Chapter 4
Bayesian Interpretation of Regularization

Abstract In the previous chapter, it has been shown that the regularization approach is particularly useful when information contained in the data is not sufficient to obtain a precise estimate of the unknown parameter vector and standard methods, such as least squares, yield poor solutions. The fact itself that an estimate is regarded as poor suggests the existence of some form of prior knowledge on the degree of acceptability of candidate solutions. It is this knowledge that guides the choice of the regularization penalty that is added as a corrective term to the usual sum of squared residuals. In the previous chapters, this design process has been described in a deterministic setting where only the measurement noises are random. In this chapter, we will see that an alternative formalization of prior information is obtained if a subjective/Bayesian estimation paradigm is adopted. The major difference is that the parameters, rather than being regarded as deterministic, are now treated as a random vector. This stochastic setting permits the definition of new powerful tools for both priors selection, e.g., through the maximum entropy principle, and for regularization parameters tuning, e.g., through the empirical Bayes approach and its connection with the concept of equivalent degrees of freedom.

4.1 Preliminaries

We have seen that the regularization approach can be used to effectively solve estimation problems that are otherwise ill-conditioned. In particular, a penalty is added as a corrective term to the usual sum of squared residuals. In this way, between two candidate solutions achieving the same squared loss, the regularizer is chosen such as to penalize candidate solutions that depart from our prior knowledge on some features of the unknown parameter vector.

It is worth noting that the regularization approach lies within a frequentist paradigm in which the observed data, affected by noise, are random variables, but the unknown parameter vector is deterministic in nature. For linear-in-parameter

G. Pillonetto et al., *Regularized System Identification*, Communications and Control Engineering, https://doi.org/10.1007/978-3-030-95860-2_4

models, regularization yields an estimate that, though biased, may be preferable to the unbiased least squares estimate in view of the smaller variance. In particular, the tuning of the regularization parameter aims at an advantageous solution of the bias-variance dilemma. By trading an excessive variance for some bias, a smaller mean squared error may be achieved, as exemplified by the James–Stein estimator. An alternative formalization of prior information is obtained if a subjective/Bayesian estimation paradigm is adopted. The major difference is that the parameters, rather than being regarded as deterministic, are now treated as a random vector.

In order to introduce the Bayesian paradigm, it can be useful to start with a simple example in which the parameters do depend on the result of a random experiment. Consider a metabolism model for which the parameter vector θ can take only two possible values, θ_h and θ_d, associated with healthy and diabetic patients, respectively. The model specifies $\mathrm{p}(Y|\theta)$, where Y are observations collected from a randomly chosen patient with 90% probability of being healthy and 10% probability of being diabetic. In this simple case, model identification amounts to deciding between θ_h and θ_d. It is also clear that θ is a discrete random variable with $\mathrm{p}(\theta = \theta_h) = 0.9$ and $\mathrm{p}(\theta = \theta_d) = 0.1$. These probabilities summarize the prior information about the unknown parameter, before any observation is collected. Once the data Y become available, the Bayes formula can be used to compute the posterior probability

$$\mathrm{p}(\theta_h|Y) = \frac{\mathrm{p}(Y|\theta_h)\mathrm{p}(\theta_h)}{\mathrm{p}(Y)} = \frac{\mathrm{p}(Y|\theta_h)\mathrm{p}(\theta_h)}{\mathrm{p}(Y|\theta_h)\mathrm{p}(\theta_h) + \mathrm{p}(Y|\theta_d)\mathrm{p}(\theta_d)}. \tag{4.1}$$

Of course, $\mathrm{p}(\theta_d|Y) = 1 - \mathrm{p}(\theta_h|Y)$. In particular, if the data Y are consistent with diabetes symptoms, it may well happen that $\mathrm{p}(\theta_d|Y) > 0.5$, in which case $\theta = \theta_d$ would be taken as the final estimate.

In the previous example, the prior probability distribution assigned to θ reflects a real experiment that is the random choice of a patient from a population where 90% of subjects are healthy, which implies a prejudice in favour of $\theta = \theta_h$. In other words, the prior distribution ranks the candidate parameters according to the available a priori knowledge. If we look at the numerator of (4.1), we see that it combines a priori information with the data through the product of the *prior probability* $\mathrm{p}(\theta_h)$ and the *likelihood* $\mathrm{p}(Y|\theta_h)$. In the example, the population was a binary one (either healthy or diabetic), but we can imagine more complex populations allowing for several countable or even uncountable possible values of θ.

In the actual Bayesian paradigm a further step is made: the parameters θ are assigned a prior probability $\mathrm{p}(\theta_h)$, even if there does not exist an underlying experiment that draws the model from a population of possible models. According to the subjective definition of probability, $\mathrm{p}(\theta = \bar{\theta})$ represents the (subjective) degree of belief that θ is going to take the value $\bar{\theta}$. In particular, in analogy with the regularization penalty, it is possible to rank the possible values of θ, assigning a low probability to values whose occurrence is deemed unlikely. In our context, the intrinsically subjective nature of the prior probability, a controversial issue in the confrontation between the frequentist and Bayesian paradigms, is specular to the subjective choice

of the regularization penalty: rather than expressing the preference for some solutions through the choice of a proper penalty, the preference is formulated by means a prior distribution.

As shown in the following, many formulas and results can be indifferently derived adopting either the regularization or the Bayesian paradigm. However, the Bayesian approach has its pros. In particular, the tuning of the regularization parameter, rather than being addressed on an ad hoc basis, can be formulated as a statistical estimation problem. Moreover, the Bayesian paradigm offers a very natural way to asses uncertainty intervals, whereas the regularization paradigm has a harder time assessing the amount of bias in the estimate. Among the cons, one may mention the need for a deeper probabilistic background in order to gain a full comprehension of all aspects.

Throughout the chapter we will mainly focus on the linear Gaussian case, but the approach is more general and some hints at generalizations will be provided. In addition, we will use θ to denote the stochastic vector that has generated the data, in contrast with the deterministic θ_0 used in the classical setting discussed in the previous chapter.

4.2 Incorporating Prior Knowledge via Bayesian Estimation

We consider the problem of estimating a parameter vector $\theta \in \mathbb{R}^n$, based on the observation vector $Y \in \mathbb{R}^N$. The two ingredients of Bayesian estimation are the prior distribution of θ, also known by short as *prior*, and the conditional distribution of Y given θ. As already observed, the basic assumption is that the parameter vector θ is not completely unknown, but rather some prior knowledge is available that is formulated in terms of *subjective probability*, specified as a probability density function:

$$\mathrm{p}(\theta) : \mathbb{R}^n \mapsto \mathbb{R}.$$

The density function $\mathrm{p}(\theta)$ is chosen by the user so as to assign a low probability to values whose occurrence is deemed unlikely. For instance, if θ is a scalar parameter whose value is believed to lie more or less around 30, hardly smaller than 20 and hardly larger than 40, this prior knowledge can be embedded in a Gaussian density with $\mathscr{E}\theta = \mu_\theta = 30$ and standard deviation $\sigma_\theta = 5$:

$$\theta \sim \mathscr{N}(30, 25).$$

In fact, under this distribution, $\mathrm{p}\left(|\theta - \mu_\theta| > 2\sigma_\theta\right) = \mathrm{p}\left(|\theta - 30| > 10\right) < 0.05$. Although not impossible, it is considered unlikely that values of θ too distant from 30 are going to occur. A natural question is how and when our prior knowledge is sufficient to specify a distribution. This crucial issue calls for the notion and role of *hyperparameters*, see Sect. 4.2.4, and for the possible use of the *maximum entropy*

principle as a way to obtain an entire probability distribution from partial knowledge relative to its moments, see Sect. 4.6.

The second ingredient is the conditional distribution of Y given θ that, when considered as a function of θ, is also known as *likelihood*:

$$L(\theta|Y) = \mathrm{p}(Y|\theta) = \frac{\mathrm{p}(Y,\theta)}{\mathrm{p}(\theta)},$$

where $\mathrm{p}(Y,\theta)$ is the joint probability distribution of the random vectors Y and θ. The likelihood is usually obtained from some mathematical model of the data. Consider, for instance, the simple model

$$Y_i = \theta\sqrt{i} + e_i, \qquad i = 1, \ldots, N,$$

where $e_i \sim \mathcal{N}(0, \sigma^2)$ are independent and identically distributed measurement errors, with known variance σ^2. Conditional on θ, i.e., assuming that θ is known, Y_i is Gaussian with

$$\mathscr{E}[Y_i|\theta] = \theta\sqrt{i}, \qquad \mathrm{Var}(Y_i|\theta) = \sigma^2$$

so that, in view of independence, the likelihood is

$$L(\theta|Y) = \mathrm{p}(Y|\theta) = \prod_{i=1}^{N} \mathrm{p}(Y_i|\theta), \qquad \mathrm{p}(Y_i|\theta) = \mathcal{N}(\theta\sqrt{i}, \sigma^2).$$

When both the prior distribution $\mathrm{p}(\theta)$ and the likelihood $\mathrm{p}(Y|\theta)$ have been specified, the Bayes formula yields the *posterior distribution*

$$\mathrm{p}(\theta|Y) = \frac{\mathrm{p}(Y|\theta)\mathrm{p}(\theta)}{\mathrm{p}(Y)}.$$

We have seen that all our prior knowledge was embedded in the prior. In a similar way, all the knowledge obtained by the combination of prior information with the new information brought by the observations is now embedded in the posterior distribution $\mathrm{p}(\theta|Y)$, denoted by short as *posterior*.

Although all the relevant information is encapsulated within the posterior, a *point estimate* is often required for practical or communication purposes. The *Maximum A Posteriori (MAP) estimate* is the value that maximizes the posterior:

$$\theta^{\mathrm{MAP}} = \arg\max_{\theta} \mathrm{p}(Y|\theta). \tag{4.2}$$

Its interpretation is simple, as it represents the most likely value, once the prior knowledge has been updated taking into account the observations. Alternatively, the mean squared error

$$\mathrm{MSE}(\hat{\theta}) = \mathscr{E}\left[\left(\hat{\theta} - \theta\right)^2 | Y\right]$$

can be used as a criterion to select the point estimate $\hat{\theta}$. Above, $\mathscr{E}(\cdot|Y)$ denotes the expected value taken with respect to the posterior distribution $\mathrm{p}(\theta|Y)$. The following classical result from estimation theory (whose proof is in Sect. 4.13.1) then holds.

Theorem 4.1 *The minimizer of the MSE*

$$\theta^{\mathrm{B}} = \arg\min_{\hat{\theta}} \mathrm{MSE}(\hat{\theta})$$

is known as Bayes estimate *and can be shown to be equal to the conditional mean:*

$$\theta^{\mathrm{B}} = \mathscr{E}[\theta|Y].$$

A third point estimate is the conditional median used especially in view of its statistical robustness when the posterior is obtained numerically via stochastic simulation algorithms, see Sect. 4.10.

When, in addition to a point estimate, an assessment of the uncertainty is needed, it can be derived from the posterior through the computation of a properly defined *credible region* $C_\gamma \in \mathbb{R}^n$ such that

$$\Pr(\theta \in C_\gamma | Y) = \gamma. \tag{4.3}$$

For example, C_γ could be taken as the smallest region such that (4.3) holds, a choice that goes under the name of highest posterior density region.

4.2.1 Multivariate Gaussian Variables

In this subsection, some basic properties and definitions of multivariate Gaussian variables are recalled. This review is instrumental to the derivation of the Bayesian estimator when observations and parameters are jointly Gaussian, see Sect. 4.2.2. In turn, this will pave the way to the analysis of the linear model under additive Gaussian measurement errors, see Sect. 4.2.3.

A random vector $Z = [Z_1 \dots Z_m]^T$ is said to be distributed according to a non-degenerate m-variate Gaussian distribution if its joint probability density function is of the type

$$\mathrm{p}(z_1, \dots, z_m) = \frac{1}{\sqrt{(2\pi)^m \det V}} \exp^{-\frac{1}{2}(z-\mu)^T V^{-1}(z-\mu)}, \tag{4.4}$$

where V is a symmetric positive definite matrix and μ is some vector in $\mathbb{R}^m$.

It can be shown that

$$\mathscr{E}(Z) = \mu, \quad \mathrm{Var}(Z) = V.$$

Then, the notation

$$Z \sim \mathcal{N}(\mu, V)$$

(already used before in the scalar case) indicates that Z is a multivariate Gaussian (Normal) random vector with mean μ and variance matrix V.

Property 4.1 *If $Z \sim \mathcal{N}(\mu, V)$ and $Y = AZ$, where $A \in \mathbb{R}^{n \times m}$, $n \leq m$, is a full-rank deterministic matrix, then*

$$Y \sim \mathcal{N}(A\mu, AVA^T).$$

In particular, it follows that the marginal distributions of the entries of Z are Gaussian:

$$Z_i \sim \mathcal{N}(\mu_i, V_{ii}).$$

Property 4.2 *Assuming $Z \sim \mathcal{N}(\mu, V)$, let $X = [Z_1 \dots Z_n]^T$, $Y = [Z_{n+1} \dots Z_m]^T$, where $1 \leq n < m$, and partition μ and V accordingly:*

$$\mu = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \quad \begin{bmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{bmatrix}.$$

Then, $\mathrm{p}(X|Y = y)$ *is a multivariate Gaussian density function with*

$$\begin{aligned} \mathscr{E}(X|Y = y) &= \mu_X + V_{XY} V_{YY}^{-1}(y - \mu_Y) \\ \mathrm{Var}(X|Y = y) &= V_{XX} - V_{XY} V_{YY}^{-1} V_{YX} \end{aligned}$$

and we can write

$$(X|Y = y) \sim \mathcal{N}\left(\mu_X + V_{XY} V_{YY}^{-1}(y - \mu_Y),\, V_{XX} - V_{XY} V_{YY}^{-1} V_{YX}\right),$$

where $X|Y = y$ stands for the random vector X conditional on $Y = y$.

4.2.2 The Gaussian Case

Let us consider the case in which the observation vector $Y \in \mathbb{R}^N$ and the unknown vector $\theta \in \mathbb{R}^n$ are jointly Gaussian:

$$\begin{bmatrix} \theta \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_\theta \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_\theta & \Sigma_{\theta Y} \\ \Sigma_{Y\theta} & \Sigma_Y \end{bmatrix}\right), \quad \Sigma_Y > 0. \tag{4.5}$$

The key idea behind Bayesian estimation is referring to the posterior distribution of θ given Y as representative of the state of knowledge about the unknown vector. It follows from Property 4.2 that such posterior is Gaussian as well:

$$\theta|Y \sim \mathcal{N}\left(\mu_\theta + \Sigma_{\theta Y}\Sigma_Y^{-1}(Y-\mu_Y),\, \Sigma_\theta - \Sigma_{\theta Y}\Sigma_Y^{-1}\Sigma_{Y\theta}\right). \tag{4.6}$$

In view of Gaussianity, θ^{MAP} coincides with the conditional expectation $\mathscr{E}(\theta|Y)$:

$$\theta^{\mathrm{B}} = \theta^{\mathrm{MAP}} = \mathscr{E}(\theta|Y) = \mu_\theta + \Sigma_{\theta Y}\Sigma_Y^{-1}(Y-\mu_Y). \tag{4.7}$$

The reliability of the estimate can be assessed by the posterior variance

$$\Sigma_{\theta|Y} = \mathrm{Var}(\theta|Y) = \Sigma_\theta - \Sigma_{\theta Y}\Sigma_Y^{-1}\Sigma_{Y\theta}$$

based on which the so-called credible intervals can be derived as explained below.

The posterior variance of θ_i is the i-th diagonal entry of the posterior covariance matrix:

$$\sigma^2_{\theta_i|Y} = \left[\Sigma_{\theta|Y}\right]_{ii}.$$

Observing that $\theta_i|Y \sim \mathcal{N}(\theta_i^{\mathrm{B}}, \sigma^2_{\theta_i|Y})$, it follows that

$$\Pr\left(\theta_i^{\mathrm{B}} - 1.96\sigma_{\theta_i|Y} \le \theta_i \le \theta_i^{\mathrm{B}} - 1.96\sigma_{\theta_i|Y}|Y\right) = 0.95 \tag{4.8}$$

so that $[\theta_i^{\mathrm{B}} - 1.96\sigma_{\theta_i|Y},\, \theta_i^{\mathrm{B}} + 1.96\sigma_{\theta_i|Y}]$ is the 95%-credible interval for the parameter θ_i, given the observation vector Y. If two or more parameters are jointly considered, the notion of credible region can be obtained in a similar way. In the Gaussian case, such regions are suitable (hyper)-ellipsoids centred in θ^{B}.

4.2.3 The Linear Gaussian Model

The Bayesian approach can be applied to the estimation of the standard linear model in matrix form

$$Y = \Phi\theta + E, \quad E \sim \mathcal{N}(0, \Sigma_E), \quad \Sigma_E > 0 \tag{4.9}$$

in which $Y \in \mathbb{R}^N$ and the parameter vector θ is no more regarded as a deterministic quantity, but as a random vector independent of E. In particular, we assume that some prior information is available which is embedded in a Gaussian prior distribution

$$\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta), \quad \Sigma_\theta > 0.$$

Since Y is the linear combination of the jointly Gaussian vectors θ and E, the vectors Y and θ are jointly Gaussian as well. Hereafter, positive definiteness of Σ_θ is assumed if not stated otherwise. The singular case, see Remark 4.1, amounts to assuming perfect knowledge of some linear combination of the unknown parameters or, equivalently, to constrain the estimated vector θ to belong to a prescribed subspace. The ability to incorporate this type of constraint is not unique to the Bayesian

approach. In the context of the deterministic regularization, an example is given by the optimal regularization matrix $P = \theta_0 \theta_0^T$, derived in Sect. 3.4.2.1.

In order to obtain the Bayes estimate according to (4.7), we need to compute $\mu_Y = \mathscr{E}(Y)$, $\Sigma_{\theta Y} = \mathrm{Cov}(\theta, Y)$, and $\Sigma_Y = \mathrm{Var}(Y)$:

$$\begin{aligned} \mu_Y &= \mathscr{E}(Y) = \Phi \mu_\theta \\ \mathrm{Var}(Y) &= \mathrm{Var}(\Phi\theta) + \mathrm{Var}(E) = \Phi \Sigma_\theta \Phi^T + \Sigma_E \\ \mathrm{Cov}(\theta, Y) &= \mathrm{Cov}(\theta, \Phi\theta) + \mathrm{Cov}(\theta, E) = \Sigma_\theta \Phi^T. \end{aligned}$$

Then, we can apply (4.7) to obtain

$$\theta^{\mathrm{B}} = \mu_\theta + \Sigma_\theta \Phi^T (\Phi \Sigma_\theta \Phi^T + \Sigma_E)^{-1} (Y - \Phi \mu_\theta) \tag{4.10}$$

$$\mathrm{Var}(\theta|Y) = \Sigma_\theta - \Sigma_\theta \Phi^T (\Phi \Sigma_\theta \Phi^T + \Sigma_E)^{-1} \Phi \Sigma_\theta. \tag{4.11}$$

The proofs of the following two classical results are reported in Sects. 4.13.2 and 4.13.3.

Theorem 4.2 (Orthogonality property)

$$\mathscr{E}\left[(\theta^{\mathrm{B}} - \theta) Y^T\right] = 0. \tag{4.12}$$

The following lemma, whose proof is in Sect. 4.13.3, is useful in order to obtain an alternative expression that proves more convenient, especially when $n \ll N$.

Lemma 4.1 *It holds that*

$$\Sigma_\theta \Phi^T (\Phi \Sigma_\theta \Phi^T + \Sigma_E)^{-1} = (\Phi^T \Sigma_E^{-1} \Phi + \Sigma_\theta^{-1})^{-1} \Phi^T \Sigma_E^{-1}.$$

By applying the previous lemma, the alternative expression of the Bayes estimate is obtained

$$\theta^{\mathrm{B}} = (\Phi^T \Sigma_E^{-1} \Phi + \Sigma_\theta^{-1})^{-1} (\Phi^T \Sigma_E^{-1} Y + \Sigma_\theta^{-1} \mu_\theta) \tag{4.13}$$

$$\mathrm{Var}(\theta|Y) = (\Phi^T \Sigma_E^{-1} \Phi + \Sigma_\theta^{-1})^{-1}. \tag{4.14}$$

As already noted, the Bayes estimate coincides with θ^{MAP}, the maximum of the posterior density:

$$\mathrm{p}(\theta|Y) \propto \mathrm{p}(Y|\theta)\mathrm{p}(\theta).$$

Recall that, in view of the assumed linear model (4.9),

$$Y|\theta \sim \mathscr{N}(\Phi\theta, \Sigma_E)$$

and note that

$$\log \mathrm{p}(\theta) = c_1 - \frac{1}{2}(\theta - \mu_\theta)^T \Sigma_\theta^{-1}(\theta - \mu_\theta) \tag{4.15}$$

$$\log \mathrm{p}(Y|\theta) = c_2 - \frac{1}{2}(Y - \Phi\theta)^T \Sigma_E^{-1}(Y - \Phi\theta), \tag{4.16}$$

where c_1 and c_2 are constants we are not concerned with. Therefore, the maximization of the posterior density can be written as

$$\begin{aligned}\theta^{\mathrm{MAP}} &= \arg\max_\theta \log \mathrm{p}(Y|\theta) + \log \mathrm{p}(\theta) \\ &= \arg\max_\theta (Y - \Phi\theta)^T \Sigma_E^{-1}(Y - \Phi\theta) + (\theta - \mu_\theta)^T \Sigma_\theta^{-1}(\theta - \mu_\theta)\end{aligned}$$

whose solution is easily shown to be given by (4.13). This shows that, under Gaussianity assumptions, the Bayes estimate of the linear model can be seen as a regularized least squares estimator with quadratic regularization term (ReLS-Q), see Sect. 3.4. In particular, if

$$\Sigma_E = \sigma^2 I_N, \quad \mu_\theta = 0, \tag{4.17}$$

the Bayes and MAP estimators,

$$\theta^{\mathrm{B}} = \theta^{\mathrm{MAP}} = \arg\min_\theta \|Y - \Phi\theta\|^2 + \theta^T P^{-1}\theta, \tag{4.18}$$

coincide with the ReLS estimator with regularization matrix $P = \Sigma_\theta/\sigma^2$. Under the further assumption $\Sigma_\theta = \lambda I_n$, the MAP estimator coincides with a ridge regression estimator with $\gamma = \sigma^2/\lambda$.

Remark 4.1 When $\Sigma_\theta = P$, where $P = P^T \geq 0$ is singular, one can still use (4.10) to obtain the Bayes estimate, while (4.13) and the quadratic problem (4.18) are no more valid due to the nonexistence of Σ_θ^{-1}. Nevertheless, by replicating the derivation in Remark 3.1, it is still possible to interpret the Bayes estimate as the solution of a constrained quadratic problem. In particular, under (4.17), we have that

$$\theta^{\mathrm{B}} = \arg\min_\theta \|Y - \Phi\theta\|_2^2 + \theta^T P^+ \theta \tag{4.19}$$

$$\text{subj. to} \quad U_2^T\theta = 0, \tag{4.20}$$

where U_2 was defined in Remark 3.1, as part of the singular value decomposition of P. The result can be interpreted as follows. A singular variance matrix means that we have perfect knowledge on some linear combination of the parameter vector. In particular,

$$\begin{aligned}\mathrm{Var}\left[U_2^T\theta\right] &= U_2^T \mathrm{Var}\,(\theta)\, U_2 \\ &= U_2^T \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Lambda_P & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_1 & U_2 \end{bmatrix}^T U_2 = 0,\end{aligned}$$

where, with reference to the SVD of P, we have exploited the fact that $U_2^T U_1 = 0$. As a consequence,

$$\Pr(U_2^T \theta = U_2 \mu_\theta) = 1,$$

thus justifying the presence of the equality constraints in the quadratic problem (4.19)–(4.20), where $\mu_\theta = 0$ is assumed. Recalling the orthogonality of U_1 and U_2, we have that $U_2^T \theta = 0$ implies that $\theta \in \text{Range}(U_1) = \text{Range}(P)$. Therefore, the constrained quadratic problem (4.19)–(4.20) can also be equivalently reformulated as

$$\theta^B = \underset{\theta \in Range(P)}{\arg\min} \quad \|Y - \Phi\theta\|^2 + \theta^T P^+ \theta. \tag{4.21}$$

One can also assess that the solution of this problem can be written as

$$\theta^{\mathrm{B}} = P\Phi^T(\Phi P \Phi^T + \Sigma_E)^+ Y,$$

an expression which does not require invertibility of any matrix.

In conclusion, the Bayes estimate always exists and is unique. In any case, it can be written as (4.7) with Σ_Y^{-1} replaced by its pseudoinverse.

The Bayesian interpretation of deterministic regularization can be exploited to obtain a guideline for the selection of the regularization matrix. The simplest case is when some statistics, e.g., based on samples coming from past problems, is available for the parameter vector θ. Then, the Bayesian interpretation suggests to select the covariance matrix of θ, divided by the error variance σ^2, as regularization matrix. If examples from the past are not available, one may rely on prior knowledge, telling that some entries of θ have smaller variance than others or that some correlation exists between the entries.

4.2.4 Hierarchical Bayes: Hyperparameters

In the cases in which prior information on the parameters is not sufficient to specify a prior, it is common to resort to hierarchical Bayesian models. Instead of fixing the prior, a family of priors is considered, parametrized by one or more *hyperparameters*. As an example, consider the case in which prior knowledge could be formalized in terms of zero-mean independent and equally distributed parameters whose absolute value is not too large. In absence of more precise information on their size, we could adopt the following prior:

$$\theta \sim \mathcal{N}(0, \lambda I_N),$$

where the scalar λ, called hyperparameter, enters the game as a further unknown quantity. More in general, the prior distribution $\mathrm{p}(\theta|\alpha)$ may depend on a hyperparameter vector α. One may also want to consider a hyperparameter vector β entering the definition of the likelihood $\mathrm{p}(Y|\theta, \beta)$. The most common example is when the

measurement variance σ^2 is not known and is therefore treated as a hyperparameter. In the following, the vector of all hyperparameters will be denoted by

$$\eta = \left[\alpha^T \beta^T\right]^T .$$

For a given η, we will denote by $\theta^{\mathrm{MAP}}(\eta)$ and $\theta^{\mathrm{B}}(\eta)$ the corresponding MAP and Bayes estimates:

$$\theta^{\mathrm{MAP}}(\eta) = \arg\max_{\theta} \mathrm{p}(\theta|Y, \eta) \tag{4.22}$$

$$\theta^{\mathrm{B}}(\eta) = \mathscr{E}(\theta|Y, \eta) = \int \theta \mathrm{p}(\theta|Y, \eta) d\theta, \tag{4.23}$$

where

$$\mathrm{p}(\theta|Y, \eta) = \frac{\mathrm{p}(Y|\theta, \beta)\mathrm{p}(\theta|\alpha)}{\int \mathrm{p}(Y|\theta, \beta)\mathrm{p}(\theta|\alpha) d\theta}. \tag{4.24}$$

4.3 Bayesian Interpretation of the James–Stein Estimator

In this section, we show that the James–Stein estimator can be seen as a particular Bayesian estimator. As seen, in Eq. (1.2), the measurements model is

$$Y = \theta + E, \quad E \sim \mathscr{N}(0, \sigma^2 I_N). \tag{4.25}$$

In a Bayesian setting, the parameter vector is regarded as a random vector, whose distribution reflects our state of knowledge. In particular, we assume

$$\theta \sim \mathscr{N}(0, \lambda I_N), \tag{4.26}$$

where λ plays the role of hyperparameter. It follows that θ and Y are zero-mean jointly Gaussian variables with

$$\Sigma_{\theta Y} = \mathscr{E}(\theta Y^T) = \mathscr{E}(\theta\theta^T) = \lambda I_N, \quad \Sigma_Y = \mathscr{E}(YY^T) = (\lambda + \sigma^2) I_N. \tag{4.27}$$

According to (4.7), the Bayes estimate is given by the conditional expectation

$$\mathscr{E}(\theta|Y) = \Sigma_{\theta Y} \Sigma_Y^{-1} Y = \frac{\lambda}{\lambda + \sigma^2} Y = \left(1 - r_{\mathrm{Bayes}}\right) Y, \tag{4.28}$$

where

$$r_{\mathrm{Bayes}} = \frac{\sigma^2}{\lambda + \sigma^2}. \tag{4.29}$$

It is apparent that the estimator (4.28) has the same structure as James–Stein's one, with r replaced by r_{Bayes}.

Since Y and θ are jointly Gaussian, $\mathscr{E}(\theta|Y) = \theta^{\text{MAP}}$, where

$$\theta^{\text{MAP}} = \arg\min_{\theta} \frac{\|Y-\theta\|^2}{\sigma^2} + \frac{\|\theta\|^2}{\lambda} = \arg\min_{\theta} \|Y-\theta\|^2 + \frac{\sigma^2}{\lambda}\|\theta\|^2$$

which highlights the fact that $\mathscr{E}(\theta|Y)$ is the solution of a regularized least squares problem, controlled by the regularization parameter σ^2/λ.

If the variances λ and σ^2 could be assigned on the basis of prior knowledge, the similarity would be only formal. Let us make a step forward, considering the case in which the variance σ^2 is given, while λ is estimated from the data. The basic idea is that the hyperparameter λ could be tuned based on the observed vector Y and plugged into (4.29) to obtain an estimate of r_{Bayes}. Alternatively, one may focus directly on finding a sensible estimate of r_{Bayes}. In this respect, we are going to show that Stein's r is an unbiased estimate of r_{Bayes} under the Gaussian model (4.25) and (4.26) [6]. For this purpose, we will exploit a property of the inverse chi-square variable.

Definition 4.1 *(chi-square random variable)* The sum of the squares of n standard Gaussian independent random variables is a nonnegative valued random variable known as *chi-square variable* with n degrees of freedom:

$$\chi_n^2 = \sum_{i=1}^{n} X_i^2, \quad X_i \sim \mathscr{N}(0,1).$$

Its mean and expectation are

$$\mathscr{E}\left(\chi_n^2\right) = n, \quad \text{Var}\left(\chi_n^2\right) = 2n.$$

The inverse of a chi-square variable is called *inverse chi-square*. For $n > 2$, its mean is

$$\mathscr{E}\left[\frac{1}{\chi_n^2}\right] = \frac{1}{n-2}. \tag{4.30}$$

Now, assume $N > 2$ and observe that

$$\frac{\|Y\|^2}{\lambda+\sigma^2} = \frac{\sum_i^n Y_i^2}{\lambda+\sigma^2} \sim \chi_N^2.$$

Recalling that the expectation of the inverse chi-square is equal to $1/(N-2)$, we have that

$$\mathscr{E}\left[\frac{\lambda+\sigma^2}{\|Y\|^2}\right] = \mathscr{E}\left[\frac{1}{\chi_N^2}\right] = \frac{1}{(N-2)}.$$

Therefore,

$$\mathscr{E}(r) = \mathscr{E}\left[\frac{(N-2)\sigma^2}{\|Y\|^2}\right] = \frac{\sigma^2}{\lambda+\sigma^2} = r_{Bayes}.$$

This means that James–Stein's shrinking coefficient r can be seen as an unbiased estimator of the shrinking coefficient r_{Bayes} appearing in the formula of the posterior expectation.

The example is instructive under several respects. First, it shows that, under suitable probabilistic assumptions, the typical structure of regularized estimators can be justified through Bayesian arguments. The second point has to do with the tuning of the regularization parameters. In the empirical Bayes approach, see Sect. 4.4, there is a preliminary step in which a point estimate of hyperparameters is obtained by standard estimation methods. Then, this point estimate is plugged into the expression of the Bayesian estimator. Although a full Bayesian approach would call for the joint estimation of parameters and hyperparameters, the two-step empirical Bayes approach not only conjugates simplicity and effectiveness but provides a probabilistic underpinning to regularized identification methods.

4.4 Full and Empirical Bayes Approaches

When the prior, and possibly the likelihood, include hyperparameters, Bayesian estimation becomes more complex and gives rise to alternative approaches. In principle, we want to obtain the posterior distribution

$$\mathrm{p}(\theta|Y) = \frac{\mathrm{p}(Y|\theta)\mathrm{p}(\theta)}{\mathrm{p}(Y)}.$$

However, if a hierarchical Bayesian model is adopted, we do not know $\mathrm{p}(\theta)$, but only $\mathrm{p}(\theta|\eta)$. At the cost of assigning a prior $\mathrm{p}(\eta)$ also to the hyperparameters, the prior $\mathrm{p}(\theta)$ can be obtained by marginalization of the joint probability density:

$$\mathrm{p}(\theta) = \int \mathrm{p}(\theta, \eta)d\eta = \int \mathrm{p}(\theta|\eta)\mathrm{p}(\eta)d\eta.$$

In general, this integral has to be computed numerically, e.g., by Monte Carlo methods. This leads to *full Bayesian* methods that compute the desired $\mathrm{p}(\theta|Y)$ regarding both parameters and hyperparameters as random variables. Some remarks on these methods will be given in Sect. 4.10.

The justification for a simpler computational scheme stems from the following reformulation of the posterior:

$$\mathrm{p}(\theta|Y) = \int \mathrm{p}(\theta, \eta|Y)d\eta = \int \mathrm{p}(\theta|\eta, Y)\mathrm{p}(\eta|Y)d\eta. \tag{4.31}$$

Observe that

$$\mathrm{p}(\eta|Y) \propto \mathrm{p}(Y|\eta)\mathrm{p}(\eta), \tag{4.32}$$

where $L(\eta|Y) = \mathrm{p}(Y|\eta)$ is the likelihood of the hyperparameter vector η. It is also called *marginal likelihood* because it is obtained from the marginalization with respect to θ of the joint density $\mathrm{p}(Y, \theta|\eta)$:

$$L(\eta|Y) = \int \mathrm{p}(Y, \theta|\eta)d\theta = \int \mathrm{p}(Y|\theta, \eta)\mathrm{p}(\theta|\eta)d\theta. \tag{4.33}$$

If data are sufficiently informative, the marginal likelihood has good chances to be unimodal and sharply peaked in a neighbourhood of the maximum likelihood estimate

$$\eta^{\mathrm{ML}} = \arg\max_{\eta} \mathrm{p}(Y|\eta).$$

When this happens and $\mathrm{p}(\eta)$ is rather uninformative (as it should be), from (4.32) it follows that $\mathrm{p}(\eta|Y)$ is peaked as well. Then, as long as the properties of $\mathrm{p}(\theta|\eta, Y)$ do not change rapidly with η near η^{ML}, the integral (4.31) can be approximated as

$$\mathrm{p}(\theta|Y) \simeq \mathrm{p}(\theta|\eta^{\mathrm{ML}}, Y) = \frac{\mathrm{p}(Y|\theta, \eta^{\mathrm{ML}})\mathrm{p}(\theta|\eta^{\mathrm{ML}})}{\mathrm{p}(Y|\eta^{\mathrm{ML}})}.$$

In practice, this suggests to compute the posterior using the prior $\mathrm{p}^*(\theta) = \mathrm{p}(\theta|\eta^{\mathrm{ML}})$ associated with the maximum likelihood estimate of hyperparameters. More in general, Empirical Bayes (EB) methods adopt a two-stage scheme. In the first step, a point estimate η^* is computed which is then kept fixed in the second step, when the posterior of the parameters is obtained, based on the prior $\mathrm{p}^*(\theta) = \mathrm{p}(\theta|\eta^*)$.

Among the advantages of the approach one may mention its simplicity, especially when there are few hyperparameters and the posterior $\mathrm{p}(\theta|Y, \eta^{\mathrm{ML}})$ is easily obtained as in the jointly Gaussian case. Moreover, the tuning of η admits an intuitive interpretation as the counterpart of model order selection in classic parametric estimation methods. The main drawback is that the EB method fails to propagate the uncertainty of the point estimate η^*.

Under the linear Gaussian model (4.9), the integral (4.33) admits a closed-form solution. In fact, since

$$Y \sim \mathcal{N}(\Phi\mu_\theta(\eta), \Sigma(\eta)), \quad \Sigma(\eta) = \Phi\Sigma_\theta(\eta)\Phi^T + \Sigma_E(\eta),$$

we have

$$\log L(\eta|Y) = -\frac{1}{2}\log(2\pi\det(\Sigma)) - \frac{1}{2}(Y - \Phi\mu_\theta)^T\Sigma^{-1}(Y - \Phi\mu_\theta), \tag{4.34}$$

where in the right-hand side dependence on η has been omitted for simplicity.

Therefore, application of Empirical Bayes estimation to the linear model (4.9) would consist of the following two steps:

Step 1:

$$\eta^* = \eta^{\mathrm{ML}} = \arg\max_{\eta} L(\eta|Y).$$

Step 2: Let $\mu_\theta = \mu_\theta(\eta^*)$, $\Sigma_E = \Sigma(\eta^*)$, $\Sigma_\theta = \Sigma_\theta(\eta^*)$ and compute the posterior expectation according to Sect. 4.2.3.

When the likelihood and the prior are such that integral (4.33) cannot be computed explicitly, an approximation is needed. In particular, one can resort to the Laplace approximation, which is based on a second-order Taylor expansion of $\log\ \mathrm{p}(Y, \theta|\eta)$ around $\theta^{\mathrm{MAP}}(\eta)$ defined in (4.22), from which an integrable approximation of $\mathrm{p}(Y, \theta|\eta)$ appearing in (4.33) is obtained. Note, however, that the Laplace approximation has to be recalculated for each evaluation of $L(\eta|Y)$ occurring during the iterative computation of η^{ML}.

4.5 Improper Priors and the Bias Space

The use of priors is most useful whenever the data alone are not sufficient to provide reliable parameter estimates but there exists some a priori knowledge that can be exploited. It may happen that for some parameters the introduction of a prior is not possible or not desirable, because their estimation can be satisfactorily performed anyway, given the information in the data. This can be accounted for by assuming that such parameters have *improper priors*.

In order to deal with the case where p parameters $\theta^P \in \mathbb{R}^p$ have a proper prior and the remaining $n - p$ parameters $\theta^I \in \mathbb{R}^{n-p}$ have an improper prior, consider the following model:

$$Y = \Phi\theta + E, \quad \Phi = \begin{bmatrix} \Omega\ \Psi \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta^P \\ \theta^I \end{bmatrix} \tag{4.35}$$

$$\theta \sim \mathcal{N}(0, \Sigma_\theta), \quad E \sim \mathcal{N}(0, \sigma^2 I_N) \tag{4.36}$$

$$\Sigma_\theta = \begin{bmatrix} \Sigma & 0 \\ 0 & a I_{n-p} \end{bmatrix}, \quad \Sigma > 0. \tag{4.37}$$

The (asymptotically) improper prior for θ^I is obtained by letting $a \to \infty$ so that θ^I has infinite variance, i.e., its density is flat. This amounts to complete lack of prior knowledge for the last $n - p$ entries of the parameter vector θ that, for simplicity, is assumed to be zero mean. The use of improper priors in a Bayesian setting has the same effect as the introduction of a *bias space* in a deterministic regularization setting. Within such a subspace, parameters are immune from regularization, a feature that could be useful to apply regularization only where needed without causing undesired distortions. The following theorem, whose proof is in Sect. 4.13.4, is analogous to a result obtained in [22] to obtain a Bayesian interpretation of smoothing splines. It

illustrates the asymptotic behaviour of posterior means and variances as a goes to infinity.

Theorem 4.3 (adapted from [22]) *If* $\text{rank}(\Phi) = n$ *and* $\text{rank}(\Omega) = n - p$, *then*

$$\begin{aligned}
\lim_{a\to\infty} \mathscr{E}(\theta^I|Y) &= (\Psi^T M^{-1}\Psi)^{-1}\Psi^T M^{-1} Y \\
\lim_{a\to\infty} \mathscr{E}(\theta^P|Y) &= \Sigma\Omega^T M^{-1}(I_n - \Psi(\Psi^T M^{-1}\Psi)^{-1}\Psi^T M^{-1})Y \\
M &= \Omega\Sigma\Omega^T + \sigma^2 I_N \\
\lim_{a\to\infty} \text{Var}\,(\theta|Y) &= \sigma^2\left(\Phi^T\Phi + \sigma^2\begin{bmatrix} 0 & 0 \\ 0 & \Sigma^{-1}\end{bmatrix}\right)^{-1}.
\end{aligned}$$

An interesting benefit of improper priors is the possibility of reducing the number of hyperparameters by treating some of them as unknowns whose prior is improper. Letting the symbol $\mathbf{1}_{n\times 1}$ denotes a column vector of ones, assume, for example, that $\theta \sim \mathcal{N}(\mu\mathbf{1}_{n\times 1}, \Sigma_\theta)$, i.e., all the scalar entries of θ share the same prior mean μ. In most cases, very little is known about μ that could be therefore regarded as a hyperparameter to be tuned by marginal likelihood maximization. It can be then treated as a deterministically known variable, according to the Empirical Bayes approach, see Sect. 4.4. By this choice, however, the hyperparameter is fixed to its point estimate and its uncertainty is not propagated, implying that the uncertainty of θ^{B} will be underestimated if assessed by (4.14).

Alternatively, μ can be treated as a further random parameter. For this purpose, define $\tilde{\theta} = \theta - \mu$ and consider the model

$$\begin{aligned}
\bar{\theta} &= \begin{bmatrix}\tilde{\theta} \\ \mu\end{bmatrix}, \qquad \Sigma_{\bar{\theta}} = \begin{bmatrix}\Sigma_\theta & 0 \\ 0 & a\end{bmatrix} \\
Y &= \bar{\Phi}\bar{\theta} + E, \quad \bar{\Phi} = \begin{bmatrix}\Phi & \Phi\mathbf{1}_{n\times 1}\end{bmatrix} \\
\bar{\theta} &\sim \mathcal{N}(0, \Sigma_{\bar{\theta}}), \quad E \sim \mathcal{N}(0, \sigma^2 I_N).
\end{aligned}$$

This formulation decreases the number of hyperparameters, without introducing prejudices (provided we let $a \to \infty$). More importantly, it is now possible to assess the joint uncertainty of the estimates of μ and $\tilde{\theta}$ through the posterior variance $\text{Var}(\bar{\theta}|Y)$.

4.6 Maximum Entropy Priors

A major appeal of the Bayesian paradigm lies in its ability to provide a rational foundation to regularization: one starts from prior knowledge and then proceeds with its formalization in terms of a probabilistic prior, from which the regularization

penalty is finally derived. However, there is a stumbling block in the way, because the available prior knowledge is often too vague to avoid arbitrariness in the choice of the prior distribution. As a matter of fact, the derivation of systematic approaches for the selection of prior distributions is a classic topic of Bayesian estimation theory. In this section, the approach based on entropy maximization is briefly reviewed.

The starting point is the observation that, even when prior information is absent or very limited, there are candidate distributions that are obviously preferable, due to symmetry arguments. Assume, for instance, that candidate values for a scalar parameter θ are known to belong to a finite set $\{\theta_i, i = 1, \ldots m\}$ and no further information is available. Then, the only reasonable prior distribution will be $\mathrm{p}(\theta = \theta_i) = 1/m$. In fact, assigning unequal probabilities would create an unjustified asymmetry, given that our prior information does not make any distinction between the m possible values of the parameter.

The case of a continuous-valued parameter θ taking values in a finite interval $[a, b]$ can be addressed in a similar way. In this case, a reasonable prior distribution is the uniform one:

$$\mathrm{p}(\theta) = \begin{cases} \frac{1}{b-a}, & a \leq \theta \leq b \\ 0, & \text{elsewhere} \end{cases}.$$

In both examples, we might say the chosen distributions are those that reflect the maximum ignorance about the unknown parameter.

The next step is to formalize this notion of maximum ignorance in contexts where some partial information about θ is available. This can be done by means of the notion of entropy of a probability distribution. For a discrete distribution $\mathrm{p}(\cdot)$ taking values $\mathrm{p}(\theta_i)$ on a numerable set $\{\theta_i\}$, the entropy H is defined as

$$H(\mathrm{p}) = -\sum_i \mathrm{p}(\theta_i) \log \mathrm{p}(\theta_i).$$

Note that the minimum possible entropy $H(\mathrm{p}) = 0$ occurs when the probability is concentrated at a unique value $\bar{\theta}$. This is the case of a maximally informative distribution such that $\mathrm{p}(\theta = \bar{\theta}) = 1$. Conversely, if the set $\{\theta_i\}$ has cardinality m, the maximum value $H(\mathrm{p}) = \log(m)$ is achieved in correspondence of the uniform distribution $\mathrm{p}(\theta = \theta_i) = 1/m, \forall i$. In other words, the larger the entropy, the less information is conveyed by the distribution.

For continuous-valued random variables, the notion of *differential entropy* $h(\mathrm{p})$ is introduced:

$$h(\mathrm{p}) = -\int_{D_\theta} \mathrm{p}(\theta) \log \mathrm{p}(\theta) d\theta,$$

where D_θ denotes the support of the distribution. Note that, among distributions with finite support, the maximum possible (differential) entropy is achieved by the uniform distribution.

The principle of *Maximum Entropy* (MaxEnt) states that the admissible distribution with largest entropy is the one that best represents the current state of knowledge.

The admissible distributions are those that satisfy a set of constraints, chosen so as to incorporate all the available prior knowledge. For instance, if the prior knowledge amounts to knowing that $\theta \in [a, b]$, the prior suggested by the MaxEnt principle is the uniform distribution. Other types of constraints are typically expressed as expectations of functions of the parameters θ. In particular, consider a random variable θ, subject to known values η_i of m expectations

$$\mathscr{E}[g_i(\theta)] = \int g_i(\theta)\mathrm{p}(\theta)d\theta = \eta_i, \quad i = \ldots, m. \tag{4.38}$$

Then, we have the following useful result.

Theorem 4.4 (General form of maximum entropy distributions, based on [12]) *Among all the distributions satisfying (4.38), the maximum entropy one is of exponential type*

$$\mathrm{p}(\theta) = A \exp(-\lambda_1 g_1(\theta) - \ldots - \lambda_m g_m(\theta)), \tag{4.39}$$

where λ_i are m constants determined from (4.38) and A is such that

$$A \int_{-\infty}^{+\infty} \exp(-\lambda_1 g_1(\theta) - \ldots - \lambda_m g_m(\theta))d\theta = 1. \tag{4.40}$$

Example 4.5 *(MaxEnt prior from information on expected absolute value)* Assume that prior knowledge is summarized by the expectation $\mathscr{E}|\theta| = \eta$. Then, the MaxEnt prior is the solution of the constrained optimization problem

$$\max_{\mathrm{p}} h(\mathrm{p}) \quad \text{s.t.} \quad \mathscr{E}|\theta| = \eta.$$

Obviously, $m = 1$ and $g_1(\theta) = |\theta|$. In view of (4.39) and (4.40), $\mathrm{p}(\theta)$ is a Laplace distribution:

$$\mathrm{p}(\theta) = 0.5\lambda e^{-\lambda|\theta|}.$$

The value of λ is found by imposing the constraint on the expectation:

$$\int_{-\infty}^{+\infty} 0.5|\theta|\lambda e^{-\lambda|\theta|}d\theta = \eta.$$

Since the constraint on the expectation is satisfied for $\lambda = 1/\eta$, the following Laplace distribution is eventually obtained:

$$\mathrm{p}(\theta) = \frac{e^{-\frac{|\theta|}{\eta}}}{2\eta}.$$

Therefore, starting from a very partial information, such as a guess on the expected absolute value of the parameter, it is possible to completely specify a prior distribu-

tion that: (i) is coherent with the prior knowledge and (ii) does not introduces undue assumptions because it is the least informative one, so far as entropy is taken as a measure of informativeness. One could object that it is scarcely realistic to assume prior knowledge of the expected absolute value of θ. However, if we adopt the empirical Bayes framework, the objection is circumvented by the possibility of treating η as a hyperparameter that will be estimated from data.

Therefore, prior knowledge may just tell that the expectation of $|\theta|$ is finite, without specifying a value for this expectation. The MaxEnt principle then suggests the functional form of the prior that incorporates a hyperparameter η, whose tuning, e.g., by marginal likelihood maximization, see Sect. 4.4, will be the first step of the actual estimation algorithm. As it will be seen in the following, this particular prior is associated with the Bayesian interpretation of the regularization penalty employed by the so-called Lasso estimator that has been already introduced in a deterministic regularization setting in Sect. 3.6.1.1. □

For our purposes, of particular interest are MaxEnt priors satisfying constraints on the second-order moments. In the scalar case, we have the following classical result, e.g., see [19].

Proposition 4.1 (based on [12]) *Let θ be a zero-mean random variable with known variance $\mathscr{E}\theta^2 = \lambda$. Then, the MaxEnt distribution is normal:*

$$\theta \sim \mathscr{N}(0, \lambda).$$

Also in this case, the necessity of specifying λ is not an issue, because the unknown variance can be regarded as a hyperparameter and tuned by marginal likelihood maximization. In other words, if the only prior knowledge is that θ has a finite, yet unknown, variance, the MaxEnt principle suggests the use of a normal prior parametrized by its variance.

When θ is a vector, a multivariate prior might be derived according to the following proposition.

Proposition 4.2 (based on [12]) *Let θ be a zero-mean n-dimensional random vector whose entries have known variances $\mathscr{E}\theta_i^2 = \lambda_i$, $i = 1, \ldots, n$. Then, the MaxEnt distribution is a multivariate normal with diagonal covariance matrix:*

$$\theta \sim \mathscr{N}(0, \Sigma_\theta), \quad \Sigma_\theta = \mathrm{diag}\{\lambda_i\}.$$

The importance of this result is twofold. First, also in the multivariate case, the least informative distribution under second moment constraints is of normal type. Moreover, if the covariances are unknown, it is seen that the MaxEnt principle yields independent distributions.

A shortcoming of the maximum entropy approach is that the resulting distributions are not invariant with respect to reparametrizations of the unknown vector. To make an example, we have already seen that the maximum entropy distribution of θ in a finite interval $[1, 2]$ is uniform. On the other hand, if the reparametrization $\psi = 1/\theta$

is adopted and the MaxEnt approach is applied to ψ, the resulting prior will be a uniform distribution for ψ in [0.5,1], which corresponds to

$$\mathrm{p}(\theta) = \begin{cases} \frac{2}{\theta^2}, & 1 \leq \theta \leq 2 \\ 0, & \text{elsewhere,} \end{cases}$$

which is obviously different from a uniform distribution. A possible way to limit arbitrariness is to specify that, before applying the MaxEnt principle, one should first identify the "object of interest". Indeed, choosing either θ or $1/\theta$ as object of interest is going to yield different MaxEnt priors.

4.7 Model Approximation via Optimal Projection ⋆

Approximate low-order models are commonly used even when there is awareness that the real data are generated by a more complex model. Motivations may range from their use for control design purposes to better interpretability of the phenomena under investigation. Unfortunately, under model misspecification, several nice properties enjoyed by standard estimators are no more valid. In particular, a naive application of the least squares may provide far less than satisfactory results. In this section, it is shown that, within the Bayesian framework, the search for an optimal approximate model can be given a rigorous formulation that admits a projection-based solution.

We assume that the data Y are distributed according to (4.9), which summarizes our state of knowledge. However, rather than resorting to Bayesian estimation of the vector θ, an approximate model, typically of low order, is searched for. For instance, if θ_i were the samples of an impulse response, one might be interested in approximating them by a parametric model:

$$\theta \simeq g(\zeta), \quad g(\zeta) = \begin{bmatrix} g_1(\zeta) & \cdots & g_n(\zeta) \end{bmatrix}^T,$$

where $\zeta = \begin{bmatrix} \zeta_1 & \cdots & \zeta_q \end{bmatrix}^T$ is the unknown parameter vector. For example, in order to approximate the sequence θ_i by means of a single exponential function, it suffices to let $q = 2$ and

$$g_i(\zeta) = \zeta_1 e^{\zeta_2 i},$$

where ζ_1 is the amplitude and ζ_2 is the rate coefficient of the exponential.

A very natural estimator is the least squares one:

$$\zeta^{LS} = \arg\min_{\zeta} \|Y - \Phi g(\zeta)\|^2.$$

Note that ζ^{LS} coincides with the maximum likelihood estimate if the following model is assumed:

$$Y = \Phi g(\zeta) + E, \quad E \sim \mathcal{N}(0, \sigma^2 I_N).$$

In the present context, however, no claim is made that reality conforms to our approximate model. It may well be that the true θ, being more complex than its parsimonious parametric model $g(\zeta)$, is better represented by the model (4.9). Nevertheless, we are interested in finding the best approximation of θ within a set $\mathscr{P} = \{g(\zeta)|\zeta \in \mathbb{R}^q, \}$ of parametric approximations.

Under model (4.9), the optimal approximate model g^* can be defined as the one that minimizes the mean squared error $\mathscr{E}\|\theta - g\|^2$. For a generic model $g = g(\zeta)$, parametrized by the vector $\zeta \in \mathbb{R}^q$, $q \leq n$, we have that

$$g^* = g(\zeta^*), \quad \zeta^* := \arg\min_{\zeta} \mathscr{E}\left[\|\theta - g(\zeta)\|^2 | Y\right], \tag{4.41}$$

where the conditional expectation is taken with reference to the probability measure specified by (4.9). The following theorem, whose proof is in Sect. 4.13.5, was first derived in the context of linear system identification [20]. It shows that the optimal approximation is the projection of the Bayes estimate θ^{B} onto the set $\mathscr{P}$.

Theorem 4.6 (Optimal approximation, based on [20]) *Assume that (4.9) holds. Then,*

$$\zeta^* = \arg\min_{\zeta} \|\theta^{\mathrm{B}} - g(\zeta)\|^2. \tag{4.42}$$

In view of the last theorem, the best approximation $g(\zeta) \in \mathscr{P}$ can be computed by a two-step procedure. First, the Bayes estimate θ^{B} is obtained and in the second step the optimal $g(\zeta^*)$ is calculated as the solution of the least squares problem (4.42).

An interesting question is whether the obtained approximation is still optimal if the goal is minimizing the error, not with respect to θ, but with respect to the noiseless output $\Phi\theta$. In other words, the goal is finding g^o that minimizes $\|\Phi\theta - \Phi g^o)\|^2$. This can be done by introducing a weighted norm in the cost function:

$$g^o = g(\zeta^o), \quad \zeta^o := \arg\min_{\zeta} \mathscr{E}\left[\|\theta - g(\zeta)\|_W^2 \big| Y\right], \tag{4.43}$$

where $\|x\|_W^2$ stands for $x^T W x$. In particular, if $W = \Phi^T\Phi$, then

$$\|\theta - g(\zeta)\|_W^2 = \|\Phi\theta - \Phi g(\zeta)\|^2.$$

By extending the proof of Theorem 4.6 to the case of a weighted norm, the following projection result is obtained.

Theorem 4.7 (Optimal weighted approximation, based on [20]) *Assume that (4.9) holds. Then,*

$$\zeta^o = \arg\min_{\zeta} \|\theta^{\mathrm{B}} - g(\zeta)\|_W^2. \tag{4.44}$$

The consequence is that different approximations g^o are obtained depending on their prospective use. If the scope is just approximating θ, then $W = I_n$, but, if the scope is predicting the outputs, then $W = \Phi^T\Phi$ and a different result is obtained.

4.8 Equivalent Degrees of Freedom

In this section, the Bayesian estimation problem for the linear model is analysed by means of a diagonalization approach. The purpose is twofold: (i) the equivalent degrees of freedom of the Bayesian estimator are introduced together with their relationship with suitable weighted squared sums of residuals and squared sums of estimated parameters; (ii) it is shown that η^{ML}, the ML estimate of the hyperparameter vector, satisfies meaningful conditions involving the degrees of freedom. Finally, the obtained results are applied to the tuning of the regularization parameter, defined as the ratio between scaling factors for the noise variance Σ_E and the parameter variance Σ_θ. For the sake of simplicity, in this section, we assume $\mu_\theta = 0$.

Let us consider the case when the hyperparameters are just two scaling factors for the covariance matrices Σ_E and Σ_θ, that is,

$$\Sigma_\theta = \lambda K, \quad \lambda > 0 \tag{4.45}$$

$$\Sigma_E = \sigma^2 \Psi, \quad \sigma^2 > 0 \tag{4.46}$$

$$\eta = \begin{bmatrix} \lambda & \sigma^2 \end{bmatrix}^T, \tag{4.47}$$

where K and Ψ are known definite positive matrices. In such a case, it is immediate to see that the Bayes estimate

$$\theta^{\mathrm{B}} = \left(\Phi^T \Psi^{-1} \Phi + \frac{\sigma^2}{\lambda} K^{-1} \right)^{-1} \Phi^T \Psi^{-1} Y$$

depends only on the ratio $\gamma = \sigma^2/\lambda$, which behaves as a deterministic regularization parameter. This means that only the ratio between the scaling factors is relevant to the computation of a point estimate, although both of them are needed to compute the posterior variance (4.14). When $\Psi = I_N$ and $K = I_n$, the above estimator provides a Bayesian interpretation to the classical ridge regression estimator. In particular, γ can be interpreted as a noise-to-signal ratio and its tuning reformulated as a statistical estimation problem.

Given a positive definite symmetric matrix S, let $S^{1/2} = \left(S^{1/2}\right)^T$ be its symmetric square root, i.e., $S^{1/2} S^{1/2} = S$. Now, consider the singular value decomposition

$$\Psi^{-1/2} \Phi K^{1/2} = U D V^T,$$

where U and V are square matrices such that $U^T U = I_N$ and $V^T V = I_n$ and $D \in \mathbb{R}^{N \times n}$ is a diagonal matrix with diagonal entries $\{d_i\}$, $i = 1, \ldots, n$, see (3.134). Moreover, define

$$\bar{Y} = U^T \Psi^{-1/2} Y$$
$$\bar{E} = U^T \Psi^{-1/2} E$$
$$\bar{\theta} = V^T K^{-1/2} \theta.$$

Observe that

$$\mathscr{E}\left(\bar{E}\bar{E}^T\right) = U^T\Psi^{-1/2}\mathscr{E}EE^T\Psi^{-1/2}U = \sigma^2 U^T U = \sigma^2 I_N.$$

Analogously, $\mathscr{E}\left(\bar{\theta}\bar{\theta}^T\right) = \lambda I_n$. Moreover,

$$\begin{aligned}\bar{Y} &= U^T\Psi^{-1/2}(\Phi\theta + E) = U^T\Psi^{-1/2}\Phi K^{1/2}VV^TK^{-1/2}\theta + \bar{E}\\ &= U^T U D V^T V\bar{\theta} + \bar{E} = D\bar{\theta} + \bar{E}.\end{aligned}$$

In view of these properties, it follows that the original Bayesian estimation problem admits the following diagonal reformulation:

$$\bar{Y} = D\bar{\theta} + \bar{E}, \quad \bar{E} \sim \mathscr{N}(0, \sigma^2 I_N), \quad \bar{\theta} \sim \mathscr{N}(0, \lambda I_n), \tag{4.48}$$

where $\bar{E}$ and $\bar{\theta}$ are independent of each other.

In view of statistical independence, we have N independent scalar models:

$$\begin{aligned}\bar{y}_i &= d_i\bar{\theta}_i + \bar{v}_i, \quad i = 1, \ldots, n\\ \bar{y}_i &= \bar{v}_i, \quad i = n+1, \ldots, N,\end{aligned}$$

where $\bar{v}_i \sim \mathscr{N}(0, \sigma^2)$, $i = 1, \ldots, N$, and $\bar{\theta}_i \sim \mathscr{N}(0, \lambda)$, $i = 1, \ldots, n$.

By (4.11), it is straightforward to see that the Bayes estimates are

$$\bar{\theta}_i^{\mathrm{B}} = \frac{\lambda d_i \bar{y}_i}{\sigma^2 + \lambda d_i^2} = \frac{d_i \bar{y}_i}{\gamma + d_i^2}, \quad i = 1, \ldots, n$$

or, in matrix form,

$$\bar{\theta}^{\mathrm{B}} = (D^T D + \gamma I_n)^{-1} D^T \bar{Y}.$$

Let the residuals be defined as $\bar{\varepsilon}_i = \bar{y}_i - \bar{d}_i\bar{\theta}_i^{\mathrm{B}}$, $i = 1, \ldots, N$, where

$$\bar{d}_i = \begin{cases} d_i, & 1 \le i \le n \\ 0, & n+1 \le i \le N \end{cases}. \tag{4.49}$$

Then, we have

$$\bar{\varepsilon}_i = y_i - \frac{\bar{d}_i^2 \bar{y}_i}{\gamma + \bar{d}_i^2} = \frac{\gamma \bar{y}_i}{\gamma + \bar{d}_i^2} \tag{4.50}$$

$$\mathscr{E}\bar{\varepsilon}_i^2 = \frac{\gamma^2 \mathscr{E}\bar{y}_i^2}{(\gamma + \bar{d}_i^2)^2} = \frac{\gamma^2(\bar{d}_i^2\lambda + \sigma^2)}{(\gamma + \bar{d}_i^2)^2} = \frac{\sigma^2\gamma}{\gamma + \bar{d}_i^2} = \sigma^2\left(1 - \frac{\bar{d}_i^2}{\gamma + \bar{d}_i^2}\right) \tag{4.51}$$

or, in matrix form,

$$\bar{\varepsilon} = \gamma(D^T D + \gamma I_N)^{-1}\bar{Y}, \quad \mathscr{E}\|\bar{\varepsilon}\|^2 = \sigma^2\left(N - \text{trace}(D(D^T D + \gamma I_N)^{-1}D^T)\right). \tag{4.52}$$

It is worth noting that the above relationships do not hold for a generic regularization parameter γ, but only when $\gamma = \sigma^2/\lambda$. In the remaining part, we present some results that were first derived in the context of Bayesian deconvolution in [5]. The proof of the following proposition is in Sect. 4.13.6.

Proposition 4.3 (based on [5]) *For a given hyperparameter vector η, let* WRSS *denote the following weighted squared sum of residuals:*

$$\text{WRSS} = (Y - \Phi\theta^{\text{B}})^T\Psi^{-1}(Y - \Phi\theta^{\text{B}}),$$

where $\theta^{\text{B}} = \mathscr{E}[\theta|Y, \eta]$. Then,

$$\mathscr{E}\,(\text{WRSS}) = \sigma^2(N - \text{trace}(H(\gamma))),$$

where

$$H(\gamma) = \Phi(\Phi^T\Psi^{-1}\Phi + \gamma K^{-1})^{-1}\Phi^T\Psi^{-1}$$

is the so-called hat matrix.

As already noted, see (3.64), when $\Sigma_E = \sigma^2 I_N$, the predicted output $\hat{Y} = \Phi\theta^{\text{B}}$ and the measured output Y are related through the hat matrix:

$$\hat{Y} = H(\gamma)Y.$$

In order to better understand the link between the hat matrix and the degrees of freedom, just consider the standard linear model $Y = \Phi\theta + E$, $\theta \in \mathbb{R}^n$, and the corresponding LS estimate $\theta^{\text{LS}} = (\Phi^T\Phi)^{-1}\Phi^T Y$. The predicted output is $\hat{Y} = H^{\text{LS}}Y$, where $H^{\text{LS}} = \Phi(\Phi^T\Phi)^{-1}\Phi^T$ enjoys the property $\text{trace}(H^{\text{LS}}) = n$.

It is this analogy that justifies the introduction of equivalent degrees of freedom which we already encountered in (3.65) as a function of the regularized estimate θ^{R} described in the deterministic context. Its definition, here derived starting from the stochastic context, is reported below stressing its dependence on the regularization parameter γ.

Definition 4.2 *(equivalent degrees of freedom)* The quantity

$$\text{dof}(\gamma) = \text{trace}(H(\gamma)), \quad 0 \le \text{dof}(\gamma) \le n \tag{4.53}$$

is called *equivalent degrees of freedom*.

In view of (4.52),

$$\text{dof}(\gamma) = \sum_{i=1}^{n} \frac{d_i^2}{d_i^2 + \gamma}$$

so that $\mathrm{dof}(\gamma)$ is a monotonically decreasing function of γ with $0 \leq \mathrm{dof}(\gamma) \leq n$. The equivalent degrees of freedom provide an easily understandable measure of the flexibility of estimator: for instance, if they are approximately equal to three, the Bayesian estimator has a flexibility comparable to a model with three parameters. For linear-in-parameter models estimated by ordinary or weighted least squares, the degrees of freedom coincide with the rank of the regressor matrix and, therefore, they can take only integer values. The equivalent degrees of freedom of the Bayesian estimator, conversely, are a nonnegative real number controlled by γ.

The next theorem establishes a connection between the degrees of freedom and the ML estimate

$$\eta^{\mathrm{ML}} = \left[\lambda^{\mathrm{ML}} \ \left(\sigma^2\right)^{\mathrm{ML}} \right]^T$$

of the hyperparameter vector. Accordingly, we define

$$\gamma^{\mathrm{ML}} = \frac{\left(\sigma^2\right)^{\mathrm{ML}}}{\lambda^{\mathrm{ML}}}.$$

Moreover, we introduce the following weighted squared sum of estimated parameters:

$$\mathrm{WPSS} = (\theta^{\mathrm{B}})^T K^{-1} \theta^{\mathrm{B}} = \left\| \bar{\theta}^{\mathrm{B}} \right\|^2 = \sum_{i=1}^{n} \frac{d_i^2 \bar{y}_i^2}{(\gamma + d_i^2)^2}. \tag{4.54}$$

The proof of the following result is in Sect. 4.13.7.

Theorem 4.8 (based on [5]) *Assume that model (4.9) holds where Σ_θ and Σ_E are as in (4.46)–(4.47). Then, the* ML *estimates of the hyperparameters satisfy the following necessary conditions:*

$$\mathrm{WRSS} = \left(\sigma^2\right)^{\mathrm{ML}} \left(N - \mathrm{dof}(\gamma^{\mathrm{ML}})\right) \tag{4.55}$$

$$\mathrm{WPSS} = \lambda^{\mathrm{ML}} \mathrm{dof}(\gamma^{\mathrm{ML}}). \tag{4.56}$$

By taking the ratio between (4.55) and (4.56), the following proposition is obtained.

Proposition 4.4 (based on [5]) *If λ^{ML} and $\left(\sigma^2\right)^{\mathrm{ML}}$ are nonnull and finite, then*

$$\gamma^{\mathrm{ML}} = \frac{\mathrm{dof}(\gamma^{\mathrm{ML}})}{N - \mathrm{dof}(\gamma^{\mathrm{ML}})} \frac{\mathrm{WRSS}}{\mathrm{WPSS}}. \tag{4.57}$$

This last corollary can be used as a simple and practical tuning procedure as it requires just a line search on the scalar γ. Of course, (4.57) relies on the necessary conditions of Theorem 4.8, so that one has to check if the solution corresponds to a maximum of the likelihood function.

4.9 Bayesian Function Reconstruction

In this section, the Bayesian estimation approach is illustrated through its application to the reconstruction of an unknown function from noisy samples. The observations will be generated by adding pseudorandom noise to a known function $g(x)$, so that the performances of alternative estimators can be directly assessed by comparison with the ground truth. The selected $g(x)$ is the same function (3.26) used in the previous chapter in order to illustrate polynomial regression:

$$g(x) = (\sin(x))^2(1 - x^2), \quad x \in [0,\ 1]. \tag{4.58}$$

Also the noise model is the same:

$$y_i = g(x_i) + e_i, \quad i = 1, \ldots, N. \tag{4.59}$$

We let $N = 40$, $x_1 = 0$, $x_{40} = 1$, and $x_2, \ldots, x_{39}$ are evenly spaced points between x_1 and x_{40}. Finally, e_i, $i = 1, \ldots, 40$, are i.i.d. Gaussian distributed with mean zero and standard deviation 0.034.

The problem of estimating $\theta_i = g(t_i)$, i.e., the samples of the unknown function, is a particular case of the linear Gaussian model (4.9) with $\Phi = I_N$, that is,

$$Y = \theta + E, \quad E \sim \mathcal{N}(0, \sigma^2 I_N). \tag{4.60}$$

Since Φ is square, in this case, the number n of unknowns coincides with the number N of observations.

The noisy data and the true function are displayed in the top left panel of Fig. 4.1. It is assumed that the available prior knowledge regards the "regularity" of $g(\cdot)$ and the knowledge that $g(0) = 0$. A possible probabilistic translation of this qualitative knowledge is assuming that θ_i is a so-called random walk:

$$\theta_i = \theta_{i-1} + w_i, \quad i = 1, \ldots, N, \quad \theta_0 = 0,$$

where $w_i \sim \mathcal{N}(0, \lambda)$ are independent random variables. In fact, under the random walk model, the first difference

$$\theta_i - \theta_{i-1} = w_i$$

has a finite variance, equal to λ. Hence, if we approximate the derivative of $g(\cdot)$ by the first difference $\theta_i - \theta_{i-1}$, this approximation is less than $1.96\sqrt{\lambda}$ with probability 0.95, which guarantees that the profile of the function cannot vary too quickly. Note that, due to the qualitative nature of the prior knowledge, the precise value of λ is unknown, so that it has to be treated as a hyperparameter. Conversely, it is assumed that the true value of σ^2 is known. Summarizing, we have

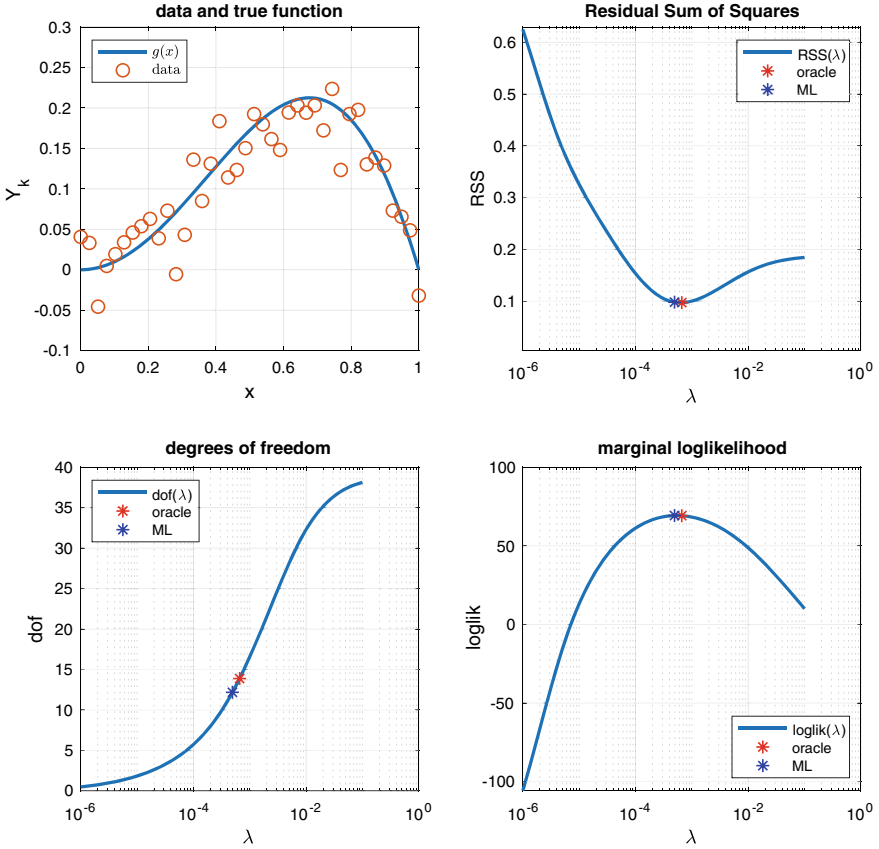

Fig. 4.1 Function reconstruction example. Top left: noisy data and true function. Top right, bottom left and bottom right: Residual sum of squares, i.e., the sum of the squared differences between the function values and their estimates, degrees of freedom and marginal loglikelihood against the hyperparameter λ. The oracle denotes the value that minimizes RSS while ML indicates the maximizer of the marginal likelihood

$$\theta_i = \sum_{j=1}^{i} w_j, \quad i = 1, \ldots, N$$

or, in matrix form,

$$\theta = Fw, \quad F = \begin{bmatrix} 1\,0\,0\,\ldots\,0 \\ 1\,1\,0\,\ldots\,0 \\ 1\,1\,1\,\ldots\,0 \\ \vdots\,\vdots\,\vdots\,\ddots\,\vdots \\ 1\,1\,1\,\ldots\,1 \end{bmatrix}, \quad w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_N \end{bmatrix}.$$

Observing that $\mathrm{Var}(w) = \lambda I_N$, the prior variance of the parameter vector is

$$\Sigma_\theta = \lambda F F^T = \lambda \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & 2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 2 & \dots & N \end{bmatrix}.$$

For a given λ, the Bayes estimate θ^{B} is obtained according to (4.10) and can be written as

$$\theta^{\mathrm{B}} = \Sigma_\theta \left(\Sigma_\theta + \sigma^2 I_N\right)^{-1} Y.$$

The corresponding equivalent degrees of freedom, obtained by (4.53), are now thought as a (monotonically nondecreasing) function of λ, i.e.,

$$\mathrm{dof}(\lambda) = \mathrm{trace}\, H(\lambda), \quad H(\lambda) = \Sigma_\theta \left(\Sigma_\theta + \sigma^2 I_N\right)^{-1}, \quad \Sigma_\theta = \lambda F F^T.$$

In the bottom left panel of Fig. 4.1, the degrees of freedom are plotted against λ. For small values of λ they are close to zero and get closer to $N = 40$ as λ goes to infinity. It is a rather general feature that the function $\mathrm{dof}(\lambda)$ is better visualized on a semilog scale. In order to tune the regularization parameter λ, one can resort to the maximization of the marginal loglikelihood:

$$\lambda^{\mathrm{ML}} = \arg\max_\lambda \left\{ -\frac{1}{2} \log(2\pi \det(\Sigma)) - \frac{1}{2} Y^T \Sigma^{-1} Y \right\}$$
$$\Sigma = \Sigma_\theta + \sigma^2 I_N = \lambda F F^T + \sigma^2 I_N.$$

It turns out that $\lambda^{\mathrm{ML}} = 4.92e-4$, the corresponding degrees of freedom being 12.17. For the sake of comparison, $\lambda = 6.61e-4$ is the best possible value, i.e., the one provided by an oracle that exploits the knowledge of the true function in order to minimize the sum of the squared reconstruction errors. This latter quantity is function of λ and here denoted by $\mathrm{RSS}(\lambda)$. As seen in the top right panel of Fig. 4.1, marginal likelihood maximization achieves $\mathrm{RSS} = 9.80e-2$, not much worse than $\mathrm{RSS} = 9.71e-2$ achieved by the oracle, whose associated degrees of freedom are 13.88. Therefore, in this specific case, the marginal likelihood criterion somehow underestimates the complexity of the model.

In Fig. 4.2, the estimates obtained in correspondence of six different values of λ are displayed. It is apparent that for $\lambda = 1e-6$ and $\lambda = 1e-5$ the estimated function is overregularized, while overfitting occurs for $\lambda = 1e-1$ and $\lambda = 1e-2$. The two bottom panels display the oracle and ML estimates, the former exhibiting a slightly more regular profile.

Finally, observing that in our case $\Sigma_{\theta Y} = \Sigma_\theta$, we have

$$\Sigma_{\theta|Y} = \mathrm{Var}(\theta|Y) = \Sigma_\theta - \Sigma_\theta \Sigma_Y^{-1} \Sigma_\theta$$

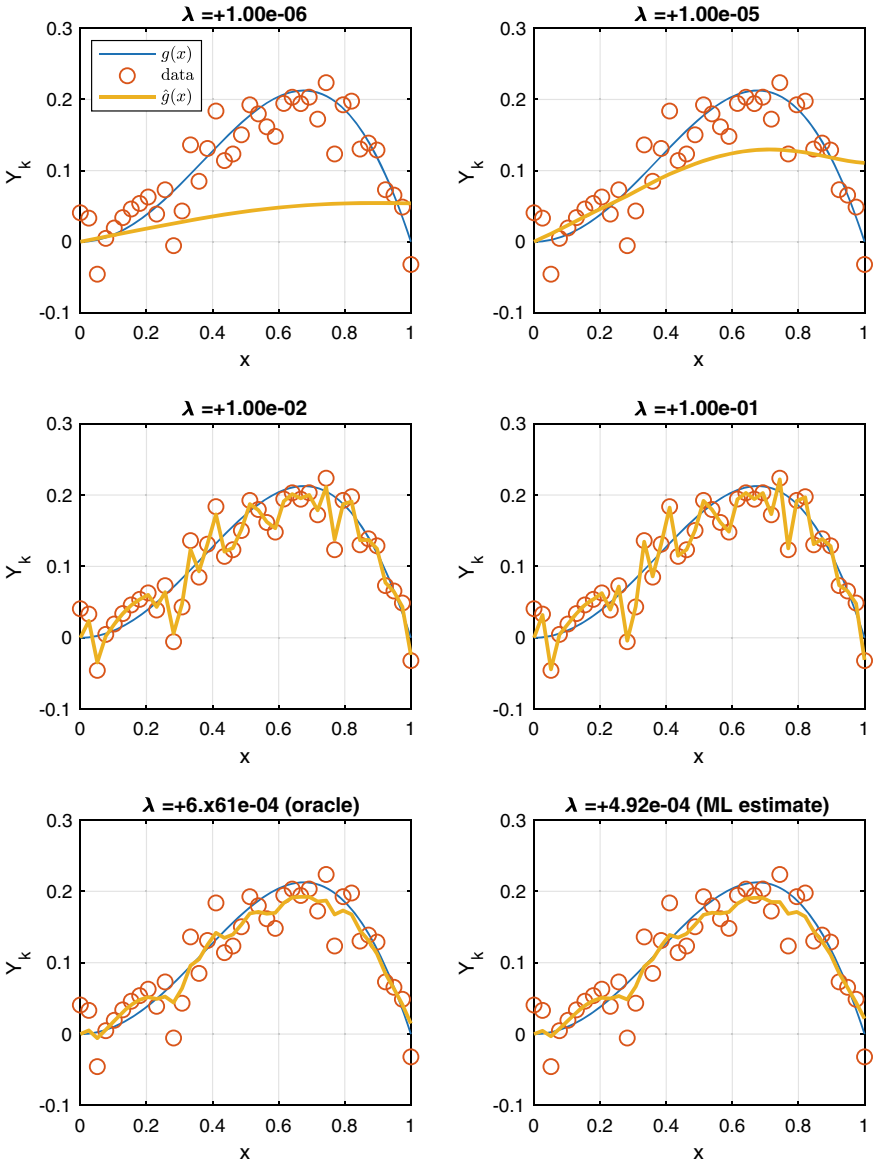

Fig. 4.2 Function reconstruction example. The panels display the Bayes estimates $\hat{g}(x)$ corresponding to six different values of the hyperparameter λ, including the one provided by the oracle and the maximum likelihood one

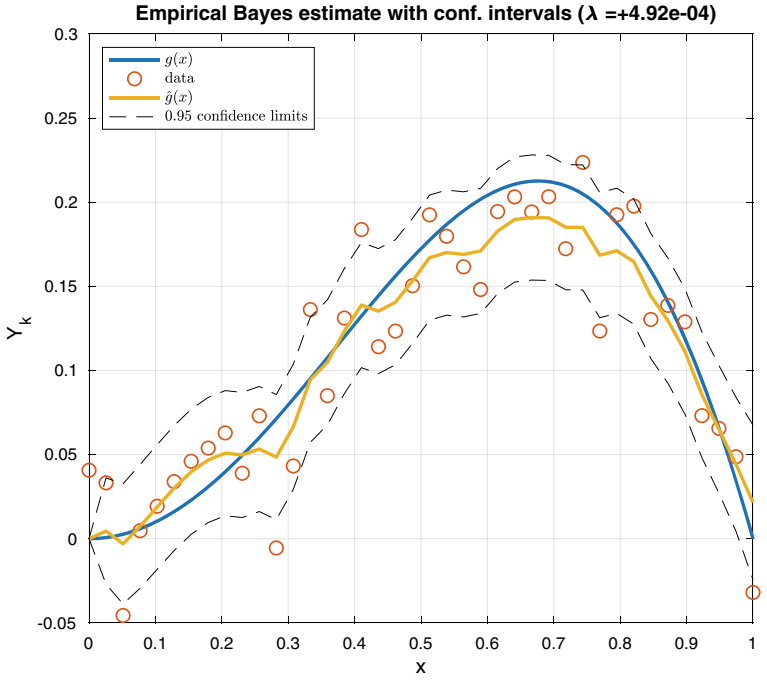

Fig. 4.3 Function reconstruction example. True function and Empirical Bayes estimate $\hat{g}(x)$ based on λ^{ML} together with its 95% Bayesian credible intervals

and we can compute the 95% Bayesian credible intervals, according to (4.8). As it can be seen from Fig. 4.3, the credible limits successfully capture the uncertainty, as demonstrated by the fact that the true function lies within the limits.

This simple example has shown that Bayesian estimation can be effectively employed in order to reconstruct an unknown function without need of assuming a specific parametric structure, e.g., polynomial or other. The key idea is the use of a smoothness prior, expressed through the assumed prior distribution of the first differences of the function. The associated variance λ is treated as a hyperparameter that can be tuned via marginal likelihood maximization. Altogether, this is a flexible Empirical Bayes scheme that can be employed as a general-purpose black-box estimator.

Of interest is also the fact that the considered function could have been the impulse response of a dynamical system. In this respect, the example highlights also the limits of the approach. A first issue, easily fixable, has to do with the insufficient smoothness of the estimate. As seen in Fig. 4.3, the true function is significantly smoother than its estimate. As a matter of fact, it is not difficult to increase the regularity of the Bayes estimate: for instance, it suffices to assume that the samples $\theta_i = g(x_i)$ are an integrated random walk:

$$\theta_i = \theta_{i-1} + \xi_i$$
$$\xi_i = \xi_{i-1} + w_i,$$

where $w_i \sim \mathcal{N}(0, \lambda)$ are again independent and identically distributed. This prior distribution is going to yield smoother profiles. Rather interestingly, the obtained estimate can be seen as the discrete-time counterpart of cubic smoothing splines, a method widely used for the nonparametric reconstruction of unknown functions.

A more serious issue regards extrapolation properties of the estimate that are in turn connected with the type of asymptotic decay shown by stable impulse responses. As it can be seen from Fig. 4.3, oscillations and credible intervals do not tend to dampen as x increases. While it would be easy to compute the Bayes estimate also for values far beyond the observation window, the result would be disappointing. Indeed, coherently with the diffusive nature of random walks, the width of the credible band would diverge, which is unnecessarily conservative when a stable impulse response is reconstructed. It appears that the task of identifying impulse responses calls for prior distributions that are specifically suited to the their features, especially the asymptotic ones. The development of these prior distributions, or equivalently the design of suitable regularization penalties, will be a central topic of the subsequent chapters.

4.10 Markov Chain Monte Carlo Estimation

As already mentioned in Sect. 4.4, in the full Bayesian approach the estimate

$$\mathrm{p}(\theta|Y) = \int \mathrm{p}(\theta, \eta|Y)d\eta = \int \mathrm{p}(\theta|\eta, Y)\mathrm{p}(\eta|Y)d\eta$$

requires a marginalization with respect to the hyperparameter vector η. In general, this integral cannot be computed analytically. Nevertheless it can be computed numerically by means of Markov Chain Monte Carlo (MCMC) methods that generate pseudorandom samples drawn from the joint posterior density $\mathrm{p}(\theta, \eta|Y)$. The Gibbs sampling (GS) algorithm is the most straightforward and popular MCMC method. Its goal is to simulate a realization of a Markov chain, whose samples, though not independent of each other, form an ergodic process whose stationary distribution coincides with the desired posterior. Hence, provided that the burn-in phase is discarded, the posterior distribution is approximated by the histogram of the samples. In order to generate the samples, at each step a random extraction is made from a proposal distribution. In the Gibbs sampler, the proposal distribution is the so-called full conditional, that is, the probability of a given element of the parameter vector given the data and the current values of all other elements.

For the linear Gaussian model (4.9), a Gibbs sampler may be implemented as follows:

1. Select initializations η^0, θ^0, and let $k = 0$.
2. Draw a sample $\eta^{(k+1)}$ from the full conditional distribution $\mathrm{p}(\eta|\theta^{(k)}, Y)$.
3. Draw a sample $\theta^{(k+1)}$ from the full conditional distribution $\mathrm{p}(\theta|\eta^{(k+1)}, Y)$.
4. If $k = k_{max}$, end, else $k = k + 1$ and go to step 2.

This stochastic simulation algorithm generates a Markov chain whose stationary distribution coincides with $\mathrm{p}(\theta, \eta|Y)$. Therefore, though correlated, the generated samples $\{\theta^{(k)}, \eta^{(k)}\}$ can be used to estimate the (joint and marginal) posterior distributions and also the posterior expectations via the proper sample averages. For example,

$$\frac{1}{N}\sum_{k=1}^{N}\theta^{(k)} \simeq \mathscr{E}(\theta|Y).$$

The choice of the prior distributions $\mathrm{p}(\theta|\eta)$ and $\mathrm{p}(\eta|Y)$ has a critical influence on the efficiency of the scheme. The priors are called *conjugate*, when for each parameter the prior and the full conditional belong to the same distribution family. This implies that the same random variable generators can be used throughout the simulation.

Consider model (4.9), where Σ_E is known and $\Sigma_\theta = \lambda K$, with λ unknown. Below, we describe a Gibbs sampling scheme for obtaining the posterior distributions of θ and $\eta = \lambda$. For θ, the prior is $\theta|\lambda \sim \mathscr{N}(0, \lambda K)$. A conjugate prior for λ is the inverse Gamma distribution:

$$\frac{1}{\lambda} \sim \Gamma(g_1, g_2), \quad g_1, g_2 > 0.$$

In other words, it is assumed that $1/\lambda$ is distributed as a Gamma random variable, so that

$$\mathrm{p}\left(\frac{1}{\lambda}\right) \propto \left(\frac{1}{\lambda}\right)^{g_1-1} e^{-\left(\frac{g_2}{\lambda}\right)}.$$

With this choice of the prior, the full conditional of $1/\lambda$ will be distributed as a suitable Gamma variable, $\forall k$. More precisely, it can be shown that, if

$$\mathrm{p}\left(\bar{\theta}|\lambda\right) \sim \mathscr{N}(0, \lambda I_N), \quad \mathrm{p}\left(\frac{1}{\lambda}\right) \sim \Gamma(g_1, g_2)$$

then

$$\mathrm{p}\left(\frac{1}{\lambda}\middle|\bar{\theta}\right) \sim \Gamma\left(g_1 + \frac{N}{2}, g_2 + \frac{\left\|\bar{\theta}\right\|^2}{2}\right). \tag{4.61}$$

Recall that the mean and variance of the Gamma random variable are g_1/g_2 and g_1/g_2^2, respectively. For the prior to be as uninformative as possible, we let g_1 and g_2 decrease to zero. Under these assumptions, the Gibbs sampler unfolds as follows:

1. Initialize λ and θ, e.g., using the empirical Bayes estimates

$$\lambda^{(0)} = \lambda^{ML}, \quad \theta^0 = \theta^{\mathrm{B}} = \mathscr{E}(\theta|\lambda^{ML}, Y)$$

and let $k = 0$.

2. Draw a sample $1/\lambda^{(k+1)}$ from the full conditional distribution

$$\mathrm{p}\left(\frac{1}{\lambda}\middle|\theta^{(k)}, Y\right) = \mathrm{p}\left(\frac{1}{\lambda}\middle|\theta^{(k)}\right) = \Gamma\left(\frac{N}{2}, \frac{\theta^{(k)T} K^{-1} \theta^{(k)}}{2}\right). \tag{4.62}$$

3. Draw a sample $\theta^{(k+1)}$ from the full conditional distribution

$$\mathrm{p}\left(\theta\middle|\lambda^{(k+1)}, Y\right) = \mathcal{N}\left(\mathscr{E}(\theta|\lambda^{(k+1)}, Y), \mathrm{Var}(\theta|\lambda^{(k+1)}, Y)\right)$$

whose mean and variance are obtained according to (4.10) or (4.13).

4. If $k = k_{max}$, end, else $k = k + 1$ and go to step 2.

Above, the expression of the full conditional (4.62) is a direct consequence of the conjugacy property (4.61), as it can be seen by letting $\bar{\theta} = K^{-1/2}\theta^{(k)}$, where $K^{-1/2}$ is a symmetric matrix such that $K^{-1/2}K^{-1/2} = K^{-1}$.

When there are other hyperparameters to tune, e.g., the noise variance σ^2, the MCMC scheme can be properly extended. Provided that they exist, conjugate priors ensure an efficient sampling from the proposal distributions that generate the random samples, although a variety of MCMC schemes are available that deal with non-conjugate priors at the cost of an increased computational effort.

The main advantage of MCMC methods is that they implement the full Bayesian framework that is only approximated by the empirical Bayes scheme. In particular, MCMC methods do not neglect the hyperparameter uncertainty which is correctly propagated to the parameter estimate. However, as already discussed in Sect. 4.4, if data are informative enough to ensure a precise estimate of the hyperparameters, the difference between MCMC and empirical Bayes estimates (and associated credible regions) may be of minor importance.

4.11 Model Selection Using Bayes Factors

As discussed in Sect. 2.6.2, one fundamental issue is the selection of the "best" model inside a class of postulated structures. In the classical setting, this can be performed using criteria like AIC (2.34) and BIC (2.36) or adopting a cross-validation strategy. We will now see that the Bayesian approach provides a powerful alternative based on the concept of posterior model probability.

Let $\mathscr{M}^i$ be a model structure parametrized by the vector x^i. In the system identification scenario discussed in Chap. 2, the structures could be ARMAX models of different complexity. Hence, each x^i would correspond to the θ^i parametrizing (2.1) and containing the coefficients of rational transfer functions of different orders. If little knowledge on them were available, poorly informative prior densities could be assigned. Another example concerns the function estimation problem illustrated in Sect. 4.9. Here, x^i could contain the samples θ^i of the unknown function g modelled

as a stochastic process. Then, the different structures could represent different covariances of g defined by a random walk or an integrated random walk. Each covariance would then depend on an unknown hyperparameter vector η^i containing the variance of the random walk increments and possibly also of the measurement noise. So, in this case, one would have $x^i = [\theta^i \ \eta^i]$. Here, η^i is a random vector with flat priors typically assigned to the variances to include just nonnegativity information.

Now, suppose that we are given m competitive structures $\mathscr{M}^i$. An important conceptual step is to interpret even them as (discrete) random variables, each having probability $\Pr(\mathscr{M}^i)$ before seeing the data Y. The selection of the best model has then a natural answer: one should select the structure having the largest posterior probability $\Pr(\mathscr{M}^i|Y)$. Using Bayes rule, one has

$$\Pr(\mathscr{M}^i|Y) = \frac{\int \mathrm{p}(Y|\mathscr{M}^i, x^i)dx^i \Pr(\mathscr{M}^i)}{\mathrm{p}(Y)}.$$

A typical choice is to think of the structures as equiprobable, so that $\Pr(\mathscr{M}^i) = 1/m$ for any i. Then, one can select the $\mathscr{M}^i$ maximizing the so-called Bayesian evidence given by

$$\mathrm{p}(Y|\mathscr{M}^i) = \int \mathrm{p}(Y|\mathscr{M}^i, x^i)dx^i.$$

Note that this corresponds to the marginal likelihood where all the parameter uncertainty connected with the i-th structure has been integrated out. Given two structures $\mathscr{M}^1$ and $\mathscr{M}^2$, the Bayes factor is also defined as follows:

$$B_{12} = \frac{\mathrm{p}(Y|\mathscr{M}^1)}{\mathrm{p}(Y|\mathscr{M}^2)}.$$

Hence, large values of B_{12} indicate that data strongly support $\mathscr{M}^1$ as opposed to $\mathscr{M}^2$.

For the computation of the Bayesian evidence, the same numerical considerations reported at the end of Sect. 4.4 then hold. In particular, when the evidence cannot be computed explicitly, approximations are needed given by the Laplace approximation. Also the BIC criterion is often adopted. In particular, in the function estimation problem one can integrate out θ. Then, one can evaluate the complexity of the model using the marginal likelihood optimized w.r.t. the hyperparameters η^i, then adding a term which penalizes the dimension of the hyperparameter vector. This will be also discussed later on in Sect. 7.2.1.1.

MCMC can be also used to compute the evidence by simulating from posterior distributions and using the harmonic mean of the likelihood values, see Sect. 4.3 in [14]. A more powerful and complex approach employs MCMC techniques able to jump between models of different dimensions, an approach known in the literature as reversible jump Markov chain Monte Carlo computation [10].

4.12 Further Topics and Advanced Reading

There is an extensive literature debating on the interpretation of probability as a quantification of personal belief and it would be impossible to give a satisfactory account of all the contributions. The reader interested in studying motivations and foundations of *subjective probability* may refer to [4, 16]. One of the merits of Bayesian probability is its efficacy in addressing ill-posed and ill-conditioned problems, including also a wide class of statistical learning problems. The connection between deterministic regularization and Bayesian estimation has been pointed out by several authors in different contexts. Two examples related to spline approximation and neural networks are given by [8, 15].

The choice and tuning of the priors is undoubtedly the crux of any Bayesian approach. It is not a surprise that the tuning of hyperparameters via the *Empirical Bayes* approach emerged early as a practical and effective way to deploy Bayesian methods in real-world contexts, see [6] for its use in the study of the James–Stein estimator. Since the 1980s, thanks to the advent of Markov chain Monte Carlo methods, *full Bayesian* approaches have become a viable alternative, motivating reflections on the pros and cons of the two approaches, see, for instance, [17]. In particular, the connection between Stein's Unbiased Risk Estimator (SURE), equivalent degrees of freedom and the robustness of marginal likelihood hyperparameter tuning has been investigated by [1, 21]. The choice of the prior distributions is somehow more controversial. In the present chapter, we exposed the principles of the maximum entropy approach, mainly following [12], but other approaches have been advocated for finding non-informative priors. A requirement could be invariance with respect to change of coordinates, enjoyed, for instance, by Jeffreys' prior [13].

It not unusual to have parameters that should be left immune from regularization. In the Bayesian approach, this corresponds to the absence of prior information, usually expressed through an infinite variance prior. Although the case could be treated by assigning large variances to some parameters, it is numerically more robust useful to use the exact formulas. Their derivation by a limit argument followed [22].

The idea of deriving approximated parametric models by a suitable projection of the Bayes estimate conforms to Hjalmarsson's advice "always first model as well as possible" [11]. The projection result has been derived in [23] for Gaussian processes and subsequently extended to general distributions in [20].

The equivalent degrees of freedom of a regularized estimator have been studied in the context of smoothing by additive [2] and spline models [3, 9], while a discussion specialized to the case of Bayesian estimation can be found in [5, 17].

Starting by the seminal paper [7], the use of stochastic simulation for computing posterior distributions according to a full Bayesian paradigm has gained a wider and wider adoption, especially when there exist conjugate priors that allow efficient sampling schemes. In particular, this is possible for the linear model discussed in this chapter, whose MCMC estimation is discussed in [18].

4.13 Appendix

4.13.1 Proof of Theorem 4.1

For simplicity, the proof is given in the scalar parameter case. We have that

$$\begin{aligned}\frac{d\text{MSE}(\hat{\theta})}{d\hat{\theta}} &= \frac{d}{d\hat{\theta}}\int_{-\infty}^{+\infty}(\hat{\theta}-\theta)^2\text{p}(\theta|Y)d\theta \\ &= 2\hat{\theta}\int_{-\infty}^{+\infty}\text{p}(\theta|Y)d\theta - 2\int_{-\infty}^{+\infty}\theta\text{p}(\theta|Y)d\theta \\ &= 2\left(\hat{\theta}-\mathscr{E}\left[\theta|Y\right]\right).\end{aligned}$$

Moreover,

$$\frac{d^2\text{MSE}(\hat{\theta})}{d\hat{\theta}^2} = 2\int_{-\infty}^{+\infty}\text{p}(\theta|Y)d\theta = 2 > 0.$$

Therefore, $\theta^{\text{B}} = E\left[\theta|Y\right]$ minimizes $\text{MSE}(\hat{\theta})$.

4.13.2 Proof of Theorem 4.2

Let $X = \theta^{\text{B}} - \theta$ denote the estimation error. Recalling that $\mathscr{E}[Y - \Phi\mu_\theta] = \mathscr{E}[E] = 0$, from (4.10) it follows that $\mathscr{E}X = 0$. Note also that X and Y are jointly Gaussian and

$$\text{Cov}(X,Y) = \mathscr{E}[X(Y-\mathscr{E}Y)^T] = \mathscr{E}[XY^T] - \mathscr{E}X\mathscr{E}Y^T = \mathscr{E}[XY^T].$$

Now, using (4.7), we have

$$\begin{aligned}\mathscr{E}[XY^T] &= \mathscr{E}\left[(\theta^{\text{B}}-\theta)Y^T\right] \\ &= \Sigma_{\theta Y}\Sigma_Y^{-1}\mathscr{E}\left[(Y-\mu_Y)Y^T\right] - \mathscr{E}\left[(\theta-\mu_\theta)Y^T\right] \\ &= \Sigma_{\theta Y}\Sigma_Y^{-1}(\mathscr{E}\left[(YY^T\right] - \mu_Y\mu_Y^T) - \mathscr{E}\left[\theta Y^T\right] - \mu_\theta\mu_Y^T \\ &= \Sigma_{\theta Y}\Sigma_Y^{-1}\Sigma_Y - \Sigma_{\theta Y} = 0.\end{aligned}$$

4.13.3 Proof of Lemma 4.1

By applying the matrix inversion lemma (3.145) and proceeding with simple matrix manipulations,

$$\begin{aligned}\Sigma_\theta \Phi^T(\Sigma_E + \Phi\Sigma_\theta\Phi^T)^{-1} &= \Sigma_\theta\Phi^T(\Sigma_E^{-1} - \Sigma_E^{-1}\Phi(\Phi^T\Sigma_E^{-1}\Phi + \Sigma_\theta^{-1})^{-1}\Phi^T\Sigma_E^{-1}) \\ &= \Sigma_\theta\Phi^T\Sigma_E^{-1} - \Sigma_\theta\Phi^T\Sigma_E^{-1}\Phi(\Phi^T\Sigma_E^{-1}\Phi + \Sigma_\theta^{-1})^{-1}\Phi^T\Sigma_E^{-1} \\ &= \Sigma_\theta(I - \Phi^T\Sigma_E^{-1}\Phi(\Phi^T\Sigma_E^{-1}\Phi + \Sigma_\theta^{-1})^{-1})\Phi^T\Sigma_E^{-1} \\ &= \Sigma_\theta(\Phi^T\Sigma_E^{-1}\Phi + \Sigma_\theta^{-1} - \Phi^T\Sigma_E^{-1}\Phi)(\Phi^T\Sigma_E^{-1}\Phi + \Sigma_\theta^{-1})^{-1}\Phi^T\Sigma_E^{-1} \\ &= (\Phi^T\Sigma_E^{-1}\Phi + \Sigma_\theta^{-1})^{-1}\Phi^T\Sigma_E^{-1}.\end{aligned}$$

4.13.4 *Proof of Theorem 4.3*

In view of (4.13), the conditional variance is

$$\mathrm{Var}(\theta|Y) = \left(\frac{\Phi^T\Phi}{\sigma^2} + \Sigma_\theta^{-1}\right)^{-1} = \sigma^2\left(\Phi^T\Phi + \sigma^2\begin{bmatrix} a^{-1}I_{n-p} & 0 \\ 0 & \Sigma^{-1}. \end{bmatrix}\right)^{-1}.$$

In view of (4.7)

$$\mathscr{E}(\theta|Y) = \Sigma_\theta\Phi^T(\Phi\Sigma_\theta\Phi^T + \sigma^2 I_n)^{-1}Y = \begin{bmatrix} \Sigma\Omega^T \\ a\Psi^T \end{bmatrix}(a\Psi\Psi^T + M)^{-1}Y.$$

By replicating the passages of Lemma 4.1

$$a\Psi(a\Psi\Psi^T + M)^{-1} = \left(\Psi^T M^{-1}\Psi + \frac{I_{n-p}}{a}\right)^{-1}\Psi^T M^{-1}.$$

Moreover, by applying the matrix inversion lemma, see (3.145),

$$\begin{aligned}(a\Psi\Psi^T + M)^{-1} &= M^{-1} - M^{-1}\Psi\left(\Psi^T M^{-1}\Psi + \frac{I_{n-p}}{a}\right)^{-1}\Psi^T M^{-1} \\ &= M^{-1} - M^{-1}\Psi(\Psi^T M^{-1}\Psi)^{-1}\left(I_{n-p} + \frac{1}{a}(\Psi^T M^{-1}\Psi)^{-1}\right)^{-1}\Psi^T M^{-1}.\end{aligned}$$

Then, letting $a \to \infty$ complete the proof. Observe that all the inverse matrices appearing in the proof exist due to the full-rank assumptions made on Φ and Ψ.

4.13.5 *Proof of Theorem 4.6*

The expectation in (4.41) can be rewritten as

$$
\begin{aligned}
&\mathscr{E}\left[\left\|\theta-\theta^{\mathrm{B}}+\theta^{\mathrm{B}}-g(\zeta)\right\|^2\middle|Y\right] \\
&=\mathscr{E}\left[\left\|\theta-\theta^{\mathrm{B}}\right\|^2+2\left(\theta-\theta^{\mathrm{B}}\right)^T\left(\theta^{\mathrm{B}}-g(\zeta)\right)+\left\|\theta^{\mathrm{B}}-g(\zeta)\right\|^2\middle|Y\right] \\
&=\mathscr{E}\left[\left\|\theta-\theta^{\mathrm{B}}\right\|^2\middle|Y\right]+\mathscr{E}\left\|\theta^{\mathrm{B}}-g(\zeta)\right\|^2.
\end{aligned}
$$

The proof follows by observing that in the last equation the first term does not depend on ζ. In the last passage, we have exploited the fact that $\theta^{\mathrm{B}}|Y$ is deterministic and equal to $\mathscr{E}(\theta|Y)$.

4.13.6 Proof of Proposition 4.3

First observe that

$$
\mathrm{WRSS}=\|\bar{\varepsilon}\|^2=\sum_{i=1}^{N}\frac{\gamma^2\bar{y}_i^2}{(\gamma+\bar{d}_i^2)^2}. \tag{4.63}
$$

Hence, in view of (4.52)

$$
\mathscr{E}\mathrm{WRSS}=\sigma^2\left(N-\mathrm{trace}(D(D^TD+\gamma I_N)^{-1}D^T)\right).
$$

On the other hand, by simple matrix manipulations, it turns out that

$$
U^T\Psi^{-1/2}H\Psi^{1/2}U=D(D^TD+\gamma I_N)^{-1}D^T.
$$

Finally, recalling that $\mathrm{trace}(AB)=\mathrm{trace}(BA)$,

$$
\mathrm{trace}(U^T\Psi^{-1/2}H\Psi^{1/2}U)=\mathrm{trace}(\Psi^{1/2}UU^T\Psi^{-1/2}H)=\mathrm{trace}(H)
$$

thus proving the thesis.

4.13.7 Proof of Theorem 4.8

Without loss of generality, the proof refers to the diagonalized Bayesian estimation problem (4.48). The marginal loglikelihood function is

$$
\sum_{i=1}^{N}\log(\bar{d}_i^2\lambda+\sigma^2)+\sum_{i=1}^{N}\frac{\bar{y}_i^2}{\bar{d}_i^2\lambda+\sigma^2}+\kappa,
$$

where κ denotes a constant we are not concerned with. By equating to zero the partial derivatives with respect to σ^2 and λ we obtain

$$\sum_{i=1}^{N} \frac{1}{\bar{d}_i^2\lambda + \sigma^2} - \sum_{i=1}^{N} \frac{\bar{y}_i^2}{(\bar{d}_i^2\lambda + \sigma^2)^2} = 0$$
$$\sum_{i=1}^{n} \frac{d_i^2}{d_i^2\lambda + \sigma^2} - \sum_{i=1}^{n} \frac{d_i^2\bar{y}_i^2}{(d_i^2\lambda + \sigma^2)^2} = 0.$$

In view of (4.54) and (4.63),

$$\sigma^2 (N - \mathrm{dof}(\gamma)) - \mathrm{WRSS} = 0$$
$$\lambda \mathrm{dof}(\gamma) - \mathrm{WPSS} = 0,$$

which concludes the proof.

References

1. Aravkin A, Burke JV, Chiuso A, Pillonetto G (2014) Convex vs non-convex estimators for regression and sparse estimation: the mean squared error properties of ARD and GLASSO. J Mach Learn Res 15(1):217–252
2. Buja A, Hastie T, Tibshirani R (1989) Linear smoothers and additive models. Ann Stat 453–510
3. Craven P, Wahba G (1979) Smoothing noisy data with spline functions. Numer Math 31:377–403
4. De Finetti B (2017) Theory of probability: a critical introductory treatment, vol 6. Wiley
5. De Nicolao G, Sparacino G, Cobelli C (1997) Nonparametric input estimation in physiological systems: problems, methods, and case studies. Automatica 33(5):851–870
6. Efron B, Morris C (1973) Stein's estimation rule and its competitors-an empirical Bayes approach. J Am Stat Assoc 68(341):117–130
7. Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6:721–741
8. Girosi F, Jones M, Poggio T (1995) Regularization theory and neural networks architectures. Neural Comput 7(2):219–269
9. Golub GH, Heath M, Wahba G (1979) Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21(2):215–223
10. Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82(4):711–732
11. Hjalmarsson H (2005) From experiment design to closed loop control. Automatica 41(3):393–438
12. Jaynes ET (1982) On the rationale of maximum-entropy methods. Proc IEEE 70(9):939–952
13. Jeffreys H (1946) An invariant form for the prior probability in estimation problems. Proc Math Phys Eng Sci 186(1007):453–461
14. Kass RE, Raftery AE (1995) Bayes factors. J Amer Statist Assoc 90:773–795
15. Kimeldorf G, Wahba G (1970) A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. Ann Math Stat 41(2):495–502
16. Lindley DV (2013) Understanding uncertainty. Wiley
17. MacKay DJC (1992) Bayesian interpolation. Neural Comput 4:415–447

18. Magni P, Bellazzi R, De Nicolao G (1998) Bayesian function learning using MCMC methods. IEEE Trans Pattern Anal Mach Intell 20(12):1319–1331
19. Papoulis A (1984) Probability, random variables and stochastic processes. Mc Graw-Hill
20. Pillonetto G, De Nicolao G (2010) A new kernel-based approach for linear system identification. Automatica 46(1):81–93
21. Pillonetto G, Chiuso A (2015) Tuning complexity in regularized kernel-based regression and linear system identification: the robustness of the marginal likelihood estimator. Automatica 58:106–117
22. Wahba G (1990) Spline models for observational data. SIAM, Philadelphia
23. Zhu H, Rohwer R (1996) Bayesian regression filters and the issue of priors. Neural Comput Appl 4:130–142