

# Chapter 2

## Selection of the Regularization Parameter

### 2.1 General Considerations

The success of all currently available regularization techniques relies heavily on the proper choice of the regularization parameter. Although many regularization parameter selection methods (RPSMs) have been proposed, very few of them are used in engineering practice. This is due to the fact that theoretically justified methods often require unrealistic assumptions, while empirical methods do not guarantee a good regularization parameter for any set of data. Among the methods that found their way into engineering applications, the most common are Morozov's Discrepancy Principle (abbreviated as MDP) [morozov84, phillips62], Mallows' CL [mallows73], generalized cross validation (abbreviated as GCV) [wahba90], and the L-curve method [hansen98]. A high sensitivity of CL and MDP to an underestimation of the noise level has limited their application to cases in which the noise level can be estimated with high fidelity [hansen98]. On the other hand, noise-estimate-free GCV occasionally fails, presumably due to the presence of correlated noise [wahba90]. The L-curve method is widely used; however, this method is nonconvergent [leonov97, vogel96]. An example of image restoration using different values of regularization parameters is shown in Figs. 2.1, 2.2, 2.3, 2.4, and 2.5. The Matlab code for this example was provided by Dr. Curt Vogel of Montana State University in a personal communication. The original image is presented in Fig. 2.1, and the observed blurred image is in Fig. 2.2.

Figures 2.3 and 2.4 represent reconstructed images with too small and too large regularization parameters, respectively. These two images illustrate the importance of the regularization parameter selection for proper image restoration. For comparison, Fig. 2.5 represents the image reconstructed using a "good" value of regularization parameter.

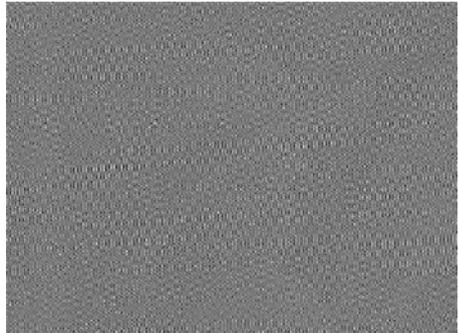
**Fig. 2.1** Original image



**Fig. 2.2** Observed blurred image



**Fig. 2.3** Reconstructed image with a very small regularization parameter ( $\lambda = 10-20$ )



## 2.2 Discrepancy Principle

The discrepancy principle is the most widely used method which requires a priori knowledge of some of the noise properties such as the power of the noise. The regularization parameter value is chosen as a solution of the equation

**Fig. 2.4** Reconstructed image with a very large regularization parameter ( $\lambda = 10$ )



**Fig. 2.5** Reconstructed image with a good value of regularization parameter ( $\lambda = 0.0007$ )



$$\|Y - Dx_\lambda\|_2 = \varepsilon, \text{ where } \|\eta\|_2 \leq \varepsilon. \quad (2.1)$$

The  $\varepsilon$  is the upper bound on the variance of the noise.

The regularization parameter  $\lambda$  is chosen such that the corresponding residual (left-hand side) of Eq. (2.1) is less than or equal to the a priori specified bound (right-hand side) for the noise level in the response. Since a smaller  $\lambda$  corresponds to less stable solutions, the  $\lambda$  for which the residual equals the specified noise level is chosen. There is no reason to expect a residual less than the noise level. In modeling from data, a residual less than the noise level in the response corresponds to overfitting, which is a term for learning noise in the training data. The regularization method with  $\lambda$  chosen according to the discrepancy principle in Eq. (2.1) is convergent and of optimal order [engl00, morozov84]. Application of the discrepancy principle requires solving the following nonlinear equation with respect to  $\lambda$  as shown in [golub97].

$$\left\| Y - D(D^T D + \lambda I)^{-1} D^T Y \right\|_2 = \varepsilon. \quad (2.2)$$

For  $\lambda > 0$ , the identity

$$I - D(D^T D + \lambda I)^{-1} D^T = \lambda (D D^T + \lambda I)^{-1} \quad (2.3)$$

holds. Hence, Eq. (2.2) can be rewritten as

$$\left\| Y \lambda (D D^T + \lambda I)^{-1} \right\|_2 = \varepsilon \quad (2.4)$$

or

$$\left[ Y \lambda (D D^T + \lambda I)^{-1} \right]^T \cdot \left[ Y \lambda (D D^T + \lambda I)^{-1} \right] = \varepsilon, \quad (2.5)$$

and after elementary matrix algebra, we arrive at the following nonlinear equation for  $\lambda$ .

$$\lambda^2 Y^T (D D^T + \lambda I)^{-2} Y = \varepsilon. \quad (2.6)$$

Since the derivative of the left-hand side is equal to

$$2\lambda Y^T D (D^T D + \lambda I)^{-3} D^T Y, \quad (2.7)$$

the function  $\lambda^2 Y^T (D D^T + \lambda I)^{-2} Y$  is strictly increasing for  $\lambda > 0$ , and Eq. (2.6) has a unique positive solution.

A very important property of the discrepancy principle is its convergence or regularity, which means that as error  $\|\eta\|_2$  in the data goes to zero, the  $\lambda$  selected by MDP goes to zero; hence, the approximated regularized solution  $x_\lambda$  converges to the exact solution or true image  $x_{\text{exact}}$ . Normally, the literature on inverse problems analyzes the rates of convergence of  $x_\lambda$  to  $x_{\text{exact}}$ . The faster the method converges, the better its behavior.

Statistical literature on the selection of a regularization parameter is more concerned about asymptotic behavior of different methods as the number of samples  $N$  goes to infinity.

To apply MDP, we must have a priori knowledge about the noise level in the response. Since the noise level is usually unknown, we use an estimate of the noise level. One of the methods for noise estimation is described in [wahba90] and consists in monitoring the function

$$\hat{\sigma}_\eta^2 = \frac{\|Y - D x_\lambda\|_2^2}{\text{trace}\left(I - D(D^T D + \lambda I)^{-1} D\right)}. \quad (2.8)$$

The plateau of this function can serve as a good estimate of the noise variance. However, in practice the variance of the residuals of the least square solution is often used as a quick estimate of the noise level.

Unfortunately, MDP is very sensitive to an underestimation of the noise level. This limits its application to cases in which the noise level can be estimated with high fidelity [hansen98]. The MDP belongs to a posteriori methods for the selection of a regularization parameter. A posteriori RPSM requires the noise level to be either known or reliably estimated. Such an estimate of the noise level can be hard to obtain.

## 2.3 L-Curve

An alternative approach to regularization parameter selection uses noise-level-free RPSMs. Noise-level-free RPSMs are also referred to as heuristic RPSMs. We are now going to consider some of these methods. The method which attracted the attention of researchers recently is the L-curve method [hansen93]. The method is based on the plot of the logarithm of the solution norm  $x_\lambda$  versus the logarithm of the norm of the residuals. In many cases, such a curve has a characteristic L shape. It is then argued that the optimal regularization parameter has to be selected at the point of maximum curvature of the curve or its “elbow.” Mathematically, the L-curve criterion seeks to maximize the curvature

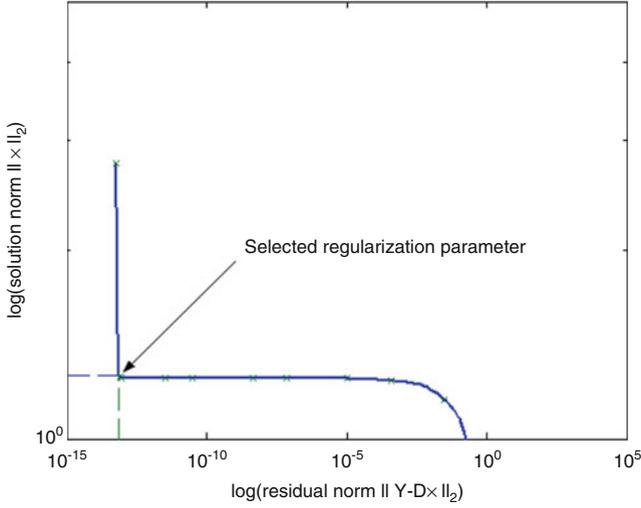
$$C_L(\lambda) = \frac{\rho' \eta'' - \rho'' \eta'}{\left((\rho')^2 + (\eta')^2\right)^{3/2}} = \max, \text{ where} \quad (2.9)$$

$$\rho(\lambda) = \log \left\| Y - D(D^T D + \lambda I)^{-1} D Y \right\|_2 = \log \left\| \lambda (D D^T + \lambda I)^{-1} Y \right\|_2 \text{ and} \quad (2.10)$$

$$\eta(\lambda) = \log \left\| (D^T D + \lambda I)^{-1} D^T Y \right\|_2. \quad (2.11)$$

The differentiation is with respect to  $\lambda$ . A typical L-curve is presented in Fig. 2.6. To obtain the curve, we used Hansen’s regularization toolbox [hansen94].

The L-curve method recently suffered major theoretical as well as practical setbacks. It was shown in [leonov97, vogel96] that the method is generally not convergent, and in practice the L-curve may not have an “elbow” or have several ones.



**Fig. 2.6** The generic form of the L-curve

## 2.4 Mallows's $C_L$

The other heuristic method which was proposed in statistical literature is the Mallows's  $C_L$  or unbiased prediction risk estimation method. In the following derivations, we closely follow [vogel02]. Let's call the difference between the regularized image and true image an estimation error

$$\varepsilon_\lambda = x_\lambda - x_{\text{true}}. \quad (2.12)$$

Obviously, this quantity is unknown and not computable due to the unavailability of the true image  $x_{\text{true}}$ . The observed image represents convolution of the true image with point spread function plus some additive noise as in

$$Y = Dx_{\text{true}} + \eta. \quad (2.13)$$

Let's define the predictive error as the difference between two quantities

$$p_\lambda = Dx_\lambda - Dx_{\text{true}}. \quad (2.14)$$

We can express the regularized image  $x_\lambda$  as

$$x_\lambda = (D^T D + \lambda I)^{-1} D^T Y \quad (2.15)$$

or

$$x_\lambda = (D^T D + \lambda I)^{-1} D^T (D x_{\text{true}} + \eta), \quad (2.16)$$

and, hence, the predictive error can be expressed as

$$p_\lambda = (H - I) D x_{\text{true}} + H \eta, \quad (2.17)$$

where  $H$  is the hat or influence matrix

$$H = D(D^T D + \lambda I)^{-1} D^T. \quad (2.18)$$

As shown in [wahba90], the mean value of the mean squared predictive error can be written as

$$E\left(\frac{1}{n} \|p_\lambda\|^2\right) = \frac{1}{n} \|(H - I) D f_{\text{true}}\|^2 + \frac{\sigma^2}{n} \text{trace}(H). \quad (2.19)$$

Notice that this value is not computable either; however, we can introduce the training error as

$$r_\lambda = D f_\lambda - Y. \quad (2.20)$$

As shown in [vogel02],

$$E\left(\frac{1}{n} \|r_\lambda\|^2\right) = E\left(\frac{1}{n} \|p_\lambda\|^2\right) - 2 \frac{\sigma^2}{n} \text{trace}(H) + \sigma^2; \quad (2.21)$$

hence,

$$E\left(\frac{1}{n} \|p_\lambda\|^2\right) = E\left(\frac{1}{n} \|r_\lambda\|^2\right) + 2 \frac{\sigma^2}{n} \text{trace}(H) - \sigma^2. \quad (2.22)$$

The  $C_L$  function is given as

$$C_L = \frac{1}{n} \|r_\lambda\|^2 + 2 \frac{\sigma^2}{n} \text{trace}(H) - \sigma^2; \quad (2.23)$$

hence, the  $C_L$  function is an unbiased estimator for the mean squared predictive error

$$E(C_L) = E\left(\frac{1}{n} \|p_\lambda\|^2\right). \quad (2.24)$$

**Fig. 2.7** Reconstructed image with regularization parameter selected with CL ( $\lambda = 0.0007$ )



In the case of a correctly specified point spread function and Gaussian noise,  $C_L$  is theoretically optimal. It should be noted that  $C_L$  requires the estimation of the noise variance

$$\sigma^2 = \text{var}(\eta), \quad (2.25)$$

and, what is even more important, the  $C_L$  performance depends heavily on the accuracy of this estimate. The image reconstructed using  $C_L$  as parameter selection method is shown in Fig. 2.7.

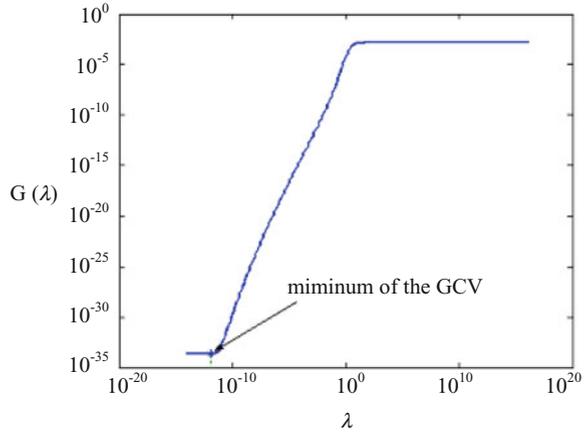
## 2.5 Generalized Cross Validation

To overcome the inconvenience of noise estimation, Wahba [wahba90] suggested a noise-free method for the selection of a regularization parameter which is currently widely used—generalized cross validation (GCV). GCV is a rotation-invariant version of ordinary leave-one-out cross validation. The ordinary cross validation is known not to be invariant under data transformations, and GCV fixes this problem. The GCV seeks for a minimum of the following function:

$$\text{GCV}(\lambda) = \frac{\frac{1}{n} \|Y - Dx_\lambda\|_2^2}{\left[ \frac{1}{n} \|\text{trace}(I - H)\| \right]^2}. \quad (2.26)$$

We can see that the GCV function is the ratio of two functions—the mean sum of squares  $\frac{1}{n} \|Y - Dx_\lambda\|_2^2$  and a penalty function  $\left[ \frac{1}{n} \|\text{trace}(I - H)\| \right]^2$  which is often called the effective degrees of freedom and is used to quantify the amount of information in ill-posed problems. The GCV function has a number of nice properties such as convergence to the optimal regularization parameter as  $n$  goes to infinity, and convergence rates for this method are also available. GCV is derived under the assumption of white Gaussian noise, and, if this condition fails to hold,

**Fig. 2.8** A typical GCV function



**Fig. 2.9** Reconstructed image using GCV for the selection of regularization parameter ( $\lambda = 0.0007$ )



the GCV may produce grossly under regularized solutions. The GCV function itself can also have very flat minima, thus leading to numerical difficulties in determining a unique value of regularization parameter. A typical GCV function is plotted in Fig. 2.8. An image reconstructed using GCV for the selection of the regularization parameter is shown in Fig. 2.9.

## 2.6 Information Approach to the Selection of a Regularization Parameter

There are two potential problems with all the methods that we considered so far. First of all, if the true relationships between the observed image  $Y$  and original image  $X$  are not linear or if the response function is not known exactly, then we have what is called functional misspecification. The second problem is the assumption of white Gaussian noise in the data which is rarely a valid assumption in image

**Fig. 2.10** Reconstructed image with regularization parameter selected with ICOMPRPS ( $\lambda = 0.0007$ )



processing applications. This second type of misspecification is called distributional misspecification. Both these misspecifications affect the estimation of the covariance matrix of the restored image which is implicitly used by many methods to select the regularization parameter. We now consider the information approach to regularization parameter selection, which is robust to the model misspecification and also robust to the underestimation of the regularization parameter. However, first, we have to consider some theoretical preliminaries such as Kullback-Leibler (abbreviated KL) distance (Fig. 2.10).

When the parameters of a specified model  $f(X_i, Y_i; b)$  are estimated by the maximum penalized likelihood (MPL) method (see Appendix D), each particular choice of the penalty operator and regularization parameter yields some approximating density  $\hat{f}_\lambda \equiv f(X_i, Y_i; \hat{b}_\lambda)$ . The closeness of this approximating density  $\hat{f}_\lambda$  to the unknown true density  $g(X_i, Y_i)$ , assuming such exists, can be evaluated by the Kullback-Leibler [kullback51] (abbreviated as KL) information (or distance) that measures the divergence between the densities

$$\text{KL}(\hat{f}_\lambda; g) \equiv E_{w,z} \left\{ \log \frac{g}{\hat{f}_\lambda} \right\} = \int \cdots \int \log \frac{g(w, z)}{f(w, z; \hat{b}_\lambda)} \cdot g(w, z) dw_1, dw_2, \dots, dw_m dz. \quad (2.27)$$

The regularization parameter can be selected to minimize the mean KL distance. The mean KL distance is the KL distance averaged over all possible data sets  $D$  which can be used to obtain the approximating density  $\hat{f}_\lambda$ .

$$\hat{\lambda}_{\text{KL}} = \arg \min_{\lambda} \{ E_D \text{KL}(\hat{f}_\lambda; g) \}. \quad (2.28)$$

Such a choice guarantees that, on the average, the corresponding approximating density will be closest among those considered in the sense of the minimum KL distance. We can decompose the mean KL distance into a “systematic error” and a “random error”:

$$\begin{aligned}
E_D \text{KL}(\hat{f}_\lambda; g) &= E_D \left\{ E_{W,Z} \log \frac{g}{\hat{f}_\lambda} \right\} \\
&= E_D \left\{ E_{W,Z} \log \frac{g f_\lambda^* f_\lambda^*}{f_\lambda^* f_\lambda^* \hat{f}_\lambda} \right\} \\
&= \underbrace{E_{W,Z} \log \frac{g}{f_\lambda^*}}_{\text{Systematic Error}} + E_{W,Z} \log \frac{f_\lambda^*}{\hat{f}_\lambda} + \underbrace{E_D \left\{ E_{W,Z} \log \frac{f_\lambda^*}{\hat{f}_\lambda} \right\}}_{\text{Random Error}},
\end{aligned} \tag{2.29}$$

where  $f_\lambda^* \equiv f(W, Z; b^*)$ , and  $b^*$  is a solution of

$$E_{W,Z} \left\{ \frac{\partial}{\partial b} \text{LL}(W, Z|b) \right\} = 0, \tag{2.30}$$

or the limiting value of the maximum likelihood (ML) estimator  $f_\lambda^* \equiv f(W, Z; b_\lambda^*)$ , and  $b_\lambda^*$  is a solution of

$$E_{W,Z} \left\{ \frac{\partial}{\partial b} \text{PLL}(W, Z|b) \right\} = 0, \tag{2.31}$$

or the limiting value of the MPL estimator.

The systematic error, which can also be termed the bias, consists of two terms. The first term represents the error of modeling and vanishes when the model is correctly specified. The second term represents the error due to using a penalization and vanishes when the maximum likelihood method of estimation is used. The random error, also called the variance, arises due to the inaccuracy of the model's parameter estimation because of a limited number of observations. When the model is correctly specified and the ML method is used, only the variance term contributes to the mean KL distance. However, as we know, the variance in a case of ill-conditioned data sets can be very large and make the approximating density useless. Although penalization introduces a bias, it also drastically reduces the variance, allowing for a trade-off which may reduce the mean KL distance. This means that, on the average, with a properly chosen regularization parameter, the penalized model can be closer to the true model.

From the definition of the KL distance, it can be seen that, since  $E_D \{ E_{W,Z} \log g \}$  does not depend on the model  $\hat{f}_\lambda$ , minimization of the mean KL distance is equivalent to maximization of the mean expected log likelihood (abbreviated as MELL) which is defined as

$$\text{MELL}(\lambda) \equiv E_D \{ E_{W,Z} \log \hat{f}_\lambda \}, \tag{2.32}$$

where, as before,  $W$  and  $Z$  have the same joint distribution as  $X_i$  and  $Y_i$  and are independent of them. That is why the mean expected log likelihood is extensively used in statistical model selection as a powerful tool for evaluating the model performance and for choosing one model from the competing models. In a pioneering work, [akaike73] introduced the MELL as a model selection method and justified the use of ML for parameter estimation.

In the Gaussian case (when  $Z|W$  is normally distributed) and with a correctly specified model, maximization of the mean expected log likelihood is equivalent to minimization of the mean predictive error (abbreviated as MPE). As with MPE, the mean expected log likelihood is not computable because of the unknown true distribution, but it can be estimated by plugging the empirical distribution into Eq. (2.32). By this means, the so-called average log likelihood (abbreviated as ALL) is obtained as follows:

$$\text{ALL}(\hat{b}_\lambda) = \frac{1}{n} \sum_{i=1}^n \log f(Y_i | X_i; \hat{b}_\lambda). \quad (2.33)$$

Despite the fact that  $\text{ALL}(b) \rightarrow \text{ELL}(b)$  as  $n \rightarrow \infty$ , due to the law of large numbers, the ALL, evaluated at MPLE  $\hat{b}_\lambda$ , is a biased estimator of the MELL of the MPL model, i.e.,  $E_D \text{ALL}(\hat{b}_\lambda) \neq \text{MELL}(\lambda)$ . This bias should be corrected when we use MELL as an RPSM. In the next section, one of the methods for bias correction is presented. This method is usually used for deriving information model selection criteria as in [akaike73, sakamoto86, bozdogan87, konishi96, shibata89].

An information-based RPSM is given as the maximization of the mean expected log likelihood (Eq. 2.32) of maximum penalized likelihood models

$$\hat{\lambda}_{\text{MELL}} = \arg \max_{\lambda} \{\text{MELL}(\lambda)\}. \quad (2.34)$$

As already mentioned, the MELL is not computable and can be estimated by the ALL. The ALL, evaluated at the MPLE, is a biased estimator of MELL. To quantify the bias of ALL in estimating the MELL, we first define the expected penalized log likelihood (abbreviated as EPLL) as

$$\text{EPLL}(b) \equiv E_{W,Z} \text{PLL}(W, Z | b) \quad (2.35)$$

and expand it in a Taylor series at  $\hat{b}_\lambda$  around  $b_\lambda^*$ , which is the limiting value of the MPLE  $\hat{b}_\lambda$  as  $n \rightarrow \infty$ .

$$\begin{aligned}
\text{EPLL}(\hat{b}_\lambda) &\approx \text{EPLL}(b_\lambda^*) + \left\{ \frac{\partial}{\partial b} \text{EPLL}(b_\lambda^*) \right\}^T (\hat{b}_\lambda - b_\lambda^*) \\
&\quad + \frac{1}{2} (\hat{b}_\lambda - b_\lambda^*)^T \left\{ \frac{\partial^2}{\partial b \partial b^T} \text{EPLL}(b_\lambda^*) \right\} (\hat{b}_\lambda - b_\lambda^*) \\
&= \text{EPLL}(b_\lambda^*) - \frac{1}{2} (\hat{b}_\lambda - b_\lambda^*)^T J (\hat{b}_\lambda - b_\lambda^*),
\end{aligned} \tag{2.36}$$

where

$$J \equiv - \frac{\partial^2}{\partial b \partial b^T} \text{EPLL}(b_\lambda^*). \tag{2.37}$$

Next, we expand the average penalized log likelihood (abbreviated as APLL) defined as

$$\text{APLL}(b) \equiv \frac{1}{n} \sum_{i=1}^n \text{LL}(X_i, Y_i | b) - \lambda p(b) \tag{2.38}$$

in a Taylor series at  $b_\lambda^*$  around  $\hat{b}_\lambda$  as

$$\begin{aligned}
\text{APLL}(b_\lambda^*) &\approx \text{APLL}(\hat{b}_\lambda) + \left\{ \frac{\partial}{\partial b} \text{APLL}(\hat{b}_\lambda) \right\}^T (b_\lambda^* - \hat{b}_\lambda) \\
&\quad + \frac{1}{2} (b_\lambda^* - \hat{b}_\lambda)^T \left\{ \frac{\partial^2}{\partial b \partial b^T} \text{APLL}(\hat{b}_\lambda) \right\} (b_\lambda^* - \hat{b}_\lambda) \\
&\approx \text{APLL}(\hat{b}_\lambda) - \frac{1}{2} (b_\lambda^* - \hat{b}_\lambda)^T J (b_\lambda^* - \hat{b}_\lambda)
\end{aligned} \tag{2.39}$$

We used the fact that

$$\frac{\partial}{\partial b} \text{APLL}(\hat{b}_\lambda) = 0 \tag{2.40}$$

and that, by the law of large numbers, as  $n \rightarrow \infty$

$$\left\{ \frac{\partial^2}{\partial b \partial b^T} \text{APLL}(b_\lambda^*) \right\} \rightarrow \left\{ \frac{\partial^2}{\partial b \partial b^T} \text{EPLL}(b_\lambda^*) \right\}, \tag{2.41}$$

and, since  $\hat{b}_\lambda \rightarrow b_\lambda^*$  as  $n \rightarrow \infty$ , we have

$$\left\{ \frac{\partial^2}{\partial b \partial b^T} \text{APLL}(\hat{b}_\lambda) \right\} \rightarrow \left\{ \frac{\partial^2}{\partial b \partial b^T} \text{EPLL}(b_\lambda^*) \right\}. \tag{2.42}$$

Using  $E_D \text{EPLL}(\hat{b}_\lambda^*) = E_D \text{APLL}(\hat{b}_\lambda^*)$  and combining Eqs. (2.36) and (2.39), we obtain

$$E_D \text{EPLL}(\hat{b}_\lambda) \approx E_D \text{APLL}(\hat{b}_\lambda) - E_D \left\{ (b_\lambda^* - \hat{b}_\lambda)^\top J (b_\lambda^* - \hat{b}_\lambda) \right\}. \quad (2.43)$$

Since

$$E_D \text{EPLL}(\hat{b}_\lambda) = E_D \text{ELL}(\hat{b}_\lambda) - \lambda E_D p(\hat{b}_\lambda) \quad \text{and} \quad (2.44)$$

$$E_D \text{APLL}(\hat{b}_\lambda) = E_D \text{ALL}(\hat{b}_\lambda) - \lambda E_D p(\hat{b}_\lambda), \quad (2.45)$$

we have

$$\begin{aligned} E_D \text{ELL}(\hat{b}_\lambda) &\approx E_D \text{ALL}(\hat{b}_\lambda) - E_D \left\{ (b_\lambda^* - \hat{b}_\lambda)^\top J (b_\lambda^* - \hat{b}_\lambda) \right\} \\ &\approx E_D \text{ALL}(\hat{b}_\lambda) - \frac{1}{n} \text{trace}(IJ^{-1}), \end{aligned} \quad (2.46)$$

where we use the asymptotic normality of the maximum penalized likelihood estimator and the trace result from Appendix D to obtain

$$E_D \left\{ (b_\lambda^* - \hat{b}_\lambda)^\top J (b_\lambda^* - \hat{b}_\lambda) \right\} = \frac{1}{n} \text{trace}(IJ^{-1}). \quad (2.47)$$

Therefore, an unbiased estimator of the mean expected log likelihood is defined as

$$T_{\text{MELL}}(\hat{b}_\lambda) \equiv \text{ALL}(\hat{b}_\lambda) - \frac{1}{n} \text{trace}(\hat{I}\hat{J}^{-1}), \quad (2.48)$$

where

$$\hat{I} = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial b} \text{PLL}(X_i, Y_i | \hat{b}_\lambda) \frac{\partial}{\partial b^\top} \text{PLL}(X_i, Y_i | \hat{b}_\lambda) \quad \text{and} \quad (2.49)$$

$$\hat{J} = -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial b \partial b^\top} \text{PLL}(X_i, Y_i | \hat{b}_\lambda), \quad (2.50)$$

and the corresponding RPSM is

$$\hat{\lambda}_{\text{MELL}} = \underset{\lambda}{\text{argmax}} \left\{ \text{ALL}(\hat{b}_\lambda) - \frac{1}{n} \text{trace}(\hat{I}\hat{J}^{-1}) \right\}. \quad (2.51)$$

A number of RPSMs can follow from this. When the model is Gaussian, correctly specified, and  $X$  is fixed, the well-known Mallows' (1973) CL method is obtained.

$$\hat{\lambda}_{\text{CL}} = \operatorname{argmin}_{\lambda} \left\{ \frac{1}{n} \|Y - X\hat{b}_{\lambda}\|^2 + \frac{2\sigma^2}{n} \operatorname{trace} \left( X^T X (X^T X + n\lambda I_m)^{-1} \right) \right\}. \quad (2.52)$$

When the model is Gaussian and  $\sigma^2$  is treated as a nuisance parameter, and  $J$  and  $I$  are estimated as

$$\hat{J} = -\frac{1}{\sigma^2} \left( \frac{1}{n} \sum_{i=1}^n X_i X_i^T + \lambda I_m \right) \quad \text{and} \quad (2.53)$$

$$\hat{I} = \frac{1}{n\sigma^4} \sum_{i=1}^n r_{\text{ols}i}^2 X_i X_i^T, \quad (2.54)$$

Shibata's (1989) regularization information criterion (abbreviated as RIC) is obtained, and the corresponding RPSM is

$$\hat{\lambda}_{\text{RIC}} = \operatorname{argmin}_{\lambda} \left\{ \frac{1}{n} \|Y - X\hat{b}_{\lambda}\|^2 + \frac{2\sigma^2}{n} \sum_{i=1}^n \frac{r_{\text{ols}i}^2}{\sigma^2} H_{ii} \right\}, \quad (2.55)$$

where  $H = X(X^T X + n\lambda I_m)^{-1} X^T$  and  $r_{\text{ols}i} = Y_i - X_i^T \hat{b}$ .

When  $\hat{b}_{\lambda}$  is an M-estimator [huber81], Konishi and Kitagawa [konishi96] propose an information criterion for choosing the regularization parameter which is similar to RIC (Eq. 2.55).

We also suggest a RPSM that uses [bozdogan96a, bozdogan96b] an informational complexity framework to account for interdependencies between parameter estimates when evaluating the bias of ALL in estimating the MELL. The resulting method, by means of a more severe penalization of the inaccuracy of estimation, produces slightly overestimated regularization parameter values as compared to that given by CL or RIC. Overestimation, however, is in a safe direction and is shown to be beneficial in situations with a limited number of observations.

Despite its simplicity, the Gaussian correctly specified case is very important, especially for the numerical solution of integral equations with a method of regularization, because  $X$  is fixed and there is no functional misspecification. In the Gaussian correctly specified case, the information RPSM (Eq. 2.51) becomes similar to CL.

The MELL RPSM (Eq. 2.51) reduces to Mallows' CL under the following conditions: the approximating distribution (model) belongs to the Gaussian family, i.e.,

$$W \sim N_m(\mu, A) \quad \text{and} \quad (2.56)$$

$$Z|W \sim N(m(W), \sigma^2), \quad (2.57)$$

and the model is correctly specified, meaning that there exists  $b_0$ , referred to as the true regression coefficients (or the true solution), such that

$$f(W, Z; b_0) = g(W, Z), \quad (2.58)$$

where  $g(W, Z)$  is the actual (true) data-generating distribution and where  $\sigma^2$ , the conditional variance of the output (or noise variance), is treated as a nuisance parameter. In particular, correct specification implies that

$$E_{Z|W}\{Z - W^T b^*\} = 0 \quad \text{and} \quad (2.59)$$

$$E_{Z|W}\{(Z - W^T b^*)(Z - W^T b^*)^T\} = \sigma^2. \quad (2.60)$$

The log likelihood in this case is

$$\begin{aligned} \log f(Z|W; b) &= \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(Z - W^T b)^T(Z - W^T b)\right) \\ &= \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(Z - W^T b)^T(Z - W^T b). \end{aligned} \quad (2.61)$$

Its derivatives with respect to  $b$  are

$$\frac{\partial}{\partial b} \log f(Z|W; b) = \frac{1}{\sigma^2} W(Z - W^T b), \quad (2.62)$$

$$\frac{\partial}{\partial b^T} \log f(Z|W; b) = \frac{1}{\sigma^2} (Z - W^T b)^T W^T, \quad \text{and} \quad (2.63)$$

$$\frac{\partial^2}{\partial b \partial b^T} \log f(Z|W; b) = -\frac{1}{\sigma^2} W W^T. \quad (2.64)$$

Using the quadratic penalty, matrix  $J$  becomes

$$\begin{aligned} J &= -\frac{\partial^2}{\partial b \partial b^T} E_{W,Z}\{\log f_\lambda^* - \lambda p(b_\lambda^*)\} \\ &= E_W\left\{\frac{1}{\sigma^2} W W^T + \lambda p'(b_\lambda^*) p'(b_\lambda^*)^T\right\} \\ &= \frac{1}{\sigma^2} E_W\{W W^T\} + \lambda p'(b_\lambda^*) p'(b_\lambda^*)^T = \frac{1}{\sigma^2} (E_W\{W W^T\} + \lambda I_m) \end{aligned} \quad (2.65)$$

and can be estimated as

$$\hat{J} = \frac{1}{\sigma^2} \left( \frac{1}{n} \sum_{i=1}^n X_i X_i^T + \lambda I_m \right) = \frac{1}{n\sigma^2} (X^T X + n\lambda I_m). \quad (2.66)$$

Matrix  $I$  becomes

$$\begin{aligned}
I &= E_{W,Z} \left\{ \frac{\partial}{\partial b} (\log f_\lambda^* - \lambda p(b_\lambda^*)) \frac{\partial}{\partial b^T} (\log f_\lambda^* - \lambda p(b_\lambda^*)) \right\} \\
&= E_{W,Z} \left\{ \frac{\partial}{\partial b} \log f_\lambda^* \frac{\partial}{\partial b^T} \log f_\lambda^* \right\} - E_{W,Z} \left\{ \frac{\partial}{\partial b^T} \log f_\lambda^* \right\} E_{W,Z} \left\{ \frac{\partial}{\partial b} \log f_\lambda^* \right\} \\
&= \frac{1}{\sigma^2} E_W \{ WW^T \} + \frac{1}{\sigma^4} E_W \left\{ WW^T (b^* - b_\lambda^*) (b^* - b_\lambda^*)^T WW^T \right\} \\
&\quad - \frac{1}{\sigma^4} E_W \{ WW^T \} (b^* - b_\lambda^*) (b^* - b_\lambda^*)^T E_W \{ WW^T \},
\end{aligned} \tag{2.67}$$

and, for a large  $n$ , it can be estimated as

$$\hat{I} = \frac{1}{n\sigma^2} \sum_{i=1}^n X_i X_i^T = \frac{1}{n\sigma^2} X^T X. \tag{2.68}$$

The trace term becomes

$$\begin{aligned}
\text{trace}(\hat{I} \hat{J}^{-1}) &= \text{trace} \left( \frac{1}{n\sigma^2} X^T X \cdot n\sigma^2 (X^T X + n\lambda I_m)^{-1} \right) \\
&= \text{trace} \left( X^T X (X^T X + n\lambda I_m)^{-1} \right) \\
&= \text{trace}(H),
\end{aligned} \tag{2.69}$$

where the hat matrix is defined as  $H \equiv X(X^T X + n\lambda I_m)^{-1} X^T$ .

The RPSM becomes

$$\hat{\lambda}_{\text{MELL}} = \arg \min_{\lambda} \left\{ \frac{1}{2\sigma^2} \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^T \hat{b}_\lambda)^2 + \frac{1}{n} \text{trace}(H) \right\} \quad \text{or} \tag{2.70}$$

$$\hat{\lambda}_{\text{MELL}} = \arg \min_{\lambda} \left\{ \frac{1}{n} \|Y - X \hat{b}_\lambda\|^2 + \frac{2\sigma^2}{n} \text{trace}(H) \right\}. \tag{2.71}$$

This is exactly CL. Therefore, CL can be viewed as an information RPSM when the model is correctly specified and is Gaussian with fixed  $X$ .

Dropping the assumption of correct model specification and using the Gaussian approximating distribution as in the previous case, a similar expression for  $J$  is obtained as

$$J = \frac{1}{\sigma^2} (E_W \{ WW^T \} + \lambda I_m) \tag{2.72}$$

and estimated as

$$\hat{J} = \frac{1}{\sigma^2} \left( \frac{1}{n} \sum_{i=1}^n X_i X_i^T + \lambda I_m \right) = \frac{1}{n\sigma^2} (X^T X + n\lambda I_m). \quad (2.73)$$

Matrix  $I$  becomes

$$\begin{aligned} I &= E_{W,Z} \left\{ \frac{\partial}{\partial b} \log f_\lambda^* \frac{\partial}{\partial b^T} \log f_\lambda^* \right\} - E_{W,Z} \left\{ \frac{\partial}{\partial b^T} \log f_\lambda^* \right\} E_{W,Z} \left\{ \frac{\partial}{\partial b^T} \log f_\lambda^* \right\} \\ &= \frac{1}{\sigma^4} E_{W,Z} \left\{ W (Z - W^T b^*)^2 W^T \right\} \end{aligned} \quad (2.74)$$

and is estimated as

$$\hat{I} = \frac{1}{\sigma^4 n} \sum_{i=1}^n X_i (Y_i - X_i^T \hat{b})^2 X_i^T. \quad (2.75)$$

The RPSM becomes

$$\hat{\lambda}_{\text{MELL}} = \arg \min_{\lambda} \left\{ \frac{1}{n} \|Y - X \hat{b}_\lambda\|^2 + \frac{2\sigma^2}{n} \text{trace}(\hat{I} \hat{J}^{-1}) \right\}. \quad (2.76)$$

This RPSM uses the Gaussian model but does not assume that the conditional mean is correctly specified. That means the choice of the regularization parameter value remains consistent even if a functional misspecification is present, i.e., when  $m(x) \equiv E\{Y_i | X_i = x\} \neq x^T b$  for any parameter  $b \in R^m$ .

As mentioned already, distributional misspecification does not affect the estimation of the location parameter  $b$ . However, when an estimate of the covariance matrix of the MLE or MPLE is needed, an estimator that is consistent under distributional misspecification must be used because the usual covariance matrix estimators are not consistent under distributional misspecification. To account for possible distributional misspecifications, the estimation of  $\sigma^2$ , treated so far as a nuisance parameter, must be considered. This allows one to account for a nonzero skewness and kurtosis in the response variable  $Z|W$ .

With a limited number of observations, the inaccuracy penalization in Eq. (2.76) becomes inadequate, and further refinement is needed. Starting from Eq. (2.76) and using Bozdogan's [[bozdogan96a](#), [bozdogan96b](#)] refinement argument, we obtain an information complexity regularization parameter selection (abbreviated as ICOMPRPS) method that behaves favorably for a limited number of observations.

Notice that the term  $\text{trace}(\hat{I} \hat{J}^{-1})$  in Eq. (2.76) can be interpreted as the effective number of parameters of a possibly misspecified model. ICOMPRPS also penalizes the interdependency between the parameter estimates. ICOMPRPS imposes a more severe penalization of estimation inaccuracy caused by the fact that the data-generating distribution is unknown.

For the MPLE method, the ICOMPRPS has the form [urmanov02]

$$\text{ICOMPRPS}(\lambda) \equiv \text{ALL}(\hat{b}_\lambda) - \frac{1}{n} \text{trace}(\hat{J} \hat{J}^{-1}) - \frac{1}{n} C_1(\hat{J}^{-1}), \quad (2.77)$$

and the corresponding RPSM is

$$\hat{\lambda}_{\text{ICOMPRPS}} = \arg \max_{\lambda} \left\{ \text{ALL}(\hat{b}_\lambda) - \frac{1}{n} \text{trace}(\hat{J} \hat{J}^{-1}) - \frac{1}{n} C_1(\hat{J}^{-1}) \right\}, \quad (2.78)$$

where  $C_1$  is the maximal covariance complexity index proposed by van Emden [emden71] to measure the degree of interdependency between parameter estimates.  $C_1$  is a function of a covariance matrix and is computed as in Eq. (2.79) using the eigenvalues of the covariance matrix. Notice that the more ill conditioned the data matrix  $X$ , the more dependent the parameter estimates become; therefore, the covariance complexity can be used to quantify ill conditioning.

Under the assumption that the vector of parameter estimates  $\hat{b}_\lambda$  is approximately normally distributed, the maximal covariance complexity reduces to

$$C_1(\hat{J}^{-1}) = \frac{m}{2} \log \frac{\bar{v}_a}{\bar{v}_g}, \quad (2.79)$$

where  $\bar{v}_a = \frac{1}{m} \sum_{j=1}^m v_j$ ,  $\bar{v}_g = \left( \prod_{j=1}^m v_j \right)^{\frac{1}{m}}$ , and  $v_j$  are the eigenvalues of  $\hat{J}^{-1}$ .

In the Gaussian case, ICOMPRPS for correctly specified models (abbreviated as ICOMPRPSCM) becomes

$$\text{ICOMPRPSCM}(\lambda) = \frac{1}{n} \|Y - X \hat{b}_\lambda\|^2 + \frac{2\sigma^2}{n} \left( \text{trace}(H) + C_1(\hat{J}^{-1}) \right), \quad (2.80)$$

and the corresponding RPSM is

$$\hat{\lambda}_{\text{ICOMPRPSCM}} = \arg \min_{\lambda} \left\{ \frac{1}{n} \|Y - X \hat{b}_\lambda\|^2 + \frac{2\sigma^2}{n} \left( \text{trace}(H) + C_1(\hat{J}^{-1}) \right) \right\}, \quad (2.81)$$

where

$$\hat{J} = X^T X + n\lambda I_m \quad \text{and} \quad (2.82)$$

$$H = X(X^T X + n\lambda I_m)^{-1} X^T. \quad (2.83)$$

There is a strong bond between the RPSMs based on maximizing the mean expected log likelihood and minimizing the mean predictive error. Namely, if the parametric family of approximating distributions (the model) is Gaussian,

$$f(Y_i|X_i; b) \equiv N(X_i^T b, \sigma^2), \quad (2.84)$$

then maximizing the MELL is equivalent to minimizing the MPE. This fact allows us to write an MPE analog of the information criterion (Eq. 2.48). Indeed, using the Gaussian model, the ALL can be written as the sum of the training error (abbreviated TE) and a constant term

$$\begin{aligned} \text{ALL}(\hat{b}_\lambda) &= \frac{1}{n} \sum_{i=1}^n \log f(X_i, Y_i | \hat{b}_\lambda) \\ &= \frac{1}{n} \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(Y_i - X_i^T \hat{b}_\lambda)^2\right) \\ &= \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{n2\sigma^2} \sum_{i=1}^n (Y_i - X_i^T \hat{b}_\lambda)^2 \\ &= \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \text{TE}(\hat{b}_\lambda), \end{aligned} \quad (2.85)$$

where the training error is defined as

$$\text{TE}(\hat{b}_\lambda) \equiv \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^T \hat{b}_\lambda)^2. \quad (2.86)$$

The expected log likelihood for the Gaussian model is

$$\begin{aligned} \text{ELL}(\hat{b}_\lambda) &= E_{W,Z} \log f(W, Z | \hat{b}_\lambda) = E_{W,Z} \left\{ \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(Z - W^T \hat{b}_\lambda)^2\right) \right\} \\ &= \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} E_{W,Z} \left\{ (Z - W^T \hat{b}_\lambda)^2 \right\} \\ &= \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} E_{W,Z} \left\{ (Z - m(W))^2 \right\} \\ &\quad - \frac{1}{2\sigma^2} E_W \left\{ (m(W) - W^T \hat{b}_\lambda)^T (m(W) - W^T \hat{b}_\lambda) \right\} \\ &= \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2} - \frac{1}{2\sigma^2} \text{PE}(\hat{b}_\lambda), \end{aligned} \quad (2.87)$$

where the predictive error is defined as

$$\text{PE}(\hat{b}_\lambda) \equiv E_W \left\{ (m(W) - W^T \hat{b}_\lambda)^T (m(W) - W^T \hat{b}_\lambda) \right\}. \quad (2.88)$$

Plugging these representations into Eq. (2.48), an MPE analog of the information RPSM is obtained. The mean predictive error is approximated as

$$E_D \text{PE}(\hat{b}_\lambda) \approx E_D \text{TE}(\hat{b}_\lambda) + \frac{2\sigma^2}{n} \text{trace}(\hat{I} \hat{J}^{-1}) - \sigma^2. \quad (2.89)$$

Therefore, an unbiased estimator of the MPE is given by

$$T_{\text{MPE}}(\lambda) \equiv \text{TE}(\hat{b}_\lambda) + \frac{2\sigma^2}{n} \text{trace}(\hat{I}\hat{J}^{-1}) - \sigma^2, \quad (2.90)$$

and the corresponding RPSM is

$$\hat{\lambda}_{\text{MPE}} = \arg \min_{\lambda} \left\{ \text{TE}(\hat{b}_\lambda) + \frac{2\sigma^2}{n} \text{trace}(\hat{I}\hat{J}^{-1}) - \sigma^2 \right\}. \quad (2.91)$$

Therefore, when the Gaussian model is used, the MELL and MPE have the same minimizer. When the model is correctly specified,  $\text{trace}(\hat{I}\hat{J}^{-1}) = \text{trace}(H)$ , and the CL method follows as

$$\text{CL}(\lambda) = \text{TE}(\hat{b}_\lambda) + \frac{2\sigma^2}{n} \text{trace}(H) - \sigma^2, \quad (2.92)$$

with the corresponding RPSM

$$\hat{\lambda}_{\text{CL}} = \arg \min_{\lambda} \left\{ \text{TE}(\hat{b}_\lambda) + \frac{2\sigma^2}{n} \text{trace}(H) - \sigma^2 \right\}. \quad (2.93)$$

We now present an example of an image reconstructed using the information approach to the selection of regularization parameter. Notice that in this example of a correctly specified model, the  $\text{C}_L$  and ICOMPERS selected identical parameters as expected from theoretical derivations.

## References

- [akaike73] H. Akaike, Information theory and an extension of the maximum likelihood principle, in *2nd International Symposium on Information Theory*, ed. by B.N. Petrov, F. Csaki (Akademiai Kiado, Budapest, 1973), pp. 267–281
- [bozdogan87] H. Bozdogan, Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* **52**(3), 345–370 (1987)
- [bozdogan88] H. Bozdogan, ICOMP: a new model selection criterion, in *Classification and Related Methods of Data Analysis*, ed. by Hans H. Bock (Elsevier Science Publishers B.V. (North-Holland), Amsterdam, 1988), pp. 599–608
- [bozdogan96a] H. Bozdogan, A new informational complexity criterion for model selection: the general theory and its applications, in *Information Theoretic Models and Inference (INFORMS)*, Washington D.C., 5–8 May 1996
- [bozdogan96b] H. Bozdogan, Informational complexity criteria for regression models, in *Information Theory and Statistics Section on Bayesian Stat. Science*, ASA Annual Meeting, Chicago, IL, 4–8 August 1996
- [engl00] H.W. Engl, M. Hanke, A. Neubauer, *Regularization of Inverse Problems* (Kluwer Academic, Dordrecht, 2000)

- [golub97] G.H. Golub, U. von Matt, Tikhonov regularization for large scale problems. Technical report SCCM-97-03, Stanford University, 1997
- [hansen93] P.C. Hansen, D.P. O’Leary, The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.* **14**, 1487–1503 (1993)
- [hansen94] P.C. Hansen, Regularization tools: a Matlab package for analysis and solution of discrete ill-posed problems. *Numer. Algorithms* **6**, 1–35 (1994)
- [hansen98] P.C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*. SIAM Monographs on Mathematical Modeling and Computation (SIAM, Philadelphia, 1998)
- [huber81] P.J. Huber, *Robust Statistics* (Wiley, New York, 1981)
- [konishi96] S. Konishi, G. Kitagawa, Generalized information criteria in model selection. *Biometrika* **83**(4), 875–890 (1996)
- [kullback51] S. Kullback, R.A. Leibler, On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951)
- [leonov97] A.S. Leonov, A.G. Yagola, The L-curve method always introduces a non removable systematic error. *Mosc. Univ. Phys. Bull.* **52**(6), 20–23 (1997)
- [mallows73] C.L. Mallows, Some comments on CP. *Technometrics* **15**(4), 661–675 (1973)
- [morozov84] V.A. Morozov, *Methods for Solving Incorrectly Posed Problems* (Springer, New York, 1984)
- [phillips62] D.L. Phillips, A technique for the numerical solution of certain integral equations of the first kind. *JACM* **9**, 84–97 (1962)
- [sakamoto86] Y. Sakamoto, *Akaike Information Criterion Statistics* (KTK Scientific publishers, Tokyo, 1986)
- [shibata89] R. Shibata, Statistical aspects of model selection, in *From Data to Model*, ed. by J.C. Willems (Springer, New York, 1989), pp. 215–240
- [urmanov02] A.M. Urmanov, A.V. Gribok, H. Bozdogan, J.W. Hines, R.E. Uhrig, Information complexity-based regularization parameter selection for solution of ill-conditioned inverse problems. *Inverse Prob.* **18**, L1–L9 (2002)
- [emden71] M.H. van Emden, An analysis of complexity, in *Mathematical Centre Tracts*, vol. 35 (Mathematisch Centrum, Amsterdam, 1971)
- [vogel96] C.R. Vogel, Non-convergence of the L-curve regularization parameter selection method. *Inverse Prob.* **12**, 535–547 (1996)
- [vogel02] C.R. Vogel, *Computational Methods for Inverse Problems*. SIAM, Frontiers in Applied Mathematics Series, vol 23 (SIAM, Philadelphia, 2002)
- [wahba90] G. Wahba, *Spline Models for Observational Data* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990)



<http://www.springer.com/978-3-319-46363-6>

Optimization Techniques in Computer Vision

Ill-Posed Problems and Regularization

Abidi, M.A.; Gribok, A.V.; Paik, J.

2016, XV, 293 p. 127 illus., 23 illus. in color., Hardcover

ISBN: 978-3-319-46363-6