var}(\boldsymbol{X})$ and $\Sigma_{\boldsymbol{Y}} = \mathrm{Var}(\boldsymbol{Y})$.*

$$\square \stackrel{\mathrm{def}}{=} \|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\|_{\infty},$$

it holds for any Δ and z

$$I\!P\bigl(\max_j X_j > z\bigr) \leq I\!P\bigl(\max_j Y_j > z - 2\Delta\bigr) + 2\Delta^{-2}\bigl\{\log(\mathbb{M}) + 1\bigr\}\, \square. \tag{8.15}$$

8.2.3 Anti-concentration for Gaussian maxima

This section explains how one can compare the distribution of two maxima using the anti-concentration bound. The obtained results allow to bound the probability $I\!P\bigl(\max_j X_j > 0\bigr)$ by a similar probability $I\!P\bigl(\max_j Y_j > -2\Delta\bigr)$ from above and $I\!P\bigl(\max_j Y_j > 2\Delta\bigr)$ from below up to the error term $2\Delta^{-2}\bigl\{\log(\mathbb{M}) + 1\bigr\}\, \square$. The next question is whether we can replace 2Δ or -2Δ by zero without an essential change of probability. In other words, we have to bound the difference

$$I\!P\bigl(\max_j Y_j > -2\Delta\bigr) - I\!P\bigl(\max_j Y_j > 0\bigr).$$

The following theorem provides bounds on the Lévy concentration function of the maximum of a Gaussian random vector in $I\!R^{\mathbb{M}}$, where the terminology is borrowed from Chernozhukov et al. (2013). The Lévy concentration function of a real valued random variable ξ is defined for $\varepsilon > 0$ as

$$\mathcal{L}(\xi, \varepsilon) = \sup_{x \in I\!R} I\!P(|\xi - x| \leq \varepsilon).$$

Theorem 8.2.3 (Anti-concentration). *Let* $(X_1, \ldots, X_{p_n})^\top$ *be a centered Gaussian random vector in* $\mathbb{R}^{\mathbb{M}}$ *with* $\sigma_j^2 = \mathbb{E}[X_j^2] > 0$ *for all* $1 \le j \le \mathbb{M}$. *Moreover, let* $\underline{\sigma} = \min_{1\le j\le \mathbb{M}} \sigma_j$, $\overline{\sigma} = \max_{1\le j\le \mathbb{M}} \sigma_j$, *and* $a_{\mathbb{M}} = \mathbb{E}[\max_{1\le j\le \mathbb{M}}(X_j/\sigma_j)]$.

1. *If the variances are all equal, namely* $\underline{\sigma} = \overline{\sigma} = \sigma$, *then for every* $\epsilon > 0$,

$$\mathcal{L}\Big(\max_{1\le j\le \mathbb{M}} X_j, \epsilon\Big) \le 4\epsilon(a_{\mathbb{M}} + 1)/\sigma;$$

2. *If the variances are not equal, namely* $\underline{\sigma} < \overline{\sigma}$, *then for every* $\epsilon > 0$,

$$\mathcal{L}\Big(\max_{1\le j\le \mathbb{M}} X_j, \epsilon\Big) \le \mathtt{C}\epsilon\{a_{\mathbb{M}} + 1 \vee \log(\underline{\sigma}/\epsilon)\}$$

where $\mathtt{C} > 0$ *depends only on* $\underline{\sigma}$ *and* $\overline{\sigma}$.

To compare the distribution of two maxima, we use the anti-concentration bound: if $\mathrm{Var}(Y_j) \equiv \sigma^2$

$$I\!P\big(\max_j Y_j > 0\big) - I\!P\big(\max_j Y_j > -2\Delta\big) \le 8\Delta(a_{\mathbb{M}} + 1)/\sigma,$$

where $a_{\mathbb{M}} \stackrel{\mathrm{def}}{=} \mathbb{E} \max_j |Y_j/\sigma| \le (2\log \mathbb{M})^{1/2}$. If the variances $\sigma_j^2 \stackrel{\mathrm{def}}{=} \mathrm{Var}(Y_j)$ are unequal then

$$I\!P\big(\max_j Y_j > 0\big) - I\!P\big(\max_j Y_j > -2\Delta\big) \le \mathtt{C}\Delta\sqrt{\log(\mathbb{M}/\Delta)}.$$

We now apply all the inequalities with the following choice: with $\Delta = b^{-1} = \beta^{-1}\log(\mathbb{M})$ and $\mathbb{Q} = \mathbb{M}/\Delta$

$$b = \square^{-1/3}\{\log(\mathbb{Q})\}^{-1/6}.$$

It follows by (8.14)

$$\begin{aligned}
&\big|I\!P\big(\max_j X_j > 0\big) - I\!P\big(\max_j Y_j > 0\big)\big| \\
&\quad \le (2\beta\Delta^{-1} + 2\Delta^{-2})\,\square + \mathtt{C}\Delta\sqrt{\log(\mathbb{Q})} \\
&\quad \le \mathtt{C}\,\square^{1/3}\log^{2/3}(\mathbb{Q}) + \mathtt{C}\,\square^{1/3}\log^{2/3}(\mathbb{Q}) \\
&\quad \le \mathtt{C}\,\square^{1/3}\log^{2/3}(\mathbb{Q}).
\end{aligned}$$

The definition yields $\mathbb{Q} = \mathbb{M}/\Delta \le \mathbb{M}/\square^{1/3}$. We conclude with the following result.

Theorem 8.2.4. *Let* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *be two zero mean Gaussian vectors in* $\mathbb{R}^{\mathbb{M}}$ *with* $\Sigma_{\boldsymbol{X}} = \mathrm{Var}(\boldsymbol{X})$ *and* $\Sigma_{\boldsymbol{Y}} = \mathrm{Var}(\boldsymbol{Y})$. *With*

$$\square \stackrel{\mathrm{def}}{=} \|\Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{Y}}\|_\infty,$$

it holds for any Δ

$$\bigl|I\!P\bigl(\max_j X_j > 0\bigr) - I\!P\bigl(\max_j Y_j > 0\bigr)\bigr| \le \mathtt{C}\,\Box^{1/3}\log^{2/3}(\mathbb{M}/\Box^{1/3}).$$

9

Penalized model selection

This chapter discusses a class of procedures which can be represented as penalized minimization of the empirical risk. We consider the linear model $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ in which the parameter dimension p can be very large. The empirical risk is just the squared norm of the difference $\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}$. The penalized procedure tries to minimize this empirical risk penalized by the complexity of the vector $\boldsymbol{\theta}$ used for prediction:

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \bigl\{ \|\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}\|^2 + \mathtt{pen}(\boldsymbol{\theta}) \bigr\} \tag{9.1}$$

for a penalty function $\mathtt{pen}(\boldsymbol{\theta})$.

The roughness penalty has been already discussed in Chapter 3. The resulting estimate is again linear and the general approach of linear model selection continues to apply. Here we consider two special choices of $\mathtt{pen}(\boldsymbol{\theta})$ which are essentially non-linear. The complexity penalty $\mathtt{pen}(\boldsymbol{\theta}) = \mathtt{C}\|\boldsymbol{\theta}\|_0$ just counts the number of non-zero $\boldsymbol{\theta}$-coefficients. The sparse penalty $\mathtt{pen}(\boldsymbol{\theta}) = \mathtt{C}\|\boldsymbol{\theta}\|_q^q$ use the q-norm of $\boldsymbol{\theta}$ for some $q < 2$. The most popular sparse penalty corresponds to $q = 1$.

9.1 Complexity penalization

This section considers the important special case of penalization by complexity. The famous Akaiki criteria is a special case of such penalization. We offer another viewpoint based on the SmA idea. As previously in Section 2.1, for a subset $\varkappa$ of the index set $\{1, \ldots, p\}$, the estimate $\widetilde{\boldsymbol{\theta}}_\varkappa$ is the corresponding projection MLE:

$$\widetilde{\boldsymbol{\theta}}_\varkappa = \bigl(\Psi_\varkappa \Psi_\varkappa^\top\bigr)^{-1} \Psi_\varkappa \boldsymbol{Y}.$$

Theorem 9.1.1. *Let* $\mathtt{pen}(\boldsymbol{\theta}) = \mathtt{C}\|\boldsymbol{\theta}\|_0$. *Then the solution* $\widehat{\boldsymbol{\theta}}$ *of the problem* (9.1) *satisfies*

$$\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{\varkappa}},$$

where

$$\widehat{\varkappa} = \operatorname*{argmin}_{\varkappa} \bigl\{ \|\widetilde{\varepsilon}_{\varkappa}\|^2 + \mathtt{C}|\varkappa| \bigr\}, \tag{9.2}$$

with

$$\widetilde{\varepsilon}_{\varkappa} = \boldsymbol{Y} - \Psi_{\varkappa}^{\top} \widetilde{\boldsymbol{\theta}}_{\varkappa} = \bigl(I_n - \Pi_{\varkappa} \bigr) \boldsymbol{Y}$$

and $|\varkappa|$ *means the cardinality of* $\varkappa$ *or, equivalently the number of coefficients in the support set* $\varkappa$*.*

The unbiased risk estimation procedure corresponds to the special choice of constant $\mathtt{C} = 2\sigma^2$. This results in selecting a proper subset $\varkappa$ which provides a reasonable fit under complexity constraint:

$$\widehat{\varkappa} = \operatorname*{argmin}_{\varkappa} \bigl\{ \|\widetilde{\varepsilon}_{\varkappa}\|^2 + 2\sigma^2 |\varkappa| \bigr\}.$$

This procedure faces two essential problems when the dimension p becomes large. One of them is algorithmic: the procedure requires to compute the MLE $\widetilde{\boldsymbol{\theta}}_{\varkappa}$ and the related empirical risk for any subset $\varkappa$; such a problem is in general NP-hard and can be solved in very special cases, e.g. if the design matrix Ψ is orthogonal. Then the procedure can be reduced to thresholding of individual Fourier coefficients $\widetilde{\theta}_j = \psi_j \boldsymbol{Y}$, where ψ_j denotes the j th row of Ψ.

Theorem 9.1.2. *Let* $n \geq p$ *and the matrix* Ψ *be orthonormal, that is,* $\Psi\Psi^T = I_p$*. Then the active set* $\widehat{\varkappa}$ *from* (9.2) *is given by hard thresholding*

$$\widehat{\varkappa} = \bigl\{ j \colon |\widetilde{\theta}_j| \geq \lambda \bigr\}$$

for a proper $\lambda > 0$*. The corresponding hard thresholding estimate* $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_j)$ *reads as*

$$\widehat{\theta}_j = \begin{cases} \psi_j \boldsymbol{Y}, & |\psi_j \boldsymbol{Y}| > \lambda, \\ 0, & \text{otherwise.} \end{cases}$$

The other problem is statistical. First of all, the procedure assumes a homogeneous noise and requires that the noise variance σ^2 is given. Second, the penalization in the form $2\sigma^2|\varkappa|$ is too mild and does not ensure a proper model selection for p large. The reason is that there are very many models to select between, this number is exponential in p, and therefore, the price for this choice has to be much larger than in the ordered case.

Below we discuss several ways of solving the statistical problem. First we consider the penalized model selection procedure (9.1) or equivalently (9.2) with a data-driven constant $\mathtt{C}$. Then we extend it to a more sophisticated non-linear choice of the penalty function $\mathtt{pen}(\boldsymbol{\theta})$. Finally we discuss the saddle-point bivariate model selection.

The approach is based on the propagation idea: if the model is "good" in the sense that it provides a reasonable data fit, it should be competitive against larger models: no reason to increase the complexity if you already have a proper prediction ability. A typical situation is as follows: we have a model-candidate $\varkappa^\circ$ which is not too complex, that is, $|\varkappa^\circ|$ is small relative to the sample size n and the total dimension p. One says in such cases that $\varkappa^\circ$ is "sparse". Further, this choice $\varkappa^\circ$ is good if the coefficients θ_j^* for $j \notin \varkappa^\circ$ are nearly zero. We aim at designing a data-driven procedure such that the criterium (9.2) keeps $\varkappa^\circ$ alive in competition with all larger models. This precisely means that

$$\|\widetilde{\varepsilon}_\varkappa\|^2 + \mathtt{C}|\varkappa| \geq \|\widetilde{\varepsilon}_{\varkappa^\circ}\|^2 + \mathtt{C}|\varkappa^\circ|$$

This inequality has to be verified for all $\varkappa > \varkappa^\circ$ with a high probability. Rearranging yields in view of $\widetilde{\varepsilon}_\varkappa = (\boldsymbol{I}_n - \Pi_\varkappa)\boldsymbol{Y}$

$$\|\widetilde{\varepsilon}_{\varkappa^\circ}\|^2 - \|\widetilde{\varepsilon}_\varkappa\|^2 = \|\Pi_{\varkappa,\varkappa^\circ}\boldsymbol{Y}\|^2 \leq \mathtt{C}\bigl(|\varkappa| - |\varkappa^\circ|\bigr).$$

If we knew the noise distribution then we can fix the constant $\mathtt{C}$ in the "pure noise" situation. Indeed, the underlying structural assumption means that there is no significant signal θ_j^* for $j \notin \varkappa^\circ$, and hence, one can simply ignore such signal and consider $\theta_j^* \equiv 0$ in the complement of $\varkappa^\circ$:

$$I\!P\Bigl(\bigcup_{\varkappa > \varkappa^\circ} \bigl\{\|\Pi_{\varkappa,\varkappa^\circ}\varepsilon\|^2 > \mathtt{C}_0\bigl(|\varkappa| - |\varkappa^\circ|\bigr)\bigr\}\Bigr) \leq \mathrm{e}^{-\mathtt{x}} \tag{9.3}$$

for a constant $\mathtt{C}_0$. It is obvious that this condition becomes stronger if the set $\varkappa^\circ$ is taken smaller. The hardest case corresponds to the empty set $\varkappa^\circ$, yielding the constraint

$$I\!P\Bigl(\bigcup_{\varkappa}\bigl\{\|\Pi_\varkappa\varepsilon\|^2 > \mathtt{C}_0|\varkappa|\bigr\}\Bigr) \leq \mathrm{e}^{-\mathtt{x}}. \tag{9.4}$$

If the dimension of $\varkappa$ only slightly higher than the dimension of $\varkappa^\circ$, the inequality $\|\Pi_{\varkappa,\varkappa^\circ}\varepsilon\|^2 > \mathtt{C}_0\bigl(|\varkappa| - |\varkappa^\circ|\bigr)$ would require a very large constant $\mathtt{C}_0$. At the same time, this constant rapidly stabilizes if $|\varkappa| - |\varkappa^\circ|$ exceeds some prescribed value. This suggests to extend the condition (9.6) (resp. (9.4)): given $\tau \geq 1$

$$I\!P\Bigl(\bigcup_{\varkappa \in \mathcal{M}_\tau(\varkappa^\circ)} \bigl\{\|\Pi_{\varkappa,\varkappa^\circ}\varepsilon\|^2 > \mathtt{C}_0\bigl(|\varkappa| - |\varkappa^\circ|\bigr)\bigr\}\Bigr) \leq \mathrm{e}^{-\mathtt{x}}. \tag{9.5}$$

Here $\mathcal{M}_\tau(\varkappa^\circ) = \bigl\{\varkappa\colon \varkappa > \varkappa^\circ, |\varkappa| \geq |\varkappa^\circ| + \tau\bigr\}$.

Now suppose that such a constant $\mathtt{C}_0 = \mathtt{C}_0(\tau)$ is fixed. Then the procedure can be applied with

$$\mathtt{C} \stackrel{\text{def}}{=} \mathtt{C}_0 + \beta,$$

where β appears in the definition of a "good" choice: $\varkappa^\circ$ is good if

$$\|\Pi_{\varkappa,\varkappa^\circ} \boldsymbol{f}^*\|^2 \le \beta\bigl(|\varkappa| - |\varkappa^\circ|\bigr).$$

To be done: An upper bound on $\mathtt{C}_0$

Apply the arguments from Section 8.1.1 and matrix Bernstein inequality from Section 1.6;

To be done: Choice of τ by concentration of $\|\Pi_{\varkappa,\varkappa^\circ}\boldsymbol{\varepsilon}\|^2$

Bound the difference between $\|\Pi_{\varkappa,\varkappa^\circ}\boldsymbol{\varepsilon}\|^2$ and its expectation; Use matrix Bernstein from Section 1.6.

To be done: Oracle inequality for $\widehat{\varkappa}$-choice

The construction ensures with a dominating probability that $\widehat{\varkappa} \le \varkappa^* + \tau$.

To be done: the quantity $\|\Pi_{\varkappa,\varkappa^\circ}\boldsymbol{Y}\|^2$ is in the Gaussian case non-central χ^2. make it accurate

Now we discuss the situation when the noise distribution is unknown. Then the relation (9.4) cannot be used for fixing the constant $\mathtt{C}_0$. Below we discuss the bootstrap based choice of this constant.

9.1.1 Bootstrap based tuning

As in Section 7.3.3 we fix some model-candidate $\varkappa^\circ$ and consider the family of estimates

$$\widetilde{\boldsymbol{\theta}}_\varkappa = \bigl(\Psi_\varkappa \Psi_\varkappa^\top\bigr)^{-1} \Psi_\varkappa \boldsymbol{Y}$$

for $\varkappa > \varkappa^\circ$. We also suppose a pilot estimate $\widetilde{\boldsymbol{\theta}}$ to be given which provides a reasonable data fit but is probably too volatile. Define the residuals $\breve{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \Psi^\top \widetilde{\boldsymbol{\theta}}$. The procedure follows the same path as in the ordered case. For each $\varkappa$ we compute and store the corresponding MLE $\widetilde{\boldsymbol{\theta}}_\varkappa$ and a collection of the bootstrap-based stockstic vectors $\boldsymbol{\zeta}^\flat_\varkappa$:

$$\boldsymbol{\zeta}^\flat_\varkappa = \bigl(\Psi_\varkappa \Psi_\varkappa^\top\bigr)^{-1} \Psi_\varkappa \mathcal{E}^\flat \breve{\boldsymbol{\varepsilon}}.$$

Further we can fix the value $\mathtt{C}_0$ using the bootstrap differences

$$\boldsymbol{\xi}^\flat_{\varkappa,\varkappa^\circ} = W\bigl(\boldsymbol{\zeta}^\flat_\varkappa - \boldsymbol{\zeta}^\flat_{\varkappa^\circ}\bigr)$$

for the weighting loss matrix W. The bootstrap critical values can be computed from these differences by the bootstrap analog of the propagation condition (9.6)

$$I\!\!P^\flat\Bigl(\bigcup_{\varkappa \in \mathcal{M}_\tau(\varkappa^\circ)} \bigl\{\|\boldsymbol{\xi}^\flat_{\varkappa,\varkappa^\circ}\|^2 > \mathtt{C}_0\bigl(|\varkappa| - |\varkappa^\circ|\bigr)\bigr\}\Bigr) \le \mathrm{e}^{-\mathtt{x}}. \tag{9.6}$$

Here $I\!P^\flat$ is to be understood as the empirical bootstrap measure.

To be done: An analytic bound on $\mathtt{C}_0$

9.2 Sparse penalty

This section briefly discusses the use of a sparse penalty based on the ℓ_1-norm of the vector $\boldsymbol{\theta}$. Below for a vector $\boldsymbol{\theta} \in I\!R^p$ we denote

$$\|\boldsymbol{\theta}\|_1 = \sum_{j=1}^p |\theta_j|, \qquad \|\boldsymbol{\theta}\|^2 = \sum_{j=1}^p \theta_j^2, \qquad \|\boldsymbol{\theta}\|_\infty = \max_{j \le p} |\theta_j|.$$

We consider the LASSO type procedure which is based on minimization of the empirical risk $\|\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}\|^2$ penalized by $\lambda \|\boldsymbol{\theta}\|_1$:

$$\widehat{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{\theta}} \jmath_\lambda(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \bigl\{ \|\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|_1 \bigr\}. \tag{9.7}$$

Note that the formulation of the problem implicitly assumes that all components $\boldsymbol{\theta}_j$ of the vector $\boldsymbol{\theta}$ have the same impact. This can be translated into the scaling condition on the matrix Ψ. Usually this matrix and thus, the coefficients θ_j are rescaled in a way that the diagonal elements of the matrix $\Psi\Psi^\top$ are equal to one:

$$\sum_{i=1}^n \psi_{i,j}^2 = 1.$$

The problem (9.7) has a closed form solution only in very special situation. One of them is when the matrix Ψ is orthogonal.

Theorem 9.2.1. *Let $n \ge p$ and $\Psi\Psi^\top = I_p$. Then $\widehat{\boldsymbol{\theta}}$ is obtained by soft-thresholding of $\widetilde{\boldsymbol{\theta}} = \Psi \boldsymbol{Y}$:*

$$\widehat{\theta}_j = \begin{cases} \bigl(\widetilde{\theta}_j - \lambda\bigr)_+ & \widetilde{\theta}_j \ge 0, \\ -\bigl(|\widetilde{\theta}_j| - \lambda\bigr)_+ & \widetilde{\theta}_j < 0 \end{cases}$$

However, compared to the complexity penalization, the problem (9.7) can be solved numerically because the objective function is convex.

Now we briefly discuss some properties of the solution. The underlying structural assumption is that the true vector $\boldsymbol{\theta}^*$ is sparse, that is, most of its entries vanish. By $\varkappa^*$ we denote the corresponding oracle support. We first consider the case when $\Psi_{\varkappa^*}$ is orthogonal to the rest of Ψ. Our first result shows that a proper choice of the parameter λ ensures a sparse solution: the non-zero coefficients of $\widehat{\boldsymbol{\theta}}$ are all located within $\varkappa^*$.

Theorem 9.2.2. *Let* $\boldsymbol{\theta}^*$ *be supported on* $\varkappa_0$, $\varkappa_0^c$ *be the complement of* $\varkappa_0$, *and*

$$\Psi_{\varkappa_0}\Psi_{\varkappa_0^c}^\top = 0. \tag{9.8}$$

If the coefficient λ *fulfills*

$$\|\Psi\boldsymbol{\varepsilon}\|_\infty \le \lambda,$$

then

$$\widehat{\varkappa} \le \varkappa_0 \tag{9.9}$$

Proof. It suffices to check for each candidate $\boldsymbol{\theta}$ that the criteria in the optimization problem (9.7) only improves if we kill all its entries which do not enter in the oracle set $\varkappa^*$. Let $\varkappa$ be the support of $\boldsymbol{\theta}$. Define $\boldsymbol{\theta}_{\varkappa_0} = \Pi_{\varkappa_0}\boldsymbol{\theta}$ as the restriction of the parameter vector $\boldsymbol{\theta}$ to $\varkappa_0$, and similarly $\boldsymbol{\theta}_{\varkappa_0^c} = \Pi_{\varkappa_0^c}\boldsymbol{\theta}$ is the projection on the complement set $\varkappa_0^c$. Obviously $\boldsymbol{\theta}_{\varkappa_0^c} = \boldsymbol{\theta} - \boldsymbol{\theta}_{\varkappa_0}$. Then the model equation $\boldsymbol{Y} = \Psi^\top\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ implies

$$\begin{aligned}\|\boldsymbol{Y} - \Psi^\top\boldsymbol{\theta}\|^2 - \|\boldsymbol{Y} - \Psi^\top\boldsymbol{\theta}_{\varkappa_0}\|^2 &= \|\boldsymbol{\varepsilon} - \Psi^\top(\boldsymbol{\theta}_{\varkappa_0} + \boldsymbol{\theta}_{\varkappa_0^c} - \boldsymbol{\theta}^*)\|^2 - \|\boldsymbol{\varepsilon} - \Psi^\top(\boldsymbol{\theta}_{\varkappa_0} - \boldsymbol{\theta}^*)\|^2 \\ &= \|\Psi^\top\boldsymbol{\theta}_{\varkappa_0^c}\|^2 - 2\big\{\boldsymbol{\varepsilon} - \Psi^\top(\boldsymbol{\theta}_{\varkappa_0} - \boldsymbol{\theta}^*)\big\}^\top\Psi^\top\boldsymbol{\theta}_{\varkappa_0^c}.\end{aligned}$$

Exercise 9.2.1. Check that

$$\big\{\Psi^\top(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)\big\}^\top\Psi^\top\boldsymbol{\theta}_{\varkappa_0^c} = 0.$$

Hint: use that $\boldsymbol{\theta}_{\varkappa_0} - \boldsymbol{\theta}^*$ is supported on $\varkappa^*$, and $\boldsymbol{\theta}_{\varkappa_0^c} = \boldsymbol{\theta} - \boldsymbol{\theta}_{\varkappa_0^c}$ on $\varkappa_0^c$. Then the result follows from (9.8).

Now $\|\boldsymbol{\theta}\|_1 - \|\boldsymbol{\theta}_{\varkappa_0}\|_1 = \|\boldsymbol{\theta}_{\varkappa_0^c}\|_1$ and

$$\begin{aligned}&\|\boldsymbol{Y} - \Psi^\top\boldsymbol{\theta}\|^2 - \|\boldsymbol{Y} - \Psi^\top\boldsymbol{\theta}_{\varkappa_0}\|^2 + \lambda(\|\boldsymbol{\theta}\|_1 - \|\boldsymbol{\theta}_{\varkappa_0}\|_1) \\ &\quad = \|\Psi^\top\boldsymbol{\theta}_{\varkappa_0^c}\|^2 + \lambda\|\boldsymbol{\theta}_{\varkappa_0^c}\|_1 - 2(\Psi\boldsymbol{\varepsilon})^\top\boldsymbol{\theta}_{\varkappa_0^c}.\end{aligned}$$

It remains to check that the condition $2\|\Psi\boldsymbol{\varepsilon}\|_\infty \le \lambda$ ensures that

$$\lambda\|\boldsymbol{\theta}_{\varkappa_0^c}\|_1 - 2(\Psi\boldsymbol{\varepsilon})^\top\boldsymbol{\theta}_{\varkappa_0^c} \ge 0.$$

Therefore, reduction of $\boldsymbol{\theta}$ to $\boldsymbol{\theta}_{\varkappa_0}$ only improves the objective function, and the result follows.

9.2.1 Basic inequality

If we drop the orthogonality condition (9.8), the wonderful oracle result (9.9) does not hold any more. However, one can establish an oracle bound on the quadratic risk in terms of the sparsity value $\|\boldsymbol{\theta}^*\|_1$. We again check the condition that $\boldsymbol{\theta}$ is better than $\boldsymbol{\theta}^*$ w.r.t. the criterion $\mathcal{J}_\lambda(\boldsymbol{\theta})$ from (9.7). It holds due to the model equation $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$

$$\begin{aligned}\|\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}\|^2 - \|\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}^*\|^2 &= \|\boldsymbol{\varepsilon} - \Psi^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 - \|\boldsymbol{\varepsilon}\|^2 \\ &= -2(\Psi\boldsymbol{\varepsilon})^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \|\Psi^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2.\end{aligned}$$

The event $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ is only possible if $\mathcal{J}_\lambda(\boldsymbol{\theta}) \le \mathcal{J}_\lambda(\boldsymbol{\theta}^*)$. This yields

$$\|\Psi^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + \lambda\|\boldsymbol{\theta}\|_1 \le 2(\Psi\boldsymbol{\varepsilon})^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \lambda\|\boldsymbol{\theta}^*\|_1 \tag{9.10}$$

Moreover, if the score $\nabla = \Psi\boldsymbol{\varepsilon}$ fulfills $\|\nabla\|_\infty \le \mathtt{C}_0$, then

$$2(\Psi\boldsymbol{\varepsilon})^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \le 2\mathtt{C}_0\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 \tag{9.11}$$

By the triangle inequality $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 \ge \|\boldsymbol{\theta}\|_1 - \|\boldsymbol{\theta}^*\|_1$, and (9.10) and (9.11) imply

$$\|\Psi^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + (\lambda - 2\mathtt{C}_0)\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 \le 2\lambda\|\boldsymbol{\theta}^*\|_1$$

If $\lambda > 2\mathtt{C}_0$, this inequality provides a number of informative messages. The prediction loss $\|\Psi^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2$ and the estimation loss $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1$ are can be bounded as

$$\begin{aligned}\|\Psi^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 &\le 2\lambda\|\boldsymbol{\theta}^*\|_1, \\ \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 &\le \frac{2\lambda}{\lambda - 2\mathtt{C}_0}\|\boldsymbol{\theta}^*\|_1 .\end{aligned}$$

The last inequality can be made more exact if we consider for any $\boldsymbol{\theta}$ the decomposition $\boldsymbol{\theta} = \boldsymbol{\theta}_{\varkappa_0} + \boldsymbol{\theta}_{\varkappa_0^c}$, where $\boldsymbol{\theta}_{\varkappa_0}$ is supported on $\varkappa^*$ and $\boldsymbol{\theta}_{\varkappa_0^c}$ on its complement $\varkappa_0^c$. Obviously $\|\boldsymbol{\theta}\|_1 = \|\boldsymbol{\theta}_{\varkappa_0}\|_1 + \|\boldsymbol{\theta}_{\varkappa_0^c}\|_1$. Denote also $\boldsymbol{u} = \boldsymbol{\theta}_{\varkappa_0} - \boldsymbol{\theta}^*$. By construction, $\boldsymbol{u}$ is supported on $\varkappa^*$. It holds

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 = \|\boldsymbol{u}\| + \|\boldsymbol{\theta}_{\varkappa_0^c}\|_1 .$$

Now (9.10) and (9.11) imply

$$\|\Psi^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + \lambda\big(\|\boldsymbol{\theta}_{\varkappa_0}\|_1 + \|\boldsymbol{\theta}_{\varkappa_0^c}\|_1\big) \le 2\mathtt{C}_0\big(\|\boldsymbol{u}\| + \|\boldsymbol{\theta}_{\varkappa_0^c}\|_1\big) + \lambda\|\boldsymbol{\theta}^*\|_1$$

and therefore, the component $\boldsymbol{\theta}_{\varkappa_0^c}$ of $\boldsymbol{\theta}$ fulfills

$$\|\Psi^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + (\lambda - 2\mathtt{C}_0)\|\boldsymbol{\theta}_{\varkappa_0^c}\|_1 \le 2\mathtt{C}_0\|\boldsymbol{u}\|_1 + \lambda\|\boldsymbol{\theta}^*\|_1 - \lambda\|\boldsymbol{\theta}_{\varkappa_0}\|_1 \le (\lambda + 2\mathtt{C}_0)\|\boldsymbol{u}\|.$$

Theorem 9.2.3. *Let* $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ *and* $\nabla = \Psi^\top \boldsymbol{\varepsilon}$ *fulfill* $\|\nabla\|_\infty \le \mathtt{C}_0$. *If* $\lambda > 2\mathtt{C}_0$, *then the estimate* $\widehat{\boldsymbol{\theta}}$ *fulfills*

$$\|\Psi^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 + (\lambda - 2\mathtt{C}_0)\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \le 2\lambda\|\boldsymbol{\theta}^*\|_1$$

In addition, if $\boldsymbol{\theta}^*$ *is supported on* $\varkappa_0$, *then the projections* $\widehat{\boldsymbol{\theta}}_{\varkappa_0}$ *and* $\widehat{\boldsymbol{\theta}}_{\varkappa_0^c}$ *of* $\widehat{\boldsymbol{\theta}}$ *to* $\varkappa^*$ *and* $\varkappa_0^c$ *can be related as*

$$\|\Psi^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 + (\lambda - 2\mathtt{C}_0)\|\widehat{\boldsymbol{\theta}}_{\varkappa_0^c}\|_1 \le (\lambda + 2\mathtt{C}_0)\|\widehat{\boldsymbol{\theta}}_{\varkappa^*} - \boldsymbol{\theta}^*\|_1 .$$

To be done: Compatibility condition

To be done: Restricted isometry and oracle bound

9.2.2 Dual problem and Danzig selector

The LASSO optimization can be viewed as minimizing the fit $\|\boldsymbol{Y} - \Psi^\top\boldsymbol{\theta}\|^2$ under the constraint on the ℓ_1-norm of $\boldsymbol{\theta}$. Then the objective (9.7) is obtained by Lagrange multiplier method. One can also consider the dual problem: minimizing ℓ_1-norm of $\boldsymbol{\theta}$ under the fit constraints. The dual problem is known as Danzig selector and reads as

$$\inf \|\boldsymbol{\theta}\|_1 \qquad \text{s.t.} \quad \|\Psi(\boldsymbol{Y} - \Psi^\top\boldsymbol{\theta})\|_\infty \le 2\lambda$$

The true value $\boldsymbol{\theta}^*$ is a natural candidate. Then $\|\Psi(\boldsymbol{Y} - \Psi^\top\boldsymbol{\theta})\|_\infty = \|\Psi\boldsymbol{\varepsilon}\|_\infty$. If λ is selected properly to ensure $\|\Psi\boldsymbol{\varepsilon}\|_\infty \le 2\lambda$, then the constraints meet, and the objective functional is equal to $\|\boldsymbol{\theta}^*\|_1$. Therefore, the solution $\widehat{\boldsymbol{\theta}}$ cannot be less sparse than $\boldsymbol{\theta}^*$ in the sense $\|\widehat{\boldsymbol{\theta}}\|_1 \le \|\boldsymbol{\theta}^*\|_1$.

9.2.3 Data-driven choice of λ

The necessary requirement to the choice of λ is that the score vector $\nabla = \Psi\boldsymbol{\varepsilon}$

$$\|\nabla\|_\infty \le 2\lambda.$$

This condition can be assessed by replacing the true noise by its bootstrap counterpart $\nabla^\flat$. With a pilot $\widetilde{\boldsymbol{\theta}}$, define $\boldsymbol{\varepsilon}^\flat = \mathcal{W}^\flat(\boldsymbol{Y} - \Psi^\top\widetilde{\boldsymbol{\theta}})$ and fix $\lambda^\flat$ by the condition

$$I\!P^\flat\big(\|\nabla^\flat\|_\infty > \lambda^\flat\big) \le \mathrm{e}^{-\mathtt{x}}.$$

The original procedure has to be applied with $\lambda = \lambda^\flat + \beta$.

References

Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Relat. Fields*, 97(1-2):113–150.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.

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.

Ibragimov, I. and Khas'minskij, R. (1981). *Statistical estimation. Asymptotic theory. Transl. from the Russian by Samuel Kotz.* New York - Heidelberg -Berlin: Springer-Verlag .

Kim, Y. (2006). The Bernstein-von Mises theorem for the proportional hazard model. *Ann. Statist.*, 34(4):1678–1700.

Kleijn, B. J. K. and van der Vaart, A. W. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic J. Statist.*, 6:354–381.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Stat.*, 21(1):255–285.

Mammen, E. (1996). Empirical process of residuals for high-dimensional linear models. *Ann. Stat.*, 24(1):307–335.

Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Stat.*, 12:1298–1309.

Portnoy, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II: Normal approximation. *Ann. Stat.*, 13:1403–1417.

Portnoy, S. (1986). Asymptotic behavior of the empiric distribution of M-estimated residuals from a regression model with many parameters. *Ann. Stat.*, 14:1152–1170.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.*, 16(1):356–366.

Shen, X. (1997). On methods of sieves and penalization. *Ann. Stat.*, 25(6):2555–2591.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Ann. Stat.*, 22(2):580–615.

Spokoiny, V. (2012). In Spokoiny (2012).

Spokoiny, V., Wang, W., and Härdle, W. (2013). Local quantile regression (with rejoinder). *J. of Statistical Planing and Inference*, 143(7):1109–1129. ArXiv:1208.5384.

Van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Stat.*, 21(1):14–44.

van de Geer, S. (2002). M-estimation using penalties or sieves. *J. Stat. Plann. Inference*, 108(1-2):55–69.

Zaitsev, A., Burnaev, E., and Spokoiny, V. (2013). Properties of the posterior distribution of a regression model based on gaussian random fields. *Automation and Remote Control*, 74(10):1645–1655.