4

Statistical inversion theory

The majority of retrieval approaches currently used in atmospheric remote sensing belong to the category of statistical inversion methods (Rodgers, 2000). The goal of this chapter is to reveal the similarity between classical regularization and statistical inversion regarding

(1) the regularized solution representation,
(2) the error analysis,
(3) the design of one- and multi-parameter regularization methods.

In statistical inversion theory all variables included in the model are absolutely continuous random variables and the degree of information concerning their realizations is coded in probability densities. The solution of the inverse problem is the a posteriori density, which makes possible to compute estimates of the unknown atmospheric profile.

In the framework of Tikhonov regularization we have considered the linear data model

$$\mathbf{y}^{\delta} = \mathbf{K}\mathbf{x} + \boldsymbol{\delta}, \tag{4.1}$$

where $\mathbf{y}^{\delta}$ is the noisy data vector and $\boldsymbol{\delta}$ is the noise vector. In statistical inversion theory all parameters are viewed as random variables, and since in statistics random variables are denoted by capital letters and their realizations by lowercase letters, the stochastic version of the data model (4.1) is

$$\mathbf{Y}^{\delta} = \mathbf{K}\mathbf{X} + \boldsymbol{\Delta}. \tag{4.2}$$

The random vectors $\mathbf{Y}^{\delta}$, $\mathbf{X}$ and $\boldsymbol{\Delta}$ represent the data, the state and the noise, respectively; their realizations are denoted by $\mathbf{Y}^{\delta} = \mathbf{y}^{\delta}$, $\mathbf{X} = \mathbf{x}$ and $\boldsymbol{\Delta} = \boldsymbol{\delta}$, respectively.

4.1 Bayes theorem and estimators

The data model (4.2) gives a relation between the three random vectors $\mathbf{Y}^{\delta}$, $\mathbf{X}$ and $\boldsymbol{\Delta}$, and therefore, their probability densities depend on each other. The following probability densities are relevant for our analysis:

(1) the a priori density $p_{\mathrm{a}}(\mathbf{x})$, which encapsulates our presumable information about $\mathbf{X}$ before performing the measurement of $\mathbf{Y}^{\delta}$;
(2) the likelihood density $p\left(\mathbf{y}^{\delta} \mid \mathbf{x}\right)$, which represents the conditional probability density of $\mathbf{Y}^{\delta}$ given the state $\mathbf{X}=\mathbf{x}$;
(3) the a posteriori density $p\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right)$, which represents the conditional probability density of $\mathbf{X}$ given the data $\mathbf{Y}^{\delta}=\mathbf{y}^{\delta}$.

The choice of the a priori density $p_{\mathrm{a}}(\mathbf{x})$ is perhaps the most important part of the inversion process. Different a priori models yield different objective functions, and in particular, the classical regularization terms correspond to Gaussian a priori models. Gaussian densities are widely used in statistical inversion theory because they are easy to compute and often lead to explicit estimators. Besides Gaussian densities other types of a priori models, as for instance the Cauchy density and the entropy density can be found in the literature (Kaipio and Somersalo, 2005).

The construction of the likelihood density $p\left(\mathbf{y}^{\delta} \mid \mathbf{x}\right)$ depends on the noise assumption. The data model (4.2) operates with additive noise, but other explicit noise models including multiplicative noise models and models with an incompletely known forward model matrix can be considered. If the noise is additive and is independent of the atmospheric state, the probability density $p_{\mathrm{n}}(\boldsymbol{\delta})$ of $\boldsymbol{\Delta}$ remains unchanged when conditioned on $\mathbf{X}=\mathbf{x}$. Thus, $\mathbf{Y}^{\delta}$ conditioned on $\mathbf{X}=\mathbf{x}$ is distributed like $\boldsymbol{\Delta}$, and the likelihood density becomes

$$p\left(\mathbf{y}^{\delta} \mid \mathbf{x}\right)=p_{\mathrm{n}}\left(\mathbf{y}^{\delta}-\mathbf{K}\mathbf{x}\right). \tag{4.3}$$

Assuming that after analyzing the measurement setting and accounting of the additional information available about all variables we have found the joint probability density $p\left(\mathbf{x}, \mathbf{y}^{\delta}\right)$ of $\mathbf{X}$ and $\mathbf{Y}^{\delta}$, then the a priori density is given by

$$p_{\mathrm{a}}(\mathbf{x})=\int_{\mathbb{R}^m} p\left(\mathbf{x}, \mathbf{y}^{\delta}\right) \mathrm{d}\mathbf{y}^{\delta},$$

while the likelihood density and the a posteriori density can be expressed as

$$p\left(\mathbf{y}^{\delta} \mid \mathbf{x}\right)=\frac{p\left(\mathbf{x}, \mathbf{y}^{\delta}\right)}{p_{\mathrm{a}}(\mathbf{x})}, \tag{4.4}$$

and

$$p\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right)=\frac{p\left(\mathbf{x}, \mathbf{y}^{\delta}\right)}{p\left(\mathbf{y}^{\delta}\right)}, \tag{4.5}$$

respectively.

The following result known as the Bayes theorem of inverse problems relates the a posteriori density to the likelihood density (cf. (4.4) and (4.5)):

$$p\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right)=\frac{p\left(\mathbf{y}^{\delta} \mid \mathbf{x}\right) p_{\mathrm{a}}(\mathbf{x})}{p\left(\mathbf{y}^{\delta}\right)}. \tag{4.6}$$

In (4.6), the marginal density $p\left(\mathbf{y}^{\delta}\right)$ computed as

$$p\left(\mathbf{y}^{\delta}\right)=\int_{\mathbb{R}^n} p\left(\mathbf{x}, \mathbf{y}^{\delta}\right) \mathrm{d}\mathbf{x}=\int_{\mathbb{R}^n} p\left(\mathbf{y}^{\delta} \mid \mathbf{x}\right) p_{\mathrm{a}}(\mathbf{x}) \mathrm{d}\mathbf{x},$$

plays the role of a normalization constant and is usually ignored. However, as we will see, this probability density is of particular importance in the design of regularization parameter choice methods.

The knowledge of the a posteriori density allows the calculation of different estimators and spreads of solution. A popular statistical estimator is the maximum a posteriori estimator

$$\widehat{\mathbf{x}}_{\mathtt{map}} = \arg\max_{\mathbf{x}} p\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right),$$

and the problem of finding the maximum a posteriori estimator requires the solution of an optimization problem. Another estimator is the conditional mean of $\mathbf{X}$ conditioned on the data $\mathbf{Y}^{\delta} = \mathbf{y}^{\delta}$,

$$\widehat{\mathbf{x}}_{\mathtt{cm}} = \int_{\mathbb{R}^n} \mathbf{x} p\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right) \, \mathrm{d}\mathbf{x}, \qquad (4.7)$$

and the problem of finding the conditional mean estimator requires to solve an integration problem. The maximum likelihood estimator

$$\widehat{\mathbf{x}}_{\mathtt{ml}} = \arg\max_{\mathbf{x}} p\left(\mathbf{y}^{\delta} \mid \mathbf{x}\right)$$

is not a Bayesian estimator but it is perhaps the most popular estimator in statistics. For ill-posed problems, the maximum likelihood estimator corresponds to solving the inverse problem without regularization, and is therefore of little importance for our analysis.

4.2 Gaussian densities

An n-dimensional random vector $\mathbf{X}$ has a (non-degenerate) Gaussian, or normal, distribution, if its probability density has the form

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{C}_{\mathrm{x}})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})^T \mathbf{C}_{\mathrm{x}}^{-1} (\mathbf{x}-\bar{\mathbf{x}})\right).$$

In the above relation,

$$\bar{\mathbf{x}} = \mathcal{E}\{\mathbf{X}\} = \int_{\mathbb{R}^n} \mathbf{x} p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \qquad (4.8)$$

is the mean vector or the expected value of $\mathbf{X}$ and

$$\mathbf{C}_{\mathrm{x}} = \mathcal{E}\left\{(\mathbf{X}-\mathcal{E}\{\mathbf{X}\})(\mathbf{X}-\mathcal{E}\{\mathbf{X}\})^T\right\} = \int_{\mathbb{R}^n} (\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})^T p(\mathbf{x}) \, \mathrm{d}\mathbf{x}$$

is the covariance matrix of $\mathbf{X}$. These parameters characterize the Gaussian density and we indicate this situation by writing $\mathbf{X} \sim \mathtt{N}(\bar{\mathbf{x}}, \mathbf{C}_{\mathrm{x}})$. In this section, we derive Bayesian estimators for Gaussian densities and characterize the solution error following the treatment of Rodgers (2000). We then discuss two measures of the retrieval quality, the degree of freedom for signal and the information content.

4.2.1 Estimators

Under the assumption that $\mathbf{X}$ and $\mathbf{\Delta}$ are independent Gaussian random vectors, characterized by $\mathbf{X} \sim \mathtt{N}\left(\mathbf{0}, \mathbf{C}_{\mathrm{x}}\right)$ and $\mathbf{\Delta} \sim \mathtt{N}\left(\mathbf{0}, \mathbf{C}_{\delta}\right)$, the a priori density can be expressed as

$$p_{\mathtt{a}}\left(\mathbf{x}\right) = \frac{1}{\sqrt{\left(2\pi\right)^{n}\det\left(\mathbf{C}_{\mathrm{x}}\right)}} \exp\left(-\frac{1}{2}\mathbf{x}^{T}\mathbf{C}_{\mathrm{x}}^{-1}\mathbf{x}\right), \tag{4.9}$$

while by virtue of (4.3), the likelihood density takes the form

$$p\left(\mathbf{y}^{\delta} \mid \mathbf{x}\right) = \frac{1}{\sqrt{\left(2\pi\right)^{m}\det\left(\mathbf{C}_{\delta}\right)}} \exp\left(-\frac{1}{2}\left(\mathbf{y}^{\delta} - \mathbf{K}\mathbf{x}\right)^{T}\mathbf{C}_{\delta}^{-1}\left(\mathbf{y}^{\delta} - \mathbf{K}\mathbf{x}\right)\right). \tag{4.10}$$

With this information, the Bayes formula yields the following expression for the a posteriori density:

$$p\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right) \propto \exp\left(-\frac{1}{2}\left(\mathbf{y}^{\delta} - \mathbf{K}\mathbf{x}\right)^{T}\mathbf{C}_{\delta}^{-1}\left(\mathbf{y}^{\delta} - \mathbf{K}\mathbf{x}\right) - \frac{1}{2}\mathbf{x}^{T}\mathbf{C}_{\mathrm{x}}^{-1}\mathbf{x}\right). \tag{4.11}$$

Setting

$$p\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right) \propto \exp\left(-\frac{1}{2}V\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right)\right),$$

where the a posteriori potential $V\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right)$ is defined by

$$V\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right) = \left(\mathbf{y}^{\delta} - \mathbf{K}\mathbf{x}\right)^{T}\mathbf{C}_{\delta}^{-1}\left(\mathbf{y}^{\delta} - \mathbf{K}\mathbf{x}\right) + \mathbf{x}^{T}\mathbf{C}_{\mathrm{x}}^{-1}\mathbf{x},$$

we see that the maximum a posteriori estimator $\widehat{\mathbf{x}}_{\mathtt{map}}$ maximizing the conditional probability density $p\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right)$ also minimizes the potential $V\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right)$, that is,

$$\widehat{\mathbf{x}}_{\mathtt{map}} = \arg\min_{\mathbf{x}} V\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right).$$

The solution to this minimization problem is given by

$$\widehat{\mathbf{x}}_{\mathtt{map}} = \widehat{\mathbf{G}}\mathbf{y}^{\delta}, \tag{4.12}$$

where

$$\widehat{\mathbf{G}} = \left(\mathbf{K}^{T}\mathbf{C}_{\delta}^{-1}\mathbf{K} + \mathbf{C}_{\mathrm{x}}^{-1}\right)^{-1}\mathbf{K}^{T}\mathbf{C}_{\delta}^{-1} \tag{4.13}$$

is known as the gain matrix or the contribution function matrix (Rodgers, 2000). Equation (4.12) reveals that the gain matrix corresponds to the regularized generalized inverse appearing in the framework of Tikhonov regularization. An alternative representation for the gain matrix can be derived from the relation

$$\left(\mathbf{K}^{T}\mathbf{C}_{\delta}^{-1}\mathbf{K} + \mathbf{C}_{\mathrm{x}}^{-1}\right)^{-1}\mathbf{K}^{T}\mathbf{C}_{\delta}^{-1} = \mathbf{C}_{\mathrm{x}}\mathbf{K}^{T}\left(\mathbf{C}_{\delta} + \mathbf{K}\mathbf{C}_{\mathrm{x}}\mathbf{K}^{T}\right)^{-1}, \tag{4.14}$$

and the result is

$$\widehat{\mathbf{G}} = \mathbf{C}_{\mathrm{x}}\mathbf{K}^{T}\left(\mathbf{C}_{\delta} + \mathbf{K}\mathbf{C}_{\mathrm{x}}\mathbf{K}^{T}\right)^{-1}. \tag{4.15}$$

To prove (4.14), we multiply this equation from the left and from the right with the matrices $\mathbf{K}^T\mathbf{C}_\delta^{-1}\mathbf{K}+\mathbf{C}_{\mathrm{x}}^{-1}$ and $\mathbf{C}_\delta+\mathbf{K}\mathbf{C}_{\mathrm{x}}\mathbf{K}^T$, respectively, and use the identity

$$\mathbf{K}^T+\mathbf{K}^T\mathbf{C}_\delta^{-1}\mathbf{K}\mathbf{C}_{\mathrm{x}}\mathbf{K}^T=\left(\mathbf{K}^T\mathbf{C}_\delta^{-1}\mathbf{K}+\mathbf{C}_{\mathrm{x}}^{-1}\right)\mathbf{C}_{\mathrm{x}}\mathbf{K}^T \tag{4.16}$$

to conclude.

The a posteriori density $p\left(\mathbf{x}\mid\mathbf{y}^\delta\right)$ can be expressed as a Gaussian density

$$p\left(\mathbf{x}\mid\mathbf{y}^\delta\right)\propto\exp\left(-\frac{1}{2}\left(\mathbf{x}-\bar{\mathbf{x}}\right)^T\widehat{\mathbf{C}}_{\mathrm{x}}^{-1}\left(\mathbf{x}-\bar{\mathbf{x}}\right)\right), \tag{4.17}$$

where the mean vector $\bar{\mathbf{x}}$ and the covariance matrix $\widehat{\mathbf{C}}_{\mathrm{x}}$ can be obtained directly from (4.11) and (4.17) by equating like terms (see, e.g., Rodgers, 2000). Equating the terms quadratic in $\mathbf{x}$ leads to the following expression for the a posteriori covariance matrix:

$$\widehat{\mathbf{C}}_{\mathrm{x}}=\left(\mathbf{K}^T\mathbf{C}_\delta^{-1}\mathbf{K}+\mathbf{C}_{\mathrm{x}}^{-1}\right)^{-1}.$$

To obtain the expression of the a posteriori mean vector, we equate the terms linear in $\mathbf{x}$ and obtain $\bar{\mathbf{x}}=\widehat{\mathbf{x}}_{\mathtt{map}}$. On the other hand, by (4.7), (4.8) and (4.17), we see that the a posteriori mean coincides with the conditional mean, and we conclude that in the purely Gaussian case there holds

$$\bar{\mathbf{x}}=\widehat{\mathbf{x}}_{\mathtt{map}}=\widehat{\mathbf{x}}_{\mathtt{cm}}.$$

Due to this equivalence and in order to simplify the writing, the maximum a posteriori estimator will be simply denoted by $\widehat{\mathbf{x}}$.

An alternative expression for the a posteriori covariance matrix follows from the identity (cf. (4.14))

$$\begin{aligned}&\mathbf{C}_{\mathrm{x}}-\mathbf{C}_{\mathrm{x}}\mathbf{K}^T\left(\mathbf{C}_\delta+\mathbf{K}\mathbf{C}_{\mathrm{x}}\mathbf{K}^T\right)^{-1}\mathbf{K}\mathbf{C}_{\mathrm{x}}\\&=\mathbf{C}_{\mathrm{x}}-\left(\mathbf{K}^T\mathbf{C}_\delta^{-1}\mathbf{K}+\mathbf{C}_{\mathrm{x}}^{-1}\right)^{-1}\mathbf{K}^T\mathbf{C}_\delta^{-1}\mathbf{K}\mathbf{C}_{\mathrm{x}}\\&=\left(\mathbf{K}^T\mathbf{C}_\delta^{-1}\mathbf{K}+\mathbf{C}_{\mathrm{x}}^{-1}\right)^{-1},\end{aligned} \tag{4.18}$$

which yields (cf. (4.15))

$$\widehat{\mathbf{C}}_{\mathrm{x}}=\mathbf{C}_{\mathrm{x}}-\widehat{\mathbf{G}}\mathbf{K}\mathbf{C}_{\mathrm{x}}=\left(\mathbf{I}_n-\mathbf{A}\right)\mathbf{C}_{\mathrm{x}} \tag{4.19}$$

with $\mathbf{A}=\widehat{\mathbf{G}}\mathbf{K}$ being the averaging kernel matrix.

For Gaussian densities with covariance matrices of the form

$$\mathbf{C}_\delta=\sigma^2\mathbf{I}_m,\ \ \mathbf{C}_{\mathrm{x}}=\sigma_{\mathrm{x}}^2\mathbf{C}_{\mathrm{nx}}=\sigma_{\mathrm{x}}^2\left(\mathbf{L}^T\mathbf{L}\right)^{-1}, \tag{4.20}$$

we find that

$$\widehat{\mathbf{x}}=\left(\mathbf{K}^T\mathbf{K}+\alpha\mathbf{L}^T\mathbf{L}\right)^{-1}\mathbf{K}^T\mathbf{y}^\delta,$$

where we have set

$$\alpha=\frac{\sigma^2}{\sigma_{\mathrm{x}}^2}.$$

As in section 3.2, σ is the white noise standard deviation, $\sigma_{\mathtt{x}}$ is the profile standard deviation, $\mathbf{C}_{\mathtt{nx}}$ is the normalized a priori covariance matrix, and α and $\mathbf{L}$ are the regularization parameter and the regularization matrix, respectively. Thus, under assumptions (4.20), the maximum a posteriori estimator coincides with the Tikhonov solution. The regularization parameter is the ratio of the noise variance to the profile variance in our a priori knowledge, and in an engineering language, α can be interpreted as the noise-to-signal ratio. We can think of our a priori knowledge in terms of ellipsoids of constant probability of the a priori, whose shape and orientation are determined by $\mathbf{C}_{\mathtt{nx}}$ and whose size is determined by $\sigma_{\mathtt{x}}^2$. The number $\sigma_{\mathtt{x}}$ then, represents the a priori confidence we have in our initial guess of the state vector, confidence being measured through the Mahalanobis norm with covariance $\mathbf{C}_{\mathtt{nx}}$. The correspondence between the Bayesian approach and Tikhonov regularization, which has been recognized by several authors, e.g., Golub et al. (1979), O'Sullivan and Wahba (1985), Fitzpatrick (1991), Vogel (2002), Kaipio and Somersalo (2005), allows the construction of natural schemes for estimating $\sigma_{\mathtt{x}}^2$.

4.2.2 Error characterization

In a semi-stochastic setting the total error in the state space has a deterministic component, the smoothing error, and a stochastic component, the noise error. In a stochastic setting, both error components are random vectors. To introduce the random errors, we express the maximum a posteriori estimator as (see (3.65))

$$\widehat{\mathbf{x}} = \widehat{\mathbf{G}}\mathbf{y}^{\delta} = \widehat{\mathbf{G}}\left(\mathbf{K}\mathbf{x}^{\dagger} + \boldsymbol{\delta}\right) = \mathbf{A}\mathbf{x}^{\dagger} + \widehat{\mathbf{G}}\boldsymbol{\delta}.$$

and find that

$$\mathbf{x}^{\dagger} - \widehat{\mathbf{x}} = \left(\mathbf{I}_n - \mathbf{A}\right)\mathbf{x}^{\dagger} - \widehat{\mathbf{G}}\boldsymbol{\delta}. \tag{4.21}$$

In view of (4.21), we define the random total error by

$$\mathbf{E} = \mathbf{X} - \widehat{\mathbf{X}} = \left(\mathbf{I}_n - \mathbf{A}\right)\mathbf{X} - \widehat{\mathbf{G}}\boldsymbol{\Delta}, \tag{4.22}$$

where

$$\widehat{\mathbf{X}} = \widehat{\mathbf{G}}\mathbf{Y}^{\delta}$$

is an estimator of $\mathbf{X}$. In (4.22), $\mathbf{X}$ should be understood as the true state, and a realization of $\mathbf{X}$ is the exact solution of the linear equation in the noise-free case.

The random smoothing error is defined by

$$\mathbf{E}_{\mathtt{s}} = \left(\mathbf{I}_n - \mathbf{A}\right)\mathbf{X},$$

and it is apparent that the statistics of $\mathbf{E}_{\mathtt{s}}$ is determined by the statistics of $\mathbf{X}$. If $\mathcal{E}\{\mathbf{X}\} = \mathbf{0}$ and $\mathbf{C}_{\mathtt{xt}} = \mathcal{E}\{\mathbf{X}\mathbf{X}^T\}$ is the covariance matrix of the true state, then the mean vector and the covariance matrix of $\mathbf{E}_{\mathtt{s}}$ become

$$\mathcal{E}\left\{\mathbf{E}_{\mathtt{s}}\right\} = \mathbf{0}, \;\; \mathbf{C}_{\mathtt{es}} = \left(\mathbf{I}_n - \mathbf{A}\right)\mathbf{C}_{\mathtt{xt}}\left(\mathbf{I}_n - \mathbf{A}\right)^T.$$

In practice, the statistics of the true state is unknown and, as in a semi-stochastic setting, the statistics of the smoothing error is unknown.

The random noise error is defined as

$$\mathbf{E}_{\mathrm{n}} = -\widehat{\mathbf{G}}\boldsymbol{\Delta}$$

and the mean vector and the covariance matrix of $\mathbf{E}_{\mathrm{n}}$ are given by

$$\mathcal{E}\left\{\mathbf{E}_{\mathrm{n}}\right\} = \mathbf{0}, \ \ \mathbf{C}_{\mathtt{en}} = \widehat{\mathbf{G}}\mathbf{C}_{\delta}\widehat{\mathbf{G}}^{T}.$$

As $\mathbf{X}$ and $\boldsymbol{\Delta}$ are independent random vectors, the random total error has zero mean and covariance

$$\mathbf{C}_{\mathtt{e}} = \mathbf{C}_{\mathtt{es}} + \mathbf{C}_{\mathtt{en}}.$$

When computing the maximum a posteriori estimator we use an ad hoc a priori covariance matrix $\mathbf{C}_{\mathtt{x}}$ because the covariance matrix of the true state $\mathbf{C}_{\mathtt{xt}}$ is not available. It should be pointed out, that only for $\mathbf{C}_{\mathtt{x}} = \mathbf{C}_{\mathtt{xt}}$, the total error covariance matrix coincides with the a posteriori covariance matrix. To prove this claim, we construct the total error covariance matrix as

$$\begin{aligned}\mathbf{C}_{\mathtt{e}} &= \left(\mathbf{I}_{n} - \widehat{\mathbf{G}}\mathbf{K}\right)\mathbf{C}_{\mathtt{x}}\left(\mathbf{I}_{n} - \widehat{\mathbf{G}}\mathbf{K}\right)^{T} + \widehat{\mathbf{G}}\mathbf{C}_{\delta}\widehat{\mathbf{G}}^{T} \\ &= \mathbf{C}_{\mathtt{x}} - \mathbf{C}_{\mathtt{x}}\mathbf{K}^{T}\widehat{\mathbf{G}}^{T} - \widehat{\mathbf{G}}\mathbf{K}\mathbf{C}_{\mathtt{x}} + \widehat{\mathbf{G}}\mathbf{K}\mathbf{C}_{\mathtt{x}}\mathbf{K}^{T}\widehat{\mathbf{G}}^{T} + \widehat{\mathbf{G}}\mathbf{C}_{\delta}\widehat{\mathbf{G}}^{T},\end{aligned}$$

and use the result (cf. (4.13) and (4.16))

$$\begin{aligned}\widehat{\mathbf{G}}\mathbf{C}_{\delta} + \widehat{\mathbf{G}}\mathbf{K}\mathbf{C}_{\mathtt{x}}\mathbf{K}^{T} &= \left(\mathbf{K}^{T}\mathbf{C}_{\delta}^{-1}\mathbf{K} + \mathbf{C}_{\mathtt{x}}^{-1}\right)^{-1}\mathbf{K}^{T}\mathbf{C}_{\delta}^{-1}\left(\mathbf{C}_{\delta} + \mathbf{K}\mathbf{C}_{\mathtt{x}}\mathbf{K}^{T}\right) \\ &= \left(\mathbf{K}^{T}\mathbf{C}_{\delta}^{-1}\mathbf{K} + \mathbf{C}_{\mathtt{x}}^{-1}\right)^{-1}\left(\mathbf{K}^{T}\mathbf{C}_{\delta}^{-1}\mathbf{K} + \mathbf{C}_{\mathtt{x}}^{-1}\right)\mathbf{C}_{\mathtt{x}}\mathbf{K}^{T} \\ &= \mathbf{C}_{\mathtt{x}}\mathbf{K}^{T}\end{aligned}$$

to obtain (cf. (4.19))

$$\mathbf{C}_{\mathtt{e}} = \mathbf{C}_{\mathtt{x}} - \widehat{\mathbf{G}}\mathbf{K}\mathbf{C}_{\mathtt{x}} = \widehat{\mathbf{C}}_{\mathtt{x}}.$$

The main conclusion which can be drawn is that an error analysis based on the a posteriori covariance matrix is correct only if the a priori covariance matrix approximates sufficiently well the covariance matrix of the true state.

4.2.3 Degrees of freedom

In classical regularization theory, the expected residual $\mathcal{E}\{\left\|\mathbf{y}^{\delta} - \mathbf{K}\mathbf{x}_{\alpha}^{\delta}\right\|^{2}\}$ and the expected constraint $\mathcal{E}\{\left\|\mathbf{L}\mathbf{x}_{\alpha}^{\delta}\right\|^{2}\}$ are important tools for analyzing discrete ill-posed problems. In statistical inversion theory, the corresponding quantities are the degree of freedom for noise and the degree of freedom for signal.

To introduce these quantities, we consider the expression of the a posteriori potential $V\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right)$ and define the random variable

$$\widehat{V} = \left(\mathbf{Y}^{\delta} - \mathbf{K}\widehat{\mathbf{X}}\right)^{T}\mathbf{C}_{\delta}^{-1}\left(\mathbf{Y}^{\delta} - \mathbf{K}\widehat{\mathbf{X}}\right) + \widehat{\mathbf{X}}^{T}\mathbf{C}_{\mathtt{x}}^{-1}\widehat{\mathbf{X}}, \tag{4.23}$$

where, as before, $\widehat{\mathbf{X}} = \widehat{\mathbf{G}}\mathbf{Y}^{\delta}$. The random variable $\widehat{V}$ is Chi-square distributed with m degrees of freedom, and therefore, the expected value of $\widehat{V}$ is equal to the number of measurements m (Appendix D). This can be divided into the degrees of freedom for signal and noise, defined by

$$d_{\mathrm{s}} = \mathcal{E}\left\{\widehat{\mathbf{X}}^T \mathbf{C}_{\mathrm{x}}^{-1} \widehat{\mathbf{X}}\right\}$$

and

$$d_{\mathrm{n}} = \mathcal{E}\left\{\left(\mathbf{Y}^{\delta} - \mathbf{K}\widehat{\mathbf{X}}\right)^T \mathbf{C}_{\delta}^{-1} \left(\mathbf{Y}^{\delta} - \mathbf{K}\widehat{\mathbf{X}}\right)\right\},$$

respectively, and evidently we have

$$d_{\mathrm{s}} + d_{\mathrm{n}} = m.$$

The degree of freedom for signal measures that part of $\mathcal{E}\{\widehat{V}\}$ corresponding to the state vector, while the degree of freedom for noise that part corresponding to the measurement.

Using the identity

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = \mathrm{trace}\left(\mathbf{x}\mathbf{x}^T \mathbf{A}\right),$$

which holds true for a symmetric matrix $\mathbf{A}$, we express the degree of freedom for signal as

$$d_{\mathrm{s}} = \mathcal{E}\left\{\widehat{\mathbf{X}}^T \mathbf{C}_{\mathrm{x}}^{-1} \widehat{\mathbf{X}}\right\} = \mathcal{E}\left\{\mathrm{trace}\left(\widehat{\mathbf{X}}\widehat{\mathbf{X}}^T \mathbf{C}_{\mathrm{x}}^{-1}\right)\right\} = \mathrm{trace}\left(\mathcal{E}\left\{\widehat{\mathbf{X}}\widehat{\mathbf{X}}^T\right\} \mathbf{C}_{\mathrm{x}}^{-1}\right),$$

where the covariance of the estimator $\widehat{\mathbf{X}}$ is related to the covariance of the data $\mathbf{Y}^{\delta}$ by the relation

$$\mathcal{E}\left\{\widehat{\mathbf{X}}\widehat{\mathbf{X}}^T\right\} = \widehat{\mathbf{G}}\mathcal{E}\left\{\mathbf{Y}^{\delta}\mathbf{Y}^{\delta T}\right\}\widehat{\mathbf{G}}^T.$$

To compute the covariance of the data, we assume that the covariance matrix of the true state is adequately described by the a priori covariance matrix, and obtain

$$\mathcal{E}\left\{\mathbf{Y}^{\delta}\mathbf{Y}^{\delta T}\right\} = \mathbf{K}\mathcal{E}\left\{\mathbf{X}\mathbf{X}^T\right\}\mathbf{K}^T + \mathcal{E}\left\{\boldsymbol{\Delta}\boldsymbol{\Delta}^T\right\} = \mathbf{K}\mathbf{C}_{\mathrm{x}}\mathbf{K}^T + \mathbf{C}_{\delta}. \tag{4.24}$$

By (4.13) and (4.15), we then have

$$\mathcal{E}\left\{\widehat{\mathbf{X}}\widehat{\mathbf{X}}^T\right\} = \mathbf{C}_{\mathrm{x}}\mathbf{K}^T\mathbf{C}_{\delta}^{-1}\mathbf{K}\left(\mathbf{K}^T\mathbf{C}_{\delta}^{-1}\mathbf{K} + \mathbf{C}_{\mathrm{x}}^{-1}\right)^{-1}, \tag{4.25}$$

whence using the identities $\mathrm{trace}\left(\mathbf{B}^{-1}\mathbf{A}\mathbf{B}\right) = \mathrm{trace}\left(\mathbf{A}\right)$ and $\mathrm{trace}\left(\mathbf{A}\right) = \mathrm{trace}\left(\mathbf{A}^T\right)$, which hold true for a square matrix $\mathbf{A}$ and a nonsingular matrix $\mathbf{B}$, we find that

$$\begin{aligned} d_{\mathrm{s}} &= \mathrm{trace}\left(\mathbf{K}^T\mathbf{C}_{\delta}^{-1}\mathbf{K}\left(\mathbf{K}^T\mathbf{C}_{\delta}^{-1}\mathbf{K} + \mathbf{C}_{\mathrm{x}}^{-1}\right)^{-1}\right) \\ &= \mathrm{trace}\left(\left(\mathbf{K}^T\mathbf{C}_{\delta}^{-1}\mathbf{K} + \mathbf{C}_{\mathrm{x}}^{-1}\right)^{-1}\mathbf{K}^T\mathbf{C}_{\delta}^{-1}\mathbf{K}\right) \\ &= \mathrm{trace}\left(\widehat{\mathbf{G}}\mathbf{K}\right) \\ &= \mathrm{trace}\left(\mathbf{A}\right). \end{aligned} \tag{4.26}$$

Hence, the degree of freedom for signal is the trace of the averaging kernel matrix. Consequently, the diagonal of the averaging kernel matrix $\mathbf{A}$ may be thought of as a measure of the number of degrees of freedom per layer (level), and thus as a measure of information, while its reciprocal may be thought of as the number of layers per degree of freedom, and thus as a measure of resolution. The degree of freedom for signal can also be interpreted as a measure of the minimum number of parameters that could be used to define a state vector without loss of information (Mateer, 1965); Rodgers, 2000).

The degree of freedom for noise can be expressed in terms of the influence matrix $\widehat{\mathbf{A}} = \mathbf{K}\widehat{\mathbf{G}}$ as (cf. (4.24))

$$\begin{aligned} d_{\mathrm{n}} &= \mathcal{E}\left\{\left(\mathbf{Y}^{\delta} - \mathbf{K}\widehat{\mathbf{X}}\right)^T \mathbf{C}_{\delta}^{-1}\left(\mathbf{Y}^{\delta} - \mathbf{K}\widehat{\mathbf{X}}\right)\right\} \\ &= \mathcal{E}\left\{\text{trace}\left(\left(\mathbf{Y}^{\delta} - \mathbf{K}\widehat{\mathbf{X}}\right)\left(\mathbf{Y}^{\delta} - \mathbf{K}\widehat{\mathbf{X}}\right)^T \mathbf{C}_{\delta}^{-1}\right)\right\} \\ &= \text{trace}\left(\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)\mathcal{E}\left\{\mathbf{Y}^{\delta}\mathbf{Y}^{\delta T}\right\}\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)^T \mathbf{C}_{\delta}^{-1}\right) \\ &= \text{trace}\left(\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)\left(\mathbf{K}\mathbf{C}_{\mathrm{x}}\mathbf{K}^T + \mathbf{C}_{\delta}\right)\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)^T \mathbf{C}_{\delta}^{-1}\right), \end{aligned} \tag{4.27}$$

whence using the identity

$$\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)\left(\mathbf{K}\mathbf{C}_{\mathrm{x}}\mathbf{K}^T + \mathbf{C}_{\delta}\right) = \mathbf{C}_{\delta}, \tag{4.28}$$

we obtain

$$d_{\mathrm{n}} = \text{trace}\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right). \tag{4.29}$$

Note that the term ‘degree of freedom for noise’ has been used by Craven and Wahba (1979) and later on by Wahba (1985) to designate the denominator of the generalized cross-validation function.

Under assumptions (4.20), we have

$$\text{trace}\left(\mathbf{A}\right) = \text{trace}\left(\widehat{\mathbf{A}}\right) = \sum_{i=1}^{n} \frac{\gamma_i^2}{\gamma_i^2 + \alpha}, \tag{4.30}$$

where γ_i are the generalized singular values of the matrix pair $(\mathbf{K}, \mathbf{L})$. By (4.26), (4.29) and (4.30), it is apparent that the degree of freedom for signal is a decreasing function of the regularization parameter, while the degree of freedom for noise is an increasing function of the regularization parameter. Thus, when very little regularization is introduced, the degree of freedom for signal is very large and approaches n, and when a large amount of regularization is introduced, the degree of freedom for noise is very large and approaches m. As in classical regularization theory, an optimal regularization parameter should balance the degrees of freedom for signal and noise.

The degree of freedom for signal can be expressed in terms of the so-called information matrix $\mathbf{R}$ defined by

$$\mathbf{R} = \mathbf{C}_{\mathrm{x}}^{\frac{1}{2}}\mathbf{K}^T\mathbf{C}_{\delta}^{-1}\mathbf{K}\mathbf{C}_{\mathrm{x}}^{\frac{1}{2}}. \tag{4.31}$$

Using the identity

$$\mathbf{A} = \mathbf{C}_{\mathrm{x}}^{\frac{1}{2}}\left(\mathbf{I}_n + \mathbf{R}\right)^{-1}\mathbf{R}\mathbf{C}_{\mathrm{x}}^{-\frac{1}{2}}, \tag{4.32}$$

we find that

$$d_{\mathrm{s}} = \mathrm{trace}\left(\mathbf{A}\right) = \mathrm{trace}\left(\left(\mathbf{I}_n + \mathbf{R}\right)^{-1}\mathbf{R}\right), \tag{4.33}$$

whence assuming the singular value decomposition of the positive definite matrix $\mathbf{R}$,

$$\mathbf{R} = \mathbf{V}_{\mathrm{r}}\Sigma_{\mathrm{r}}\mathbf{V}_{\mathrm{r}}^T, \;\; \Sigma_{\mathrm{r}} = \left[\mathrm{diag}\left(\omega_i\right)_{n\times n}\right], \tag{4.34}$$

we obtain the representation

$$d_{\mathrm{s}} = \sum_{i=1}^{n} \frac{\omega_i}{\omega_i + 1}.$$

The degree of freedom for signal d_{s} remains unchanged under linear transformations of the state vector or of the data vector, and as a result, d_{s} is an invariant of the retrieval. Purser and Huang (1993) showed that the degree of freedom for signal, regarded as a real-valued function over sets of independent data, obeys a positive monotonic subadditive algebra. In order to understand these properties from a practical point of view, we consider a set of m_1 data $\mathbf{Y}_1^\delta = \mathbf{y}_1^\delta$, and an independent set of m_2 data $\mathbf{Y}_2^\delta = \mathbf{y}_2^\delta$. For the ith set of measurements, the data model is

$$\mathbf{Y}_i^\delta = \mathbf{K}_i\mathbf{X} + \mathbf{\Delta}_i, \;\; i = 1, 2,$$

and the maximum a posteriori estimator is computed as

$$\widehat{\mathbf{x}}_i = \arg\min_{\mathbf{x}}\left(\left(\mathbf{y}_i^\delta - \mathbf{K}_i\mathbf{x}\right)^T\mathbf{C}_{\delta i}^{-1}\left(\mathbf{y}_i^\delta - \mathbf{K}_i\mathbf{x}\right) + \mathbf{x}^T\mathbf{C}_{\mathrm{x}}^{-1}\mathbf{x}\right).$$

The corresponding information matrix and the degree of freedom for signal are given by

$$\mathbf{R}_i = \mathbf{C}_{\mathrm{x}}^{\frac{1}{2}}\mathbf{K}_i^T\mathbf{C}_{\delta i}^{-1}\mathbf{K}_i\mathbf{C}_{\mathrm{x}}^{\frac{1}{2}}$$

and

$$d_{\mathrm{s}i} = \mathrm{trace}\left(\left(\mathbf{I}_n + \mathbf{R}_i\right)^{-1}\mathbf{R}_i\right),$$

respectively. For the full set of $m_{12} = m_1 + m_2$ measurements, we consider the data model

$$\begin{bmatrix} \mathbf{Y}_1^\delta \\ \mathbf{Y}_2^\delta \end{bmatrix} = \begin{bmatrix} \mathbf{K}_1 \\ \mathbf{K}_2 \end{bmatrix}\mathbf{X} + \begin{bmatrix} \mathbf{\Delta}_1 \\ \mathbf{\Delta}_2 \end{bmatrix},$$

and compute the maximum a posteriori estimator as

$$\begin{aligned}\widehat{\mathbf{x}}_{12} = &\arg\min_{\mathbf{x}}\left(\left(\mathbf{y}_1^\delta - \mathbf{K}_1\mathbf{x}\right)^T\mathbf{C}_{\delta 1}^{-1}\left(\mathbf{y}_1^\delta - \mathbf{K}_1\mathbf{x}\right)\right.\\ &\left. + \left(\mathbf{y}_2^\delta - \mathbf{K}_2\mathbf{x}\right)^T\mathbf{C}_{\delta 2}^{-1}\left(\mathbf{y}_2^\delta - \mathbf{K}_2\mathbf{x}\right) + \mathbf{x}^T\mathbf{C}_{\mathrm{x}}^{-1}\mathbf{x}\right).\end{aligned}$$

When the data are treated jointly, the information matrix and the degree of freedom for signal are given by

$$\mathbf{R}_{12} = \mathbf{C}_{\mathrm{x}}^{\frac{1}{2}}\left(\mathbf{K}_1^T\mathbf{C}_{\delta 1}^{-1}\mathbf{K}_1 + \mathbf{K}_2^T\mathbf{C}_{\delta 2}^{-1}\mathbf{K}_2\right)\mathbf{C}_{\mathrm{x}}^{\frac{1}{2}} = \mathbf{R}_1 + \mathbf{R}_2$$

and

$$d_{\mathrm{s}12} = \mathrm{trace}\left(\left(\mathbf{I}_n + \mathbf{R}_1 + \mathbf{R}_2\right)^{-1}\left(\mathbf{R}_1 + \mathbf{R}_2\right)\right),$$

respectively. In this context, the monotonicity of the degree of freedom for signal means that $d_{\mathrm{s}12}$ of the full m_{12} measurements is never less than either $d_{\mathrm{s}1}$ or $d_{\mathrm{s}2}$, i.e.,

$$d_{\mathrm{s}12} \geq \max\left(d_{\mathrm{s}1}, d_{\mathrm{s}2}\right), \tag{4.35}$$

while the subadditivity means that $d_{\mathrm{s}12}$ can never exceed $d_{\mathrm{s}1} + d_{\mathrm{s}2}$, i.e.,

$$d_{\mathrm{s}12} \leq d_{\mathrm{s}1} + d_{\mathrm{s}2}. \tag{4.36}$$

These assertions are the result of the following theorem: considering a monotonic, strictly increasing, and strictly concave function $f(x)$ with $f(0) = 0$, and defining the associated scalar function F of $\mathbf{R} \in \mathcal{S}_n$ by

$$F(\mathbf{R}) = \sum_{i=1}^{n} f(\omega_i),$$

where $\mathcal{S}_n$ is the set of all semi-positive definite matrices of order n, and ω_i are the singular values of $\mathbf{R}$, we have

$$\mathbf{R}_2 \geq \mathbf{R}_1 \Rightarrow F(\mathbf{R}_2) \geq F(\mathbf{R}_1) \quad \text{(monotonicity)}, \tag{4.37}$$

and

$$F(\mathbf{R}_1) + F(\mathbf{R}_2) \geq F(\mathbf{R}_1 + \mathbf{R}_2) \quad \text{(subadditivity)}, \tag{4.38}$$

for all $\mathbf{R}_1, \mathbf{R}_2 \in \mathcal{S}_n$. Here, we write $\mathbf{R}_2 \geq \mathbf{R}_1$ if $\mathbf{R}_2 - \mathbf{R}_1 \in \mathcal{S}_n$. Since the degree of freedom for signal d_{s} can be expressed in terms of the information matrix $\mathbf{R}$ as a scalar function $F(\mathbf{R})$ with $f(x) = x/(1+x)$, (4.37) and (4.38) yield (4.35) and (4.36), respectively. A rigorous proof of this theorem has been given by Purser and Huang (1993) by taking into account that $F(\mathbf{R})$ is invariant to orthogonal transformations. However, (4.35) and (4.36) can simply be justified when

$$m_1 = m_2 = m, \quad \mathbf{K}_1 = \mathbf{K}_2, \quad \mathbf{C}_{\delta 1} = \mathbf{C}_{\delta 2}. \tag{4.39}$$

In this case, we obtain

$$\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{R}, \quad \mathbf{R}_{12} = 2\mathbf{R},$$

and further,

$$d_{\mathrm{s}1} = d_{\mathrm{s}2} = \sum_{i=1}^{n} \frac{\omega_i}{\omega_i + 1}, \quad d_{\mathrm{s}12} = \sum_{i=1}^{n} \frac{2\omega_i}{2\omega_i + 1}.$$

Then, from

$$\frac{2\omega_i}{2\omega_i + 1} \geq \frac{\omega_i}{\omega_i + 1}, \quad \frac{2\omega_i}{2\omega_i + 1} \leq \frac{2\omega_i}{\omega_i + 1}, \quad i = 1, \ldots, n,$$

the conclusions are apparent. The deficit $m_{12} - d_{\mathrm{s}12}$ may be interpreted as the internal redundancy of the set of data, while the deficit $d_{\mathrm{s}1} + d_{\mathrm{s}2} - d_{\mathrm{s}12}$ may be thought as the mutual redundancy between two pooled sets.

Another statistics of a linear retrieval is the effective data density. Whereas the degree of freedom for signal is a measure that indicates the number of independent pieces of information, the effective data density is a measure that indicates the density of effectively independent pieces of information. The data density at the ith layer of thickness $\triangle z_i$ is defined by

$$\rho_i = \frac{[\mathbf{A}]_{ii}}{\triangle z_i}, \tag{4.40}$$

and it is apparent that the 'integral' of the effective data density is the degree of freedom for signal,

$$d_{\mathrm{s}} = \sum_{i=1}^{n} \rho_i \triangle z_i .$$

This estimate together with the degree of freedom for signal can be used to interprete the quality of the retrieval and the effectiveness of the measurements.

4.2.4 Information content

An alternative criterion for the quality of a measurement is the information content or the incremental gain in information. The information content is defined in terms of the change in entropy that expresses a priori and a posteriori knowledge of the atmospheric state. This measure of performance has been proposed in the context of retrieval by Peckham (1974) and has also been discussed by Rodgers (1976) and Eyre (1990).

In information theory, the Shannon entropy or the absolute entropy is a measure of uncertainty associated with a random variable. The Shannon entropy of a discrete random vector $\mathbf{X}$, which can take the values $\mathbf{x}_1, \ldots, \mathbf{x}_n$, is defined by

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i , \tag{4.41}$$

where the probability mass function of $\mathbf{X}$ is given by

$$p(\mathbf{x}) = \begin{cases} p_i, & \mathbf{X} = \mathbf{x}_i, \\ 0, & \text{otherwise}, \end{cases} \quad \sum_{i=1}^{n} p_i = 1.$$

H is positive and attains its global maximum $H_{\max} = \log n$ for a uniform distribution, i.e., when all p_i are the same. On the other hand, the lowest entropy level, $H_{\min} = 0$, is attained when all probabilities p_i but one are zero. Shannon (1949) showed that $H(p)$ defined by (4.41) satisfies the following desiderata:

(1) H is continuous in $(p_1, \ldots, p_n)$ (continuity);
(2) H remains unchanged if the outcomes $\mathbf{x}_i$ are re-ordered (symmetry);
(3) if all the outcomes are equally likely, then H is maximal (maximum);
(4) the amount of entropy is the same independently of how the process is regarded as being divided into parts (additivity).

These properties guarantee that the Shannon entropy is well-behaved with regard to relative information comparisons. For a continuous density $p(\mathbf{x})$, the following entropy formula also satisfies the properties enumerated above:

$$H(p) = -\int_{\mathbb{R}^n} p(\mathbf{x}) \log p(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{4.42}$$

For a Gaussian random vector with covariance matrix $\mathbf{C}$, the integral in (4.42) can be analytically computed and the result is

$$H(p) = \frac{n}{2}\log(2\pi\mathrm{e}) + \frac{1}{2}\log(\det(\mathbf{C})).$$

As the a priori density $p_{\mathtt{a}}(\mathbf{x})$ describes knowledge before a measurement and the a posteriori density $p\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right)$ describes it afterwards, the information content of the measurement $\triangle H$ is the reduction in entropy (e.g., Rodgers, 2000)

$$\triangle H = H(p_{\mathtt{a}}(\mathbf{x})) - H\left(p\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right)\right).$$

For Gaussian densities with the a priori and the a posteriori covariance matrices $\mathbf{C}_{\mathtt{x}}$ and $\widehat{\mathbf{C}}_{\mathtt{x}}$, respectively, the information content then becomes

$$\triangle H = -\frac{1}{2}\log\left(\det\left(\widehat{\mathbf{C}}_{\mathtt{x}}\mathbf{C}_{\mathtt{x}}^{-1}\right)\right) = -\frac{1}{2}\log(\det(\mathbf{I}_n - \mathbf{A})).$$

By virtue of (4.32), which relates the information matrix $\mathbf{R}$ and the averaging kernel matrix $\mathbf{A}$, we obtain the representation

$$\triangle H = \frac{1}{2}\log(\det(\mathbf{I}_n + \mathbf{R})),$$

and further

$$\triangle H = \frac{1}{2}\sum_{i=1}^{n}\log(1 + \omega_i).$$

Similar to the degree of freedom for signal, the information content obeys a positive monotonic subadditive algebra (Huang and Purser, 1996). By 'monotonic' we mean that the addition of independent data does not decrease (on average) the information content, while by 'subadditive' we mean that any two sets of data treated jointly never yield more of the information content than the sum of the amounts yielded by the sets treated singly. These results follow from (4.37) and (4.38) by taking into account that the information content $\triangle H$ can be expressed in terms of the information matrix $\mathbf{R}$ as a scalar function $F(\mathbf{R})$ with $f(x) = (1/2)\log(1 + x)$, or, in the simple case (4.39), they follow from the obvious inequalities

$$\log(1 + 2\omega_i) \geq \log(1 + \omega_i),\ \ \log(1 + 2\omega_i) \leq 2\log(1 + \omega_i),\ \ i = 1, \ldots, n.$$

A density of information can be defined by employing the technique which has been used to define the effective data density. For this purpose, we seek an equivalent matrix $\mathbf{A}_{\mathtt{h}}$, whose trace is the information content $\triangle H$, so that the diagonal elements of this matrix can be used as in (4.40) to define the density of information at each layer,

$$\rho_{\mathtt{h}i} = \frac{[\mathbf{A}_{\mathtt{h}}]_{ii}}{\triangle z_i}.$$

The matrix $\mathbf{A}_{\mathtt{h}}$ is chosen as

$$\mathbf{A}_{\mathtt{h}} = \mathbf{V}_{\mathtt{r}}\Sigma_{\mathtt{ah}}\mathbf{V}_{\mathtt{r}}^{T},$$

where $\mathbf{V}_{\mathtt{r}}$ is the orthogonal matrix in (4.34) and

$$\Sigma_{\mathtt{ah}} = \left[\operatorname{diag}\left(\frac{1}{2} \log\left(1+\omega_i\right) \right)_{n\times n} \right].$$

The information content is used as a selection criterion in the framework of the so-called information operator method. Assuming (4.20) and considering a generalized singular value decomposition of the matrix pair $(\mathbf{K}, \mathbf{L})$, the maximum a posteriori estimator and the information content of the measurement can be expressed as

$$\widehat{\mathbf{x}}_{\mathtt{map}} = \sum_{i=1}^{n} f_\alpha\left(\gamma_i^2\right) \frac{1}{\sigma_i} \left(\mathbf{u}_i^T \mathbf{y}^\delta\right) \mathbf{w}_i,$$

and

$$\triangle H = \frac{1}{2} \sum_{i=1}^{n} \log\left(1 + \frac{\gamma_i^2}{\alpha}\right),$$

respectively, where

$$f_\alpha\left(\gamma_i^2\right) = \frac{\gamma_i^2}{\gamma_i^2+\alpha}, \quad i = 1, \ldots, n,$$

are the filter factors for Tikhonov regularization and $\alpha = \sigma^2/\sigma_{\mathtt{x}}^2$. In the information operator method, only the generalized singular values γ_i larger than $\sqrt{\alpha}$ are considered to give a relevant contribution to the information content. Note that α should not be regarded as a regularization parameter whose value should be optimized; rather α is completely determined by the profile variance $\sigma_{\mathtt{x}}^2$ which we take to be fixed. The state space spanned by the singular vectors associated with the relevant singular values gives the effective state space accessible with the measurement (Kozlov, 1983; Rozanov, 2001). If p is the largest index i so that

$$\gamma_i^2 \geq \alpha = \frac{\sigma^2}{\sigma_{\mathtt{x}}^2}, \quad i = 1, \ldots, p,$$

then the information operator solution can be expressed as

$$\widehat{\mathbf{x}}_{\mathtt{io}} = \sum_{i=1}^{p} f_\alpha\left(\gamma_i^2\right) \frac{1}{\sigma_i} \left(\mathbf{u}_i^T \mathbf{y}^\delta\right) \mathbf{w}_i.$$

Essentially, the filter factors of the information operator method are given by

$$f_\alpha\left(\gamma_i^2\right) = \begin{cases} \gamma_i^2, & \gamma_i^2 \geq \alpha, \\ 0, & \gamma_i^2 < \alpha, \end{cases}$$

and we see that the information operator method has sharper filter factors than Tikhonov regularization.

4.3 Regularization parameter choice methods

Under assumptions (4.20), the Bayesian approach is equivalent to Tikhonov regularization in the sense that the maximum a posteriori estimator simultaneously minimizes the potential

$$V\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right) = \frac{1}{\sigma^2}\left\|\mathbf{y}^{\delta} - \mathbf{K}\mathbf{x}\right\|^2 + \frac{1}{\sigma_{\mathtt{x}}^2}\left\|\mathbf{L}\mathbf{x}\right\|^2,$$

and the Tikhonov function

$$\mathcal{F}_{\alpha}\left(\mathbf{x}\right) = \sigma^2 V\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right) = \left\|\mathbf{y}^{\delta} - \mathbf{K}\mathbf{x}\right\|^2 + \alpha\left\|\mathbf{L}\mathbf{x}\right\|^2, \ \ \alpha = \frac{\sigma^2}{\sigma_{\mathtt{x}}^2}.$$

When the profile variance $\sigma_{\mathtt{x}}^2$ is unknown, it seems to be justified to ask for a reliable estimator $\widehat{\sigma}_{\mathtt{x}}^2$ of $\sigma_{\mathtt{x}}^2$, or equivalently, for a plausible estimator $\widehat{\alpha}$ of α. For this reason, in statistical inversion theory, a regularization parameter choice method can be regarded as an approach for estimating $\sigma_{\mathtt{x}}^2$.

4.3.1 Expected error estimation method

In a semi-stochastic setting, the expected error estimation method has been formulated in the following way: given the exact profile $\mathbf{x}^{\dagger}$, compute the optimal regularization parameter $\overline{\alpha}_{\mathtt{opt}}$ as the minimizer of the expected error $\mathcal{E}\{\left\|\mathbf{x}^{\dagger} - \mathbf{x}_{\alpha}^{\delta}\right\|^2\}$, with $\mathbf{x}_{\alpha}^{\delta}$ being the Tikhonov solution of regularization parameter α. In statistical inversion theory, an equivalent formulation may read as follows: given the covariance matrix of the true state $\mathbf{C}_{\mathtt{xt}}$, compute the profile variance $\sigma_{\mathtt{x}}^2$ as the minimizer of the expected error

$$\mathcal{E}\left\{\left\|\mathbf{E}\right\|^2\right\} = \mathrm{trace}\left(\left(\mathbf{I}_n - \mathbf{A}\right)\mathbf{C}_{\mathtt{xt}}\left(\mathbf{I}_n - \mathbf{A}\right)^T\right) + \sigma^2\,\mathrm{trace}\left(\widehat{\mathbf{G}}\widehat{\mathbf{G}}^T\right), \tag{4.43}$$

where the a priori covariance matrix in the expressions of $\widehat{\mathbf{G}}$ and $\mathbf{A}$ is given by $\mathbf{C}_{\mathtt{x}} = \sigma_{\mathtt{x}}^2\mathbf{C}_{\mathtt{nx}}$. If the covariance matrix of the true state is expressed as $\mathbf{C}_{\mathtt{xt}} = \sigma_{\mathtt{xt}}^2\mathbf{C}_{\mathtt{nx}}$, then the minimization of the expected error (4.43), yields $\sigma_{\mathtt{x}} = \sigma_{\mathtt{xt}}$. To prove this result under assumptions (4.20), we take $\mathbf{L} = \mathbf{I}_n$, and obtain

$$E\left(\alpha\right) = \mathcal{E}\left\{\left\|\mathbf{E}\right\|^2\right\} = \sigma_{\mathtt{xt}}^2\sum_{i=1}^{n}\left[\left(\frac{\alpha}{\sigma_i^2 + \alpha}\right)^2 + \alpha_{\mathtt{t}}\left(\frac{\sigma_i}{\sigma_i^2 + \alpha}\right)^2\right], \ \ \alpha_{\mathtt{t}} = \frac{\sigma^2}{\sigma_{\mathtt{xt}}^2}.$$

Setting $E'\left(\alpha\right) = 0$ gives

$$\sum_{i=1}^{n}\left[\frac{\alpha\sigma_i^2}{\left(\sigma_i^2 + \alpha\right)^3} - \frac{\alpha_{\mathtt{t}}\sigma_i^2}{\left(\sigma_i^2 + \alpha\right)^3}\right] = 0,$$

which further implies that $\alpha = \alpha_{\mathtt{t}}$, or equivalently that $\sigma_{\mathtt{x}} = \sigma_{\mathtt{xt}}$. Thus, the maximum a posteriori estimator is given by $\widehat{\mathbf{x}}_{\mathtt{map}} = \widehat{\mathbf{G}}\mathbf{y}^{\delta}$ with

$$\widehat{\mathbf{G}} = \left(\mathbf{K}^T\mathbf{C}_{\delta}^{-1}\mathbf{K} + \mathbf{C}_{\mathtt{xt}}^{-1}\right)^{-1}\mathbf{K}^T\mathbf{C}_{\delta}^{-1}. \tag{4.44}$$

The selection rule based on the minimization of (4.43) simply states that if the covariance matrix of the true state is known, then this information should be used to construct the a priori density.

In statistical inversion theory, the minimization of the expected error is not formulated in terms of the profile variance (or the regularization parameter), but rather in terms of the inverse matrix $\mathbf{G}$. The resulting method, which is known as the minimum variance method, possesses the following formulation: if the statistics of the true state is known,

$$\mathcal{E}\left\{\mathbf{X}\right\} = \mathbf{0},\ \ \mathcal{E}\left\{\mathbf{X}\mathbf{X}^T\right\} = \mathbf{C}_{\mathtt{xt}}, \tag{4.45}$$

then for the affine estimation rule $\widehat{\mathbf{x}} = \mathbf{G}\mathbf{y}^{\delta}$, the matrix $\widehat{\mathbf{G}}$ minimizing the expected error

$$\widehat{\mathbf{G}} = \arg\min_{\mathbf{G}} \mathcal{E}\left\{\left\|\mathbf{X} - \mathbf{G}\mathbf{Y}^{\delta}\right\|^2\right\} \tag{4.46}$$

is given by (4.44), and the minimum variance estimator $\widehat{\mathbf{x}}_{\mathtt{mv}} = \widehat{\mathbf{G}}\mathbf{y}^{\delta}$ coincides with the maximum a posteriori estimator $\widehat{\mathbf{x}}_{\mathtt{map}}$. To justify this claim, we look at the behavior of the expected error when $\mathbf{G}$ is replaced by a candidate solution $\mathbf{G} + \mathbf{H}$. Using the result

$$\begin{aligned}
&\left\|\mathbf{X} - \left(\mathbf{G} + \mathbf{H}\right)\mathbf{Y}^{\delta}\right\|^2 \\
&= \left\|\mathbf{X} - \mathbf{G}\mathbf{Y}^{\delta}\right\|^2 - 2\left(\mathbf{X} - \mathbf{G}\mathbf{Y}^{\delta}\right)^T \mathbf{H}\mathbf{Y}^{\delta} + \left\|\mathbf{H}\mathbf{Y}^{\delta}\right\|^2 \\
&= \left\|\mathbf{X} - \mathbf{G}\mathbf{Y}^{\delta}\right\|^2 - 2\,\mathrm{trace}\left(\mathbf{H}\mathbf{Y}^{\delta}\left(\mathbf{X} - \mathbf{G}\mathbf{Y}^{\delta}\right)^T\right) + \left\|\mathbf{H}\mathbf{Y}^{\delta}\right\|^2,
\end{aligned}$$

we obtain

$$\begin{aligned}
&\mathcal{E}\left\{\left\|\mathbf{X} - \left(\mathbf{G} + \mathbf{H}\right)\mathbf{Y}^{\delta}\right\|^2\right\} \\
&= \mathcal{E}\left\{\left\|\mathbf{X} - \mathbf{G}\mathbf{Y}^{\delta}\right\|^2\right\} - 2\,\mathrm{trace}\left(\mathbf{H}\,\mathcal{E}\left\{\mathbf{Y}^{\delta}\left(\mathbf{X} - \mathbf{G}\mathbf{Y}^{\delta}\right)^T\right\}\right) + \mathcal{E}\left\{\left\|\mathbf{H}\mathbf{Y}^{\delta}\right\|^2\right\}.
\end{aligned}$$

The trace term vanishes for the choice

$$\mathbf{G} = \widehat{\mathbf{G}} = \mathcal{E}\left\{\mathbf{X}\mathbf{Y}^{\delta T}\right\}\left(\mathcal{E}\left\{\mathbf{Y}^{\delta}\mathbf{Y}^{\delta T}\right\}\right)^{-1}, \tag{4.47}$$

since

$$\mathcal{E}\left\{\mathbf{Y}^{\delta}\left(\mathbf{X} - \widehat{\mathbf{G}}\mathbf{Y}^{\delta}\right)^T\right\} = \mathcal{E}\left\{\mathbf{Y}^{\delta}\mathbf{X}^T\right\} - \mathcal{E}\left\{\mathbf{Y}^{\delta}\mathbf{Y}^{\delta T}\right\}\widehat{\mathbf{G}}^T = \mathbf{0}.$$

Under assumptions (4.45), we find that

$$\mathcal{E}\left\{\mathbf{X}\mathbf{Y}^{\delta T}\right\} = \mathcal{E}\left\{\mathbf{X}\mathbf{X}^T\right\}\mathbf{K}^T = \mathbf{C}_{\mathtt{xt}}\mathbf{K}^T,$$

whence using (4.24), (4.47) becomes

$$\widehat{\mathbf{G}} = \mathbf{C}_{\mathtt{xt}}\mathbf{K}^T\left(\mathbf{K}\mathbf{C}_{\mathtt{xt}}\mathbf{K}^T + \mathbf{C}_{\delta}\right)^{-1} = \left(\mathbf{K}^T\mathbf{C}_{\delta}^{-1}\mathbf{K} + \mathbf{C}_{\mathtt{xt}}^{-1}\right)^{-1}\mathbf{K}^T\mathbf{C}_{\delta}^{-1}.$$

Hence, we have

$$\mathcal{E}\left\{\left\|\mathbf{X}-\left(\widehat{\mathbf{G}}+\mathbf{H}\right)\mathbf{Y}^{\delta}\right\|^{2}\right\}=\mathcal{E}\left\{\left\|\mathbf{X}-\widehat{\mathbf{G}}\mathbf{Y}^{\delta}\right\|^{2}\right\}+\mathcal{E}\left\{\left\|\mathbf{H}\mathbf{Y}^{\delta}\right\|^{2}\right\}$$
$$\geq\mathcal{E}\left\{\left\|\mathbf{X}-\widehat{\mathbf{G}}\mathbf{Y}^{\delta}\right\|^{2}\right\}$$

for any $\mathbf{H}\in\mathbb{R}^{n\times m}$, and therefore, $\mathcal{E}\{\|\mathbf{X}-\mathbf{G}\mathbf{Y}^{\delta}\|^{2}\}$ is minimal for $\mathbf{G}=\widehat{\mathbf{G}}$.

The minimum variance estimator minimizes the expected error, which represents the trace of the a posteriori covariance matrix. Instead of minimizing the trace of the a posteriori covariance matrix we may formulate a minimization problem involving the entire a posteriori covariance matrix. For this purpose, we define the random total error

$$\mathbf{E}=\mathbf{X}-\mathbf{G}\mathbf{Y}^{\delta}=\left(\mathbf{I}_{n}-\mathbf{G}\mathbf{K}\right)\mathbf{X}-\mathbf{G}\boldsymbol{\Delta},$$

for some $\mathbf{G}\in\mathbb{R}^{n\times m}$. The covariance matrices of the smoothing and noise errors $\mathbf{E}_{\mathrm{s}}=\left(\mathbf{I}_{n}-\mathbf{G}\mathbf{K}\right)\mathbf{X}$ and $\mathbf{E}_{\mathrm{n}}=-\mathbf{G}\boldsymbol{\Delta}$, can be expressed in terms of the matrix $\mathbf{G}$, as

$$\mathbf{C}_{\mathtt{es}}=\left(\mathbf{I}_{n}-\mathbf{G}\mathbf{K}\right)\mathbf{C}_{\mathtt{xt}}\left(\mathbf{I}_{n}-\mathbf{G}\mathbf{K}\right)^{T}$$

and

$$\mathbf{C}_{\mathtt{en}}=\mathbf{G}\mathbf{C}_{\delta}\mathbf{G}^{T},$$

respectively. Then, it is readily seen that the minimizer of the error covariance matrix

$$\widehat{\mathbf{G}}=\arg\min_{\mathbf{G}}\left(\mathbf{C}_{\mathtt{es}}+\mathbf{C}_{\mathtt{en}}\right), \tag{4.48}$$

solves the equation

$$\frac{\partial}{\partial\mathbf{G}}\left(\mathbf{C}_{\mathtt{xt}}-\mathbf{C}_{\mathtt{xt}}\mathbf{K}^{T}\mathbf{G}^{T}-\mathbf{G}\mathbf{K}\mathbf{C}_{\mathtt{xt}}+\mathbf{G}\mathbf{K}\mathbf{C}_{\mathtt{xt}}\mathbf{K}^{T}\mathbf{G}^{T}+\mathbf{G}\mathbf{C}_{\delta}\mathbf{G}^{T}\right)=\mathbf{0} \tag{4.49}$$

and is given by (4.44).

Because in statistical inversion theory, the conventional expected error estimation method is not beneficial, we design a regularization parameter choice method by looking only at the expected value of the noise error. Under assumptions (4.20), the noise error covariance matrix is given by (cf. (3.38))

$$\mathbf{C}_{\mathtt{en}}=\sigma^{2}\widehat{\mathbf{G}}^{T}\widehat{\mathbf{G}}=\sigma^{2}\mathbf{W}\Sigma_{\mathrm{n}\alpha}\mathbf{W}^{T},$$

with

$$\Sigma_{\mathrm{n}\alpha}=\left[\operatorname{diag}\left(\left(\frac{\gamma_{i}^{2}}{\gamma_{i}^{2}+\alpha}\frac{1}{\sigma_{i}}\right)^{2}\right)_{n\times n}\right],$$

and the expected value of the noise error (cf. (3.41)),

$$\mathcal{E}\left\{\left\|\mathbf{E}_{\mathrm{n}}\right\|^{2}\right\}=\operatorname{trace}\left(\mathbf{C}_{\mathtt{en}}\right)=\sigma^{2}\sum_{i=1}^{n}\left(\frac{\gamma_{i}^{2}}{\gamma_{i}^{2}+\alpha}\frac{1}{\sigma_{i}}\right)^{2}\left\|\mathbf{w}_{i}\right\|^{2},$$

is a decreasing function of α. To improve the degree of freedom for signal we need to chose a small value of the regularization parameter. But when the regularization parameter is too small, the noise error may explode. Therefore, we select the smallest regularization parameter so that the expected value of the noise error is below a specific level. Recalling that $\mathbf{x}$ is the deviation of the retrieved profile from the a priori profile $\mathbf{x}_{\mathtt{a}}$, we define the regularization parameter for noise error estimation $\widehat{\alpha}_{\mathtt{ne}}$ as the solution of the equation

$$\mathcal{E}\left\{\left\|\mathbf{E}_{\mathtt{n}}\right\|^{2}\right\} = \varepsilon_{\mathtt{n}}\left\|\mathbf{x}_{\mathtt{a}}\right\|^{2},$$

for some relative error level $\varepsilon_{\mathtt{n}}$. In atmospheric remote sensing, the expected noise error estimation method has been successfully applied for ozone retrieval from nadir sounding spectra measured by the Tropospheric Emission Spectrometer (TES) on the NASA Aura platform (Steck, 2002).

4.3.2 Discrepancy principle

In a semi-stochastic setting, the discrepancy principle selects the regularization parameter as the solution of the equation

$$\left\|\mathbf{r}_{\alpha}^{\delta}\right\|^{2} = \tau m \sigma^{2}. \tag{4.50}$$

Under assumptions (4.20), equation (4.50) reads as

$$\sum_{i=1}^{m}\left(\frac{\alpha}{\gamma_i^2+\alpha}\right)^{2}\left(\mathbf{u}_i^T\mathbf{y}^{\delta}\right)^{2} = \tau m \sigma^{2}, \tag{4.51}$$

with the convention $\gamma_i = 0$ for $i = n+1,\ldots,m$.

The regularization parameter choice method (4.50) with $\tau = 1$ is known as the constrained least squares method (Hunt, 1973; Trussel, 1983; Trussel and Civanlar, 1984). It has been observed and reported by a number of researchers, e.g., Demoment (1989), that the constrained least squares method yields an oversmooth solution. To ameliorate this problem, Wahba (1983), and Hall and Titterington (1987) proposed, in analogy to regression, the equivalent degree of freedom method. In a stochastic setting, this method takes into account that the expected value of the residual is equal to the trace of the matrix $\mathbf{I}_m - \widehat{\mathbf{A}}$, that is, (cf. (4.27) and (4.29))

$$\mathcal{E}\left\{\left(\mathbf{Y}^{\delta}-\mathbf{K}\widehat{\mathbf{X}}\right)^{T}\mathbf{C}_{\delta}^{-1}\left(\mathbf{Y}^{\delta}-\mathbf{K}\widehat{\mathbf{X}}\right)\right\} = \operatorname{trace}\left(\mathbf{I}_m-\widehat{\mathbf{A}}\right).$$

The resulting equation for computing the regularization parameter is then given by

$$\left(\mathbf{y}^{\delta}-\mathbf{K}\widehat{\mathbf{x}}\right)^{T}\mathbf{C}_{\delta}^{-1}\left(\mathbf{y}^{\delta}-\mathbf{K}\widehat{\mathbf{x}}\right) = \operatorname{trace}\left(\mathbf{I}_m-\widehat{\mathbf{A}}\right),$$

or equivalently, by

$$\sum_{i=1}^{m}\left(\frac{\alpha}{\gamma_i^2+\alpha}\right)^{2}\left(\mathbf{u}_i^T\mathbf{y}^{\delta}\right)^{2} = \sigma^{2}\sum_{i=1}^{m}\frac{\alpha}{\gamma_i^2+\alpha}.$$

On the other hand, the random variable $\hat{V}$, defined by (4.23), is Chi-square distributed with m degrees of freedom. In this regard, we may choose the regularization parameter as the solution of the equation

$$\left(\mathbf{y}^{\delta} - \mathbf{K}\widehat{\mathbf{x}}\right)^{T} \mathbf{C}_{\delta}^{-1} \left(\mathbf{y}^{\delta} - \mathbf{K}\widehat{\mathbf{x}}\right) + \widehat{\mathbf{x}}^{T}\mathbf{C}_{\mathrm{x}}^{-1}\widehat{\mathbf{x}} = m,$$

that is,

$$\sum_{i=1}^{m} \left(\frac{\alpha}{\gamma_i^2 + \alpha}\right) \left(\mathbf{u}_i^T \mathbf{y}^{\delta}\right)^2 = m\sigma^2.$$

As compared to (4.51), the factors multiplying the Fourier coefficients $\mathbf{u}_i^T\mathbf{y}^{\delta}$ converge more slowly to zero as α tends to zero, and therefore, this selection rule yields a larger regularization parameter than the discrepancy principle with $\tau = 1$.

4.3.3 Hierarchical models

In the Bayesian framework, all unknown parameters of the model are included in the retrieval and this applies also for parameters describing the a priori density. The resulting model is then known as hierarchical or hyperpriori model (Kaipio and Somersalo, 2005).

For the a priori covariance matrix $\mathbf{C}_{\mathrm{x}} = \sigma_{\mathrm{x}}^2\mathbf{C}_{\mathrm{nx}}$, we suppose that the a priori density is conditioned on the knowledge of σ_{x}, i.e.,

$$p_{\mathrm{a}}\left(\mathbf{x} \mid \sigma_{\mathrm{x}}\right) = \frac{1}{\sqrt{\left(2\pi\sigma_{\mathrm{x}}^2\right)^n \det\left(\mathbf{C}_{\mathrm{nx}}\right)}} \exp\left(-\frac{1}{2\sigma_{\mathrm{x}}^2}\mathbf{x}^T\mathbf{C}_{\mathrm{nx}}^{-1}\mathbf{x}\right). \tag{4.52}$$

For the parameter σ_{x}, we assume the Gaussian density

$$p_{\mathrm{a}}\left(\sigma_{\mathrm{x}}\right) = \frac{1}{\sqrt{2\pi}\triangle\sigma_{\mathrm{x}}} \exp\left(-\frac{1}{2\triangle\sigma_{\mathrm{x}}^2}\left(\sigma_{\mathrm{x}} - \bar{\sigma}_{\mathrm{x}}\right)^2\right),$$

where the mean $\bar{\sigma}_{\mathrm{x}}$ and the variance $\triangle\sigma_{\mathrm{x}}^2$ are considered to be known. The joint probability density of $\mathbf{X}$ and σ_{x} is then given by

$$\begin{aligned} p_{\mathrm{a}}\left(\mathbf{x}, \sigma_{\mathrm{x}}\right) &= p_{\mathrm{a}}\left(\mathbf{x} \mid \sigma_{\mathrm{x}}\right) p_{\mathrm{a}}\left(\sigma_{\mathrm{x}}\right) \\ &\propto \frac{1}{\left(\sigma_{\mathrm{x}}^2\right)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma_{\mathrm{x}}^2}\mathbf{x}^T\mathbf{C}_{\mathrm{nx}}^{-1}\mathbf{x} - \frac{1}{2\triangle\sigma_{\mathrm{x}}^2}\left(\sigma_{\mathrm{x}} - \bar{\sigma}_{\mathrm{x}}\right)^2\right), \end{aligned}$$

the Bayes formula conditioned on the data $\mathbf{Y}^{\delta} = \mathbf{y}^{\delta}$ takes the form

$$\begin{aligned} p\left(\mathbf{x}, \sigma_{\mathrm{x}} \mid \mathbf{y}^{\delta}\right) \propto & \frac{1}{\left(\sigma_{\mathrm{x}}^2\right)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\left(\mathbf{y}^{\delta} - \mathbf{Kx}\right)^T \mathbf{C}_{\delta}^{-1}\left(\mathbf{y}^{\delta} - \mathbf{Kx}\right)\right. \\ & \left. -\frac{1}{2\sigma_{\mathrm{x}}^2}\mathbf{x}^T\mathbf{C}_{\mathrm{nx}}^{-1}\mathbf{x} - \frac{1}{2\triangle\sigma_{\mathrm{x}}^2}\left(\sigma_{\mathrm{x}} - \bar{\sigma}_{\mathrm{x}}\right)^2\right), \end{aligned}$$

and the maximum a posteriori estimators $\widehat{\mathbf{x}}$ and $\widehat{\sigma}_{\mathrm{x}}$ are found by minimizing the a posteriori potential

$$\begin{aligned} V\left(\mathbf{x}, \sigma_{\mathrm{x}} \mid \mathbf{y}^{\delta}\right) = & \left(\mathbf{y}^{\delta} - \mathbf{Kx}\right)^T \mathbf{C}_{\delta}^{-1}\left(\mathbf{y}^{\delta} - \mathbf{Kx}\right) \\ & + \frac{1}{\sigma_{\mathrm{x}}^2}\mathbf{x}^T\mathbf{C}_{\mathrm{nx}}^{-1}\mathbf{x} + \frac{1}{\triangle\sigma_{\mathrm{x}}^2}\left(\sigma_{\mathrm{x}} - \bar{\sigma}_{\mathrm{x}}\right)^2 + n\log\sigma_{\mathrm{x}}^2. \end{aligned}$$

4.3.4 Maximum likelihood estimation

In the Bayes theorem

$$p\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right)=\frac{p\left(\mathbf{y}^{\delta} \mid \mathbf{x}\right) p_{\mathrm{a}}(\mathbf{x})}{p\left(\mathbf{y}^{\delta}\right)}, \tag{4.53}$$

the denominator $p\left(\mathbf{y}^{\delta}\right)$ gives the probability that the data $\mathbf{Y}^{\delta}=\mathbf{y}^{\delta}$ is observed. The marginal density $p\left(\mathbf{y}^{\delta}\right)$ is obtained by integrating the joint probability density $p\left(\mathbf{x}, \mathbf{y}^{\delta}\right)$ with respect to $\mathbf{x}$, that is,

$$p\left(\mathbf{y}^{\delta}\right)=\int_{\mathbb{R}^{n}} p\left(\mathbf{x}, \mathbf{y}^{\delta}\right) \mathrm{d} \mathbf{x}=\int_{\mathbb{R}^{n}} p\left(\mathbf{y}^{\delta} \mid \mathbf{x}\right) p_{\mathrm{a}}(\mathbf{x}) \mathrm{d} \mathbf{x}. \tag{4.54}$$

By (4.53) and (4.54), we see that $p\left(\mathbf{x} \mid \mathbf{y}^{\delta}\right)$ integrates to 1 as all legitimate probability densities should and that the marginal density $p\left(\mathbf{y}^{\delta}\right)$ is nothing more than a normalization constant. Despite of this fact, $p\left(\mathbf{y}^{\delta}\right)$ plays an important role in the design of regularization parameter choice methods and in particular, of the maximum likelihood estimation.

Assuming that the likelihood density $p\left(\mathbf{y}^{\delta} \mid \mathbf{x}\right)$ and the a priori density $p_{\mathrm{a}}(\mathbf{x})$ depend on additional parameters, which can be cast in the form of a parameter vector $\boldsymbol{\theta}$, we express the marginal density $p\left(\mathbf{y}^{\delta} ; \boldsymbol{\theta}\right)$ as

$$p\left(\mathbf{y}^{\delta} ; \boldsymbol{\theta}\right)=\int_{\mathbb{R}^{n}} p\left(\mathbf{y}^{\delta} \mid \mathbf{x} ; \boldsymbol{\theta}\right) p_{\mathrm{a}}(\mathbf{x} ; \boldsymbol{\theta}) \mathrm{d} \mathbf{x}. \tag{4.55}$$

The marginal density $p\left(\mathbf{y}^{\delta} ; \boldsymbol{\theta}\right)$ is also known as the marginal likelihood function and the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ is defined by

$$\widehat{\boldsymbol{\theta}}=\arg \max _{\boldsymbol{\theta}} \log p\left(\mathbf{y}^{\delta} ; \boldsymbol{\theta}\right).$$

Let us derive the maximum likelihood estimator for Gaussian densities with the covariance matrices (4.20) when σ^{2} and $\alpha=\sigma^{2} / \sigma_{\mathrm{x}}^{2}$ are unknown, that is, when $\boldsymbol{\theta}$ is of the form $\boldsymbol{\theta}=\left[\theta_{1}, \theta_{2}\right]^{T}$ with $\theta_{1}=\sigma^{2}$ and $\theta_{2}=\alpha$. The a priori density $p_{\mathrm{a}}\left(\mathbf{x} ; \sigma^{2}, \alpha\right)$ and the conditional probability density $p\left(\mathbf{y}^{\delta} \mid \mathbf{x} ; \sigma^{2}\right)$ are given by (cf. (4.9) and (4.10))

$$p_{\mathrm{a}}\left(\mathbf{x} ; \sigma^{2}, \alpha\right)=\frac{1}{\sqrt{\left(2 \pi \sigma^{2}\right)^{n} \operatorname{det}\left(\left(\alpha \mathbf{L}^{T} \mathbf{L}\right)^{-1}\right)}} \exp \left(-\frac{\alpha}{2 \sigma^{2}}\|\mathbf{L} \mathbf{x}\|^{2}\right)$$

and

$$p\left(\mathbf{y}^{\delta} \mid \mathbf{x} ; \sigma^{2}\right)=\frac{1}{\sqrt{\left(2 \pi \sigma^{2}\right)^{m}}} \exp \left(-\frac{1}{2 \sigma^{2}}\left\|\mathbf{y}^{\delta}-\mathbf{K} \mathbf{x}\right\|^{2}\right), \tag{4.56}$$

respectively. Taking into account that

$$\left\|\mathbf{y}^{\delta}-\mathbf{K} \mathbf{x}\right\|^{2}+\alpha\|\mathbf{L} \mathbf{x}\|^{2}=(\mathbf{x}-\widehat{\mathbf{x}})^{T}\left(\mathbf{K}^{T} \mathbf{K}+\alpha \mathbf{L}^{T} \mathbf{L}\right)(\mathbf{x}-\widehat{\mathbf{x}})+\mathbf{y}^{\delta T}\left(\mathbf{I}_{m}-\widehat{\mathbf{A}}\right) \mathbf{y}^{\delta},$$

where $\widehat{\mathbf{x}} = \widehat{\mathbf{G}}\mathbf{y}^{\delta}$ and $\widehat{\mathbf{A}} = \mathbf{K}\widehat{\mathbf{G}}$, we express the integrand in (4.55) as

$$\begin{aligned}
&p\left(\mathbf{y}^{\delta} \mid \mathbf{x}; \sigma^2\right) p_{\mathsf{a}}\left(\mathbf{x}; \sigma^2, \alpha\right) \\
&= \frac{1}{\sqrt{\left(2\pi\sigma^2\right)^{n+m} \det\left(\left(\alpha\mathbf{L}^T\mathbf{L}\right)^{-1}\right)}} \exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{x}-\widehat{\mathbf{x}}\right)^T \left(\mathbf{K}^T\mathbf{K}+\alpha\mathbf{L}^T\mathbf{L}\right)\left(\mathbf{x}-\widehat{\mathbf{x}}\right)\right) \\
&\quad \times \exp\left(-\frac{1}{2\sigma^2}\mathbf{y}^{\delta T}\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)\mathbf{y}^{\delta}\right).
\end{aligned}$$

Using the normalization condition

$$\begin{aligned}
&\int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\left(\mathbf{x}-\widehat{\mathbf{x}}\right)^T \left[\sigma^2\left(\mathbf{K}^T\mathbf{K}+\alpha\mathbf{L}^T\mathbf{L}\right)^{-1}\right]^{-1}\left(\mathbf{x}-\widehat{\mathbf{x}}\right)\right) d\mathbf{x} \\
&= \sqrt{\left(2\pi\sigma^2\right)^{n} \det\left(\left(\mathbf{K}^T\mathbf{K}+\alpha\mathbf{L}^T\mathbf{L}\right)^{-1}\right)}
\end{aligned}$$

we obtain

$$\begin{aligned}
p\left(\mathbf{y}^{\delta}; \sigma^2, \alpha\right) &= \int_{\mathbb{R}^n} p\left(\mathbf{y}^{\delta} \mid \mathbf{x}; \sigma^2\right) p_{\mathsf{a}}\left(\mathbf{x}; \sigma^2, \alpha\right) d\mathbf{x} \\
&= \sqrt{\frac{\det\left(\left(\mathbf{K}^T\mathbf{K}+\alpha\mathbf{L}^T\mathbf{L}\right)^{-1}\right)}{\left(2\pi\sigma^2\right)^{m} \det\left(\left(\alpha\mathbf{L}^T\mathbf{L}\right)^{-1}\right)}} \exp\left(-\frac{1}{2\sigma^2}\mathbf{y}^{\delta T}\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)\mathbf{y}^{\delta}\right).
\end{aligned}$$

Taking the logarithm and using the identity

$$\frac{\det\left(\left(\mathbf{K}^T\mathbf{K}+\alpha\mathbf{L}^T\mathbf{L}\right)^{-1}\right)}{\det\left(\left(\alpha\mathbf{L}^T\mathbf{L}\right)^{-1}\right)} = \det\left(\left(\mathbf{K}^T\mathbf{K}+\alpha\mathbf{L}^T\mathbf{L}\right)^{-1}\alpha\mathbf{L}^T\mathbf{L}\right) = \det\left(\mathbf{I}_n - \mathbf{A}\right),$$

yields

$$\begin{aligned}
&\log p\left(\mathbf{y}^{\delta}; \sigma^2, \alpha\right) \\
&= -\frac{m}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2}\log\left(\det\left(\mathbf{I}_n - \mathbf{A}\right)\right) - \frac{1}{2\sigma^2}\mathbf{y}^{\delta T}\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)\mathbf{y}^{\delta}. \qquad (4.57)
\end{aligned}$$

Computing the derivative of (4.57) with respect to σ^2 and setting it equal to zero gives

$$\widehat{\sigma}^2 = \frac{1}{m}\mathbf{y}^{\delta T}\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)\mathbf{y}^{\delta}. \qquad (4.58)$$

Substituting (4.58) back into (4.57), and using the result

$$\det\left(\mathbf{I}_n - \mathbf{A}\right) = \det\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right) = \prod_{i=1}^{n}\frac{\alpha}{\gamma_i^2+\alpha},$$

we find that

$$\log p\left(\mathbf{y}^{\delta} \mid \widehat{\sigma}^2, \alpha\right) = -\frac{m}{2}\left[\log\left(\mathbf{y}^{\delta T}\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)\mathbf{y}^{\delta}\right) - \frac{1}{m}\log\left(\det\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)\right)\right] + c,$$

where c does not depend on α. Thus, the regularization parameter $\hat{\alpha}_{\texttt{mle}}$ which maximizes the log of the marginal likelihood function also minimizes the maximum likelihood function

$$\lambda_{\alpha}^{\delta} = \frac{\mathbf{y}^{\delta T}\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)\mathbf{y}^{\delta}}{\sqrt[m]{\det\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)}},$$

and we indicate this situation by writing

$$\widehat{\alpha}_{\texttt{mle}} = \arg\min_{\alpha} \lambda_{\alpha}^{\delta}.$$

The numerical simulations performed in the preceding chapter have shown that the maximum likelihood estimation is superior to the generalized cross-validation method in the sense that the minimum of the objective function is not very flat and the estimated regularization parameter is closer to the optimum.

4.3.5 Expectation minimization

The Expectation Minimization (EM) algorithm is an alternative to the maximum likelihood estimation in which the negative of the log of the marginal likelihood function is minimized by an iterative approach. The formulation of the expected minimization as a regularization parameter choice method has been provided by Fitzpatrick (1991), while a very general development can be found in Dempster et al. (1977), and McLachlan and Krishnan (1997). In this section we present a version of the EM algorithm by following the analysis of Vogel (2002).

Taking into account that the a posteriori density $p\left(\mathbf{x} \mid \mathbf{y}^{\delta}; \boldsymbol{\theta}\right)$ is normalized,

$$\int_{\mathbb{R}^n} p\left(\mathbf{x} \mid \mathbf{y}^{\delta}; \boldsymbol{\theta}\right)\, \mathrm{d}\mathbf{x} = 1, \tag{4.59}$$

and representing the joint probability density $p\left(\mathbf{x}, \mathbf{y}^{\delta}; \boldsymbol{\theta}\right)$ as

$$p\left(\mathbf{x}, \mathbf{y}^{\delta}; \boldsymbol{\theta}\right) = p\left(\mathbf{x} \mid \mathbf{y}^{\delta}; \boldsymbol{\theta}\right) p\left(\mathbf{y}^{\delta}; \boldsymbol{\theta}\right),$$

we see that for any fixed $\boldsymbol{\theta}_0$, the negative of the log of the marginal likelihood function can be expressed as

$$\begin{aligned} -\log p\left(\mathbf{y}^{\delta}; \boldsymbol{\theta}\right) &= -\log p\left(\mathbf{y}^{\delta}; \boldsymbol{\theta}\right) \int_{\mathbb{R}^n} p\left(\mathbf{x} \mid \mathbf{y}^{\delta}; \boldsymbol{\theta}_0\right)\, \mathrm{d}\mathbf{x} \\ &= -\int_{\mathbb{R}^n} p\left(\mathbf{x} \mid \mathbf{y}^{\delta}; \boldsymbol{\theta}_0\right)\, \log p\left(\mathbf{y}^{\delta}; \boldsymbol{\theta}\right)\, \mathrm{d}\mathbf{x} \\ &= -\int_{\mathbb{R}^n} p\left(\mathbf{x} \mid \mathbf{y}^{\delta}; \boldsymbol{\theta}_0\right)\, \log\left(\frac{p\left(\mathbf{x}, \mathbf{y}^{\delta}; \boldsymbol{\theta}\right)}{p\left(\mathbf{x} \mid \mathbf{y}^{\delta}; \boldsymbol{\theta}\right)}\right) \mathrm{d}\mathbf{x} \\ &= Q\left(\mathbf{y}^{\delta}, \boldsymbol{\theta}, \boldsymbol{\theta}_0\right) - H\left(\mathbf{y}^{\delta}, \boldsymbol{\theta}, \boldsymbol{\theta}_0\right) \end{aligned}$$

with

$$Q\left(\mathbf{y}^{\delta},\boldsymbol{\theta},\boldsymbol{\theta}_0\right)=-\int_{\mathbb{R}^n}p\left(\mathbf{x}\mid\mathbf{y}^{\delta};\boldsymbol{\theta}_0\right)\,\log p\left(\mathbf{x},\mathbf{y}^{\delta};\boldsymbol{\theta}\right)\,\mathrm{d}\mathbf{x}$$

and

$$H\left(\mathbf{y}^{\delta},\boldsymbol{\theta},\boldsymbol{\theta}_0\right)=-\int_{\mathbb{R}^n}p\left(\mathbf{x}\mid\mathbf{y}^{\delta};\boldsymbol{\theta}_0\right)\,\log p\left(\mathbf{x}\mid\mathbf{y}^{\delta};\boldsymbol{\theta}\right)\,\mathrm{d}\mathbf{x}.$$

To evaluate the difference

$$H\left(\mathbf{y}^{\delta},\boldsymbol{\theta},\boldsymbol{\theta}_0\right)-H\left(\mathbf{y}^{\delta},\boldsymbol{\theta}_0,\boldsymbol{\theta}_0\right)=-\int_{\mathbb{R}^n}p\left(\mathbf{x}\mid\mathbf{y}^{\delta};\boldsymbol{\theta}_0\right)\,\log\left(\frac{p\left(\mathbf{x}\mid\mathbf{y}^{\delta};\boldsymbol{\theta}\right)}{p\left(\mathbf{x}\mid\mathbf{y}^{\delta};\boldsymbol{\theta}_0\right)}\right)\,\mathrm{d}\mathbf{x},$$

we use the Jensen inequality

$$\int\varphi\left(g\left(\mathbf{x}\right)\right)f\left(\mathbf{x}\right)\,\mathrm{d}\mathbf{x}\geq\varphi\left(\int g\left(\mathbf{x}\right)f\left(\mathbf{x}\right)\,\mathrm{d}\mathbf{x}\right)$$

for the convex function $\varphi\left(u\right)=-\log u$, that is,

$$-\int_{\mathbb{R}^n}p\left(\mathbf{x}\mid\mathbf{y}^{\delta};\boldsymbol{\theta}_0\right)\,\log\left(\frac{p\left(\mathbf{x}\mid\mathbf{y}^{\delta};\boldsymbol{\theta}\right)}{p\left(\mathbf{x}\mid\mathbf{y}^{\delta};\boldsymbol{\theta}_0\right)}\right)\,\mathrm{d}\mathbf{x}\geq-\log\left(\int_{\mathbb{R}^n}p\left(\mathbf{x}\mid\mathbf{y}^{\delta};\boldsymbol{\theta}\right)\,\mathrm{d}\mathbf{x}\right)=0,$$

and obtain

$$-H\left(\mathbf{y}^{\delta},\boldsymbol{\theta},\boldsymbol{\theta}_0\right)\leq-H\left(\mathbf{y}^{\delta},\boldsymbol{\theta}_0,\boldsymbol{\theta}_0\right).$$

Assuming that $\boldsymbol{\theta}$ is such that

$$Q\left(\mathbf{y}^{\delta},\boldsymbol{\theta},\boldsymbol{\theta}_0\right)\leq Q\left(\mathbf{y}^{\delta},\boldsymbol{\theta}_0,\boldsymbol{\theta}_0\right),$$

it follows that

$$-\log p\left(\mathbf{y}^{\delta};\boldsymbol{\theta}\right)\leq-\log p\left(\mathbf{y}^{\delta};\boldsymbol{\theta}_0\right).$$

The EM algorithm seeks to minimize $-\log p\left(\mathbf{y}^{\delta};\boldsymbol{\theta}\right)$ by iteratively applying the following two steps:

(1) *Expectation step*. Calculate the function $Q\left(\mathbf{y}^{\delta},\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_k\right)$ for the a posteriori density under the current estimator $\widehat{\boldsymbol{\theta}}_k$,

$$Q\left(\mathbf{y}^{\delta},\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_k\right)=-\int_{\mathbb{R}^n}p\left(\mathbf{x}\mid\mathbf{y}^{\delta};\widehat{\boldsymbol{\theta}}_k\right)\,\log\left(p\left(\mathbf{y}^{\delta}\mid\mathbf{x};\boldsymbol{\theta}\right)p_{\mathrm{a}}\left(\mathbf{x};\boldsymbol{\theta}\right)\right)\,\mathrm{d}\mathbf{x}.$$

(2) *Minimization step*. Find the parameter vector $\widehat{\boldsymbol{\theta}}_{k+1}$ which minimizes this function, that is,

$$\widehat{\boldsymbol{\theta}}_{k+1}=\arg\min_{\boldsymbol{\theta}}Q\left(\mathbf{y}^{\delta},\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_k\right).$$

Two main peculiarities of the EM algorithm can be evidenced:

(1) Even if the algorithm has a stable point, there is no guarantee that this stable point is a global minimum of $-\log p\left(\mathbf{y}^{\delta};\boldsymbol{\theta}\right)$, or even a local minimum. If the function $Q\left(\mathbf{y}^{\delta},\boldsymbol{\theta},\boldsymbol{\theta}'\right)$ is continuous, convergence to a stationary point of $-\log p\left(\mathbf{y}^{\delta};\boldsymbol{\theta}\right)$ is guaranteed.

(2) The solution generally depends on the initialization.

To illustrate how the EM algorithm works, we consider Gaussian densities with the covariance matrices (4.20), and choose the parameter vector $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$ with $\theta_1 = \sigma_{\mathrm{x}}^2$ and $\theta_2 = \sigma^2$. The a priori density $p_{\mathrm{a}}\left(\mathbf{x}; \sigma_{\mathrm{x}}^2\right)$ and the conditional probability density $p\left(\mathbf{y}^\delta \mid \mathbf{x}; \sigma^2\right)$ are given by (4.52) and (4.56), respectively. Using the results

$$\frac{\partial}{\partial \sigma_{\mathrm{x}}^2} \log\left(p\left(\mathbf{y}^\delta \mid \mathbf{x}; \sigma^2\right) p_{\mathrm{a}}\left(\mathbf{x}; \sigma_{\mathrm{x}}^2\right)\right) = -\frac{n}{2\sigma_{\mathrm{x}}^2} + \frac{1}{2\sigma_{\mathrm{x}}^4}\mathbf{x}^T \mathbf{C}_{\mathrm{nx}}^{-1}\mathbf{x},$$
$$\frac{\partial}{\partial \sigma^2} \log\left(p\left(\mathbf{y}^\delta \mid \mathbf{x}; \sigma^2\right) p_{\mathrm{a}}\left(\mathbf{x}; \sigma_{\mathrm{x}}^2\right)\right) = -\frac{m}{2\sigma^2} + \frac{1}{2\sigma^4}\left\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\right\|^2,$$

we deduce that the EM iteration step yields the recurrence relations

$$\widehat{\sigma}_{\mathrm{x}k+1}^2 = \frac{1}{n}\int_{\mathbb{R}^n} \mathbf{x}^T \mathbf{C}_{\mathrm{nx}}^{-1}\mathbf{x}\, p\left(\mathbf{x} \mid \mathbf{y}^\delta; \widehat{\sigma}_{\mathrm{x}k}^2, \widehat{\sigma}_k^2\right) \mathrm{d}\mathbf{x}, \tag{4.60}$$
$$\widehat{\sigma}_{k+1}^2 = \frac{1}{m}\int_{\mathbb{R}^n} \left\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\right\|^2 p\left(\mathbf{x} \mid \mathbf{y}^\delta; \widehat{\sigma}_{\mathrm{x}k}^2, \widehat{\sigma}_k^2\right) \mathrm{d}\mathbf{x}. \tag{4.61}$$

To compute the n-dimensional integrals in (4.60) and (4.61) we may use the Monte Carlo method (Tarantola, 2005). As the a posteriori density under the current estimator is Gaussian, the integration process involves the following steps:

(1) for $\widehat{\sigma}_{\mathrm{x}k}^2$ and $\widehat{\sigma}_k^2$, compute the maximum a posteriori estimator $\widehat{\mathbf{x}}_k$ and the a posteriori covariance matrix $\widehat{\mathbf{C}}_{\mathrm{x}k}$;
(2) generate a random sample $\{\mathbf{x}_{ki}\}_{i=\overline{1,N}}$ of a Gaussian distribution with mean vector $\widehat{\mathbf{x}}_k$ and covariance matrix $\widehat{\mathbf{C}}_{\mathrm{x}k}$;
(3) estimate the integrals as

$$\int_{\mathbb{R}^n} \mathbf{x}^T \mathbf{C}_{\mathrm{nx}}^{-1}\mathbf{x}\, p\left(\mathbf{x} \mid \mathbf{y}^\delta; \widehat{\sigma}_{\mathrm{x}k}^2, \widehat{\sigma}_k^2\right) \mathrm{d}\mathbf{x} \approx \frac{1}{N}\sum_{i=1}^N \mathbf{x}_{ki}^T \mathbf{C}_{\mathrm{nx}}^{-1}\mathbf{x}_{ki},$$
$$\int_{\mathbb{R}^n} \left\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\right\|^2 p\left(\mathbf{x} \mid \mathbf{y}^\delta; \widehat{\sigma}_{\mathrm{x}k}^2, \widehat{\sigma}_k^2\right) \mathrm{d}\mathbf{x} \approx \frac{1}{N}\sum_{i=1}^N \left\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{ki}\right\|^2.$$

This integration process is quite demanding, and as a result, the method may become very time-consuming.

4.3.6 A general regularization parameter choice method

In this section we present a general technique for constructing regularization parameter choice methods in statistical inversion theory. Our analysis follows the treatment of Neumaier (1998) and enables us to introduce the generalized cross-validation method and the maximum likelihood estimation in a natural way.

Assuming Gaussian densities with the covariance matrices (4.20) and considering a generalized singular value decomposition of the matrix pair $(\mathbf{K}, \mathbf{L})$, i.e., $\mathbf{K} = \mathbf{U}\Sigma_1\mathbf{W}^{-1}$ and $\mathbf{L} = \mathbf{V}\Sigma_2\mathbf{W}^{-1}$, we express the covariance matrix of the data $\mathbf{Y}^\delta$ as (cf. (4.24))

$$\mathcal{E}\left\{\mathbf{Y}^\delta \mathbf{Y}^{\delta T}\right\} = \mathbf{K}\mathbf{C}_{\mathrm{x}}\mathbf{K}^T + \mathbf{C}_\delta = \sigma_{\mathrm{x}}^2 \mathbf{K}\left(\mathbf{L}^T\mathbf{L}\right)^{-1}\mathbf{K}^T + \sigma^2 \mathbf{I}_m = \mathbf{U}\Sigma_{\mathrm{y}}\mathbf{U}^T,$$

where

$$\begin{aligned}\Sigma_{\mathrm{y}} &= \sigma_{\mathrm{x}}^2 \Sigma_1 \left(\Sigma_2^T \Sigma_2\right)^{-1} \Sigma_1^T + \sigma^2 \mathbf{I}_m \\ &= \begin{bmatrix} \operatorname{diag}\left(\sigma_{\mathrm{x}}^2 \gamma_i^2 + \sigma^2\right)_{n \times n} & \mathbf{0} \\ \mathbf{0} & \operatorname{diag}\left(\sigma^2\right)_{(m-n)\times(m-n)} \end{bmatrix}.\end{aligned}$$

Next, we define the scaled data

$$\bar{\mathbf{Y}}^\delta = \mathbf{U}^T \mathbf{Y}^\delta,$$

and observe that $\bar{\mathbf{Y}}^\delta$ has a diagonal covariance matrix, which is given by

$$\mathcal{E}\left\{\bar{\mathbf{Y}}^\delta \bar{\mathbf{Y}}^{\delta T}\right\} = \mathcal{E}\left\{\mathbf{U}^T \mathbf{Y}^\delta \mathbf{Y}^{\delta T} \mathbf{U}\right\} = \Sigma_{\mathrm{y}}.$$

If σ_{x} and σ correctly describe the covariance matrix of the true state and the instrumental noise covariance matrix, respectively, we must have

$$\mathcal{E}\left\{\bar{Y}_i^{\delta 2}\right\} = \sigma_{\mathrm{x}}^2 \gamma_i^2 + \sigma^2, \ \ i = 1, \ldots, m, \tag{4.62}$$

where $\bar{Y}_i^\delta = \mathbf{u}_i^T \mathbf{Y}^\delta$ for $i = 1, \ldots, m$, and $\gamma_i = 0$ for $i = n+1, \ldots, m$. If σ_{x} and σ are unknown, we can find the estimators $\widehat{\sigma}_{\mathrm{x}}$ and $\widehat{\sigma}$ from the equations

$$\mathcal{E}\left\{\bar{Y}_i^{\delta 2}\right\} = \widehat{\sigma}_{\mathrm{x}}^2 \gamma_i^2 + \widehat{\sigma}^2, \ \ i = 1, \ldots, m. \tag{4.63}$$

However, since only one realization of the random vector $\bar{\mathbf{Y}}^\delta$ is known, the calculation of these estimators may lead to erroneous results and we must replace (4.63) by another selection criterion. For this purpose, we set (cf. (4.62))

$$a_i(\boldsymbol{\theta}) = \theta_1 \gamma_i^2 + \theta_2, \tag{4.64}$$

with $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$, $\theta_1 = \sigma_{\mathrm{x}}^2$ and $\theta_2 = \sigma^2$, and define the function

$$f\left(\bar{\mathbf{Y}}^\delta, \boldsymbol{\theta}\right) = \sum_{i=1}^m \psi\left(a_i(\boldsymbol{\theta})\right) + \psi'\left(a_i(\boldsymbol{\theta})\right)\left[\bar{Y}_i^{\delta 2} - a_i(\boldsymbol{\theta})\right],$$

with ψ being a strictly concave function. The expected value of f is given by

$$\mathcal{E}\left\{f\left(\bar{\mathbf{Y}}^\delta, \boldsymbol{\theta}\right)\right\} = \sum_{i=1}^m \psi\left(a_i(\boldsymbol{\theta})\right) + \psi'\left(a_i(\boldsymbol{\theta})\right)\left[\mathcal{E}\left\{\bar{Y}_i^{\delta 2}\right\} - a_i(\boldsymbol{\theta})\right],$$

whence, defining the estimator $\widehat{\boldsymbol{\theta}}$ through the relation

$$\mathcal{E}\left\{\bar{Y}_i^{\delta 2}\right\} = a_i\left(\widehat{\boldsymbol{\theta}}\right), \ \ i = 1, \ldots, m, \tag{4.65}$$

$\mathcal{E}\{f\}$ can be expressed as

$$\mathcal{E}\left\{f\left(\bar{\mathbf{Y}}^\delta, \boldsymbol{\theta}\right)\right\} = \sum_{i=1}^m \psi\left(a_i(\boldsymbol{\theta})\right) + \psi'\left(a_i(\boldsymbol{\theta})\right)\left[a_i\left(\widehat{\boldsymbol{\theta}}\right) - a_i(\boldsymbol{\theta})\right].$$

Then, we obtain

$$\mathcal{E}\left\{f\left(\bar{\mathbf{Y}}^{\delta},\boldsymbol{\theta}\right)\right\}-\mathcal{E}\left\{f\left(\bar{\mathbf{Y}}^{\delta},\widehat{\boldsymbol{\theta}}\right)\right\}=\sum_{i=1}^{m}\psi\left(a_i\left(\boldsymbol{\theta}\right)\right)-\psi\left(a_i\left(\widehat{\boldsymbol{\theta}}\right)\right)$$
$$+\psi'\left(a_i\left(\boldsymbol{\theta}\right)\right)\left[a_i\left(\widehat{\boldsymbol{\theta}}\right)-a_i\left(\boldsymbol{\theta}\right)\right].$$

Considering the second-order Taylor expansion

$$\psi\left(a_i\left(\boldsymbol{\theta}\right)\right)-\psi\left(a_i\left(\widehat{\boldsymbol{\theta}}\right)\right)+\psi'\left(a_i\left(\boldsymbol{\theta}\right)\right)\left[a_i\left(\widehat{\boldsymbol{\theta}}\right)-a_i\left(\boldsymbol{\theta}\right)\right]=-\frac{1}{2}\psi''\left(\xi_i\right)\left[a_i\left(\widehat{\boldsymbol{\theta}}\right)-a_i\left(\boldsymbol{\theta}\right)\right]^2$$

with some ξ_i between $a_i\left(\boldsymbol{\theta}\right)$ and $a_i\left(\widehat{\boldsymbol{\theta}}\right)$, and taking into account that ψ is strictly concave, we deduce that each term in the sum is non-negative and vanishes only for $a_i\left(\boldsymbol{\theta}\right)=a_i\left(\widehat{\boldsymbol{\theta}}\right)$. Thus, we have

$$\mathcal{E}\left\{f\left(\bar{\mathbf{Y}}^{\delta},\boldsymbol{\theta}\right)\right\}\geq\mathcal{E}\left\{f\left(\bar{\mathbf{Y}}^{\delta},\widehat{\boldsymbol{\theta}}\right)\right\},$$

for all $\boldsymbol{\theta}$. If, in addition, $\widehat{\boldsymbol{\theta}}$ is determined uniquely by (4.65), then $\widehat{\boldsymbol{\theta}}$ is the unique global minimizer of $\mathcal{E}\{f\left(\bar{\mathbf{Y}}^{\delta},\boldsymbol{\theta}\right)\}$, and we propose a regularization parameter choice method in which the estimator $\widehat{\boldsymbol{\theta}}$ is computed as

$$\widehat{\boldsymbol{\theta}}=\arg\min_{\theta}\mathcal{E}\left\{f\left(\bar{\mathbf{Y}}^{\delta},\boldsymbol{\theta}\right)\right\}. \tag{4.66}$$

Different regularization parameter choice methods can be obtained by choosing the concave function ψ in an appropriate way.

Generalized cross-validation

For the choice

$$\psi\left(a\right)=1-\frac{1}{a},$$

we obtain

$$\mathcal{E}\left\{f\left(\bar{\mathbf{Y}}^{\delta},\boldsymbol{\theta}\right)\right\}=m+\sum_{i=1}^{m}\left[\frac{\mathcal{E}\left\{\bar{Y}_i^{\delta 2}\right\}}{a_i\left(\boldsymbol{\theta}\right)^2}-\frac{2}{a_i\left(\boldsymbol{\theta}\right)}\right]. \tag{4.67}$$

As $\widehat{\boldsymbol{\theta}}$ is the unique global minimizer of $\mathcal{E}\{f\left(\bar{\mathbf{Y}}^{\delta},\boldsymbol{\theta}\right)\}$, the gradient

$$\nabla\mathcal{E}\left\{f\left(\bar{\mathbf{Y}}^{\delta},\boldsymbol{\theta}\right)\right\}=-2\sum_{i=1}^{m}\left[\frac{\mathcal{E}\left\{\bar{Y}_i^{\delta 2}\right\}}{a_i\left(\boldsymbol{\theta}\right)^3}-\frac{1}{a_i\left(\boldsymbol{\theta}\right)^2}\right]\nabla a_i\left(\boldsymbol{\theta}\right)$$

vanishes at $\widehat{\boldsymbol{\theta}}$. Thus,

$$\widehat{\boldsymbol{\theta}}^T\nabla\mathcal{E}\left\{f\left(\bar{\mathbf{Y}}^{\delta},\widehat{\boldsymbol{\theta}}\right)\right\}=0, \tag{4.68}$$

and since (cf. (4.64))

$$\boldsymbol{\theta}^T\nabla a_i\left(\boldsymbol{\theta}\right)=a_i\left(\boldsymbol{\theta}\right),$$

we deduce that

$$\sum_{i=1}^{m}\left[\frac{\mathcal{E}\left\{\bar{Y}_i^{\delta 2}\right\}}{a_i\left(\widehat{\boldsymbol{\theta}}\right)^2}-\frac{1}{a_i\left(\widehat{\boldsymbol{\theta}}\right)}\right]=0. \tag{4.69}$$

Equation (4.69) together with the relation

$$a_i\left(\widehat{\boldsymbol{\theta}}\right)=\widehat{\sigma}_{\mathrm{x}}^2\gamma_i^2+\widehat{\sigma}^2,$$

gives

$$\widehat{\sigma}_{\mathrm{x}}^2=\frac{p\left(\widehat{\alpha}\right)}{q\left(\widehat{\alpha}\right)},\ \ \widehat{\sigma}^2=\widehat{\alpha}\,\widehat{\sigma}_{\mathrm{x}}^2, \tag{4.70}$$

where

$$p\left(\alpha\right)=\sum_{i=1}^{m}\frac{\mathcal{E}\left\{\bar{Y}_i^{\delta 2}\right\}}{\left(\gamma_i^2+\alpha\right)^2}$$

and

$$q\left(\alpha\right)=\sum_{i=1}^{m}\frac{1}{\gamma_i^2+\alpha}.$$

From (4.70), it is apparent that $\widehat{\sigma}_{\mathrm{x}}^2$ and $\widehat{\sigma}^2$ are expressed in terms of the single parameter $\widehat{\alpha}$, and by (4.67) and (4.69), we find that

$$-\mathcal{E}\left\{f\left(\bar{\mathbf{Y}}^{\delta},\widehat{\boldsymbol{\theta}}\right)\right\}+m=\sum_{i=1}^{m}\frac{1}{a_i\left(\widehat{\boldsymbol{\theta}}\right)}=\frac{1}{\widehat{\sigma}_{\mathrm{x}}^2}q\left(\widehat{\alpha}\right)=\frac{q\left(\widehat{\alpha}\right)^2}{p\left(\widehat{\alpha}\right)}. \tag{4.71}$$

Now, if $\widehat{\alpha}$ minimizes the function

$$\upsilon_{\alpha}=\frac{p\left(\alpha\right)}{q\left(\alpha\right)^2}=\frac{\sum_{i=1}^{m}\left(\frac{\alpha}{\gamma_i^2+\alpha}\right)^2\mathcal{E}\left\{\bar{Y}_i^{\delta 2}\right\}}{\left(\sum_{i=1}^{m}\frac{\alpha}{\gamma_i^2+\alpha}\right)^2},$$

then by (4.71), $\widehat{\alpha}$ maximizes $-\mathcal{E}\left\{f\right\}$, or equivalently, $\widehat{\alpha}$ minimizes $\mathcal{E}\left\{f\right\}$. In practice, the expectation $\mathcal{E}\{\bar{Y}_i^{\delta 2}\}$ cannot be computed since only a single realization $\bar{y}_i^{\delta}=\mathbf{u}_i^T\mathbf{y}^{\delta}$ of $\bar{Y}_i^{\delta}$ is known. To obtain a practical regularization parameter choice method, instead of υ_{α} we consider the function

$$\upsilon_{\alpha}^{\delta}=\frac{\sum_{i=1}^{m}\left(\frac{\alpha}{\gamma_i^2+\alpha}\right)^2\left(\mathbf{u}_i^T\mathbf{y}^{\delta}\right)^2}{\left(\sum_{i=1}^{m}\frac{\alpha}{\gamma_i^2+\alpha}\right)^2}=\frac{\left\|\mathbf{y}^{\delta}-\mathbf{K}\widehat{\mathbf{x}}\right\|^2}{\left[\mathrm{trace}\left(\mathbf{I}_m-\widehat{\mathbf{A}}\right)\right]^2},$$

which represents the generalized cross-validation function discussed in Chapter 3.

Note that for $\psi(a) = (1 - 1/a^q)/q$ with $q > -1$ and $q \neq 0$, we obtain

$$\mathcal{E}\left\{f\left(\bar{\mathbf{Y}}^{\delta}, \boldsymbol{\theta}\right)\right\} = \frac{m}{q} + \sum_{i=1}^{m} \left[\frac{\mathcal{E}\left\{\bar{Y}_i^{\delta 2}\right\}}{a_i(\boldsymbol{\theta})^{q+1}} - \frac{1 + \dfrac{1}{q}}{a_k(\boldsymbol{\theta})^{q}} \right],$$

and we are led to a generalization of the cross-validation function of the form

$$\upsilon_{\alpha q}^{\delta} = \frac{\displaystyle\sum_{i=1}^{m} \left(\frac{\alpha}{\gamma_i^2 + \alpha}\right)^{q+1} \left(\mathbf{u}_i^T \mathbf{y}^{\delta}\right)^{2q}}{\left[\displaystyle\sum_{i=1}^{m} \left(\frac{\alpha}{\gamma_i^2 + \alpha}\right)^{q}\right]^{q+1}}.$$

Maximum likelihood estimation

For the choice

$$\psi(a) = \log a,$$

we obtain

$$\mathcal{E}\left\{f\left(\bar{\mathbf{Y}}^{\delta}, \boldsymbol{\theta}\right)\right\} = -m + \sum_{i=1}^{m} \left[\frac{\mathcal{E}\left\{\bar{Y}_i^{\delta 2}\right\}}{a_i(\boldsymbol{\theta})} + \log a_i(\boldsymbol{\theta}) \right], \tag{4.72}$$

and the minimization condition (4.68) yields

$$\sum_{i=1}^{m} \frac{\mathcal{E}\left\{\bar{Y}_i^{\delta 2}\right\}}{a_i\left(\widehat{\boldsymbol{\theta}}\right)} = m. \tag{4.73}$$

As before, equation (4.73) implies that $\widehat{\sigma}_{\mathrm{x}}^2$ and $\widehat{\sigma}^2$ can be expressed in terms of the single parameter $\widehat{\alpha}$ through the relations

$$\widehat{\sigma}_{\mathrm{x}}^2 = \frac{1}{m} \sum_{i=1}^{m} \frac{\mathcal{E}\left\{\bar{Y}_i^{\delta 2}\right\}}{\gamma_i^2 + \widehat{\alpha}}, \quad \widehat{\sigma}^2 = \widehat{\alpha}\,\widehat{\sigma}_{\mathrm{x}}^2, \tag{4.74}$$

and we find that

$$\begin{aligned}
\mathcal{E}\left\{f\left(\bar{\mathbf{Y}}^{\delta}, \widehat{\boldsymbol{\theta}}\right)\right\} + m \log m &= m \log m + \sum_{i=1}^{m} \log a_i\left(\widehat{\boldsymbol{\theta}}\right) \\
&= m \log m + m \log \widehat{\sigma}_{\mathrm{x}}^2 + \sum_{i=1}^{m} \log\left(\gamma_i^2 + \widehat{\alpha}\right) \\
&= m \log\left(\sum_{i=1}^{m} \frac{\mathcal{E}\left\{\bar{Y}_i^{\delta 2}\right\}}{\gamma_i^2 + \widehat{\alpha}}\right) + \sum_{i=1}^{m} \log\left(\gamma_i^2 + \widehat{\alpha}\right) \\
&= m \left[\log\left(\sum_{i=1}^{m} \frac{\mathcal{E}\left\{\bar{Y}_i^{\delta 2}\right\}}{\gamma_i^2 + \widehat{\alpha}}\right) - \frac{1}{m} \log\left(\prod_{i=1}^{m} \frac{1}{\gamma_i^2 + \widehat{\alpha}}\right) \right].
\end{aligned}$$

Hence, if $\widehat{\alpha}$ minimizes the function

$$\lambda_\alpha = \frac{\displaystyle\sum_{i=1}^{m} \frac{\mathcal{E}\left\{\bar{Y}_i^{\delta 2}\right\}}{\gamma_i^2 + \alpha}}{\sqrt[m]{\displaystyle\prod_{i=1}^{m} \frac{1}{\gamma_i^2 + \alpha}}},$$

then $\widehat{\alpha}$ minimizes $\mathcal{E}\left\{f\right\}$. In practice, we replace $\mathcal{E}\{\bar{Y}_i^{\delta 2}\}$ by $(\mathbf{u}_i^T \mathbf{y}^\delta)^2$ and minimize the maximum likelihood function

$$\lambda_\alpha^\delta = \frac{\displaystyle\sum_{i=1}^{m} \frac{\left(\mathbf{u}_i^T \mathbf{y}^\delta\right)^2}{\gamma_i^2 + \alpha}}{\sqrt[m]{\displaystyle\prod_{i=1}^{m} \frac{1}{\gamma_i^2 + \alpha}}} = \frac{\mathbf{y}^{\delta T}\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)\mathbf{y}^\delta}{\sqrt[m]{\det\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)}}. \tag{4.75}$$

An equivalent interpretation of the maximum likelihood estimation can be given as follows. Let us consider the scaled data $\bar{\mathbf{Y}}^\delta = \mathbf{U}^T \mathbf{Y}^\delta$ and let us compute the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ as

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\theta} \log p\left(\bar{\mathbf{y}}^\delta; \boldsymbol{\theta}\right),$$

with

$$p\left(\bar{\mathbf{y}}^\delta; \boldsymbol{\theta}\right) = \frac{1}{\sqrt{(2\pi)^m \det\left(\Sigma_{\mathrm{y}}(\boldsymbol{\theta})\right)}} \exp\left(-\frac{1}{2}\bar{\mathbf{y}}^{\delta T}\Sigma_{\mathrm{y}}^{-1}(\boldsymbol{\theta})\,\bar{\mathbf{y}}^\delta\right),$$

and

$$\Sigma_{\mathrm{y}}(\boldsymbol{\theta}) = \begin{bmatrix} \mathrm{diag}\left(\theta_1\gamma_i^2 + \theta_2\right)_{n\times n} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}\left(\theta_2\right)_{(m-n)\times(m-n)} \end{bmatrix}.$$

Then, taking into account that

$$\begin{aligned}\log p\left(\bar{\mathbf{y}}^\delta; \boldsymbol{\theta}\right) &= -\frac{1}{2}\bar{\mathbf{y}}^{\delta T}\Sigma_{\mathrm{y}}^{-1}(\boldsymbol{\theta})\,\bar{\mathbf{y}}^\delta - \frac{1}{2}\log\left(\det \Sigma_{\mathrm{y}}(\boldsymbol{\theta})\right) + c \\ &= -\frac{1}{2}\left[\sum_{i=1}^{m} \frac{\bar{y}_i^{\delta 2}}{\theta_1\gamma_i^2 + \theta_2} + \log\left(\prod_{i=1}^{m}\left(\theta_1\gamma_i^2 + \theta_2\right)\right)\right] + c,\end{aligned}$$

where c does not depend on $\boldsymbol{\theta}$, we see that the maximization of $\log p\left(\bar{\mathbf{y}}^\delta; \boldsymbol{\theta}\right)$ is equivalent to the minimization of $f\left(\bar{\mathbf{Y}}^\delta, \boldsymbol{\theta}\right)$ as in (4.72).

4.3.7 Noise variance estimators

In a semi-stochastic setting, we have estimated the noise variance by looking at the behavior of the residual norm in the limit of small α. This technique considers the solution of the inverse problem without regularization and requires an additional computational step.

In this section we present methods for estimating the noise variance, which do not suffer from this inconvenience.

In the analysis of the generalized cross-validation method and the maximum likelihood estimation we considered the parameter vector $\boldsymbol{\theta}$, whose components depend on the regularization parameter α and the noise variance σ^2. In fact, these methods are ideal candidates for estimating both the regularization parameter and the noise variance.

In the the generalized cross-validation method, the second relation in (4.70) gives the noise variance estimator

$$\widehat{\sigma}_{\mathtt{gcv}}^2 = \widehat{\alpha}_{\mathtt{gcv}} \frac{p\left(\widehat{\alpha}_{\mathtt{gcv}}\right)}{q\left(\widehat{\alpha}_{\mathtt{gcv}}\right)} \approx \frac{\sum_{i=1}^{m} \left(\frac{\widehat{\alpha}_{\mathtt{gcv}}}{\gamma_i^2 + \widehat{\alpha}_{\mathtt{gcv}}}\right)^2 \left(\mathbf{u}_i^T \mathbf{y}^\delta\right)^2}{\sum_{i=1}^{m} \frac{\widehat{\alpha}_{\mathtt{gcv}}}{\gamma_i^2 + \widehat{\alpha}_{\mathtt{gcv}}}} = \frac{\left\|\mathbf{y}^\delta - \mathbf{K}\widehat{\mathbf{x}}\right\|^2}{\mathrm{trace}\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)}, \tag{4.76}$$

where $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{A}}$ are computed for the regularization parameter $\widehat{\alpha}_{\mathtt{gcv}}$. The noise variance estimator (4.76) has been proposed by Wahba (1983) and numerical experiments presented by a number of researchers support the choice of this estimator (Fessler, 1991; Nychka, 1988; Thompson et al., 1991).

In the maximum likelihood estimation, a noise variance estimator can be constructed by using (4.58); the result is

$$\widehat{\sigma}_{\mathtt{mle}}^2 = \frac{1}{m} \mathbf{y}^{\delta T} \left(\mathbf{I}_m - \widehat{\mathbf{A}}\right) \mathbf{y}^\delta,$$

where $\widehat{\mathbf{A}}$ is computed for the regularization parameter $\widehat{\alpha}_{\mathtt{mle}}$. Numerical experiments where this estimator is tested has been reported by Galatsanos and Katsaggelos (1992).

An estimator which is similar to (4.76) can be derived in the framework of the unbiased predictive risk estimator method. This selection criterion chooses the regularization parameter $\widehat{\alpha}_{\mathtt{pr}}$ as the minimizer of the function

$$\pi_\alpha^\delta = \sum_{i=1}^{m} \left(\frac{\alpha}{\gamma_i^2 + \alpha}\right)^2 \left(\mathbf{u}_i^T \mathbf{y}^\delta\right)^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\gamma_i^2}{\gamma_i^2 + \alpha} - m\sigma^2.$$

Taking the derivative of π_α^δ with respect to α, and setting it equal to zero gives

$$\sigma^2 \sum_{i=1}^{n} \frac{\gamma_i^2}{\left(\gamma_i^2 + \alpha\right)^2} = \sum_{i=1}^{n} \frac{\alpha\gamma_i^2}{\left(\gamma_i^2 + \alpha\right)^3} \left(\mathbf{u}_i^T \mathbf{y}^\delta\right)^2. \tag{4.77}$$

By straigthforward calculation we find that

$$\mathrm{trace}\left(\widehat{\mathbf{A}}\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)\right) = \sum_{i=1}^{n} \frac{\alpha\gamma_i^2}{\left(\gamma_i^2 + \alpha\right)^2}$$

and that

$$\mathbf{y}^{\delta T} \left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)^T \widehat{\mathbf{A}} \left(\mathbf{I}_m - \widehat{\mathbf{A}}\right) \mathbf{y}^\delta = \sum_{i=1}^{n} \frac{\alpha^2\gamma_i^2}{\left(\gamma_i^2 + \alpha\right)^3} \left(\mathbf{u}_i^T \mathbf{y}^\delta\right)^2.$$

Now, taking into account that $\widehat{\alpha}_{\mathtt{pr}}$ and $\widehat{\alpha}_{\mathtt{gcv}}$ are asymptotically equivalent, equation (4.77) can be used to estimate the noise variance; we obtain

$$\widehat{\sigma}_{\mathtt{pr}}^{2} = \frac{\mathbf{y}^{\delta T}\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)^T \widehat{\mathbf{A}}\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)\mathbf{y}^{\delta}}{\operatorname{trace}\left(\widehat{\mathbf{A}}\left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)\right)},$$

where $\widehat{\mathbf{A}}$ is computed for the generalized cross-validation parameter $\widehat{\alpha}_{\mathtt{gcv}}$. Since

$$\mathbf{y}^{\delta} - \mathbf{K}\widehat{\mathbf{x}} = \left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)\mathbf{y}^{\delta},$$

we see that this estimator is similar to (4.76); the only difference is the multiplication with the influence matrix in both the numerator and denominator.

4.4 Marginalizing method

In a stochastic setting, a two-component data model reads as

$$\mathbf{Y}^{\delta} = \mathbf{K}_1\mathbf{X}_1 + \mathbf{K}_2\mathbf{X}_2 + \boldsymbol{\Delta}, \tag{4.78}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are assumed to be independent Gaussian random vectors characterized by $\mathbf{X}_1 \sim \mathtt{N}\left(\mathbf{0}, \mathbf{C}_{\mathrm{x}1}\right)$ and $\mathbf{X}_2 \sim \mathtt{N}\left(\mathbf{0}, \mathbf{C}_{\mathrm{x}2}\right)$. The dimensions of the random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ are n_1 and n_2, respectively, and we have $n_1 + n_2 = n$. The maximum a posteriori estimator $\widehat{\mathbf{x}}$ of the state

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

is obtained from the Bayes theorem

$$p\left(\mathbf{x}_1, \mathbf{x}_2 \mid \mathbf{y}^{\delta}\right) = \frac{p\left(\mathbf{y}^{\delta} \mid \mathbf{x}_1, \mathbf{x}_2\right) p_{\mathtt{a}}\left(\mathbf{x}_1, \mathbf{x}_2\right)}{p\left(\mathbf{y}^{\delta}\right)} = \frac{p\left(\mathbf{y}^{\delta} \mid \mathbf{x}_1, \mathbf{x}_2\right) p_{\mathtt{a}}\left(\mathbf{x}_1\right) p_{\mathtt{a}}\left(\mathbf{x}_2\right)}{p\left(\mathbf{y}^{\delta}\right)}, \tag{4.79}$$

where the a priori densities and the likelihood density are given by

$$p_{\mathtt{a}}\left(\mathbf{x}_i\right) \propto \exp\left(-\frac{1}{2}\mathbf{x}_i^T \mathbf{C}_{\mathrm{x}i}^{-1}\mathbf{x}_i\right), \quad i = 1, 2, \tag{4.80}$$

and

$$p\left(\mathbf{y}^{\delta} \mid \mathbf{x}_1, \mathbf{x}_2\right) \propto \exp\left(-\frac{1}{2}\left(\mathbf{y}^{\delta} - \mathbf{K}_1\mathbf{x}_1 - \mathbf{K}_2\mathbf{x}_2\right)^T \mathbf{C}_{\delta}^{-1}\left(\mathbf{y}^{\delta} - \mathbf{K}_1\mathbf{x}_1 - \mathbf{K}_2\mathbf{x}_2\right)\right), \tag{4.81}$$

respectively.

To show the equivalence between classical regularization and statistical inversion, we assume Gaussian densities with covariance matrices of the form

$$\mathbf{C}_{\delta} = \sigma^2\mathbf{I}_m, \quad \mathbf{C}_{\mathrm{x}i} = \sigma_{\mathrm{x}i}^2\mathbf{C}_{\mathrm{nx}i} = \sigma_{\mathrm{x}i}^2\left(\mathbf{L}_i^T\mathbf{L}_i\right)^{-1}, \quad i = 1, 2, \tag{4.82}$$

and write the penalty term in the expression of $\sigma^2 V\left(\mathbf{x}_1, \mathbf{x}_2 \mid \mathbf{y}^\delta\right)$ as

$$\sigma^2\left(\frac{1}{\sigma_{\mathrm{x}1}^2}\left\|\mathbf{L}_1\mathbf{x}_1\right\|^2+\frac{1}{\sigma_{\mathrm{x}2}^2}\left\|\mathbf{L}_2\mathbf{x}_2\right\|^2\right)=\alpha\left[\omega\left\|\mathbf{L}_1\mathbf{x}_1\right\|^2+(1-\omega)\left\|\mathbf{L}_2\mathbf{x}_2\right\|^2\right].$$

Then, it is readily seen that the regularization parameter α and the weighting factor ω are given by

$$\alpha=\frac{\sigma^2}{\sigma_{\mathrm{x}}^2},\ \ \omega=\frac{\sigma_{\mathrm{x}}^2}{\sigma_{\mathrm{x}1}^2}, \tag{4.83}$$

where

$$\frac{1}{\sigma_{\mathrm{x}}^2}=\frac{1}{\sigma_{\mathrm{x}1}^2}+\frac{1}{\sigma_{\mathrm{x}2}^2}.$$

In the framework of classical regularization theory we discussed multi-parameter regularization methods for computing α and ω, or equivalently, for estimating $\sigma_{\mathrm{x}1}$ and $\sigma_{\mathrm{x}2}$. An interesting situation occurs when the statistics of $\mathbf{X}_2$ is known, and only $\sigma_{\mathrm{x}1}$ is the parameter of the retrieval. In this case we can reduce the dimension of the minimization problem by using the so-called marginalizing technique. The idea is to formulate a minimization problem for the first component of the state vector by taking into account the statistics of the second component. The maximum a posteriori estimator for the first component of the state vector is defined as

$$\widehat{\mathbf{x}}_1=\arg\max_{\mathbf{x}_1} p\left(\mathbf{x}_1 \mid \mathbf{y}^\delta\right).$$

To compute the marginal a posteriori density $p\left(\mathbf{x}_1 \mid \mathbf{y}^\delta\right)$, we must integrate the density $p\left(\mathbf{x}_1, \mathbf{x}_2 \mid \mathbf{y}^\delta\right)$ over $\mathbf{x}_2$,

$$p\left(\mathbf{x}_1 \mid \mathbf{y}^\delta\right)=\int_{\mathbb{R}^{n_2}} p\left(\mathbf{x}_1, \mathbf{x}_2 \mid \mathbf{y}^\delta\right)\,\mathrm{d}\mathbf{x}_2=\frac{p_{\mathrm{a}}\left(\mathbf{x}_1\right)}{p\left(\mathbf{y}^\delta\right)}\int_{\mathbb{R}^{n_2}} p\left(\mathbf{y}^\delta \mid \mathbf{x}_1, \mathbf{x}_2\right) p_{\mathrm{a}}\left(\mathbf{x}_2\right)\,\mathrm{d}\mathbf{x}_2, \tag{4.84}$$

where the a priori densities and the likelihood density are given by (4.80) and (4.81), respectively. To evaluate the integral, we have to arrange the argument of the exponential function as a quadratic function in $\mathbf{x}_2$. For this purpose, we employ the technique which we used to derive the mean vector and the covariance matrix of the a posteriori density $p\left(\mathbf{x} \mid \mathbf{y}^\delta\right)$ in the one-parameter case, that is,

$$\begin{aligned}&\left[\left(\mathbf{y}^\delta-\mathbf{K}_1\mathbf{x}_1\right)-\mathbf{K}_2\mathbf{x}_2\right]^T \mathbf{C}_\delta^{-1}\left[\left(\mathbf{y}^\delta-\mathbf{K}_1\mathbf{x}_1\right)-\mathbf{K}_2\mathbf{x}_2\right]+\mathbf{x}_2^T\mathbf{C}_{\mathrm{x}2}^{-1}\mathbf{x}_2\\&=\left(\mathbf{x}_2-\bar{\mathbf{x}}_2\right)^T\widehat{\mathbf{C}}_{\mathrm{x}2}^{-1}\left(\mathbf{x}_2-\bar{\mathbf{x}}_2\right)+\left(\mathbf{y}^\delta-\mathbf{K}_1\mathbf{x}_1\right)^T\left(\mathbf{K}_2\mathbf{C}_{\mathrm{x}2}\mathbf{K}_2^T+\mathbf{C}_\delta\right)^{-1}\left(\mathbf{y}^\delta-\mathbf{K}_1\mathbf{x}_1\right),\end{aligned}$$

with

$$\bar{\mathbf{x}}_2=\mathbf{G}_2\left(\mathbf{y}^\delta-\mathbf{K}_1\mathbf{x}_1\right),\ \ \mathbf{G}_2=\left(\mathbf{K}_2^T\mathbf{C}_\delta^{-1}\mathbf{K}_2+\mathbf{C}_{\mathrm{x}2}^{-1}\right)^{-1}\mathbf{K}_2^T\mathbf{C}_\delta^{-1},$$

and

$$\widehat{\mathbf{C}}_{\mathrm{x}2}=\left(\mathbf{K}_2^T\mathbf{C}_\delta^{-1}\mathbf{K}_2+\mathbf{C}_{\mathrm{x}2}^{-1}\right)^{-1}.$$

Using the normalization condition for the Gaussian density

$$\exp\left(-\frac{1}{2}\left(\mathbf{x}_2-\bar{\mathbf{x}}_2\right)^T\widehat{\mathbf{C}}_{\mathrm{x}2}^{-1}\left(\mathbf{x}_2-\bar{\mathbf{x}}_2\right)\right),$$

we obtain

$$\begin{aligned}&p\left(\mathbf{x}_1 \mid \mathbf{y}^\delta\right)\\&\propto \exp\left(-\frac{1}{2}\left(\mathbf{y}^\delta - \mathbf{K}_1\mathbf{x}_1\right)^T \left(\mathbf{K}_2\mathbf{C}_{\mathrm{x}2}\mathbf{K}_2^T + \mathbf{C}_\delta\right)^{-1}\left(\mathbf{y}^\delta - \mathbf{K}_1\mathbf{x}_1\right) - \frac{1}{2}\mathbf{x}_1^T\mathbf{C}_{\mathrm{x}1}^{-1}\mathbf{x}_1\right),\end{aligned}$$

and it is apparent that $\widehat{\mathbf{x}}_1$ is given by (4.12) and (4.13), with $\mathbf{K}$ replaced by $\mathbf{K}_1$ and $\mathbf{C}_\delta$ replaced by

$$\mathbf{C}_{\delta_{\mathrm{y}}} = \mathbf{C}_\delta + \mathbf{K}_2\mathbf{C}_{\mathrm{x}2}\mathbf{K}_2^T. \tag{4.85}$$

Thus, when retrieving the first component of the state vector we may interpret the data error covariance matrix as being the sum of the instrumental noise covariance matrix plus a contribution due to the second component (Rodgers, 2000).

Actually, the marginalizing method can be justified more simply as follows: express the data model (4.78) as

$$\mathbf{Y}^\delta = \mathbf{K}_1\mathbf{X}_1 + \mathbf{\Delta}_{\mathrm{y}},$$

where the random data error $\mathbf{\Delta}_{\mathrm{y}}$ is given by

$$\mathbf{\Delta}_{\mathrm{y}} = \mathbf{K}_2\mathbf{X}_2 + \mathbf{\Delta},$$

and use the result $\mathcal{E}\{\mathbf{\Delta}_{\mathrm{y}}\} = \mathbf{0}$ to conclude that the covariance matrix $\mathbf{C}_{\delta_{\mathrm{y}}} = \mathcal{E}\{\mathbf{\Delta}_{\mathrm{y}}\mathbf{\Delta}_{\mathrm{y}}^T\}$ is given by (4.85). In the state space, the marginalizing method yields the random model parameter error

$$\mathbf{E}_{\mathtt{mp}} = -\widehat{\mathbf{G}}\mathbf{K}_2\mathbf{X}_2,$$

characterized by

$$\mathcal{E}\left\{\mathbf{E}_{\mathtt{mp}}\right\} = \mathbf{0},\ \ \mathbf{C}_{\mathtt{emp}} = \widehat{\mathbf{G}}\mathbf{K}_2\mathbf{C}_{\mathrm{x}2}\mathbf{K}_2^T\widehat{\mathbf{G}}^T.$$

Finally, we present a general derivation of the marginalizing method, which is not restricted to a stochastic setting. The maximum a posteriori estimator, written explicitly as

$$\begin{aligned}\begin{bmatrix}\widehat{\mathbf{x}}_1\\ \widehat{\mathbf{x}}_2\end{bmatrix} &= \left(\begin{bmatrix}\mathbf{K}_1^T\\ \mathbf{K}_2^T\end{bmatrix}\mathbf{C}_\delta^{-1}\left[\mathbf{K}_1, \mathbf{K}_2\right] + \begin{bmatrix}\mathbf{C}_{\mathrm{x}1}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}_{\mathrm{x}2}^{-1}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{K}_1^T\\ \mathbf{K}_2^T\end{bmatrix}\mathbf{C}_\delta^{-1}\mathbf{y}^\delta\\ &= \begin{bmatrix}\mathbf{K}_1^T\mathbf{C}_\delta^{-1}\mathbf{K}_1 + \mathbf{C}_{\mathrm{x}1}^{-1} & \mathbf{K}_1^T\mathbf{C}_\delta^{-1}\mathbf{K}_2\\ \mathbf{K}_2^T\mathbf{C}_\delta^{-1}\mathbf{K}_1 & \mathbf{K}_2^T\mathbf{C}_\delta^{-1}\mathbf{K}_2 + \mathbf{C}_{\mathrm{x}2}^{-1}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{K}_1^T\\ \mathbf{K}_2^T\end{bmatrix}\mathbf{C}_\delta^{-1}\mathbf{y}^\delta,\end{aligned} \tag{4.86}$$

is equivalent to the Tikhonov solution under assumptions (4.82). Setting

$$\mathbf{A} = \mathbf{K}_1^T\mathbf{C}_\delta^{-1}\mathbf{K}_1 + \mathbf{C}_{\mathrm{x}1}^{-1},\ \ \mathbf{B} = \mathbf{K}_1^T\mathbf{C}_\delta^{-1}\mathbf{K}_2,\ \ \mathbf{C} = \mathbf{K}_2^T\mathbf{C}_\delta^{-1}\mathbf{K}_2 + \mathbf{C}_{\mathrm{x}2}^{-1},$$

we compute the inverse matrix in (4.86) by using the following result (Tarantola, 2005): if $\mathbf{A}$ and $\mathbf{C}$ are symmetric matrices, then

$$\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & \mathbf{C}\end{bmatrix}^{-1} = \begin{bmatrix}\tilde{\mathbf{A}} & \tilde{\mathbf{B}}\\ \tilde{\mathbf{B}}^T & \tilde{\mathbf{C}}\end{bmatrix},$$

with

$$\tilde{\mathbf{A}} = \left(\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T\right)^{-1},\ \tilde{\mathbf{C}} = \left(\mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1},\ \tilde{\mathbf{B}} = -\tilde{\mathbf{A}}\mathbf{B}\mathbf{C}^{-1} = -\mathbf{A}^{-1}\mathbf{B}\tilde{\mathbf{C}}.$$

The first component of the state vector is then given by

$$\widehat{\mathbf{x}}_1 = \tilde{\mathbf{A}}\mathbf{K}_1^T\mathbf{C}_\delta^{-1}\mathbf{y}^\delta - \tilde{\mathbf{A}}\mathbf{B}\mathbf{C}^{-1}\mathbf{K}_2^T\mathbf{C}_\delta^{-1}\mathbf{y}^\delta = \tilde{\mathbf{A}}\left(\mathbf{K}_1^T - \mathbf{B}\mathbf{C}^{-1}\mathbf{K}_2^T\right)\mathbf{C}_\delta^{-1}\mathbf{y}^\delta.$$

By straightforward calculation we obtain

$$\begin{aligned}
&\tilde{\mathbf{A}}\left(\mathbf{K}_1^T - \mathbf{B}\mathbf{C}^{-1}\mathbf{K}_2^T\right)\mathbf{C}_\delta^{-1} \\
&= \left(\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T\right)^{-1}\left(\mathbf{K}_1^T - \mathbf{B}\mathbf{C}^{-1}\mathbf{K}_2^T\right)\mathbf{C}_\delta^{-1} \\
&= \Big\{\mathbf{K}_1^T\mathbf{C}_\delta^{-\frac{1}{2}}\Big[\mathbf{I}_m - \mathbf{C}_\delta^{-\frac{1}{2}}\mathbf{K}_2\left(\mathbf{K}_2^T\mathbf{C}_\delta^{-1}\mathbf{K}_2 + \mathbf{C}_{\mathtt{x2}}^{-1}\right)^{-1}\mathbf{K}_2^T\mathbf{C}_\delta^{-\frac{1}{2}}\Big]\mathbf{C}_\delta^{-\frac{1}{2}}\mathbf{K}_1 \\
&\quad + \mathbf{C}_{\mathtt{x1}}^{-1}\Big\}\mathbf{K}_1^T\mathbf{C}_\delta^{-\frac{1}{2}}\Big[\mathbf{I}_m - \mathbf{C}_\delta^{-\frac{1}{2}}\mathbf{K}_2\left(\mathbf{K}_2^T\mathbf{C}_\delta^{-1}\mathbf{K}_2 + \mathbf{C}_{\mathtt{x2}}^{-1}\right)^{-1}\mathbf{K}_2^T\mathbf{C}_\delta^{-\frac{1}{2}}\Big]\mathbf{C}_\delta^{-\frac{1}{2}}
\end{aligned}$$

and

$$\mathbf{I}_m - \mathbf{C}_\delta^{-\frac{1}{2}}\mathbf{K}_2\left(\mathbf{K}_2^T\mathbf{C}_\delta^{-1}\mathbf{K}_2 + \mathbf{C}_{\mathtt{x2}}^{-1}\right)^{-1}\mathbf{K}_2^T\mathbf{C}_\delta^{-\frac{1}{2}} = \mathbf{C}_\delta^{\frac{1}{2}}\left(\mathbf{C}_\delta + \mathbf{K}_2\mathbf{C}_{\mathtt{x2}}\mathbf{K}_2^T\right)^{-1}\mathbf{C}_\delta^{\frac{1}{2}},$$

which then yields

$$\widehat{\mathbf{x}}_1 = \left(\mathbf{K}_1^T\mathbf{C}_{\delta_\mathrm{y}}^{-1}\mathbf{K}_1 + \mathbf{C}_{\mathtt{x1}}^{-1}\right)^{-1}\mathbf{K}_1^T\mathbf{C}_{\delta_\mathrm{y}}^{-1}\mathbf{y}^\delta,$$

with $\mathbf{C}_{\delta_\mathrm{y}}$ as in (4.85). This derivation clearly shows that the solution for the full state vector will give the same results for each of the partial state vectors as their individual solutions.