INSTITUTE OF PHYSICS PUBLISHING INVERSE PROBLEMS

Inverse Problems **18** (2002) L1–L9 PII: S0266-5611(02)29622-7

LETTER TO THE EDITOR

Information complexity-based regularization parameter selection for solution of ill conditioned inverse problems

A M Urmanov[1], A V Gribok[1], H Bozdogan[1,2], J W Hines[1] and R E Uhrig[1]

[1] The University of Tennessee Nuclear Engineering Department, Knoxville, TN 37996, USA
[2] The University of Tennessee Statistics Department, Knoxville, TN 37996, USA

E-mail: urmanov@utk.edu, agribok@utk.edu, bozdogan@utk.edu, hines@utkux.utk.edu and ruhrig@utk.edu

Received 5 October 2001
Published 25 February 2002
Online at stacks.iop.org/IP/18/L1

Abstract

We propose an information complexity-based regularization parameter selection method for solution of ill conditioned inverse problems. The regularization parameter is selected to be the minimizer of the *Kullback–Leibler (KL) distance* between the unknown data-generating distribution and the fitted distribution. The *KL distance* is approximated by an *information complexity criterion* developed by Bozdogan. The method is not limited to the white Gaussian noise case. It can be extended to correlated and non-Gaussian noise. It can also account for possible model misspecification. We demonstrate the performance of the proposed method on a test problem from Hansen's regularization tools.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

In a statistical framework, one solves an ill conditioned inverse problem $y = Ax + \varepsilon$, with $A \in R^{n\times m}$, $y \in R^n$, $x \in R^m$ and $n \geqslant m$, by specifying a parametric family of conditional probability distributions with densities $f(y|A, x)$, choosing a *penalty* $p(x)$ that assesses the physical plausibility of the solution, choosing the value of the *regularization parameter* α, and by using the *maximum penalized likelihood (MPL)* method

$$\hat{x}_\alpha = \arg\max_x\{pl(x)\}, \qquad \text{where } pl(x) = \log f(y|A, x) - \alpha^2 p(x). \tag{1.1}$$

For a detailed description of the properties of the *MPL* method, see Cox and O'Sullivan (1990). If we consider uncorrelated Gaussian noise with variance σ^2 and use the penalty $p(x) = \frac{1}{2\sigma^2}\|\Omega x\|_2^2$, the MPL solution is the same as the Tikhonov (1963) regularized solution

$$\hat{x}_{T,\alpha} = \arg\min_x\{\|Ax - y\|_2^2 + \alpha^2\|\Omega x\|_2^2\}.$$

The success of currently available regularization techniques relies heavily on the proper choice of the regularization parameter. Although many methods 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 have found their way into engineering practice, the most common are the *discrepancy principle (DP)* (Morozov 1984, Phillips 1962), Mallows' (1973) *CL*, *generalized cross validation (GCV)* (Wahba 1990) and the *L-curve method* (Hansen 1998). A high sensitivity of *CL* and *DP* to an underestimation of the noise level has limited their application to cases in which the noise level can be estimated with high fidelity (Hansen 1998). On the other hand, noise-estimate-free *GCV* occasionally fails, presumably due to the presence of correlated noise (Wahba 1990). The *L-curve* method is widely used; however, this method is nonconvergent (Vogel 1996, Leonov and Yagola 1997). From this standpoint, the search for new methods to choose a valid regularization parameter is well justified from both theoretical and practical points of view.

We approach the regularization parameter selection problem in the statistical model selection framework. Given a family of conditional probability density functions $f(y|A, x)$ where x is a vector of parameters, we choose the regularization parameter value corresponding to the density function from the specified family that matches the unknown data-generating density function most closely.

When the unknown parameters are estimated by the MPL method (1.1), each particular choice of the smoothing operator Ω and regularization parameter α yields some fitted density $f(y|A, \hat{x}_\alpha)$. The closeness of any two probability densities g and f can be evaluated by the Kullback and Leibler (1951) *(KL) distance (or information)* that measures the divergence between the densities

$$I(g(z);\, f(z)) = E_G\{\log g(z)\} - E_G\{\log f(z)\}.$$

The regularization parameter is selected to be the minimizer of the *KL distance* between the unknown data-generating density $g(y|A)$ and the fitted density $f(y|A, \hat{x}_\alpha)$

$$\hat{\alpha} = \arg\min_\alpha\{I(g(y|A);\, f(y|A, \hat{x}_\alpha))\} = \arg\min_\alpha\{E_G\{\log g(y|A)\} - E_G\{\log f(y|A, \hat{x}_\alpha)\}\}.$$

From the definition of the *KL distance* we see that its minimization is equivalent to maximization of the *expected log likelihood* $E_G\{\log f(y|A, \hat{x}_\alpha)\}$. In statistical learning theory, the *expected log likelihood* is also known as the *predictive risk*. In the Gaussian case, maximization of the expected log likelihood is equivalent to minimization of the *expected squared predictive error*. Since the data-generating distribution $G(y|A)$ is unknown, the empirical distribution is used instead, and the expected log likelihood is estimated by the mean log likelihood $\frac{1}{n}\sum_{i=1}^{n} \log f(y_i|A_i, \hat{x}_\alpha)$. It is well known (see e.g. Akaike 1973, Bozdogan 1987, Konishi and Kitagawa 1996) that such an approximation introduces a *bias* in estimating the expected log likelihood that should be corrected. Starting with the pioneering work of Akaike (1973), a number of bias-corrected information criteria that estimate the expected log likelihood have been proposed (Takeuchi 1976, Schwaz 1978, Bozdogan 1987, 2000, Shibata 1989, Konishi and Kitagawa 1996). We use the results of Takeuchi (1976) and Konishi and Kitagawa (1996) to calculate an *unbiased estimate* of the *expected log likelihood* for the *penalized maximum likelihood* estimation case and apply Bozdogan's (1996) refinement argument to arrive at *information complexity (ICOMP)* criteria for choosing the regularization parameter value.

2. Information complexity-based regularization parameter selection

As discussed by Shibata (1989), Konishi and Kitagawa (1996) and Bozdogan (2000), when we assume that the true distribution may not belong to the specified family, the asymptotic

bias of the mean log likelihood in the estimation of the expected log likelihood is given by

$$\text{Bias} = E_G\left\{E_G\{\log f(y|A,\hat{x}_\alpha)\} - \frac{1}{n}\sum_{i=1}^{n}\log f(y_i|A_i,\hat{x}_\alpha)\right\} = \frac{1}{n}b_1(G) + \text{o}\left(\frac{1}{n}\right)$$

where $b_1(G) = \text{tr}\,\{F_\alpha^{-1}R_\alpha\}$, and matrices F_α and R_α, following the notation of Bozdogan (2000), are defined as

$$F_\alpha = -E_G\left\{\left.\frac{\partial^2 pl(x)}{\partial x\,\partial x^T}\right|_{\hat{x}_\alpha}\right\} \qquad \text{and} \qquad R_\alpha = E_G\left\{\left.\frac{\partial pl(x)}{\partial x}\frac{\partial f(y|A,x)}{\partial x^T}\right|_{\hat{x}_\alpha}\right\}.$$

In the literature, $\text{tr}\,\{F_\alpha^{-1}R_\alpha\}$ is also known as the *Lagrange-multiplier test statistic*.

In the white Gaussian noise case, with correct model specification and σ^2 treated as a nuisance parameter, $b_1(G)$ reduces to the trace of the *hat* or *influence matrix*, $b_1 = \sum_{i=1}^{n} h_{ii}$, or to the *effective number of parameters*. The effective number of parameters, also known as the *effective rank*, was introduced by Gilliam *et al* (1990). In the same situation but with possible model misspecification the bias is given by $b_1 = \frac{1}{\sigma^2}\sum_{i=1}^{n} h_{ii}\hat{e}_i^2$, where $\hat{e}_i^2$ are the maximum likelihood ($\alpha = 0$) residuals (Shibata 1989). When the estimation is done by the maximum likelihood method, $\text{tr}\,\{F_0^{-1}R_0\}$ reduces to the number of parameters, yielding Akaike's information criterion (AIC) (Akaike 1973).

Notice that the asymptotic bias is of order $1/n$ and should be estimated by using the empirical distribution. For a small number of data points, an error in bias estimation can be very significant (see e.g. Konishi and Kitagawa 1996) so that the bias estimate should be refined. As argued by Bozdogan (1987, 1996), in many statistical model selection situations considering only the number of parameters is inadequate. He developed an ICOMP framework that can be used to refine the information criteria that use only the number of parameters. The information criterion that also penalizes the *interdependence between the parameter estimates* is termed the *ICOMP* criterion (Bozdogan 1996, 2000). *ICOMP* criteria impose a more severe penalization of estimation inaccuracy caused by the fact that the data-generating distribution is unknown, by taking into account interdependences between parameter estimates. For the *MPL* estimation method, the *ICOMP* criterion estimates a minus 2*n expected log likelihood* and has the following form:

$$ICOMP(\alpha) = -2\log f(y|A,\hat{x}_\alpha) + 2\,\text{tr}\,\{F_\alpha^{-1}R_\alpha\} + 2C_1(F_\alpha^{-1}) \tag{2.1}$$

where C_1 is the *maximal covariance complexity index* proposed by van Emden (1971) to measure the degree of interdependence between parameter estimates. Notice that the more ill conditioned the data matrix A, the more dependent the parameter estimates become, and therefore the covariance complexity can be used to quantify ill-conditioning. Under the assumption that the vector of parameter estimates $\hat{x}$ is approximately normally distributed, the maximal covariance complexity reduces to

$$C_1(F_\alpha^{-1}) = \frac{m}{2}\log\frac{\bar{\nu}_a}{\bar{\nu}_g} \qquad \text{where } \bar{\nu}_a = \frac{1}{m}\sum_{j=1}^{m}\nu_j,\ \bar{\nu}_g = \left(\prod_{j=1}^{m}\nu_j\right)^{1/m},$$

and ν_j are the singular values of F_α^{-1}. The regularization parameter is chosen as the minimizer of *ICOMP*

$$\hat{\alpha}_{ICOMP} = \arg\min_{\alpha}\{\,ICOMP(\alpha)\}. \tag{2.2}$$

For uncorrelated Gaussian noise, $\varepsilon \sim N(0,\sigma^2 I)$, quadratic penalty, $p(x) = (1/2\sigma^2)x^T x$, the MPL solution is given by $\hat{x}_\alpha = (A^T A + \alpha^2 I)^{-1}A^T y$. In this situation, the *ICOMP* criterion takes the form

$$ICOMP(\alpha) = \frac{1}{\hat{\sigma}^2}\|A\hat{x}_\alpha - y\|_2^2 + 2\,\text{tr}\,\{H_\alpha\} + 2C_1(F_\alpha^{-1}) \tag{2.3}$$

where $F_\alpha^{-1} = (A^T A + \alpha^2 I)^{-1}$ and $H_\alpha = A(A^T A + \alpha^2 I)^{-1} A^T$ is the hat matrix that defines the effective number of parameters as $k_{eff} = \mathrm{tr}\,\{H_\alpha\}$.

Notice that Mallows' (1973) *CL* given by

$$CL(\alpha) = \frac{1}{\hat{\sigma}^2}\|A\hat{x}_\alpha - y\|_2^2 + 2\,\mathrm{tr}\,\{H_\alpha\}$$

has the same bias-corrected form as the information criterion. Therefore, CL can be considered as a bias-corrected information criterion for the white Gaussian noise case with a correctly specified model.

3. Simulation results

To demonstrate the proposed method we apply the ICOMP criterion (2.3) to select the regularization parameter value for solution of the Fredholm integral equation of the first kind in the discretized form

$$\int_a^b K(s,t)\,f(t)\,\mathrm{d}t \approx I_n(s) = \sum_{i=1}^{n} w_i K(s,t_i)\,f(t_i),$$

$$\left.\begin{array}{l} I_n(s) \to y_s \\ w_i K(s,t_i) \to A_{s,i} \\ f(t_i) \to x_i \end{array}\right\} \to \underline{y = Ax + \varepsilon}\,.$$

In particular, the Phillips problem is considered from Hansen's regularization tools toolbox (Hansen 1994). This problem is mildly ill conditioned (the condition number is 10^5 for $n = 64$) and is known to have a regularized solution for the smoothing operator $\Omega = I$.

We performed simulations with different noise levels and compared the regularization parameter values chosen by *DP*, *CL*, and *ICOMP* criteria against the optimal value that minimizes $\|x_{true} - \hat{x}_\alpha\|_2$. Since the optimal value depends on a particular realization of the noise vector, it is also a random value. Estimated probability densities of the chosen parameters for a given noise level (approximately 1%) are shown in figure 1. The solid curve represents the optimal parameter. The variation of the regularization parameter chosen by *ICOMP* is smaller than those chosen by *CL* and *DP*. *CL* consistently underestimates the regularization parameter value in this problem. This may be due to poor bias approximation for such a small number of observations.

Notice that the above result is for the case in which we know the exact noise level. On the other hand, if we overestimate or underestimate the noise level, the performance of the criteria changes drastically. In figures 2 and 3 the estimated densities for the regularization parameter values are shown for a noise level that is overestimated and underestimated by 50%. In the case of overestimation, the chosen value is biased but its variance is reduced.

A difficult situation for the criteria is underestimation. *DP* and *CL* are known to be very sensitive to underestimation, and that is demonstrated in figure 3. The chosen values are scattered all over the possible range, making any choice meaningless. However, *ICOMP* can sustain such underestimation and still provide a very good regularization parameter value.

We also looked at the performance for different noise levels, as shown in figure 4. For each noise level the shown value is the average over ten noise realizations.

With an exact noise level, *DP* provides the average value closest to the optimal one. *ICOMP* and *CL* tend to underestimate the regularization parameter for all noise levels. However if we again overestimate or underestimate the noise level by 50%, the performance gets much worse, as shown in figures 5 and 6.

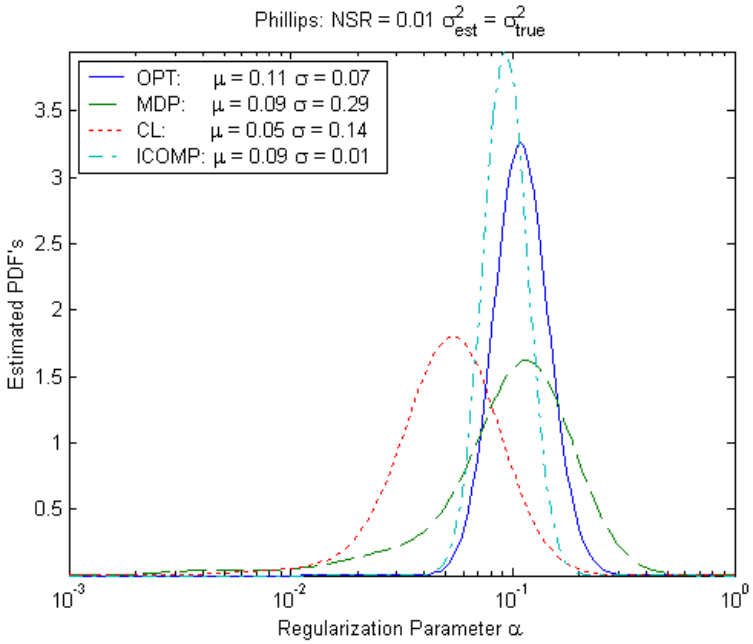

Figure 1. Estimated distributions of the regularization parameter selected by different methods with known noise level.

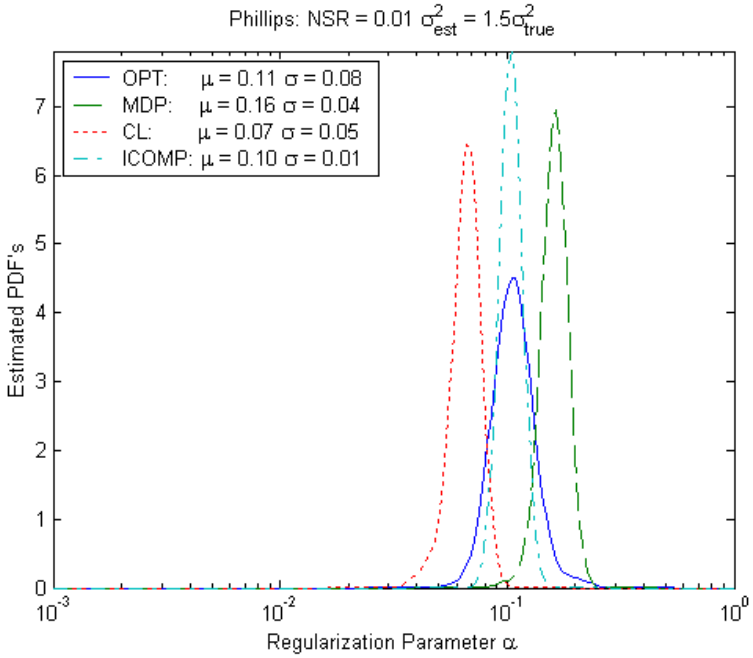

Figure 2. Estimated distributions of the regularization parameter selected by different methods with overestimated noise level.

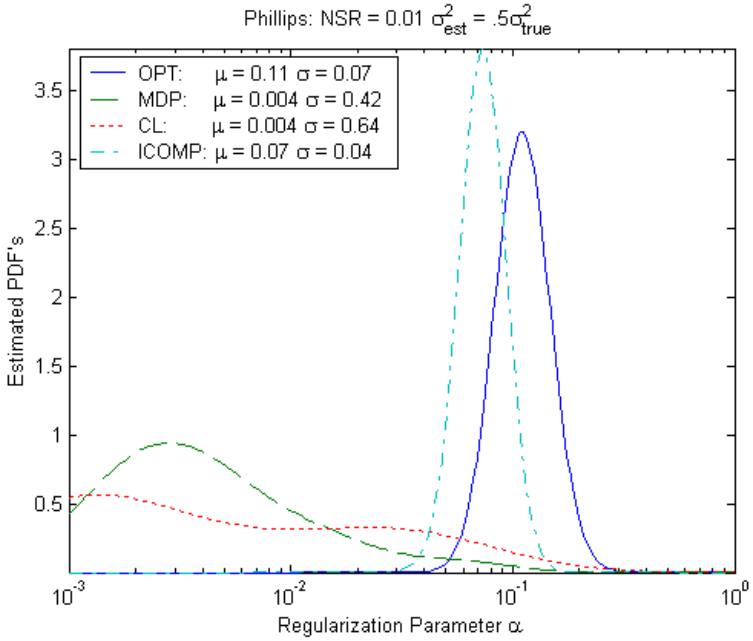

Figure 3. Estimated distributions of the regularization parameter selected by different methods with underestimated noise level.

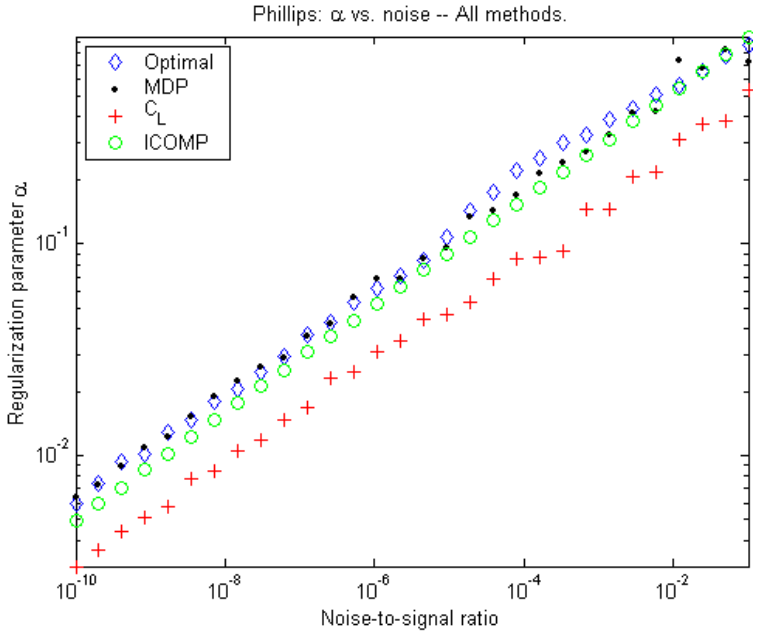

Figure 4. Regularization parameter values selected by different methods for different noise levels.

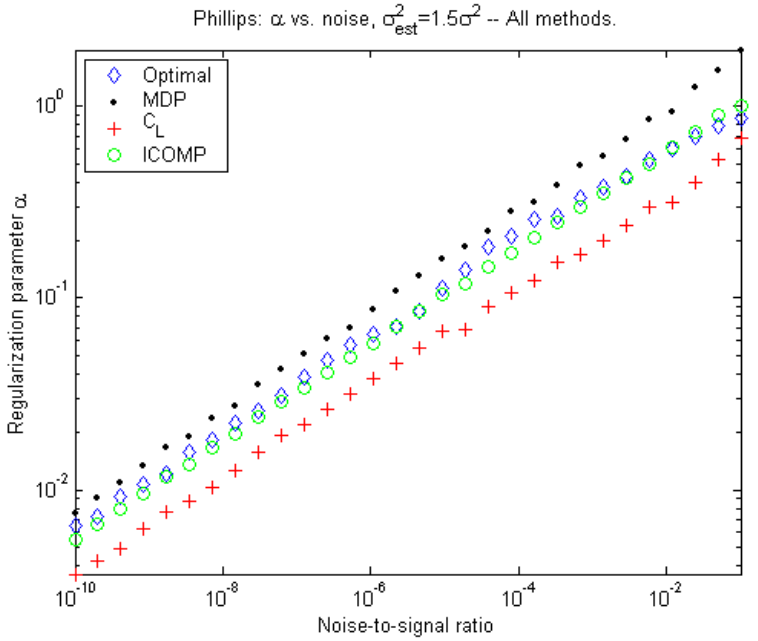

Figure 5. Regularization parameter values selected by different methods for different noise levels with overestimated noise level.

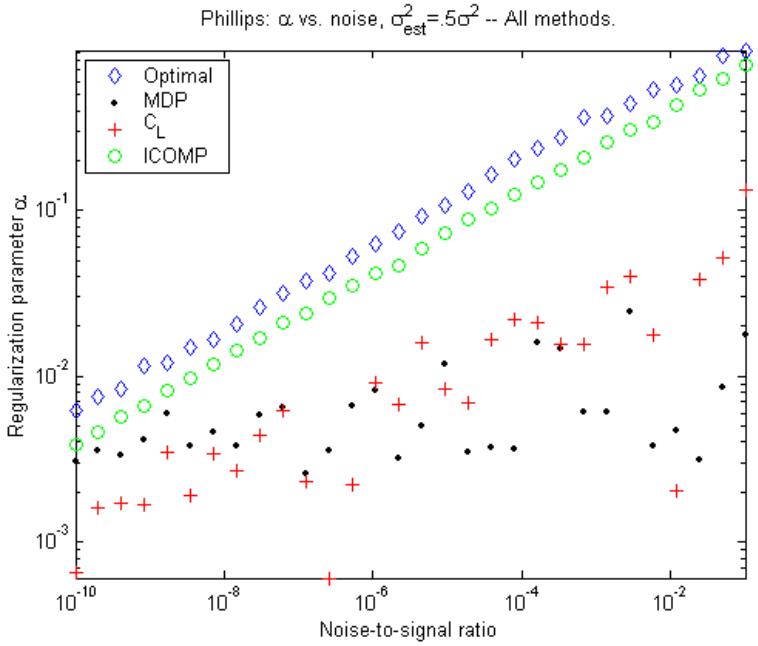

Figure 6. Regularization parameter values selected by different methods for different noise levels with underestimated noise level.

The underestimation is a killing condition for *DP* and *CL*, whereas *ICOMP* still does a good job. This may serve as an illustration that the bias estimated solely by the number of parameters is grossly underestimated in situations with a small number of observations and should be refined. *ICOMP* suggests one possible refinement by accounting for interdependences between the parameter estimates.

4. Conclusions

We have proposed an ICOMP-based method for the regularization parameter selection. The method is not limited to the case of white Gaussian noise, but can be adjusted to correlated and non-Gaussian noise. *CL* can be considered similar to the proposed method under very restrictive conditions that limit its practical application. The method is demonstrated to outperform *DP* and *CL* in the very important case of the underestimated noise level.

The random nature of noise introduces some variability in the chosen parameter. Depending on the particular problem and the degree of ill-conditioning, the sensitivity of the solution to the parameter value can vary drastically. Therefore, it is useful to have some means to evaluate the variability of the chosen parameter, because given only one set of data the chosen parameter may lie far away from the optimal one. Having knowledge about the precision of the parameter estimate allows us to study the sensitivity of the solution by varying the parameter in the range of its standard deviation. If the solution is fairly stable in this range, we can consider the problem solved; if not, further work is obviously needed.

References

Akaike H 1973 Information theory and an extension of the maximum likelihood principle *2nd Int. Symp. on Information Theory* ed B N Petrov and F Csaki (Budapest: Akademiai Kiado) pp 267–81

Bozdogan H 1987 Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions *Psychometrika* **52** 345–70

Bozdogan H 1988 ICOMP: A new model selection criterion *Classification and Related Methods of Data Analysis* ed H H Bock (Amsterdam: Elsevier) pp 599–608

Bozdogan H 1990 On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models *Commun. Stat. Theory Methods* **19** 221–78

Bozdogan H 1994 Mixture model cluster analysis using model selection criteria and a new informational measure of complexity *Multivariate Statistical Modeling* vol 2, ed H Bozdogan (Dordrecht: Kluwer) pp 69–113

Bozdogan H 1996 A new informational complexity criterion for model selection: the general theory and its applications *Information Theoretic Models and Inference, INFORMS (Washington, DC, 1996)*

Bozdogan H 2000 Akaike's information criterion and recent developments in information complexity *J. Math. Psychol.* **44** 62–91

Cox D D and O'Sullivan F 1990 Asymptotic analysis of penalized likelihood and related estimators *Ann. Stat.* **18** 1676–95

Gilliam D S, Lund J R and Vogel C R 1990 Quantifying information content for ill-posed problems *Inverse Problems* **6** 725–36

Hansen P C 1994 Regularization tools: a Matlab package for analysis and solution of discrete ill-posed problems *Numer. Algorithms* **6** 1–35

Hansen P C 1998 Rank-deficient and discrete ill-posed problems *SIAM Monographs on Mathematical Modeling and Computation* (Philadelphia, PA: SIAM)

Konishi S and Kitagawa G 1996 Generalized information criteria in model selection *Biometrika* **83** 875–90

Kullback S and Leibler R A 1951 On information and sufficiency *Ann. Math. Stat.* **22** 79–86

Leonov A S and Yagola A G 1997 The L-curve method always introduces a nonremovable systematic error *Moscow Univ. Phys. Bull.* **52** 20–3

Mallows C L 1973 Some comments on C_P *Technometrics* **15** 661–75

Morozov V A 1984 *Methods for Solving Incorrectly Posed Problems* (New York: Springer)

Phillips D L 1962 A technique for the numerical solution of certain integral equations of the first kind *J. Assoc. Comput. Mach.* **9** 84–97

Schwaz R 1978 Estimating the dimension of a model *Ann. Stat.* **6** 461–4

Shibata R 1989 Statistical aspects of model selection *From Data to Models* ed J C Willems (New York: Springer) pp 215–40

Takeuchi K 1976 Distribution of information statistics and a criterion of model fitting *Suri-Kagaku (Mathematical Sciences)* **153** 12–18 (in Japanese)

Tikhonov A N 1963 Solution of incorrectly formulated problems and regularization method *Sov. Math.–Dokl.* **4** 1035–8

van Emden M H 1971 *An Analysis of Complexity (Mathematical Centre Tracts vol 35)* (Amsterdam: Mathmatisch Centrum Amsterdam)

Vogel C R 1996 Non-convergence of the L-curve regularization parameter selection method *Inverse Problems* **12** 535–47

Wahba G 1990 *Spline Models for Observational Data* (Philadelphia, PA: SIAM)