8-2002

An Information Approach to Regularization Parameter Selection for the Solution of Ill-Posed Inverse Problems Under Model Misspecification

Aleksey M. Urmanov
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a dissertation written by Aleksey M. Urmanov entitled "An Information Approach to Regularization Parameter Selection for the Solution of Ill-Posed Inverse Problems Under Model Misspecification." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Nuclear Engineering.

Robert E. Uhrig, Major Professor

We have read this dissertation and recommend its acceptance:

J. Wesley Hines, Hamparsum Bozdogan, Peter Groer, Belle R. Upadhyaya, Andrei Gribok

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a dissertation written by Aleksey M. Urmanov entitled "An Information Approach to Regularization Parameter Selection for the Solution of Ill-Posed Inverse Problems Under Model Misspecification." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Nuclear Engineering.

Robert E. Uhrig, Major Professor

We have read this dissertation
and recommend its acceptance:

J. Wesley Hines

Hamparsum Bozdogan

Peter Groer

Belle R. Upadhyaya

Andrei Gribok

Accepted for the Council:

Anne Mayhew
Vice Provost and Dean of
Graduate Studies

(Original signatures are on file with official student records.)

AN INFORMATION APPROACH TO REGULARIZATION PARAMETER SELECTION FOR THE SOLUTION OF ILL-POSED INVERSE PROBLEMS UNDER MODEL MISSPECIFICATION

A Dissertation
Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Aleksey M. Urmanov
August 2002

ACKNOWLEDGMENTS

I wish to thank all those who helped me in completing the Doctor of Philosophy in Nuclear Engineering. I thank Dr. Uhrig and Dr Gribok for scientific supervision and for introducing me to the field of ill-posed inverse problems in engineering. I also wish to thank Dr. Hines, Dr. Groer and Dr. Upadhyaya for helpful discussions and useful criticism. I would like to thank Dr. Dodds for fostering a friendly and creative environment in the department.

I would like to acknowledge Dr. Bozdogan for introducing me to the field of information approaches to statistical model selection and to his information complexity framework of modeling. I also would like to thank Dr. Curt Vogel of Montana State University for useful and encouraging discussions and for providing an image-deblurring test problem.

I would like to acknowledge Florida Power Corporation for providing the data for simulations, and Electric Power Research Institute and the Department of Defense for financial supporting of this work.

I thank my family and friends, whose suggestions and encouragement made this work possible.

ABSTRACT

Engineering problems are often ill-posed, i.e. cannot be solved by conventional data-driven methods such as parametric linear and nonlinear regression or neural networks. A method of regularization that is used for the solution of ill-posed problems requires an a priori choice of the regularization parameter. Several regularization parameter selection methods have been proposed in the literature, yet, none is resistant to model misspecification. Since almost all models are incorrectly or approximately specified, misspecification resistance is a valuable option for engineering applications.

Each data-driven method is based on a statistical procedure which can perform well on one data set and can fail on other. Therefore, another useful feature of a data-driven method is robustness. This dissertation proposes a methodology of developing misspecification-resistant and robust regularization parameter selection methods through the use of the information complexity approach.

The original contribution of the dissertation to the field of ill-posed inverse problems in engineering is a new robust regularization parameter selection method. This method is misspecification-resistant, i.e. it works consistently when the model is misspecified. The method also improves upon the information-based regularization parameter selection methods by correcting inadequate penalization of estimation inaccuracy through the use of the information complexity framework. Such an improvement makes the proposed regularization parameter selection method robust and reduces the risk of obtaining grossly underregularized solutions.

A method of misspecification detection is proposed based on the discrepancy between the proposed regularization parameter selection method and its correctly specified version. A detected misspecification indicates that the model may be inadequate for the particular problem and should be revised.

The superior performance of the proposed regularization parameter selection method is demonstrated by practical examples. Data for the examples are from Carolina Power & Light's Crystal River Nuclear Power Plant and a TVA fossil power plant. The results of applying the proposed regularization parameter selection method to the data demonstrate that the method is robust, i.e. does not produce grossly underregularized solutions, and performs well when the model is misspecified. This enables one to implement the proposed regularization parameter selection method in autonomous diagnostic and monitoring systems.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ALL	Average Log Likelihood
CM	Correctly specified Model
CL	Mallows' CL method
ELL	Expected Log Likelihood
EPLL	Expected Penalized Log Likelihood
FFR	Feedwater Flow Rate
GCV	Generalized Cross Validation
ICOMPRPS	Information COMPlexity Regularization Parameter Selection method
ICP	Information Criterion for Penalized models
KL	Kullback-Leibler (1951)
MDP	Morozov's Discrepancy Principle
MELL	Mean Expected Log Likelihood
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimator
MOR	Method Of Regularization
MPE	Mean Predictive Error
MPL	Maximum Penalized Likelihood
MPLE	Maximum Penalized Likelihood Estimator
MSE	Mean Square Error
OLS	Ordinary Least Squares
PR	Prediction Risk
PWR	Pressurized Water Reactor
RBF	Radial Basis Function neural network
RIC	Regularization Information Criterion
RPSM	Regularization Parameter Selection Method
SSR	Sum of Squared Residuals
SVD	Singular Value Decomposition
TE	Training Error
URE	Unbiased Risk Estimator

CHAPTER 1
INTRODUCTION

This dissertation is about data-driven modeling methods used for diagnostics, monitoring, and fault detection in industrial applications. The methodology developed in this dissertation falls into the field of statistical learning from data. The practical application of the methodology is for the solution of ill-posed inverse problems in engineering. This work was motivated by the lack of robust and resistant to model misspecification Regularization Parameter Selection Methods (abbreviated RPSM's) for the solution of ill-posed problems. This lack has limited the applicability of data-driven methods in the industry and indicated a necessity for developing robust and misspecification-resistant RPSM's which are more suitable for autonomous diagnostic and control systems. In this dissertation we develop a systematic way of constructing robust and misspecification-resistant RPSM's and demonstrate their superior performance.

1.1 Motivation

Data-driven modeling is widely used in industrial diagnostics and control. Based on results of modeling, the personnel make operational and safety-related decisions. Therefore, a modeling method that leaves the personnel to wonder whether the method produced a reasonable result or not is of no practical use. The majority of the latest data-driven methods used in diagnostics and surveillance are so complicated that only very few people, including the developers and specialists in the area, are aware of all the crucial conditions and assumptions under which the methods perform reliably or know

the meaning of the parameters to be tuned and their influence on the methods' performance. The users of the methods may have no knowledge of the fundamental limitations and theory behind the development and have no tools for assessing how reasonable the produced result is, unless, of course, it is so inappropriate that does not make common sense.

From the practical point of view, a method should provide the user with a result that can be taken without further analysis and used in decision-making. Since many decisions are not only operational but also safety related, they must be conservative. This means that a method should provide a result that is guaranteed to be either correct or conservative. There is no statistical method that can guarantee a correct result in an arbitrary situation. This is the downside of any statistical procedure. A statistical procedure can perform perfectly on one data set and fail miserably on another data set. Since data-driven methods employ statistical procedures, one must be very careful in implementing them.

Despite the downside of statistical procedures, the need for data-driven methods is beyond argument. They are of great value. In many situations, data-driven methods are the only option available. Therefore, the main focus in developing a data-driven method is to make sure that the statistical procedures it employs give either correct or conservative results. After all, the users may have no knowledge of the smoothing properties of the learning operator, and may not know that learning from data is possible only if the underlying relationship is smooth. They only design a model of an engineering system and want to use the model to make a correct decision. Users want automatically-tuned methods that choose the parameters they need, provide safe results, and can be manually tuned further to maximize economical benefits or to meet some specific goals.

Any data-driven modeling method that is going to be implemented in an industrial application should be resistant to any kind of violation of the assumptions and should

provide users with a safe result. It is well known that many methods use unrealistic assumptions that rarely occur in real life. Therefore, the main challenge is to develop methods that are resistant to violations of the assumptions and provide users with reasonable results. A "smart" method that gives users no result if the assumptions are violated and indicates that it does not work under the present conditions is of little value because the decision that was supposed to be made on the basis of the anticipated result still needs to be made. Moreover, some, if not all, modeling assumptions are usually violated to some degree; and thus smart methods would never work.

Resistance to assumption-violation as a valuable option is well recognized in such fields as econometrics in which one can rarely claim that a model is correctly specified. Incorrectly specified models are also common in engineering. After all, a model by definition gives an approximate description of the data-generating process under consideration. Since the actual data-generating process is usually unknown, the model can be easily misspecified. Misspecification means that the model cannot be tuned to describe the data-generating process exactly even when the process is known. The problem is to obtain a method that works properly and provides best possible approximations under model-misspecification. An example of usual assumptions is a well-conditioned data set with white Gaussian noise in the response. Not all real data sets can satisfy this assumption. Therefore, to be practically useful, any method derived under that assumption, because of the mathematical simplicity of the analysis, must be resistant to the violation of the assumption.

For example, a linear regression model provides the best (in the least squares sense) linear approximation to the nonlinear relationship. When the data set is ill-conditioned, ridge regression with a properly chosen regularization parameter value should be used to obtain a regularized (stable) solution. If the RPSM we use was derived assuming correct model specification, will it provide a proper regularization parameter

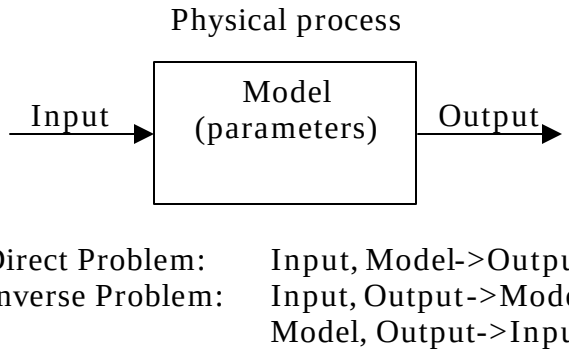

Figure 1.1. Direct and inverse problems in modeling.

value when the model is incorrect? Will we able to get a stable approximation in this case? These are practical questions that need to be answered. Therefore, a practically valuable method must behave properly under violation of the assumptions and consistently provide the best possible approximations when the model is misspecified.

In this dissertation we address the issue of developing a RPSM that is resistant to assumption-violation and that consistently provides reasonable and safe results. This method is an important part of any autonomous diagnostic and surveillance system that operates under real conditions and provides results that can be safely used for decision-making.

1.2 Modeling as an Ill-Posed Problem

Many problems solved in applied science and engineering are inverse problems. An inverse problem consists of finding unknown causes of known consequences. In contrast, solving a direct problem is finding unknown consequences of known causes. In terms of mathematical models of physical processes, the inverse problem illustrated in Figure 1.1 is to determine the model parameters, given the observed input and output. This problem is also known as the identification problem. The direct problem is to find the output of the model, given the input and the model parameters.

We name some examples of inverse problems and fields in which inverse problems are found:

- Nuclear transport
- Heat and mass transfer
- Fluid and solid mechanics
- Acoustics
- Electromagnetism
- Geophysics
- Vibrations and structural dynamics
- Inverse design
- Optimum experimental design
- System identification
- Sensor validation
- Restoration or deconvolution of signals in signal processing
- Signal deconvolution
- Evaluation of derivatives of a noisy signal
- Property estimation
- Imaging
- Image deblurring in astronomy
- Computed tomography
- Tomography and inverse scattering
- Statistical learning from data
- Artificial intelligence techniques
- Backward prior specification in Bayesian inference
- Retrospective reasoning in history
- Evolution theories.

Many inverse problems are ill-posed. Hadamard (1902) first introduced the notion of well- and ill-posed problems. A problem is well-posed if the following conditions are satisfied:

1. A solution of the problem exists;
2. The solution is unique;
3. The solution is stable.

If a problem is ill-posed, its solution has no practical use. For example, solution (model parameter) instability implies that any insignificant change in the input and output, due to noise or possible outliers, would result in a completely different solution that produces significantly different predictions. This is a situation which no engineer would like to see, especially when crucial decisions based on the model output are to be made. In engineering applications, the data may not contain the information required to solve the problem. We usually assume that a solution exists because we use only approximate models of physical processes, although the true solution may not be among the approximate ones under consideration. Moreover, model parameters may not have a physical interpretation at all, i.e. they may be unobservable.

Inverse problems may violate all the above conditions. Usually the solution is not stable. If a direct problem is smoothing, its corresponding inverse problem is roughening and, as a result, has a highly unstable solution. The roughening mapping tends to amplify noise in the observed output and produces very unstable solutions. This effect is most pronounced when the output is known only approximately due to noise corruption, modeling error, or discretization error.

1.3 A Method of Regularization

Ill-posed problems can be solved by using a Method of Regularization (abbreviated MOR). MOR is a method of finding approximate solutions to ill-posed problems, which are stable under small perturbations of the data. Basically, instead of solving an ill-posed problem we solve a set of well-posed problems that approximate the original ill-posed problem. Though ill-posed problems were encountered by Hadamard as early as 1902, a systematic way of solving them was not developed until 1963 when Tikhonov introduced the method of regularization. The method provides approximate solutions to ill-posed problems which are stable under small perturbations of the data. Getting stable solutions is extremely important in engineering applications because stable solutions provide a reliable source of information for decision-making. Stability also corresponds to the repeatability of the results, which is an important requirement for the result to be scientifically valid.

Ill-posed problems can be continuous and discrete. In the continuous case the solution of the problem is a continuous function of some variables. We are most interested in cases, when the solution is discrete, because the estimation of parameters of parametric models falls into this case, and data are usually collected and stored in the digitized form and are processed with numerical methods on computers. In the discrete case, we estimate a finite number of parameters from a finite amount of data (or observations).

Many discrete ill-posed problems can be reduced to the solution of a simple linear equation

$$Y = Xb + e\,, \tag{1.1}$$

where Y is an $n \times 1$ vector of noisy output signals of a system or process under consideration called the response, X is an $n \times m$ matrix representing n observations or

measurements of m independent variables called the predictors, b is a vector of m parameters called the regression coefficients, and e is an unknown noise vector that represents the measurement error, the modeling error, and the true stochastic noise.

The problem (1.1) is ill-posed when the matrix of second moments of X is singular or near singular. It becomes singular or near singular because of the inclusion of linearly related variables as in the case of prediction from correlated sensor values. As a result, the estimate $X^T X / n$ of the matrix of second moments becomes ill-conditioned, or has a very large condition number. The main implication is a highly unstable least squares solution which becomes very sensitive to particular noise realizations in the observed response. The Ordinary Least Squares (abbreviated OLS) solution is given by

$$b_{OLS} = \left(X^T X\right)^{-1} X^T Y . \tag{1.2}$$

Because $X^T X$ is ill-conditioned, its inverse drastically amplifies the noise component in the response and makes the solution hypersensitive to particular realizations of that noise component. In applications, we desire the opposite. We want a solution be insensitive to noise in the response, because the noise is a noninformative component that contributes nothing to the problem solution. Any influence of noise on the solution is highly undesirable.

We can also write the OLS solution in terms of a Singular Value Decomposition (abbreviated SVD) of matrix X given by

$$X = U \operatorname{diag}(s_i) V^T , \tag{1.3}$$

where U is an $n \times m$ column-orthogonal matrix, V is an $m \times m$ orthogonal matrix, and s is an m-vector of positive or zero elements called the singular values, as

$$\hat{b}_{OLS} = V \operatorname{diag}\left(\frac{1}{s_i}\right) U^T Y = \sum_{i=1}^{m} \frac{r_i}{s_i} v_i . \tag{1.4}$$

It is usually true that the last few components are noise components in the data. These noise components, multiplied by r_i / s_i, contribute to the solution. Even if the cross-

correlation $r_i = u_i^T Y$ between the i-th noise component and the response is fairly small, a very small singular value makes the noise component contribute noticeably to the solution, making the solution unstable and hyper sensitive to the noise component in the data.

The method of regularization suggests using a regularization operator $R(Y, l)$ to obtain a regularized solution

$$b_l = R(Y, l), \tag{1.5}$$

where l is the regularization parameter, which determines the proper degree of regularization depending on the amount of noise in the response. An important property of that operator is that it gives the exact solution of (1.1) when the amount of noise in the response goes to zero. The form of the regularization operator depends on the specifics of the particular problem. It is usually chosen so that the corresponding regularized solutions are physically plausible. For problem (1.1), a regularization operator that produces solutions with small variance is reasonable because it is the large variance of the OLS solution that makes it useless and very sensitive to the noise component.

The most common choice of the regularization operator for problem (1.1) is

$$R(Y, l) \equiv \left(X^T X + l^2 \Omega^T \Omega\right)^{-1} X^T Y \tag{1.6}$$

which, for $\Omega \equiv I_m$, produces the well-known ridge regression (Hoerl, 1970) coefficients. This regularized solution is also known as a minimum energy solution (Hansen, 1998) because large values of the regression coefficients are being penalized. Notice that the introduction of $l^2 \Omega^T \Omega$ causes matrix $X^T X + l^2 \Omega^T \Omega$ become well-conditioned. As a result, the regularized solution has smaller variance and becomes much more stable to the noise component in the response.

In terms of the SVD of X, the regularized solution is written as

$$\hat{b}_l = V \operatorname{diag}\left(\frac{s_i}{s_i^2 + l^2}\right) U^T Y = \sum_{i=1}^{m} \frac{s_i}{s_i^2 + l^2} r_i v_i . \tag{1.7}$$

The value of l which is larger than the singular values corresponding to the noise components prevents these noise components from contributing to the solution because the correlation coefficients are no longer divided by very small singular values as in the OLS case.

Ω in (1.6) is usually called a penalty operator because it is used to penalize undesirable properties of the solution. We also note that making the regularization parameter a function of the noise level in the response, which becomes zero when the noise level goes to zero, would guarantee that the regularization operator would give the exact solution to the problem (1.1) when the noise level is zero.

There are other useful choices of the penalty operator. If it is a matrix that approximates the first derivative operator, the regularized solution is a maximum flatness solution (Hansen, 1998). If it approximates the second derivative operator, the regularized solution is a smooth solution. Since the method of regularization was developed to solve operator equations in which the desired solution is a smooth function, it originally used the second derivative operator as the penalty operator to produce smooth regularized solutions.

The only obstacle to applying these methods is the selection of a proper regularization parameter value. As mentioned already, setting it to zero produces an OLS solution which is unstable, and setting it nonzero produces a regularized solution which has smaller amplitude, greater flatness, or smoothness depending on the penalty operator. The regularization parameter must be a function of the true noise level to guarantee the convergence property of the regularization method. Since the true noise level is almost always unknown in real applications, selection of a proper value of the regularization operator is a very important and challenging problem in itself.

A number of methods for choosing an optimal regularization parameter are proposed in the literature. However, none of them has the desirable properties needed in

engineering practice. They perform poorly with colored noise and in the important case of model misspecification. From this standpoint, the search for new, more powerful regularization parameter selection methods that are data-driven and misspecification-resistant is well justified from both theoretical and practical points of view.

1.4 Originality of the Proposed Work

The main original contribution of the proposed work is a new information complexity-based RPSM and a new method for detection of possible model-misspecification. The information complexity-based RPSM is used for the solution of ill-posed inverse engineering problems. The misspecification-detection method is based on a discrepancy between two versions of the proposed RPSM: one is misspecification-resistant and the other is not.

Unlike the existing methods, the proposed RPSM works consistently for misspecified models, i.e. it is misspecification-resistant, and reduces the risk of obtaining grossly underregularized solutions, i.e. reduces variability of the chosen regularization parameter. Misspecification-resistance makes solutions that use the proposed RPSM robust to modeling errors while the reduced variability of the chosen regularization parameter makes the system robust to peculiarities in the noise components of the response. None of the existing methods combines both of these features. As a result, the existing RPSM's can perform well only in certain situations when the crucial assumptions under which the methods were derived are satisfied. The proposed RPSM is resistant to violation of the assumptions and performs well in the very important case of a small number of observations. Misspecification-resistance and robustness are of great value in building reliable autonomous diagnostic and monitoring systems.

The superior performance of the proposed RPSM is demonstrated using various examples starting from building an inferential system for venturi meter drift detection, through building a sensor validation system and solving integral equations to image restoration and prior distribution specification in Bayesian inference.

1.5 Dissertation Organization

The rest of the dissertation is organized as follows. CHAPTER 2 is a literature survey of current methods for choosing the regularization parameter value. CHAPTER 3 describes the information approach in the context of maximum penalized likelihood estimation, which is used for the solution of ill-posed problems in the stochastic setting. It also presents a new extension of the information approach in the context of penalized estimation and develops a new RPSM which is misspecification-resistant and more robust in real world applications than information-based RPSM's because of an extra penalization of estimation inaccuracy. CHAPTER 4 contains a number of examples that cover a wide spectrum of practical applications from sensor validation using data from a nuclear power plant to image restoration and learning from data. In the last chapter we draw conclusions and mark possible future work and further improvements.

CHAPTER 2
PREVIOUS WORK ON REGULARIZATION PARAMETER SELECTION

There are two major approaches to regularization parameter selection: deterministic and stochastic. The stochastic approach exploits the statistical nature of the noise component in the response whereas the deterministic approach completely ignores it. In either approach there are methods that require different types of input information for producing a proper value of the regularization parameter for a particular problem. Figure 2.1 demonstrates a possible classification of the RPSM's. The "Heuristic" and "Error Free" methods do not require an estimate of the noise level in the response; the others do.

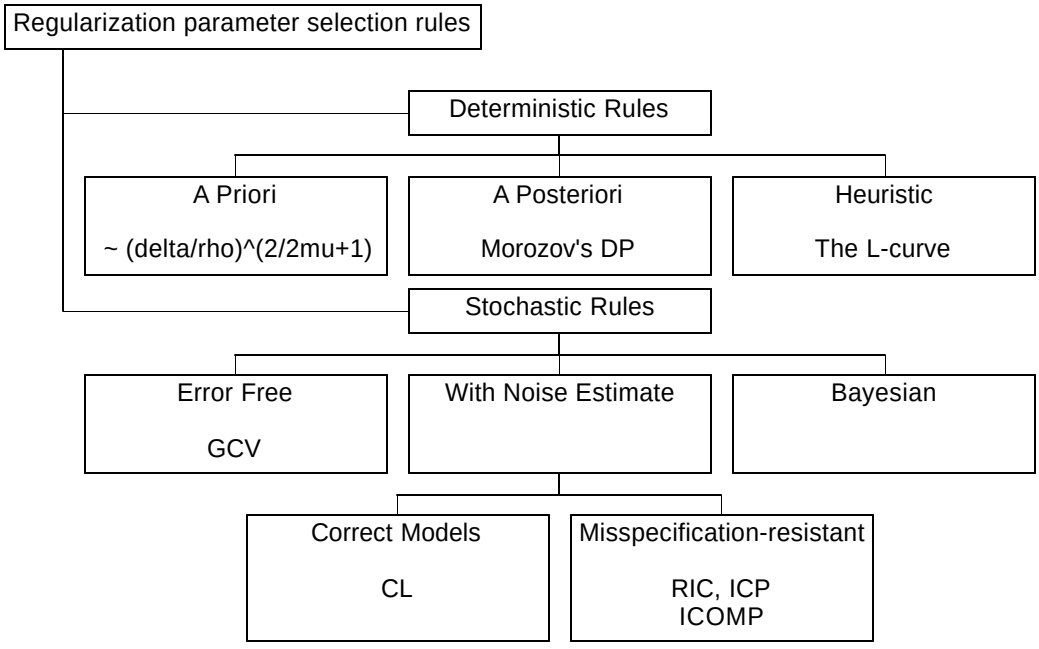

Figure 2.1. Regularization parameter selection method classification.

2.1 Deterministic RPSM's

A priori RPSM's require, as their name implies, a priori information about the true solution and the true noise level in the response. Since neither is available in practical applications, especially when parameters have no physical interpretation at all, these methods are of little interest for practical implementations. They are important from the theoretical point of view because they establish optimal convergence rates. A particular regularization method is convergent when the error between the regularized solution obtained using this method and the true solution goes to zero as the noise in the response goes to zero. The convergence rates are useful in the theoretical analysis of the regularization methods and in comparing different RPSM's. RPSM's with faster convergence would provide more accurate solutions for a given noise level and, thus, are preferable.

2.1.1 A priori RPSM's

When the noise level, denoted as d, is known and, for some $m > 0$, $b = \left(X^T X\right)^m w$, where $\|w\| \le u$, i.e. b has a source representation, the regularization method is of optimal order with the following a priori RPSM (Engle, 2000),

$$l \sim \left(\frac{d}{u}\right)^{\frac{2}{2m+1}}. \tag{2.1}$$

This result is for the deterministic setting. The source representation can be seen as a condition on the decay rate of the correlation coefficients r_i between Y and u_i. For problem (1.1) to have a regularized solution, the correlation coefficients r_i arranged in decreasing order of the singular values must decay faster then the singular values of $X^T X / n$. For larger m, this condition becomes more severe. Namely, the correlation coefficients must decay faster than the singular values raised to the $2 + 4m$ power. If this

is fulfilled for larger m, the convergence of the regularized solution to the true one will be faster.

For most real-world applications neither m nor u is known, and, as a result, it is impossible to construct an a priori RPSM of optimal order. Therefore, a number of a posteriori RPSM's that depend on the data have been proposed.

2.1.2 A posteriori RPSM's

The a posteriori RPSM that is most widely used is Morozov's (1984) Discrepancy Principle (abbreviated MDP). The regularization parameter value is chosen as a solution of the following equation

$$\left\| Xb_l - Y \right\| \leq d \,. \tag{2.2}$$

The regularization parameter l is chosen such that the corresponding residual (left hand side of (2.2)) is less than or equal to the a priori specified bound (right hand side) for the noise level in the response. Since a smaller l corresponds to less stable solutions, the l 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 then the noise level in the response corresponds to overfitting, which is a term for learning noise in the training data. The regularization method with l chosen according to the discrepancy principle (2.2) is convergent and of optimal order (Morozov 1984; Engle, 2000).

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. 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 (Hansen, 1998). An improved a posteriori method (Engle, 2000; Raus, 1984) outperforms MDP in that it is of optimal order for a wider range of m than MDP.

A posteriori RPSM's require the noise level to be either known or reliably estimated. Such a noise level can be hard to obtain. An alternative approach to regularization parameter selection uses noise-level-free RPSM's. Noise-level-free RPSM's are also referred to as heuristic RPSM's. Heuristic RPSM's provide a regularization parameter value without knowledge of the noise level. However, due to the result of Bakushinskii (1984), a noise-level-free RPSM cannot provide a convergent regularization method. Therefore, heuristic RPSM's are nonconvergent. Despite that, in practical applications, heuristic RPSM's may demonstrate very good performance in reconstructing the solution of ill-posed problems (e.g. Hanke, 1993).

2.1.3 The L-curve method

The most widely-used heuristic method is the L-curve method (Hansen, 1998). In this method, the residual norm is plotted versus the regularized solution norm and the regularization parameter value corresponding to the corner of the L-shape curve is chosen. The corner occurs where the curve has its maximum curvature. The L-curve method has been shown to be nonconvergent (Vogel, 1996; Leonov, 1997). For some problems, it is extremely difficult to locate the corner; for others, the L-curve may have several corners. The L-curve method can be also used in the stochastic setting.

2.2 Stochastic RPSM's

In a stochastic setting, a distributional model of the noise component e in the response is specified. Usually, white Gaussian noise is assumed, i.e. the noise component has a multivariate normal distribution denoted as $e \sim N_n\left(0, s^2 I_n\right)$, where e is a random noise n-vector whose components are independent and normally distributed with zero mean and common variance s^2. I_n denotes the $n \times n$ identity matrix. A RPSM is

obtained so that it minimizes the mean predictive error estimated from the data. Therefore, all RPSM's in the stochastic setting use an estimator of the mean predictive error and select the regularization parameter value that minimizes the corresponding estimator.

2.2.1 Generalized Cross Validation

Probably the most widely used noise-level-free RPSM is Generalized Cross Validation (abbreviated GCV) (Wahba, 1990). According to this method, the regularization parameter is chosen such that it minimizes the GCV function given by

$$GCV(l) = \frac{\|Xb_l - Y\|^2 / n}{(trace(I - H_l)/n)^2}, \tag{2.3}$$

where $H_l = X\left(X^T X + lI\right)^{-1} X^T$ is called the hat or projection matrix. GCV does not require prior knowledge of the noise level and works with the white Gaussian noise model for the noise component. GCV occasionally fails, presumably due to the presence of correlated noise (Wahba, 1990). GCV can also produce grossly underregularized solutions (Wahba, 1993).

2.2.2 Mallows' CL method

Other widely used RPSM's are Mallows' (1973) CL and the Unbiased Risk Estimator (abbreviated URE) (Eubank, 1988), which is similar to CL. CL is derived as an estimator of the mean predictive error, in which the noise level is treated as a nuisance parameter and components e_i of the noise vector are assumed to be normally distributed with zero mean and common variance s^2. CL is given by

$$CL(l) = \frac{\|Xb_l - Y\|^2}{n} + \frac{2s^2}{n} trace(H_l) - s^2. \tag{2.4}$$

CL can be considered as an information criterion as shown in Section 3.3.

CL must be accompanied by either an a priori noise level as in the deterministic setting or by a reliable estimate of the noise level. CL is very sensitive to an underestimation of the noise level and may fail to provide a regularization parameter value corresponding to an admissible regularized solution. CL was derived for the white Gaussian noise case and, hence, may not work reliably if that assumption is violated.

2.2.3 Information Criteria

GCV and CL methods are defined for uncorrelated Gaussian noise case and cannot be easily extended to more realistic cases. In real applications, the distribution of noise can be non-Gaussian with non-zero values of skewness (be asymmetric) and excess (be narrower or wider than Gaussian). Data can contain outliers and can be generated by a mixture of distributions. The level or variance of the noise may not be stationary but can vary. The noise may also be correlated. Finally, the statistical model of the noise can be misspecified, and the results obtained without taking this fact into account can be invalid. None of the above methods can be generalized to any of these conditions.

To be able to deal with noise and model-misspecification and to construct misspecification-resistant RPSM's, we should consider the information approach which became widely-used in statistical model selection due to the works of Akaike (1973), Takeuchi (1976), Bozdogan (1987-2001), Murata (1994), and others. Unfortunately, information-based criteria such as the Regularization Information Criterion (abbreviated RIC) proposed by Shibata (1989) and the Information Criterion for Penalized models (abbreviated ICP) proposed by Konishi and Kitagawa (1996) have not been widely used as RPSM's for the solution of ill-posed problems. The RPSM's derived using the information approach are described in CHAPTER 3.

The main advantage of the information approach is that it accounts for possible functional and distributional misspecifications of the models in a very natural way. While

misspecification may not be an issue when solving integral equations, it plays a crucial role in engineering applications based on black-box and data-driven techniques where the very notion of a true model is arguable and usually not discussed though the existence of one is silently assumed. A similar situation arises with econometric models in which, in contrast to engineering, misspecification-detection and misspecification-resistant estimation have been extensively used. For a detailed treatment of misspecification in modeling and further references on misspecification testing we refer to the works of White (1981-1994). In these situations, methods that are consistent under possible misspecifications are valuable because they automatically guard against the unrealistic assumption of correct model specification.

Criteria such as CL, RIC, and ICP evaluate the generalization (or prediction) error using the training error and an additional term. This additional term penalizes the inaccuracy of parameter estimation and can be interpreted as the effective number of parameters of correctly specified models (for CL) or incorrectly specified models (for RIC and ICP).

With a limited number of observations, penalization of the number of parameters alone becomes inadequate. This additional term cannot be computed exactly because of the dependence on the unknown true distribution and should be estimated from the same data set. As a result, the selected regularization parameter value is often underestimated and produces grossly underregularized or inadmissible solutions. An additional penalization of the parameter estimation inaccuracy, taking into account the interdependencies between the parameter estimates as in the Information Complexity RPSM (abbreviated ICOMPRPS) proposed in Urmanov and et. al. (2002) can drastically reduce the risk of regularization parameter value underestimation and make such a choice more suitable for black-box modeling. Such an "overestimation", or more precisely correction, of the inadequate penalization of inaccuracy is beneficial for engineering

applications in which the regularization parameter value should be chosen automatically during model building, and there is no means for assessing the proper amount of regularization.

In CHAPTER 3, a systematic way of deriving information-based RPSM's, which are misspecification-resistant, is presented. The problem of obtaining grossly underregularized solutions because of large variability of the chosen regularization parameter is discussed. A method of reducing the risk of obtaining grossly underregularized solutions, using information complexity-based RPSM's is proposed. A method of misspecification detection is proposed.

CHAPTER 3
REGULARIZATION PARAMETER SELECTION: AN INFORMATION APPROACH

3.1 Introduction

We introduce the information approach to regularization parameter selection in the linear case, though the approach is not limited to this case and is very general. Consider linear models of the form

$$Y_i = X_i^T b + u_i,\ i = 1...n, \tag{3.1}$$

where Y_i is a dependent variable (or response), X_i is an independent m-vector variable (or predictors), b is an unknown m-vector of regression coefficients or parameters to be estimated from observed data, and u_i's are random (noise) variables with the following properties

$$E(u_i) = 0,\ E(u_i^2) = s_u^2, \text{ and } E(X_i u_i) = 0. \tag{3.2}$$

There are two potential problems with such model specifications: (3.1)-(3.2). If the true relationship $m(x) \equiv E\{Y_i \mid X_i = x\}$ between Y_i and X_i, also referred to as the true model, is not linear, or some relevant predictors are missing, we have functional misspecification. In this case the error term $u_i \equiv m(X_i) - X_i^T b + e_i$ includes both the error of approximation, $m(X_i) - X_i^T b$, and the true stochastic error e_i. This means that X_i and u_i are no longer independent. For example, the usual covariance matrix of the ridge regression coefficients is obtain as

$$\begin{aligned}\Sigma &= Cov\left(\hat{b}_l\right) \\ &= Cov\left(\left(X^T X + nl\right)^{-1} X^T y\right) \\ &= \left(X^T X + nl\right)^{-1} X^T Cov\left(yy^T\right) X\left(X^T X + nl\right)^{-1} \\ &= s^2\left(X^T X + nl\right)^{-1} X^T X\left(X^T X + nl\right)^{-1}\end{aligned} \tag{3.3}$$

and can be estimated as

$$\hat{\Sigma} = \hat{s}^2\left(X^T X + nl\right)^{-1} X^T X\left(X^T X + nl\right)^{-1}. \tag{3.4}$$

It becomes inconsistent for misspecified models (White 1980). This inconsistency may make RPSM's such as Mallows' CL and URE, which implicitly use this covariance matrix estimator, inconsistent as well. In the maximum likelihood or Ordinary Least Squares (abbreviated OLS) framework, one can use an improved covariance matrix estimator, which is consistent under functional misspecification (White, 1980). In the ridge regression framework, a modified covariance matrix estimator can also be used to cope with possible functional misspecifications.

The second problem is that the distributional assumption on the stochastic error (usually normality with zero mean and constant variance) is not fulfilled. Luckily, this type of model misspecification does not affect estimation of the regression coefficients b. However, the covariance matrix estimator again becomes inconsistent. This may destroy the performance of a RPSM that uses the estimator (3.4). In the OLS framework, one can use an estimator which is consistent under distributional misspecifications, as in White (1982), and in the ridge framework we can also use a modified estimator that accounts for possible distributional misspecifications.

Perhaps a third and more serious problem with model (3.1-3.2) is the assumption of independent observations or uncorrelated noise. If the true noise happens to be correlated, a RPSM that uses the uncorrelated noise assumption will most probably fail to select a plausible value of the regularization parameter. To cope with correlated noise,

one can consider an autoregressive noise model and work out an information criterion in a similar manner.

To solve the problem (3.1) for the regression coefficients b, we assume that we have n observations (X_i, Y_i) and rewrite the model in a matrix form

$$Y = Xb + u\,, \tag{3.5}$$

where Y is an $n \times 1$ vector, X is an $n \times m$ matrix, and u is an $n \times 1$ vector of random errors in the response. The OLS solution minimizes the Sum of Squared Residuals (abbreviated SSR)

$$SSR \equiv \sum_{i=1}^{n} \left(Y_i - X_i^T b\right)^2 = \left\|Y - Xb\right\|^2 \to \min \tag{3.6}$$

and is given by

$$b_{OLS} = \left(X^T X\right)^{-1} X^T Y\,. \tag{3.7}$$

When the data matrix X is ill-conditioned (due to collinear predictor variables), the OLS solution is unstable (or statistically insignificant) and has no practical use. To obtain a stable solution, one can proceed with using a Method Of Regularization (abbreviated MOR). The common choice is Tikhonov (1963) regularization, which uses universal prior information of smoothness to obtain plausible solutions. In particular, a penalty term is added to the sum of squared residuals that assesses the physical plausibility of solutions. The resulting regularized solution is given as a solution of the minimization problem

$$\left\|Xb - Y\right\|^2 + l\left\|\Omega b\right\|^2 \to \min \tag{3.8}$$

which is minimized for

$$b_l = \left(X^T X + l\,\Omega^T \Omega\right)^{-1} X^T Y\,, \tag{3.9}$$

where Ω is a penalty operator (matrix), and l is the regularization parameter that controls the amount of penalty. When $\Omega = I_m$, minimum energy solutions are preferred (Hansen, 1998). These solutions correspond to the well known ridge regression (Hoerl,

1970) solution. When matrix Ω is an approximation to the second derivative operator, smooth solutions are obtained. The regularized solution (3.9) is biased. This means that the expected value of the regularized regression coefficients is not equal to the true value of the regression coefficients if such value exists. When regression coefficients have no physical interpretation and the main goal is to predict future observations, the bias cannot be considered as a drawback as long as it results in improved prediction accuracy. Though biased, the regularized solution (3.9) for a suitably chosen l is useful and reduces the mean estimation error by significantly reducing the variance of the regularized solution as compared to the OLS solution (Hoerl, 1970).

As already mentioned, the proper choice of the regularization parameter value is critical for obtaining a useful regularized solution, and many different RPSM's have been proposed. The rest of the chapter is dedicated to describing an information approach that can naturally account for possible model misspecification. RPSM's based on that approach are shown to be robust against such misspecifications. In addition, we argue that for a limited number of observations, a slight "overestimation" of the regularization parameter value is beneficial from the practical point of view, and that a RPSM derived in the information complexity framework provides such a refinement.

The information approach uses the Maximum Penalized Likelihood (abbreviated MPL) estimation framework and properties of the MPL Estimators (abbreviated MPLE). Therefore, we briefly review the maximum penalized likelihood method in the context of ridge regression. For a more in depth description of the maximum penalized likelihood method, see Good and Gaskins (1971), Silverman (1985), Green (1987) and for asymptotic analysis of the MPL method, see Cox (1990).

3.2 Maximum Penalized Likelihood Method

It is well known that the only way to overcome a lack of information is to bring some. When the amount of information contained in an observed data set is not sufficient for obtaining a useful solution additional information must be brought from outside the observed data. This is naturally implemented in the Bayesian approach by combining prior information with data to make an inference. An alternative approach is to use a penalized likelihood that also exploits prior information but in a narrower way than the Bayesian approach. Specifically, maximum penalized likelihood estimation corresponds to a maximum a posteriori procedure in Bayesian analysis (Leonard, 1978).

When solving an ill-conditioned problem as in (3.1) we often obtain inadmissible results by using the Maximum Likelihood (abbreviated ML) method because of violations of the assumptions under which the ML method is valid. In the ill-conditioned case, the collinearity makes the solution underdetermined, and additional information must be used to further constrain the solution. This additional constraint is in the form of a penalization operator. The use of the penalization operator reduces undesirable properties of the solution. The idea of using smoothness as additional information, proposed by Tikhonov (1963), is a good example of prior information that works successfully in numerous engineering and scientific problems.

We will now consider a more general model than (3.1). Assume that there exists an unknown true joint cumulative distribution function (abbreviated c.d.f.) of X_i, which is a random m-vector, and Y_i, which is a random variable dependent on X_i,

$$G(X_i, Y_i) \text{ with density } g_{X_i,Y_i}(x, y). \tag{3.10}$$

The problem of modeling an observed data set $D = \{(X_i, Y_i)\}_{i=1}^{n}$ is comprised of specifying a parametric family of approximating distributions called the model

$$F(X_i, Y_i; b) \text{ with density } f_{X_i,Y_i}(x, y; b) = f_{X_i}(x) f_{Y_i|X_i}(y \mid x; b) \tag{3.11}$$

and estimating the parameter vector b from the observed data set D.

The model (3.11) is said to be correctly specified if there exists b_0 such that $F(X_i, Y_i; b_0) = G(X_i, Y_i)$; otherwise, the model is said to be misspecified. Several forms of misspecification are possible. Functional misspecification occurs when the conditional mean of Y_i is misspecified, i.e. $m(x) \equiv E(Y_i \mid X_i = x) \neq x^T b$ for any $b \in R^m$. Distributional misspecification occurs when the true distribution (3.10) does not belong to the specified family of approximating distributions (3.11). A detail discussion of misspecification in statistical modeling can be found in White (1994).

To estimate b from the observed data D, the maximum likelihood method is used. For a given sample of n independent identically distributed (abbreviated i.i.d.) observations, the likelihood is defined as

$$L(D \mid b) \equiv \prod_{i=1}^{n} f(X_i, Y_i; b). \tag{3.12}$$

The likelihood represents the joint probability of the observations, regarded as a function of an unknown parameter. The log likelihood function of an observation is defined as

$$LL(X_i, Y_i \mid b) \equiv \log f(Y_i \mid X_i; b). \tag{3.13}$$

The log likelihood function is given by $\log f(X_i, Y_i) = \log f(X_i) + \log f(Y_i \mid X_i; b)$. However, since the first term does not depend on b, it will not affect the estimation of b. Therefore, in the following we refer to (3.13) as the log likelihood function of an observation. The log likelihood is defined as

$$LL(D \mid b) \equiv \log L(D \mid b) = \sum_{i=1}^{n} LL(X_i, Y_i \mid b). \tag{3.14}$$

The value of b that maximizes the log likelihood (3.14) is called the Maximum Likelihood Estimator (abbreviated MLE) of b and denoted as $\hat{b}$. The use of the maximum likelihood estimation method was first suggested by Fisher (1921) and has become one of the most extensively-used tools in statistical analysis.

However, when the maximum likelihood method is applied to ill-posed problems, it does not produce a valuable result. The ML method can be modified, for example by introducing a penalty and "converting" it into a maximum penalized likelihood method. The penalized log likelihood function of an observation is defined as

$$PLL(X_i, Y_i \mid b) = LL(X_i, Y_i \mid b) - l\,p(b), \tag{3.15}$$

where $p(b)$ is a penalty, and l is the regularization parameter as in (3.8). In the statistical literature, the penalized likelihood method was first proposed by Good and Gaskins (1971). Different penalties are discussed in Green (1987). In particular, consider the quadratic penalty in the form

$$p(b) \equiv \frac{1}{2s^2}(\Omega b)^T \Omega b = \frac{1}{2s^2}\|\Omega b\|^2, \tag{3.16}$$

where Ω is a penalty operator (an $(m \times m)$ matrix) with $\Omega = I_m$ corresponding to ridge regression. Given a sample of n observations (X_i, Y_i), the Maximum Penalized Likelihood Estimator (abbreviated MPLE) of b is obtained as a solution of the following problem:

$$\begin{aligned} \hat{b}_l &= \arg\max_b \frac{1}{n}\sum_{i=1}^{n} PLL(X_i, Y_i \mid b) \\ &= \arg\max_b \frac{1}{n}\sum_{i=1}^{n} (\log f(Y_i \mid X_i; b) - l\,p(b)) \end{aligned} \tag{3.17}$$

When we specify a normal distribution for the dependent variable

$$Y_i \mid X_i \sim N(X_i^T b, s^2) = \frac{1}{\sqrt{2p}\,s}\exp\left(-\frac{1}{2s^2}(Y_i - X_i^T b)^2\right) \tag{3.18}$$

and a quadratic penalty of form (3.16), the maximum penalized likelihood estimator (3.17) is exactly a ridge estimator (in matrix notation) (Hoerl, 1970):

$$\hat{b}_l = \left(X^T X + n l I_m\right)^{-1} X^T Y\,. \tag{3.19}$$

Two results concerning the asymptotic properties of the MPLE defined in (3.17) are stated below without proof and will be used later for deriving misspecification-resistant RPSM's. These results can be proved following the same steps as in White

(1981). For more information on asymptotic properties of penalized likelihood estimators in the case of correctly-specified model, see Cox (1990) and Knight (1998), and on asymptotic properties of maximum likelihood estimators under model misspecification, see White (1980).

R.1 $\hat{b}_l$ is a consistent estimator of b_l^* which is the unique solution of

$$E_{W,Z}\left\{\frac{\partial}{\partial b} PLL(W, Z \mid b)\right\} = 0 . \tag{3.20}$$

R.2 With a large enough n, $\hat{b}_l$ is approximately normally distributed, $\sqrt{n}\left(\hat{b}_l - b_l^*\right) \sim N_m\left(0, J^{-1} I J^{-1}\right)$, where matrices J and I are defined as

$$J \equiv -E_{W,Z}\left\{\frac{\partial^2}{\partial b \partial b^T} PLL\left(W, Z \mid b_l^*\right)\right\} \text{ and}$$

$$I \equiv E_{W,Z}\left\{\frac{\partial}{\partial b} PLL\left(W, Z \mid b_l^*\right) \cdot \frac{\partial}{\partial b^T} PLL\left(W, Z \mid b_l^*\right)\right\} . \tag{3.21}$$

W and Z are random variables that have the same joint distribution as X_i and Y_i and are independent from X_i and Y_i. E in (3.20) and (3.21) stands for the expectation operator with expectation is taken with respect to the true joint distribution of W and Z. In the maximum likelihood case, when $l = 0$, the matrices (3.21) are called Fisher information matrices in the Hessian (outer product) and inner product form respectively. In the maximum likelihood case, these matrices are equal ($J = I$) when the model is correctly specified, and they are different ($J \neq I$) when the model is misspecified (White, 1980). This property is extensively used for misspecification detection of econometric models.

In the Gaussian case with quadratic penalty (3.16), (3.20) is minimized for

$$b_l^* = \left(E_W\left\{WW^T\right\} + l I_m\right)^{-1} E_{W,Z}\{WZ\} = \left(E_W\left\{WW^T\right\} + l I_m\right)^{-1} E_W\left\{WW^T\right\} b^* \tag{3.22}$$

which is the limiting value of the MPLE $\hat{b}_l$ as $n \to \infty$. b^* is the solution of

$$E_{W,Z}\left\{\frac{\partial}{\partial b} LL(W, Z \mid b)\right\} = 0 \tag{3.23}$$

or the limiting value of the maximum likelihood estimator. In the case of correct model specification $b^* = b_0$ (White, 1980). When the matrix $E_W\{WW^T\}$ of second moments of W is near singular, its estimator $X^T X / n$ will be ill-conditioned and, as a result, the maximum likelihood estimator given by

$$\hat{b} = \left(X^T X\right)^{-1} X^T Y \tag{3.24}$$

will have a very large variance and no practical value.

Since matrices J and I depend on the unknown true distribution (3.10) they are not computable and should be estimated in practice. Estimation is done by substituting the empirical distribution in (3.21). This results in the following estimators for the matrices

$$\hat{J} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial b \partial b^T} PLL\left(X_i, Y_i \mid \hat{b}_l\right) \text{ and} \tag{3.25}$$

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial b} PLL\left(X_i, Y_i \mid \hat{b}_l\right) \cdot \frac{\partial}{\partial b^T} PLL\left(X_i, Y_i \mid \hat{b}_l\right). \tag{3.26}$$

These matrices are used to estimate the asymptotic covariance matrix of the MPLE $\hat{b}_l$ as

$$\hat{\Sigma}\left(\hat{b}_l\right) = \hat{J}^{-1} \hat{I} \hat{J}^{-1}. \tag{3.27}$$

In analogy with the ML case, this covariance matrix estimator can be shown to be consistent under model misspecification.

In the following section, an information approach is developed for evaluating different competing models whose parameters are estimated by the described maximum penalized likelihood method. The information criterion can also be used to choose the regularization parameter value that minimizes this criterion. Optimal selection of penalized models can be performed in two steps. For each competing model the regularization parameter that minimizes the information criterion is chosen and then the minimized values of the criterion are compared to select the best model. The chosen

model is the one that is closest to the true one in the sense of minimum of the Kullback-Leibler (1951) distance, which will be discussed in the next section.

3.3 Information Approach to Regularization Parameter Selection

When the parameters of a specified model $f(X_i, Y_i; b)$ are estimated by the MPL method, each particular choice of the penalty operator and regularization parameter yields some approximating density $\hat{f}_l \equiv f(X_i, Y_i; \hat{b}_l)$. The closeness of this approximating density $\hat{f}_l$ to the unknown true density $g(X_i, Y_i)$, assuming such exists, can be evaluated by the Kullback-Leibler (1951) (abbreviated KL) information (or distance) that measures the divergence between the densities

$$KL(\hat{f}_l ; g) \equiv E_{W,Z}\left\{\log \frac{g}{\hat{f}_l}\right\} = \int \ldots \int \log \frac{g(w,z)}{f(w,z;\hat{b}_l)} \cdot g(w,z) dw_1 dw_2 \ldots dw_m dz \,. \quad (3.28)$$

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 $\hat{f}_l$

$$\hat{l}_{KL} = \arg\min_l \{E_D KL(\hat{f}_l ; g)\}. \quad (3.29)$$

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":

$$\begin{aligned} E_D KL(\hat{f}_l ; g) &= E_D\left\{E_{W,Z} \log \frac{g}{\hat{f}_l}\right\} \\ &= E_D\left\{E_{W,Z} \log \frac{g}{f^*} \frac{f^*}{f_l^*} \frac{f_l^*}{\hat{f}_l}\right\} \quad (3.30) \\ &= \underbrace{E_{W,Z} \log \frac{g}{f^*} + E_{W,Z} \log \frac{f^*}{f_l^*}}_{SystematicError} + \underbrace{E_D\left\{E_{W,Z} \log \frac{f_l^*}{\hat{f}_l}\right\}}_{RandomError} \end{aligned}$$

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

$$E_{W,Z}\left\{\frac{\partial}{\partial b} LL(W, Z \mid b)\right\} = 0$$

or the limiting value of the ML estimator; $f_l^* \equiv f(W, Z; b_l^*)$ and b_l^* is a solution of

$$E_{W,Z}\left\{\frac{\partial}{\partial b} PLL(W, Z \mid b)\right\} = 0$$

or the limiting value of the MPL estimator.

The systematic error, which can be also termed as 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 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 tradeoff 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 $\hat{f}_l$, minimization of the mean KL distance is equivalent to maximization of the Mean Expected Log Likelihood (abbreviated MELL) which is defined as

$$MELL(l) \equiv E_D\{E_{W,Z} \log \hat{f}_l\}, \tag{3.31}$$

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, Akaike (1973) 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 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 (3.31). By this means, the so-called Average Log Likelihood (abbreviated ALL) is obtained:

$$ALL\left(\hat{b}_l\right)=\frac{1}{n}\sum_{i=1}^{n}\log f\left(Y_i \mid X_i;\hat{b}_l\right). \tag{3.32}$$

Despite the fact that $ALL(b)\to ELL(b)$ as $n\to\infty$, due to the law of large numbers, the ALL, evaluated at MPLE $\hat{b}_l$, is a biased estimator of the MELL of the MPL model i.e. $E_D ALL\left(\hat{b}_l\right)\neq MELL(l)$. This bias should be corrected when we use MELL as a 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 Akaike (1973), Sakamoto (1986), Bozdogan (1987-2001), Konishi (1996), and Shibata (1989).

3.3.1 Maximum mean expected log likelihood parameter choice

An information-based RPSM is given as the maximization of the mean expected log likelihood (3.31) of maximum penalized likelihood models

$$\hat{l}_{MELL}=\arg\max_{l}\{MELL(l)\}. \tag{3.33}$$

As already mentioned, the MELL is not computable and can be estimated by the ALL (3.32). 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 EPLL) as

$$EPLL(b) \equiv E_{W,Z} PLL(W, Z \mid b) \tag{3.34}$$

and expand it in a Taylor series at $\hat{b}_l$ around b_l^*, which is the limiting value of the MPLE $\hat{b}_l$ as $n \to \infty$

$$\begin{aligned} EPLL(\hat{b}_l) &\approx EPLL(b_l^*) + \left\{ \frac{\partial}{\partial b} EPLL(b_l^*) \right\}^T (\hat{b}_l - b_l^*) \\ &+ \frac{1}{2} (\hat{b}_l - b_l^*)^T \left\{ \frac{\partial^2}{\partial b \partial b^T} EPLL(b_l^*) \right\} (\hat{b}_l - b_l^*) \\ &= EPLL(b_l^*) - \frac{1}{2} (\hat{b}_l - b_l^*)^T J (\hat{b}_l - b_l^*) \end{aligned} \tag{3.35}$$

where

$$J \equiv -\frac{\partial^2}{\partial b \partial b^T} EPLL(b_l^*).$$

Next, we expand the Average Penalized Log Likelihood (abbreviated APLL) defined as

$$APLL(b) \equiv \frac{1}{n} \sum_{i=1}^{n} LL(X_i, Y_i \mid b) - lp(b) \tag{3.36}$$

in a Taylor series at b_l^* around $\hat{b}_l$

$$\begin{aligned} APLL(b_l^*) &\approx APLL(\hat{b}_l) + \left\{ \frac{\partial}{\partial b} APLL(\hat{b}_l) \right\}^T (b_l^* - \hat{b}_l) \\ &+ \frac{1}{2} (b_l^* - \hat{b}_l)^T \left\{ \frac{\partial^2}{\partial b \partial b^T} APLL(\hat{b}_l) \right\} (b_l^* - \hat{b}_l) \\ &\approx APLL(\hat{b}_l) - \frac{1}{2} (b_l^* - \hat{b}_l)^T J (b_l^* - \hat{b}_l) \end{aligned} \tag{3.37}$$

We used the fact that

$$\frac{\partial}{\partial b} APLL(\hat{b}_l) = 0$$

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

$$\left\{\frac{\partial^2}{\partial b \partial b^T} APLL\left(b_l^*\right)\right\} \to \left\{\frac{\partial^2}{\partial b \partial b^T} EPLL\left(b_l^*\right)\right\},$$

and, since $\hat{b}_l \to b_l^*$ as $n \to \infty$ due to (R.1), we have

$$\left\{\frac{\partial^2}{\partial b \partial b^T} APLL\left(\hat{b}_l\right)\right\} \to \left\{\frac{\partial^2}{\partial b \partial b^T} EPLL\left(b_l^*\right)\right\}.$$

Using $E_D EPLL\left(b_l^*\right) = E_D APLL\left(b_l^*\right)$ and combining (3.35) and (3.37) we obtain

$$E_D EPLL\left(\hat{b}_l\right) \approx E_D APLL\left(\hat{b}_l\right) - E_D\left\{\left(b_l^* - \hat{b}_l\right)^T J\left(b_l^* - \hat{b}_l\right)\right\} \tag{3.38}$$

and since

$$E_D EPLL\left(\hat{b}_l\right) = E_D ELL\left(\hat{b}_l\right) - l E_D p\left(\hat{b}_l\right) \text{ and } E_D APLL\left(\hat{b}_l\right) = E_D ALL\left(\hat{b}_l\right) - l E_D p\left(\hat{b}_l\right)$$

we have

$$\begin{aligned} E_D ELL\left(\hat{b}_l\right) &\approx E_D ALL\left(\hat{b}_l\right) - E_D\left\{\left(b_l^* - \hat{b}_l\right)^T J\left(b_l^* - \hat{b}_l\right)\right\} \\ &\approx E_D ALL\left(\hat{b}_l\right) - \frac{1}{n} trace\left(IJ^{-1}\right) \end{aligned} \tag{3.39}$$

where we use the asymptotic normality of the maximum penalized likelihood estimator (R.2), and the trace result from Appendix A.1

$$E_D\left\{\left(b_l^* - \hat{b}_l\right)^T J\left(b_l^* - \hat{b}_l\right)\right\} = \frac{1}{n} trace\left(IJ^{-1}\right). \tag{3.40}$$

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

$$T_{MELL}\left(\hat{b}_l\right) \equiv ALL\left(\hat{b}_l\right) - \frac{1}{n} trace\left(\hat{I}\hat{J}^{-1}\right), \tag{3.41}$$

where

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial b} PLL\left(X_i, Y_i \mid \hat{b}_l\right) \frac{\partial}{\partial b^T} PLL\left(X_i, Y_i \mid \hat{b}_l\right) \tag{3.42}$$

and

$$\hat{J} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial b \partial b^T} PLL\left(X_i, Y_i \mid \hat{b}_l\right), \tag{3.43}$$

and the corresponding RPSM is

$$\hat{l}_{MELL} = \arg\max_{l} \left\{ ALL\left(\hat{b}_l\right) - \frac{1}{n} trace\left(\hat{I}\hat{J}^{-1}\right) \right\}. \tag{3.44}$$

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

$$\hat{l}_{CL} = \arg\min_{l} \left\{ \frac{1}{n} \left\| Y - X\hat{b}_l \right\|^2 + \frac{2s^2}{n} trace\left(X^T X \left(X^T X + n l I_m \right)^{-1} \right) \right\}. \tag{3.45}$$

When the model is Gaussian and s^2 is treated as a nuisance parameter and J and I are estimated as

$$\hat{J} = -\frac{1}{s^2} \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i^T + l I_m \right) \text{ and } \hat{I} = \frac{1}{n s^4} \sum_{i=1}^{n} r_{ols\,i}^2 X_i X_i^T \tag{3.46}$$

Shibata's (1989) Regularization Information Criterion (abbreviated RIC) is obtained and the corresponding RPSM is

$$\hat{l}_{RIC} = \arg\min_{l} \left\{ \frac{1}{n} \left\| Y - X\hat{b}_l \right\|^2 + \frac{2s^2}{n} \sum_{i=1}^{n} \frac{r_{ols\,i}^2}{s^2} H_{ii} \right\}, \tag{3.47}$$

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

When $\hat{b}_l$ is an M-estimator (Huber, 1981), Konishi and Kitagawa (1996) propose an information criterion for choosing the regularization parameter which is similar to RIC (3.47).

We also suggest a RPSM that uses Bozdogan's (1996) 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. We give a brief description of the informational complexity RPSM in Section 3.4.

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 (3.44) becomes similar to CL.

3.3.2 Gaussian, correctly specified case

The MELL RPSM (3.44) reduces to Mallows' (1973) CL under the following conditions: the approximating distribution (model) belongs to the Gaussian family, i.e.

$$W \sim N_m(m, A) \text{ and } Z \mid W \sim N\left(m(W), s^2\right) \tag{3.48}$$

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

$$f(W, Z; b_0) = g(W, Z), \tag{3.49}$$

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

$$E_{Z|W}\left\{Z - W^T b^*\right\} = 0 \text{ and } E_{Z|W}\left\{\left(Z - W^T b^*\right)\left(Z - W^T b^*\right)^T\right\} = s^2. \tag{3.50}$$

The log likelihood in this case is

$$\begin{aligned} \log f(Z \mid W; b) &= \log \frac{1}{\sqrt{2p\, s^2}} \exp\left(-\frac{1}{2s^2}\left(Z - W^T b\right)^T \left(Z - W^T b\right)\right) \\ &= \log \frac{1}{\sqrt{2p\, s^2}} - \frac{1}{2s^2}\left(Z - W^T b\right)^T \left(Z - W^T b\right) \end{aligned} \tag{3.51}$$

Its derivatives with respect to b are

$$\frac{\partial}{\partial b} \log f(Z \mid W; b) = \frac{1}{s^2} W\left(Z - W^T b\right) \text{ and} \tag{3.52}$$

$$\frac{\partial}{\partial b^T} \log f(Z \mid W; b) = \frac{1}{s^2}\left(Z - W^T b\right)^T W^T \tag{3.53}$$

and

$$\frac{\partial^2}{\partial b \partial b^T} \log f(Z \mid W; b) = -\frac{1}{s^2} W W^T . \tag{3.54}$$

Using the quadratic penalty (3.16), matrix J becomes

$$\begin{aligned} J &= -\frac{\partial^2}{\partial b \partial b^T} E_{W,Z}\left\{\log f_l^* - l p\left(b_l^*\right)\right\} \\ &= E_W\left\{\frac{1}{s^2} W W^T + l p'\left(b_l^*\right) p'\left(b_l^*\right)^T\right\} \\ &= \frac{1}{s^2} E_W\left\{W W^T\right\} + l p'\left(b_l^*\right) p'\left(b_l^*\right)^T \\ &= \frac{1}{s^2}\left(E_W\left\{W W^T\right\} + l I_m\right) \end{aligned} \tag{3.55}$$

and can be estimated as

$$\hat{J} = \frac{1}{s^2}\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T + l I_m\right) = \frac{1}{n s^2}\left(X^T X + n l I_m\right). \tag{3.56}$$

Matrix I becomes

$$\begin{aligned} I &= E_{W,Z}\left\{\frac{\partial}{\partial b}\left(\log f_l^* - l p\left(b_l^*\right)\right)\frac{\partial}{\partial b^T}\left(\log f_l^* - l p\left(b_l^*\right)\right)\right\} \\ &= E_{W,Z}\left\{\frac{\partial}{\partial b}\log f_l^* \frac{\partial}{\partial b^T}\log f_l^*\right\} - E_{W,Z}\left\{\frac{\partial}{\partial b^T}\log f_l^*\right\} E_{W,Z}\left\{\frac{\partial}{\partial b^T}\log f_l^*\right\} \\ &= \frac{1}{s^2} E_W\left\{W W^T\right\} + \frac{1}{s^4} E_W\left\{W W^T\left(b^* - b_l^*\right)\left(b^* - b_l^*\right)^T W W^T\right\} \\ &\quad - \frac{1}{s^4} E_W\left\{W W^T\right\}\left(b^* - b_l^*\right)\left(b^* - b_l^*\right)^T E_W\left\{W W^T\right\} \end{aligned} \tag{3.57}$$

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

$$\hat{I} = \frac{1}{n s^2}\sum_{i=1}^{n} X_i X_i^T = \frac{1}{n s^2} X^T X . \tag{3.58}$$

The trace term becomes

$$\begin{aligned} trace\left(\hat{I}\hat{J}^{-1}\right) &= trace\left(\frac{1}{n s^2} X^T X \cdot n s^2\left(X^T X + n l I_m\right)^{-1}\right) \\ &= trace\left(X^T X\left(X^T X + n l I_m\right)^{-1}\right) \\ &= trace(H) \end{aligned} \tag{3.59}$$

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

The RPSM becomes

$$\hat{l}_{MELL} = \arg\min_{l} \left\{ \frac{1}{2s^2} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - X_i^T \hat{b}_l \right)^2 + \frac{1}{n} trace(H) \right\} \tag{3.60}$$

or

$$\hat{l}_{MELL} = \arg\min_{l} \left\{ \frac{1}{n} \left\| Y - X\hat{b}_l \right\|^2 + \frac{2s^2}{n} trace(H) \right\}. \tag{3.61}$$

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.

3.3.3 Gaussian, misspecified case

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

$$J = \frac{1}{s^2} \left(E_W \left\{ WW^T \right\} + lI_m \right) \tag{3.62}$$

and estimated as

$$\hat{J} = \frac{1}{s^2} \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i^T + lI_m \right) = \frac{1}{ns^2} \left(X^T X + nlI_m \right). \tag{3.63}$$

Matrix I becomes

$$\begin{aligned} I &= E_{W,Z} \left\{ \frac{\partial}{\partial b} \log f_l^* \frac{\partial}{\partial b^T} \log f_l^* \right\} - E_{W,Z} \left\{ \frac{\partial}{\partial b^T} \log f_l^* \right\} E_{W,Z} \left\{ \frac{\partial}{\partial b^T} \log f_l^* \right\} \\ &= \frac{1}{s^4} E_{W,Z} \left\{ W \left(Z - W^T b^* \right)^2 W^T \right\} \end{aligned} \tag{3.64}$$

and is estimated as

$$\hat{I} = \frac{1}{s^4 n} \sum_{i=1}^{n} X_i \left(Y_i - X_i^T \hat{b} \right)^2 X_i^T \ . \tag{3.65}$$

The RPSM becomes

$$\hat{l}_{MELL} = \arg\min_{l} \left\{ \frac{1}{n} \left\| Y - X\hat{b}_l \right\|^2 + \frac{2s^2}{n} trace\left(\hat{I}\hat{J}^{-1} \right) \right\}. \tag{3.66}$$

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 $m(x) \equiv E\{Y_i \mid X_i = x\} \neq x^T b$ for any parameter $b \in R^m$.

3.3.4 Distributional misspecification

As mentioned already, distributional misspecification does not affect the estimation of the location parameter b. However, whenever 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 (3.3) are not consistent under distributional misspecification. To account for possible distributional misspecifications, the estimation of s^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 $Z \mid W$.

3.4 Information Complexity RPSM

With a limited number of observations, the inaccuracy penalization in (3.44) becomes inadequate and further refinement is needed. Starting from (3.44) and using Bozdogan's (1996) refinement argument, we obtain an Information Complexity Regularization Parameter Selection method (abbreviated ICOMPRPS) that behaves favorably for a limited number of observations.

Notice that the term $trace\left(IJ^{-1}\right)$ in (3.44) 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 (Urmanov and et. al., 2002)

$$ICOMPRPS(l) \equiv ALL(\hat{b}_l) - \frac{1}{n} trace(\hat{I}\hat{J}^{-1}) - \frac{1}{n} C_1(\hat{J}^{-1}) \tag{3.67}$$

and the corresponding RPSM is

$$\hat{l}_{ICOMPRPS} = \arg\max_l \left\{ ALL(\hat{b}_l) - \frac{1}{n} trace(\hat{I}\hat{J}^{-1}) - \frac{1}{n} C_1(\hat{J}^{-1}) \right\}, \tag{3.68}$$

where C_1 is the maximal covariance complexity index proposed by Emden (1971) to measure the degree of interdependency between parameter estimates. C_1 is a function of a covariance matrix and is computed as in (3.69) 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 $\hat{b}_l$ is approximately normally distributed, the maximal covariance complexity reduces to

$$C_1(\hat{J}^{-1}) = \frac{m}{2} \log \frac{\bar{n}_a}{\bar{n}_g}, \text{ where } \bar{n}_a = \frac{1}{m} \sum_{j=1}^{m} n_j, \; \bar{n}_g = \left(\prod_{j=1}^{m} n_j \right)^{\frac{1}{m}}, \text{ and} \tag{3.69}$$

$?_j$ are the eigenvalues of $\hat{J}^{-1}$.

In the Gaussian case, ICOMPRPS for Correctly specified Models (abbreviated ICOMPRPS-CM) becomes

$$ICOMPRPSCM(l) = \frac{1}{n} \left\| Y - X\hat{b}_l \right\|^2 + \frac{2s^2}{n} \left(trace(H) + C_1(\hat{J}^{-1}) \right) \tag{3.70}$$

and the corresponding RPSM is

$$\hat{l}_{ICOMPRPSCM} = \arg\min_l \left\{ \frac{1}{n} \left\| Y - X\hat{b}_l \right\|^2 + \frac{2s^2}{n} \left(trace(H) + C_1(\hat{J}^{-1}) \right) \right\}, \tag{3.71}$$

where

$$\hat{J} = X^T X + n l I_m \text{ and } H = X \left(X^T X + n l I_m \right)^{-1} X^T.$$

3.5 Minimum Mean Predictive Error RPSM

There is a strong bond between the RPSM's 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,

$$f(Y_i \mid X_i; b) \equiv N\left(X_i^T b, s^2\right), \tag{3.72}$$

then maximizing the MELL is equivalent to minimizing the MPE. This fact allows us to write an MPE analog of the information criterion (3.41). Indeed, using the Gaussian model, the ALL can be written as the sum of the training error (abbreviated TE) and a constant term

$$\begin{aligned} ALL\left(\hat{b}_l\right) &= \frac{1}{n}\sum_{i=1}^{n} \log f\left(X_i, Y_i \mid \hat{b}_l\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \log \frac{1}{\sqrt{2p}\, s^2} \exp\left(-\frac{1}{2s^2}\left(Y_i - X_i^T \hat{b}_l\right)^2\right) \\ &= \log \frac{1}{\sqrt{2p}\, s^2} - \frac{1}{n2s^2}\sum_{i=1}^{n}\left(Y_i - X_i^T \hat{b}_l\right)^2 \\ &= \log \frac{1}{\sqrt{2p}\, s^2} - \frac{1}{2s^2} TE\left(\hat{b}_l\right) \end{aligned} \tag{3.73}$$

where the training error is defined as

$$TE\left(\hat{b}_l\right) \equiv \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i^T \hat{b}_l\right)^2 . \tag{3.74}$$

The expected log likelihood for the Gaussian model is

$$
\begin{aligned}
ELL\left(\hat{b}_l\right) &= E_{W,Z}\ \log\ f\left(W,Z\mid\hat{b}_l\right)\\
&= E_{W,Z}\left\{\log\frac{1}{\sqrt{2p}\,s^{\,2}}\exp\left(-\frac{1}{2s^{\,2}}\left(Z-W^T\hat{b}_l\right)^2\right)\right\}\\
&= \log\frac{1}{\sqrt{2p}\,s^{\,2}}-\frac{1}{2s^{\,2}}E_{W,Z}\left\{\left(Z-W^T\hat{b}_l\right)^2\right\}\\
&= \log\frac{1}{\sqrt{2p}\,s^{\,2}}-\frac{1}{2s^{\,2}}E_{W,Z}\left\{\left(Z-m(W)\right)^2\right\}\\
&\quad -\frac{1}{2s^{\,2}}E_W\left\{\left(m(W)-W^T\hat{b}_l\right)^T\left(m(W)-W^T\hat{b}_l\right)\right\}\\
&= \log\frac{1}{\sqrt{2p}\,s^{\,2}}-\frac{1}{2}-\frac{1}{2s^{\,2}}PE\left(\hat{b}_l\right)
\end{aligned}
\tag{3.75}
$$

where the predictive error is defined as

$$
PE\left(\hat{b}_l\right)\equiv E_W\left\{\left(m(W)-W^T\hat{b}_l\right)^T\left(m(W)-W^T\hat{b}_l\right)\right\}. \tag{3.76}
$$

Plugging these representations into (3.41) an MPE analog of the information RPSM is obtained. The mean predictive error is approximated as

$$
E_D PE\left(\hat{b}_l\right)\approx E_D TE\left(\hat{b}_l\right)+\frac{2s^{\,2}}{n}trace\left(\hat{I}\hat{J}^{-1}\right)-s^{\,2}. \tag{3.77}
$$

Therefore, an unbiased estimator of the MPE is given by

$$
T_{MPE}(l)\equiv TE\left(\hat{b}_l\right)+\frac{2s^{\,2}}{n}trace\left(\hat{I}\hat{J}^{-1}\right)-s^{\,2} \tag{3.78}
$$

and the corresponding RPSM is

$$
\hat{l}_{MPE}=\arg\min_l\left\{TE\left(\hat{b}_l\right)+\frac{2s^{\,2}}{n}trace\left(\hat{I}\hat{J}^{-1}\right)-s^{\,2}\right\}. \tag{3.79}
$$

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

$$
CL(l)=TE\left(\hat{b}_l\right)+\frac{2s^{\,2}}{n}trace(H)-s^{\,2} \tag{3.80}
$$

with the corresponding RPSM:

$$\hat{l}_{CL} = \arg\min_{l} \left\{ TE\left(\hat{b}_l\right) + \frac{2s^2}{n} trace(H) - s^2 \right\}. \qquad (3.81)$$

3.6 Variability of Chosen Parameter

In the previous sections we chose the regularization parameter to maximize the MELL or minimize the MPE. It is important to realize that such a regularization parameter value maximizes the estimate of the MELL, not the MELL itself. As a result, for a given data set there is no guarantee that the chosen regularization parameter value will produce an admissible solution for the problem. A chosen regularization parameter value is an estimate of the 'true' parameter value that maximizes the MELL. The variability of that estimate may be large and diminish its usefulness. If the variability is too large, it questions any use of such a method for a practical application in which the true solution is unknown. In this case, there is no way to evaluate the validity of the regularized solution corresponding to the chosen regularization parameter value.

A big problem with RPSM's is underestimation. The chosen parameter value is often too small, presumably due to inadequate penalization of estimation inaccuracy for small data sets when correcting the bias of ALL in estimation of the MELL, and the corresponding solution is not 'smooth' enough and is still unstable. This makes it reasonable to consider other RPSM's that can provide estimates with smaller variance.

It seems a gain in precision can come only at the expense of introducing a bias. This is not a problem as long as the method remains convergent though it can lose its optimality in some sense. Besides, as with any regularized solution, the probability that a regularization parameter chosen with some RPSM will hit in a certain vicinity of the 'true' regularization parameter value may be larger with a RPSM that produces a biased regularization parameter estimate. In light of underestimation, a bias in a safe direction of

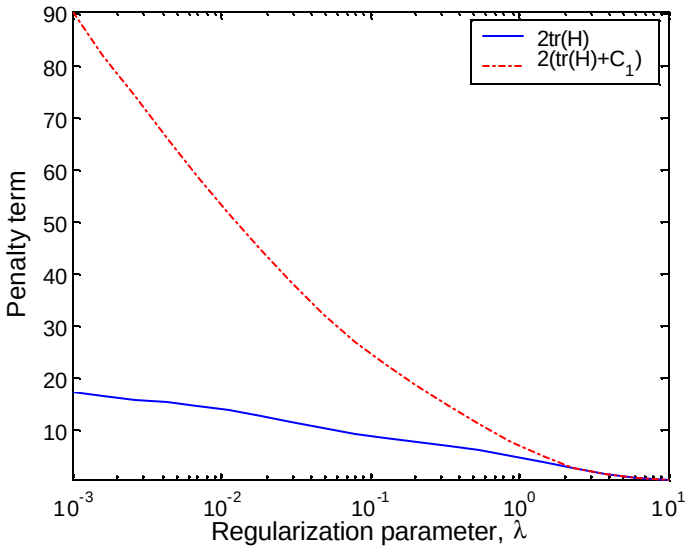

Figure 3.1. The trace part of the RPSM.

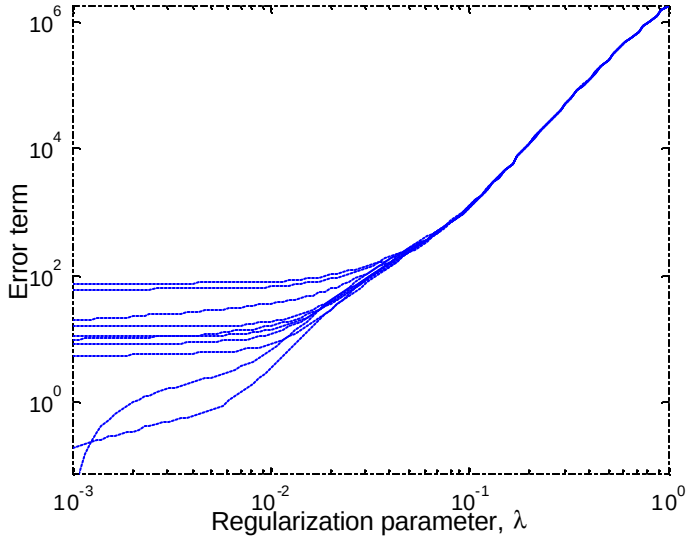

Figure 3.2. The SSR part of the RPSM's.

lager values can help reduce the risk of obtaining an underregularized solution. This is particularly important when the solution has no physical interpretation as in many data-driven black-box techniques and it would be extremely difficult to assess a proper degree of regularization.

A valuable advantage of ICOMPRPS as a RPSM is the additional penalty that significantly reduces the risk of obtaining grossly underregularization solutions. $C_1\left(\hat{J}^{-1}\right)$ is a monotonically decreasing function of the regularization parameter l. In addition, it decreases faster than $trace\left(\hat{I}\hat{J}^{-1}\right)$, so for small values of l, the correction is significant, while for larger l, the correction is negligible. This is reasonable, since for small values of l, the corresponding regularized solution is close to the OLS solution which has large variance, and penalization of that inaccuracy with only the trace term is inadequate. With decreasing l, the effective number of parameters approaches the number of estimated parameters in the model or, equivalently, the number of observations per effective parameter is reduced so that the asymptotic results used in the derivations of the RPSM's become less applicable. That is why for ill-conditioned problems with limited number of observations, a much stronger penalization of estimation inaccuracy is beneficial in many respects. The following example demonstrates such an extra penalization of estimation inaccuracy.

Figure 3.1 shows the effect of introducing C_1 on the choice of the regularization parameter value. Since for fixed X, the trace term shown in Figure 3.1 does not depend on a particular realization of the noise vector in the response variable, it does not contribute to the variability of the regularization parameter estimate $\hat{l}$. The only contribution comes from the sum of squared residuals term shown in Figure 3.2. C_1 corrects the trace part only for small values of l and reduces the variability of the chosen regularization parameter, reducing the chance of selecting smaller values as shown in Figure 3.3 and Figure 3.4.

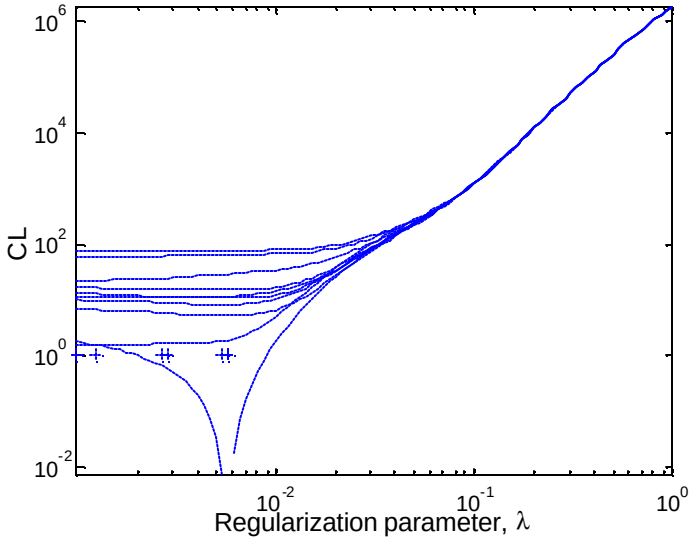

Figure 3.3. CL vs. regularization parameter for 10 realizations of noise.

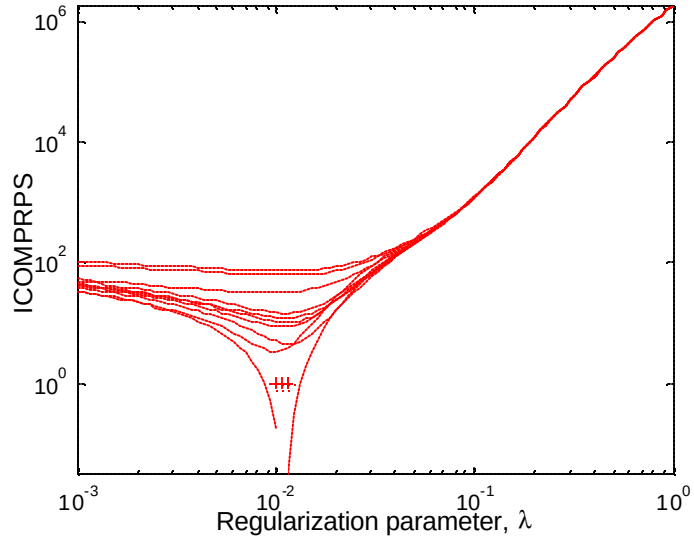

Figure 3.4. ICOMPRPS vs. regularization parameter for 10 realizations of noise.

For 10 realizations of the noise component in the response, CL shown in Figure 3.3 and ICOMPRPS shown in Figure 3.4 were computed as functions of the regularization parameter. The crosses mark the regularization parameter values at which the minimum of CL and ICOMPRPS occurs for each noise realization. For CL, the chosen values are spread from zero to 10^{-2}. For ICOMPRPS, all 10 chosen values are concentrated around 10^{-2}. This demonstrates the much lower variability of the ICOMPRPS-chosen regularization parameter value compared with that chosen by CL. The lower variability drastically reduces the risk of underregularization, though some optimal properties of such methods as CL may not be shared.

3.7 Regularization Parameter Selection For Misspecified Models

In this section we show that the choice of the regularization parameter using CL or ICOMPRPS-CM becomes inconsistent whenever the model is misspecified while the choice using misspecification-resistant methods such as RIC and ICOMPRPS remains consistent and produces the closest approximating model.

Given the simple data-generating process

$$y_t = x_t + x_t^2 + e_t \tag{3.82}$$

we assume that x is measured with 3 redundant sensors with some measurement errors so that the data set

$$X = \left\{\left(x_{1t}, x_{2t}, x_{3t}, x_{1t}^2, x_{2t}^2, x_{3t}^2\right)\right\}_{t=1}^{n}, \tag{3.83}$$

where x_1, x_2, x_3 are noisy copies of x, is ill-conditioned due to highly collinear variables. A simple reason for keeping all the redundant variables in the data set is to increase robustness or to make more reliable predictions when one or more sensors fails.

Due to the presence of collinearity, the OLS solution is highly oscillatory and will not produce stable predictions on future observations, especially in situations when the

collinearity pattern changes due to a failing sensor. To avoid such instability and hypersensitivity, regularization (ridge regression) is used to obtain "low-energy" regression coefficients that produce very stable predictions. As usual, to obtain a good regularized solution (regression coefficients) a proper regularization parameter value must be chosen.

Two cases of choosing the regularization parameter value for correctly and incorrectly specified models using CL, ICOMPRPS-CM, RIC, and ICOMPRPS are examined. The two competing models are:

$$y_t = a_1 x_{1t} + a_2 x_{2t} + a_3 x_{3t} + h_t, \tag{3.84}$$

$$y_t = a_1 x_{1t} + a_2 x_{2t} + a_3 x_{3t} + b_1 x_{1t}^2 + b_2 x_{2t}^2 + b_3 x_{3t}^2 + e_t. \tag{3.85}$$

Model (3.84) is misspecified while model (3.85) is correctly specified. The misspecified model (referred to as linear) lacks quadratic terms and can only give a linear approximation to the true relationship. The error term of this model includes the true stochastic noise and the modeling error as well. The correct model (referred to as quadratic) has no functional misspecification and can suffer only from the collinearity. The error term of this model includes only the random error in the response.

If least squares is used to solve for the coefficients for both models and these OLS solutions are used to produce predictions on future observations, very unstable results are obtained as expected. Figure 3.5 demonstrates the OLS predictions by the misspecified model. Figure 3.6 demonstrates the OLS predictions by the correct model. The generated data set is mildly ill-conditioned with a condition number of 10^4. In Figure 3.5, the OLS predictions for the training data (the dash-dot line) are quite good and give a fairly good linear approximation to the true quadratic relationship (the solid line). However, predictions for new, unseen data (the noisy dashed line) are very unstable as a result of highly oscillatory solutions.

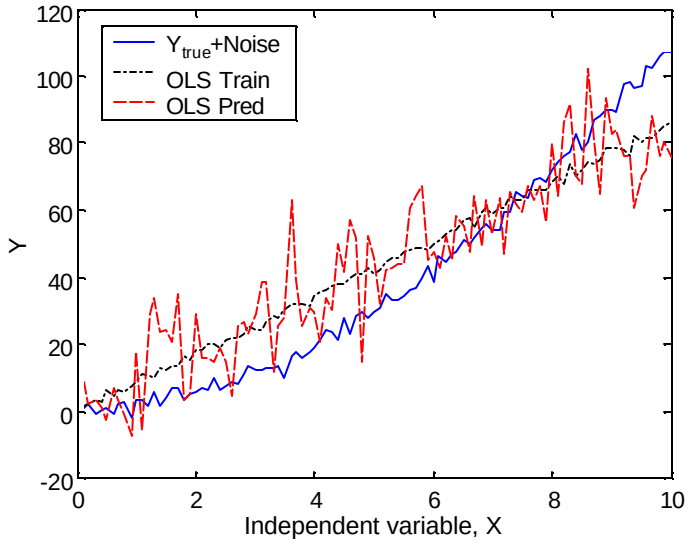

Figure 3.5. OLS predictions by the misspecified model.

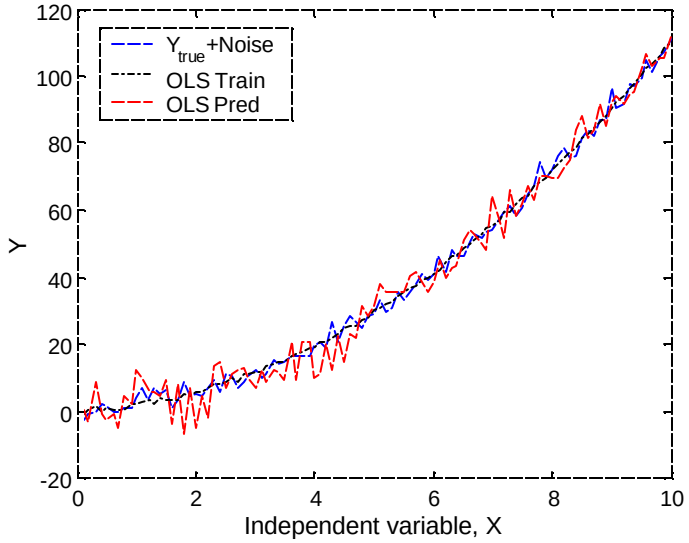

Figure 3.6. OLS predictions by the correct model.

The OLS regression coefficient estimates for both models are given below.

Linear	609	283	-884			
Quadratic	87	-507	420	7.9	16.6	-23.6.

In Figure 3.6, the OLS predictions (the dash-dot line) of the correct model are very good for the train data; however, for the new data, predictions (the dashed line) are still very noisy. Regardless of correct specification, the OLS predictions are not stable and are of little practical value.

To improve the situation and get more stable predictions we use ridge regression and select the regularization parameter value using Mallows' CL, ICOMPRPS-CM, RIC, and ICOMPRPS. The behavior of these RPSM's for the misspecified model is shown in Figure 3.7. For the misspecified model, CL and RIC are very different and have different minimizers. ICOMPRPS-CM and ICOMPRPS are different as well. For the correct model, CL and RIC, shown in Figure 3.8, are almost identical and have the same minimizer. ICOMPRPS-CM and ICOMPRPS are also identical for the correct model.

Since CL and ICOMPRPS-CM are not misspecification-resistant, their choice of the regularization parameter is no longer valid for the misspecified model. For the correct model, CL and RIC are almost identical. This fact could be used for detection of possible model misspecification. Basically, an identical behavior of CL and RIC can indicate that the model (the functional relationship between the predictors and the response) is correct; otherwise, the model might be misspecified. For example, some relevant variables might be missing. The ICOMPRPS-CM and ICOMPRPS pair can be used for misspecification detection as well.

As a noise level, we use an estimated noise variance ($\hat{s}^2_{NOISE} = 3.4$) from the response which is much less than the OLS estimate ($\hat{s}^2_{OLS} = 132$) using the linear model, because the OLS estimate with linear model also includes the modeling error.

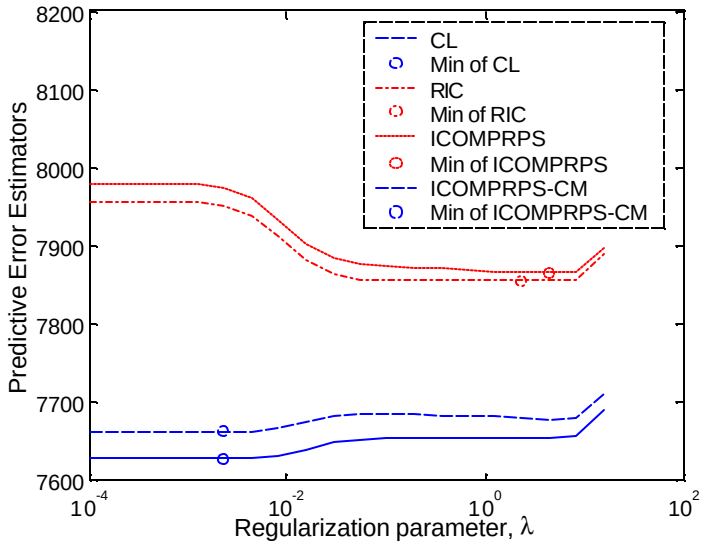

Figure 3.7. Behavior of the RPSM's for the misspecified model.

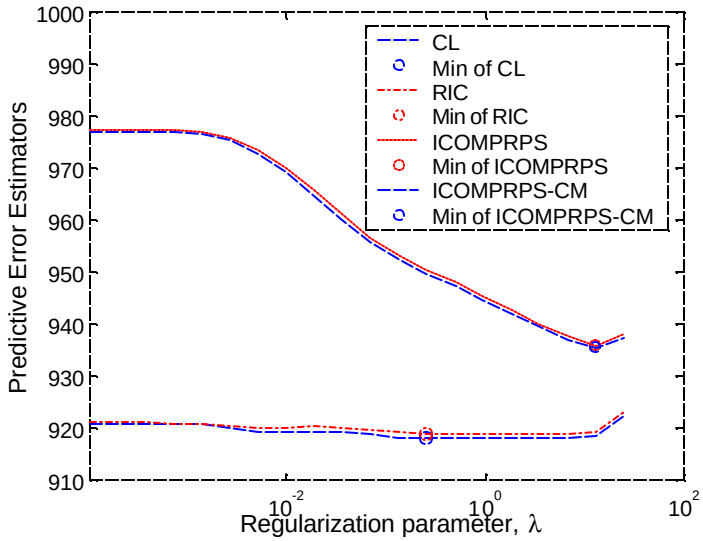

Figure 3.8. Behavior of the RPSM's for the correct model.

Table 1. Simulation results for the misspecified model.

RPSM	CL	RIC	ICOMPRPS-CM	ICOMPRPS
Chosen λ	0.002	2.4	0.002	4.4
Solution	575 273 -840	2.8 2.8 2.8	575 273 -840	2.8 2.8 2.8
Prediction MSE	227	111	227	111

Table 2. Simulation results for the correct model.

RPSM	CL	RIC	ICOMPRPS-CM	ICOMPRPS
Chosen λ	0.25	0.25	13	13
Solution	0.4 -0.2 0.5 2.6 -3.8 1.6	0.4 -0.2 0.5 2.6 -3.8 1.6	0.21 0.21 0.21 0.35 0.35 0.35	0.21 0.21 0.21 0.35 0.35 0.35
Prediction MSE	0.29	0.29	0.16	0.16

Since RIC and ICOMPRPS were derived under possible model misspecification, they can detect the discrepancy between the supplied noise level and the one estimated using the current model (which is done implicitly in RIC and ICOMPRPS), and use this discrepancy to their advantage.

As a result of model misspecification, in this particular example, CL and ICOMPRPS-CM consistently failed to choose a reasonable regularization parameter value while RIC and ICOMPRPS performed well and delivered the regularization parameter values that produce good linear approximations (the dashed line in Figure 3.9) to the true relationship. The predictions corresponding to the values chosen by CL (the dash-dot line) and ICOMPRPS-CM are almost as bad as OLS predictions. The regularization parameter values chosen by the RPSM's, the corresponding regularized regression coefficients, and the mean square error on the test data for the misspecified model are shown in Table 1.

For the correct model, all the methods give regularization parameter values, which correspond to stable predictions shown in Figure 3.10 that are almost identical to the true relationship. The regularization parameter values chosen by the RPSM's, the corresponding regularized regression coefficients, and the mean square error on the test data for the correct model are shown in Table 2.

This result, though on artificial data, demonstrates the superior performance of the information-based RPSM's, which are much more reliable than CL and perform well in more realistic situations in which models are rarely correctly specified. Also, in many engineering applications correctly-specified models are extremely unusual; therefore, misspecification-resistant RPSM's are of great value. They consistently deliver proper regularization parameter values regardless of misspecification and produce good fits when the model is correct or good approximations when the model is not correct.

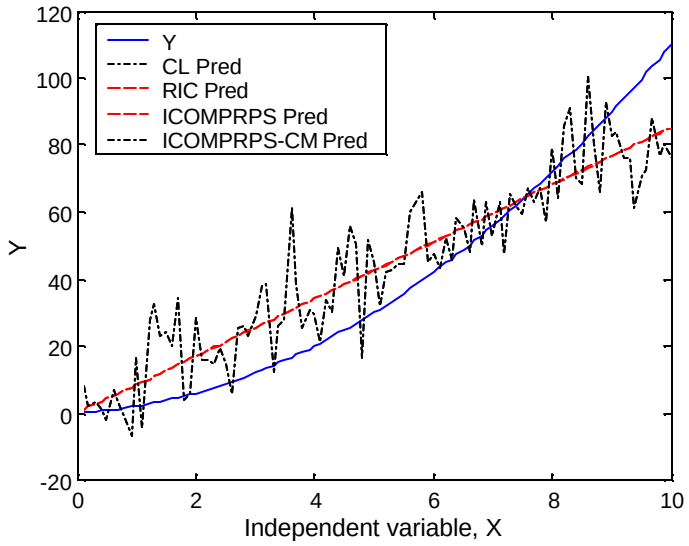

Figure 3.9. Regularized predictions by the misspecified model.

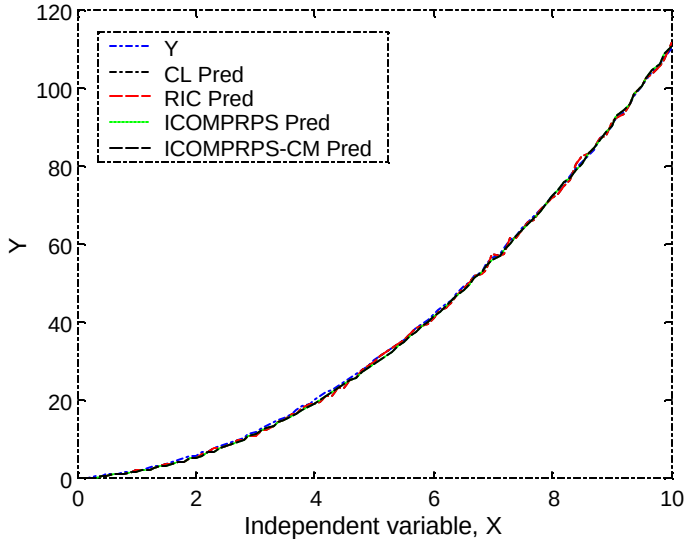

Figure 3.10. Regularized predictions by the correct model.

CHAPTER 4
PRACTICAL APPLICATIONS

We present several examples of practical applications using the proposed ICOMPRPS method and comparing it against CL and other RPSM's. The examples are

1. Venturi Meter Drift Prediction
2. Sensor Validation System
3. Statistical Learning from Data (Radial Basis Function Neural Network)
4. Numerical Solution of an Integral Equation
5. Image Reconstruction
6. Specification of Prior Distribution in Bayesian Inference.

The first example is a construction of an inferential system for venturi meter drift prediction. The data are from Carolina Power & Light's Crystal River Nuclear Power Plant. Because the data set is composed of correlated sensor measurements, the OLS solution is very unstable. Regularization is required to build a reliable inferential system that uses measurements of other sensors to infer the value of the venturi meter. A linear regression model used in this example may be misspecified due to missing variables or possible nonlinear relationships between sensors.

The second example is a construction of a sensor validation system. Measurements of 83 sensors from a TVA fossil power plant are available. The sensors represent different plant variables. The problem is to build an inferential model that uses 82 sensors as predictors to infer the value of the remaining sensor and see if the predicted values are significantly different from the actual measured values. The main problem in building such a model is the hypersensitivity of the solution to noise in the data because

of ill-conditioning. As a result, if one of the sensors fails, the inference about the sensor being monitored becomes invalid. Regularization is required to obtain a stable solution. A proper value of the regularization parameter must be found.

The third example is statistical learning from data. Building a parametric or non-parametric model is an inverse problem. In many situations, it is also an ill-posed problem. This example demonstrates a use of nonparametric technique such as a radial basis function neural network, which will be described, to fit an unknown underlying relationship using only the observed data. The OLS solution, the network's weights, in this case is inadmissible because, due to the nature of the technique, the OLS solution only minimizes the SSR and overfits the data. As a result, the solution is highly unstable and produces useless predictions. On the other hand, regularization produces very stable solutions which, for a suitably chosen regularization parameter value, are very close to the true relationship.

The fourth example is the numerical solution of an integral equation that arises in many practical applications such as image restoration, heat-transfer calculations, and others. The RPSM's are tested for different noise levels in the response and for the case of noise level underestimation. The ICOMPRPS choice method is shown to be superior in the case of noise level underestimation. The example also demonstrates that CL becomes inconsistent when the model is misspecified while the ICOMPRPS method produces regularization parameter values corresponding to admissible solutions.

The fifth example is a test image-deblurring problem provided by Dr. Vogel of Montana State University. The OLS image reconstruction does not produce anything even remotely resembling the original image. Regularization produces admissible solutions which can be recognized visually as similar to the original image.

The sixth example is the specification of prior distribution in Bayesian inference. One of the major components of Bayesian analysis is the prior distribution of the

parameters which, combined with the likelihood, is used to produce the posterior distribution of the parameters and then the predictive distribution of the output. In some applications, parameters are unobservable and have no physical interpretation. In this case, it is difficult to assign, a priori, any prior distribution to the parameters and justify that choice. However, the output is observable and information about it is available prior to modeling. This information can be used for assigning the prior distribution to the unobserved parameters by solving an inverse ill-posed problem.

4.1 Venturi Meter Drift Prediction

The majority of Pressurized Water Reactors (abbreviated PWR) utilize venturi meters to measure feedwater flow rate, which is used to estimate reactor thermal power. However, venturi meters are susceptible to measurement-drift due to corrosion products building up near the meter's orifice. This fouling increases the measured pressure drop across the meter, which in turn results in an over-estimation of the flow rate. Consequently, reactor thermal power is also overestimated. To stay within regulatory limits, reactor operators have to derate their plants or justify a compensating process. On average, the amount of derating varies from insignificant to 3% of full power. For example, a derating of 2% in an 800 MWe unit will cost the utility ~$20,000 per day.

To overcome this problem, an inferential sensing system has been developed at the University of Tennessee to infer the true Feedwater Flow Rate (abbreviated FFR) (Upadhyaya, 1994; Gribok, 2001). Twenty-four (24) plant variables have been selected as predictor variables based on engineering judgment and their high correlation with feedwater flow rate, and a linear regression model has been chosen as a predictive tool. The predicted value of feedwater flow rate is not affected by fouling because the linear regression model is built on the data corresponding to the initial operation time period

before fouling starts. The difference between the predicted value and the actual measurement defines a drift due to fouling.

The data set contains 24 plant variables highly correlated with FFR. The description of the variables is given in Appendix A.2. Before modeling, the data shown in Figure 4.1 were preprocessed (median filtered, centered and range-scaled). The first 601 data points represent 12 days of operation starting with a new fuel cycle; the first 8640 data points represent 6 months of operation. FFR is shown in Figure 4.2. The first 601 data points are used to build an inferential model. Although fouling can occur as soon as on the 3rd day of operation, these 12-day measurements are assumed to be free of fouling.

After an inferential model was build using Procedure 1, the model was used to predict (infer) future values of FFR and compare them with actually measured ones. At some point, the predicted FFR begins to deviate from the measured FFR, indicating the beginning of fouling. The difference between the predicted and measured FFR, averaged over a range of 100 data points to eliminate the random component and pick up only the systematic deviation, defines the drift value. The reported drift is estimated in the range 8601-8700; which is about 6 months into the fuel cycle.

Without using any regularization, the solution is very unstable and produces drift values ranging from negative as shown in Figure 4.3 to positive as shown in Figure 4.4, and no drift as shown in Figure 4.5. Such unstable drift predictions are obtained because the data matrix is composed of measurements of correlated sensors is ill-conditioned, and the resulting OLS predictions are hyper sensitive to the number of training points and to the filter window width. Using a regularization method, stable feedwater flow rate estimation can be achieved, which is stable with respect to the number of training points and filter window width.

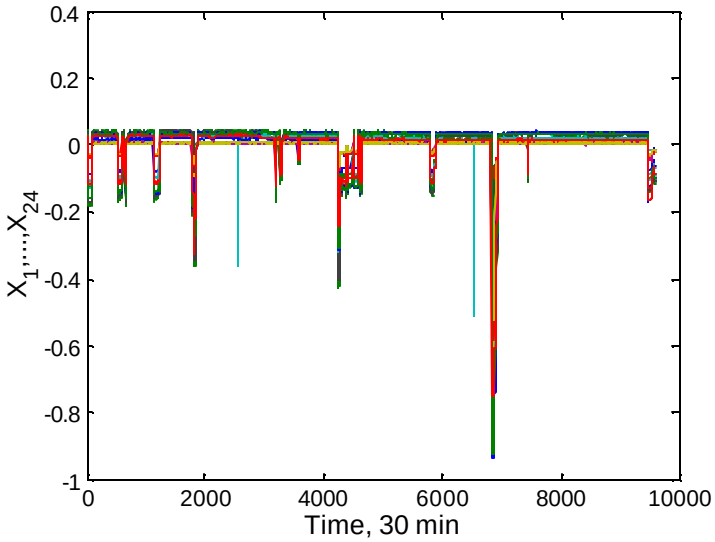

Figure 4.1. 24 preprocessed predictor variables.

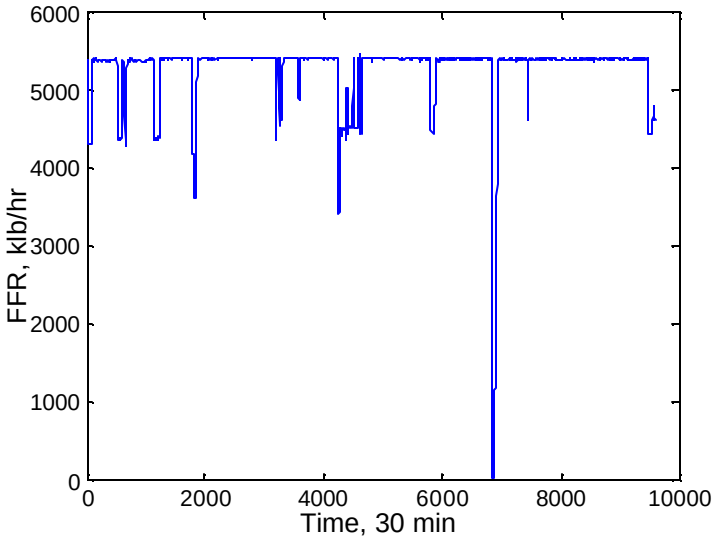

Figure 4.2. FFR filtered measurements.

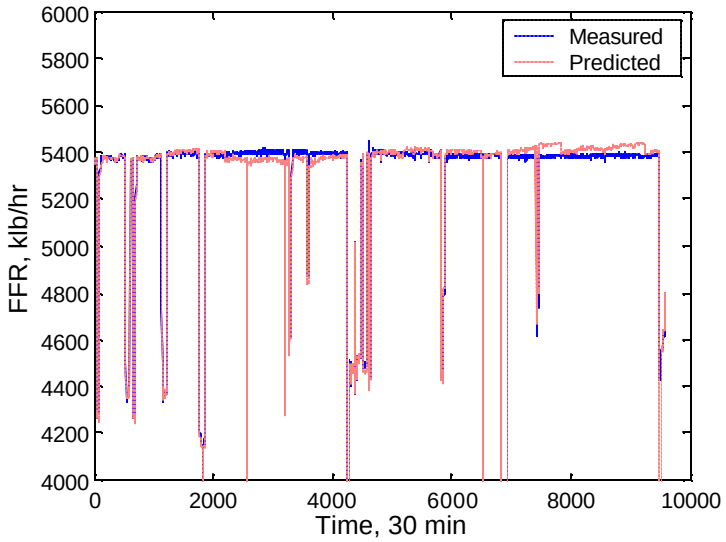

Figure 4.3. Drift prediction by the OLS method. (Negative drift of 31 klb/hr).

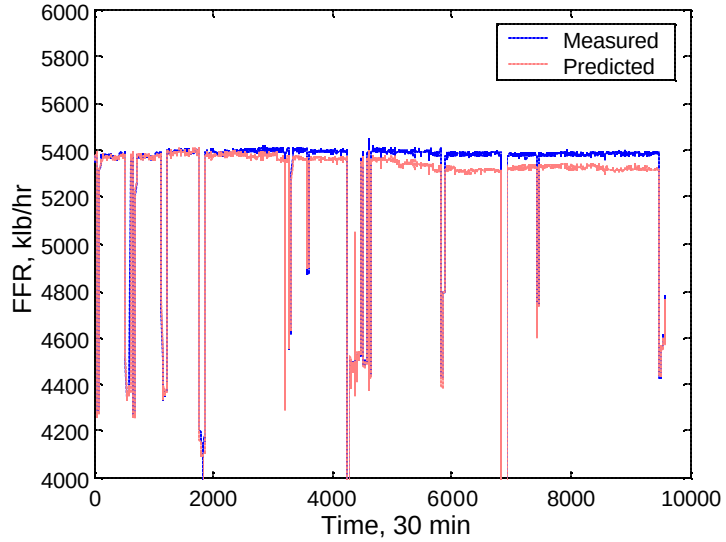

Figure 4.4. Drift prediction by the OLS method (Positive drift of 69 klb/hr).

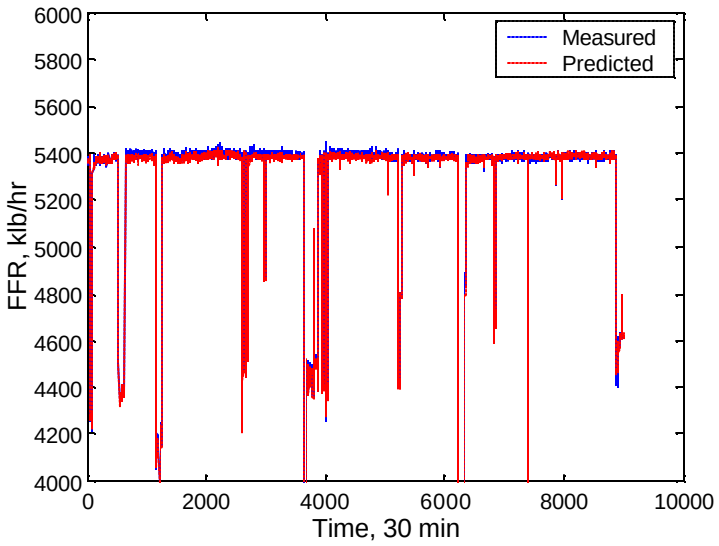

Figure 4.5. Drift prediction by the OLS method (Zero drift).

In Figure 4.6, estimated probability densities are shown for the unstable and stabilized feedwater flow rate estimation. The densities are obtained using the bootstrap method in which we resample N times a certain number of observations from the same data set, obtain N drift values, and estimate the drift's probability density using these N values. The standard deviation of the OLS (unstable) drift prediction is 20 times standard deviation of the ridge (stabilized using ridge regression) drift prediction. The ridge prediction is extremely stable, however it seems biased from the OLS prediction. However, according to the ridge theorem (Hoerl, 1970), the prediction error using ridge regression coefficients with a suitably chosen ridge parameter is less than the prediction error using OLS regression coefficients. This means that the ridge drift prediction lies closer to the true drift than OLS drift prediction.

Although the drift prediction has been stabilized, there is still some uncertainty involved regarding the proper variable subset to be used for predictions. Different variable subsets used as predictors can produce different regularized drift values. Each variable subset should be evaluated by some criterion that estimates the prediction error. The subset with the lowest value of the criterion is chosen to be the best in the sense of

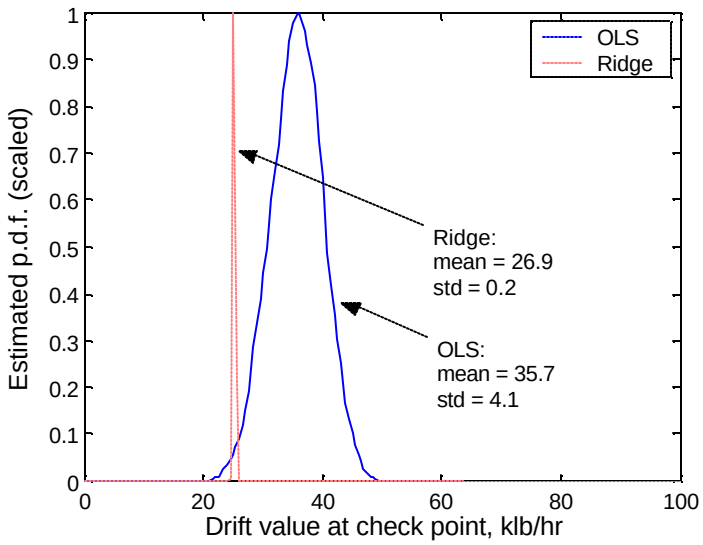

Figure 4.6. Unstable and stabilized predictions of the venturi meter drift.

having minimum prediction error or minimum prediction risk. Since the considered criteria such as CL, RIC, and ICOMPRPS compare MPL models, i.e. stabilized models, the value of the ridge parameter should be chosen for each subset individually by using one of the available ridge parameter selection methods.

It is not unusual to encounter a situation in which several models are equally good according to a chosen model selection criterion. For example, several models can have approximately the same estimated prediction risk. If we are interested in selecting the best prediction model we cannot prefer one model to the others. In this situation one can use model averaging procedures. A most common averaging procedure is Bayesian model averaging (Leamer, 1978) in which each model's prediction is weighted according to its posterior distribution. A naïve approach one can entertain when using non-Bayesian model selection is to weight them according to the criterion value. However, since we have to average models with approximately the same prediction risk (abbreviated PR) value, which can be computed as the mean predictive error, we give them the same weight and take the average prediction over all models as our final prediction of the drift:

$$p_f = \frac{1}{n}\sum_{i=1}^{24} p_i \,. \tag{4.1}$$

Since there is some uncertainty s_i associated with each individual prediction, the uncertainty in the final prediction will be

$$s_f = \frac{1}{m}\sqrt{\sum_{i=1}^{m} s_i^2} \,. \tag{4.2}$$

The model (variable subset) selection procedure employed here is summarized as follows:

Procedure 1

- For each number of predictor variables j=1…24
    - For each subset of j variables
        - Find the optimal value of the ridge parameter l ,
        - Assign the corresponding estimated PR value to that subset,
        - Update the best (with lowest PR) model of j variables.
    - End
    - Save the best model with j variables
- End.
- Choose the best model (with lowest PR) as the best predictive model.
- If several models have approximately the same PR, perform averaging.

In performing model selection with this technique, no information is used other than that extracted from the available data by various statistical methods. This approach is oriented to a situation in which there is no method to obtain additional information about the process under consideration. It is possible that a successful inferential system may be built using additional information about the specific operating environment.

The potential predictor variables must be preselected based on their physical relevance to FFR and on other factors, as partially is done in the following example; it might be necessary to consider a compensating effect of the control system on the plant

variables and take that into account. If such information is available, it can be successfully incorporated into the inferential system by means of Bayesian approaches. For example, if it is known from historical operation that the venturi meter drift lies in the range 1-3%, this can be used to dispose of the models that predict drift outside that range.

A prior distribution of the parameters could also be specified to incorporate other available information into the estimation. One can argue that since the parameters are usually unobservable, it would be difficult to justify any choice of the prior distribution. However, the information about the possible drift range can be transferred to the prior distribution of parameters by using the backward prior specification method proposed by Gribok et. al. (2002). This could also reduce uncertainty in the predictions.

After applying Procedure 1 to model-building, we obtain the following results summarized in Table 3. In the table, [ ] shows variables excluded from the subset, and ( ) shows a different variable chosen by CL. The drift uncertainty shown in the table accounts only for the uncertainty due to inaccurate estimation of parameters.

Each subset in the j-th row represents the best subset of j variables. All possible subsets of j variables form a model group; inside each group ICOMPRPS and CL vary significantly. For example, for 3-variable subsets ICOMPRPS varies from 6226.6 to 6204.0, making subset selection inside the group meaningful (or significant). We found that the best models from all 24 groups have approximately the same value of ICOMPRPS equal to 6204 and CL equal to 6200. The exceptions are the two extremes, 1- and 24-variable subsets, which have significantly different ICOMPRPS and CL values. The one-variable case should be disregarded because the predicted drift value is negative, which is impossible.

Table 3. Variable subsets evaluation results.

	Best subsets of j variables	ICOMP RPS	CL	α	Drift +/- 3σ
1	21	6201.9	6201.9	0.01	-3.4 +/- 0.6
2	15 21	6204.1	6199.8	0.14	20.9 +/- 0.9
3	1 15 21	6204.0	6199.5	0.14	21.0 +/- 0.9
4	1 4(12) 15 21	6204.1	6199.6	0.17	11.7 +/- 1.2
5	1 3(7) 4 15 21	6204.1	6199.6	0.17	12.1 +/- 1.2
6	1 3 4 7 15 21	6204.2	6199.7	0.17	12.6 +/- 1.2
7	1 3 4 5 7 15 21	6204.2	6199.7	0.17	15.5 +/- 1.5
8	1 3 4 5 7 10 15 21	6204.3	6199.8	0.18	14.4 +/- 1.2
9	1 3(22) 4 5 7 10 12 15 21	6204.3	6199.8	0.18	13.2 +/- 1.2
10	1 3(13) 4 5 7 9(22) 10 12 15 21	6204.4	6199.9	0.18	11.9 +/- 1.2
11	1 3(22) 4 5 7 9 10 12 13 15 21	6204.4	6199.9	0.18	10.8 +/- 1.2
12	1 3(22) 4 5 7 9 10 11 12 13 15 21	6204.5	6199.9	0.18	9.6 +/- 1.5
13	1 3(8) 4 5 7 9 10 11 12 13 15 21 22	6204.6	6200.0	0.22	6.1 +/- 1.2
14	1 3(17) 4 5 7 8 9 10 11 12 13 15 21 22	6204.6	6200.0	0.22	5.4 +/- 1.2
15	1 3 4 5 7 8 9 10 11 12 13 15 17 21 22	6204.7	6200.1	0.22	5.7 +/- 1.2
16	1 3 4 5 6(20) 7 8 9 10 11 12 13 15 17 21 22	6204.7	6200.2	0.22	6.2 +/- 1.2
17	1 3 4 5 7 8 9 10 11 12 13 15 17 18(6) 19 21	6204.8	6200.2	0.24	17.4 +/- 1.5
18	1 3 4 5 7 8 9 10 11 12 13 15 17 18(6) 19 20 21 22	6204.8	6200.3	0.24	17.3 +/- 1.5
19	[2 6 16 18 23]	6204.9	6200.4	0.29	31.8 +/- 1.2
20	[2 16 17 18]	6204.9	6200.4	0.32	28.3 +/- 1.2
21	[2 16 18]	6205.0	6200.5	0.32	28.3 +/- 1.2
22	[2 16]	6205.2	6200.7	0.32	32.8 +/- 1.5
23	[2]	6205.6	6201.1	0.35	35.8 +/- 1.5
24	All	6206.7	6202.1	0.35	39.9 +/- 1.5
	Average excluding subsets 1 and 24				16.8 +/- 0.3

The negative drift value demonstrates that it is useful to incorporate available information in the inference either in the form of constrains or in the form of prior distributions, to automatically remove meaningless models from consideration. Approximately equal values of the criteria mean that all 22 models are equally good in prediction: they have the same estimated prediction error or prediction risk. One possible explanation is that the effective number of parameters (variables) for all 24 models is about 1, meaning that all models essentially contain the same amount of information extracted from the corresponding input variables. This means that they all have equal performance and should be averaged.

After averaging, the resulting predicted drift value is 16.8 +/- 0.3 klb/hr. Notice that we obtain a stable and unique drift estimation, which is robust to the number of variables used in the model and the particular noise realization in the data. Such drift estimation corresponds to the predicted FFR values of 5369.5 klb/hr whereas the measured FFR is 5386.3. Notice that the detected drift value is less than the accuracy of the measurement instrument (the venturi meter), which is about 3%. This additional accuracy is possible because before modeling, the random component in FFR is partially filtered out with a median filter. The systematic component (the drift) is not affected by median filtering and is still present in the data.

4.2 Sensor Validation

A sensor validation system usually employs an inferential model that uses measurements of correlated sensors to predict the value of the sensor being monitored. To take into account all available information, it is necessary to include all available sensors that are correlated with the one being predicted. This results in an ill-conditioned data matrix of predictors, due to the inclusion of correlated predictors. A linear regression

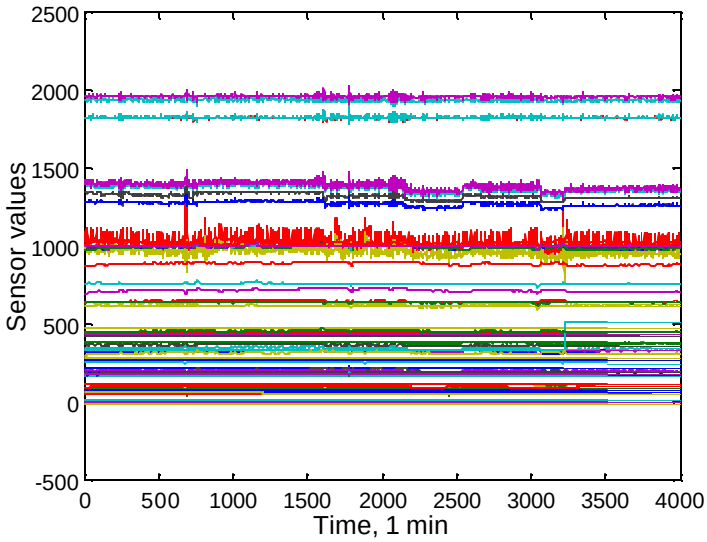

Figure 4.7. 82 Sensors used as predictors.

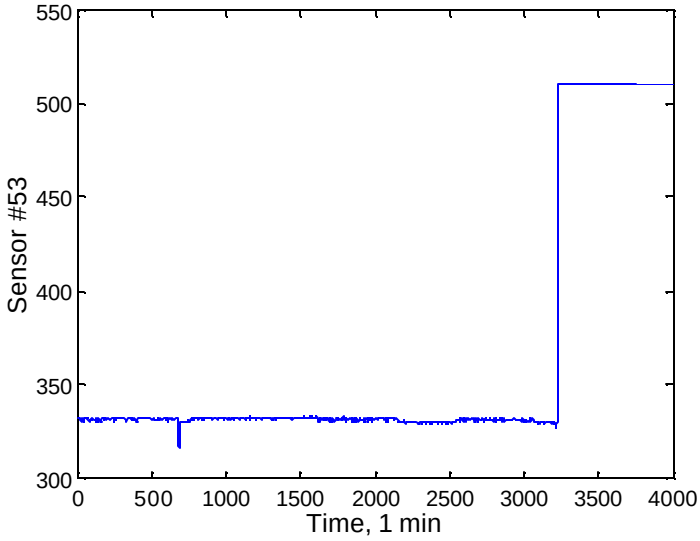

Figure 4.8. Sensor #53.

model is usually used. The OLS solution (or the OLS regression coefficients) in this case may be unstable and statistically insignificant. Predictions using the OLS solution are also unstable and unreliable. If one of the sensors fails, the entire system is destroyed and wrong results are produced that could lead to wrong conclusions. Regularization in the form of ridge regression produces more stable solutions and improves the reliability of the sensor validation system.

The data set contains measurements of 83 sensors from a Fossil Power Plant. The sensors represent various plant variables and must be monitored to detect sensor failures. To simplify the problem, we consider monitoring of only 1 sensor, namely sensor number 1, and use the other 82 sensors as predictors in a linear regression model whose coefficients are estimated using the ridge estimator. Figure 4.7 shows all 82 sensors.

Sensor number 53 which failed at about the 3300^{th} measurement is shown in Figure 4.8. The condition number of the data matrix is 547920. The problem is ill-conditioned, and regularization is required to obtain a stable solution. The singular values of the data matrix are plotted in Figure 4.9. The singular value spectrum decays gradually. Therefore, it is difficult to decide on a cut-off number of principal components to be retained in the model. Instead, ridge regression is used, in which all the components are retained, but with different filter factors. For ridge regression, the proper value of the regularization parameter must be selected. The ridge parameter is chosen using four different estimators of the mean predictive error. Namely, CL, RIC, ICOMPRPS, and ICOMPRPS-CM are used. The noise variance of 0.1 used in the RPSM's was roughly estimated as the variance of sensor measurements during a short period of time.

In Figure 4.10, the values of the CL, RIC, and ICOMPRPS are plotted versus the regularization parameter value.

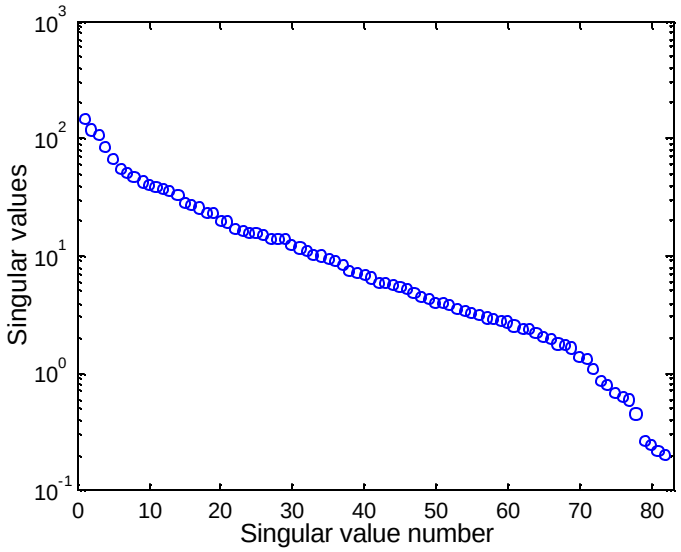

Figure 4.9. Singular values of the data matrix.

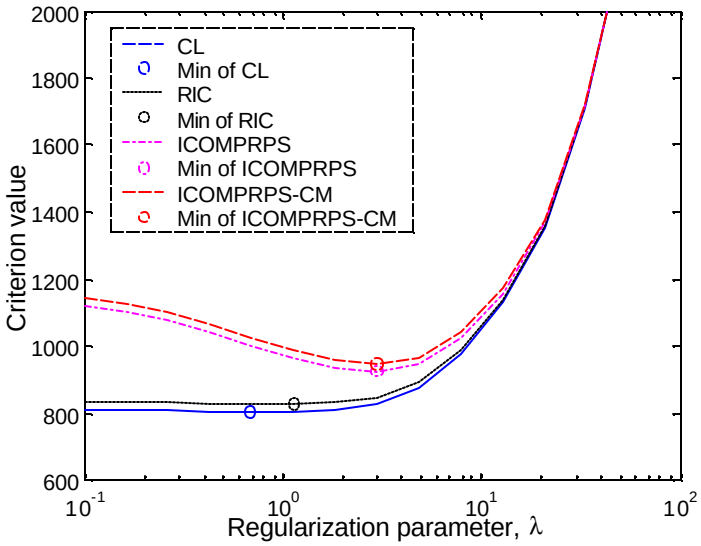

Figure 4.10. Regularization parameter selection.

Table 4. Results of regularization parameter selection using different methods.

Method	Regularization parameter, λ
CL	0.6951
RIC	1.1288
ICOMPRPS	2.9764
ICOMPRPS-CM	2.9764

The values of the regularization parameter that correspond to the minimum of the criteria are chosen as proper regularization parameter values. These values are shown in Table 4. The ICOMPRPS-chosen value indicates that about the first 50 components were passed and the rest were dumped. RIC and CL passed many more components than 50. The corresponding solutions in Figure 4.11 show that the OLS solution is highly oscillatory. For such solutions, unstable predictions are expected. The CL and RIC solutions also have fairy large values, so predictions using these solutions are expected to be unstable as well.

The sensor value to be monitored is shown in Figure 4.12. In Figure 4.13, predictions using the OLS solution are shown. As expected, predictions using the OLS solution are unstable and become irrelevant at the point where one of the sensors used as a predictor failed. This occurred because OLS does not use the available information optimally; instead, it overuses and underuses certain inputs. The failure of the 53rd sensor resulted in an invalid inference about the sensor being monitored. The prediction inaccuracy increased following the failing 53^{rd} sensor. This happened because of large OLS regression coefficients that make predictions sensitive to minor changes of the input variables. Once the collinearity pattern was destroyed (one of the sensors went bad) predictions became irrelevant. If we had not known that the 53^{rd} sensor had failed we would have drawn a wrong conclusion that the sensor being monitored had failed.

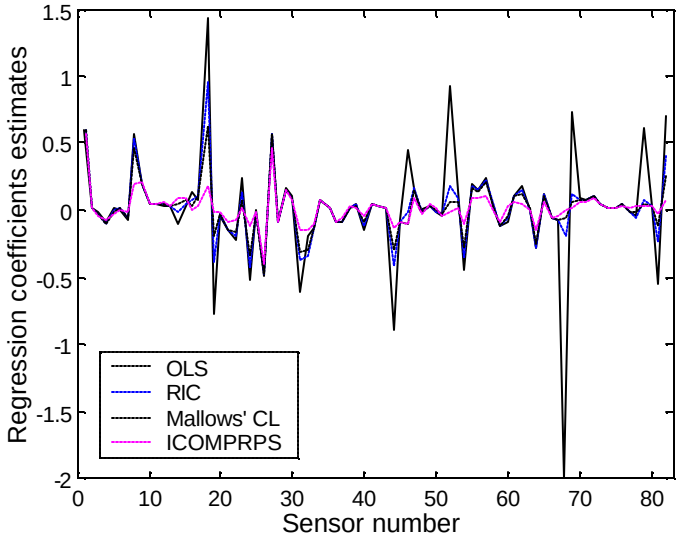

Figure 4.11. Solutions without (the solid line) and with (the dotted lines) regularization.

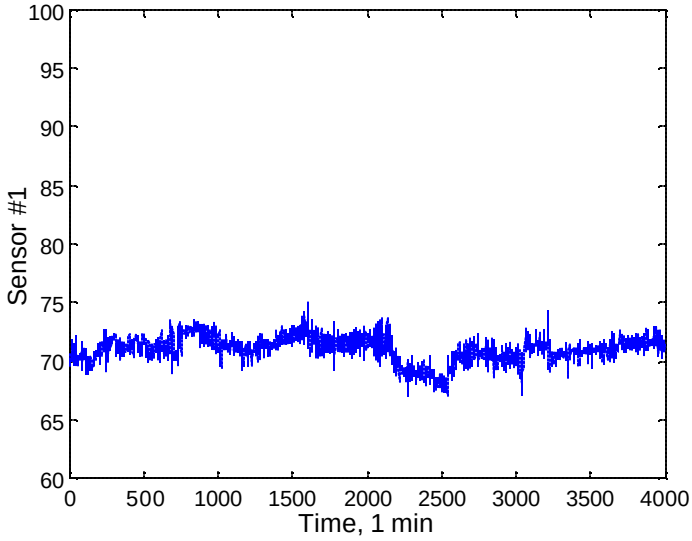

Figure 4.12. Sensor #1 to be predicted.

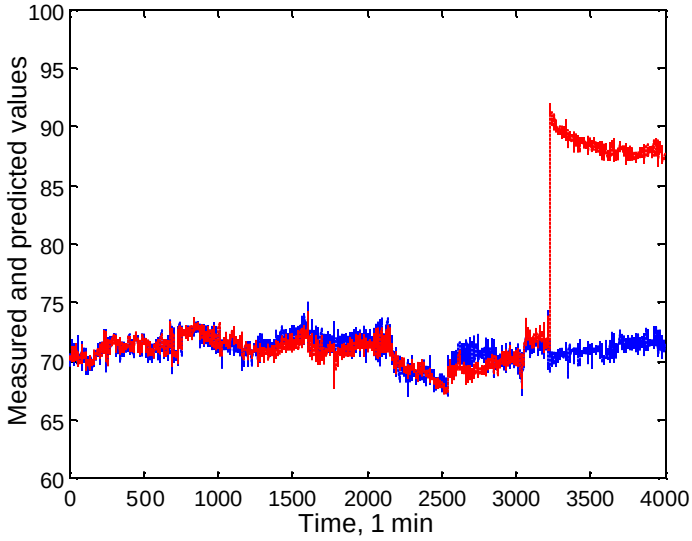

Figure 4.13. Prediction of sensor #1 using the OLS solution.

This was obviously not the case and shows the danger of using unstable solutions and ignoring the ill-posed nature of this problem.

Using a regularization parameter chosen by CL and RIC, predictions shown in Figure 4.14 and Figure 4.15 are obtained. In Figure 4.14, predictions for CL are shown. As with the OLS solution, the regularized solution corresponding to the CL-chosen parameter is not stable enough and as a result, the predictions are invalid once the 53rd sensor fails. This is another example in which CL produces a solution that is too optimistic or underregularized.

In Figure 4.15, predictions for RIC are shown. Despite the fact that RIC chose a larger value of the regularization parameter than CL, that value is still not enough to prevent failure of the predictive modeling when one of the sensors fails. The slight difference between CL and RIC indicate that the model might be misspecified. More likely, some relevant variables might be missing. Since in the case of sensor validation, the very notion of the true model does not exist, the difference simply indicates that there is probably not enough information in the sensors used as predictors to predict the sensor being monitored.

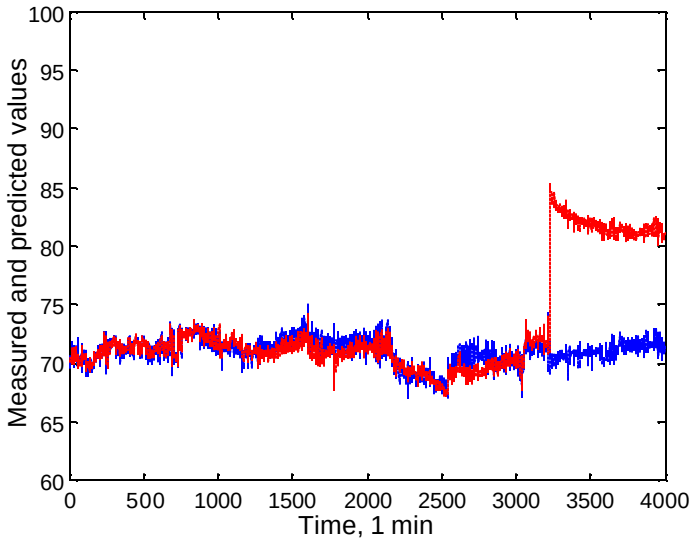

Figure 4.14. Prediction of sensor #1 using the parameter chosen by CL.

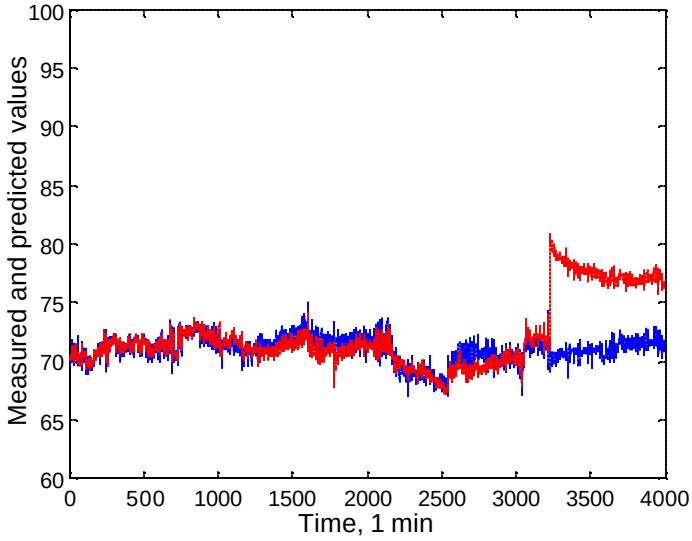

Figure 4.15. Prediction of sensor #1 using the parameter chosen by RIC.

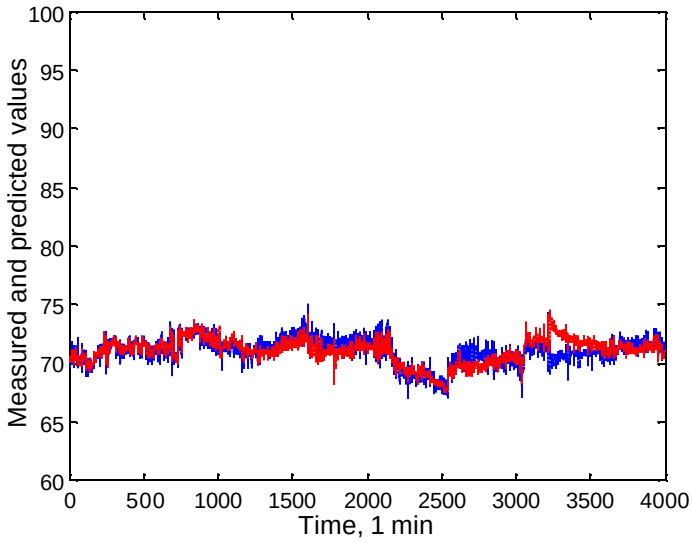

Figure 4.16. Prediction of sensor #1 using the parameter chosen by ICOMPRPS-CM (for correctly specified models).

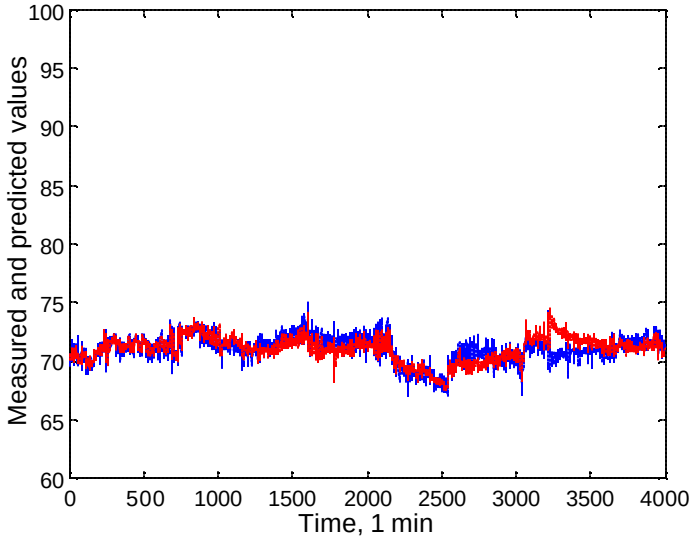

Figure 4.17. Prediction of sensor #1 using the parameter chosen ICOMPRPS.

Predictions made using the regularized solution corresponding to the ICOMPRPS-chosen parameter value are shown in Figure 4.16. The predictions in this case are much more stable and are not destroyed by the failed sensor. The regularization parameter chosen by ICOMPRPS is large enough to produce a regularized solution that is stable to at least one failing sensor. This is possible due to the extra penalization of the estimation inaccuracy. In many engineering applications, as in this example, it is beneficial to be more conservative because otherwise the solution is useless and cannot be used for building a reliable sensor validation system.

4.3 Statistical Learning from Data

A simple example of fitting a number of noisy observations demonstrates how difficult the solution of simple inverse problems can be. The true relationship between X and Y is

$$Y = X + 0.3\sin(2pX). \tag{4.3}$$

In Figure 4.18, data points generated from the true relationship (4.3) and corrupted by noise are shown. The goal is to use these noisy observations to find (learn) the true relationship or best approximation to it.

A Radial Basis Function (abbreviated RBF) neural network is used to fit the data. The RBF network computes its output according to the relationship

$$f_{RBF}(X) = \sum_{i=1}^{n} w_i a_i(X), \tag{4.4}$$

where the transfer function of the hidden units is given by

$$a_i(X) = \frac{1}{r}\sqrt{r^2 + (X - c)^2}\ . \tag{4.5}$$

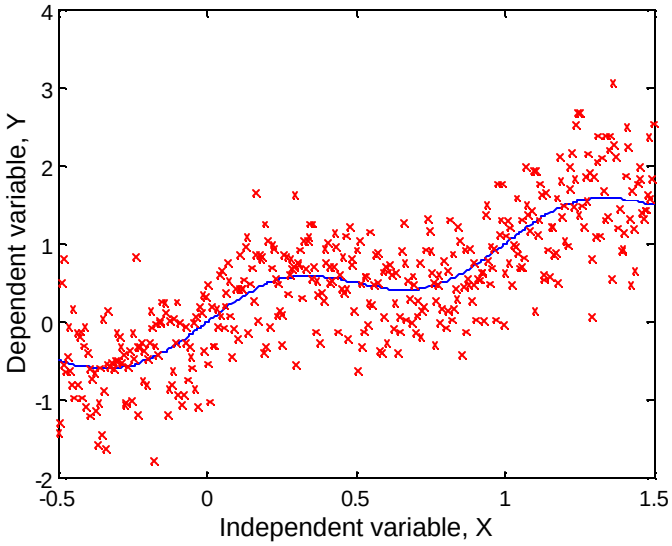

Figure 4.18. The true relationship (the solid line) and observed noisy data (the crosses).

When the network has n hidden nodes with the centers c_i equal to the input data points X_i, an optimal set of the weights w_i can be found by minimizing the sum of square residuals given by

$$SSR_{RBF} = \sum_{i=1}^{n} \left(Y_i - f_{RBF}\left(X_i\right)\right)^2 . \tag{4.6}$$

The weights that minimize (4.6) are called the OLS solution and are given by

$$w_{OLS} = A^{-1}Y , \tag{4.7}$$

where A is a square matrix composed of $a_i\left(X_j\right),\ i, j = 1...n$. The relationship produced by the RBF network with the OLS weights (4.7) is called the OLS fit and is shown in Figure 4.19. Since the number of weights equals the number of the data points, the OLS fit achieves a zero SSR. The OLS fit is very oscillatory and passes through each data point, i.e. the RBF network learned the noise component in the data.

It is obvious that the OLS fit is of no practical use because it is very different from the true relationship. Predictions based on the OLS fit are meaningless.

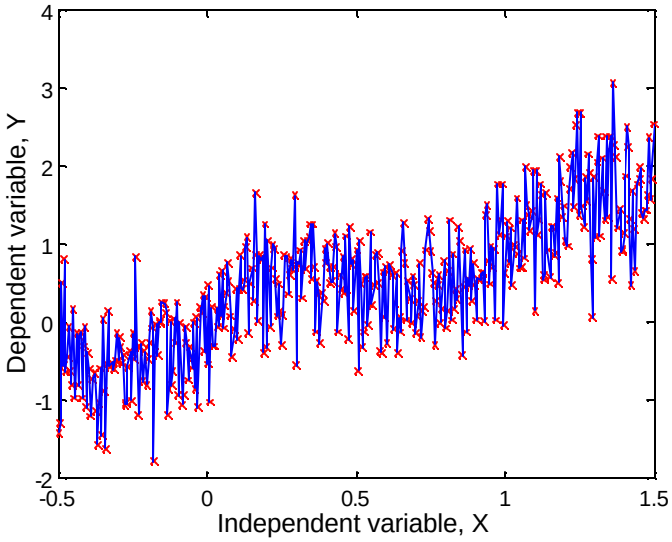

Figure 4.19. The OLS fit to the data.

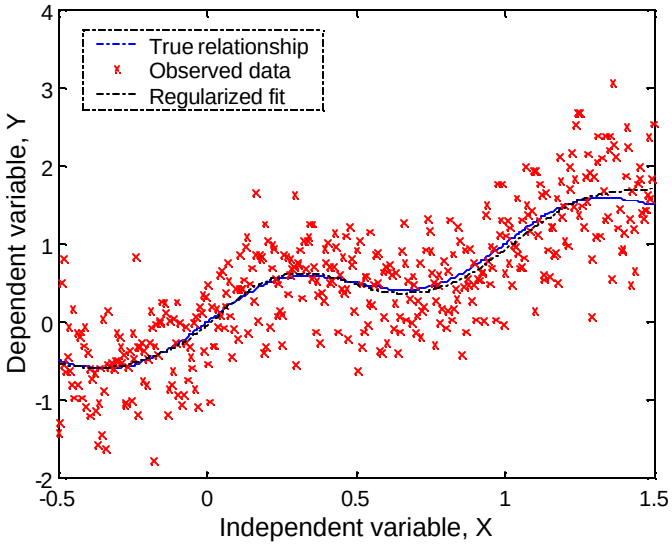

Figure 4.20. The true relationship (the solid line) and the regularized fit (the dash-dot line) to the data.

The reason for such poor performance is ill-conditioning of matrix A. The method of regularization in the form of ridge regression can be used to solve this problem. The regularized solution is obtained as

$$w_l = (A + lI_n)^{-1} Y. \tag{4.8}$$

The regularized fit to the data using the regularization parameter l that minimizes the GCV function (2.3) is shown in Figure 4.20. The regularized fit is close to the true relationship. It is useful and can be successfully used for predictions.

4.4 Numerical Solution of an Integral Equation

In this example we apply the RPSM's to select the regularization parameter value for solution of the Fredholm integral equation of the first kind in the discretized form

$$\int_a^b K(s,t) f(t) dt \approx I_n(s) = \sum_{i=1}^{n} w_i K(s,t_i) f(t_i). \tag{4.9}$$

In particular, the one-dimensional image restoration model studied by Shaw (1972) is considered from Hansen's (1994) Regularization Tools Matlab toolbox. This problem is severely ill-conditioned (the condition number is 10^{19} for n = 64) and is known to have a regularized solution for the penalty operator $\Omega = I_m$. The true solution and regularized solutions corresponding to regularization parameters chosen by different methods using the exact noise level as an noise level estimate are shown in Figure 4.21.

The regularized solutions corresponding to the different methods are almost identical, which is the result of selecting almost the same regularization parameter value by all the methods. For the true noise level, the mean square error between the true and regularized solutions is summarized in Table 5.

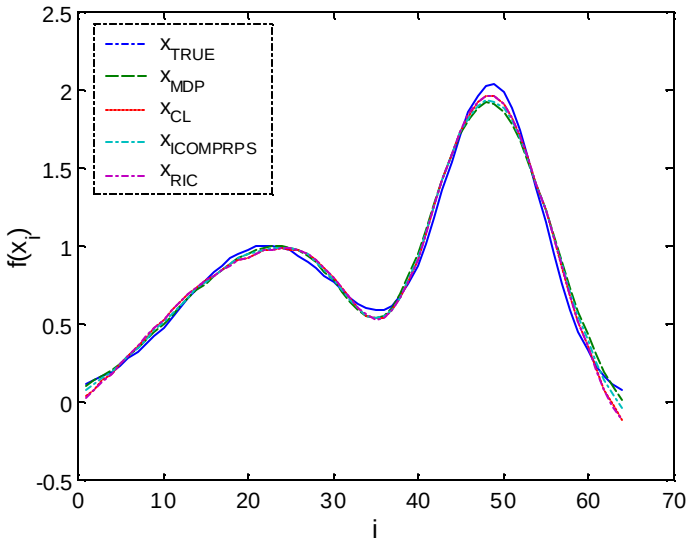

Figure 4.21. Solutions for the true noise level.

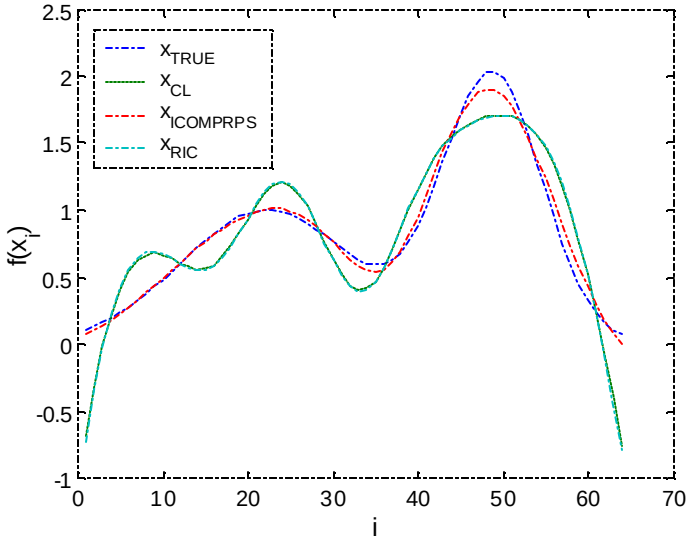

Figure 4.22. Solutions for the underestimated noise level.

Table 5. Mean square error for the true noise level.

RPSM	Mean square error
MDP	0.0033
CL	0.0028
ICOMPRPS	0.0026
RIC	0.0029

Table 6. Mean square error for the underestimated noise level.

RPSM	Mean square error
MDP	1.127088e+024
CL	0.0703
ICOMPRPS	0.0036
RIC	0.0771

The regularized solutions corresponding to regularization parameters chosen by the methods using the noise level artificially underestimated by 50% are shown in Figure 4.22. CL and RIC choose small regularization parameter values that produce undersmoothed solutions. MDP, with noise level underestimation, produces an unreasonable solution. The ICOMPRPS-based choice is more robust to an underestimated noise level and produces a good solution. For the underestimated true noise level, the mean square error between the true and regularized solutions is summarized in Table 6. Underestimation of the noise level is not unusual in engineering applications, where it is commonly estimated with a fairly large uncertainty.

If the regularization parameter selection is repeated with different noise realizations in the response, the sampling density of the chosen regularization parameter value can be estimated for different methods and the parameter variability can be compared. Figure 4.23 shows the sampling distribution of the chosen parameters, which were obtained using the true noise value.

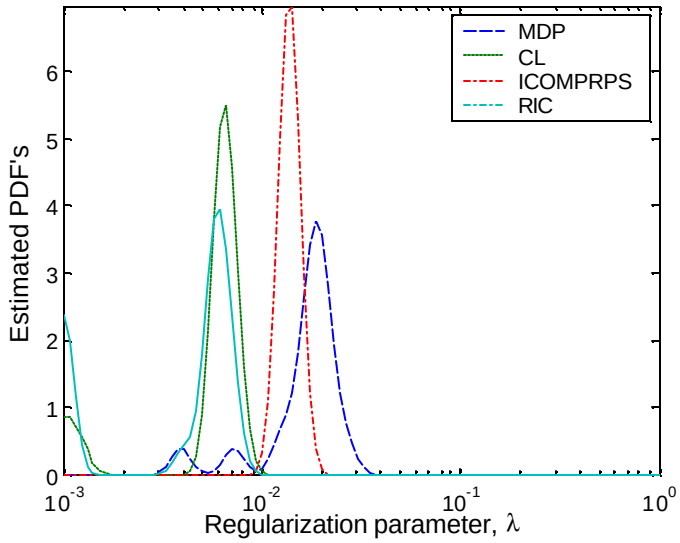

Figure 4.23. Sampling distributions of the chosen regularization parameter (NSR=0.003).

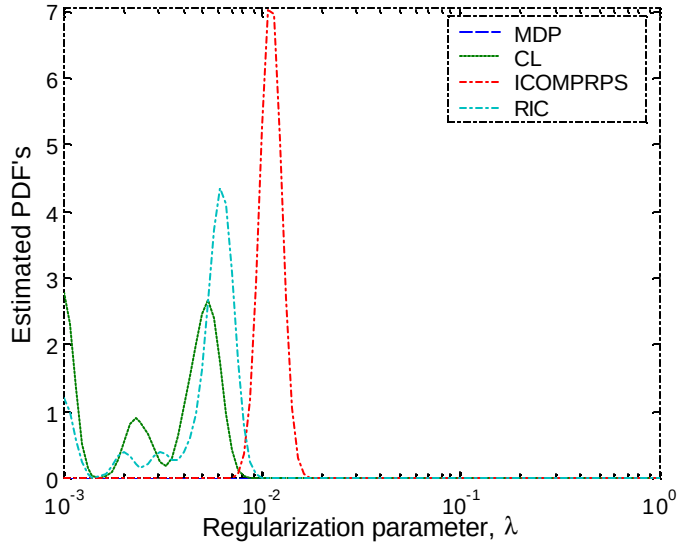

Figure 4.24. Sampling distributions of the chosen regularization parameter for the underestimated noise level by 50% (NSR=0.003).

Note that CL and RIC behave almost identically. This is reasonable because there is no functional misspecification in this example and accounting for possible misspecification has not resulted in any improvement. Despite the fact that most of the chosen values lie around 0.008, some very small values chosen for some noise realizations produce grossly underregularized solutions. This effect is more pronounced when the noise level is underestimated as shown in Figure 4.24.

The number of small values are much larger, and, as a result, the probability of getting grossly underregularized solutions increases drastically. The ICOMPRPS method does not fail; the chosen values are still concentrated around 0.01. However, when the noise level is underestimated even more, ICOMPRPS eventually fails. As discussed above, such behavior is due to the introduction of an additional term that results in the decreased variability of the regularization parameter.

Noise underestimation makes MDP and CL useless, whereas ICOMPRPS 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. ICOMPRPS suggests one possible refinement by accounting for interdependencies between the parameter estimates.

4.5 Image Reconstruction

An original image, registered by a measuring device, is convolved with the point-spread function of the measuring device. As a result, the registered image is a blurred version of the original. Notice that convolution is a smoothing process. When the inverse problem is solved, i.e. when the original image is reconstructed from its blurred version, deconvolution is used. Deconvolution is a roughening process and the solution (reconstructed original image) is unstable and must be regularized.

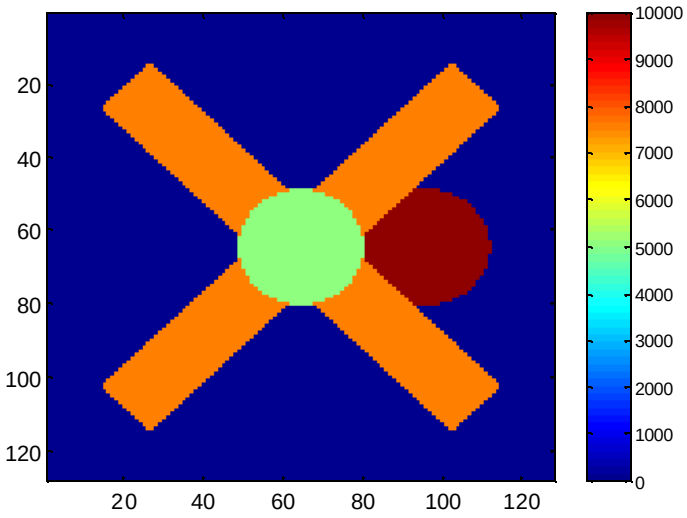

Figure 4.25. Original image.

The Matlab code for this example was provided by Dr. Curt Vogel of Montana State University in a personal communication.

The original image is shown in Figure 4.25. The process of registering the image is modeled by its convolution with a point-spread function. The blurred image registered by the device is shown in Figure 4.26.

If the ill-posedness of the problem is ignored, and the image is reconstructed by using standard deconvolution, the reconstructed image, shown in Figure 4.27, has nothing in common with the original image and has no practical value.

If regularization is applied, the reconstructed images are similar to the original one. The reconstructed image for a small regularization parameter is shown in Figure 4.28; for the ICOMPRPS regularization parameter in Figure 4.29; for the optimal regularization parameter in Figure 4.30; and for a large regularization parameter in Figure 4.31. We see that the value of the regularization parameter plays an important role in the solution of the inverse problem; thus it is a necessity to be able to choose it properly.

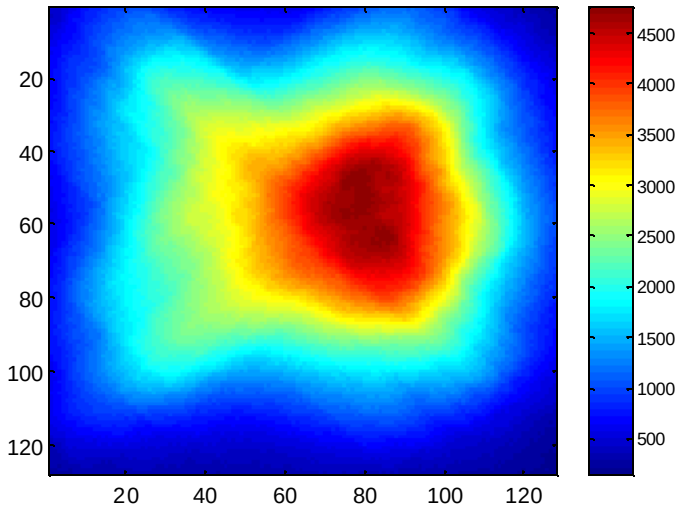

Figure 4.26. Observed blurred image.

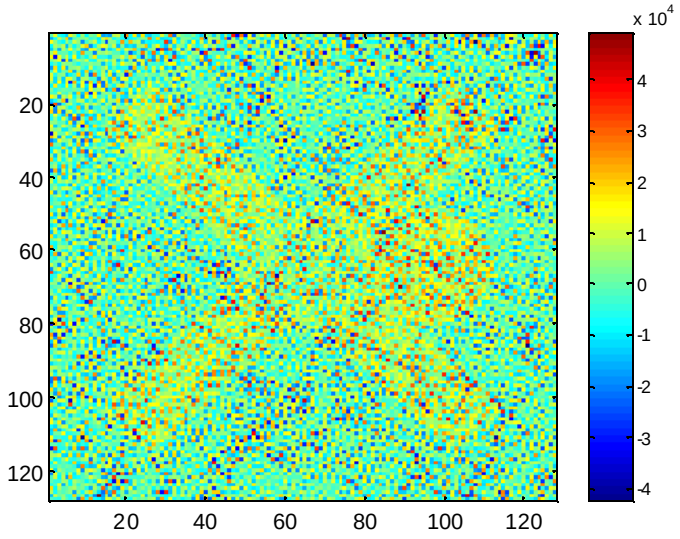

Figure 4.27. Reconstructed image without using any regularization (l=1e-20).

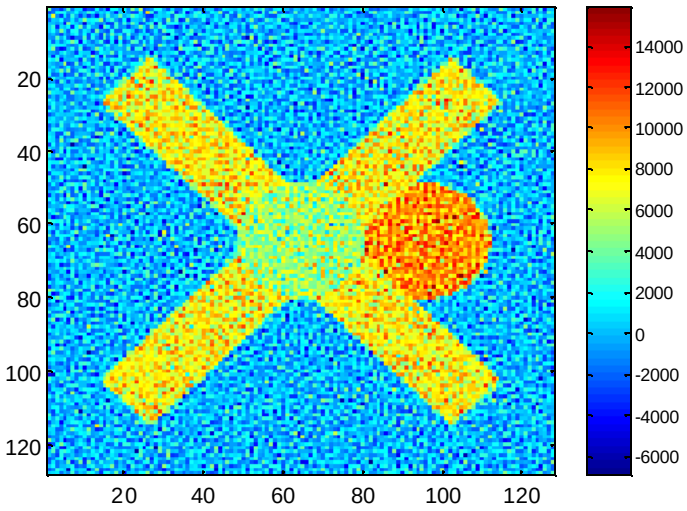

Figure 4.28. Reconstructed image with a too small regularization parameter value (l=1.3e-6).

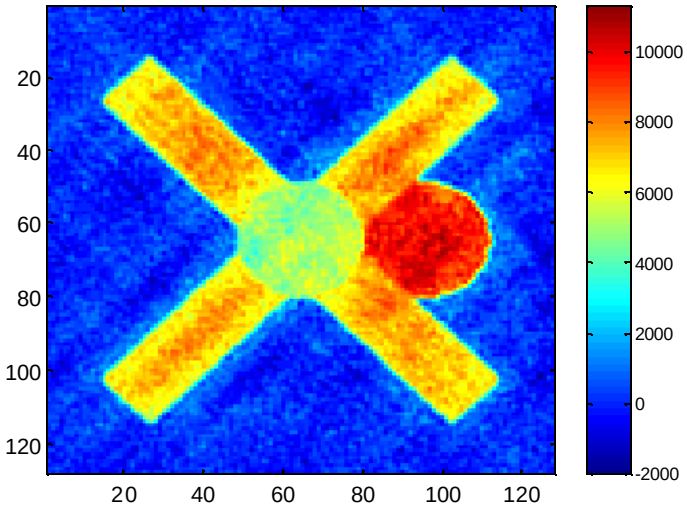

Figure 4.29. Reconstructed image with the regularization parameter value chosen by ICOMPRPS (l=0.000136).

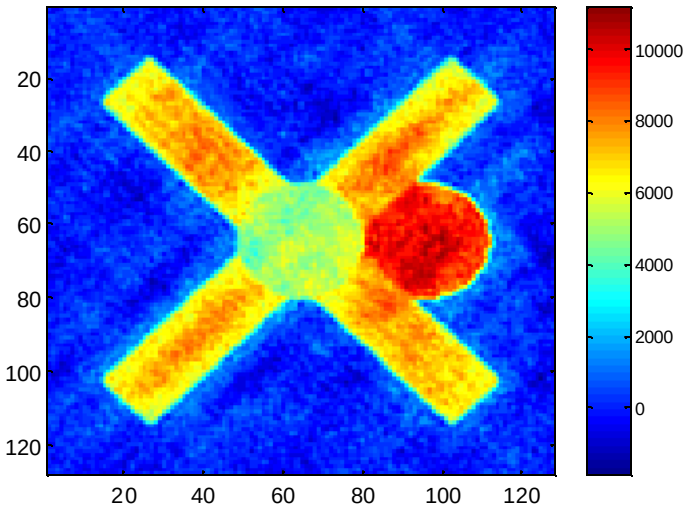

Figure 4.30. Reconstructed image with the optimal regularization parameter value (l=0.000179).

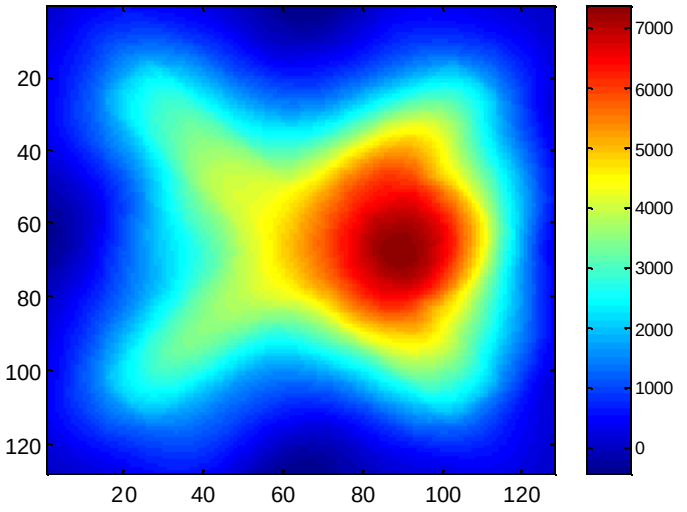

Figure 4.31. Reconstructed image with a too large regularization parameter value (l=0.032).

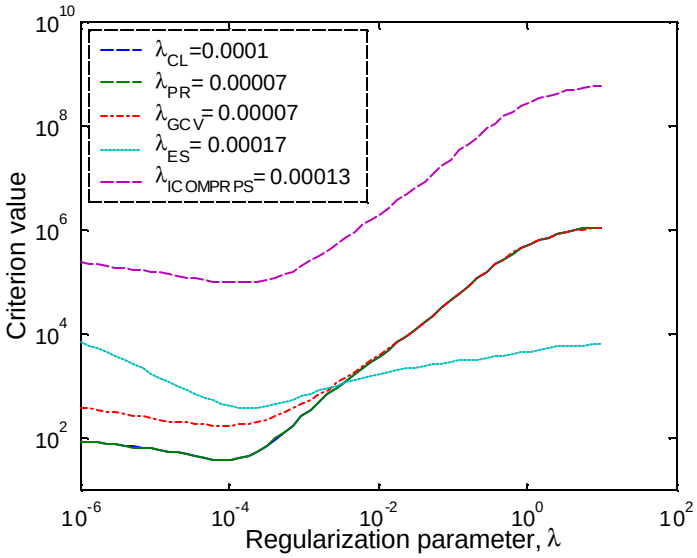

Figure 4.32. Regularization parameter selection methods.

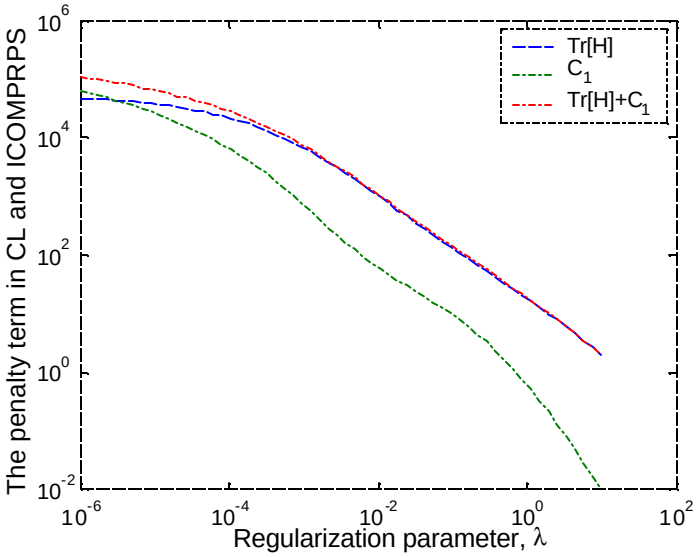

Figure 4.33. Using C1 to refine the estimation of the bias in estimating the prediction error.

In Figure 4.32 the results of several RPSM's are plotted. ICOMPRPS, derived in the information complexity framework, outperforms CL and GCV. ICOMPRPS chooses the value 0.0004, which is closer to the optimal value of 0.0007 than those chosen by GCV and CL. Figure 4.33 demonstrates how the complexity measure (C1) refines the bias estimation, imposing a more severe penalization of the estimation inaccuracy by means of a penalization of interdependencies among parameter estimates.

4.6 Specification of Prior Distribution in Bayesian Inference

In Bayesian analysis, we start with assigning a prior distribution $p(q)$ to the parameters q of a model. Given a data set X, we build the likelihood $L(X|q)$ and calculate the posterior distribution $p(q|X)$ of the parameters. The posterior distribution of the parameters in the proportional form is defined as

$$p(q|X) \propto L(X|q)p(q). \tag{4.10}$$

Using the posterior distribution we can build the predictive distribution of the new data set Y given the old data X which was used for parameter estimation. The predictive distribution is defined as

$$p(Y|X) \propto \int L(Y|q)p(q|X)dq, \tag{4.11}$$

which is used to make an inference about future observations of Y.

When there is no prior information, a common method is to use a non-informative prior to performing the analysis. This produces a tool to analyze the uncertainty in the predictions. However, it does not solve the problem. For example, performing a regression analysis of an ill-posed problem using a noninformative prior produces OLS coefficients which are useless. To be able to obtain a useful solution, an informative prior must be used which can help penalize undesirable properties of the solution.

If some information about the output is available, such as a range of its values, its sign, or a rough approximation to the predictive density, this information can be used to solve the above integral equation with the approximately known left hand side $\tilde{p}_Y$ for the prior

$$\tilde{p}_Y \propto \int L(Y|q)p_q\,dq \tag{4.12}$$

This is a typical ill-posed inverse problem, which is also severely ill-conditioned in a discretized form.

This approach can be demonstrated using a coin example. The goal is to infer the value of the parameter that describes the probability of getting the head when tossing a coin. We observe a sequence of heads and tails. In order to build a predictive distribution of the fraction of heads in the sequence, it is necessary to assign a prior distribution to the parameter, which in this particular case is known to be binomial. Since the parameter is unobservable (because we don't know whether the coin is fair), we try to use some information about the observed output (the fraction of heads in the sequence) in the form of a rough predictive density. This is used in the left hand side of the integral equation to solve for the prior. The OLS solution (prior density of the parameter) is shown in Figure 4.34. The OLS solution is very oscillatory. It is not a proper probability density because of the negative values and is useless.

However, the regularized solutions shown in Figure 4.35 are very similar to the true probability density of the parameter (the solid line). The regularized solutions are not proper density functions because of the negative values in their tails. This can be corrected by using a different smoothing operator to produce only positive solutions. The regularized prior distribution of the parameter is obtained using only prior information about the observable output and the universal smoothness prior in regularization. This prior can be used to obtain the posterior distribution of the parameter and then, using the observed sequence, the predictive distribution of the output.

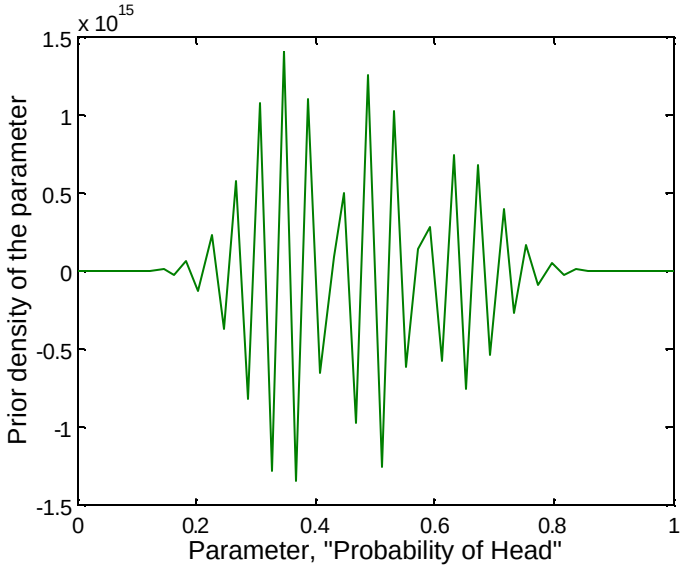

Figure 4.34. OLS solution.

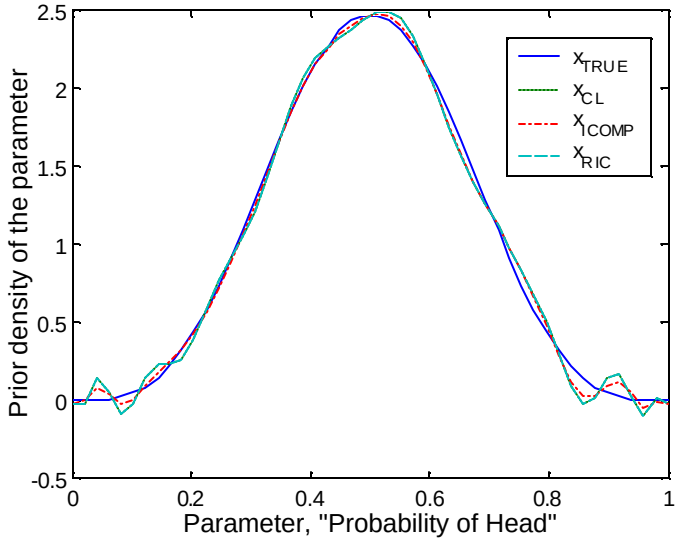

Figure 4.35. Regularized solution.

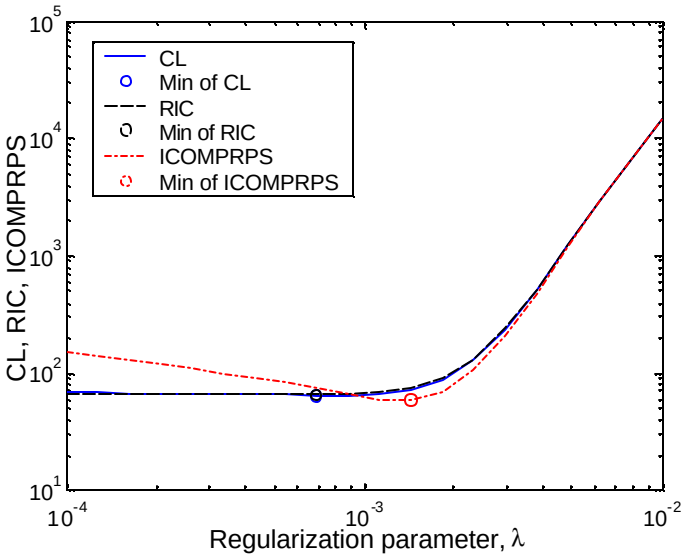

Figure 4.36. Regularization parameter selection.

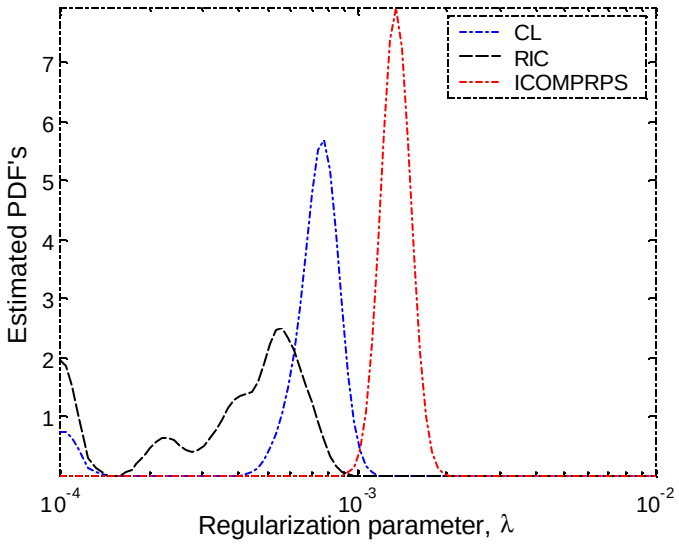

Figure 4.37. Variability of the regularization parameter value chosen by different RPSM's.

We now compare several RPSM's. A typical behavior of the RPSM's is shown in Figure 4.36. Since there is no model misspecification in this example, CL and RIC methods behave almost identically. However, both CL and RIC produce slightly underregularized solutions.

The estimated sample densities of the regularization parameter chosen by different RPSM's is shown in Figure 4.37. The ICOMPRPS method produces a much more stable parameter estimate than CL and RIC. Due to the inadequate estimation inaccuracy penalization, CL and RIC very often choose smaller values of the regularization parameter. The corresponding solutions are undersmoothed and useless. ICOMPRPS always chooses parameter values that correspond to useful solutions. ICOMPRPS drastically reduces the risk of obtaining grossly underregularized solutions. The described approach for the prior specification may have a very wide range of applications varying from regression analysis to Bayesian regularization of neural networks.

CHAPTER 5
CONCLUSION AND SUGGESTIONS FOR FUTURE WORK

Many engineering problems are ill-posed. The failure to realize this fact can lead to unsuccessful attempts to build a data-driven method which is reliable and stable. An ill-posed problem is not solvable by conventional methods because the assumptions under which the methods were derived are violated. For example, it is impossible to build a stable sensor validation system using the OLS method. The OLS solution in the case of highly collinear predictors is extremely unstable and hypersensitive to small perturbations and particular realizations of the noise component. This is exactly the opposite property a data-driven method should possess to be of a practical value.

Special techniques, such as regularization methods, must be employed to obtain stable solutions. The resulting regularized solution may be considered to be a solution of a well-posed problem that approximates the given ill-posed problem. Using a method of regularization alone does not automatically guarantee a good solution. Even if the penalization operator is properly chosen according to the physical interpretation of the solution (if such is possible), a proper regularization parameter must still be determined. The proper choice of the parameter is a difficult problem because it requires prior information about the sought solution and knowledge about the noise level in the response. In almost all practical situations, such information is unavailable.

Several different RPSM's have been proposed in the literature. Yet none is misspecification-resistant. They assume that the specified model is correct and do not guarantee reliable solutions when the model is misspecified. For many applications, it is extremely difficult to find a basis for arguing that the model is correct. Most probably the

model will be misspecified. Misspecification can be functional, in which the functional relationship between the predictors and the response is not correct or some relevant variables are missing. Misspecification can also be distributional, in which the distributional model of the data is not correct. For example, instead of a normal distribution, the noise can have a skewed distribution and/or a distribution with heavy tails. There also may be outliers present in the data. Each of these is usual and is present in real data sets.

In the dissertation we propose a misspecification-resistant RPSM that not only behaves consistently under possible model misspecification, but also has a significantly smaller risk of producing grossly underregularized solutions. Common rules such as CL or RIC, because of their statistical nature, can perform well on one data set and fail miserably on another. Stable, reliable methods are needed in autonomous applications. The ICOMPRPS method, due to the extra penalization of estimation inaccuracy, significantly reduces the risk of obtaining grossly underregularized solutions. This method can be reliably implemented in autonomous diagnostic and monitoring systems.

The ICOMPRPS method combines two powerful theoretical approaches. One is the information approach to regularization parameter selection and the other is the information complexity approach. The information approach enables one to build an estimator of the mean expected log likelihood, which is consistent under model misspecification. This approach involves asymptotics, i.e. the results are guaranteed when the number of data points goes to infinity. For actual problems, this is impossible to fulfill. Small sample properties of the information methods require investigation. ICOMPRPS, by means of also using the information complexity approach, is able to compensate for the inadequate penalization of estimation inaccuracy in the information methods for a limited number of observations. This extra penalization is more severe for smaller data sets and for more ill-conditioned problems. As a result, the proposed RPSM

can handle possible model misspecification and significantly reduces the risk of obtaining grossly underregularized solutions. Both of these properties are of exceptional value for engineering applications.

The superior performance of the ICOMPRPS method was demonstrated by several practical applications including a sensor validation system, an inferential drift prediction system and other examples.

The topic of the dissertation and presented material have been peer-reviewed and published in several conference and journal papers. These papers are attached in Appendix A.3. The new method presented is an important contribution to the field of ill-posed inverse problems in engineering.

5.1 Future Work and Further Improvement

Future work may include an extension of the information complexity approach to other modeling paradigms such as support vector machines, which also use regularization and require regularization parameter selection.

Further improvements may include the generalization of the ICOMPRPS method to the correlated (colored) noise case. Although this case is very important, many existing RPSM's fail when the noise is colored.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, In 2nd International Symposium on Information Theory, Ed. B.N. Petrov and F. Csaki, pp. 267-281. Budapest: Akademiai Kiado.

Bakushinskii, A.B. (1984). Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion, USSR Comp. Math. Math. Phys. 24, 4, pp.: 181-182.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions, Psychometrika, 52, (3), pp. 345-370.

Bozdogan, H. (1988). ICOMP: A new model selection criterion. In Hans H. Bock (Ed.), Classification and related methods of data analysis, pp 599-608. Amsterdam: Elsevier Science Publishers B.V. (North-Holland).

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. Communications in statistics theory and methods, 19 (1), pp. 221-278.

Bozdogan, H. (1994). Mixture model cluster analysis using model selection criteria and a new informational measure of complexity. In Bozdogan H. (Ed.) Multivariate statistical modeling, Vol. 2, Kluwer Academic Publishers, Dordrecht, the Netherlands, pp. 69-113.

Bozdogan, H. (1996). A new informational complexity criterion for model selection: The general theory and its applications. Information Theoretic Models & Inference (INFORMS), Washington D.C., May 5-8.

Bozdogan, H. (1996). Informational complexity criteria for regression models, Information Theory and Statistics Section on Bayesian Stat. Science. ASA Annual Meeting, Chicago, IL, Aug. 4-8.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. Journal of Mathematical Psychology, 44, pp. 62-91.

Bozdogan, H. and Magnus, J. (2001). Misspecification resistant model selection using information complexity. Invited paperpresented at the IMPS-2001 Conference at Osaka University, Osaka, Japan, July 15-19.

Cox, D.D., O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators, Annals of Statistics, 18, No. 4, pp. 1676-1695.

Engl, H.W., Hanke, M. and Neubauer, A., (2000) Regularization of inverse problems. Kluwer Academic Publishers.

Eubank, R.L. (1988). Spline smoothing and nonparametric regression. New York : M. Dekker.

Fisher, R.A. (1921). On the mathematical foundations of theoretical statistics. Phil. Trans., (A), 222, 309.

Good, I.J. and Gaskins, R.A. (1971), Nonparametric roughness penalties for probability densities. Biometrika, 58, 255-277.

Green, P.J. (1987), Penalized likelihood for general semi-parametric regression models. Int. Statist. Rev. 55, 255-259.

Gribok, A.V., I. Attieh, J.W. Hines and R.E. Uhrig (2001), Regularization of Feedwater Flow Rate Evaluation for Venturi Meter Fouling Problems in Nuclear Power Plants, Nuclear Technology vol.134 pp.3-14.

Gribok, A.V., Urmanov, A.M., Hines, W.J., Uhrig, R.E. (2002). Backward Specification of Prior Distribution in Bayesian Inference as an Inverse Problem. Submitted to Inverse Problems in Engineering.

Hadamard, J. (1902). Sur les problemes aux derivees partielies et leur signification physique, Bull. Univ. Princeton, Vol. 13, pp.: 49-52.

Hanke, M. and Hansen, P.C. (1993) Regularization methods for large scale problems. Surveys math. Indust. 3, pp.: 253-315.

Hansen, P.C. (1998). Rank-deficient and discrete ill-posed problems, SIAM monographs on mathematical modeling and computation, Philadelphia.

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

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, Vol. 12, No. 1, pp.: 55-82.

Huber, P.J. (1981), Robust statistics. John Wiley & Sons, Inc.

Knight, K. (1998). Asymptotics for L1 regression estimators under general conditions. The Annals of Statistics, V. 26, No. 2.

Konishi, S., Kitagawa, G. (1996). Generalized information criteria in model selection, Biometrika, 83, No. 4, pp. 875-890.

Kullback, S., Leibler, R.A. (1951). On information and sufficiency, Ann. Math. Statist., 22, pp. 79-86.

Leamer, E. E. (1978). Specification Searches. Wiley, New York.

Leonard, T. (1978), Density estimation, stochastic processes and prior information. J. Roy. Statist. Soc. Ser. B, 40, 113-146.

Leonov, A.S., Yagola, A.G. (1997). The L-curve method always introduces a nonremovable systematic error, Moscow University Physics Bulletin, 52, No. 6, pp. 20-23.

Mallows, C.L. (1973). Some comments on CP, Technometrics, 15, No. 4, pp. 661-675.

Morozov, V.A. (1984). Methods for solving incorrectly posed problems, Springer-Verlag New York Inc.

Murata, N., Yoshizawa, S., Amari, S. (1994). Network information criterion – determining the number of hidden units for an artificial neural network model. IEEE Transactions on Neural Networks, Vol. 5 No. 6, pp.: 865-872.

Raus, T. (1984) Residue principle for ill-posed problems. Acta et comment. Univers. Tartuensis 672 (in Russian).

Sakamoto, Y. (1986), Akaike Information Criterion Statistics. KTK Scientific publishers.

Schwaz, R. (1978). Estimating the dimension of a model, Ann. Statist. 6, 461-464.

Shaw, C.B. Jr. (1972). Improvements of the resolution of an instrument by numerical solution of an integral equation, J. Math. Anal. Appl. 37, 83-112.

Shibata, R. (1989), Statistical aspects of model selection. In From data to model, Ed. J.C. Willems, pp. 215-240. New York: Springer-Verlag.

Silverman, B.W. (1985), Penalized maximum likelihood estimation. Encyclopedia of Statistical Sciences, 6, 664-667.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. Mathematical Sciences, No. 153, pp.: 12-18 (in Japanese)

Tikhonov, A.N. (1963). Solution of incorrectly formulated problems and regularization method. Soviet Math. Dokl. 4, 1035-1038, USSR Academy of Science.

Upadhyaya, B.R., Kavaklioglu, K., (1994). Monitoring feedwater flow rate and component thermal performance of pressurized water reactors by means of artificial neural networks. Nuclear Technology 107.

Urmanov, A.M., Gribok, A.V., Bozdogan, H., Hines, J.W., and Uhrig, R.E (2002), Information Complexity-Based Regularization Parameter Selection for Solution of Ill-Conditioned Inverse Problems. Inverse Problems 18, L1-L9.

van Emden, M.H. (1971). An analysis of complexity. Mathematical Centre Tracts, 35, Amsterdam.

Vogel, C.R. (1996). Non-convergence of the L-curve regularization parameter selection method, Inverse Problems, 12, pp. 535-547.

Wahba, G. (1990). Spline models for observational data. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

Wahba, G. (1993). Behavior near zero of the distribution of GCV smoothing parameter estimates, Technical Report No. 910, Dept. of Statistics, University of Wisconsin.

White, H. (1980). Using least squares to approximate unknown regression functions. International Economic Review, Vol. 21, No. 1, pp.: 149-170.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models. JASA, Vol. 76, No. 374, pp.: 419-433.

White, H. (1994). Estimation, inference and specification analysis. Cambridge University press.

APPENDIX

A.1
The Trace Result

Consider a random m-vector b normally distributed as $\sqrt{n}b \sim N(0,\Sigma)$. The expected value of $\left(b^T Ab\right)$ is given by

$$E\left(b^T Ab\right) = \frac{1}{n} trace(A\Sigma).$$

Indeed, the expected value can be calculated using the properties of the expectation operator as

$$\begin{aligned} E\left(b^T Ab\right) &= E\left(\begin{pmatrix} b_1 & \cdots & b_m \end{pmatrix} \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mm} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \right) \\ &= E\left(\sum_{i=1}^{m} b_1 b_i a_{i1} + \sum_{i=1}^{m} b_2 b_i a_{i2} + \ldots + \sum_{i=1}^{m} b_m b_i a_{im} \right) \\ &= \sum_{k=1}^{m}\sum_{i=1}^{m} a_{ik} E\left(b_k b_i\right) = \frac{1}{n}\sum_{k=1}^{m}\sum_{i=1}^{m} a_{ik} s_{ki} = \frac{1}{n} trace(A\Sigma) \end{aligned}$$

This proofs the result.

A.2
Plant Variables for Example 1

Table 7. 24 plant variables used as predictors to evaluate feedwater flow rate.

Var. Num.	Description	Range	Units
1	FWP speed	0-7500	rpm
2	'A' OTSG efic high level	0-100	percent
3	Feedwater pump A speed	0-7500	rpm
4	Linear power CH NI-6	0-125	percent
5	Heater 3A inlet cond temp	40-300	degf
6	Heater 3B outlet cond temp.	40-350	degf
7	Dearator inlet cond temp	40-350	degf
8	Heater 6A inlet FW temp	40-500	degf
9	FWP A discharge temp	40-500	degf
10	FWP A suction temp	40-500	degf
11	Heater 5B outlet FW temp	40-500	degf
12	Steam gen B inlet FW temp	40-600	degf
13	Heater 6B outlet FW temp	40-600	degf
14	Steam gen A level (op)	0-100	percent
15	Steam gen A level (full)	40-640	inches
16	Steam gen A level (start up)	0-250	inches
17	Steam gen B inlet FW temp	0-500	degf
18	Steam gen B level (start up)	0-250	inches
19	Steam gen A inlet FW temp	40-600	degf
20	Steam gen B inlet FW temp	40-600	degf
21	Reheater A cold reheat press.	0-200	psig
22	Reheater D cold reheat press.	0-200	psig
23	Reheater C cold reheat press.	0-200	psig
24	No. 2A extr LP turb pressure	0-20	psia

A.3

List of Papers and Book Chapters Partially Based on the Material of the Dissertation

1. Urmanov, A.M., Bozdogan, H., Gribok, A.V., Hines, J.W., Uhrig, R.E. (2002) ICOMP-Based Regularization Parameter Selection for Solution of Ill-Conditioned Inverse Problems. *Inverse Problems*, Vol. 18, No. 2.
2. Gribok A.V., Hines J.W., Urmanov A.M., and Uhrig R.E., (2002). Regularization of Ill-Posed Surveillance and Diagnostic Measurements. Chapter in *Power Plant Surveillance and Diagnostics - Modern Approaches and Advanced Applications*, Editors: Da Ruan and Paolo F. Fantoni, Springer to be printed in 2002.
3. Gribok A.V., Hines J.W., Urmanov A.M., and Uhrig R.E. (2002). Heuristic, Systematic, and Informational Regularization for Process Monitoring. To Appear in a special issue of *Intelligent Systems for Plant Surveillance and Diagnostics.*
4. Urmanov, A.M., Gribok, A.V., Hines, J.W., and Uhrig, R.E. (2002). An Information Approach to Regularization Parameter Selection Under Model Misspecification. To Appear in *Inverse Problems.*
5. Urmanov, A.M., Gribok, A.V., Hines, J.W., Uhrig, R.E. (2001). Information-based approaches to model (variable) selection for the venturi meter drift detection in a nuclear power plant. Submitted to *Nuclear Technology*.
6. Gribok, A.V., Urmanov, A.M., Hines, J.W., Uhrig, R.E. (2002). Backward Specification of Prior Distribution in Bayesian Inference as an Inverse Problem. To appear in *Inverse Problems in Engineering*.
7. Hines, J.W., Urmanov, A.M., Gribok, A.V. and Buckner, M.A. (2002). Selection of Multiple Regularization Parameters in Local Ridge Regression Using Evolutionary Algorithms and Prediction Risk Optimization. To appear in *Inverse Problems in Engineering*.

VITA

Aleksey Urmanov, originally from Obninsk, Russia, earned a Bachelor of Science degree in Applied Mathematics and a Master of Science degree in Systems Science from Moscow Institute of Physics and Engineering, Obninsk, Russia in May, 1994 and February, 1996 respectively. He joined the Institute of Physics and Power Engineering, Obninsk, Russia in March, 1996 and worked on a variety of engineering problems including real-time diagnostics and fault detection in nuclear power plants. He entered the University of Tennessee to work toward a Doctor of Philosophy degree in Nuclear Engineering in the spring of 1999. While at the university, he conducted research on anticipatory control and management of complex distributed systems, statistical learning from data, and on the solution of industrial ill-posed inverse problems. He received his Ph.D. in Nuclear Engineering in May, 2002.