SMOOTHING TECHNIQUES IN UNDERDETERMINED LINEAR MODELS

By

William Gary Sneddon

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
DALHOUSIE UNIVERSITY
HALIFAX, NOVA SCOTIA
OCTOBER 1997

0-612-36562-X

DALHOUSIE UNIVERSITY

FACULTY OF GRADUATE STUDIES

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "Smoothing Techniques in Underdetermined Linear Models"

by William Gary Sneddon

in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dated: October 3, 1997

External Examiner [redacted]

Research Supervisor

Examining Committee

DALHOUSIE UNIVERSITY

Date: **October 1997**

Author: **William Gary Sneddon**

Title: **Smoothing Techniques in Underdetermined Linear Models**

Department: **Mathematics Statistics and Computing Science**

Degree: **Ph.D.** Convocation: **May** Year: **1998**

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

[redacted]

Signature of Author

THE AUTHOR RESERVES OTHER PUBLICATION RIGHTS, AND NEITHER THE THESIS NOR EXTENSIVE EXTRACTS FROM IT MAY BE PRINTED OR OTHERWISE REPRODUCED WITHOUT THE AUTHOR'S WRITTEN PERMISSION.

THE AUTHOR ATTESTS THAT PERMISSION HAS BEEN OBTAINED FOR THE USE OF ANY COPYRIGHTED MATERIAL APPEARING IN THIS THESIS (OTHER THAN BRIEF EXCERPTS REQUIRING ONLY PROPER ACKNOWLEDGEMENT IN SCHOLARLY WRITING) AND THAT ALL SUCH USE IS CLEARLY ACKNOWLEDGED.

Contents

List of Tables

List of Figures

Acknowledgments

I have been working on this degree for a long time, so it seems only fitting that I have a lengthy list of people to acknowledge. Read on at your own peril.

I begin with expressing my heartfelt thanks to Dr. Chris Field. His tireless efforts and patience with my progress, and the countless hours he spent in making this thesis become a reality, will not be forgotten by me. I would also like to thank Dr. Francis Zwiers for serving as the external examiner for my thesis. His pointed questions and constructive comments led to many improvements in the quality of the material in this thesis and its presentation. I also thank Dr. Bruce Smith, who gave of his own time to meet with Chris and I in various sessions as this work progressed, and always offered useful comments and suggestions. Dr. Keith Thompson also played a significant role in my work, as he introduced me to the role that statistics can play in studies of the earth sciences, which became the motivation of this thesis. He also provided tremendous help and arranged for access to the data sets that are analysed in this thesis.

At this point these acknowledgements become like an Academy Award speech. I am going to mention a lot of names, but I just know some people are going to be left out. So I apologize in advance to the people who fall in the latter category. This will also tell me who actually read these acknowledgments.

In recognizing people I've met in my years here, I have to start with Bob. Bob was around for virtually all the time that this piece of work was being put together, and we'll both freely admit we spent way too much time around each other in the math building. Fortunately we also got to know each other very well outside of our office space, and it was definitely the highlight of my time at Dal to count Bob as one of my friends. It can't all be detailed here, but my time at Dalhousie would not have been the same without him around.

Of course, there are other important people. I can't forget Clyde and Doreen, who wish they also came from Cape Breton (joking, guys!). They have served me well as friends over the past few years, with Clyde having the patience to put up with my endless computer questions, while Doreen endured our less than assertive personalities.

I certainly can't forget about Kadre, whom I had the pleasure of meeting in the past year. A partner of mine in procrastination, she always took the time to listen, and tell lots of stories, which helped to keep what's left of my mind intact during some crucial stages of writing this book. Now I just have to do a PhD in anthropology to repay her.

There were others, too. People like Ted, Jonathan, Mike (what's that boy holding back there?), Jane, Larry and anyone else who showed up at the Friday Grad House functions. These entertaining afternoons showed that math and stats people don't sit around and play chess all the time. I would be in big trouble if I didn't mention Olga, who became a terrific friend to me in the past year. She brought a new perspective, a certain spark, to things around here, which I found very refreshing.

I had some great officemates during my time at Dal. Most of the time Hossein, Robert and I shared the same space, and it was a pleasure to get to know them, even if I never really knew what cosmology was all about. Then George appeared at one of the desks, a true whirlwind of energy. George taught me the secret of time management-don't sleep.

Finally, in all this talk about people at Dal who played a role in getting this thesis to where it is, I must not forget my three special friends in Bedford. They showed me there was more to life than school, and helped keep my mind out of the books for a small part of each week.

Is anyone still awake, yet feeling adventurous? Then turn the page!

Abstract

In many physical processes linear models can arise from the discretization of a continuous process in order to describe behavior over an entire field. An important consequence is the models may have more parameters than observations. To alleviate this problem, one can impose a smoothness constraint on the parameters that reflects some prior knowledge of the physical process in order to obtain sensible estimates.

A linear model is developed that has random explanatory and response variables, and a smoothness penalty is imposed based on the signal-to-noise ratio of the model. Results are presented assuming the value of the ratio is fixed, and when a procedure for estimating its value is used. The estimates perform well using a prediction based criterion in both situations. Robust estimation procedures for the model are also developed.

The methods are applied to modelling temperature and salinity data in the California Current, with the goal of using shallow water observations to predict deep ocean readings.

Chapter 1

Introduction

1.1 Introduction

The use of multiple linear regression is widespread in the study of many problems in the physical sciences, such as chemistry, oceanography and climatology. A problem that can arise in the modelling of many physical or dynamic processes is having explanatory variables that are not fixed, but represent measurements of the same quantity at various locations in either space or time. These could be temperatures at different ocean depths, air pressures collected over time or plasma absorption levels at various frequencies collected from an infrared (IR) spectrum. Since the variables are measured with some error, often over a fixed grid, it is natural to think of them as realizations of a continuous random process. Usual linear regression modelling deals with fixed explanatory variables. We are describing what are often referred to as errors-in-variables, or measurement error models. Fuller (1987) gives thorough coverage of the topic, while more recent developments are reviewed by Van Huffel and Vandewalle (1991) and Van Huffel (1997), while Carroll, Ruppert, and Stefanski (1995) describe extensions to nonlinear models.

The more important complication arises out of the fact that the values can correspond to grid points. In this case it is quite possible to have more parameters than observations, or an *underdetermined* model. Our focus will be to study linear models

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon \tag{1.1.1}$$

under these conditions.

Since the model is underdetermined most classical estimation methods will not be appropriate. Instead, we will make the initial assumption that there is some inherent smoothness in the behavior of $\boldsymbol{\beta}$ due to the fact that the measurements in $\mathbf{X}$ represent the same quantity over different times and locations. This can be seen in the following examples. We will discuss two of these problems in more detail later in this chapter.

A question that can be of interest in ocean studies is estimating the sources of water temperature changes. We can suppose there are two main sources for heating or cooling of a water column: heat entering vertically from the atmosphere, or an advection of warm or cool water coming from deeper locations. Given an $\mathbf{X}$ matrix that contains the vertical heating, and temperatures in the $\mathbf{y}$ vector, we could estimate the horizontal advection. A derivation of this model will be given in section 1.2. The $\boldsymbol{\beta}$ vector represents the horizontal advection in this problem. It would be reasonable to assume some inherent smoothness in this parameter from a physical point of view.

A more complicated oceanographic problem, as discussed by Dowd and Thompson (1996), is interpretation of data obtained from a ship-borne current profiler because of the presence of tides in the record. Their method utilizes a system of differential equations, referred to as shallow water equations, which are discretized to give a linear model. They use the current data to estimate tidal flows across the open boundary of the model. The tidal flows comprise $\boldsymbol{\beta}$, and an assumption of smoothness in the tidal flows is appropriate.

In general, the imposition of constraints in oceanographic and weather prediction models is important because the models are extremely large, but there are only a limited number of observations available. Therefore it is appropriate to penalize departures from spatial or temporal smoothness. These methods can be used effectively in data assimilation (Thacker and Long 1988), a procedure which blends new observations with the results of a previous forecast to predict a future state.

Problems of this type also arise in chemometrics. Boswell-Purdy (1995) studied infrared absorption data of glucose-spiked plasma, the dependent variable being glucose concentration and the explanatory variables being absorption levels. The absorption levels are measured at various wavenumbers. It is reasonable to think of the absorption levels being sampled from a continuous spectrum, and there should be some smoothness in the $\boldsymbol{\beta}$ vector which relates the absorption levels to the glucose concentration.

Finally, consider a linear model in which we wish to predict deep water characteristics, such as density or temperature, using shallow water observations. In this case our $\mathbf{X}$ matrix would contain the upper water observations, and we would expect a natural smoothness in the parameter that related these observations to the density or temperature in the deep ocean.

The underdetermined regression model can be considered as an example of an ill-posed inverse problem (O'Sullivan 1986, Hansen 1992). These are problems in which, to quote O'Sullivan, "classical solutions may be unacceptably sensitive to slight perturbations in the data." Inverse problems and methods arise frequently in geophysics and related areas (Vogel, Ofoegbu, Gorenflo, and Ursin 1990, Bennett 1992) and are reviewed by Tarantola (1987).

In the mathematics literature, the term regularization is often used to describe the imposition of a smoothness assumption on the model parameters (Tikhonov and Arsenin 1977). Regardless of the term used, the basic principle in these problems is to find a solution that is consistent with the observations and *a priori* beliefs about the behavior of the parameters. These beliefs often arise from the physical phenomenon in question.

There are several ways to introduce a smoothness constraint on the estimator arising from (1.1.1); these will be discussed in chapter 2. Our approach will be to consider (1.1.1) as representing a model in which

$$\textit{observation} = \textit{signal} + \textit{noise} \tag{1.1.2}$$

where both the signal and noise are random quantities. We will then impose our smoothing based on the signal-to-noise ratio of the model. We note that what we are calling smoothing will not be based on smoothing in spatial dimensions. Instead the smoothing, or shrinkage, of the estimators will rely on the covariance structure of the explanatory variables. This point will be discussed in greater detail in sections 2.2 and 3.3.2.

A priori knowledge of the signal-to-noise ratio, or at least a possible range for its value, may be available in some dynamic models. Although this will be case-specific, a general argument in favour of this is as follows. We can think of the noise term as being comprised of model error and measurement, or instrumental error. We can often have a good idea of the instrumental error variability, such as the accuracy in temperature recordings. We can use this knowledge to help set a bound on the noise variability, which can lead to a bound being placed on the signal-to-noise ratio. A similar assumption is often made in measurement error models (Casella and Berger 1990, pg. 587). This will be discussed in more detail in chapter 3.

The form (1.1.2) is similar to that used in objective analysis, or optimal interpolation. As described by Bretherton, Davis, and Fandry (1976), objective analysis is a method which can be used to interpolate between observations taken on an array. This allows for the construction of an estimate for an entire field from scattered observations. It is used in both meteorology and oceanography. We present a simplified version of objective analysis at this point to illustrate the method, and to briefly indicate how it compares with the method we will propose.

Symbolically we can write (1.1.2) as

$$y_i = \theta_i + \epsilon_i$$

where each observation $y_i = y_i(\mathbf{x}_i)$ depends on its location $\mathbf{x}_i$. We assume that

$$\begin{aligned} \mathrm{E}(\theta_i) &= \mathrm{E}(\epsilon_i) \;=\; 0 \\ \mathrm{E}(\epsilon_i\epsilon_j) &= \begin{cases} \sigma_\epsilon^2 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \\ \mathrm{E}[\theta(\mathbf{x})\theta(\mathbf{x}+\Delta\mathbf{x})] &= \sigma_\theta^2 r(\Delta\mathbf{x}) \end{aligned}$$

$$\mathrm{E}[\epsilon_i \theta_i(\mathbf{x}_i)] = 0 \quad \text{(error, signal uncorrelated)}$$

where $r(\Delta\mathbf{x})$ is the correlation function. So the signal covariance only depends on the distance between the locations.

We want to estimate $\theta(\mathbf{x})$ at an arbitrary location $\mathbf{x}$, using the available observations $y_1, \ldots y_n$. We assume that $\hat{\theta}$ is formed as a linear combination of the y_i values:

$$\theta(\mathbf{x}) = \alpha_1 y_1 + \ldots + \alpha_n y_n + \text{error} .$$

We then estimate $\boldsymbol{\alpha}$ by minimizing

$$\mathrm{E}[(y(\mathbf{x}) - \boldsymbol{\alpha}'\mathbf{y})^2] .$$

Given σ_ϵ^2, σ_θ^2 and $r(\Delta\mathbf{x})$ we find

$$\hat{\boldsymbol{\alpha}} = \begin{bmatrix} 1+q & r(\mathbf{x}_1 - \mathbf{x}_2) & \ldots & r(\mathbf{x}_1. - \mathbf{x}_{n-1}) \\ r(\mathbf{x}_2 - \mathbf{x}_1) & 1+q & \ldots & r(\mathbf{x}_2 - \mathbf{x}_{n-2}) \\ \vdots & \vdots & \vdots & \vdots \\ r(\mathbf{x}_{n-1} - \mathbf{x}_1) & r(\mathbf{x}_{n-1} - \mathbf{x}_2) & \ldots & 1+q \end{bmatrix}^{-1} \begin{bmatrix} r(\mathbf{x} - \mathbf{x}_1) \\ r(\mathbf{x} - \mathbf{x}_2) \\ \vdots \\ r(\mathbf{x} - \mathbf{x}_n) \end{bmatrix}$$

so $\hat{\theta} = \hat{\boldsymbol{\alpha}}'\mathbf{y}$, where $q^{-1} = \sigma_\theta^2/\sigma_\epsilon^2$ can be thought of as the signal-to-noise ratio.

Our model will also utilize knowledge of the signal-to-noise ratio, but there are differences in the approaches. Our model will be more general because it will allow the explanatory and response variables to measure different quantities. Objective analysis, as described above, assumes the signal and noise measure the same quantity. The presence of explanatory variables in our model will also mean we will be able to combine information in the $\mathbf{X}$ matrix in our model with assumptions about the covariance structure.

1.2 Motivating Examples

We now discuss two distinct problems that were alluded to in section 1.1, one of which will be analyzed in chapter 5.

1.2.1 Discretization of a Dynamic Model

Consider the example of water temperature change described in section 1.1. The following modified diffusion equation can be used to describe water temperature changes over time and depth:

$$\frac{\partial T}{\partial t} = k\frac{\partial^2 T}{\partial z^2} + \Gamma(z,t)\ , \tag{1.2.1}$$

where k represents a diffusion coefficient for heat entering vertically, and $\Gamma(z,t)$ represents horizontal advection, which we can think of as warm or cold water entering the water from deeper locations. Our goal is to solve for $\Gamma(z,t)$ at all locations and times, given k and scattered temperature measurements. To achieve this, we demonstrate a procedure to discretize (1.2.1) to yield a linear model of the form (1.1.1).

We assume a constant time step Δt and constant spatial distance Δz, and discretize (1.2.1) by writing

$$\frac{T_i^{t+1} - T_i^t}{\Delta t} = \frac{k}{(\Delta z)^2}\left(T_{i+1}^t - 2T_i^t + T_{i-1}^t\right) + \Gamma_i^{t+1}\ ,$$

where $i = 1,\ldots,n$ represents spatial location and $t = 1,\ldots,N$ represents temporal location. We can rewrite this model as

$$T_i^{t+1} = \alpha T_{i+1}^t + (1-2\alpha)T_i^t + \alpha T_{i-1}^t + \Delta t \Gamma_i^{t+1}$$

where $\alpha = k\Delta t/(\Delta z)^2$. We will use the initial and boundary conditions $T_0^t = T_{n+1}^t = 0$ and $T_i^0 = 0$ for notational purposes.

Taking all values at time t together in a vector gives us

$$\mathbf{T}^{t+1} = \mathbf{A}\mathbf{T}^t + \Delta t\mathbf{\Gamma}^{t+1} \tag{1.2.2}$$

where

$$\mathbf{A} = \begin{pmatrix} 1-2\alpha & \alpha & 0 & \mathbf{0}' & \ldots & \mathbf{0}' \\ \alpha & 1-2\alpha & \alpha & \mathbf{0}' & \ldots & \mathbf{0}' \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \ldots & 0 & \alpha & 1-2\alpha \end{pmatrix} .$$

If we stack all our values of $\mathbf{T}^t$ together, we find

$$\begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ -\mathbf{A} & \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & -\mathbf{A} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{T}^1 \\ \mathbf{T}^2 \\ \vdots \\ \mathbf{T}^N \end{pmatrix} = \Delta t \begin{pmatrix} \mathbf{\Gamma}^1 \\ \mathbf{\Gamma}^2 \\ \vdots \\ \mathbf{\Gamma}^N \end{pmatrix}$$

or

$$\mathbf{D}_1\mathbf{T} = \mathbf{D}_2\mathbf{\Gamma} \tag{1.2.3}$$

where $\mathbf{D}_2 = \Delta t\mathbf{I}$.

Let $\mathbf{y}$ represent the scattered observed temperatures. These will be related to the true temperatures by

$$\mathbf{y} = \mathbf{G}\mathbf{T} + \epsilon\ .$$

In the simplest situation the matrix $\mathbf{G}$ could be a matrix of 1's and 0's, indicating at which grid points we have observations. We can rewrite $\mathbf{y}$ as a function of $\mathbf{\Gamma}$ using (1.2.3):

$$\mathbf{y} = \mathbf{G}\mathbf{D}_1^{-1}\mathbf{D}_2\mathbf{\Gamma} + \epsilon\ .$$

We now see we have transformed our model (1.2.1) to the regression model (1.1.1), where $\mathbf{X} = \mathbf{H}\mathbf{D}_1^{-1}\mathbf{D}_2$. We have used an explicit differencing scheme, which may be numerically unstable for large values of Δt. It could be replaced by an implicit scheme which would allow large time steps. The only thing in this derivation that would change is the matrix $\mathbf{D}_1$; each identity matrix would be replaced by another matrix, say $\mathbf{B}$, which would be similar in structure to $\mathbf{A}$.

The $\mathbf{X}$ matrix contains the diffusion coefficient k, which we have treated to be constant. However, we could treat that as a term that is measured, with error, at our grid points. We would still obtain a linear model from the discretization, but the $\mathbf{A}$ matrix would then change over time, and its elements would include $\alpha_i^t = (\Delta t/(\Delta z)^2)k_i^t$. This example shows how dynamic models can be treated as linear models in many cases, allowing us to use statistical procedures that do not ignore important dynamics that are known about the problem.

1.2.2 Prediction of Deep Water Measurements

Our main example involves modelling deep water temperature and salinity using a multiple linear regression model, where shallow water temperatures or salinity readings are the explanatory variables. The two data sets consist of measurements collected from deep CTD (conductivity-temperature-depth) stations off Point Sur, California during the summer months. The data are described in detail by Haney, Hale, and Collins (1995). Because some of the stations are close together it is not clear that the samples collected are independent. To deal with this Haney et al. (1995) removed neighbouring CTD stations that had a correlation greater than 0.5 with a station previously selected for analysis. This left a data set of 64 stations, where the average distance between stations was 25 km. Haney et al. (1995) felt this left a data set in which locations could be treated as independent, and we will use the same assumption in our analyses. At each station we have a collection of 200 salinity and temperature measurements, taken from the sea surface to a depth of 2000 meters, at 10 meter intervals.

This is an important practical problem in oceanography for many reasons. Clearly, far more observations are available in the upper ocean. Also, new ocean equipment can carry out measurements that are nearly synoptic in time, but only if the measurements are restricted to a shallow water range (Haney et al. 1995). Therefore it is important to discover the extent to which an upper ocean survey can describe features in the deep ocean.

The linear model we will study uses the readings in the upper 800 m as our explanatory variables. This means we have 80 parameters and only 64 observations, so the model is underdetermined. The example also has the other characteristics in which we are interested. It is reasonable to assume our explanatory variables are random, naturally thought of as realizations of a random process, with the observations measured with error. Finally, it may be reasonable to have some *a priori* knowledge of the signal-to-noise ratio. At the very least, we would expect it to be small. This is because it will be the variability of the noise that will play a major role, particularly if we attempt to use very deep observations as our response.

1.3 Outline

This thesis will be broken down in the following manner. In chapter 2 we review the manner in which we can introduce a smoothness constraint in linear regression models, and some of the present methods used for choosing the smoothing parameter. We will see some of them are not applicable in underdetermined models. In chapter 3 we derive our model and estimators, first under the assumption of a known value of the signal-to-noise ratio, followed by a method for estimating its value. This chapter will include simulation results on the performance of the method, based on its predictive ability, even though it is not derived from a prediction-based criterion. In chapter 4 we discuss some robust extensions, which give protection against outliers in the residuals and influential observations in the explanatory variables. Chapter 5 contains analyses of the California Current data sets described in section 1.2. Finally, chapter 6 will summarize our results and discuss some possible further research.

Chapter 2

Background

2.1 Introduction

In this chapter we will give an overview of the use of smoothing techniques in statistics, focusing on the linear model. This will include detail on the need, and desire, for smoothing in many situations, and approaches by which we can impose smoothness constraints in problems. We will discuss some of the well known methods for choosing the smoothing parameter, and how they are often not applicable to the underdetermined regression model. This will lead us into the proposed methods of the next chapter.

2.2 Notion of Smoothing

As mentioned in the introduction to this chapter, we begin with some motivation on why we may want to use smoothed estimators. A smoothed estimator is one which attempts to achieve a balance between a good fit of the model and observations and the regularity we wish to impose on our solution. In most applications a smoothed estimator has greater bias than, say, the maximum likelihood estimator (MLE), but has lower variability. In multiple linear regression much of the original motivation for the use of smoothed estimators was the presence of multicollinearity in the data. If we consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.2.1}$$

where $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$, multicollinearity implies that our least squares estimator

$$\hat{\boldsymbol{\beta}}_{ls} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{2.2.2}$$

although unbiased, would be highly variable because $\mathbf{X}'\mathbf{X}$ is ill-conditioned. This led to the development of ridge regression (Hoerl and Kennard 1970), which uses

$$\hat{\boldsymbol{\beta}}_{\lambda} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{C}'\mathbf{C})^{-1}\mathbf{X}'\mathbf{y}\ . \tag{2.2.3}$$

In many cases $\mathbf{C} = \mathbf{I}$ is chosen. We can think of λ as the smoothing parameter. We will outline some different approaches for deriving this estimator in the next section. This estimator is clearly biased for $\lambda > 0$, but has smaller variability than the least squares estimator (Weisberg 1985, pg. 255). Hoerl and Kennard (1970) show that, for nonstochastic λ and $\mathbf{C} = \mathbf{I}$, we can find a $\lambda > 0$ for which

$$\mathrm{E}[(\hat{\boldsymbol{\beta}}_{\lambda} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{\lambda} - \boldsymbol{\beta})] \leq \mathrm{E}[(\hat{\boldsymbol{\beta}}_{ls} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{ls} - \boldsymbol{\beta})]\ .$$

An assumption of smoothness of $\hat{\boldsymbol{\beta}}$ is often reasonable, as indicated in chapter 1.

Before we continue we will make a comment on terminology to be used in this thesis. In the statistics literature estimator (2.2.3) is called a smoothed estimator regardless of the choice of $\mathbf{C}$. In much of the work in ocean or climate studies, for example, (2.2.3) would not be referred to a smoothed estimator if $\mathbf{C} = \mathbf{I}$. In that case it may only be described as a shrinkage estimator, because all that has been penalized is the squared length of $\boldsymbol{\beta}$. "Smoothness", in the physical sciences, often refers to spatial smoothing, *i.e.* penalizing specific local or global properties of $\boldsymbol{\beta}$, which can only be achieved with $\mathbf{C} \neq \mathbf{I}$. However, throughout this thesis, whenever we refer to a smoothed estimator, it will mean an estimator of the form (2.2.3) with $\mathbf{C} = \mathbf{I}$ permitted.

We should note that the use of smoothed estimators occurs in many other types of problems, as discussed by Titterington (1985). This includes nonparametric density estimation, smoothing splines and multinomial smoothing. In all cases the following theme is present: find an estimator which achieves a reasonable tradeoff between a good fit to the data and smoothness of $\hat{\boldsymbol{\beta}}$.

2.3 Methods of Smoothing

There are several ways to derive the ridge, or smoothed, estimator (2.2.3). We will outline several methods, which arise from the frequentist and Bayesian perspectives. All of these results assume λ is fixed, as is the $\mathbf{X}$ matrix.

2.3.1 Penalized Least Squares

Perhaps the simplest way to derive the estimator (2.2.3) is to minimize the sum of squares function subject to a constraint on $\boldsymbol{\beta}$ by introducing a Lagrange multiplier term:

$$\min_{\boldsymbol{\beta}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\lambda\boldsymbol{\beta}'\mathbf{C}'\mathbf{C}\boldsymbol{\beta}$$

By differentiating with respect to $\boldsymbol{\beta}$ and setting the result equal to $\mathbf{0}$ we obtain (2.2.3). This method can also be seen as penalized likelihood, if we assume that ϵ is normally distributed. This will be the approach used to find our estimators in chapter 3.

2.3.2 Augmented Data

Another method is the use of augmented data (Askin and Montgomery 1980). In this method we assume that $\boldsymbol{\zeta} = -\sqrt{\lambda}\mathbf{C}'\boldsymbol{\beta}$, where $\boldsymbol{\zeta} \sim \mathrm{N}(\mathbf{0},\mathbf{I})$ is independent of $\boldsymbol{\epsilon}$. We then augment our $\mathbf{X}$, $\mathbf{y}$ and ϵ matrices as follows:

$$\begin{aligned}\mathbf{y}_{aug} &= \mathbf{X}_{aug}\boldsymbol{\beta}+\boldsymbol{\epsilon}_{aug}\\ \begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix} &= \begin{bmatrix}\mathbf{X}\\ \sqrt{\lambda}\mathbf{C}'\end{bmatrix}\boldsymbol{\beta}+\begin{bmatrix}\boldsymbol{\epsilon}\\ \boldsymbol{\zeta}\end{bmatrix}\end{aligned}$$

Then minimizing

$$(\mathbf{y}_{aug}-\mathbf{X}_{aug}\boldsymbol{\beta})'(\mathbf{y}_{aug}-\mathbf{X}_{aug}\boldsymbol{\beta})$$

with respect to $\boldsymbol{\beta}$ gives us the solution (2.2.3). Thacker (1988) refers to this augmentation procedure as introducing "bogus" data into the problem, which are hypothetical observations of slope, curvature or other indications of a variable's smoothness.

2.3.3 Bayesian Approach

We can also derive (2.2.3) from a Bayesian perspective. Suppose that the prior distribution of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \sim \mathrm{N}\left(\mathbf{0}, \frac{\sigma^2}{\lambda}(\mathbf{C}'\mathbf{C})^{-1}\right)$$

where $\mathbf{C}'\mathbf{C}$ is nonsingular, and a noninformative prior is placed on λ. If $\mathbf{C}'\mathbf{C}$ is of rank $k < p$ we can use

$$\pi(\boldsymbol{\beta}) \propto \left(\frac{\sigma^2}{\lambda}\right)^{k/2} \exp\left(-\frac{\lambda}{\sigma^2}\boldsymbol{\beta}'\mathbf{C}'\mathbf{C}\boldsymbol{\beta}\right)$$

which is an improper prior. There are cases when it may be appropriate to choose $\mathbf{C}'\mathbf{C}$ as singular, such as when penalizing approximate derivatives. In either case a posterior distribution can be computed.

We find the posterior distribution is proportional to

$$\begin{aligned} \pi(\boldsymbol{\beta}|\mathbf{y}) &\propto \pi(\boldsymbol{\beta})f(\mathbf{y}|\boldsymbol{\beta}) \\ &\propto \sigma^{-(n+p)}\lambda^{p/2}\exp[-(1/2\sigma^2)Q(\boldsymbol{\beta})] \end{aligned} \qquad (2.3.1)$$

where $Q(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\mathbf{C}'\mathbf{C}\boldsymbol{\beta}$. The posterior mean of (2.3.1) is (2.2.3).

Barry (1995) considers a joint prior density

$$\pi(\boldsymbol{\beta}, \sigma^2, \lambda) \propto \pi(\sigma^2, \lambda)\pi(\boldsymbol{\beta}|\sigma^2, \lambda)$$

and derives Jeffreys' prior for σ^2 and λ from the likelihood function. See Lindley and Smith (1972) for more discussion of the Bayesian linear model.

2.3.4 Other Approaches

There are other ways to use biased estimators in linear regression, although they are not often referred to as smoothed estimators. We will discuss two of them at this point: principal component regression and partial least squares.

Principal Component Regression

In principal component regression (PCR) we wish to eliminate the dimensions of $\mathbf{X}$ which are causing a multicollinearity problem (Weisberg 1985, pg. 257) by examining the principal components of $\mathbf{X}$.

Let the singular value decomposition (SVD) of $\mathbf{X}$ be

$$\mathbf{X} = \mathbf{UDV}' \tag{2.3.2}$$

where $\mathbf{U}$ is a $n \times n$ orthogonal matrix, $\mathbf{V}$ is a $p \times p$ orthogonal matrix, and $\mathbf{D}$ has the singular values of $\mathbf{X}$ along its main diagonal and zeros elsewhere (Golub and Van Loan 1989, pg. 71).

We use (2.3.2) to rewrite (2.2.1) as

$$\begin{aligned} \mathbf{y} &= \mathbf{XVV}'\boldsymbol{\beta} + \epsilon \\ &= \mathbf{P}\boldsymbol{\alpha} + \epsilon \end{aligned}$$

where $\boldsymbol{\alpha} = \mathbf{V}'\boldsymbol{\beta}$ and $\mathbf{P} = \mathbf{XV}$. Then, provided $\mathbf{X}'\mathbf{X}$ is invertible, $\hat{\boldsymbol{\alpha}} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{y} = \mathbf{V}'\hat{\boldsymbol{\beta}}_{ls}$. We then eliminate the dimensions for which the singular values of $\mathbf{X}$ are small. The number of principal components to retain may be determined by cross validation, which will be described in detail in the next section. PCR assumes the $\mathbf{X}$ matrix is centered and scaled.

In the underdetermined model, assuming rank($\mathbf{X}$) $= n$, there would be $p - n$ singular values equal to 0. Simulation results (Frank and Friedman 1993) suggest PCR performs worse than ridge regression.

Partial Least Squares

Partial least squares (PLS) is a commonly used technique in chemometrics. It is described in detail by Helland (1988) and Frank and Friedman (1993). PLS forms a relationship between $\mathbf{y}$ and $\mathbf{X}$ by constructing new explanatory variables, each of them being a linear combination of the rows of $\mathbf{X}$, $\mathbf{x}_1', \ldots, \mathbf{x}_n'$. It differs from PCR, however, in that the values of both $\mathbf{x}_i'$ and y influence the new variables formed.

Following the notation of Helland (1988), suppose we can write

$$\begin{aligned} \mathbf{X} &= \sum_{i=1}^{a} \mathbf{t}_i \mathbf{p}_i' + \mathbf{E}_a \\ \mathbf{y} &= \sum_{i=1}^{a} \mathbf{t}_i q_i + \mathbf{f}_a \end{aligned}$$

where the $\mathbf{t}_i$ vectors may be considered latent variables. The goal is to find the $\mathbf{p}_i'$ and q_i values using both of these equations to get a good fit. The values for $\mathbf{t}_i$, $\mathbf{p}_i'$ and q_i are determined by induction, where each $\mathbf{t}_i$ is determined as a linear combination of the $\mathbf{x}'$-residuals from the previous step in the inductive process. There are different, but equivalent, algorithms, to do this. The number of terms a to be used must also be found; cross-validation is often used to determine its value.

All three methods shrink the least squares estimator (2.2.2), in the sense that the length of the resulting estimator is shorter. However, ridge regression and PCR will shrink (2.2.2) in all eigendirections, while this may not be the case for PLS (Frank and Friedman 1993).

Stone and Brooks (1990) developed the method of continuum regression, in which they consider a range of possible estimators, each associated with a parameter γ and a vector $\mathbf{c}_\gamma$. They show that ridge regression, PCR and PLS are all special cases of the method. However, Björkström and Sundberg (1996) illustrate that the predictor $\mathbf{x}'\hat{\boldsymbol{\beta}}$ may not change continuously as γ is varied. In these cases, the correspondence between the continuum regression estimators and the ridge regression predictors is not one-to-one.

Frank and Friedman (1993) and Breiman and Friedman (1997) discuss extensive simulation results to compare these methods, and find that ridge regression tends to outperform the other two.

2.4 Choice of Smoothing Parameter

Up to now we have assumed that the smoothing parameter was fixed. In most instances it will have to be estimated from the data. We will now discuss some of the methods that have been proposed in the literature, and how there are problems with many of them in underdetermined models. They may not be applicable or they require very strict model assumptions. Many of these methods are discussed in Titterington (1985) and Hall and Titterington (1987).

Many methods try and choose the smoothing parameter that will yield a good estimator of the minimizer of a mean square error criterion, such as

$$\mathrm{E}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)] \text{ or } \mathrm{E}[(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)'(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)]$$

2.4.1 Review of Methods

Lawless (1978) summarizes many of the early methods proposed for choosing ridge parameters. The methods are motivated by using the SVD of $\mathbf{X}$ in (2.3.2). In our discussion of principal component regression, recall that we defined $\boldsymbol{\alpha} = \mathbf{V}'\boldsymbol{\beta}$. Hoerl and Kennard (1970) show that choosing $\lambda < \sigma^2/\alpha^2_{max}$, where $\alpha^2_{max} = \max(\alpha_1^2, \ldots, \alpha_p^2)$, will ensure

$$\mathrm{E}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)] < \mathrm{E}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ls})'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ls})]$$

under the assumption that $\mathbf{C} = \mathbf{I}$. This is used as the basis for choosing (Lawless 1978)

$$\hat{\lambda} = \frac{p\hat{\sigma}^2}{\sum_{i=1}^p \hat{\alpha}_i^2}$$

where $\hat{\sigma}^2$ is an unbiased estimator of σ^2.

This choice presents a problem for us, because we do not have the least squares solution to use. Other methods presented by Lawless (1978) also have the disadvantage of depending on the ordinary least squares solution. Hoerl and Kennard (1970) replace $\sum_{i=1}^{p} \hat{\alpha}_i^2$ with $\max(\hat{\alpha}_1^2, \ldots, \hat{\alpha}_p^2)$, but the same problem mentioned above exists. By their argument it also seems that this choice of $\hat{\lambda}$ may not impose enough smoothing in the model. The reason is that, although $\lambda = \sigma^2/\alpha_{\max}^2$ will give us smaller MSE than the least squares estimator, this choice of λ is not necessarily close to the one that minimizes the MSE; it may be a great deal smaller.

Since

$$\mathrm{E}[(y - \mathbf{X}\beta)'(y - \mathbf{X}\beta)] = n\sigma^2$$

Hall and Titterington (1987) suggest we could try to find λ such that

$$(y - \mathbf{X}\hat{\beta}_\lambda)'(y - \mathbf{X}\hat{\beta}_\lambda) = n\sigma^2$$

The drawback to this approach is it requires a good estimator of σ^2.

Wahba (1983) suggests

$$\frac{(y - \mathbf{X}\hat{\beta}_\lambda)'(y - \mathbf{X}\hat{\beta}_\lambda)}{n - \mathrm{tr}(\mathbf{H}_\lambda)}$$

could be a good estimator of σ^2, where $\mathbf{H}_\lambda = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{C}'\mathbf{C})^{-1}\mathbf{X}'$. This leads us to consider λ that solves

$$(y - \mathbf{X}\hat{\beta}_\lambda)'(y - \mathbf{X}\hat{\beta}_\lambda) = \hat{\sigma}^2[n - \mathrm{tr}(\mathbf{H}_\lambda)]$$

Wahba (1983) refers to $(n - \mathrm{tr}(\mathbf{H}_\lambda))$ as the "equivalent degrees of freedom for error". This requires a consistent estimator of σ^2, which is not available in our context.

An approach that has been used in ill-posed problems for choosing a smoothing parameter is the L-curve method (Hansen 1992). It examines the plot of $\hat{\beta}_\lambda'\mathbf{C}'\mathbf{C}\hat{\beta}_\lambda$ versus $(\mathbf{y}-\mathbf{X}\hat{\beta}_\lambda)'(\mathbf{y}-\mathbf{X}\hat{\beta}_\lambda)$ as a function of λ, often on the log-log scale. The criterion proposed is to choose the value of λ that corresponds to the point of maximum curvature, or the corner point, on the L-curve. The justification for this choice is a smaller value of λ will give a larger length of $\hat{\beta}_\lambda$ with only a marginally smaller residual, while a larger λ will produce a larger residual and only a marginally shorter $\hat{\beta}_\lambda$. Clements, Carroll, and Horáček (1996) have used the method in studying an inverse problem in electrocardiography.

2.4.2 Bayesian Approach

We return to our Bayesian derivation of (2.2.3), which was the mean of the posterior distribution (2.3.1). Using a fully Bayes approach, λ is a hyperparameter with a noninformative prior. First, we rewrite the exponent $Q(\boldsymbol{\beta})$ in (2.3.1) as

$$Q(\boldsymbol{\beta}) = Q^*(\hat{\boldsymbol{\beta}}_\lambda) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)'\mathbf{M}^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)$$

where

$$Q^*(\hat{\boldsymbol{\beta}}_\lambda) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda) + (\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)$$

and $\mathbf{M} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{C}'\mathbf{C})$. Then we can write (2.3.1) as

$$\begin{aligned} \pi(\boldsymbol{\beta}, \sigma^2, \lambda) &\propto \frac{\sigma^{-n}\lambda^{p/2}}{|\mathbf{X}'\mathbf{X} + \lambda\mathbf{C}'\mathbf{C}|^{1/2}} \exp\left[\frac{-Q^*(\hat{\boldsymbol{\beta}}_\lambda)}{2\sigma^2}\right] \\ &\quad \times \sigma^{-p}|\mathbf{X}'\mathbf{X} + \lambda\mathbf{C}'\mathbf{C}|^{1/2} \exp\left[\frac{-1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)'\mathbf{M}^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)\right] \\ &\propto \pi(\sigma^2, \lambda)\pi(\boldsymbol{\beta}|\sigma^2, \lambda) \end{aligned}$$

We then find the marginal posterior of λ by integrating $\pi(\sigma^2, \lambda)$ over σ^2. This eventually yields the marginal posterior, post(λ):

$$\text{post}(\lambda) \propto \frac{\lambda^{p/2}}{|\mathbf{X}'\mathbf{X} + \lambda\mathbf{C}'\mathbf{C}|^{1/2}}[Q^*(\hat{\boldsymbol{\beta}}_\lambda)]^{1-n/2}$$

We could choose the mode of this distribution as our choice of $\hat{\lambda}$.

If we do not take this approach, empirical Bayes methods must be used. Nebebe and Stroud (1986) define the distinction between full Bayes and empirical Bayes as depending on one's willingness to assign a prior distribution to the hyperparameter λ and integrate with respect to λ, or to use a point estimate for λ as if it was known.

2.4.3 Crossvalidatory Choice

One of the most common methods of choosing $\hat{\lambda}$ is by some type of crossvalidatory choice; see Stone (1974) for one of the original overviews of the topic. It usually involves omitting one data point at a time, calculating estimators using the remaining

$n-1$ data points, and seeing how well we predict the value of the omitted observation. This is repeated for each observation. Mathematically, if we write our estimate of $\boldsymbol{\beta}$ using $n-1$ observations as

$$\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'_{(i)}\mathbf{X}_{(i)} + \lambda\mathbf{C}'\mathbf{C})^{-1}\mathbf{X}'_{(i)}\mathbf{y}_{(i)}$$

where $\mathbf{X}_{(i)}, \mathbf{y}_{(i)}$ are the $\mathbf{X}$ and $\mathbf{y}$ matrices with the *ith* row removed, we could choose $\hat{\lambda}$ as the value which minimizes

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}'_i\hat{\boldsymbol{\beta}}_{(i)})^2 \tag{2.4.1}$$

We can rewrite (2.4.1) in terms of the full $\mathbf{X}$ and $\mathbf{y}$ matrices, modifying the results of Weisberg (1985, pg. 293). Using (2.4.1) is often referred to as leave-1-out, or ordinary cross validation (CV).

The idea of ordinary CV can be extended to leaving out groups of data at once, sometimes called multifold CV. This method has often been used in model selection procedures, *e.g.* Shao (1993).

The bootstrap can also be used for choosing λ, again using a prediction based criterion. Delaney and Chaterjee (1986) outline an algorithm for its use, which involves measuring the loss in predicting points not selected in the bootstrap sample.

Although ordinary CV has an appealing form, and is based on a prediction criterion, which is often the focus of regression modelling, it can break down in certain situations. Golub, Heath, and Wahba (1979) show that if $\mathbf{X}$ is non-zero only along its main diagonal, then (2.4.1) does not have a unique minimizer in λ. For this reason they proposed to modify ordinary CV, giving it the appropriate name of generalized cross validation, or GCV. Since this method is used in many contexts, such as ridge regression, nonparametric regression and spline models, we will discuss many of the properties of GCV estimators in the next section.

2.5 Generalized Cross Validation

As previously mentioned, GCV can be viewed as a generalized, or weighted version, of ordinary CV. It chooses λ to minimize

$$\mathrm{GCV}(\lambda) = \frac{n[(\mathbf{I}-\mathbf{H})\mathbf{y}]'[(\mathbf{I}-\mathbf{H})\mathbf{y}]}{[\mathrm{tr}(\mathbf{I}-\mathbf{H})]^2} \tag{2.5.1}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{C}'\mathbf{C})^{-1}\mathbf{X}'$. Golub et al. (1979) refer to GCV as a rotation-invariant form of ordinary CV. If we think of $\mathbf{X}$ as a mapping from an arbitrary space $\mathcal{E}_p$ to $\mathcal{E}_n$, Wahba (1978) describes the GCV estimate as coming from rotating the $\mathcal{E}_n$ coordinate system to a new design matrix $\tilde{\mathbf{X}}$, such that $\tilde{\mathbf{X}}$ is circulant (each column of $\tilde{\mathbf{X}}$ is equal to the previous column rotated downwards by one element), and doing ordinary CV in the new system.

Craven and Wahba (1979) introduce GCV for choosing the smoothing parameter in spline models. GCV has been used in many applications in atmospheric models (O'Sullivan and Wahba 1985) and in larger scale models, such as numerical weather prediction (Wahba, Johnson, Gao, and Gong 1995).

Let $\hat{\lambda}_{GCV}$ denote the value of λ which minimizes (2.5.1). Golub et al. (1979) propose that $\hat{\lambda}_{GCV}$ is a good estimator of λ which minimizes the following expected predictive mean square error (EPMSE):

$$\begin{aligned} \mathrm{E}[\mathrm{PMSE}(\lambda)] &= \frac{1}{n}\mathrm{E}[(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)'(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)] \\ &= \frac{1}{n}(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}-\mathbf{H})^2\mathbf{X}\boldsymbol{\beta} + \frac{\sigma^2}{n}\mathrm{tr}(\mathbf{H}^2) \end{aligned} \tag{2.5.2}$$

We will let $\hat{\lambda}_{EPMSE}$ denote the λ value which minimizes (2.5.2).

The value $\hat{\lambda}_{GCV}$ is random, since it depends on $\mathbf{y}$. However, much of the original work on GCV does not deal with $\hat{\lambda}_{GCV}$ but $\hat{\lambda}_{EGCV}$, which minimizes the expected value of (2.5.1):

$$\mathrm{E}[\mathrm{GCV}(\lambda)] = n\left[\frac{(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}-\mathbf{H})^2\mathbf{X}\boldsymbol{\beta} + \sigma^2[n - 2\mathrm{tr}(\mathbf{H}) + \mathrm{tr}(\mathbf{H}^2)]}{[\mathrm{tr}(\mathbf{I}-\mathbf{H})]^2}\right] \tag{2.5.3}$$

Therefore $\hat{\lambda}_{EGCV}$ is a nonrandom quantity. The following results, proven by Golub et al. (1979), establish how we expect $\hat{\lambda}_{EGCV}$ to behave relative to $\hat{\lambda}_{EPMSE}$, and the assumptions required:

Result 2.1: Define

$$h = \left[\frac{2\mathrm{tr}(\mathbf{H})}{n} + \frac{[\mathrm{tr}(\mathbf{H})]^2}{n\mathrm{tr}(\mathbf{H}^2)}\right] \frac{1}{(1-\mathrm{tr}(\mathbf{H})/n)^2}$$

Then $\hat{\lambda}_{EPMSE}$ *and* $\hat{\lambda}_{EGCV}$ *satisfy*

$$\mathrm{I}_r = \frac{\mathrm{E[PMSE}(\hat{\lambda}_{EGCV})]}{\mathrm{E[PMSE}(\hat{\lambda}_{EPMSE})]} \le \frac{1 + h(\hat{\lambda}_{EPMSE})}{1 - h(\hat{\lambda}_{EGCV})} \quad \square$$

However, this result holds if and only if $1 - h(\hat{\lambda}_{EGCV}) > 0$, and it can be shown that this will not be true if $n < p$. Wahba (1990, pg. 57) makes this point in a different context.

Golub et al. (1979) prove the following result in the underdetermined case, given certain assumptions on the $\mathbf{X}$ matrix. Specifically, we assume that the sum of the squared elements in $\mathbf{X}$ remains bounded as the number of columns of $\mathbf{X}$ increases. We also make an assumption on the behavior of the eigenvalues of $\mathbf{XX}'$, as described in the following result.

Result 2.2. Suppose

$$\sum_{j=1}^{\infty} x_{ij}^2 \le k_1 < \infty \text{ for all } i \ , \ \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{\infty} x_{ij}^2 = k_3 < \infty$$

and the eigenvalues of $\mathbf{XX}'$, λ_v, $v = 1, \ldots, n$ *satisfy* $\lambda_v \simeq nv^{-m}$, *where* $m > 1$. *This means that* $k_3 = \sum_{v=1}^{\infty} v^{-m}$. *Then*

$$\frac{\mathrm{tr}(\mathbf{H})}{n} \to 0 \ , \quad \frac{[\mathrm{tr}(\mathbf{H})]^2}{n\mathrm{tr}(\mathbf{H}^2)} \to 0 \quad \text{if } n\lambda^{1/m} \to \infty \ .$$

Result 2.1 may now be applied to show $\mathrm{I}_r \downarrow 1$ *as* $n \to \infty$ □

Craven and Wahba (1979) give similar results for smoothing splines. Wahba (1977) gives a slight simplification of the results in Golub et al. (1979). If $n > p$, then as $n \to \infty$,

$$\frac{\mathrm{E[PMSE}(\hat{\lambda}_{EGCV})]}{\mathrm{E[PMSE}(\hat{\lambda}_{EPMSE})]} = 1 + O\left(\frac{p}{n}\right)$$

If $n, p \to \infty$, $\sum \beta_v^2 < \infty$, $\lambda_v = O(nv^{-m})$, where $m > 1$, $v = 1, \ldots, n$, then

$$\frac{\mathrm{E}[\mathrm{PMSE}(\hat{\lambda}_{EGCV})]}{\mathrm{E}[\mathrm{PMSE}(\hat{\lambda}_{EPMSE})]} = 1 + o(1)$$

As we said above, these results do not make any claims about the behavior of $\hat{\lambda}_{GCV}$.

Kay (1992) extends some of these results on the optimality of $\hat{\lambda}_{EGCV}$. We include them at this point to illustrate the strict conditions we need to place on the structure of the model, especially the $\mathbf{X}$ matrix.

Kay (1992) uses the generalized singular value decomposition (GSVD) (Van Loan 1976) to express (2.2.1) in terms of the singular values of $\mathbf{X}$ and $\mathbf{C}$. The GSVD of $\mathbf{X}$ and $\mathbf{C}$ is as follows:

$$\mathbf{X} = \mathbf{U}\mathbf{D}_1\mathbf{P} \;,\; \mathbf{C} = \mathbf{V}\mathbf{D}_2\mathbf{P} \tag{2.5.4}$$

where

$$\mathbf{D}_1 = \begin{cases} (\mathbf{M}_1', \mathbf{0}')' & \text{if } n > p \\ (\mathbf{M}_1, \mathbf{0}) & \text{if } n < p \\ \mathbf{M}_1 & \text{if } n = p \end{cases} ,\; \mathbf{D}_2 = (\mathbf{C}_1, \mathbf{0}) \text{ if } q \leq p$$

and $\mathbf{M}_1 = \mathrm{diag}(m_1, \ldots, m_t)$, $\mathbf{C}_1 = \mathrm{diag}(c_1, \ldots, c_q)$, where $t = \min(n, p)$. The matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ satisfy $\mathbf{D}_1'\mathbf{D}_1 + \mathbf{D}_2'\mathbf{D}_2 = \mathbf{I}$ (Golub and Van Loan 1989, pg. 471). The $\mathbf{U}$ and $\mathbf{V}$ matrices are orthogonal, while $\mathbf{P}$ is non-singular. The matrices $\mathbf{M}_1$ and $\mathbf{C}_1$ may not be square; the m_i and c_i values may run along the main diagonal of a non-square matrix. Kay (1992) uses the GSVD to obtain the following result:

Result 2.3. Assume $p = O(n)$, $q = O(n)$, *n is large and that the following are true:*

1. *Assume* $m_i^2 \sim ni^{-\mu}/d_1$, $c_i^2 \sim d_2 i^{\kappa}$ *as* $n \to \infty$, *with constants* $d_1 > 0$, $d_2 > 0$, $\nu = \mu + \kappa > 1$ *and* $\mu, \kappa \geq 0$.

2. *Assume* $\lambda \to 0$ *as* $n \to \infty$, *while* $n\lambda^{1/\nu} \to \infty$.

3. *Assume*

$$b_1 = d_1 d_2^2 \sum_{i=1}^{\infty} i^{\nu+\kappa} \beta_i^2 < \infty$$

Then

$$\hat{\lambda}_{EGCV} = \left[\frac{\sigma^2 k_2}{2/(n\nu b_1)}\right]^{\nu/(2\nu+1)} (1+o(1))$$

where $o(1) \to 0$ as $n \to \infty$ and

$$k_2 = (d_1/d_2)^{-1/v} \int_0^\infty \frac{\mathrm{d}x}{(1+x^\nu)^2}$$

If we use $\hat{\lambda}_{EPMSE}$ as defined previously, then as a corollary we obtain

$$\frac{\hat{\lambda}_{EGCV}}{\hat{\lambda}_{EPMSE}} \sim 1$$

where

$$\hat{\lambda}_{EPMSE} = \left(\frac{\sigma^2 k_3}{n b_1}\right)^{\nu/(2\nu+1)} (1+o(1)) \ , \ k_3 = (d_1/d_2)^{-1/v} \int_0^\infty \frac{x^\nu \mathrm{d}x}{(1+x^\nu)^3} \quad \square$$

The corollary is an extension of Result 2.2, with the assumption on the behavior of the m_i values being similar to the assumption on the eigenvalues made in the previous result. The assumption made in point 3 of the result, involving the elements of the unknown parameter vector $\boldsymbol{\beta}$, is an additional strict assumption that is needed to achieve the desired optimality.

Li (1986) also gives convergence results based on the singular values of $\mathbf{X}$, but assumes $\mathbf{C} = \mathbf{I}$ in (2.2.3). Li (1986) proves that

$$\frac{\text{PMSE}(\hat{\lambda}_{GCV})}{\inf_{\lambda \geq 0} \text{PMSE}(\lambda)} \xrightarrow{P} 1 \text{ as } n \to \infty$$

This requires that

$$\inf_{\lambda \geq 0} \text{E}[(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)'(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)] \to \infty \text{ as } n \to \infty$$

and

$$\left(\frac{1}{n} \sum_{i=m+1}^{n} \lambda_i\right)^2 \Big/ \left(\frac{1}{n} \sum_{i=m+1}^{n} \lambda_i^2\right) \to 0 \text{ as } n \to \infty$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ are the eigenvalues of $\mathbf{X}'\mathbf{X}$, and m is chosen such that $m/n \to 0$. Li (1986) states this means the coefficient of variation of the λ_i's tends to infinity as $n \to \infty$. Hence we have a convergence in probability result for $\hat{\lambda}_{GCV}$.

This final result does show we have an asymptotic optimality result for $\hat{\lambda}_{GCV}$, but it requires some strong assumptions on the $\mathbf{X}$ matrix. It does hold, though, if $n < p$. This result is also derived under the assumption that the smoothing matrix $\mathbf{C}'\mathbf{C} = \mathbf{I}$. It is not clear if the result is easily extended to the non-identity matrix case. The selection of the smoothing matrix will be the subject of the next section. Andrews (1991) states that, if the errors are correlated, GCV will not be asymptotically optimal, in general.

Thompson, Kay, and Titterington (1989) give some cautionary remarks on the use of GCV. They point out that there is no guarantee that (2.5.1) will have a unique, easily identified minimum, and the derivative of (2.5.1) as $\lambda \rightarrow 0$ may not be negative if $n < p$. Delaney and Chaterjee (1986) show several examples in which GCV performs poorly. This was especially noticeable when the condition number of $\mathbf{X}$,

$$K(\mathbf{X}) = \gamma_{max}/\gamma_{min}$$

where γ_{max} and γ_{min} are the largest and smallest singular values of $\mathbf{X}$ respectively, was large. This is the type of situation we have in the underdetermined model, since our minimum singular value is 0.

2.6 Choice of Smoothing Matrix

The majority of the ridge regression literature focuses on using (2.2.3) with $\mathbf{C}'\mathbf{C} = \mathbf{I}$, so the penalty is based on the squared length of $\boldsymbol{\beta}$. However, there may be situations where the appropriate penalty is based on an assumption of spatial smoothness, typically of the form $\boldsymbol{\beta}'\mathbf{C}'\mathbf{C}\boldsymbol{\beta}$. In spline smoothing (Wahba and Wendelberger 1980, Wahba 1983 plus many of the references therein) the matrix C is often used to penalize finite second differences of the parameter, which means using a $(p-2) \times p$ matrix C of the form

$$\mathbf{C} = \begin{bmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \dots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & -2 & 1 \end{bmatrix}$$

We can interpret $\mathbf{C}'\mathbf{C}$ as penalizing mean square curvature of the solution, which is an intuitive penalty. However, little work has gone into finding more appropriate choices based on the data. Wahba and Wendelberger (1980) use GCV to estimate both the degree of derivative penalty and the smoothing parameter. A more natural specification of $\mathbf{C}$ arises if we use a Bayesian formulation, for in that case $\mathbf{C}'\mathbf{C}$ corresponds to the *a priori* covariance matrix of $\boldsymbol{\beta}$.

In the next chapter we will outline a procedure for finding a smoothed estimator which will utilize the random structure of $\mathbf{X}$ to specify the form of $\mathbf{C}'\mathbf{C}$.

2.7 Conclusions

We have given an overview of the structure of smoothed estimators in the linear regression model, and discussed some of their properties. We have described several data-driven methods for choosing a smoothing parameter, and how they are either not applicable in the underdetermined model, or require very strong assumptions.

In the next chapter we will describe our method for imposing a smoothness constraint on the linear model with randomness in $\mathbf{X}$ and ϵ. The method will utilize the fact that there is randomness in $\mathbf{X}$ to impose a reasonable smoothness penalty, and construct estimators that will have desirable properties, particularly in relation to predictive ability.

Chapter 3

Model and Estimation

3.1 Introduction

In chapter 2 we outlined several methods used to choose the smoothing parameter in multiple linear regression models, and some of their drawbacks in the underdetermined case. In this chapter we will introduce a method which takes advantage of the random explanatory variable structure, which is very appropriate in models discussed in chapter 1, and introduce a smoothing method based on the signal-to-noise ratio of the model. We will outline parameter estimation with this ratio fixed, and discuss a method to estimate this ratio. The method will be used in the analysis of one of our motivating examples in chapter 5.

3.2 The Model and Parameter Estimation

We assume the data follow the linear model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{3.2.1}$$

where $\mathbf{X}$ is an $n \times p$ design matrix, $\mathbf{y}$ is an $n \times 1$ vector of observations, and

$$\epsilon \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

Since we are interested in models in which the explanatory variables are random, as would be the case in the realization of a continuous process, we will write $\mathbf{X}$ as

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} .$$

with

$$\mathbf{x}_i' \sim \mathrm{N}(\mathbf{0}', \Sigma_X) \ , \ i = 1, \dots, n$$

and assume the $\mathbf{x}_i'$ values are independent of each other and ϵ. This implies that

$$\mathbf{y} \sim \mathrm{N}(\mathbf{0}, (\boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta} + \sigma^2)\mathbf{I})$$

The columns of $\mathbf{X}$ are correlated, with the strength of the correlation depending on the elements of Σ_X. Note that we can write (3.2.1) as

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\Sigma_X^{-1/2}\Sigma_X^{1/2}\boldsymbol{\beta} + \epsilon \\ &= \mathbf{Z}\boldsymbol{\alpha} + \epsilon \end{aligned}$$

where $\mathbf{z}_i' \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$, *i.e.* $\mathbf{Z}$ contains *i.i.d.* random variables.

We are assuming that $n < p$, so that the model is underdetermined. This means we must use an alternative to the classical least squares estimator for $\boldsymbol{\beta}$. We will propose a method of deriving our estimators based on the signal-to-noise ratio.

From our assumptions on (3.2.1) we have

$$\mathrm{Var}(\mathbf{x}_i'\boldsymbol{\beta}) = \mathrm{Var(signal)} = \boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta}$$

$$\mathrm{Var}(\epsilon_i) = \mathrm{Var(noise)} = \sigma^2 \ .$$

We denote the signal-to-noise ratio as

$$\frac{\mathrm{Var(signal)}}{\mathrm{Var(noise)}} = \frac{\boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta}}{\sigma^2} = q \ . \qquad (3.2.2)$$

We assume q is known, which may be given prior knowledge in some dynamic models, as mentioned previously. We will then use (3.2.2) to constrain our estimator of $\boldsymbol{\beta}$.

We will introduce a method for estimating q later in the chapter. We will also assume that σ^2 and Σ_X are known at this point. These assumptions will be relaxed in later sections.

We now present two ways to estimate $\boldsymbol{\beta}$. A first approach is to minimize the sum of squares subject to (3.2.2), which is a quadratic constraint on $\boldsymbol{\beta}$:

$$\min_{\boldsymbol{\beta}} \left[\frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \left(\frac{\boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta}}{\sigma^2} - q \right) \right] = \min_{\boldsymbol{\beta}} SS_p \; . \tag{3.2.3}$$

In the optimization λ acts as a Lagrange multiplier. To find $\hat{\boldsymbol{\beta}}$ we differentiate (3.2.3) with respect to $\boldsymbol{\beta}$,

$$\frac{\partial SS_p}{\partial \boldsymbol{\beta}} = \frac{-2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} + \frac{2\lambda\Sigma_X\boldsymbol{\beta}}{\sigma^2}$$

set this equal to $\mathbf{0}$ and solve for $\boldsymbol{\beta}$, yielding

$$\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}'\mathbf{X} + \lambda\Sigma_X)^{-1}\mathbf{X}'\mathbf{y} \tag{3.2.4}$$

The value for λ must be chosen to satisfy the constraint (3.2.2).

We can also derive $\hat{\boldsymbol{\beta}}_\lambda$ from a penalized likelihood approach. We begin with the joint density of $\mathbf{y}$ and $\mathbf{x}_1', \ldots, \mathbf{x}_n'$, which can be written as the product of a marginal and a conditional density:

$$f(\mathbf{y}, \mathbf{x}_1', \ldots, \mathbf{x}_n') = g(\mathbf{y}|\mathbf{x}_1', \ldots, \mathbf{x}_n')h(\mathbf{x}_1', \ldots, \mathbf{x}_n') \; .$$

Now, $(\mathbf{y}|\mathbf{x}_1', \ldots, \mathbf{x}_n') \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and $\mathbf{x}_i' \sim N(\mathbf{0}', \Sigma_X)$ so we can write

$$\begin{aligned} f(\mathbf{y}, \mathbf{x}_1', \ldots, \mathbf{x}_n') \;\propto\; & (\sigma^2)^{-n/2}|\Sigma_X|^{-n/2} \\ & \times \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\sum_{i=1}^{n} \mathbf{x}_i'\Sigma_X\mathbf{x}_i \right] \end{aligned}$$

We use the joint density to construct the log-likelihood

$$l = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log|\Sigma_X| - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\sum_{i=1}^{n} \mathbf{x}_i'\Sigma_X\mathbf{x}_i \; . \tag{3.2.5}$$

We maximize (3.2.5) with respect to $\boldsymbol{\beta}$ and σ^2 subject to the constraint (3.2.2) by the use of a Lagrange multiplier. We should note that, from (3.2.5), the constraint forces the estimate of $\boldsymbol{\beta}$ to depend on the distribution of $\mathbf{X}$. In the notation of chapter 2, $\Sigma_X = \mathbf{C}'\mathbf{C}$, so this approach also makes the choice of the smoothing matrix straightforward, and very appropriate, given a fixed signal-to-noise ratio.

To find $\hat{\boldsymbol{\beta}}$ we differentiate

$$\begin{aligned} l_p = & -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log|\Sigma_X| \\ & -\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\sum_{i=1}^{n}\mathbf{x}_i'\Sigma_X\mathbf{x}_i - \frac{\lambda}{2}\left(\frac{\boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta}}{\sigma^2} - q\right) \end{aligned} \qquad (3.2.6)$$

with respect to $\boldsymbol{\beta}$, set the result equal to $\mathbf{0}$, and solve for $\boldsymbol{\beta}$. This gives us (3.2.4).

The next step is to replace $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}_\lambda$ in (3.2.2),

$$\frac{\hat{\boldsymbol{\beta}}_\lambda'\Sigma_X\hat{\boldsymbol{\beta}}_\lambda}{\sigma^2} = q \qquad (3.2.7)$$

and find λ which satisfies (3.2.7). We denote this value $\hat{\lambda}_t$.

So our procedure (assuming σ^2 is known) is the following:

1. Solve (3.2.7) to find $\hat{\lambda}_t$.

2. Use $\hat{\lambda}_t$ to find $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$, where

$$\hat{\boldsymbol{\beta}}_{\hat{\lambda}} = (\mathbf{X}'\mathbf{X} + \hat{\lambda}_t\Sigma_X)^{-1}\mathbf{X}'\mathbf{y} \qquad (3.2.8)$$

In Appendix A we establish the conditions under which $\hat{\lambda}_t$ is unique. By writing the SVD of $\mathbf{X}\Sigma^{-1/2} = \mathbf{U}\mathbf{D}\mathbf{V}'$, we show that

$$\sum_{i=1}^{n}\left(\frac{w_i}{d_i}\right)^2 > q\sigma^2$$

is required to ensure a solution exists, where $\mathbf{w} = \mathbf{U}'\mathbf{y}$ and d_i is the *ith* diagonal element of $\mathbf{D}$.

It is natural to compare this model with errors-in-variables models, discussed by Fuller (1987). We will briefly describe these models in the simple and multiple regression cases, and see they share some features with our model.

In simple linear regression, we can set up the errors-in-variables model as

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \ , \quad \epsilon_i \sim \mathrm{N}(0, \sigma_\epsilon^2)$$

$$X_i = x_i + \delta_i \ , \quad \delta_i \sim \mathrm{N}(0, \sigma_\delta^2)$$

where ϵ_i and δ_i are independent, and we observe (X_i, Y_i), $i = 1, \ldots, n$.

If x_i is treated as fixed, the likelihood function $L(\beta_0, \beta_1, \mathbf{x}, \sigma_\epsilon^2, \sigma_\delta^2)$ does not have a finite maximum (Casella and Berger 1990, pg. 587). A similar problem occurs if x_i is treated as random; the system of equations we need to solve to find the MLE's is indeterminate (Casella and Berger 1990, pg. 591). To alleviate these problems, it is assumed that $\sigma_\delta^2 = \gamma\sigma_\epsilon^2$, where $\gamma > 0$ is fixed and known. We are making a similar assumption, but our main reason is to place a smoothness constraint on our estimator.

Now we will focus on the multiple regression case, using a formulation that is similar to the simple linear regression case. We will then compare the resulting estimators to those found by our method.

We assume (Fuller 1987, pg. 124) that the model has the form

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta} + e_i$$

$$\mathbf{X}_i = \mathbf{x}_i + \mathbf{u}_i$$

where we observe the vectors $\mathbf{Z}_i' = (Y_i, \mathbf{X}_i')$, $i = 1, \ldots, n$. We also assume that

$$\boldsymbol{\epsilon}_i = \begin{pmatrix} e_i \\ \mathbf{u}_i \end{pmatrix} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Upsilon}_{\epsilon\epsilon}), \quad \boldsymbol{\Upsilon}_{\epsilon\epsilon} = \begin{bmatrix} \sigma_{ee} & \boldsymbol{v}_{eu} \\ \boldsymbol{v}_{ue} & \boldsymbol{\Upsilon}_{uu} \end{bmatrix}, \quad \boldsymbol{\Upsilon}_{\epsilon\epsilon} \text{ is known.}$$

Fuller (1987, pg. 124) shows that

$$\hat{\boldsymbol{\beta}}_e = (\mathbf{X}'\mathbf{X} - \hat{\gamma}\boldsymbol{\Upsilon}_{uu})^{-1}(\mathbf{X}'\mathbf{y} - \hat{\gamma}\boldsymbol{v}_{ue}) \tag{3.2.9}$$

where $\hat{\gamma}$ is the smallest root of $|\mathbf{Z}'\mathbf{Z} - \gamma\boldsymbol{\Upsilon}_{\epsilon\epsilon}| = 0$.

The estimator (3.2.9) has a similar form to (3.2.8), but there is a key difference. The required $\hat{\gamma}$ will be 0 if $n < p$. Then we will be left with finding the inverse of $\mathbf{X}'\mathbf{X}$. Since this is a singular matrix, the method breaks down in an underdetermined system.

3.2.1 Estimation of σ^2

In our derivation of the estimators in the previous section we assumed that σ^2 was known. This assumption is probably not realistic, unless we had replicated observations, or information from a previous experiment. Therefore we need a method to estimate σ^2.

We propose two estimators of σ^2. One will be derived from the log-likelihood (3.2.5), while the second will not depend on the model used.

Our first estimator of σ^2, denoted by $\hat{\sigma}^2_{lik(\lambda)}$, will be based on the log-likelihood. We differentiate (3.2.6) with respect to σ^2:

$$\frac{\partial l_p}{\partial \sigma^2} = -n\sigma^2 + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta}$$

We must now solve the system of equations $\partial l_p/\partial\sigma^2 = 0$ and $\partial l_p/\partial\boldsymbol{\beta} = \mathbf{0}$. This gives us (3.2.4) and

$$\hat{\sigma}^2_{lik(\lambda)} = \frac{1}{n}\left[(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda) + \lambda\hat{\boldsymbol{\beta}}_\lambda'\Sigma_X\hat{\boldsymbol{\beta}}_\lambda\right] \tag{3.2.10}$$

Let us consider the bias of this estimate. We cannot find $\mathrm{E}(\hat{\sigma}^2_{lik(\lambda)})$ easily because it depends on two random quantities, $\mathbf{X}$ and $\mathbf{y}$. If we condition on $\mathbf{X}$, and let $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\Sigma_X)^{-1}\mathbf{X}'$, we find

$$\begin{aligned}\mathrm{E}_{y|X}(\hat{\sigma}^2_{lik(\lambda)}) &= \frac{1}{n}\left[(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}-\mathbf{H})^2\mathbf{X}\boldsymbol{\beta} + \sigma^2\mathrm{tr}(\mathbf{I}-\mathbf{H})^2\right] \\ &+\frac{\lambda}{n}\left[(\mathbf{X}\boldsymbol{\beta})'\mathbf{X}(\mathbf{X}'\mathbf{X}+\lambda\Sigma_X)^{-1}\Sigma_X(\mathbf{X}'\mathbf{X}+\lambda\Sigma_X)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right] \\ &+\frac{\lambda}{n}\left[\sigma^2\mathrm{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X}+\lambda\Sigma_X)^{-1}\Sigma_X(\mathbf{X}'\mathbf{X}+\lambda\Sigma_X)^{-1}\mathbf{X}']\right].\end{aligned}$$

This shows the bias of $\hat{\sigma}^2_{lik(\lambda)}$ is very complicated, and is difficult to calculate beyond this expression. Some numerical results on the bias are included in the simulation results given in the next section.

Our second estimator of σ^2 can be considered a method of moments estimator, since it will equate sample and population moments. It relies on the fact that $\mathrm{E}(y_i)$ is the same for each y_i. From our model assumptions

$$\mathrm{E}(y_i) = 0\ ,\ \mathrm{Var}(y_i) = \boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta} + \sigma^2$$

Since the y_i values have a common mean we can estimate $\mathrm{Var}(y_i)$ by the sample variance s_y^2. From our signal-to-noise ratio we know

$$\boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta} = q\sigma^2$$

$$\boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta} + \sigma^2 = \sigma^2(q+1)$$

So we can say $\hat{\sigma}^2(q+1) = s_y^2$, hence

$$\hat{\sigma}_a^2 = \frac{1}{(q+1)(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{s_y^2}{q+1} \qquad (3.2.11)$$

This estimate is independent of the model chosen, and is unaffected by the fact that $n < p$. We can easily find the expectation of σ_a^2:

$$\begin{aligned} \mathrm{E}(\hat{\sigma}_a^2) &= \frac{1}{q+1}\mathrm{E}(s_y^2) = \frac{1}{q+1}\mathrm{Var}(y_i) \\ &= \frac{1}{q+1}(\boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta} + \sigma^2) \\ &= \frac{\sigma^2(q+1)}{q+1} = \sigma^2, \end{aligned}$$

so σ_a^2 is unbiased. This should come as no surprise since, from the regression literature (Neter, Wasserman, and Kutner 1985, pg. 127), if we have repeated $\mathbf{x}$ observations we can construct $\hat{\sigma}^2$ based on the pure error sum of squares and obtain an unbiased estimator, regardless of the model. The key is that the y values corresponding to the repeated $\mathbf{x}$ values have the same mean.

Now that we are estimating σ^2, our procedure for estimating $\boldsymbol{\beta}$ is modified:

1. Find λ to solve either

$$\frac{\hat{\boldsymbol{\beta}}_\lambda'\Sigma_X\hat{\boldsymbol{\beta}}_\lambda}{\hat{\sigma}_a^2} = q$$

   or

$$\frac{\hat{\boldsymbol{\beta}}_\lambda'\Sigma_X\hat{\boldsymbol{\beta}}_\lambda}{\hat{\sigma}_{lik(\lambda)}^2} = q. \qquad (3.2.12)$$

   Call the λ choice $\hat{\lambda}_a$ or $\hat{\lambda}_{lik}$.

2. Use $\hat{\lambda}_a$ or $\hat{\lambda}_{lik}$ to find $\boldsymbol{\beta}_{\hat{\lambda}}$ in (3.2.8).

3. If using $\hat{\lambda}_{lik}$, use $\hat{\lambda}_{lik}$ and $\hat{\beta}_{\hat{\lambda}}$ to find $\hat{\sigma}^2_{lik(\hat{\lambda})}$, where

$$\hat{\sigma}^2_{lik(\hat{\lambda})} = \frac{1}{n}\left[(\mathbf{y} - \mathbf{X}\hat{\beta}_{\hat{\lambda}})'(\mathbf{y} - \mathbf{X}\hat{\beta}_{\hat{\lambda}}) + \hat{\lambda}_{lik}\hat{\beta}'_{\hat{\lambda}}\Sigma_X\hat{\beta}_{\hat{\lambda}}\right] \qquad (3.2.13)$$

In Appendix A we derive the conditions under which $\hat{\lambda}_a$ and $\hat{\lambda}_{lik}$ exist. The key result is that $\hat{\lambda}_{lik}$ always exists, and is unique, given q.

3.3 Evaluation of Estimators

We now wish to examine some of the properties of the estimates of σ^2 and λ values introduced in the previous section. We will also compare the estimators with those obtained using the GCV, E(GCV) and E(PMSE) criteria, which were defined in chapter 2. We will do this using simulation studies under a variety of situations.

In the simulations we must ensure the data satisfies (3.2.2). To ensure this we do the following. We choose β and Σ_X, so that we can evaluate $\beta'\Sigma_X\beta$. Once Σ_X is chosen, we generate the rows of $\mathbf{X}$ from a $N(\mathbf{0}', \Sigma_X)$ distribution. We then choose σ^2, and generate ϵ from a $N(\mathbf{0}, \sigma^2\mathbf{I})$ distribution. Now that we have $\beta'\Sigma_X\beta$ and σ^2 we know q from (3.2.2). Once we have $\mathbf{X}$, β and ϵ we generate our $\mathbf{y}$ vector according to (3.2.1).

The results use different values of q, and different true β vectors. The elements of β were generated from either a U(0, 1) distribution or a N(0, 4) distribution. In all cases we used $n = 50$ and $p = 80$. We used the same $\mathbf{X}$ matrix for each simulated data set, even though it would be more appropriate to use a different matrix each time since $\mathbf{X}$ is random. However, the results do not change a great deal if a different $\mathbf{X}$ matrix is used each time. Hence the randomness in the simulation comes from the error term ϵ.

The results we present in the chapter will use a model where the rows of $\mathbf{X}$ are generated from the AR(2) process

$$x_{ij} = 0.4x_{i,j-1} + 0.2x_{i,j-2} + \eta_{ij}\ , \quad j = 1, \ldots, p \qquad (3.3.1)$$

with $\eta_{ij} \sim \mathrm{N}(0, 1)$. We are focusing on generating the rows of $\mathbf{X}$ from AR processes since this is often an appropriate assumption to make about the types of data sets in which we are interested. It will also make it easier for us to discuss estimation of Σ_X later in the chapter. Results for other choices of $\mathbf{X}$, n and p are given in Appendix B.

We will first examine the estimates of σ^2, then discuss the λ values chosen, and compare these values to those given by other methods.

3.3.1 Comparison of σ^2 Estimates

Table 3.1 summarizes the $\hat{\sigma}^2$ values found by each of the two methods described in the previous section. The estimates were calculated in 1000 simulated data sets. The results in the first column used $\boldsymbol{\beta}$ which was generated from a N(0, 4) distribution, while the other three columns used $\boldsymbol{\beta}$ which was generated from a U(0, 1) distribution.

		q			
		17.37	7.21	1.31	0.094
σ^2		25	16	88.36	1225
$\hat{\sigma}^2_{lik(\hat{\lambda})}$	mean	14.45	10.72	57.15	930.6
	median	14.35	10.63	56.40	919.2
	var	1.846	1.284	87.41	3.51×10^4
$\hat{\sigma}^2_a$	mean	22.91	19.81	101.7	1253
	median	22.88	19.71	100.3	1237
	var	2.419	3.077	263.43	6.48×10^4

Table 3.1: Summary of estimates of σ^2 when rows of $\mathbf{X}$ are generated from an AR(2) process. Models use $n = 50$, $p = 80$

We see that, for the four models presented, $\hat{\sigma}^2_a$ has smaller bias than $\hat{\sigma}^2_{lik(\lambda)}$, as we would expect. For each value of q used the $\hat{\sigma}^2_{lik(\hat{\lambda})}$ values were biased downwards. This bias was, on average, approximately $\hat{\sigma}^2_{lik(\hat{\lambda})} = 0.6\sigma^2$. The bias is expected because in a linear model with $n > p$ the maximum likelihood estimator of the error variance, SSE/n, is biased downwards. We also observed that the $\hat{\sigma}^2_{lik(\hat{\lambda})}$ values had smaller variability across the simulations. Figures 3.1 and 3.2 show the distribution of $\hat{\sigma}_a^2$ and $\hat{\sigma}^2_{lik(\hat{\lambda})}$ in the $q = 7.21$ and $q = 1.31$ models. These plots show the distributions are skewed to the right, but the $\hat{\sigma}^2_{lik(\hat{\lambda})}$ values have smaller variability.

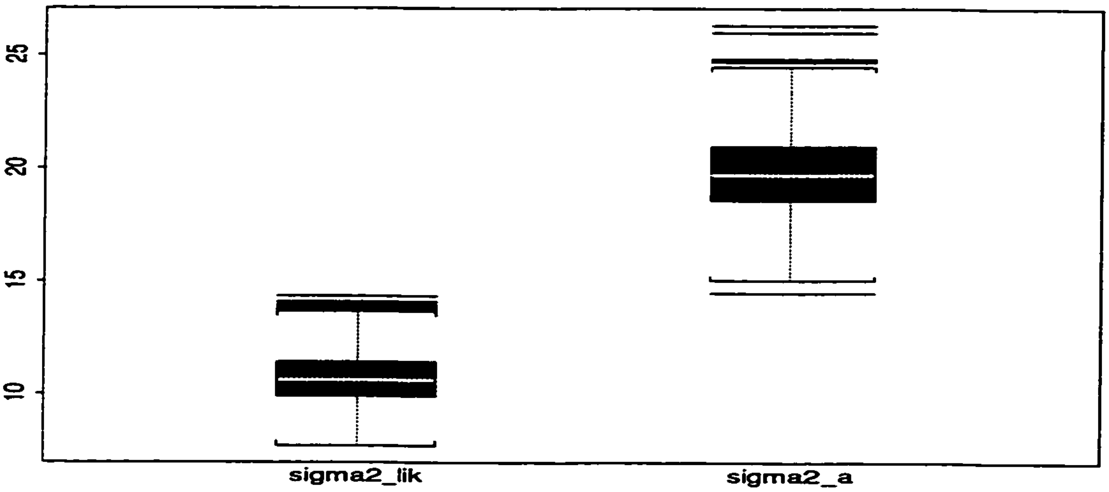


Figure 3.1: Boxplots of $\hat{\sigma}^2_{lik}$ and $\hat{\sigma}_a^2$ estimates when $q = 7.21$, $\sigma^2 = 16$

3.3.2 Comparison of λ Values

We now wish to compare the $\hat{\lambda}$ values found using the three approaches, with the $\hat{\lambda}_{GCV}$, $\hat{\lambda}_{EGCV}$ and $\hat{\lambda}_{EPMSE}$ values found for the same simulated data sets. Note that these other three methods assume that $\mathbf{X}$ is nonstochastic, while $\hat{\lambda}_{EGCV}$ and $\hat{\lambda}_{EPMSE}$ depend on the true $\boldsymbol{\beta}$, so can only be found in simulation studies.

Table 3.2 summarizes the $\hat{\lambda}$ values found using these methods. Since we used the same $\mathbf{X}$ matrix for all 1000 data sets, we only have one value of $\hat{\lambda}_{EPMSE}$ and $\hat{\lambda}_{EGCV}$.

We first identify some general trends. We see, except for $\hat{\lambda}_{GCV}$ and $\hat{\lambda}_{EGCV}$, the $\hat{\lambda}$ values decrease as q increases. For those chosen by the signal-to-noise criterion, this can be justified in two ways. First, it makes intuitive sense that if less of the variability is due to the noise term, there is less need to smooth. Second, we can use

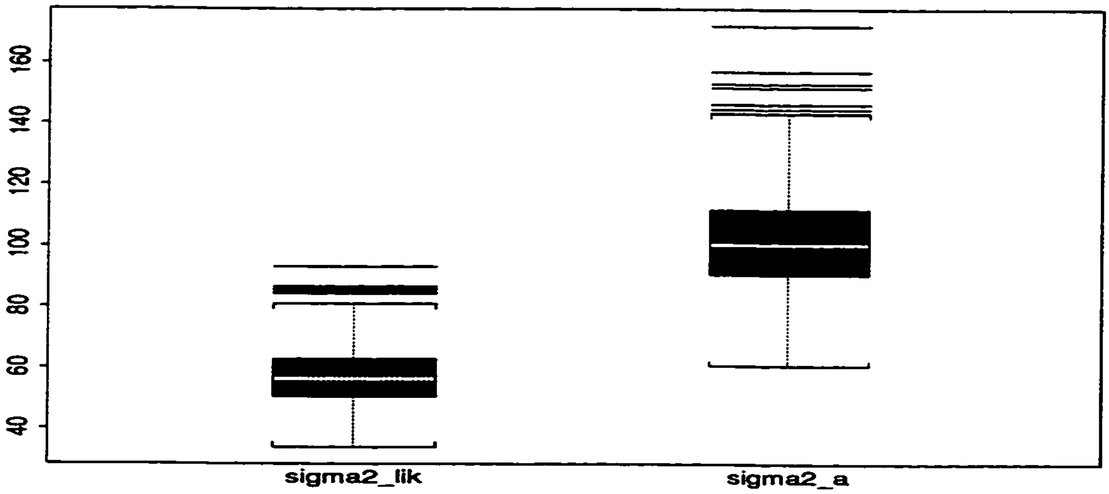


Figure 3.2: Boxplots of $\hat{\sigma}^2_{lik}$ and $\hat{\sigma}^2_a$ estimates when $q = 1.31$, $\sigma^2 = 88.36$

implicit differentiation to show that $\partial\lambda/\partial q < 0$, indicating λ is a decreasing function of q.

An important point must be made about the $\hat{\lambda}_t$ and $\hat{\lambda}_a$ values. Recall $\hat{\lambda}_t$ is found assuming σ^2 is known, while $\hat{\lambda}_a$ is found using the unbiased estimator of σ^2. We see in table 3.2 that their median values are 0 at the $q = 17.37$ and $q = 7.21$ models. This is because the constraint equation is not satisfied in most data sets, so the λ value is being set to 0. We also see that $\hat{\lambda}_{lik}$ continues to give larger values than $\hat{\lambda}_t$ or $\hat{\lambda}_a$.

We see that $\hat{\lambda}_{EPMSE}$, $\hat{\lambda}_{EGCV}$ and $\hat{\lambda}_{GCV}$ are larger than the $\hat{\lambda}_{lik}$ values for small values of q. We stated in section 3.2 that $\hat{\lambda}_{lik}$ is the unique solution to (3.2.12) but we have no guarantee that the GCV function will have a unique minimum. The other important result, which we will discuss in more detail, is the $\hat{\lambda}_{lik}$ values are much less variable than the $\hat{\lambda}_{GCV}$ values.

We begin with $q = 17.37$ in table 3.2. We see that the mean of the $\hat{\lambda}_{lik}$ and $\hat{\lambda}_{GCV}$ values are larger than $\hat{\lambda}_{EPMSE}$. We see the variance of the $\hat{\lambda}_{lik}$ values is less than the variance of the $\hat{\lambda}_{GCV}$ values by a factor of 10^3. The same difference appears

		q			
		17.37	7.21	1.31	0.094
λ_{EPMSE}		4.657	10.914	51.689	643.324
$\hat{\lambda}_{EGCV}$		7.385	4.854	45.256	581.990
$\hat{\lambda}_{lik}$	mean	2.637	5.957	23.949	144.820
	median	2.641	5.973	24.003	144.428
	var	1.44×10^{-3}	0.0191	1.759	139.906
$\hat{\lambda}_a$	mean	3.77×10^{-4}	4.06×10^{-3}	6.95	112.7
	median	0	0	6.973	112.4
	var	7.81×10^{-5}	1.90×10^{-3}	4.373	47.67
$\hat{\lambda}_t$	mean	2.81×10^{-5}	0.301	6.95	112.7
	median	0	0	9.619	113.2
	var	2.81×10^{-5}	0.398	15.98	448.8
$\hat{\lambda}_{GCV}$	mean	7.078	5.379	62.612	8.247×10^{9}
	median	5.063	1.696	33.971	470.350
	var	62.00	56.761	2.50×10^{4}	3.74×10^{20}

Table 3.2: Comparison of $\hat{\lambda}$ values when rows of $\mathbf{X}$ generated from an AR(2) process. Models use $n = 50$, $p = 80$

when $q = 7.21$. When we use a model with $q = 1.31$ we find the mean of the $\hat{\lambda}_{GCV}$ values exceeds $\hat{\lambda}_{EPMSE}$. For this value of q the $\hat{\lambda}_{lik}$ values are less variable by a factor of about 10^4. For the 1000 simulated sets is this case, the $\hat{\lambda}_{GCV}$ values are within $[0, 1.58 \times 10^4]$, while the $\hat{\lambda}_{lik}$ values never exceed 27.62. Finally, when $q = 0.094$ the mean of the $\hat{\lambda}_{GCV}$ values is very large, as is the variance. Approximately 30% of the $\hat{\lambda}_{GCV}$ values exceed 10^4. These values would clearly be oversmoothing. Once again the $\hat{\lambda}_{lik}$ values have a variability that is several orders of magnitude smaller. If we ignore those values of $\hat{\lambda}_{GCV}$ which exceed 10^4, the mean of the remaining values is 575.46, and the median is 107.24. In this model we also get cases of clear undersmoothing with GCV, as it chooses $\hat{\lambda}_{GCV} = 0$ about 11% of the time.

Figures 3.3 and 3.4 display, on the log scale, the distribution of the $\hat{\lambda}$ values in the $q = 7.21$ and $q = 1.31$ models. These plots illustrate the problems with $\hat{\lambda}_t$ and $\hat{\lambda}_a$ at the large values of q, and the high variability of $\hat{\lambda}_{GCV}$ at low values of q.

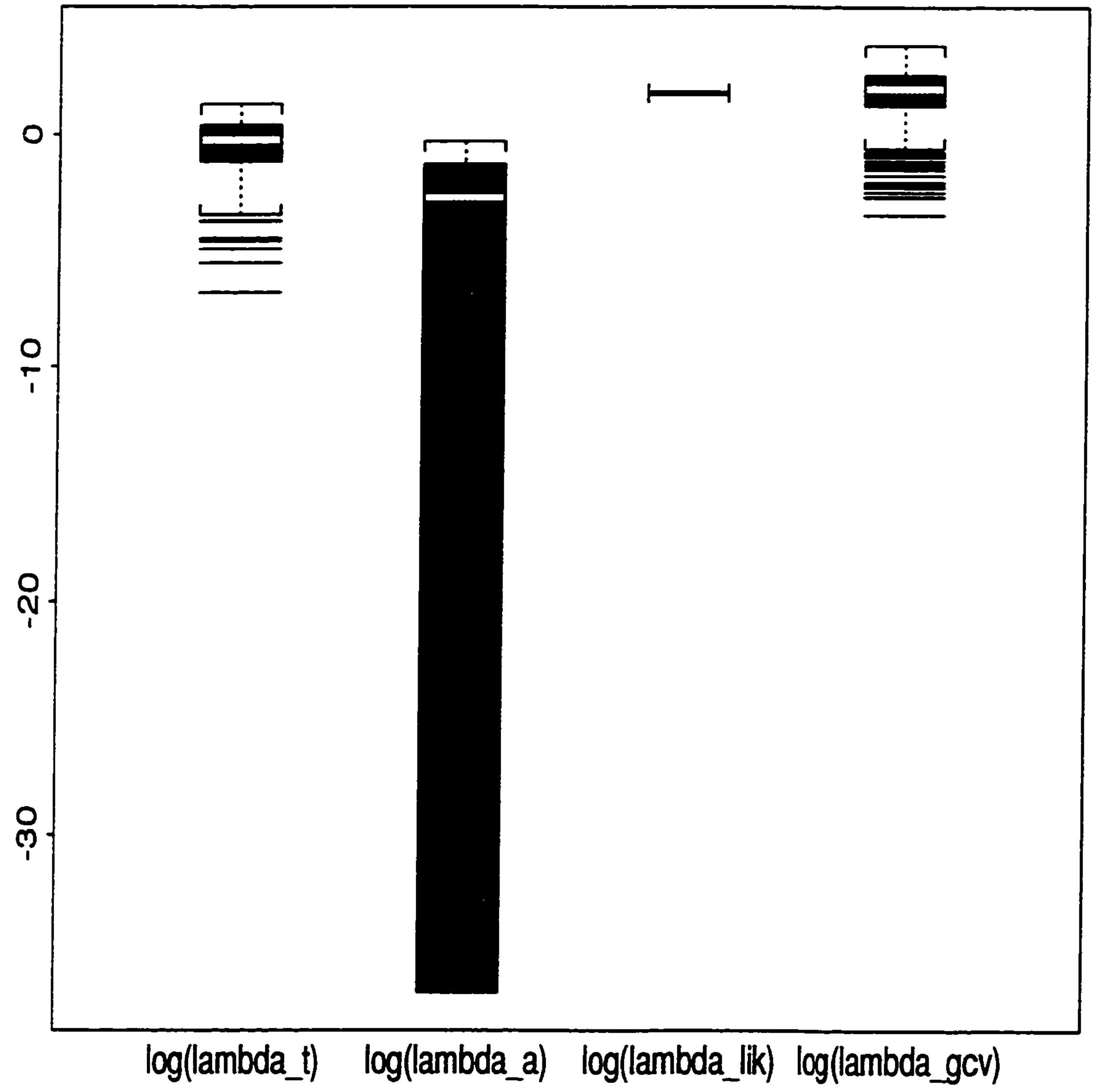


Figure 3.3: Plots of $\log(\hat{\lambda})$ values in the $q = 7.21$ model

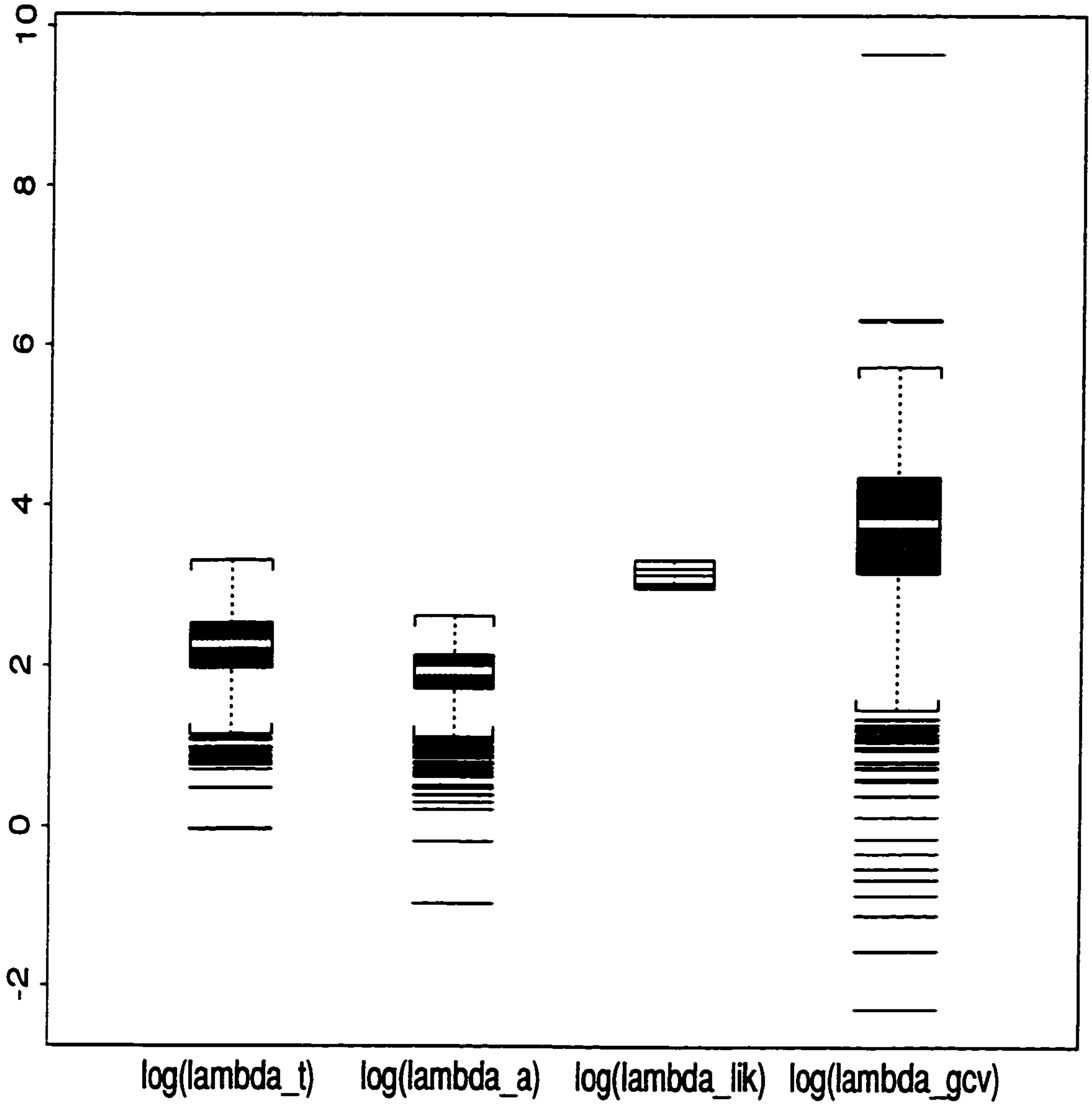


Figure 3.4: Plots of $\log(\hat{\lambda})$ values in the $q = 1.31$ model

Therefore, we can conclude that the signal-to-noise method yields values of $\hat{\lambda}_{lik}$ which are much less variable than the $\hat{\lambda}_{GCV}$ values, with few extreme values found.

We have demonstrated that the signal-to-noise method yields smoothing parameters with a smaller variance than GCV, especially in the case where we must estimate σ^2. We now wish to compare the methods based on their predictive ability. Recall from our previous discussion that the $\hat{\lambda}_{GCV}$ estimates are supposed to be chosen using a prediction criterion. The signal-to-noise estimators do not make this claim.

We will examine the predictive quality of the estimators by conducting the following simulation study:

1. Generate $\mathbf{y}_o = \mathbf{X}\boldsymbol{\beta} + \epsilon$, where $\mathbf{X}$ and ϵ are described at the beginning of this section.

2. Select λ by each of the following methods:

   (a) The solution of $\hat{\boldsymbol{\beta}}_\lambda' \boldsymbol{\Sigma}_X \hat{\boldsymbol{\beta}}_\lambda / \sigma^2 = q$.

   (b) The solution of $\hat{\boldsymbol{\beta}}_\lambda' \boldsymbol{\Sigma}_X \hat{\boldsymbol{\beta}}_\lambda / \hat{\sigma}_a^2 = q$.

   (c) The solution of $\hat{\boldsymbol{\beta}}_\lambda' \boldsymbol{\Sigma}_X \hat{\boldsymbol{\beta}}_\lambda / \hat{\sigma}_{lik}^2 = q$.

   (d) Minimize GCV over λ.

   (e) Minimize E(GCV) over λ.

   (f) Minimize E(PMSE) over λ.

3. Construct $\hat{\boldsymbol{\beta}}_{o,\hat{\lambda}} = (\mathbf{X}'\mathbf{X} + \hat{\lambda}\boldsymbol{\Sigma}_X)^{-1}\mathbf{X}'\mathbf{y}_o$ using each of the $\hat{\lambda}$ values computed in step 2.

4. Generate a validation set $\mathbf{y}_1 = \mathbf{X}\boldsymbol{\beta} + \epsilon_1$ of size n.

5. Evaluate the prediction sum of squares

$$PSS_{\hat{\lambda}} = (\mathbf{y}_1 - \mathbf{X}\hat{\boldsymbol{\beta}}_{o,\hat{\lambda}})'(\mathbf{y}_1 - \mathbf{X}\hat{\boldsymbol{\beta}}_{o,\hat{\lambda}})/n$$

   for each $\hat{\boldsymbol{\beta}}_{o,\hat{\lambda}}$ value from step 2.

6. Repeat steps 1–5 1000 times.

For each method of estimating λ the simulation gives us 1000 values of $SS_{pred.\hat{\lambda}}$. Table 3.3 summarizes the results over the range of q values used in table 3.2.

		q			
		17.37	7.21	1.31	0.094
PSS_{EPMSE}	mean	47.004	28.469	132.329	1346.211
	median	46.043	28.116	130.697	1328.581
	var	88.445	34.239	744.025	7.44×10^4
PSS_{EGCV}	mean	47.497	29.057	132.471	1346.277
	median	46.650	28.629	131.170	1328.556
	var	89.627	35.165	748.533	7.45×10^4
PSS_{lik}	mean	47.383	28.813	136.843	1420.428
	median	46.443	28.421	134.594	1395.035
	var	90.023	34.761	802.267	8.54×10^4
PSS_a	mean	49.854	31.899	153.021	1456.152
	median	49.013	31.368	150.981	1435.391
	var	99.712	40.805	150.981	9.02×10^4
PSS_t	mean	49.855	31.472	147.615	1452.649
	median	49.013	31.027	145.864	1429.880
	var	92.728	39.269	145.864	8.99×10^4
PSS_{GCV}	mean	50.381	30.400	146.893	1552.309
	median	49.428	29.981	144.713	1446.378
	var	113.011	40.865	1186.598	2.37×10^5

Table 3.3: Prediction SS values when rows of $\mathbf{X}$ generated from an AR(2) process. Models use $n = 50$, $p = 80$

We would first like to make some comparisons among the three signal-to-noise estimators, in terms of their predictive ability. We see that $\hat{\lambda}_{lik}$, although using the estimate with the greatest bias in its σ^2 estimate, leads to better prediction than using either $\hat{\lambda}_t$ or $\hat{\lambda}_a$. In addition, we pointed out previously there were many cases, when q was large, that the conditions required to find $\hat{\lambda}_t$ or $\hat{\lambda}_a$ were not satisfied. For these reasons, we will no longer consider these two estimators, but will compare the predictive ability of $\hat{\lambda}_{lik}$ with $\hat{\lambda}_{GCV}$, $\hat{\lambda}_{EGCV}$ and $\hat{\lambda}_{EPMSE}$.

We will begin discussing the model with $q = 17.37$. We see that the PSS_{lik} values are smaller than the PSS_{GCV} values, on average. The PSS_{lik} values also have smaller variability, and an approximate 95% confidence interval for the difference in the PSS values indicates a significant difference. When comparing the 1000 generated sets, we found PSS_{lik} was smaller 75% of the time.

Next we consider the case with $q = 7.21$. The PSS_{lik} values are again smaller, by at least two standard errors, than the PSS_{GCV} values. We also see the variability of the PSS_{lik} is similar to that of PSS_{EGCV} and PSS_{EPMSE}. We found that PSS_{lik} was less than PSS_{GCV} in 77.2% of the cases.

Results moved even more in favour of the signal-to-noise method as q decreased to 1.31. If we recall, GCV begins to give some very large smoothing parameter estimates at this point, but also can do no smoothing. There is a greater relative difference between PSS_{lik} and PSS_{GCV}, with the difference easily exceeding two standard errors. When comparing the 1000 generated sets, we found PSS_{lik} was smaller 61% of the time.

The model using $q = 0.094$ yields the greatest difference between the two methods, as shown in table 3.3. There is a very large difference in the average PSS values, and we see that the PSS_{GCV} values are much more variable in this situation. We also notice that both data-based methods give larger PSS values than PSS_{EGCV} and PSS_{EPMSE}. On an individual basis, we found that PSS_{lik} was less than PSS_{GCV} in 48.5% of the cases. Therefore, as q has decreased, the signal-to-noise method appears to improve on average, but not in terms of the number of individual cases where it outperforms GCV. This appears to be a reflection of the fact that there are often cases where PSS_{lik} is much smaller than PSS_{GCV}, while the reverse scenario is rarely true.

To illustrate this point, figures 3.5 and 3.6 contain boxplots of the values

$$SS_D = PSS_{ratio} - PSS_{MSE}$$

where *ratio* denotes we are using either $\hat{\lambda}_t$, $\hat{\lambda}_a$ or $\hat{\lambda}_{lik}$, while MSE denotes we are using either $\hat{\lambda}_{GCV}$, $\hat{\lambda}_{EGCV}$, or $\hat{\lambda}_{EPMSE}$. If $SS_D < 0$ then the signal-to-noise ratio method has better predictive ability. Figure 3.5 uses the model with $q = 17.37$ while figure 3.6 uses $q = 0.094$.

We see from figure 3.5 that the median of the SS_D values is positive when we use PSS_{EPMSE} and PSS_{EGCV}, but is negative when we use PSS_{GCV}. This corresponds to what we see in table 3.3. But we are primarily interested in comparing the two data-based estimators. Figure 3.6 also highlights the skewness in the distribution of differences, showing us the signal-to-noise method can often do much better, and rarely does much worse, than GCV.

We now have two reasons to recommend choosing a smoothing parameter based on the signal-to-noise ratio. First, the $\hat{\lambda}_{lik}$ values have very low variability. Second, they perform well based on predictive ability, even though the method is not designed specifically to be optimal in this sense.

Although we have primarily discussed predictive ability up to this point, there may be situations where we are concerned with the $\boldsymbol{\beta}$ vector. It may have certain local or global characteristics in mapping $\mathbf{X}$ to $\mathbf{y}$ that we wish our estimates to capture. We investigate two examples of this.

The first will use a true $\boldsymbol{\beta}$ that we will refer to as "spike": it has $\beta_i = 2$ for one element in $\boldsymbol{\beta}$ and $\beta_i = 0$ for the remaining $p - 1$ elements. The second case will take $\boldsymbol{\beta}$ as "flat": $\beta_i = 0.5$ for all i. The values of q and σ^2 used are given in table 3.4. The models used $n = 50$ and $p = 80$, while the AR(2) process (3.3.1) was used to generate the rows of the $\mathbf{X}$ matrix. Table 3.4 summarizes the results over 1000 simulated data sets. For simplicity we only include the results using GCV and the signal-to-noise ratio. Since we are now interested in the behaviour of the $\hat{\boldsymbol{\beta}}$ values, we also include in the table the values of

$$SS(\hat{\boldsymbol{\beta}}_\lambda) = (\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})/p$$

averaged across the 1000 cases.

We see in all four cases that the mean value of $SS(\hat{\boldsymbol{\beta}}_{lik})$ is less than that of $SS(\hat{\boldsymbol{\beta}}_{GCV})$. The improvement ranges from about 5% to 18%. However, this is not the case for the median values when $\boldsymbol{\beta}$ is the "spike" vector. When we use the "flat"

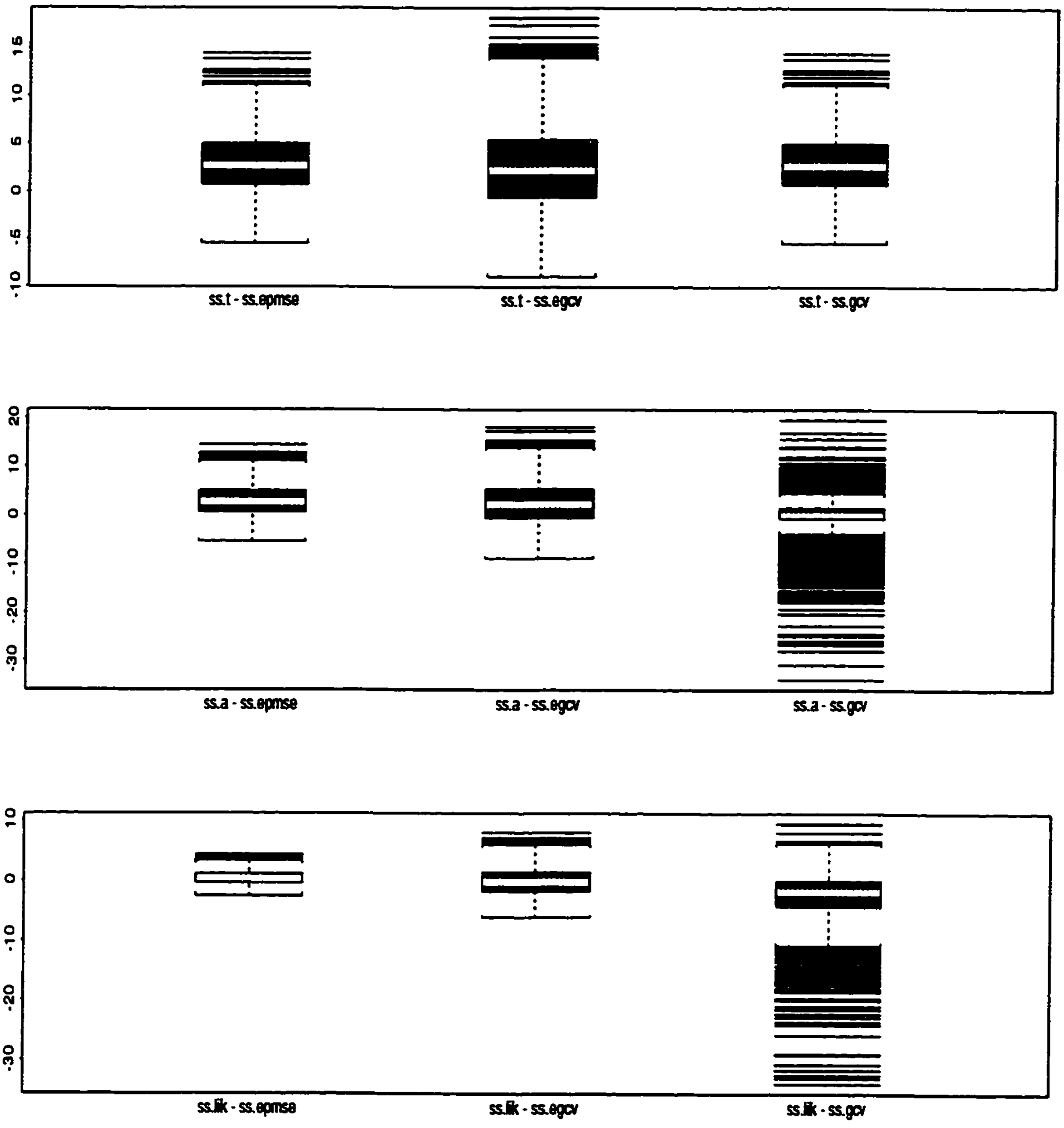


Figure 3.5: PSS differences in the $q = 17.37$ model, where the rows of $\mathbf{X}$ are generated from an AR(2) process

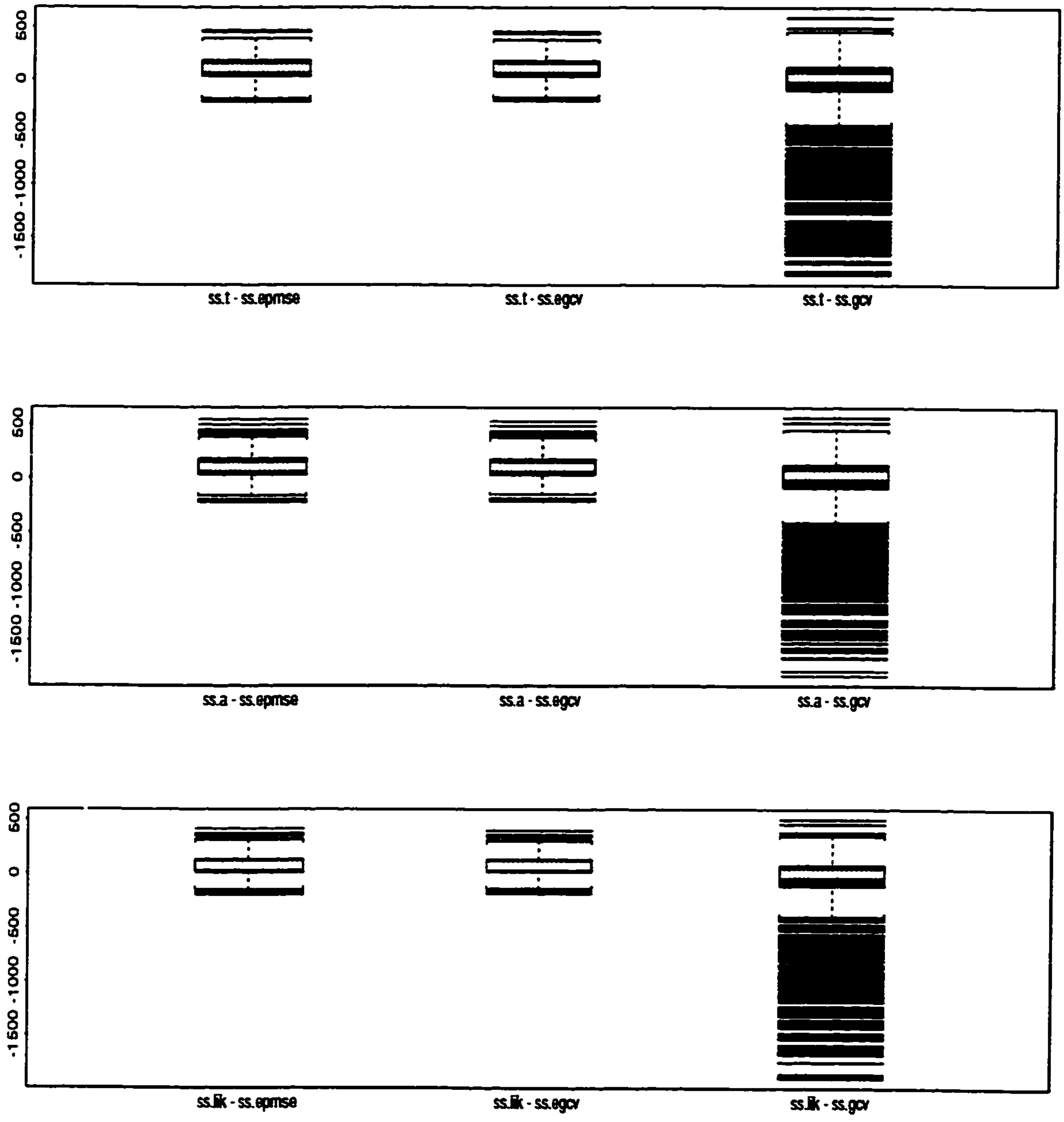


Figure 3.6: PSS differences in the $q = 0.094$ model, where the rows of $\mathbf{X}$ are generated from an AR(2) process

		$\boldsymbol{\beta}$ = "flat"		$\boldsymbol{\beta}$ = "spike"	
		$q = 1.49$	$q = 7.56$	$q = 1.39$	$q = 5.56$
		$\sigma^2 = 81$	$\sigma^2 = 16$	$\sigma^2 = 4$	$\sigma^2 = 1$
$\hat{\lambda}_{lik}$	mean	21.476	5.677	22.707	7.397
	median	21.537	5.692	22.778	7.403
	var	1.345	0.0180	1.578	0.049
$\hat{\lambda}_{GCV}$	mean	54.200	6.069	2.23×10^6	11.269
	median	32.057	2.288	38.872	8.986
	var	4.99×10^4	71.865	2.50×10^{15}	146.69
$SS(\hat{\boldsymbol{\beta}}_{lik})$	mean	0.798	0.582	0.052	0.038
	median	0.785	0.576	0.050	0.037
	var	1.61×10^{-2}	4.69×10^{-3}	3.93×10^{-5}	1.38×10^{-5}
$SS(\hat{\boldsymbol{\beta}}_{GCV})$	mean	0.971	0.645	0.060	0.040
	median	0.641	0.639	0.045	0.037
	var	0.555	0.032	1.00×10^{-3}	6.17×10^{-5}
PSS_{lik}	mean	126.60	28.739	6.171	1.755
	median	124.56	28.239	6.133	1.725
	var	727.74	36.240	1.681	0.132
PSS_{GCV}	mean	136.29	30.306	6.587	1.842
	median	135.45	29.816	6.440	1.818
	var	1014.0	41.047	2.556	0.160

Table 3.4: Results for structured $\boldsymbol{\beta}$ vector

$\boldsymbol{\beta}$ we find $SS(\hat{\boldsymbol{\beta}}_{lik}) < SS(\hat{\boldsymbol{\beta}}_{GCV})$ in 40.2% of cases when $q = 1.49$ and in 62.6% of the cases when $q = 7.56$. When the true $\boldsymbol{\beta}$ vector is the "spike" vector, we find $SS(\hat{\boldsymbol{\beta}}_{lik}) < SS(\hat{\boldsymbol{\beta}}_{GCV})$ in 36.2% of cases when $q = 1.39$ and in 45.8% of cases when $q = 5.56$. We do find the PSS values favour the signal-to-noise ratio in all situations on average and in the majority of individual cases. We also note that GCV still tends to do a considerable amount of undersmoothing in many cases. For example, we found $\hat{\lambda}_{GCV} = 0$ in 30% of the data sets in the "flat" $\boldsymbol{\beta}$ case with $q = 1.49$.

Figure 3.7 contains plots of the true $\boldsymbol{\beta}$ vector, along with $\hat{\boldsymbol{\beta}}_{lik}$ averaged across the 1000 simulations. We see that, for the "flat" case, the β_i values are highly variable, so $\hat{\boldsymbol{\beta}}_{lik}$ is doing a poor job of estimating the true $\boldsymbol{\beta}$. In the "spike" model we see the mean $\hat{\boldsymbol{\beta}}_{lik}$ vector captures the behavior of the true $\boldsymbol{\beta}$ well. We find a peak at the appropriate element of $\hat{\boldsymbol{\beta}}_{lik}$ and the rest of the elements stay close to zero.

These results indicate the method is not performing well with one situation of a highly structured $\boldsymbol{\beta}$ vector. This gives us the opportunity to revisit the notion of smoothing. as a spatial operation. and what the method based on the signal-to-noise ratio is actually doing.

The signal-to-noise ratio method clearly yields a shrinkage estimator. However. there is no guarantee that the estimator is imposing any spatial smoothing. In particular. we do not have any guarantee that smoothness in the rows of $\mathbf{X}$. as described by $\boldsymbol{\Sigma}_X$. will lead to a similar smoothness pattern in $\boldsymbol{\beta}$. This is exemplified in our last two examples. Therefore. since there may be situations where we have *a priori* knowledge of the spatial structure. we outline a possible way to incorporate this with the signal-to-noise constraint.

We could introduce this spatial smoothing constraint by appending another term to the sum of squares (3.2.3) to penalize spatial smoothness:

$$\frac{1}{\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda_1\left(\frac{\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta}}{\sigma^2} - q\right) + \frac{\lambda_2}{\sigma^2}\boldsymbol{\beta}'\mathbf{C}'\mathbf{C}\boldsymbol{\beta}$$

The choice of $\mathbf{C}$ could impose a derivative penalty. as discussed in section 2.6. with $\mathbf{C}'\mathbf{C}$ being a positive semi-definite matrix in this case.

The estimator for $\boldsymbol{\beta}$ would now be

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda_1\boldsymbol{\Sigma}_X + \lambda_2\mathbf{C}'\mathbf{C})^{-1}\mathbf{X}'\mathbf{y}$$

The value for λ_2 would have to be supplied. The larger it is chosen. the greater the amount of spatial smoothness we are imposing on the estimator. We also note it is possible to choose $\boldsymbol{\beta}'\mathbf{C}'\mathbf{C}\boldsymbol{\beta} = 0$ if $\mathbf{C}'\mathbf{C}$ is positive semi-definite. In the final two examples we discussed in this section. we would expect an estimator that incorporated a spatial smoothing term to outperform the method based on the signal-to-noise ratio exclusively.

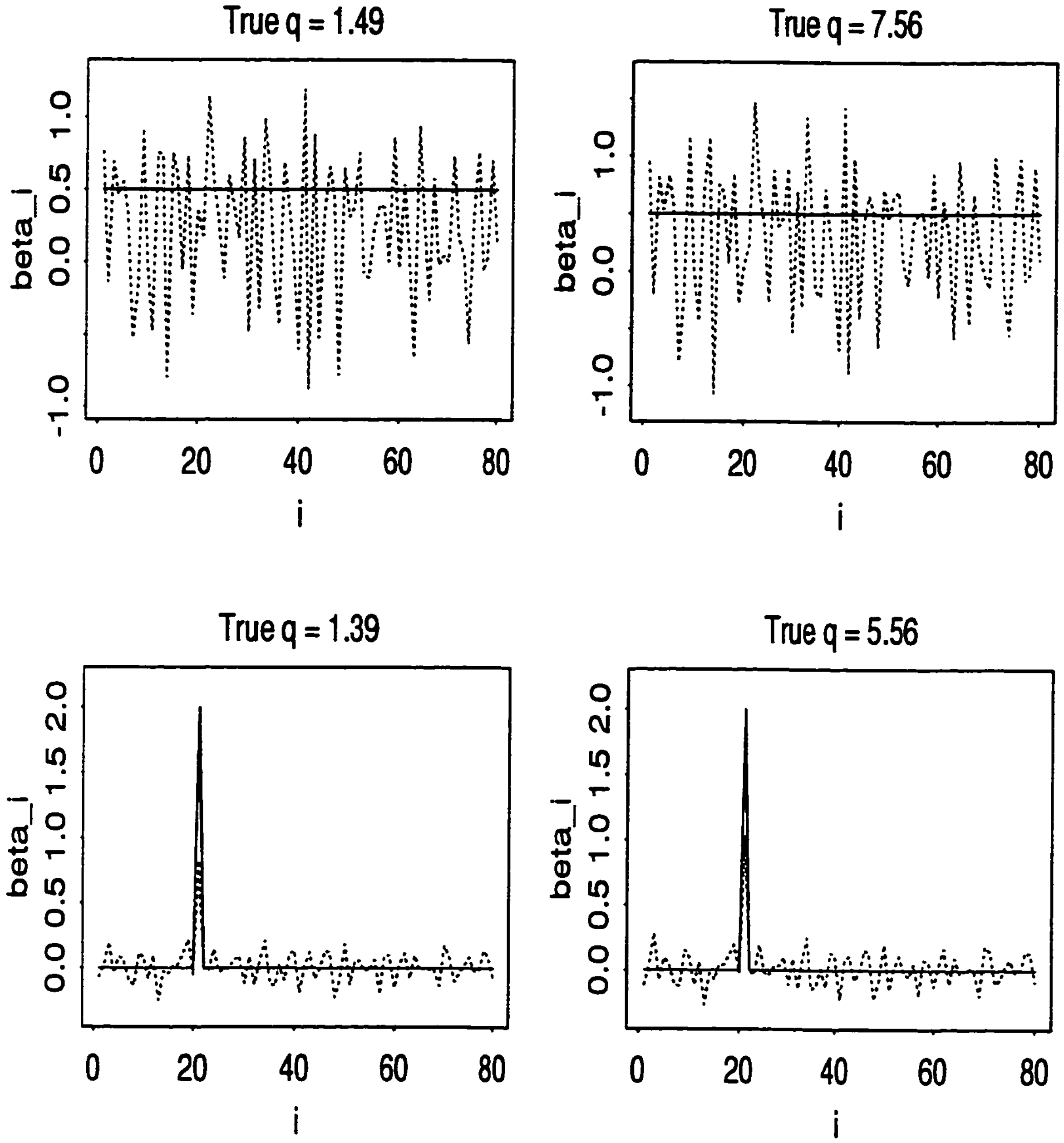


Figure 3.7: Plots of β (solid line) and mean $\hat{\beta}_{lik}$ (broken line) vector for the "flat" and "spike" situations

3.4 Estimation of Σ_X

Up to this point we have assumed that Σ_X was known. We will now deal with estimating this covariance matrix. Since its estimation will not depend on the choice of λ, we hope it should not have a major effect on our values for $\hat{\lambda}$ and $\hat{\beta}$.

There are two approaches for estimating Σ_X. The first would be to use the nonparametric estimate $\mathbf{X}'\mathbf{X}/n$ (if $\mathrm{E}(\mathbf{x}_i') = \mathbf{0}'$), or $\sum_{i=1}^n(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'/(n-1)$. The second way would be to assume the form of Σ_X was known, and we had to estimate a few parameters in the matrix.

The first method will not work in the $n < p$ case. To see this, assume that $\mathrm{E}(\mathbf{x}_i') = \mathbf{0}'$. Then $\mathbf{X}'\mathbf{X}/n$ is an unbiased estimator of Σ_X. However, if we use this estimator in (3.2.4) we get

$$(\mathbf{X}'\mathbf{X} + \lambda\hat{\Sigma})^{-1} = \left(\mathbf{X}'\mathbf{X} + \frac{\lambda}{n}\mathbf{X}'\mathbf{X}\right)^{-1} = \left(\frac{(n+\lambda)}{n}\mathbf{X}'\mathbf{X}\right)^{-1}$$

which is a singular matrix. So we have to use our second proposed method.

We assume Σ_X has the following form:

$$\Sigma_X = \begin{bmatrix} r(0) & r(1) & \dots & r(p-1) \\ r(1) & r(1) & \dots & r(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & \dots & r(0) \end{bmatrix} \qquad (3.4.1)$$

where (assuming $\mathrm{E}(\mathbf{x}_i') = \mathbf{0}'$)

$$r(k) = \mathrm{E}(x_{ij}x_{i,j+k}), \quad k = 0, \dots, p-1$$

and $r(-k) = r(k)$. This assumption of the form of Σ means that we are assuming the covariance between observations within $\mathbf{x}_i'$ depends only on the difference in location, or lag, of the points. This assumption makes sense in many dynamic models, where the relationship between points will be weaker as the distance between them increases.

For each row $\mathbf{x}_i'$ we can estimate $r(k)$ as

$$\hat{r}_i(k) = \frac{1}{p}\sum_{t=k+1}^{p} x_{it}x_{i,t-k}, \quad i = 1, \dots, n$$

(Harvey 1993, pg. 11). We use the factor $1/p$ rather than $1/(p-k)$ to ensure

$$\hat{\Sigma}_X = \begin{bmatrix} \hat{r}(0) & \hat{r}(1) & \dots & \hat{r}(p-1) \\ \hat{r}(1) & \hat{r}(0) & \dots & \hat{r}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{r}(p-1) & \hat{r}(p-2) & \dots & \hat{r}(0) \end{bmatrix}$$

is non-negative definite (Brockwell and Davis 1991, pg. 29).

Since the rows of $\mathbf{X}$ are independent, each with covariance matrix Σ_X, we can estimate $r(k)$ from the sample mean of the $\hat{r}_i(k)$ values:

$$\hat{r}(k) = \frac{1}{np}\sum_{i=1}^{n}\sum_{t=k+1}^{p} x_{it}x_{i,t-k} \ . \tag{3.4.2}$$

We will come back to this estimator later in this section. First, we will simplify things by assuming the rows of $\mathbf{X}$ are generated from an AR process of known order, and we have to estimate the parameters that govern the process. This means we will have to estimate $k+1$ parameters if we assume an AR(k) process.

We begin by assuming that each row of $\mathbf{X}$ follows an AR(2) process,

$$x_{ij} = \phi_1 x_{i,j-1} + \phi_2 x_{i,j-2} + \eta_{ij} \ , \quad j = 1, \dots, p$$

where $\eta_{ij} \sim \mathrm{N}(0, \sigma_\eta^2)$. Then

$$r(k) = \phi_1 r(k-1) + \phi_2 r(k-2), \quad k = 1, 2, \dots$$

$$r(0) = \left(\frac{1-\phi_2}{1+\phi_2}\right)\frac{\sigma_\eta^2}{[(1-\phi_2)^2 - \phi_1^2]}$$

(Harvey 1993, pg. 22-23). We must estimate ϕ_1, ϕ_2 and σ_η^2. We use the maximum likelihood estimates of these parameters from the n independent realizations of the time series to construct $\hat{\Sigma}_X$ and repeat the simulations described in section 3.3. Table 3.5 summarizes some of the results. The true AR(2) model is (3.3.1).

We see these results do not differ greatly from those in table 3.2. The $\hat{\sigma}_{lik}^2$ have increased slightly, but most values of $\hat{\lambda}$ and PSS have changed very little. The only exception is $\hat{\lambda}_{GCV}$ in the $q = 1.31$ model, whose mean value increased dramatically.

		q			
		17.37	7.21	1.31	0.094
$\hat{\sigma}^2_{lik}$	mean	15.023	11.282	58.551	932.845
	median	14.917	11.181	57.628	920.600
	var	1.992	1.413	92.141	3.53×10^4
$\hat{\lambda}_{EPMSE}$		4.329	9.222	48.702	658.467
$\hat{\lambda}_{EGCV}$		7.694	4.323	45.762	636.338
$\hat{\lambda}_{lik}$	mean	2.631	5.938	23.853	144.317
	median	2.635	5.954	23.893	144.112
	var	1.43×10^{-3}	0.0178	1.667	136.885
$\hat{\lambda}_{GCV}$	mean	7.505	5.123	1.57×10^7	8.71×10^9
	median	5.568	1.245	33.581	497.893
	var	65.685	56.657	2.47×10^{17}	3.86×10^{20}
PSS_{lik}	mean	47.434	28.931	137.565	1422.79
	median	46.524	28.519	135.399	1397.078
	var	90.236	35.042	809.556	8.56×10^4
PSS_{EGCV}	mean	47.934	29.279	133.751	1348.116
	median	46.947	28.838	132.318	1329.301
	var	90.949	35.615	759.249	7.46×10^4
PSS_{EPMSE}	mean	47.141	28.701	133.725	1348.132
	median	46.164	28.318	132.062	1329.249
	var	88.979	34.719	757.319	7.45×10^4
PSS_{GCV}	mean	50.884	30.593	149.005	1556.532
	median	46.619	30.229	146.706	1446.378
	var	119.962	40.987	1203.677	2.42×10^5

Table 3.5: Results when rows of $\mathbf{X}$ generated from AR(2) process, Σ_X correctly estimated as having AR(2) form. Model uses $n = 50$, $p = 80$

This shows the methods maintain their performance if we assume we know the correct covariance structure of $\mathbf{x}_i'$.

A more relevant problem involves the estimation of Σ_X when we do not know the true underlying process generating the rows of $\mathbf{X}$. We examined this problem in two situations. First we looked at overestimating the order of the AR process that was generating the $\mathbf{x}_i'$ values. The true AR process is (3.3.1), but we incorrectly assumed that the $\mathbf{x}_i'$ values were generated from an AR($p-1$) process. We choose such a large order to be less restrictive in the form we assume for the elements of (3.4.1). Table 3.6 summarizes the results of 1000 simulated data sets in this situation. The results should be compared to those in table 3.2.

		q			
		17.37	7.21	1.31	0.094
$\hat{\lambda}_{EPMSE}$		5.209	27.986	91.704	561.956
$\hat{\lambda}_{EGCV}$		2.829	15.430	60.398	377.871
$\hat{\lambda}_{lik}$	mean	2.659	6.017	24.164	147.798
	median	2.664	6.046	24.248	147.374
	var	1.20×10^{-3}	0.0409	2.636	162.190
$\hat{\lambda}_{GCV}$	mean	3.677	12.83	54.434	4.51×10^{9}
	median	0.157	11.597	52.125	326.700
	var	29.685	127.596	1624.932	2.72×10^{20}

Table 3.6: Summary of $\hat{\lambda}$ values when rows of $\mathbf{X}$ are generated from an AR(2) process, and incorrectly estimated as coming from an AR($p-1$) process. Models use $n = 50$, $p = 80$

We see that $\hat{\lambda}_{EPMSE}$ and $\hat{\lambda}_{EGCV}$ increased by a large amount in two situations. However, we see the $\hat{\lambda}_{lik}$ remained quite close to the values obtained when the true Σ_X was known. The $\hat{\lambda}_{GCV}$ did not follow a regular pattern as q changed. We also see the smoothing parameters often became more variable when Σ_X was estimated.

Next we generated the rows of $\mathbf{X}$ from the AR(2) process (3.3.1), but incorrectly assumed that the $\mathbf{x}_i'$ values were generated from an AR(1) process. So we were underfitting, *i.e.* we estimated 2 parameters instead of 3. Table 3.7 summarizes the results of this simulation, and these values should be compared to those in table 3.2.

		q			
		17.37	7.21	1.31	0.094
$\hat{\lambda}_{EPMSE}$		5.375	13.197	60.574	600.258
$\hat{\lambda}_{EGCV}$		6.078	6.740	48.00	478.326
$\hat{\lambda}_{lik}$	mean	2.635	5.938	23.703	143.568
	median	2.639	5.956	23.788	143.183
	var	1.73×10^{-3}	0.0260	2.129	156.028
$\hat{\lambda}_{GCV}$	mean	5.620	6.641	47.323	7.90×10^{9}
	median	2.990	3.758	39.736	405.734
	var	44.132	63.736	1943.38	4.39×10^{20}

Table 3.7: Summary of $\hat{\lambda}$ values, when rows of $\mathbf{X}$ are generated from an AR(2) Process, and incorrectly estimated as coming from an AR(1) process. Models use $n = 50$, $p = 80$

We see that the $\hat{\lambda}_{lik}$ values change by very little, while $\hat{\lambda}_{GCV}$, $\hat{\lambda}_{EGCV}$ and $\hat{\lambda}_{EPMSE}$ tend to increase. The size of the change depends on q. We do see that the methods still perform reasonably well with incorrect estimates of Σ_X.

We will now try to give some informal justification as to why our results remain similar if we replace Σ_X with $\hat{\Sigma}_X$ in (3.2.4). It will focus on writing

$$\hat{\Sigma}_X = \Sigma_X + \mathbf{R}$$

where it is hoped $\mathbf{R}$ will be of a small order. This will mean that

$$(\mathbf{X}'\mathbf{X} + \lambda\Sigma_X)^{-1} \quad \text{and} \quad (\mathbf{X}'\mathbf{X} + \lambda\Sigma_X + \lambda\mathbf{R})^{-1}$$

should be similar, yielding little change in $\hat{\boldsymbol{\beta}}$.

Let us assume we estimate the elements of $\hat{\Sigma}_X$ using (3.4.2). Then it is clear that

$$\mathrm{E}[\hat{r}(k)] = \left(1 - \frac{k}{p}\right) r(k) \ .$$

It is also true that

$$\begin{aligned} \mathrm{E}(\hat{r}^2(k)) &= \frac{1}{np^2}\left[n(p-k)^2 r^2(k) + (p-k)(r^2(0)+r^2(k))\right] \\ &\quad + \frac{1}{np^2}\left[2\sum_{w=0}^{p-k-1}(p-k-w)[r(w-k)r(w+k)+r^2(w)]\right] \end{aligned}$$

If we consider a single time series with p observations, Anderson (1971, pg. 463) shows that the bias of $\hat{r}_i(k)$ is of order $1/p$ and Priestley (1981, pg. 324) states that the variance of $\hat{r}_i(k)$ is of order $1/p$. In our case $\hat{r}(k)$ is the mean of n independent estimates, so it should also gave a bias and variance of the same order. We can use Chebyshev's inequality (Casella and Berger 1990, pg. 184) to write

$$\begin{aligned} \mathrm{P}(|\hat{r}(k)-r(k)| \geq \epsilon) &= \mathrm{P}(|\hat{r}_1(k)-r(k)+\cdots+\hat{r}_n(k)-r(k)| \geq n\epsilon) \\ &\leq \frac{1}{n^2\epsilon^2}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left[(\hat{r}_i(k)-r(k))(\hat{r}_j(k)-r(k))\right]\right] . \end{aligned}$$

The $i=j$ terms in the sum will give us n terms that are $O_p(1/p)$, and the $i \neq j$ terms will yield $n(n-1)$ bias terms that are $O_p(1/p)$. So we may now say

$$\begin{aligned} \mathrm{P}(|\hat{r}(k)-r(k)| \geq \epsilon) &\leq \frac{1}{n^2\epsilon^2}\left[nO(1/p)+n(n-1)O(1/p)\right] \\ &\leq \frac{1}{\epsilon^2}O(1/p) . \end{aligned}$$

So it appears that, for fixed k, $\hat{r}(k) = r(k) + O_p(p^{-1/2})$. Since we will be able to write $\hat{\Sigma}_X$ as the sum of the true value Σ_X plus a term of small order, we can use $\hat{\Sigma}_X$ in place of Σ_X in our procedure and still obtain reasonable estimates.

Let us discuss for a moment the implications of n and p increasing. If n increases with p fixed, then our model is no longer underdetermined. This is not the case in which we are interested. If we let p increase, we can interpret this as collecting our explanatory variables along a finer grid (every 10 km rather that every 20 km, for example). This retains the underdetermined nature of the problem, plus gives us a better physical interpretation. It may be relatively easy to increase the sampling of our $\mathbf{X}$ values, but we may be restricted in the number of response variables we have available. However, if we let $p \longrightarrow \infty$ we change the covariance structure in the problem. In applications we may not be able to let p increase for a given covariance structure.

Since we will have to estimate Σ_X in the analysis of the California Current data, we will return to this topic at that point.

3.5 Examination of Range of q Values

We have shown that using the signal-to-noise penalty performs well in selecting a smoothing parameter when the ratio q is known. We now wish to relax this assumption, first by studying the effect of specifying a value of q, say q^*, which is different than the true q.

We begin by examining the sensitivity of the results to the choice of q, if we are given a range of values to examine. Our simulations will use the same model that was described in the discussion of tables 3.1 through 3.3. However, we generate our $\mathbf{X}$ and $\mathbf{y}$ values using the true values of q and σ^2, and find λ to solve

$$\frac{\hat{\boldsymbol{\beta}}'_{\lambda}\Sigma_X\hat{\boldsymbol{\beta}}_{\lambda}}{\hat{\sigma}^2_{lik(\lambda)}} = q^* \tag{3.5.1}$$

where q^* comes from a range of possible values of q. In the examples we will present we allow q^* to range from 0.001 to 1000. Below $q^* = 0.001$ the model is dominated by noise, so the model will have little predictive ability. Values above $q^* = 1000$ would indicate the model gives near perfect predictability. Therefore we see no need to examine values of q^* outside of this range. However, our tables of summary statistics will only contain a subset of the values of q^* used. We will let $\hat{\lambda}^*_{lik}$ be the solution to (3.5.1). We then construct the estimate $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$.

Next, we generate a new set $\mathbf{y}_{new}$ using the same $\mathbf{X}$ matrix, and see how well we predict $\mathbf{y}_{new}$ by calculating the prediction sum of squares

$$PSS = (\mathbf{y}_{new} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\hat{\lambda}})'(\mathbf{y}_{new} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\hat{\lambda}})/n \; .$$

For a particular q^*, we repeat this process for 100 simulated data sets.

Tables 3.8 through 3.11 summarize some of the results for a variety of true values of q. The term *std. error* refers to the standard error of the mean of the PSS_{lik} values. For comparison we have included the results when using $\hat{\lambda}_{GCV}$.

mean($\hat{\lambda}_{GCV}$)		mean(PSS_{GCV})	
7.128		49.953	
q^*	mean($\hat{\lambda}_{lik}$)	mean(PSS_{lik})	std. error(mean(PSS_{lik}))
1	32.97	67.60	1.126
2	18.50	54.09	0.988
5	8.28	47.16	0.890
12	3.73	46.22	0.868
20	2.31	46.58	0.871
25	1.87	46.81	0.874
30	1.57	46.99	0.877
40	1.19	47.28	0.882

Table 3.8: Prediction SS values over a range of q values. True $q = 17.37$

We will first discuss the changes in $\hat{\lambda}_{lik}$ as q^* increases. We see that $\hat{\lambda}_{lik}$ decreases as q^* increases. which is no surprise. We have seen this result before when we treated q as known. We also see that the changes in $\hat{\lambda}_{lik}$ are smaller when we move between larger values of q^*.

A more interesting comparison involves the PSS_{lik} values. In table 3.8 we see PSS_{lik} changes noticeably for q^* between 1 and 5, but then changes very little for larger q^* values. We also note that PSS_{lik} achieves its minimum around $q^* = 12$, which is close to the true value of $q = 17.37$. We can see that we could use a wide choice of q^* values (from 5 to 40) and still achieve a smaller PSS than by using GCV. When we take the standard errors into account, we see that PSS_{GCV} is not within two standard errors of PSS_{lik} at $q^* = 12$.

In table 3.9 we are using a smaller true value of q, but we see a similar pattern. The PSS_{lik} achieves a minimum around $q^* = 5$ (with the true $q = 7.21$), and the PSS_{lik} values drop sharply before the minimum, and increase gradually after it. We outperform GCV for q^* choices between 2 and 25. At $q^* = 5$ we see that PSS_{GCV} is not within two standard errors of PSS_{lik}.

mean(λ_{GCV})		mean(PSS_{GCV})	
4.57		29.78	
q^*	mean(λ_{lik})	mean(PSS_{lik})	std. error(mean(PSS_{lik}))
1	33.03	32.33	0.646
2	18.51	28.85	0.591
5	8.30	27.98	0.565
12	3.74	28.80	0.561
20	2.32	29.40	0.565
25	1.87	29.65	0.567
30	1.58	29.84	0.568
40	1.20	30.10	0.570

Table 3.9: Prediction SS values over a range of q values. True $q = 7.21$

In table 3.10, where the true $q = 1.31$ is much smaller, the pattern is altered slightly. We see the PSS_{lik} values change by a greater amount at the large values of q^*. The PSS_{lik} is minimized at $q^* = 0.5$, which is less than half of the true q. This is poorer performance than what we observed in tables 3.8 and 3.9. A choice of q^* from 0.5 to 2 results in better predictive ability than GCV. In this case, we also see that PSS_{GCV} is not within two standard errors of PSS_{lik} at $q^* = 0.5$, $q^* = 1$ and $q^* = 2$.

Table 3.11 continues the pattern of table 3.10. With the true $q = 0.094$ we find our PSS_{lik} is minimized at $q^* = .007$, and the PSS_{lik} values continue to change sharply for large values of q^*. We also have a smaller range of q^* values where the signal-to-noise method outperforms GCV, based on predictive ability. Because of the poorer performance of GCV in this situation, we can see there are several choices of q^* where PSS_{GCV} is not within two standard errors of PSS_{lik}.

We can draw a few conclusions at this point. For the larger values of q, we see it would be more important to have a good lower bound for q than a good upper bound. This is because the PSS_{lik} values do not change a great deal for large choices of q^*. Plots of the elements of $\partial\hat{\beta}_i/\partial q$ and $\partial\hat{y}_i/\partial q$ give the same information; the plots are quite flat for $q > 30$. The PSS_{lik} values were minimized around the true value of q,

mean($\hat{\lambda}_{GCV}$)		mean(PSS_{GCV})	
38.13		145.29	
q^*	mean($\hat{\lambda}_{lik}$)	mean(PSS_{lik})	std. error(mean(PSS_{lik}))
0.1	166.38	145.43	2.88
0.5	51.95	130.66	2.69
1	29.99	132.90	2.70
2	16.92	138.88	2.78
5	7.72	149.02	2.92
10	4.18	156.25	3.01
20	2.23	162.67	3.08
50	0.94	167.19	3.15

Table 3.10: Prediction SS values over a range of q values. True $q = 1.31$

and there was a wide range of q^* values where this method had lower predictive error than GCV.

The results were different when the true q was quite small. The PSS_{lik} values were minimized at a smaller value than the true q, which would lead to doing more smoothing. We also do not have the same pattern of the PSS_{lik} values changing little for large values of q^*. However, we do seem to have reasonable ranges of q^* where $PSS_{lik} < PSS_{GCV}$.

These results show that PSS_{lik} is minimized, on average, near (but usually below) the true value of q. We now wish to see how we do in the individual data sets used. We want to see for what value of q^* do we minimize PSS_{lik} for each simulated data set. This will tell us how many of the individual data sets suggest choosing q in the correct range of values.

In table 3.8 we see the mean of the PSS_{lik} values is minimized at $q^* = 12$. We chose $q^* = 2$ in one data set, $q^* = 12$ in 21 cases and $q^* = 15$ in 8 cases. We chose the maximum q^* allowed, $q^* = 1000$, once.

Table 3.9 shows that choosing $q^* = 5$ minimized the mean of the PSS_{lik} values. The smallest value of q^*, $q^* = 1$, was chosen twice, $q^* = 4$ was chosen 52 times and $q^* = 8$ was chosen 15 times. The largest q^* allowed, $q^* = 1000$, was chosen once.

mean($\hat{\lambda}_{GCV}$)		mean(PSS_{GCV})	
5.639×10^9		1557	
q^*	mean($\hat{\lambda}_{lik}$)	mean(PSS_{lik})	std. error(mean(PSS_{lik}))
.003	1122.68	1324.06	26.80
.007	706.10	1321.24	26.85
.03	303.53	1337.30	27.27
.1	141.16	1404.59	28.40
1	26.47	1759.00	34.40
5	7.18	2064.93	39.72
10	3.98	2169.26	41.38
20	2.16	2249.43	42.57

Table 3.11: Prediction SS values over a range of q Values. True $q = 0.094$

Table 3.10 tells us that the mean of the PSS_{lik} values is minimized at $q^* = 0.5$. In the individual sets we chose $q^* = 0.1$ in 10 sets and $q^* = 0.5$ in 57 sets. We never chose $q^* > 20$.

With $q = 0.094$ table 3.11 shows the mean of the PSS_{lik} values is minimized at $q^* = 0.007$. In the individual sets $q^* = 0.001$ (the smallest value allowed in the simulations) was chosen 31 times, $q^* = 0.007$ was chosen 5 times, and we never chose $q^* > 1$.

These results indicate that for some values of q our method performs well not only on average, but on a case-by-case basis as well. With a small value of q we are not seeing any tendency to do a great deal of undersmoothing. The same cannot be said for GCV. There are some situations where the true q is small and GCV will still do no smoothing. The results also show we may be able to get a reasonable estimate of q from the data using a predictive loss criterion.

3.6 Use of Data-Splitting to Estimate q

The results in the previous section suggest that the signal-to-noise method for choosing λ can still perform well within a range of the true value of q. Since this performance was based on how well we predicted a new data set, it suggests that we may be able to use a data-splitting technique to estimate q.

We will first outline the general data-splitting procedure before discussing the methods by which we will assess the procedure.

1. Split the data set of n observations randomly into a construction set $(\mathbf{X}_c, \mathbf{y}_c)$ of n_c observations and a validation set $(\mathbf{X}_v, \mathbf{y}_v)$ of n_v observations. A reasonable method to use is a 50–50 split, so $n_c = n_v = n/2$.

2. Find λ to solve
$$\frac{\hat{\boldsymbol{\beta}}_c' \boldsymbol{\Sigma}_X \hat{\boldsymbol{\beta}}_c}{\hat{\sigma}^2_{lik(\lambda,c)}} = q^*$$
for a fixed choice of q^*, where
$$\hat{\boldsymbol{\beta}}_c = (\mathbf{X}_c'\mathbf{X}_c + \lambda\boldsymbol{\Sigma}_X)^{-1}\mathbf{X}_c'\mathbf{y}_c$$
and $\hat{\sigma}^2_{lik(\lambda,c)}$ is found using (3.2.10), but replacing $(\mathbf{X}, \mathbf{y})$ with $(\mathbf{X}_c, \mathbf{y}_c)$. Call this value $\hat{\lambda}$.

3. Compute
$$\hat{\boldsymbol{\beta}}_{c,\hat{\lambda}} = (\mathbf{X}_c'\mathbf{X}_c + \hat{\lambda}\boldsymbol{\Sigma}_X)^{-1}\mathbf{X}_c'\mathbf{y}_c$$
using $\hat{\lambda}$ from step 2.

4. Compute
$$SS_{v,i} = (\mathbf{y}_v - \mathbf{X}_v\hat{\boldsymbol{\beta}}_{c,\hat{\lambda}})'(\mathbf{y}_v - \mathbf{X}_v\hat{\boldsymbol{\beta}}_{c,\hat{\lambda}})/n_v \qquad (3.6.1)$$
as the measure of loss in predicting $\mathbf{y}_v$ for the *ith* random split of the data.

5. Repeat steps 1–4 for k random splits of the data, and average (3.6.1) over the k splits, finding
$$SS_v = \frac{1}{k}\sum_{i=1}^{k} SS_{v,i} \qquad (3.6.2)$$
Reasonable choices for k would be $k = 50$ or $k = 100$.

6. Repeat steps 1–5 over a range of q^* values, and choose $\hat{q}^*$ as the value which gives us the minimum value of (3.6.2).

7. Take $\hat{q}^*$ from step 6 and find $\hat{\lambda}_{lik}$ using (3.2.12) , with $\hat{q}^*$ replacing q.

8. Calculate $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ in (3.2.8) using $\hat{\lambda}_{lik}$ from step 7.

9. Use $\hat{\lambda}_{lik}$ and $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ to find $\hat{\sigma}^2_{lik(\hat{\lambda})}$ in (3.2.13).

To investigate the preceding method for estimating q we have performed the following simulations. We use model (3.2.1) with $p = 80$ and $n = 50$. The data is simulated as described in section 3.3, using the same choices for q and generating the rows of $\mathbf{X}$ from the AR(2) process defined in (3.3.1). Once the data is generated we use the above procedure to find our required estimates. In the procedure we used $n_c = n_v = 25$, $k = 50$, and q^* values ranging from 0.001 to 1000. We also calculated $\hat{\lambda}_{GCV}$ for the data set.

We then generated a new data set of $n = 50$ observations, and evaluated the prediction sums of squares

$$PSS_s = \frac{1}{n}(\mathbf{y}_{new} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\hat{\lambda}})'(\mathbf{y}_{new} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\hat{\lambda}}) \tag{3.6.3}$$

and

$$PSS_G = \frac{1}{n}(\mathbf{y}_{new} - \mathbf{X}\hat{\boldsymbol{\beta}}_{GCV})'(\mathbf{y}_{new} - \mathbf{X}\hat{\boldsymbol{\beta}}_{GCV}) \ . \tag{3.6.4}$$

where $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ and $\hat{\boldsymbol{\beta}}_{GCV}$ were found using the original 50 observations.

This procedure was repeated 100 times. We now wish to discuss some of the results, many of which are summarized in table 3.12. In the table we define $PSS_d = PSS_s - PSS_G$.

We first examine the $q = 17.37$ case. We see that the mean of the $\hat{q}$ values is large. Of our 100 data sets, we chose $\hat{q} = 1000$, the maximum allowed in our simulations, 24 times. But we see the median of the $\hat{q}$ values is below the true q. This led to $\hat{\lambda}_{lik}$ values which were larger than the $\hat{\lambda}_{GCV}$ values, on average. We also found that $\hat{\lambda}_{GCV} = 0$ in 32 data sets.

		q			
		17.37	7.21	1.31	0.094
$\hat{q}$	mean	261.145	140.485	0.7488	0.0697
	median	10	5	0.5	0.008
$\hat{\lambda}_{lik}$	mean	7.724	11.633	82.513	988.730
	median	4.419	8.340	53.054	644.830
	var	91.928	113.122	3.98×10^3	6.07×10^5
$\hat{\lambda}_{GCV}$	mean	7.128	4.574	38.137	5.64×10^9
	median	4.932	0.980	28.346	311.988
	var	61.083	50.006	2.17×10^3	3.29×10^{20}
PSS_d	mean	0.423	−0.513	−6.857	−188.302
	min	−11.547	−7.427	−58.615	−2201.499
	max	21.200	12.065	26.217	45.278

Table 3.12: Estimates of q when rows of $\mathbf{X}$ generated from an AR(2) process. Models use $n = 50$, $p = 80$

Let us now look at the PSS_d values. Recall that table 3.3 gives the relative magnitudes of the PSS values. We see the two methods perform similarly, with GCV coming out with the advantage, on average. Out of the 100 simulated data sets, we found PSS_d was negative 58 times. So the signal-to-noise method comes out with the advantage in a majority of the individual cases.

We are also interested in the association between the differences $\hat{\lambda}_{lik} - \hat{\lambda}_{GCV}$ and the corresponding values of PSS_d. We would like to find out if the larger values of PSS_d tend to be associated with very high values of $\hat{\lambda}_{GCV}$, for example. Figure 3.8 is a plot of PSS_d versus $\hat{\lambda}_{lik} - \hat{\lambda}_{GCV}$ when $q = 17.37$.

In figure 3.8 points above the horizontal line indicate cases where GCV gave us a smaller prediction error. We see that $\hat{\lambda}_{lik}$ can do too much smoothing in this situation, and consequently has poorer predictive ability.

With the true $q = 7.21$ the mean $\hat{q}$ value overestimates q again, while the median value underestimates it. In this case there were 13 situations where we found $\hat{q} = 1000$. We continue to see the average of the $\hat{\lambda}_{GCV}$ values is smaller than the average of the $\hat{\lambda}_{lik}$ values.

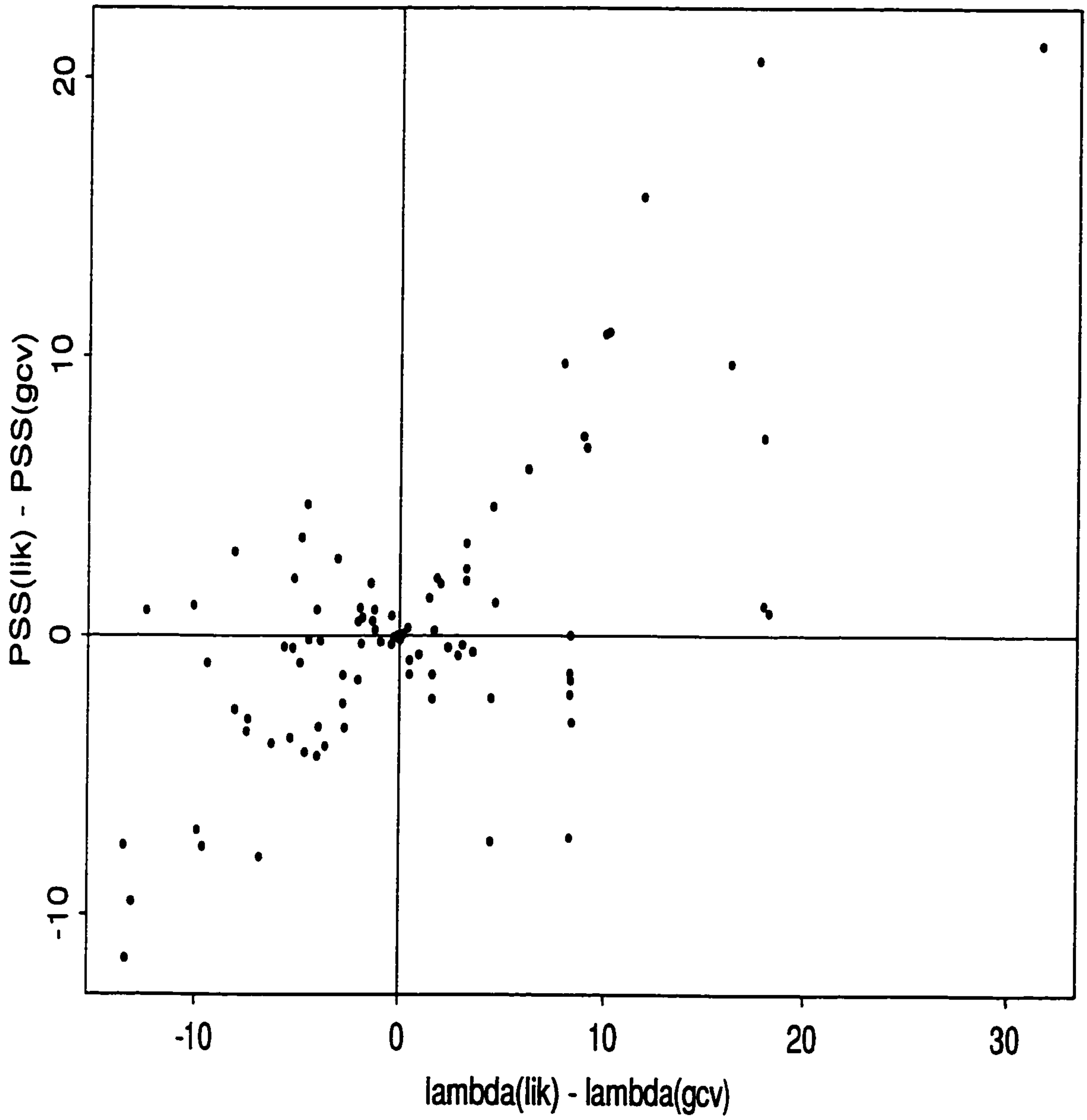


Figure 3.8: Comparison of λ estimates and resulting PSS values when estimating q. True $q = 17.37$ and the rows of $\mathbf{X}$ are generated from an AR(2) process

When we compare the PSS_d values we find that the advantage now goes to the signal-to-noise method. The mean of the PSS_d values is negative, but the two methods still perform similarly. For the individual data sets, PSS_d is negative in 62 of 100 cases. So with the decrease in q the results move slightly in favour of the signal-to-noise method, with GCV tending to undersmooth more often. However, the prediction results are comparable.

Things move more in favour of the signal-to-noise method when $q = 1.31$. The mean and median of the $\hat{q}$ values are similar, and both are biased downwards. We note that, even for this small value of q, we find $\hat{\lambda}_{GCV} = 0$ in 24 sets.

The PSS values start to differ by a greater amount in this situation, as seen in table 3.12. If we look at the individual cases, we find PSS_d is negative in 58 of 100 cases. Also, there are 32 cases where $PSS_d < -10$, while only 9 cases where $PSS_d > 10$. So there are more situations where GCV does much worse, and few where it does much better. We will come back to this point in more detail.

Finally let's examine our results when $q = 0.094$. The downward bias in the mean of our $\hat{q}$ values is becoming larger as q decreases. The method chose $\hat{q} = 0.001$, the smallest allowed, in 42 cases. But we do see GCV gave highly variable estimates for this model, as we had seen before.

When we examine the PSS_d values we see many cases where the signal-to-noise method does substantially better than GCV, and never much worse. The mean of the PSS_d strongly favours the signal-to-noise method. Out of the 100 simulated sets, PSS_d is negative 74 times. To illustrate the situations where the signal-to-noise method does much better, there were 33 cases where $PSS_d < -50$.

Figure 3.9 plots PSS_d versus $\hat{\lambda}_{lik} - \hat{\lambda}_{GCV}$ for the $q = 0.094$ case. For clarity the plot omits the cases where $\hat{\lambda}_{lik} - \hat{\lambda}_{GCV} < -2000$. In this plot we see a larger number of points in the lower quadrant of the figure. This indicates there are many cases where $\hat{\lambda}_{lik} - \hat{\lambda}_{GCV}$ is very large, but PSS_d is very much in favour of the signal-to-noise method. We seem to be observing that GCV can do too little smoothing when the true value of q is small, and the consequence is very poor predictive ability when compared to the signal-to-noise method.

These results show we can gain noticeable improvements over GCV when the true value of q is small, with few situations where GCV does a great deal better. The methods give similar results, based on the $\hat{\lambda}$ values and predictive ability, when q is large. Therefore we seem to have established that in the most general case (no *a priori* knowledge of q) we will not do any worse than GCV. Any knowledge of the signal-to-noise ratio we do have from the particular problem will only improve the results from the signal-to-noise method.

The numerical results indicate the procedure has a downward bias in estimating q. To investigate this we looked at the value of q^* which minimizes

$$\mathrm{E}(PSS_{lik}) = \frac{1}{n}\mathrm{E}\left(\mathbf{y}_{new} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda\right)'\left(\mathbf{y}_{new} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda\right)$$

where the expectation is over $\mathbf{X}$ and ϵ. If we first condition on ϵ it can be shown that

$$\mathrm{E}(PSS_{lik}) = \sigma^2(q+1) + \frac{1}{n}\mathrm{E}_X[(\mathbf{X}\boldsymbol{\beta})'\mathbf{H}^2\mathbf{X}\boldsymbol{\beta} - 2(\mathbf{X}\boldsymbol{\beta})'\mathbf{H}\mathbf{X}\boldsymbol{\beta} + \sigma^2\mathrm{tr}(\mathbf{H}^2)] \qquad (3.6.5)$$

where $\mathbf{H}$ is defined on page 20, and q is the true signal-to-noise ratio.

Because this expectation is complicated we evaluate (3.6.5) by Monte Carlo sampling from the distribution of $\mathbf{X}$. We want to see if the value of q^* which minimizes (3.6.5) is close to the true q. To assess this we have carried out the following simulation.

First we generate a matrix $\mathbf{X}_i$, where the rows of $\mathbf{X}_i$ are generated from a N($\mathbf{0}$, Σ_X) distribution. Then, for a fixed q^*, we find λ to solve

$$\mathrm{E}_{\epsilon|X}\left(\frac{\hat{\boldsymbol{\beta}}_\lambda'\Sigma_X\hat{\boldsymbol{\beta}}_\lambda}{\sigma^2}\right) = q^* .$$

Call this value $\hat{\lambda}_q$. Next we evaluate

$$(\mathbf{X}_i\boldsymbol{\beta})'\mathbf{H}_i^2\mathbf{X}_i\boldsymbol{\beta} - 2(\mathbf{X}_i\boldsymbol{\beta})'\mathbf{H}_i\mathbf{X}_i\boldsymbol{\beta} + \sigma^2\mathrm{tr}(\mathbf{H}_i^2)$$

for this choice of $\mathbf{X}_i$ and $\hat{\lambda}_q$. We then repeat this procedure, for fixed q^*, over w randomly generated $\mathbf{X}_i$ matrices, and approximate (3.6.5) by

$$\mathrm{E}(PSS_{lik}) \approx \sigma^2(q+1) + \frac{1}{nw}\sum_{i=1}^{w}[(\mathbf{X}_i\boldsymbol{\beta})'\mathbf{H}_i^2\mathbf{X}_i\boldsymbol{\beta} - 2(\mathbf{X}_i\boldsymbol{\beta})'\mathbf{H}_i\mathbf{X}_i\boldsymbol{\beta} + \sigma^2\mathrm{tr}(\mathbf{H}_i^2)] \quad (3.6.6)$$

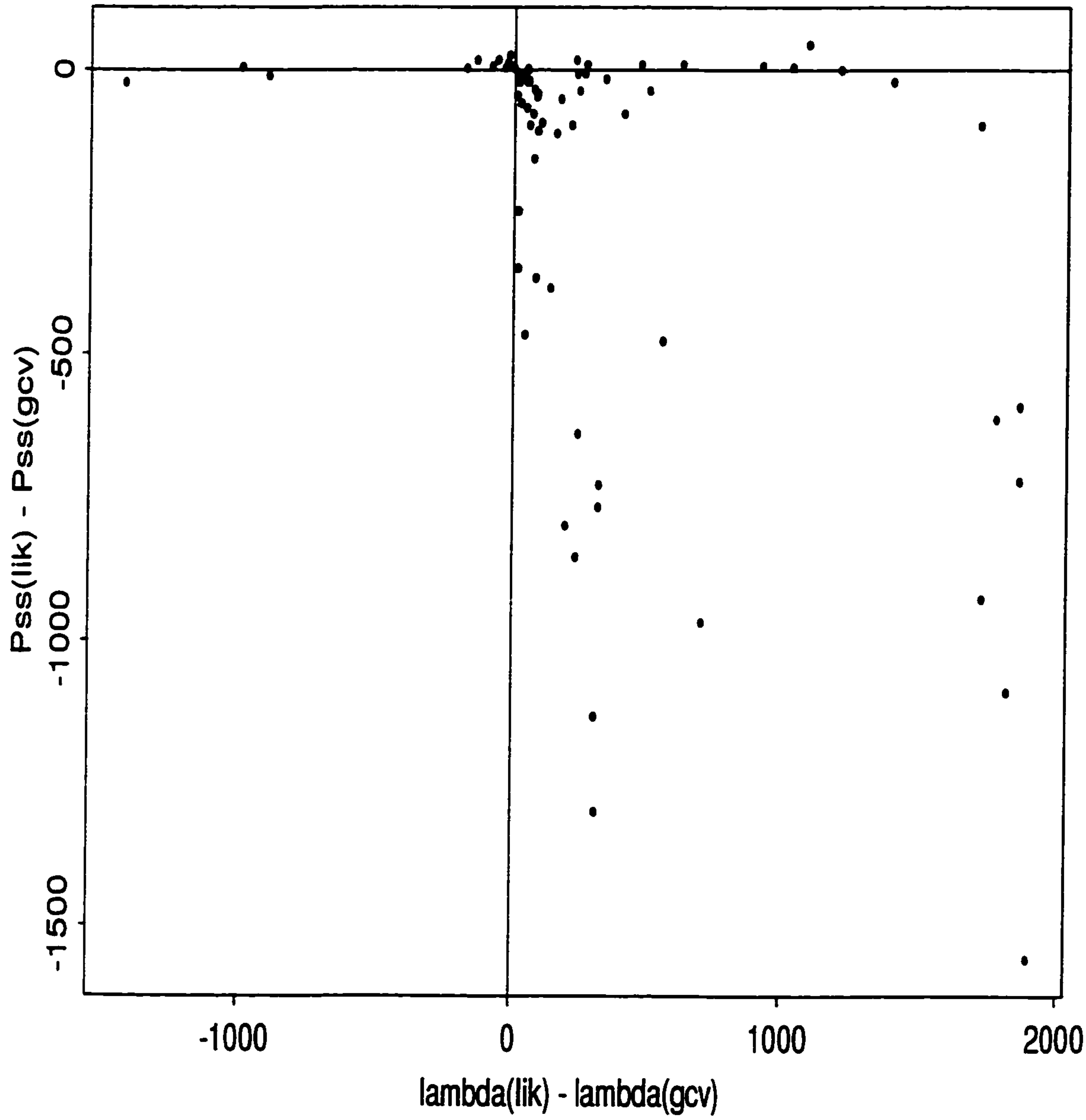


Figure 3.9: Comparison of λ estimates and resulting PSS values when estimating q: True $q = 0.094$ and the rows of $\mathbf{X}$ are generated from an AR(2) process

This is repeated over a grid of q^* values to see what choice of q^* minimizes (3.6.6).

The model used is similar to those presented earlier: we used $n = 50$, $p = 80$ and the rows of each $\mathbf{X}_i$ were generated using the AR(2) process (3.3.1). We used $w = 500$ $\mathbf{X}_i$ matrices in evaluating (3.6.6). The values of q we used ranged from 0.094 to 1.997.

We found (3.6.6) was minimized by $q^* = 0.005$ when $q = 0.094$. When $q = 0.597$ we found $q^* = 0.11$ minimized (3.6.6). We found (3.6.6) was minimized by $q^* = 0.5$ when $q = 1.306$. Finally, When $q = 1.997$ we found $q^* = 0.6$. It seems that (3.6.5) is minimized by a value of q^* that is smaller than the true q. The amount of this bias increased as the true q is decreased. It also appears that the $\hat{q}$ values that we have calculated are close to this minimizing q^*.

3.7 Conclusions

The use of the signal-to-noise ratio appears to hold promise in the underdetermined regression model for choosing a smoothing parameter. It imposes a degree of structure in the model, and allows for a method of choosing λ that is not difficult to implement. It also permits us to explicitly use the covariance structure of our explanatory variables in the estimation procedure. We have shown that knowledge of the true ratio leads to estimates that have very small variability, and have good predictive ability in many situations. Although the results are specific to the models used in the simulations studies, they should apply more generally because we have looked at cases over a range of signal-to-noise ratio values, which appears to be the important factor in determining how the method performs.

With the signal-to-noise ratio unknown, we have developed an estimation procedure based on data-splitting that yields reasonable estimates and is less likely to give very poor results, especially if the true signal-to-noise ratio is small. Any *a priori* knowledge of the ratio that can be supplied will only improve the estimates.

Chapter 4

Extension to Robust Estimation

4.1 Introduction

In the previous chapter we introduced the linear model with random explanatory variables, developed a method of estimation when the model is underdetermined, and discussed a procedure for estimating the signal-to-noise ratio of the model. We will now deal with some robust alternatives to the estimation procedure. We will begin with an overview of robust estimation in multiparameter models, followed by a discussion of robust estimation in linear models. Then we will introduce some robust alternatives that have been proposed in ridge regression. This will lead us into our proposed procedures for robustly estimating the model parameters with the signal-to-noise ratio fixed, followed by combining this with a robust procedure for estimating the ratio.

4.2 Robust Estimation: Multiparameter Models

We begin with an overview of some important concepts and results in robust estimation before dealing with linear models. Most of these results are found in Hampel, Ronchetti, Rousseeuw, and Stahel (1986) and the references therein. We begin by focusing on multiparameter models.

4.2.1 Influence Function

Suppose we have a parameter space $\mathbf{\Theta} \in \mathbf{R}^p$, and T is an estimate that can be viewed as a functional defined on a suitable subset of the set of probability measures on a sample space $\mathcal{X}$, taking values in $\mathbf{\Theta}$. Then the p dimensional influence function of T at a distribution F is

$$\mathrm{IF} = \mathrm{IF}(x; T, F) = \lim_{h \downarrow 0} \frac{T[(1-h)F + h\Delta_x] - T[F]}{h} \qquad (4.2.1)$$

where Δ_x is the probability measure which puts mass 1 at the point x (Hampel et al. 1986, pg. 226). It measures the sensitivity of T to the single point x. To see this, consider the one dimensional location problem and the sample mean $\bar{x}$. It is well-known that $\bar{x}$ is susceptible to outliers in the data. In fact, the influence function of $\bar{x}$ is x, so its IF in unbounded (Hampel et al. 1986, pg. 89). This tells us that a single outlier can have a dramatic effect on $\bar{x}$.

Under certain regularity conditions we have

$$\int \mathrm{IF} dF(x) = \mathbf{0} \quad \text{(Fisher consistency)}$$

and T is asymptotically normal with covariance matrix

$$\mathbf{V}(T, F) = \int \mathrm{IF}(\mathrm{IF})' dF(x)$$

The gross error sensitivity of an estimator T measures its sensitivity to outliers. It is defined as a function of the IF, in the unstandardized and standardized cases, as follows:

$$\gamma_u^*(T, F) = \sup_x(||\mathrm{IF}||) \quad \text{(unstandardized)}$$

$$\gamma_s^*(T, F) = \sup_x(\mathrm{IF}\mathbf{V}^{-1}\mathrm{IF}')^{-1/2} \quad \text{(standardized)}$$

where $||\mathbf{x}|| = (\mathbf{x}'\mathbf{x})^{1/2}$.

4.2.2 M-Estimators

There are different classes of robust estimators that can be used. One of the most common robust estimators, and the one that generalizes most easily to the multiparameter case, is the M-estimator, which is defined as the solution to

$$\sum_{i=1}^{n} \mathbf{\Psi}(x_i, \boldsymbol{\theta}) = \mathbf{0} \tag{4.2.2}$$

For the MLE, $\mathbf{\Psi}(x, \boldsymbol{\theta}) = \partial \ln f / \partial \boldsymbol{\theta} = \mathrm{s}(x, \boldsymbol{\theta})$ is the score function (Hampel et al. 1986, pg. 230). The idea is to replace s with a function that is less sensitive to outliers.

Hampel et al. (1986, pg. 231) show that, for the M-estimator,

$$\mathrm{IF} = \mathbf{M}^{-1}\mathbf{\Psi}, \quad M = \left[-\frac{\partial}{\partial \boldsymbol{\theta}} \int \mathbf{\Psi}(x_i, \boldsymbol{\theta}) dF(x)\right]^{-1}$$

This leads us to say that, under certain conditions, the M-estimator is Fisher consistent and is asymptotically normal with covariance matrix

$$\mathbf{V}(\boldsymbol{\theta}, F) = \mathbf{MQ}(\mathbf{M}')^{-1}$$

with

$$\mathbf{Q} = \int \mathbf{\Psi}(x, \boldsymbol{\theta})[\Psi(x, \boldsymbol{\theta})]' dF(x)$$

(Hampel et al. 1986, pg. 231).

4.2.3 Optimal Estimators

If the score function $\mathrm{s}(x, \boldsymbol{\theta})$ is unbounded in x, the estimate will have an unbounded influence function and be sensitive to outliers. Therefore we would like to choose Ψ so that it is less sensitive to outliers than $\mathrm{s}(x, \boldsymbol{\theta})$. A common choice is Huber's Ψ function,

$$\Psi(\mathbf{x}) = \mathbf{x} \min\left(1, \frac{c}{||\mathbf{x}||}\right)$$

This choice transforms each point outside a p dimensional sphere to the nearest point on it, and leaves values inside unchanged.

We now turn to the question of the properties we desire of the M-estimator that solves (4.2.2). One common choice is to find an estimator that is optimal in some sense, but the influence of an observation x is bounded. The optimality criterion is usually minimizing the trace of the asymptotic variance $\mathbf{V}$, subject to a bound on the influence function, or gross error sensitivity (Hampel et al. 1986, pg. 238).

This leads to the idea of optimal B-robust estimators (Hampel et al. 1986, pg. 238). These estimators are constructed to be as similar as possible to the MLE at nonoutlying points, but downweight possible outliers. These estimators are found by an iterative procedure, which find optimal estimators subject to Fisher consistency and bounded IF. Algorithms are given by Hampel et al. (1986, pg. 247–252) and Victoria-Feser and Ronchetti (1994).

There may be cases where minimizing the asymptotic variance may be difficult to achieve. Field and Smith (1994) give an alternate approach, where the robust estimators are derived by downweighting the ML score function, but on the probability scale. It leads to estimators that are Fisher consistent and asymptotically normal, and should be fairly close to optimal.

4.3 Robust Estimation: Linear Models

We now use some of the above results to study linear models, which is our main focus. As in previous chapters, we assume the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and each row of $\mathbf{X}$ is distributed as $\mathrm{N}(\mathbf{0}, \Sigma_X)$.

4.3.1 M-Estimators

Following Hampel et al. (1986, pg. 315), we define an M-estimator $\hat{\boldsymbol{\beta}}_M$ implicitly by the vector equation

$$\sum_{i=1}^{n} \eta\left(\mathbf{x}_i, \frac{y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_M}{\sigma}\right)\mathbf{x}_i = \mathbf{0} \qquad (4.3.1)$$

with various conditions placed on the function η. This also assumes that $n > p$. This is similar to (4.2.2) except η depends on $\hat{\boldsymbol{\beta}}_M$ only through the residual $\hat{\epsilon}_i$.

The function η is usually chosen of the form

$$\eta(\mathbf{x}, r) = w(\mathbf{x})\Psi(rv(\mathbf{x})) \tag{4.3.2}$$

where $r = y - \mathbf{x}'\hat{\boldsymbol{\beta}}_M$. Hampel et al. (1986, pg. 315) derive many of their results assuming a model with random $\mathbf{x}_i$ for convenience, which is exactly the model which we are considering.

By making the assumption that $\sigma = 1$ we can use the multiparameter results to derive the influence function for the M-estimator in the multiple linear regression model.

It can be shown that the M-estimators are consistent and asymptotically normal with covariance matrix

$$\mathbf{V} = \mathbf{M}^{-1}\mathbf{Q}\mathbf{M}^{-1}, \text{ where } \mathbf{M} = \mathrm{E}\left(\frac{\partial\eta(\mathbf{x}, r)}{\partial r}\mathbf{x}\mathbf{x}'\right), \ \mathbf{Q} = \mathrm{E}(\eta^2(\mathbf{x}, r)\mathbf{x}\mathbf{x}') \ .$$

(Hampel et al. 1986, pg. 317). We can then define the gross error sensitivities as

$$\gamma^* = \sup_{\mathbf{x},y} |\eta|(||\mathbf{M}^{-1}\mathbf{x}||) \quad \text{(unstandardized)}$$

$$\gamma_s^* = \sup_{\mathbf{x},y} |\eta|(\mathbf{x}'\mathbf{Q}^{-1}\mathbf{x})^{-1/2} \quad \text{(standardized)}$$

The form of η is motivated by the criterion used to choose the estimators. Focus is again on minimizing the asymptotic variance, subject to a bound on the gross-error sensitivity. This is often referred to as bounded influence regression (Krasker and Welsch 1982), with various forms for the estimators given by Hampel et al. (1986, pgs. 319–321).

In linear regression we would like to bound the influence of outliers in both the $\mathbf{X}$ and the residual spaces. Hampel et al. (1986, pg. 313) refers to this as bounding the influence of the residual and the influence of position in factor space. Some proposed methods only bound the influence in the residuals, some bound the effects separately, *i.e.* large downweighting of an influential $\mathbf{x}$ regardless of the size of the residual, while others downweight leverage values only if the corresponding residual is large.

Most of these results depend on the asymptotic variance of the estimators, which is a difficult issue for us, since we have $n < p$, and we do not want n increasing independently of p. Huber (1973) studied asymptotic results for $n \rightarrow \infty$ and increasing p at various rates, but did not deal specifically with a situation in which we would be interested, such as $\lim_{n,p \rightarrow \infty} p/n = m \neq 0$, where m is a constant.

Computationally the solution of (4.3.1) is found using iterative reweighted least squares. This means a starting value is needed for the estimators. One choice would be the least squares estimator (which is not applicable in underdetermined models), but a better choice is one with a high breakdown point (Ronchetti, Field, and Blanchard 1997).

Equation (4.3.1) assumes that σ^2 is known. In practice it will also have to be estimated. Two of the usual methods used are to solve (4.3.1) simultaneously with a supplementary equation for σ, or to use the median absolute deviation estimator

$$\hat{\sigma}_{mad} = b \text{ median}(|y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_M|) \tag{4.3.3}$$

where $b = 1.345$ is often chosen to achieve approximately 95% efficiency at the correct model Marazzi (1993, pg. 54–55). We will discuss the estimation of σ^2 later in the chapter.

The choice of the value of c in Huber's Ψ function affects efficiency. Typically c is chosen such that 95% efficiency is obtained at the correct model. The efficiency is usually defined as the ratio of the asymptotic variances of the estimators. This poses a problem in the underdetermined case, as mentioned above. Instead, we will use simulation results at the true model to help determine the value of c to use.

4.4 Robust Ridge Regression

There have been a few proposals to combine the properties of robust estimators with ridge estimators, principally motivated by multicollinearity problems combined with long-tailed error distributions.

Askin and Montgomery (1980) approach the problem as follows. They propose using the solution to:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)$$

subject to $\boldsymbol{\beta}'\boldsymbol{\beta}/2 \leq d^2$, where ρ is a robust function which we want to minimize.

If $\Psi(t) = \rho'(t)$ we need to find $\boldsymbol{\beta}$ to solve

$$-\mathbf{X}'\Psi(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta} = \mathbf{0}$$

or, by defining

$$\mathbf{W} = \mathrm{diag}(w_i, \ldots w_n), \quad w_i = \frac{\Psi(y_i - \mathbf{x}_i'\boldsymbol{\beta})}{y_i - \mathbf{x}_i'\boldsymbol{\beta}}$$

we can say

$$-\mathbf{X}'\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta} = \mathbf{0}\ .$$

Askin and Montgomery (1980) reformulate the problem as an augmented least squares problem, as discussed in chapter 2. We form

$$\mathbf{X}_{aug} = \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix} \quad \mathbf{y}_{aug} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}$$

so the equation we need to solve is

$$-\mathbf{X}_{aug}' \begin{pmatrix} \mathbf{W} & \mathbf{0} \\ \mathbf{0}' & \mathbf{I} \end{pmatrix} (\mathbf{y}_{aug} - \mathbf{X}_{aug}\boldsymbol{\beta}) = \mathbf{0}$$

This means the weights on the augmented observations are set to 1. This leads to a weighted ridge regression estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} \tag{4.4.1}$$

The Ψ functions they suggest are based only on outliers in the residuals, not on influential $\mathbf{x}$ observations.

They estimate λ using

$$\hat{\lambda} = \frac{p\hat{\sigma}^2}{\sum_i \alpha_i^2}$$

where $\hat{\sigma}^2$ is a nonrobust estimator and α_i is defined in section 2.3.4. Lawrence and Marsh (1984) use a similar estimator in their analysis of U.S. coal mining data.

Pfaffenberger and Dielman (1990) suggest another robust alternative. They propose

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \tag{4.4.2}$$

where

$$\hat{\lambda} = ps_L^2/\hat{\boldsymbol{\beta}}_L'\hat{\boldsymbol{\beta}}_L, \quad s_L^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_L)'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_L)}{n-p}$$

where $\hat{\boldsymbol{\beta}}_L$ is the least absolute value estimator. So they are not downweighting outliers, but a robust estimator is being used to find the smoothing parameter. This method also does not take influential $\mathbf{x}$ values into account. The simulation results of Pfaffenberger and Dielman (1990) suggest that (4.4.2) performs significantly better than (4.4.1).

Silvapulle (1991) takes perhaps the most formal approach to the problem. By using the transformation given in the derivation of the principal components estimator in section 2.3.4, Silvapulle (1991) first constructs an M-estimator for $\boldsymbol{\alpha} = \mathbf{P}\boldsymbol{\beta}$, which we write as $\hat{\boldsymbol{\alpha}}_M$. This relies on having $n > p$. We then modify the M-estimator by multiplying by a ridge, or smoothing matrix:

$$\hat{\boldsymbol{\alpha}}_{M,\lambda} = \mathbf{V}'(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}\mathbf{V}\hat{\boldsymbol{\alpha}}_M \ . \tag{4.4.3}$$

Silvapulle (1991) uses the asymptotic properties of M-estimators to develop two methods for choosing λ. The resulting choices are

$$\hat{\lambda}_a = \frac{p\hat{a}^2}{||\hat{\boldsymbol{\alpha}}||^2}, \qquad \hat{\lambda}_b = \frac{p\hat{a}^2}{\sum_i \lambda_i \hat{\alpha}_{Mi}^2}$$

where

$$(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, a^2\Lambda^{-1}), \qquad a = \frac{s_o^2 \mathrm{E}[\Psi^2(\epsilon/s_o)]}{\mathrm{E}[\Psi'(\epsilon/s_o)]}$$

and $s_o = \mathrm{plim}(s)$, where s is an estimator of σ. The value a^2 can be estimated using

$$\hat{a}^2 = s^2 \left[\frac{\sum_i \Psi^2(r_i/s)}{n-p} \Big/ \frac{\sum_i \Psi'(r_i/s)}{n} \right]$$

with r_i being the *ith* observed residual. Silvapulle (1991) uses a nonrobust estimator as s^2; there does not seem to be any reason why it could not be replaced by a robust estimator, such as (4.3.3). Simulation results show substantial reduction in MSE when this estimator is compared to the ordinary ridge regression estimator when the errors have long tails, with only a small increase in MSE when the errors are normally distributed.

4.5 Robust Ridge Regression: Downweighting of Residuals

The previous methods just described focus on the use of M-estimators being modified for ridge regression. We will take the same approach in developing our estimators. However, these methods find $\hat{\lambda}$ using methods that are not applicable in underdetermined models, as pointed out in chapter 2. Therefore we will need to use a different method for choosing λ. It will still depend on the value of our signal-to-noise ratio. As in the above methods, we will begin by focusing on outliers that arise from contaminated errors, with a fixed value of q.

4.5.1 Bounding Influence of Residuals

We are using the model (3.2.1) and its assumptions, as given in section 3.2. Our procedure will be similar to that of Askin and Montgomery (1980), in that we will find an estimator to solve

$$\sum_{i=1}^{n} \Psi_c\left(\frac{r_i}{\hat{\sigma}}\right) \mathbf{x}_i = \mathbf{0} \tag{4.5.1}$$

subject to our signal-to-noise constraint (3.2.2), where

$$\Psi_c\left(\frac{r_i}{\hat{\sigma}}\right) = r_i \min\left(1, \frac{c}{|r_i/\hat{\sigma}|}\right)$$

and $\hat{\sigma}$ is a robust estimator of σ.

With q fixed, the procedure will be as follows:

1. Find $\hat{\lambda}_{lik}$, $\hat{\boldsymbol{\beta}}_{lik}$ and $\hat{\sigma}^2_{lik}$ using the procedure described in section 3.2. This means we are starting with non-robust estimators.

2. For given $\hat{\lambda}_{lik}$ in step 1, find $\hat{\boldsymbol{\beta}}_{M,\lambda}$ which solves

$$\sum_{i=1}^{n} \Psi_c\left(\frac{r_i}{\hat{\sigma}}\right) \mathbf{x}_i = \mathbf{0}$$

subject to (3.2.2). This estimator is found using iterative reweighted least squares, and at each iteration is of the form

$$\hat{\boldsymbol{\beta}}_{M,\lambda} = (\mathbf{X}'\mathbf{W}\mathbf{X} + \hat{\lambda}_{lik}\Sigma_X)^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

and

$$\mathbf{W} = \text{diag}(w_1, \ldots, w_n), \quad w_i = \min\left(1, \frac{c}{|r_i/\hat{\sigma}|}\right) \tag{4.5.2}$$

3. Using $\hat{\boldsymbol{\beta}}_{M,\lambda}$ update $\hat{\lambda}$ to a more robust estimator by finding λ to solve

$$\frac{\hat{\boldsymbol{\beta}}'_{M,\lambda}\Sigma_X\hat{\boldsymbol{\beta}}_{M,\lambda}}{\hat{\sigma}^2} = q \, . \tag{4.5.3}$$

Call the solution $\hat{\lambda}_m$.

4. Use $\hat{\lambda}_m$ and the $\mathbf{W}$ matrix from the final step of the iteration scheme in step 2 to find

$$\hat{\boldsymbol{\beta}}_m = (\mathbf{X}'\mathbf{W}\mathbf{X} + \hat{\lambda}_m\Sigma_X)^{-1}\mathbf{X}'\mathbf{W}\mathbf{y},$$

and to update $\hat{\sigma}^2$. Note that we have not yet specified the estimator we are using for $\hat{\sigma}^2$. This will be discussed in the simulation results.

The choice of c in (4.5.2) will be determined by simulation at the correct model, at which we will try to achieve as high efficiency as possible, based on one of two measures described in the next section. Note that as $c \to \infty$ we find $w_i \to 1$, so we would expect to approach the non-robust estimator.

An important question we need to address is whether the above method bounds the influence of the residuals. We will rely on some empirical evidence, to be presented in the next section, to support the claim that the influence function of the procedure is bounded.

4.5.2 Empirical Studies

The models we will study are similar to those introduced in section 3.3 and subsequent sections. We will use $n = 50$ observations and $p = 80$ parameters, and generate $\boldsymbol{\beta}$ from either a U(0, 1) or a N(0, 4) distribution. The rows of the $\mathbf{X}$ matrix will be generated from the AR(2) process (3.3.1) and, in the first models, we will correctly assume that $\epsilon \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. After discussing results on the behavior of the influence function, and the efficiency of the procedure at the correct model, we will study our robust procedure by generating errors from a contaminated normal distribution.

We will look at two measures to determine efficiency. We have focused on predictive ability in previous results, so we will begin by considering a measure of the MSE of the fitted values:

$$\mathrm{MSE}(\hat{\mathbf{y}}) = \frac{1}{k}\sum_{i=1}^{k}(\hat{\mathbf{y}}_i - \mathbf{X}\boldsymbol{\beta})'(\hat{\mathbf{y}}_i - \mathbf{X}\boldsymbol{\beta})$$

where k is the number of simulated data sets. However, this is a nonrobust measure, and will penalize methods that fit the majority of the data well, but do not fit the outlying points. A more robust measure is

$$\mathrm{MSE}_r(\hat{\mathbf{y}}) = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n}\rho(\hat{e}_{i,j}) \tag{4.5.4}$$

where $\rho(\hat{e}_{i,j}) = \min(\hat{e}^2_{i,j}, d^2\hat{\sigma}^2_{\hat{e},i})$,

$$\hat{\sigma}_{\hat{e},i} = 1.483 \text{ med }_l|\hat{e}_{i,l} - \text{ med }_m(\hat{e}_{i,m})|$$

is the median absolute deviation of the prediction errors and $\hat{e}_{i,j}$ is the *jth* element of $(\hat{\mathbf{y}}_i - \mathbf{X}\boldsymbol{\beta})$. This robust criterion is suggested Ronchetti et al. (1997) in the context of robust model selection. The choice of $d = 1.345$ is often made, and is the value we will use in our procedure.

Using (4.5.4) we then define efficiency as

$$\mathrm{eff}(\hat{\mathbf{y}}_m) = \frac{\mathrm{MSE}_r(\hat{\mathbf{y}}_{lik})}{\mathrm{MSE}_r(\hat{\mathbf{y}}_m)} \tag{4.5.5}$$

In the case of studying a robust procedure, it may make more sense to change our focus to now look at distance from the true $\boldsymbol{\beta}$ vector. We begin with finding the approximate MSE of our $\hat{\boldsymbol{\beta}}$ vector:

$$\text{MSE}(\hat{\boldsymbol{\beta}}) = \frac{1}{k}\sum_{i=1}^{k}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}) \tag{4.5.6}$$

and defining efficiency as

$$\text{eff}(\hat{\boldsymbol{\beta}}_m) = \frac{\text{MSE}(\hat{\boldsymbol{\beta}}_{lik})}{\text{MSE}(\hat{\boldsymbol{\beta}}_m)} \tag{4.5.7}$$

We will also examine the predictive ability of the estimators in the following fashion. For each data set generated, we will calculate $\hat{\boldsymbol{\beta}}_m$ and $\hat{\boldsymbol{\beta}}_{lik}$, then generate a validation set $\mathbf{y}_1 = \mathbf{X}\boldsymbol{\beta} + \epsilon_1$ where ϵ_1 will follow the same distribution as ϵ. Then we will evaluate the robust prediction sum of squares using the function defined in (4.5.4):

$$PSS_{lik} = \frac{1}{n}\sum_{i=1}^{n}\rho(\hat{e}_i) \tag{4.5.8}$$

where $\hat{e}_i$ is the *ith* element of $\mathbf{y}_1 - \mathbf{X}\hat{\boldsymbol{\beta}}_{lik}$, and

$$PSS_r = \frac{1}{n}\sum_{i=1}^{n}\rho(\hat{e}_i) \tag{4.5.9}$$

where $\hat{e}_i$ is the *ith* element of $\mathbf{y}_1 - \mathbf{X}\hat{\boldsymbol{\beta}}_m$.

We begin by presenting some results that compare the robust and nonrobust estimation procedure, when we use $\hat{\sigma}^2_{mad}$ as our robust estimator of σ^2. Table 4.1 gives results over 500 simulated data sets for four values of q.

There are two noticeable features present in table 4.1. First, we see that $\hat{\sigma}^2_{mad}$ is biased downwards by a greater amount than $\hat{\sigma}^2_{lik}$. More importantly, the $\hat{\lambda}_m$ values are much more variable than the $\hat{\lambda}_{lik}$ values. This difference is of several orders of magnitude at large values of q. Other simulation results suggest that using $\hat{\sigma}^2_{mad}$, and letting $w_i = 1$ for all i in our estimation procedure, can still give results that are quite different from the nonrobust procedure. For these reasons, we will use a different robust estimate of σ^2, which we will introduce below.

		q			
		7.21	17.37	27.14	57.44
	σ^2	16	25	16	7.56
$\hat{\sigma}^2_{mad}$	mean	7.126	9.85	6.523	3.373
	median	7.163	9.791	6.538	3.384
	var	0.559	1.106	0.423	0.0542
$\hat{\sigma}^2_{lik}$	mean	10.79	14.42	9.317	4.457
	median	10.67	14.33	9.268	4.439
	var	1.236	1.841	0.492	0.0656
$\hat{\lambda}_m$	mean	20.00	12.92	10.09	6.571
	median	20.16	12.84	10.01	6.648
	var	8.826	4.469	2.734	0.8627
$\hat{\lambda}_{lik}$	mean	5.966	2.639	1.735	0.8446
	median	5.988	2.646	1.739	0.8451
	var	0.016	1.39×10^{-3}	3.01×10^{-4}	1.32×10^{-5}
b		1.345	1.345	1.345	1.345
c		10	5	3	3

Table 4.1: Results comparing nonrobust and robust procedures: Correct model situation, using $\hat{\sigma}^2_{mad}$ as variance estimator

4.5.3 Improved Estimation of σ^2

The nonrobust estimator $\hat{\sigma}^2_{lik}$ given in equation (3.2.13) is the sum of two terms. Since one of the terms is a sum of squares, it is reasonable to derive a robust estimator by replacing this term with a more robust loss function. We use the ρ-function introduced in (4.5.4) to construct

$$\hat{\sigma}^2_m = \frac{1}{n} \sum_{w=1}^{n} \rho(\hat{e}_w) + \frac{\lambda}{n} \hat{\boldsymbol{\beta}}'_m \Sigma_X \hat{\boldsymbol{\beta}}_m \tag{4.5.10}$$

where $\hat{e}_w = y_w - \mathbf{x}'_w \hat{\boldsymbol{\beta}}_m$.

Results using this new estimator are given in table 4.3, and we discuss them briefly here. We see that the $\hat{\sigma}^2_m$ values are less biased than those given by $\hat{\sigma}^2_{mad}$ in table 4.1 with comparable variance. We also see that the smoothing parameters are much less variable when we use $\hat{\sigma}^2_m$.

We will use this new estimator in our procedure outlined in the previous section. We will begin with some empirical results on the behavior of the influence functions of our robust and nonrobust estimates.

4.5.4 Empirical Measure of Influence Functions

We began this chapter with a discussion of the influence function of an estimator and the role it plays in identifying if the estimator will be strongly affected by one outlying value in the data set. Since we have proposed a robust method, we would like to establish, at least empirically, that it has a bounded influence function. We will do this by the following simulation study.

We generate data from the regression model as before, and calculate our estimates by the robust and nonrobust methods we have developed. Then we multiply one value in $\mathbf{y}$ by a value $s > 0$, and recalculate our estimates. This is repeated for a range of increasing values of s. If the estimators increase greatly, we have empirical evidence that their influence functions are not bounded, since they are susceptible to changes in one data point. Table 4.2 contains results for the $q = 7.21$ model using 50 simulated data sets, where s ranges from 1 to 100. Results for other models are given in Appendix C.

We see that $\hat{\sigma}^2_{lik}$ and $\hat{\boldsymbol{\beta}}_{lik}$ change dramatically as we increase a single y_i value, giving strong evidence that the influence functions are unbounded. However, we see that $\hat{\sigma}^2_m$ and $\hat{\boldsymbol{\beta}}_m$ continue to be stable over the range of s values, suggesting their influence functions are bounded. The weights that are assigned to the outlying $\mathbf{y}$ values decrease quickly to 0 as s increased, providing evidence that the robust method is giving us sensible estimates for the bulk of the data. We also see that the $\hat{\lambda}$ values are not greatly affected.

This empirical study does not prove that the influence functions of our robust estimators are bounded. To establish it theoretically we would have to derive the influence function of $\hat{\boldsymbol{\beta}}_m$ using (4.2.1), under the assumption that σ^2 was known and λ was fixed. It may also be possible to begin with the results on the influence

		s				
		1	20.8	40.6	60.4	100
$\hat{\sigma}^2_m$	mean	11.883	14.020	14.162	14.211	14.279
	median	11.923	13.352	13.944	14.103	14.096
$\hat{\sigma}^2_{lik}$	mean	14.000	118.47	420.89	921.26	2515.8
	median	13.132	50.961	163.92	351.08	947.97
$\hat{\lambda}_m$	mean	6.033	6.085	6.087	6.089	6.091
	median	6.088	6.116	6.117	6.120	6.119
$\hat{\lambda}_{lik}$	mean	5.689	5.421	5.329	5.295	5.272
	median	5.686	5.485	5.381	5.349	5.354
$(\hat{\beta}_m - \beta)'(\hat{\beta}_m - \beta)$	mean	67.604	84.222	85.285	85.641	86.103
	median	66.203	83.672	84.240	84.531	85.034
$(\hat{\beta}_{lik} - \beta)'(\hat{\beta}_{lik} - \beta)$	mean	84.091	981.24	3633.9	8041.8	2.21×10^4
	median	80.339	389.89	1370.9	3015.0	8291.9
c		2	2	2	2	2

Table 4.2: Empirical measure of change in influence function when one y_i value modified. Model using $q = 7.21$ and robust method which downweights residuals

function of the M-estimator given earlier in the chapter, although the presence of the smoothing parameter in $\hat{\beta}_m$ may mean these results are no longer applicable.

4.5.5 Simulation Results

Now that we have empirically established that our robust estimation procedure has a bounded influence function, we move on to comparing the estimators obtained by the two methods. We begin by simulating 500 data sets at the correct model, and calculating eff($\hat{\mathbf{y}}_m$) and eff($\hat{\beta}_m$). For each data set generated we use the estimators to predict a new data set, using the robust measure defined in (4.5.4). These results are presented in table 4.3.

Several general trends can be seen in table 4.3. As we had hoped, the values of $\hat{\lambda}_m$ are much less variable using $\hat{\sigma}^2_m$ defined in (4.5.10). The $\hat{\lambda}$ values are very similar for both the robust and nonrobust methods. The estimates of σ^2 are also very similar.

		q			
		7.21	17.37	27.14	57.44
$\hat{\sigma}^2_m$	mean	9.930	11.803	8.640	3.707
	median	9.862	11.677	8.622	3.691
	var	1.222	1.328	0.426	0.0494
$\hat{\sigma}^2_{lik}$	mean	10.789	14.417	9.317	4.457
	median	10.672	14.326	9.268	4.439
	var	1.236	1.562	0.492	0.066
$\hat{\lambda}_m$	mean	6.252	2.725	1.773	0.854
	median	6.281	2.729	1.778	0.856
	var	0.020	1.99×10^{-3}	4.23×10^{-4}	2.10×10^{-5}
$\hat{\lambda}_{lik}$	mean	5.966	2.639	1.735	0.845
	median	5.989	2.646	1.739	0.845
	var	0.016	1.64×10^{-3}	3.01×10^{-4}	1.32×10^{-5}
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	44.494	213.83	205.46	193.63
	median	43.326	211.98	203.64	192.95
	var	27.876	152.25	90.430	31.445
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	46.124	219.92	209.59	198.18
	median	44.912	217.09	208.06	197.64
	var	30.444	177.27	102.58	39.890
eff($\hat{\boldsymbol{\beta}}_m$)		1.037	1.028	1.020	1.023
eff($\hat{\mathbf{y}}_m$)		0.970	.906	0.946	0.859
PSS_r	mean	20.189	35.574	22.175	11.579
PSS_{lik}	mean	19.656	32.913	21.249	10.163
c		1	0.5	0.5	0.3

Table 4.3: Results comparing robust and nonrobust procedures: Correct model situation, using robust method which downweights residuals

We see, at the correct model, that our robust estimators are achieving around 100% efficiency, using criterion (4.5.7), while the efficiency ranges from 86% to 97% using criterion (4.5.5). If we compare the PSS values found using (4.5.8) and (4.5.9) we see the nonrobust method does better in all four models, with the best performance in the $q = 57.44$ model.

We will now look at the case where model (3.2.1) is not correct, specifically in terms of the error structure. The errors will be generated from a "wild" distribution, where 90% of the ϵ_i are generated from a $\mathrm{N}(0,\ \sigma^2)$ distribution, and the remaining 10% from a $\mathrm{N}(0,\ 25\sigma^2)$ distribution. Before discussing the results, let us examine how these new errors affect the true signal-to-noise ratio. Recall that

$$q = \frac{\mathrm{Var(signal)}}{\mathrm{Var(noise)}} = \frac{\mathrm{Var}(\mathbf{x}_i'\boldsymbol{\beta})}{\mathrm{Var}(\epsilon_i)}$$

but we no longer have $\mathrm{Var}(\epsilon_i) = \sigma^2$. If we define

$$\epsilon_i = \begin{cases} z_i \sim \mathrm{N}(0, \sigma^2) & \text{w.p. } p \\ t_i \sim \mathrm{N}(0, k\sigma^2) & \text{w.p. } (1-p) \end{cases}$$

Then $\mathrm{E}(\epsilon_i) = 0$ and

$$\begin{aligned} \mathrm{Var}(\epsilon_i) &= \mathrm{E}(\epsilon_i^2) \\ &= \mathrm{E}[\mathrm{E}(\epsilon_i^2 | z_i, t_i)] \\ &= \mathrm{E}[p z_i^2 + (1-p) t_i^2] \\ &= p\sigma^2 + (1-p)k\sigma^2 = \sigma^2[p + k(1-p)] \end{aligned}$$

So the "true" ratio is

$$\frac{\boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta}}{\sigma^2[p + k(1-p)]}$$

If $p = 1$ then the denominator reduces to σ^2, as before.

In our specific example, $p = 0.9$ and $k = 25$, so $[p + k(1-p)] = 3.4$. This means that the true ratio is smaller than what we are assuming, and is the reason we are studying some cases with larger values of $\boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta}/\sigma^2$. All other features of the models

		q			
		7.21	17.37	27.14	57.44
$\hat{\sigma}^2_m$	mean	12.683	13.029	9.461	3.851
	median	12.530	13.066	9.397	3.871
	var	5.973	2.589	0.974	0.118
$\hat{\sigma}^2_{lik}$	mean	15.195	18.352	11.121	4.923
	median	14.781	17.780	10.803	4.881
	var	17.162	10.762	2.717	0.276
$\hat{\lambda}_m$	mean	6.107	2.697	1.758	0.852
	median	6.115	2.701	1.760	0.853
	var	0.033	3.02×10^{-3}	6.95×10^{-4}	3.20×10^{-5}
$\hat{\lambda}_{lik}$	mean	5.701	2.591	1.717	0.841
	median	5.749	2.597	1.720	0.842
	var	0.052	3.78×10^{-3}	7.50×10^{-4}	3.52×10^{-5}
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	69.938	246.27	237.69	208.85
	median	67.936	243.17	233.14	206.65
	var	265.94	1133.6	869.91	301.07
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	86.661	293.68	261.58	225.49
	median	80.014	285.09	254.95	222.26
	var	916.37	3991.1	2049.8	616.79
eff($\hat{\boldsymbol{\beta}}_m$)		1.239	1.193	1.101	1.080
eff($\hat{\mathbf{y}}_m$)		1.009	0.907	0.948	0.854
PSS_r	mean	28.734	49.548	31.457	16.502
PSS_{lik}	mean	28.931	46.997	30.499	14.707
c		1	0.5	0.5	0.3

Table 4.4: Results comparing robust and nonrobust procedures: Wild error distribution, using robust method which downweights residuals

are the same as used to generate the results in table 4.3. Table 4.4 summarizes results over 500 simulated data sets.

We begin with some of the general features of the estimators. We see that the $\hat{\sigma}^2_m$ values tend to be smaller than the $\hat{\sigma}^2_{lik}$ values, while the $\hat{\lambda}_m$ values tend to be larger than their nonrobust counterparts. The smoothing parameters are still very similar.

We now focus on the gains in using the robust estimator, beginning with (4.5.7). We see we have large gains in efficiency, on average, in each of our models. The improvements are larger when the true value of q is smaller. The results using efficiency measure (4.5.5) differ slightly from these results. We are obtaining anywhere from 85% to 101% efficiency, with the loss in efficiency greater as q increases. If we compare how well we do in predicting a new data set, we see the PSS_{lik} values are smaller for three of the four choices of q.

In addition to considering the overall gain in efficiency, we can investigate differences in individual data sets. We can do this by comparing

$$\begin{aligned} SS(\hat{\boldsymbol{\beta}}_m) &= (\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}) \\ SS(\hat{\boldsymbol{\beta}}_{lik}) &= (\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta}) \end{aligned}$$

and by comparing the PSS_r and PSS_{lik} values for each simulated set. Table 4.5 gives the percentage of cases that fall within each of the four possibilities, based on our two comparison criteria.

We see that $SS(\hat{\boldsymbol{\beta}}_m) < SS(\hat{\boldsymbol{\beta}}_{lik})$ in virtually all situations, so this favours the robust procedure. This is not the case for the PSS measures. Here we see the nonrobust method is preferred in a majority of cases at each $q > 7.21$, with the percentage of cases favouring the nonrobust method increasing with q. We are not sure why there is this difference in performance in the two measures, but it is repeated in some of the later results.

		$PSS_{lik} < PSS_r$	$PSS_{lik} > PSS_r$
$q = 7.21$	$SS(\hat{\boldsymbol{\beta}}_m) > SS(\hat{\boldsymbol{\beta}}_{lik})$	1	0
	$SS(\hat{\boldsymbol{\beta}}_m) < SS(\hat{\boldsymbol{\beta}}_{lik})$	43	56
$q = 17.37$	$SS(\hat{\boldsymbol{\beta}}_m) > SS(\hat{\boldsymbol{\beta}}_{lik})$	2	0
	$SS(\hat{\boldsymbol{\beta}}_m) < SS(\hat{\boldsymbol{\beta}}_{lik})$	65	33
$q = 27.14$	$SS(\hat{\boldsymbol{\beta}}_m) > SS(\hat{\boldsymbol{\beta}}_{lik})$	2	0
	$SS(\hat{\boldsymbol{\beta}}_m) < SS(\hat{\boldsymbol{\beta}}_{lik})$	64	34
$q = 57.44$	$SS(\hat{\boldsymbol{\beta}}_m) > SS(\hat{\boldsymbol{\beta}}_{lik})$	2	0
	$SS(\hat{\boldsymbol{\beta}}_m) < SS(\hat{\boldsymbol{\beta}}_{lik})$	77	21

Table 4.5: Percentage of cases in which robust and nonrobust methods preferred, based on two different criteria. Wild error distribution, using robust method which downweights residuals

4.6 Robust Ridge Regression: Proposal Two

We return to the M-estimator, defined as the solution to (4.3.1), with η defined in (4.3.2). The terms $w(\mathbf{x})$ and $v(\mathbf{x})$ allow us to bound the influence of values in the $\mathbf{X}$ space. These are often referred to as influential observations.

The usual approach to bounding influence in the $\mathbf{X}$-space is similar to that of bounding the influence of residuals; downweight $\mathbf{x}$ values that are far from the origin in some sense. This typically means using weights that are decreasing functions of the distances $||\mathbf{Ax}||$ and $\mathbf{A}$ is a transformation matrix that satisfies

$$\frac{1}{n}\sum_{i=1}^{n} u(||\mathbf{Ax}_i||)\mathbf{Ax}_i\mathbf{x}_i'\mathbf{A}' = \mathbf{I}$$

where u is a given function. The matrix $\mathbf{A}$ usually is determined by an iterative procedure (Marazzi 1993, pg. 56).

These approaches have a drawback in underdetermined models. This is because it is more difficult to discuss the concept of distance, since we have n vectors in a p dimensional space, where $p > n$. Many of the methods for determining $\mathbf{A}$ involve matrices that use $(\mathbf{X}'\mathbf{X})^{-1}$, which is singular in the underdetermined model. Therefore we will attempt to use a method similar to one given by Krasker and Welsch (1982).

If $n > p$ we define the hat matrix as $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and its diagonal elements h_i as the leverage values. From regression theory we know that $0 \leq h_i \leq 1$ (Neter et al. 1985, pg. 402), and large leverage values are associated with points that would be suspected to be influential. Based on the leverage values, Krasker and Welsch (1982) suggest finding estimators which minimize

$$\sum_{i=1}^{n} \sigma^2 v^2(\mathbf{x}_i)\rho[(y_i - \mathbf{x}_i'\beta)/(\sigma v(\mathbf{x}_i)]$$

where $v(\mathbf{x}_i) = v_i = (1 - h_i)^{1/2}$. If we choose ρ such that Huber's Ψ function is its derivative, we seek an estimator which solves

$$\sum_i r_i w_i \mathbf{x}_i = 0$$

where r_i is the *ith* residual and $w_i = \min\{1, c/|r_i/(\sigma v_i)|\}$.

We will use a procedure based on this method to find our robust smoothed estimator, with the following modification. We will use

$$\mathbf{H}_\lambda = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\Sigma_X)^{-1}\mathbf{X}'$$

and define our leverage values to be the diagonal elements of $\mathbf{H}_\lambda$.

The use of these weights will provide a bound on the influence of $\mathbf{x}$, but its effect will be cancelled out if $|r_i/(\sigma v_i)|$ is small. This is similar to the Hampel-Krasker estimator, which downweights leverage values only if the corresponding residual is large (Hampel et al. 1986, pg. 322).

This leads to a new robust estimation procedure for our model. It is identical to the method outlined in section 4.5.1, except we replace the weights in (4.5.2) with

$$w_i = \min\left(1, \frac{c}{|r_i/(v_i\hat{\sigma})|}\right) \qquad (4.6.1)$$

We begin with an empirical examination of the influence functions using this new procedure. Since we are now downweighting in the $\mathbf{X}$ and residual spaces, we need to examine if the influence function of our procedure is bounded in both spaces.

		s				
		1	20.8	40.6	60.4	100
$\hat{\sigma}^2_m$	mean	9.237	9.917	10.042	10.1388	10.292
	median	9.129	9.918	10.075	10.131	10.267
$\hat{\sigma}^2_{lik}$	mean	10.789	168.12	625.58	1383.0	3798.0
	median	10.673	92.237	333.15	728.54	2002.1
$\hat{\lambda}_m$	mean	6.240	6.228	6.229	6.231	6.235
	median	6.254	6.251	6.251	6.251	6.259
$\hat{\lambda}_{lik}$	mean	5.966	5.340	5.253	5.223	5.200
	median	5.988	5.353	5.292	5.281	5.263
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	43.441	47.447	47.520	47.650	47.892
	median	42.846	45.861	46.243	46.370	46.618
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	46.124	1349.1	5209.0	1.11×10^4	3.21×10^4
	median	44.912	779.33	2921.1	6470.7	1.79×10^4
c		2	2	2	2	2

Table 4.6: Empirical measure of change in influence function when one value in $\mathbf{y}$ modified, using second robust method, $q = 7.21$

4.6.1 Empirical Assessment of Influence Function

In this second robust estimation proposal we are concerned with a bounded influence function with respect to outliers in the $\mathbf{X}$ and ϵ spaces. Therefore we conduct two simulation studies. The first is the one described in section 4.5.4, but using the robust procedure described in this section. In the second simulation we multiply one row of $\mathbf{X}$ by a value $s > 0$, calculate the estimators, and repeat this over a range of increasing values of s. Tables 4.6 and 4.7 contain results for the $q = 7.21$ model using 50 simulated data sets. Results for other models are given in Appendix C.

We see that the robust estimators appear to have bounded influence functions in both cases. The nonrobust estimators do not appear to have bounded influence functions when we increase a value in $\mathbf{y}$, as we saw previously. However, the results suggest the nonrobust method does have a bounded influence function when we have outliers in the $\mathbf{X}$ space. This may be a consequence of adding the term $\lambda\Sigma_X$ to the $\mathbf{X}'\mathbf{X}$ matrix. Therefore, when we present results that only involve a contaminated $\mathbf{X}$

		s					
		1	20.8	40.6	60.4	80.2	100
$\hat{\sigma}^2_m$	mean	9.237	9.244	9.248	9.252	9.255	9.272
	median	9.129	9.174	9.171	9.172	9.172	9.173
$\hat{\sigma}^2_{lik}$	mean	10.789	10.817	10.818	10.818	10.818	10.818
	median	10.673	10.663	10.664	10.665	10.665	10.665
$\hat{\lambda}_m$	mean	6.240	6.267	6.267	6.267	6.268	6.268
	median	6.254	6.284	6.286	6.287	6.287	6.287
$\hat{\lambda}_{lik}$	mean	5.966	5.982	5.982	5.982	5.982	5.982
	median	5.988	6.005	6.005	6.004	6.004	6.004
$(\hat{\beta}_m - \beta)'(\hat{\beta}_m - \beta)$	mean	43.441	43.298	43.302	43.307	43.310	43.331
	median	42.846	42.552	42.542	42.539	42.537	42.536
$(\hat{\beta}_{lik} - \beta)'(\hat{\beta}_{lik} - \beta)$	mean	46.124	46.016	46.018	46.018	46.018	46.019
	median	44.912	44.836	44.820	44.814	44.811	44.809
c		2	2	2	2	2	2

Table 4.7: Empirical measure of change in influence function when one row in **X** modified, using second robust method, $q = 7.21$

matrix. we should expect the nonrobust method to continue to do well.

4.6.2 Simulation Results

We now wish to investigate the properties of this new robust procedure. We begin with an examination of its performance in a model with no outliers in the errors or influential values in **X**. The model being used is the one described in section 4.5.2. Table 4.8 summarizes some of the simulation results in the correct model situation. In all results presented in this section, the tables are a summary of findings over 200 simulated data sets.

Some general features appear in table 4.8. We continue to see $\hat{\sigma}^2_m < \hat{\sigma}^2_{lik}$ and $\hat{\lambda}_m > \hat{\lambda}_{lik}$ in three models. The variability of the $\hat{\lambda}_m$ values continues to be of the same order as the $\hat{\lambda}_{lik}$ values. We are also achieving approximately 100% efficiency with the robust estimation procedure at the correct model, based on (4.5.7).

		q			
		7.21	17.37	27.14	57.44
$\hat{\sigma}^2_m$	mean	9.237	12.640	8.296	4.071
	median	9.129	12.647	8.307	4.073
	var	1.021	1.415	0.431	0.056
$\hat{\sigma}^2_{lik}$	mean	10.789	14.417	9.317	4.457
	median	10.673	14.326	9.268	4.439
	var	1.236	1.562	0.492	0.066
$\hat{\lambda}_m$	mean	6.240	2.723	1.774	0.854
	median	6.255	2.728	1.778	0.855
	var	0.019	1.97×10^{-3}	3.89×10^{-4}	2.03×10^{-5}
$\hat{\lambda}_{lik}$	mean	5.966	2.639	1.735	0.845
	median	5.988	2.646	1.739	0.845
	var	0.016	1.64×10^{-3}	3.01×10^{-4}	1.32×10^{-5}
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	43.441	214.55	204.88	194.53
	median	42.846	211.62	202.94	193.75
	var	21.923	92.206	87.860	30.911
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	46.124	219.92	209.59	198.18
	median	44.912	217.09	208.06	197.64
	var	30.444	177.27	102.58	39.890
eff($\hat{\boldsymbol{\beta}}_m$)		1.062	1.025	1.023	1.019
eff($\hat{\mathbf{y}}_m$)		0.967	0.926	0.924	0.911
PSS_r	mean	20.585	34.821	22.727	10.974
PSS_{lik}	mean	19.656	32.913	21.249	10.163
c		2	2	2	2

Table 4.8: Results comparing robust and nonrobust procedures: Correct model situation, using second robust method

If we use the efficiency measure (4.5.5), we obtain anywhere from 91% to 97% efficiency at the true model. A comparison of the *PSS* values in table 4.8 also slightly favours the nonrobust procedure.

We now move to the cases of contaminated $\mathbf{X}$ and ϵ values. The errors are generated from the wild distribution introduced previously. We take the following approach to introducing leverage values in the $\mathbf{X}$ matrix. We generate the elements of all but two of our $\mathbf{x}_i'$ vectors using the AR(2) process (3.3.1). The remaining two rows of $\mathbf{X}$ are generated from the AR(1) process

$$x_{ij} = 0.9x_{i,j-1} + \eta_t, \quad \eta_t \sim \mathrm{N}(0, 25) \tag{4.6.2}$$

We can examine how this affects the true variance of the signal in our model. Suppose $\mathbf{x}_i' \sim \mathrm{N}(\mathbf{0}', \mathbf{\Sigma}_z)$ with probability p_1 and $\mathbf{x}_i' \sim \mathrm{N}(\mathbf{0}', \mathbf{\Sigma}_t)$ with probability $1 - p_1$. Then we find

$$\mathrm{Var}(\mathbf{x}_i') = p_1\mathbf{\Sigma}_z + (1 - p_1)\mathbf{\Sigma}_t$$

Therefore with contaminated $\mathbf{X}$ and ϵ we have a true signal-to-noise ratio of

$$\frac{\boldsymbol{\beta}'[p_1\mathbf{\Sigma}_z + (1 - p_1)\mathbf{\Sigma}_t]\boldsymbol{\beta}}{\sigma^2[p + k(1 - p)]}$$

Table 4.9 summarizes the results of the scenario in which there are no influential observations in the $\mathbf{X}$ matrix, but the errors are generated from the "wild" distribution.

The results in table 4.9 are very similar to those in table 4.4. We see that the robust method yields estimators that are better than the nonrobust estimators in all four models considered, based on eff($\hat{\boldsymbol{\beta}}_m$). The improvements are greater at smaller values of q. If we use eff($\hat{\mathbf{y}}_m$) we find the methods do not differ by as much, on average, and the advantage still lies with the nonrobust method for $q > 7.21$. The same conclusion holds if we compare the mean PSS_{lik} and PSS_r values.

Table 4.10 gives the results on the percentage of times the robust or nonrobust method has the superior performance, based on each criteria. The results are very similar to those in table 4.5; we see the robust method is favoured in most cases when

		q			
		7.21	17.37	27.14	57.44
$\hat{\sigma}^2_m$	mean	10.805	13.971	8.926	4.227
	median	10.635	13.952	8.860	4.284
	var	3.228	2.769	0.911	0.123
$\hat{\sigma}^2_{lik}$	mean	15.195	18.352	11.121	4.923
	median	14.781	17.779	10.803	4.881
	var	17.162	10.762	2.717	0.276
$\hat{\lambda}_m$	mean	6.119	2.692	1.761	0.851
	median	6.118	2.701	1.766	0.852
	var	0.029	3.19×10^{-3}	6.38×10^{-4}	3.69×10^{-5}
$\hat{\lambda}_{lik}$	mean	5.701	2.591	1.717	0.841
	median	5.749	2.597	1.720	0.842
	var	0.052	3.78×10^{-3}	7.50×10^{-4}	3.53×10^{-5}
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	59.435	250.538	233.53	211.80
	median	57.845	247.385	230.92	208.63
	var	128.19	1219.4	787.66	313.61
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	86.661	293.68	261.58	225.49
	median	80.014	285.09	254.95	222.26
	var	916.37	3991.1	2049.8	616.79
eff($\hat{\boldsymbol{\beta}}_m$)		1.458	1.172	1.120	1.065
eff($\hat{\mathbf{y}}_m$)		1.008	0.931	0.920	0.899
PSS_r	mean	28.265	48.798	32.111	15.611
PSS_{lik}	mean	28.931	46.997	30.499	14.707
c		2	2	2	2

Table 4.9: Results comparing robust and nonrobust procedures: Wild error distribution, using second robust method

comparing $SS(\hat{\beta}_m)$ and $SS(\hat{\beta}_m)$, while the nonrobust method is preferred based on comparing PSS_{lik} and PSS_r values at $q > 7.21$.

		$PSS_{lik} < PSS_r$	$PSS_{lik} > PSS_r$
$q = 7.21$	$SS(\hat{\beta}_m) > SS(\hat{\beta}_{lik})$	2	0
	$SS(\hat{\beta}_m) < SS(\hat{\beta}_{lik})$	45	53
$q = 17.37$	$SS(\hat{\beta}_m) > SS(\hat{\beta}_{lik})$	2	0
	$SS(\hat{\beta}_m) < SS(\hat{\beta}_{lik})$	61	37
$q = 27.14$	$SS(\hat{\beta}_m) > SS(\hat{\beta}_{lik})$	3	0
	$SS(\hat{\beta}_m) < SS(\hat{\beta}_{lik})$	66	31
$q = 57.44$	$SS(\hat{\beta}_m) > SS(\hat{\beta}_{lik})$	3	0
	$SS(\hat{\beta}_m) < SS(\hat{\beta}_{lik})$	71	26

Table 4.10: Percentage of cases in which robust and nonrobust methods preferred, based on two different criteria. Wild error distribution, using second robust method

We now deal with the situation where the errors are not contaminated, but we introduce two influential observations in our **X** matrix, as described previously. These results are summarized in table 4.11.

As we mentioned when discussing the influence functions of the procedures, we suspected the nonrobust method would do well in this case, because the estimators appeared to have a bounded influence function with respect to **X**. The results in table 4.11 confirm this belief. The results are very similar to those given in the correct model situation in table 4.8.

Table 4.12, which summarizes the percentage of cases in which each method is preferred in the model with two leverage values, shows us some interesting results. We continue to see $SS(\hat{\beta}_m) < SS(\hat{\beta}_{lik})$ in almost all situations. However, except for the $q = 7.21$ model, we see that $PSS_r < PSS_{lik}$ in the majority of cases, indicating the robust method does better on both criteria.

Finally, in table 4.13 we present results from the situation where we have errors generated from our "wild" distribution and there are two leverage values in the **X** matrix.

		q			
		7.21	17.37	27.14	57.44
$\hat{\sigma}^2_m$	mean	9.819	16.049	10.446	5.058
	median	9.777	16.017	10.390	5.044
	var	1.550	1.024	0.323	0.044
$\hat{\sigma}^2_{lik}$	mean	11.116	16.761	10.830	5.202
	median	11.074	16.704	10.765	5.162
	var	1.261	1.478	0.485	0.074
$\hat{\lambda}_m$	mean	6.315	2.740	1.780	0.855
	median	6.330	2.739	1.779	0.854
	var	0.019	1.07×10^{-3}	2.02×10^{-4}	9.92×10^{-6}
$\hat{\lambda}_{lik}$	mean	6.123	2.661	1.740	0.843
	median	6.148	2.667	1.741	0.843
	var	0.021	3.69×10^{-3}	8.45×10^{-4}	4.94×10^{-5}
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	40.330	155.39	143.88	131.14
	median	39.695	152.85	142.03	130.56
	var	22.423	241.75	135.41	42.857
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	43.266	157.99	145.95	132.68
	median	42.443	154.57	143.95	132.73
	var	27.750	245.96	136.35	43.748
eff($\hat{\boldsymbol{\beta}}_m$)		1.073	1.017	1.014	1.012
eff($\hat{\mathbf{y}}_m$)		0.932	0.984	0.985	0.992
PSS_r	mean	21.062	32.423	21.104	10.147
PSS_{lik}	mean	19.959	48.333	31.100	14.881
c		2	2	2	2

Table 4.11: Results comparing robust and nonrobust procedures: Two large leverages in $\mathbf{X}$, using second robust method

		$PSS_{lik} < PSS_r$	$PSS_{lik} > PSS_r$
$q = 7.21$	$SS(\hat{\beta}_m) > SS(\hat{\beta}_{lik})$	2	0
	$SS(\hat{\beta}_m) < SS(\hat{\beta}_{lik})$	74	24
$q = 17.37$	$SS(\hat{\beta}_m) > SS(\hat{\beta}_{lik})$	2	9.5
	$SS(\hat{\beta}_m) < SS(\hat{\beta}_{lik})$	10	78.5
$q = 27.14$	$SS(\hat{\beta}_m) > SS(\hat{\beta}_{lik})$	2	16
	$SS(\hat{\beta}_m) < SS(\hat{\beta}_{lik})$	11	71
$q = 57.44$	$SS(\hat{\beta}_m) > SS(\hat{\beta}_{lik})$	3	25.5
	$SS(\hat{\beta}_m) < SS(\hat{\beta}_{lik})$	11	60.5

Table 4.12: Percentage of cases in which robust and nonrobust methods preferred, based on two different criteria. Two large leverage values in **X**, using second robust method

We see that the results are very similar to those found when there were only contaminated errors in the model. The efficiency measure (4.5.7) strongly favours the robust estimators. When we examine the efficiency measure (4.5.5) we find the robust and nonrobust methods perform nearly identically on average. The mean values of PSS_r and PSS_{lik} are also very similar.

Table 4.14 summarizes the results found in the individual data sets when we use the "wild" error distribution and two leverage values. The values of PSS_{lik} and PSS_r indicate that both the robust and nonrobust methods are preferred in 45% to 55% of the individual cases, depending on q. Meanwhile $SS(\hat{\beta}_m) < SS(\hat{\beta}_{lik})$ in most cases, as we saw in all other situations.

We can draw several conclusions from these results. When we base the comparison of the procedures on measure (4.5.7) the robust procedure does as well as the nonrobust method when the model assumptions are correct, and outperforms it when the errors originate from the "wild" error distribution. The improvements are more dramatic as q decreases. The methods perform similarly if there are only influential values in the **X** space, because the influence functions of both procedures appear to be bounded in this case.

		q			
		7.21	17.37	27.14	57.44
$\hat{\sigma}^2_m$	mean	11.339	17.774	11.323	5.307
	median	11.187	17.788	11.337	5.316
	var	3.243	3.615	1.023	0.132
$\hat{\sigma}^2_{lik}$	mean	15.237	20.558	12.581	5.658
	median	14.212	19.994	12.320	5.571
	var	14.263	11.624	2.944	0.311
$\hat{\lambda}_m$	mean	6.264	2.719	1.771	0.853
	median	6.273	2.720	1.770	0.853
	var	0.024	2.10×10^{-3}	3.77×10^{-4}	1.71×10^{-5}
$\hat{\lambda}_{lik}$	mean	5.878	2.622	1.726	0.841
	median	5.916	2.634	1.733	0.842
	var	0.046	5.46×10^{-3}	1.24×10^{-3}	7.71×10^{-5}
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	54.956	197.51	178.37	152.94
	median	53.794	191.44	171.04	146.19
	var	102.21	1153.6	816.47	439.65
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	79.462	232.82	199.58	161.77
	median	72.578	217.12	188.55	155.23
	var	729.04	5266.7	2949.8	976.24
eff($\hat{\boldsymbol{\beta}}_m$)		1.446	1.179	1.119	1.058
eff($\hat{\mathbf{y}}_m$)		0.998	1.000	0.993	0.991
PSS_r	mean	29.425	48.211	31.283	14.962
PSS_{lik}	mean	29.348	48.325	31.110	14.890
c		2	2	2	2

Table 4.13: Results Comparing robust and nonrobust procedures: Wild error distribution, two large leverage values in **X**, using second robust method

		$PSS_{lik} < PSS_r$	$PSS_{lik} > PSS_r$
$q = 7.21$	$SS(\hat{\beta}_m) > SS(\hat{\beta}_{lik})$	1	0
	$SS(\hat{\beta}_m) < SS(\hat{\beta}_{lik})$	53	46
$q = 17.37$	$SS(\hat{\beta}_m) > SS(\hat{\beta}_{lik})$	3	0
	$SS(\hat{\beta}_m) < SS(\hat{\beta}_{lik})$	48	49
$q = 27.14$	$SS(\hat{\beta}_m) > SS(\hat{\beta}_{lik})$	1.5	2
	$SS(\hat{\beta}_m) < SS(\hat{\beta}_{lik})$	47.5	49
$q = 57.44$	$SS(\hat{\beta}_m) > SS(\hat{\beta}_{lik})$	0	7
	$SS(\hat{\beta}_m) < SS(\hat{\beta}_{lik})$	44	49

Table 4.14: Percentage of cases in which robust and nonrobust methods preferred, based on two different criteria. Wild error distribution, two large leverage values in **X**, using second robust method

The conclusions are somewhat different if we use a robust measure of prediction for comparison, such as (4.5.5). The robust method does better using this criterion when there are influential observations in the **X** space, but worse if there are outliers in the residuals. If both types of contamination are present the methods perform similarly.

4.7 Robust Estimation of q

The work in the previous sections assumed the value $\beta'\Sigma_X\beta/\sigma^2$ was known. We now wish to modify our procedure in chapter 3 to allow for more robust estimation of q, and combine it with the procedures presented in this chapter.

Equation (3.6.1) can be thought of as a measure of predictive error. We propose to use a new measure which does not heavily penalize a choice of q^* which fits most of the data, but gives large prediction errors at a few outlying points. This is what a robust procedure should do in the presence of outliers. With this in mind, we propose to use the robust ρ-function defined in (4.5.4) to measure predictive loss:

$$\sum_{w=1}^{n_v} \rho(\hat{e}_w) \tag{4.7.1}$$

where $\rho(\hat{e}_w) = \min(\hat{e}_w^2, d^2\hat{\sigma}_{\hat{e}}^2)$,

$$\hat{\sigma}_{\hat{e}} = 1.483 \text{ med }_{k\in I_v}|\hat{e}_k - \text{ med }_{j\in I_v}(\hat{e}_j)|$$

is the median absolute deviation of the prediction errors for the validation set I_v, and $\hat{e}_w$ is the *wth* element of $(\mathbf{y}_v - \mathbf{X}_v\hat{\boldsymbol{\beta}}_{c,\hat{\lambda}})$. From our previous work $d = 1.345$ while $\hat{\boldsymbol{\beta}}_{c,\hat{\lambda}}$ is the estimator of $\boldsymbol{\beta}$ using the construction set.

This leads to the following robust procedure. Much of the notation is the same as that used in our nonrobust method for estimating q, described in section 3.6.

1. Split the data into a construction set $(\mathbf{X}_c, \mathbf{y}_c)$ and validation set $(\mathbf{X}_v, \mathbf{y}_v)$.

2. Use $(\mathbf{X}_c, \mathbf{y}_c)$ to find λ to solve

$$\frac{\hat{\boldsymbol{\beta}}_c'\boldsymbol{\Sigma}_X\hat{\boldsymbol{\beta}}_c}{\hat{\sigma}^2_{lik(\lambda,c)}} = q^*$$

   for a fixed choice of q^*. Call the result $\hat{\lambda}$.

3. Use $\hat{\lambda}$ from step 2 and $(\mathbf{X}_c, \mathbf{y}_c)$ to find the ridge M-estimator that solves (4.5.1) subject to the constraint (3.2.2), and using $\hat{\sigma}_m^2$ to estimate σ^2. Call this estimator $\hat{\boldsymbol{\beta}}_{m1,c}$. The weights can be calculated using either (4.5.2) or (4.6.1).

4. Use $\hat{\boldsymbol{\beta}}_{m1,c}$ to find λ to solve

$$\frac{\hat{\boldsymbol{\beta}}_{m1,c}'\boldsymbol{\Sigma}_X\hat{\boldsymbol{\beta}}_{m1,c}}{\hat{\sigma}_m^2} = q^*$$

   Call this value $\hat{\lambda}_c$.

5. Compute an updated robust estimator for $\boldsymbol{\beta}$ using the weights from step 3 and $\hat{\lambda}_c$ from step 4. Call it $\hat{\boldsymbol{\beta}}_{m,c}$.

6. Compute (4.7.1) as the measure of loss for the *ith* split of the data.

7. Repeat steps 1–6 for k random splits of the data, and average (4.7.1) over the k splits, yielding
$$\frac{1}{k}\sum_{j=1}^{k}\sum_{w=1}^{n_v}\rho(\hat{e}_{w,j}) \qquad (4.7.2)$$

8. Repeat steps 1–7 over a range of q^* values, and choose the estimate as the value which gives us the minimum of (4.7.2). Call this $\hat{q}_r$.

9. Using $\hat{q}_r$ as fixed, calculate $\hat{\lambda}_m$, $\hat{\boldsymbol{\beta}}_m$ and $\hat{\sigma}^2_m$ using one of the robust methods proposed in this chapter.

4.7.1 Simulation Results

We present simulation results on the robust estimation of q in models where ϵ is generated using the "wild" error distribution, and there are two influential values in $\mathbf{X}$. We will use our robust estimation procedure with the weights defined in (4.6.1). All other features of the model are the same as those described earlier in the chapter. Table 4.15 summarizes the results over 50 simulated data sets.

In the $q = 7.21$ case we see that median($\hat{q}_r$) $< q$. The median of the $\hat{q}_r$ values is a more sensible estimate than the median of the $\hat{q}_{lik}$ values. We do see both estimates are highly variable. When comparing the $SS(\hat{\boldsymbol{\beta}}_m)$ and $SS(\hat{\boldsymbol{\beta}}_{lik})$ we see that $\hat{\boldsymbol{\beta}}_m$ does a great deal better on average. We also see that the mean PSS_r and PSS_{lik} values favour the nonrobust method, but the median values are very similar. If we compare the individual cases, we find $SS(\hat{\boldsymbol{\beta}}_m) < SS(\hat{\boldsymbol{\beta}}_{lik})$ in 98% of cases, while $PSS_m < PSS_{lik}$ in 57% of cases.

When $q = 17.37$ we see poor estimation of q by both methods, and the robust method chooses a smaller smoothing parameter than the nonrobust procedure. We see the robust method is preferred, on average, when comparing $SS(\hat{\boldsymbol{\beta}}_m)$ and $SS(\hat{\boldsymbol{\beta}}_{lik})$ and when we compare PSS_r and PSS_{lik}. The individual cases also favour the robust method, with $SS(\hat{\boldsymbol{\beta}}_m) < SS(\hat{\boldsymbol{\beta}}_{lik})$ in 70% of cases and $PSS_m < PSS_{lik}$ in 67% of cases.

		q		
		7.21	17.37	27.14
$\hat{q}_r$	mean	64.473	410.47	660
	median	2	50	500
	var	4.00×10^4	1.90×10^6	2.00×10^6
$\hat{q}_{lik}$	mean	961.81	211.03	507.45
	median	1000.0	7.5	525.00
	var	3.21×10^4	1.65×10^6	2.55×10^5
$\hat{\lambda}_m$	mean	57.909	2.552	0.893
	median	20.570	0.985	0.050
	var	8266.5	25.901	2.537
$\hat{\lambda}_{lik}$	mean	0.591	10.249	3.465
	median	0.048	6.417	0.506
	var	9.086	183.93	32.273
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	42.148	191.98	180.89
	median	37.471	186.71	174.83
	var	326.44	954.40	1260.8
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	127.79	208.44	195.31
	median	111.90	206.11	181.71
	var	2884.0	1756.2	1521.7
PSS_r	mean	33.623	43.739	29.473
	median	29.721	42.526	28.480
PSS_{lik}	mean	30.898	54.163	31.751
	median	29.513	50.943	30.440
c		2	2	2

Table 4.15: Robust and nonrobust estimation of q: Wild error distribution, two large leverage values in $\mathbf{X}$

In the $q = 27.14$ model we see that both methods overestimate q by a large amount, and we see the robust method imposes less smoothing than the nonrobust method. In this case we continue to find the robust method favoured based on the measure (4.5.7), while the PSS_r and PSS_{lik} values are very similar on average. When we compare the individual cases, we find $SS(\hat{\beta}_m) < SS(\hat{\beta}_{lik})$ in 66% of cases, while $PSS_m < PSS_{lik}$ in 53% of cases.

We can conclude that the results of the robust estimation of q are satisfactory in the $q = 7.21$ model, but give large overestimation of q in our other two cases. The problem may be in the use of (4.7.2) to estimate q. It uses the mean absolute deviation of the prediction errors, and we saw earlier in the chapter that using $\hat{\sigma}^2_{mad}$ to estimate σ^2 gave poor results. The same problem may be present here. We have conducted some tests on the use of two alternatives to (4.7.2). These were using the sum of the medians of the differences

$$\sum_i (y_{v,i} - \mathbf{x}'_{i,v}\hat{\beta}_c)$$

and using a loss function similar to (4.5.10), but involving the construction and validation sets. However, neither of these two approaches yielded significant improvements in the results.

4.8 Conclusions

We have introduced a method for extending our estimation procedure to be less sensitive to the presence of outliers in the data. For a fixed value of q the method can give noticeable gains in efficiency, based on (4.5.7), particularly with moderate sized values of q. The extension of the work to allow for robust estimation of q seems to work reasonably well at small values of q, and the gains in performance observed with fixed q continue to be present.

Chapter 5

Analysis of Ocean Data

5.1 Introduction

In this chapter we return to the California Current data introduced in chapter 1, and use the methods developed in chapter 3 to derive multiple linear regression models for predicting deep ocean measurements from shallow water readings. We will also use the robust procedure developed in chapter 4 to analyse the data sets and compare our results to the analysis done by Haney et al. (1995).

5.2 Description of Data

As described in chapter 1, each data set can be thought of as a set of 64 independent vectors, representing temperature or salinity observations at 64 locations in the California Current during the summer months. The readings are taken at approximately 10 metre (m) intervals, from the sea surface to 2000 m. It was argued in chapter 1 why we can treat the 64 locations as approximately independent. Figures 5.1 and 5.2 display scatterplots of temperature and salinity as functions of depth. We see temperature decreases with depth, while salinity increases. To see the relationship between the two variables, a temperature versus salinity plot is presented in figure 5.3. We

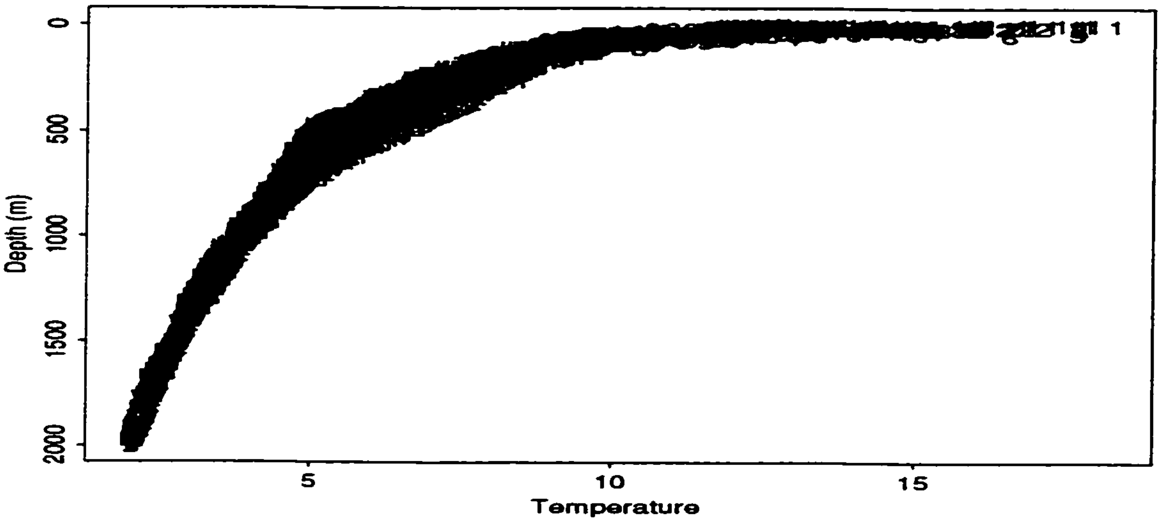


Figure 5.1: Plot of temperature versus depth

see greater variability in shallow water than in deep water, *i.e.* in the high temperature, low salinity region. The shallow water is warmed or cooled more easily by the atmosphere, so we would expect the temperatures to be more variable in this region.

There is the possibility that both of our processes are nonstationary. This suspicion arises in two ways: the variability of the observations may not be constant in depth, and the correlation between values within a vector may not only depend on the distance between them. This could be an important issue in our modelling procedure because we need to estimate Σ_X. To assess this, figure 5.4 contains plots of the sample variances for each data set as a function of depth, while figures 5.5 and 5.6 present contour plots of the correlations as a function of depth. If the processes are stationary we would expect to see approximately flat curves in figure 5.4 and to see constant values along the diagonals of figures 5.5 and 5.6. The plots of the variability at each depth certainly suggests processes are more variable near the surface than in deep water. We return to this issue in the next section.

Our aim is to construct an underdetermined model which will use upper ocean readings as the explanatory variables, and a deep ocean observation as the response.

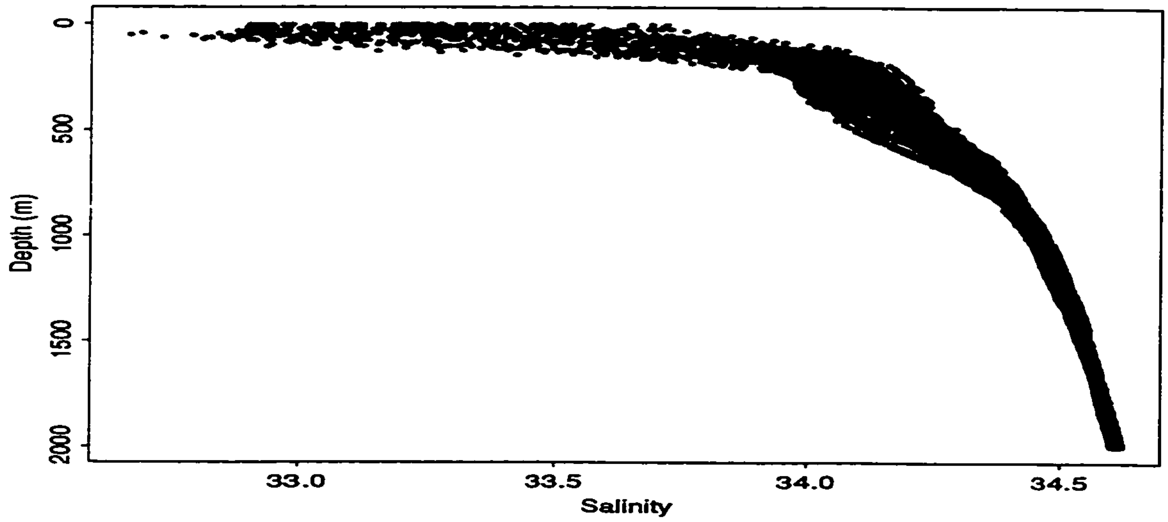


Figure 5.2: Plot of salinity versus depth

Therefore our explanatory and response variables are measurements of the same quantity. We will use a model where the explanatory variables are the upper $p = 80$ measurements in the ocean, while there are $n = 64$ observations. We are using (3.2.1) as our model, where

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_{64}' \end{pmatrix}$$

and $\mathbf{x}_i'$ contains the temperatures (or salinity readings) from the surface to 800 m at the *ith* location. Both $\mathbf{X}$ and $\mathbf{y}$ have been centered to satisfy the assumption that $\mathrm{E}(\mathbf{x}_i') = \mathbf{0}'$ and $\mathrm{E}(\mathbf{y}) = \mathbf{0}$. Since we are assuming the rows of $\mathbf{X}$ are independent, and figures 5.1 and 5.2 do not suggest a great deal of variability between the 64 vectors in each data set, it is reasonable to assume a common covariance structure Σ_X is appropriate.

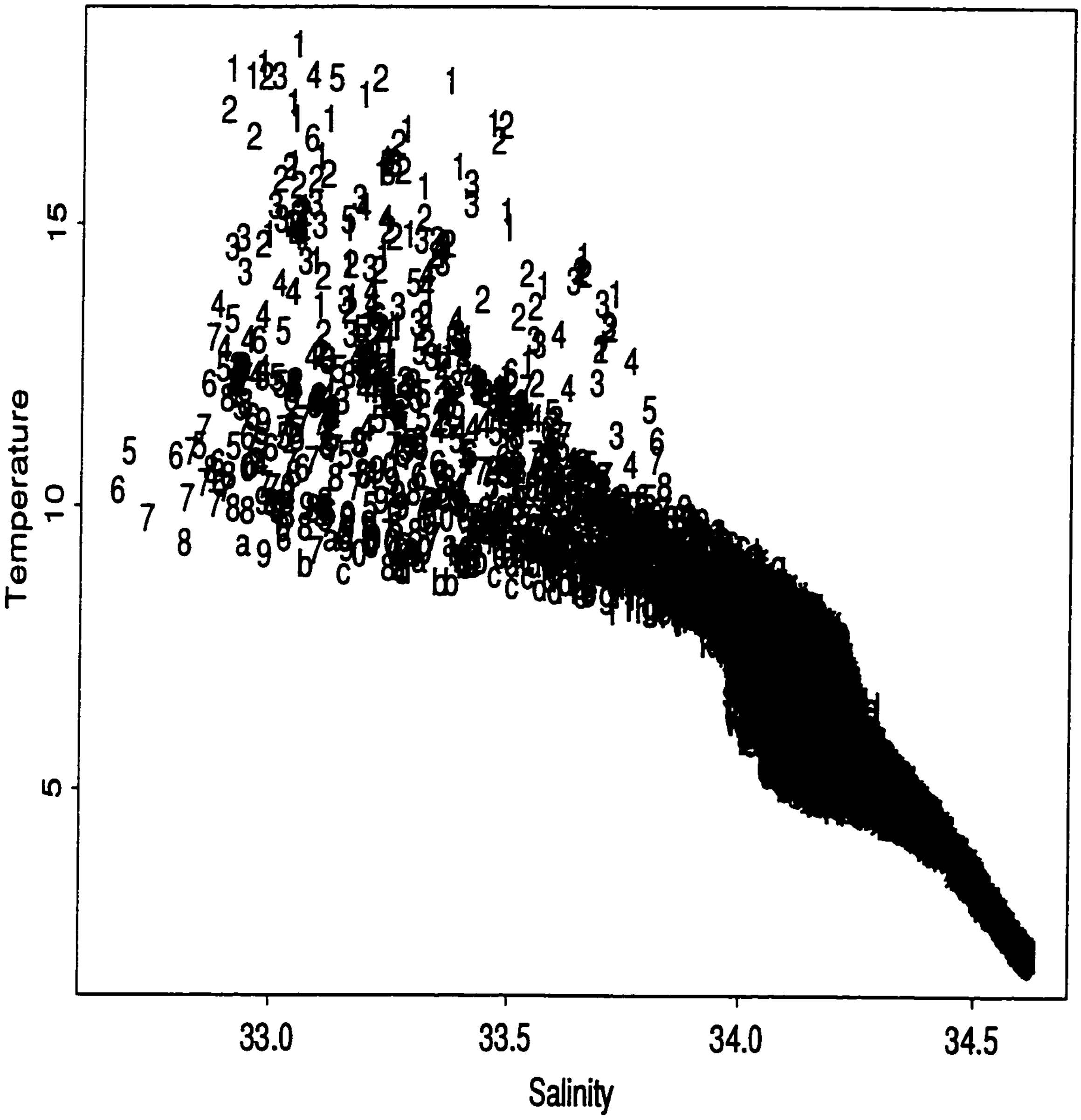


Figure 5.3: Plot of temperature versus salinity

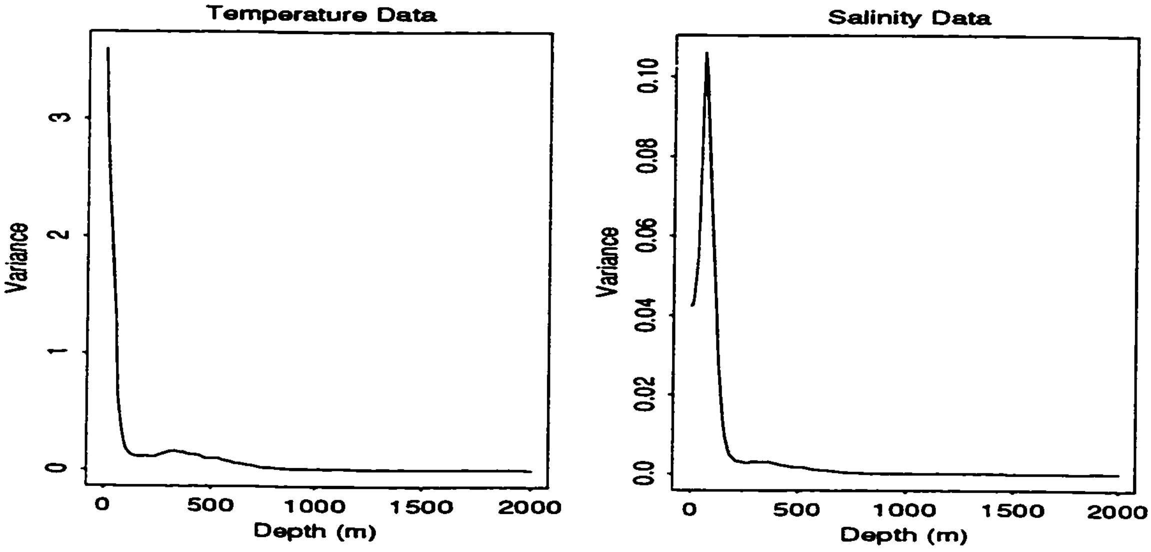


Figure 5.4: Plots of sample variance versus depth for each data set

5.3 Data Analysis

We will use our estimation procedure of chapter 3, assuming we have no *a priori* knowledge of q, so it must be estimated. We will also have to estimate Σ_X, as mentioned previously. This will be the first issue we will discuss, followed by a presentation of results under various choices for the response $\mathbf{y}$, and some data-splitting results to compare the performance of the signal-to-noise ratio with GCV on these data sets. We will then include an examination of models which use temperature and salinity readings in the $\mathbf{X}$ matrix, and finally repeat some of the analyses using our robust procedure.

5.3.1 Estimation of Σ_X

We presented results in chapter 3 on estimating Σ_X, with q fixed. We saw that $\hat{\lambda}_{lik}$ was not greatly affected by using $\hat{\Sigma}_X$, even when the form of Σ_X was incorrectly specified.

We are assuming that the observations are stationary with depth, and construct

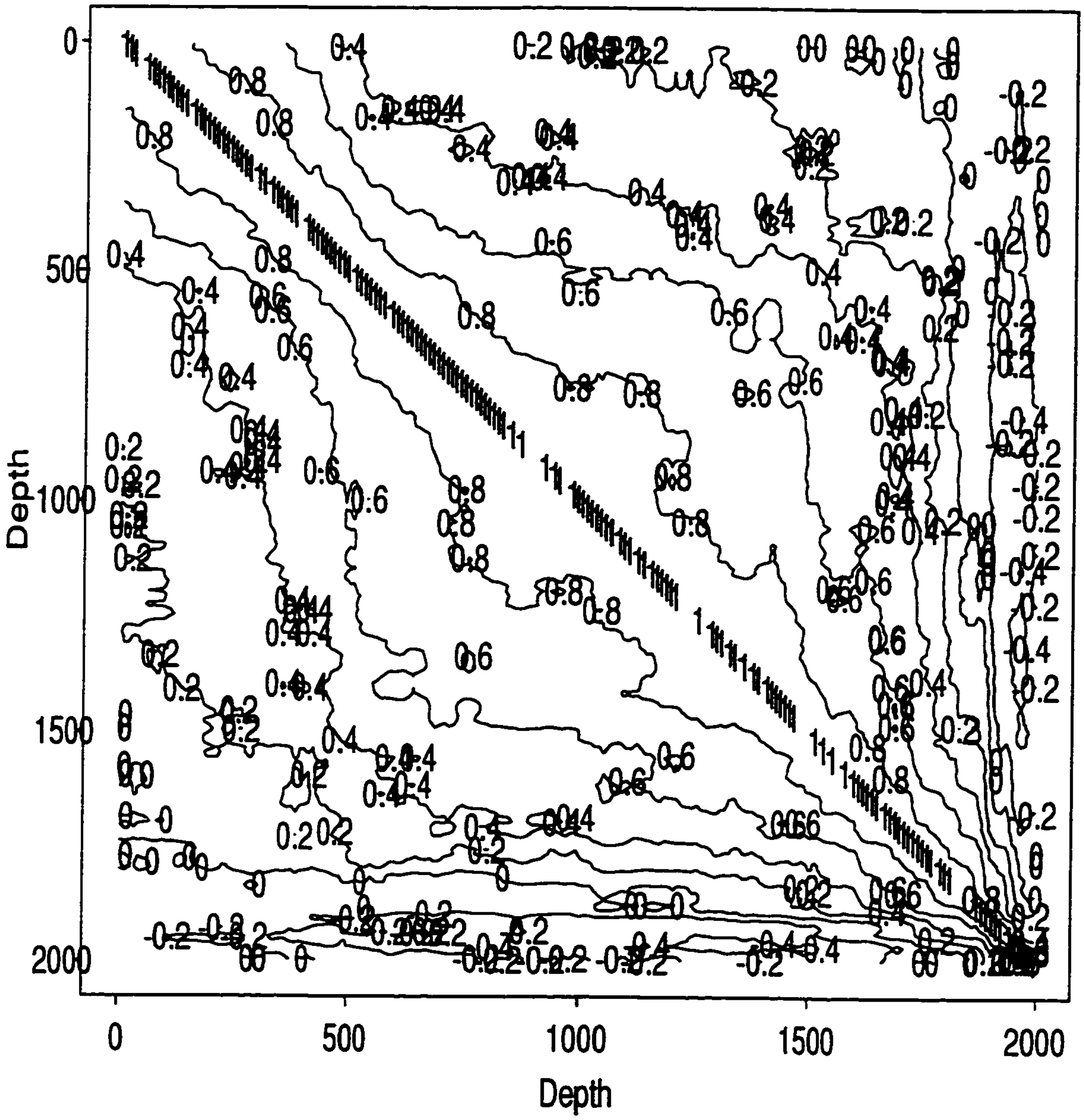


Figure 5.5: Contour plot of correlations: Temperature data

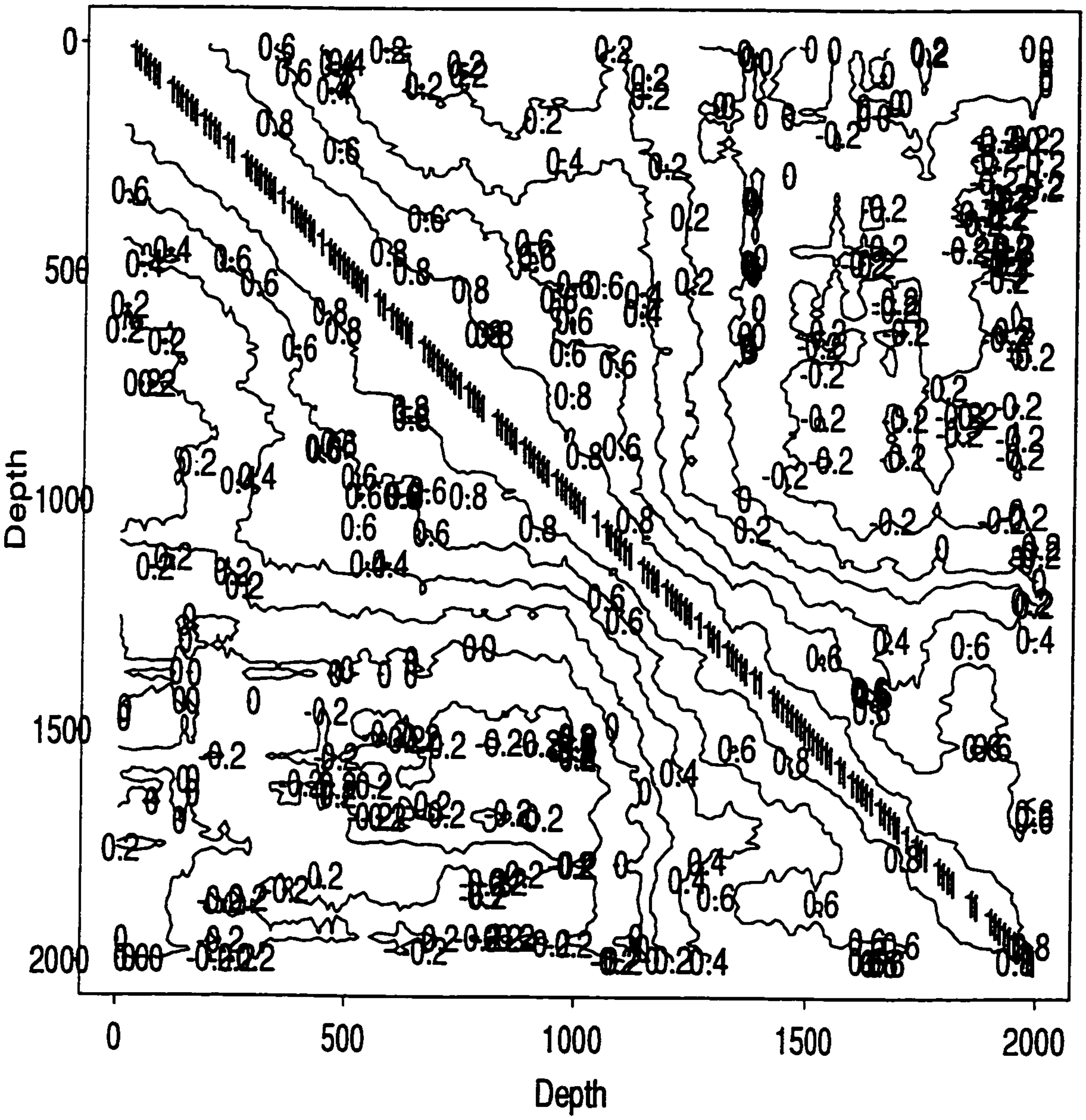


Figure 5.6: Contour plot of correlations: Salinity data

$\hat{\Sigma}_X$ under the assumption that an AR form (3.4.1) holds. We consider two methods for finding $\hat{\Sigma}_X$. The first is to treat each $\mathbf{x}_i'$ as a time series, and use traditional time domain estimation methods to estimate the order of the AR process, and its parameters. This could be done by examining ACF and PACF plots, and more formally using the AIC criterion, for example.

The ACF plots of the salinity and temperature series, for most $\mathbf{x}_i'$, showed a gradual decline as the lag increased. The PACF plots decayed quickly after lag one, indicating that an AR(1) structure appears reasonable. Plots of selected ACF and PACF plots are given in figure 5.7 and 5.8.

We also used the AIC criterion, introduced in chapter 3, to estimate the order of the AR process for each $\mathbf{x}_i'$ series. For the temperature data an AR(1) model was chosen for 58 of the 64 series, with the largest model chosen being AR(5). For the salinity data an AR(1) model was chosen for 51 of the 64 series, with an AR(8) model being the largest chosen. This is further evidence that an AR(1) structure for Σ_X appears reasonable. However, if we use this model we find the parameter estimate $\hat{\phi} = 0.9561$ for the temperature data and $\hat{\phi} = 0.9615$ for the salinity data. These estimates are close to the region of nonstationary of an AR(1) process.

Our second approach is less restrictive. We simply use the sample autocovariances (3.4.2) in $\hat{\Sigma}_X$, which is equivalent to assuming an AR$(p-1)$ form for the covariance structure. This will not give good estimate of $r(k)$ at high lags, but as we saw in chapter 3, it did not make a great deal of difference to the $\hat{\lambda}_{lik}$ values.

As a final point, we also performed some analyses with the data both centered and scaled, so the observations at a particular depth were divided by the standard deviation of the values at that depth after being centered. However, it did not make a great deal of difference to the final results.

5.3.2 Estimation and Prediction

We begin with our analysis of the temperature data. Tables 5.1 and 5.2 present the values of $\hat{q}$, $\hat{\lambda}_{lik}$ and $\hat{\lambda}_{GCV}$ for various choices of $\mathbf{y}$ and the two choices for $\hat{\Sigma}_X$. In

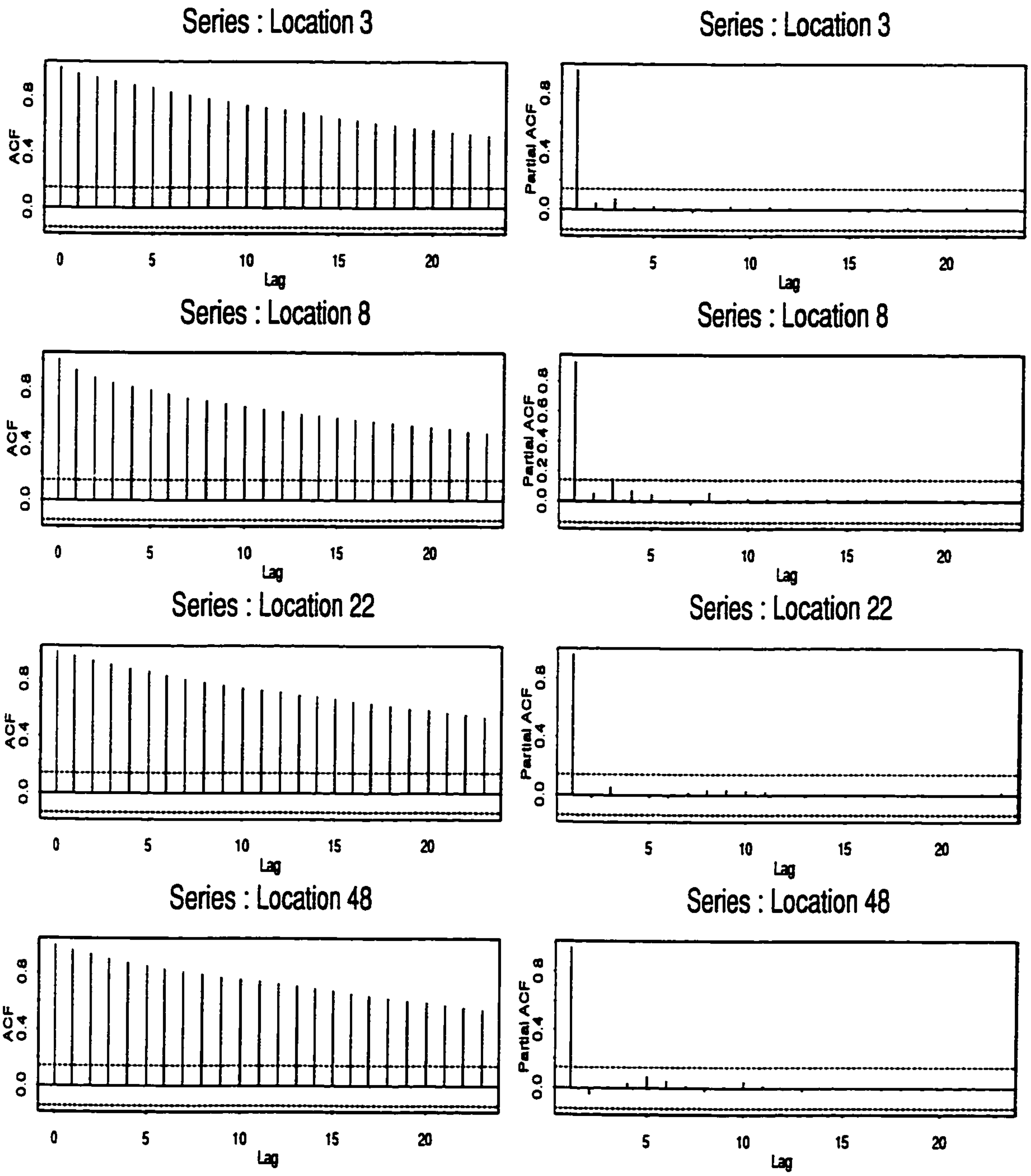


Figure 5.7: ACF and PACF plots of four temperature series

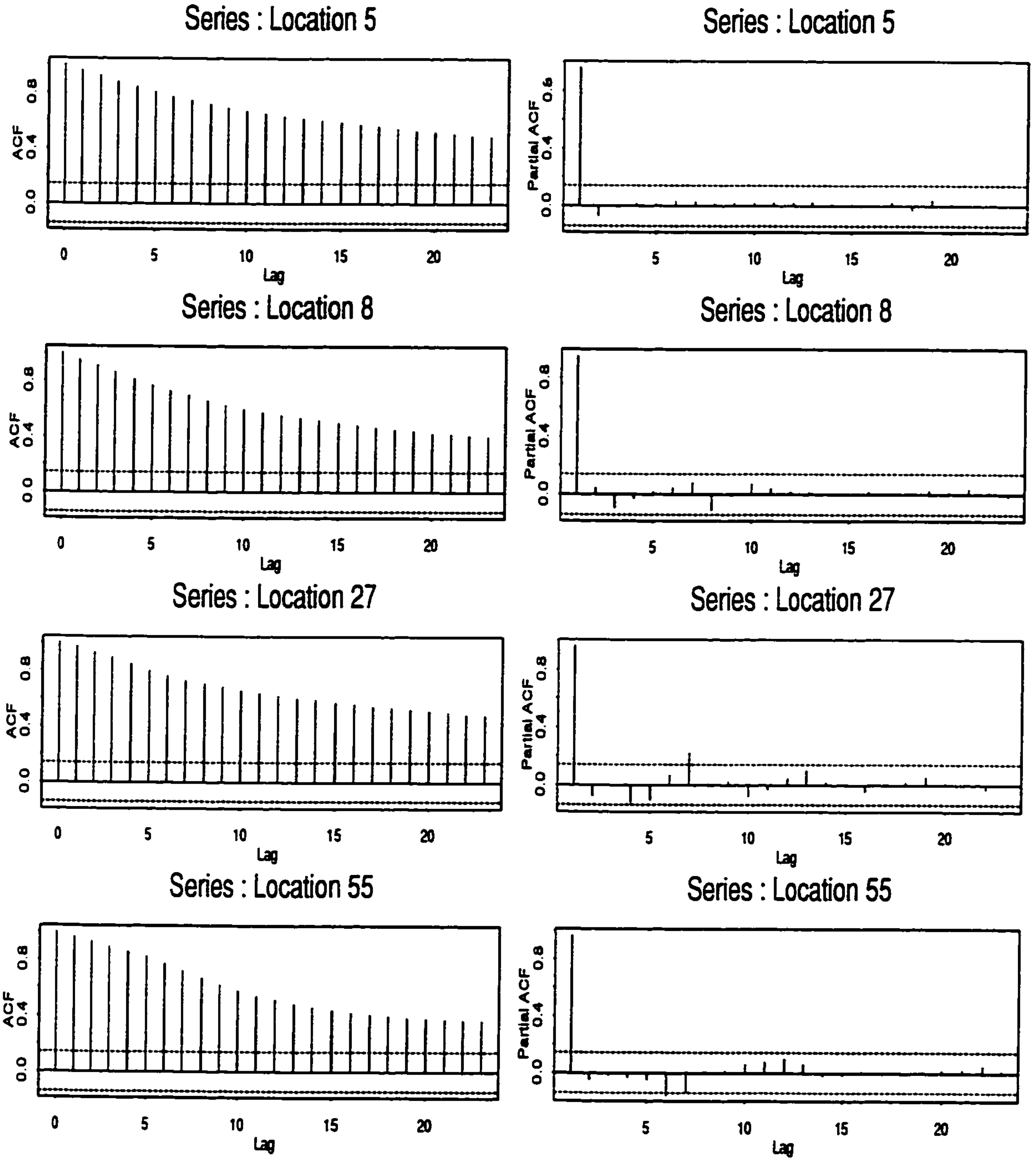


Figure 5.8: ACF and PACF plots of four salinity series

estimating q, we will use $n_c = n_v = 32$, do 50 random splits of the data, and search over q values from 0.0001 to 50.

Using $\mathbf{y}$ as the temperatures at 810 m we see $\hat{q}$ is very large, which is reasonable. We would expect variability in the signal to dominate, since $\mathbf{X}$ includes observations down to 800 m. However, this case is not very interesting, because the temperature reading at 800 m would be the most sensible prediction of the temperature at 810 m.

	y Value		
	810 m	900 m	1500 m
$\hat{q}$	50	5	5×10^{-4}
$\hat{\lambda}_{lik}$	0.943	5.97	3.1×10^{3}
$\hat{\lambda}_{GCV}$	0	4.63	34.41

Table 5.1: Values of $\hat{q}$ and $\hat{\lambda}$ for various choices of $\mathbf{y}$, assuming AR(1) covariance structure for rows of $\mathbf{X}$: Temperature Data

We see that, as we used $\mathbf{y}$ as the readings deeper in the ocean, the $\hat{q}$ values tend to decrease, so the $\hat{\lambda}$ values increase. The two methods perform similarly in many cases, but there are some models where the signal-to-noise ratio method does a great deal more smoothing. There are also more differences between the results using the two different forms for $\hat{\Sigma}_X$ than we might have expected. When we use $\mathbf{y}$ as our observations at 1900 m, both methods do a great deal of smoothing. The signal-to-noise method chooses the smallest value of $\hat{q}$ allowed, while GCV is totally eliminating

	y Value		
	900 m	1500 m	1900 m
$\hat{q}$	10	0.01	5×10^{-4}
$\hat{\lambda}_{lik}$	2.85	394.24	1640.70
$\hat{\lambda}_{GCV}$	2.51	111.77	4.75×10^{10}

Table 5.2: Values of $\hat{q}$ and $\hat{\lambda}$ for various choices of $\mathbf{y}$, using sample autocovariances to estimate Σ_X: Temperature Data

the role of $\mathbf{X}$ in the model. This tells us that the model we are using cannot predict values well that are a great distance from the $\mathbf{X}$ observations.

We now turn to the salinity data results. The estimation trends are somewhat different here. In table 5.3 we see the signal-to-noise ratio tends to impose more smoothing, with a large different in results when $\mathbf{y}$ contains the readings at 810 m. The large amount of smoothing being imposed by the choice $\hat{q}$ does not seem appropriate here. An examination of $\hat{\boldsymbol{\beta}}_{lik}$ indicates the largest element in the vector goes with the readings at 40 m, but the smallest elements are assigned to the readings at 30 m and 50 m. The largest value in $\hat{\boldsymbol{\beta}}_{GCV}$ goes with the readings at 630 m, while many of the other large elements correspond to the readings near 800 m. Similar results can be seen in table 5.4.

	y Value		
	810 m	1000 m	1500 m
$\hat{q}$	0.005	0.005	0.005
$\hat{\lambda}_{lik}$	666.08	854.66	840.14
$\hat{\lambda}_{GCV}$	0	678.9	446.16

Table 5.3: Values of $\hat{q}$ and $\hat{\lambda}$ for various choices of $\mathbf{y}$, assuming AR(1) covariance structure for rows of $\mathbf{X}$: Salinity Data

	y Value		
	1000 m	1500 m	1900 m
$\hat{q}$	0.0007	0.02	0.0003
$\hat{\lambda}_{lik}$	2.9×10^3	267.74	2.1×10^3
$\hat{\lambda}_{GCV}$	535	544	2.97×10^9

Table 5.4: Values of $\hat{q}$ and $\hat{\lambda}$ for various choices of $\mathbf{y}$, using sample autocovariances to estimate Σ_X: Salinity Data

Our primary focus in chapter 3 was the predictive ability of the estimates. To have a better comparison of GCV and the signal-to-noise ratio methods, we have

	1500 m		1900 m	
	drop 10	drop 32	drop 10	drop 32
mean(RAT_{PSS})	0.8374	0.9334	0.9966	1.0050
median(RAT_{PSS})	0.9022	0.9944	0.9958	0.9999

Table 5.5: Ratios of Predictive Sums of Squares for California Current Temperature Data

done the following.

We have taken our $n = 64$ observations and randomly chosen 10 or 32 to set aside. We then take the remaining 54 or 32 values, and select our smoothing parameters as above. Then we use these values to estimate β and predict the values left aside. Finally, we examine the error in prediction using (3.6.3) and (3.6.4), with n replaced by the number of observations set aside. This was repeated 10 times. Table 5.5 displays the results that compare (3.6.3) and (3.6.4) using the ratio

$$RAT_{PSS} = PSS_{lik}/PSS_{GCV}$$

for the temperature data. If $RAT_{PSS} < 1$ then the signal-to-noise ratio method is performing better at predicting the deleted observations. All of these results used the sample autocovariances to estimate Σ_X.

At 1500 m. and leaving aside 10 values, we see the mean ratio is less than one, so on average the signal-to-noise method does better at prediction. Even though we only have 10 cases, there is an indication of skewness in this distribution because the mean is less than the median. There were five cases where the method did much better than GCV (at least 17% better), and no cases where it did worse. The cases with superior prediction were usually linked with GCV grossly undersmoothing. When we left 32 values aside, there was little difference between the methods, based on the median. But there were two cases where GCV did much worse, because $\hat{\lambda}_{GCV} = 0$. This is reflected in the mean ratio being less than the median.

At 1900 m, and leaving out 10 values, we see little difference between the methods. Both are indicating the model is a poor predictor of temperature in this case. The same conclusions hold for leaving 32 values aside, except for one case where GCV does about 10% better at predicting the left out values.

Table 5.6 contains similar results for the salinity data. At 1500 m, and leaving aside 10 values, we see the signal-to-noise ratio does better at prediction. It never does worse in any of the 10 cases, and in 2 of the cases does at least 10% than GCV. The results are very similar when we leave 32 values aside, and $\mathbf{y}$ contains the readings at 1500 m.

At 1900 m, the methods perform similarly, both on average and in the individual cases.

	1500 m		1900 m	
	drop 10	drop 32	drop 10	drop 32
mean(RAT_{PSS})	0.9540	0.9991	1.005	1.022
median(RAT_{PSS})	0.9646	0.9996	1.012	1.014

Table 5.6: Ratios of Predictive Sums of Squares for California Current Salinity Data

Although tables 5.5 and 5.6 compare how well the signal-to-noise ratio method and GCV do relative to each other in prediction, they do not indicate how well the models do relative to using the sample mean of the left out values, $\bar{y}_{new}$. We calculate the following approximate R^2 value as a measure of the overall predictive ability of a model:

$$R^2_{new} = 1 - \frac{\sum_{i=1}^{w}(y_{new,i} - \hat{y}_{new,i})^2}{\sum_{i=1}^{w}(y_{new,i} - \bar{y}_{new})^2}$$

where $w = 10$ or $w = 32$, depending on how many points we leave aside. Tables 5.7 and 5.8 give the mean of the R^2_{new} values of the 10 splits used in constructing tables 5.5 and 5.6. We see the models are not performing well relative to $\bar{y}_{new}$ in general.

We also investigated the use of a model which used both temperature and salinity readings in the upper 800 m as the explanatory variables. The new model now has $p = 160$ parameters and $n = 64$ observations. We set the model up as follows:

$$\mathbf{y} = (\mathbf{X}_t|\mathbf{X}_s)\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

		1500 m		1900 m	
	Method	drop 10	drop 32	drop 10	drop 32
mean(R^2_{new})	signal-to-noise	0.039	0.017	0.025	2.19×10^{-3}
	GCV	0.033	0.016	1.14×10^{-6}	3.33×10^{-7}

Table 5.7: Mean of R^2_{new} values associated with left out data points: Temperature data

		1500 m		1900 m	
	Method	drop 10	drop 32	drop 10	drop 32
mean(R^2_{new})	signal-to-noise	0.184	0.034	4.64×10^{-3}	9.24×10^{-5}
	GCV	0.140	0.021	1.97×10^{-8}	4.79×10^{-10}

Table 5.8: Mean of R^2_{new} values associated with left out data points: Salinity data

where $\mathbf{X}_t$ and $\mathbf{X}_s$ represent the temperature and salinity measurements respectively. We still need to estimate Σ_X, which now has the form

$$\Sigma_X = \begin{bmatrix} \Sigma_t & \Sigma_{ts} \\ \Sigma_{st} & \Sigma_s \end{bmatrix}$$

where Σ_{ts} represents the covariance between temperature and salinity. To be consistent with our previous work we estimate Σ_t and Σ_s using the sample autocovariances and Σ_{ts} and Σ_{st} using the sample crosscovariance function. The results of these analyses are given in tables 5.9 and 5.10.

We see in table 5.9 that we chose a smaller value of $\hat{q}$ at 1000 m than was chosen when we only used temperature readings in $\mathbf{X}$, and the response was the temperature at 900 m. When we use $\mathbf{y}$ as the temperature readings at 1500 m we find $\hat{q}$ is equal to that given in table 5.2. At 1900 m the use of salinity and temperature as explanatory variables has not had a strong affect on the estimate of q. We also see that the $\hat{\lambda}_{GCV}$ values have either increased or stayed approximately the same as those given in table 5.2.

When we use the salinity data as the response variable, we find the choice of $\hat{q}$ has

	y Value		
	1000 m	1500 m	1900 m
$\hat{q}$	0.1	0.01	10^{-4}
$\hat{\lambda}_{lik}$	108.13	491.58	3994.0
$\hat{\lambda}_{GCV}$	64.541	333.12	1.08×10^{11}

Table 5.9: Values of $\hat{q}$ and $\hat{\lambda}$ when using temperature and salinity as explanatory variables, temperature as response

	y Value		
	1000 m	1500 m	1900 m
$\hat{q}$	0.005	0.001	5×10^{-4}
$\hat{\lambda}_{lik}$	858.85	1927.4	1564.66
$\hat{\lambda}_{GCV}$	422.13	734.55	1.02×10^{9}

Table 5.10: Values of $\hat{q}$ and $\hat{\lambda}$ when using temperature and salinity as explanatory variables, salinity as response

either increased, or stayed approximately the same, when compared to the values in table 5.4. We also see that $\hat{\lambda}_{GCV}$ decreased in this new model when the response is the salinity at 1000 m, increased at 1500 m, and continued to be extremely large at the 1900 m depth.

5.3.3 Robust Estimation

The exploratory plots given at the beginning of this chapter do not suggest that outliers are a major concern in these data sets. In this case we would expect the nonrobust method developed in chapter 4 to yield similar results to those already presented in this chapter. We implement the robust procedure for estimating q and β given in section 4.7 , using the weights defined in (4.6.1) and $c = 2$ in the weight function. Table 5.11 presents the values of $\hat{q}$ and $\hat{\lambda}$ for the robust analyses of the temperature and salinity data sets. We have used the sample autocovariances to estimate Σ_X.

	y Value: Temperature		y Value: Salinity	
	1500 m	1900 m	1500 m	1900 m
$\hat{q}$	0.1	0.005	0.05	10^{-4}
λ_{lik}	80.917	456.57	210.78	4709.1

Table 5.11: Values of $\hat{q}_r$ and $\hat{\lambda}_m$ for various choices of y, using sample autocovariances to estimate Σ_X

We begin with an examination of the temperature data results. We see in the model that uses readings at 1500 m as the response that the value of $\hat{q}_r$ is larger than $\hat{q}$ given in table 5.2, so less smoothing is being done. We find that five of the w_i values given by (4.6.1) are not equal to one. The smallest w_i is 0.802. In the 1900 m model we again see that $\hat{q}_r > \hat{q}$, while nine of 64 observations are assigned $w_i < 1$, with the smallest $w_i = 0.497$.

The use of the salinity data as the explanatory variables still has the same pattern in the estimation of $\hat{q}$. The value of $\hat{q}_r$ is either larger than, or similar to, the $\hat{q}$ value given in table 5.4. The model which uses y as the salinity readings at 1500 m sets 13 of the w_i less than 1, with two of the values less than 0.5, and the smallest $w_i = 0.400$. The 1900 m model downweights only six of the 64 observations, with the smallest $w_i = 0.540$.

5.4 Comparison to Alternate Method

It is natural to compare the results we have obtained with those found by Haney et al. (1995), who used a qualitatively different approach to the analysis. We will match their notation as closely as possible in this section.

They denote the temperature or salinity profile at a location z by $\Theta(z)$. This is a vector that contains all the readings at the location z. They express this profile as

$$\Theta(z) = \bar{\Theta}(z) + \Theta^*(z)$$

where $\bar{\Theta}(z)$ is the mean of the 64 vectors of either temperature or salinity readings. They model $\Theta^*(z)$ as

$$\hat{\Theta}(z) = \sum_{i=1}^{N} A_i \Theta_i(z)$$

where $\Theta_i(z)$, $i = 1, \ldots, N$ are the empirical vertical modes (EOF's) and the A_i values are found by performing a successive least squares fit of $\Theta^*(z)$ to the first N empirical vertical modes above a specified depth. Their examination of the data to see what proportion of the variance is accounted for by each mode suggested a choice of $N = 7$ was appropriate.

They begin with the original data set and use the EOF's found using the upper ocean readings to predict the entire profile in the water. We will focus on the case where they used the upper 1000 m readings to construct the vertical modes, since this is similar to the upper 800 m readings we used to construct our regression model. They found the correlation between observed and predicted values at 1500 m to be around 0.5, while the correlation is around 0.30 between observed and predicted values at 1900 m. These values are better than those found in tables 5.7 and 5.8.

They do another analysis that is probably more appropriate to compare to our results. They obtain a new data set of 34 stations from a location approximately 100 km away from the sites of the original stations. These 34 vectors are also assumed to be independent based on the arguments presented in chapter 1. They use the upper ocean readings from the original data to construct the EOF's, then see how well they predict the readings in the 34 new stations. This is similar to what we have done in constructing tables 5.7 and 5.8, where we left locations out of our estimation procedure.

They found the correlation between observed and predicted values was very low at depths below 1000 m. When the upper 1000 m of the original data is used, the correlation between observed and predicted responses at 1500 m in the new data set is around 0, and the correlation is negative when readings at 1900 m are considered. These poor results are similar to the R^2_{new} values we found using the signal-to-noise ratio method.

5.5 Conclusions

The signal-to-noise ratio method has performed comparably to GCV in the selection of a smoothing parameter in many of the situations when all data are utilized in model construction. In attempts to compare predictive ability of the models for each data set, the signal-to-noise ratio method tended to yield better results, especially for the temperature data. However, the predictive ability of the models was quite poor, based on the R^2_{new} values. The inclusion of both salinity and temperature readings as explanatory variables did not have a large effect on the values of the smoothing parameters chosen. We did note, however, that Haney et al. (1995) also obtained poor results in their attempt to predict a new data set. We did observe that the robust method tended to yield larger estimates of q that a nonrobust analysis, but the downweighting of values did not appear to suggest that outliers were a major concern in these data sets.

Chapter 6

Conclusions

6.1 Introduction

In this concluding chapter we will summarize the results that have been presented in this thesis, followed by an outline of some possible further directions in which this work could be taken.

6.2 Summary of Findings

We motivated our study of underdetermined regression models by describing several problems in which it was reasonable to consider the explanatory variable as samples from a continuous process. This opened up the possibility that the models could be underdetermined, so we needed to use smoothed estimators.

We then discussed several ways to derive a smoothed estimator in linear regression models, and various methods that have been used to empirically select a smoothing parameter. Having found many of these methods did not perform well in underdetermined models, we proposed a method to find a smoothed estimator based on the signal-to-noise ratio of the model. This allowed us to explicitly introduce the random structure of the explanatory variables in the estimation procedure.

The method gave smoothing parameters with low variability when the value of the ratio was fixed, and we found good predictive ability over a wide range of values of the ratio. We then developed a data-splitting procedure to estimate the signal-to-noise ratio, and the method continued to perform well on a prediction-based criterion.

We extended the method to permit robust estimation of the parameters. We used two methods, one which downweighted outliers in the residual space, and a second which downweighted influential values in the residual and explanatory variable spaces. Both methods had bounded influence functions and performed well in estimation based on the efficiency measure (4.5.7). We also combined these procedures with robust estimation of the signal-to-noise ratio and obtained satisfactory results, particularly at small values of the ratio.

We then used the above methods to analyze temperature and salinity data taken from the California Current. The methods performed similarly to each other and to GCV, but the predictive ability of the models was often poor.

The general conclusions that can be drawn from the thesis are the following. The use of a data-splitting procedure instead of generalized cross validation appears to result in a more stable method when dealing with underdetermined models. We see that the method is less likely to select grossly undersmoothed underestimates when the true value of the ratio is small. It is in these situations that the imposition of a reasonable amount of smoothing is needed. Finally, the assumption of a random structure in our explanatory variables allows for sensible specification of the smoothing matrix, instead of relying on a more subjective choice.

6.3 Further Directions

There appear to be two natural extensions to the methods developed in this thesis: allowing for correlated errors in the model and an extension to nonlinear models. We will also briefly discuss another approach that seems appropriate when the rows of the $\mathbf{X}$ matrix can be considered samples from an underlying continuous process.

Our underlying model throughout was

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where we assumed $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and each row of $\mathbf{X}$, $\mathbf{x}_i'$, was independent of each other and $\boldsymbol{\epsilon}$. These will often be unreasonable assumptions, particularly in problems in oceanography or climate studies. Since the data is often collected through time, or at neighbouring locations, correlated observations will be the norm rather than the exception.

The choice of a smoothing parameter in the presence of correlated errors is a growing area of research. Diggle and Hutchinson (1988) state that making an incorrect assumption of uncorrelated errors in cross-validated smoothing spline estimates will lead to gross undersmoothing because most of the random fluctuations in the errors will be ascribed to the estimate of the signal. Altman (1990) describes an extension of GCV, in kernel smoothing, to allow for correlated errors. Kohn, Ansley, and Wong (1992) estimate the unknown parameters in a nonparametric spline regression model, where the errors are modelled using an ARMA process. The reader is referred to Chiu (1989), Chu and Marron (1991), Hart (1991), van der Linde (1994) and the cited references in these works for more detail on this topic.

We first consider how our methods may be extended to allow for correlated errors. Suppose we now assume that $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$. The way we define the signal-to-noise ratio of the model remains unchanged, but using it as a penalty will not take the correlation between the errors into account. Two possible modifications are to replace the variance of the error term with either the total variance or the generalized variance of $\boldsymbol{\epsilon}$:

$$\frac{\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta}}{\mathrm{tr}(\boldsymbol{\Sigma}_\epsilon)} \quad \text{or} \quad \frac{\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta}}{|\boldsymbol{\Sigma}_\epsilon|} \; .$$

The second may be more appropriate, since it does not ignore the covariance between ϵ_i and ϵ_j.

Our method could allow for the $\mathbf{x}_i'$ values to be correlated. We could define

$\Sigma_{ij} = \mathrm{Cov}(\mathbf{x}_i', \mathbf{x}_j')$ and write

$$\Sigma_X = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \dots & \Sigma_{1n} \\ \Sigma_{21} & \dots & \dots & \Sigma_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \Sigma_{n1} & \dots & \dots & \Sigma_{nn} \end{bmatrix}$$

Now we would have Var(signal) $= \boldsymbol{\beta}'\Sigma_{ii}\boldsymbol{\beta}$ for the *ith* observation. If we wanted to incorporate Σ_X into the penalty imposed on $\boldsymbol{\beta}$ we could attempt to use the average of the Σ_{ij} matrices:

$$\frac{\boldsymbol{\beta}'\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\Sigma_{ij}\right)\boldsymbol{\beta}}{n^2\sigma^2}.$$

If the errors were also correlated, but independent of the $\mathbf{x}_i'$ values, these two approaches could be combined.

The most complicated scenario is to allow the signal and noise to be correlated. In this case we would have a more complicated log-likelihood than the one given in chapter 3. We would not be able to find a closed form expression for $\hat{\boldsymbol{\beta}}$ in this case. Once this issue is combined with finding a reasonable definition of the signal-to-noise ratio we are left with a very challenging problem.

All of these ideas also require knowledge of the Σ_ϵ and Σ_{ij} matrices defined above. In practice these matrices would have to be estimated, probably assuming a low order AR structure, as we have done in this thesis to estimate Σ_X.

The second issue is an extension of the method to nonlinear models. Once again, nonlinear models are often necessary if we want a more realistic description of the dynamic process. A simple case of this was introduced in water temperature change model of chapter 1. We derived a linear model from the discretization of a differential equation, under the assumption that the diffusion coefficient k was not a parameter. If we now consider k as a parameter the model is now nonlinear in one of its parameters. We then have to consider alternatives such as Taylor approximations to the objective functions we wish to minimize. This would have to be combined with the smoothness penalty imposed on the parameters.

A final approach to these problems is to try and take more advantage of the fact that the explanatory variables are realizations of a (smooth) continuous process. Hastie and Mallows (1993) suggest thinking of each explanatory variable as $x_{ji} = f_i(t_j)$, where f_i is the function associated with the *ith* observation, sampled at points $t_j \in D$, where D is the domain of f_i for all i. They then describe the regression model as an approximation to a linear functional:

$$\begin{aligned} \mathrm{E}(y_i|f_i) &= \int_D f_i(t)\alpha(t)dt \\ &\approx \sum_{j=1}^{p} f_i(t_j)\alpha(t_j) \\ &= \sum_{j=1}^{p} x_{ji}\alpha_j \end{aligned}$$

where $\alpha(t)$ is a coefficient function. A reasonable way to model the coefficients smoothly is to express them in terms of a basis expansion of smooth functions b_k:

$$\alpha(t) = \sum_{k=1}^{K} b_k(t)\theta_k$$

If the underlying process is smooth we should be able to express $\alpha(t)$ using a small number of basis functions. In effect, this is reducing the dimension of the model. A related idea is used by O'Sullivan and Wahba (1985) in an inversion problem to estimate the atmospheric temperature profile from upwelling radiance measurements. Manchester and Haller (1997) have extended these ideas in a fisheries problem, where the x_{ji} values are realizations of two dimensional air pressure fields over the North Atlantic Ocean. They express each y_i value as

$$y_i = \int_D H_i(x)d\mu_x$$

where $H_i(x)$ is a thin plate spline interpolant of the *ith* row of $\mathbf{X}$, and μ_x is approximated by some suitably chosen collection of basis functions.

Appendix A

Existence and Uniqueness of Smoothing Parameters

A.1 Introduction

In this appendix we will derive the conditions under which $\hat{\lambda}_t$, $\hat{\lambda}_a$ and $\hat{\lambda}_{lik}$, introduced in chapter 3, exist and are unique solutions to their respective constraint equations.

Each case will use the SVD to re-express $\mathbf{X}\boldsymbol{\Sigma}_X^{-1/2}$:

$$\mathbf{X}\boldsymbol{\Sigma}_X^{-1/2} = \mathbf{UDV}' \text{ , } \mathbf{D} = [\mathbf{D}_1|\mathbf{0}]$$

where $\mathbf{U}$ is an $n \times n$ orthogonal matrix, $\mathbf{V}$ is a $p \times p$ orthogonal matrix and $\mathbf{D}_1 = \text{diag}(d_1, d_2, \ldots, d_n)$, with $d_1 \geq \ldots \geq d_n \geq 0$.

A.1.1 Case where σ^2 Known

Returning to the results of chapter 3, we replace $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}_\lambda$ in (3.2.2) and solve for λ. We obtain

$$\begin{aligned} q\sigma^2 &= \hat{\boldsymbol{\beta}}'_\lambda \boldsymbol{\Sigma}_X \hat{\boldsymbol{\beta}}_\lambda \\ &= [(\mathbf{X}'\mathbf{X} + \lambda\boldsymbol{\Sigma}_X)^{-1}\mathbf{X}'\mathbf{y}]'\boldsymbol{\Sigma}_X(\mathbf{X}'\mathbf{X} + \lambda\boldsymbol{\Sigma}_X)^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{y}'\mathbf{X}\boldsymbol{\Sigma}_X^{-1/2}[(\boldsymbol{\Sigma}_X^{-1/2}\mathbf{X}'\mathbf{X}\boldsymbol{\Sigma}_X^{-1/2} + \lambda\mathbf{I}]^{-1} \\ &\quad \times[(\boldsymbol{\Sigma}_X^{-1/2}\mathbf{X}'\mathbf{X}\boldsymbol{\Sigma}_X^{-1/2} + \lambda\mathbf{I}]^{-1}\boldsymbol{\Sigma}_X^{-1/2}\mathbf{X}'\mathbf{y} \,. \end{aligned}$$

We now use the SVD of $\mathbf{X}\boldsymbol{\Sigma}_X^{-1/2}$ to write

$$\begin{aligned} q\sigma^2 &= \mathbf{y}'\mathbf{U}\mathbf{D}\mathbf{V}'(\mathbf{V}\mathbf{D}'\mathbf{D}\mathbf{V}' + \lambda\mathbf{I})^{-2}\mathbf{V}\mathbf{D}'\mathbf{U}'\mathbf{y} \\ &= \mathbf{y}'\mathbf{U}\mathbf{D}(\mathbf{D}'\mathbf{D} + \lambda\mathbf{I})^{-2}\mathbf{D}'\mathbf{U}'\mathbf{y} \\ &= \mathbf{y}'\mathbf{U}\begin{bmatrix} d_1^2/(\lambda+d_1^2)^2 & 0 & \ldots & 0 \\ 0 & d_2^2/(\lambda+d_2^2)^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & d_n^2/(\lambda+d_n^2)^2 \end{bmatrix}\mathbf{U}'\mathbf{y} \,. \end{aligned}$$

Let $\mathbf{w} = \mathbf{U}'\mathbf{y}$. Then we must find λ to solve

$$\begin{aligned} q\sigma^2 &= \mathbf{w}'\begin{bmatrix} d_1^2/(\lambda+d_1^2)^2 & 0 & \ldots & 0 \\ 0 & d_2^2/(\lambda+d_2^2)^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & d_n^2/(\lambda+d_n^2)^2 \end{bmatrix}\mathbf{w} \\ &= \sum_{i=1}^{n} \frac{w_i^2 d_i^2}{(d_i^2+\lambda)^2} = C(\lambda) \end{aligned} \tag{A.1.1}$$

We denote the value of λ which satisfies this equation as $\hat{\lambda}_t$.

Next we wish to establish the conditions under which $\hat{\lambda}_t$ is unique. We'll do this based on the arguments of Golub and Van Loan (1989, pg. 562–564) and Björck (1990, pg. 596–600). To show it is unique we need the following to hold:

$$C(0) > q\sigma^2 \,, \; \lim_{\lambda\to\infty} C(\lambda) < q\sigma^2 \,.$$

It is obvious that $C(\lambda)$ is monotone decreasing in λ. When we find $C(0)$ we are using a generalized inverse of $\mathbf{X}'\mathbf{X}$ to give us a solution.

First we note that the limit of $C(\lambda)$, as $\lambda \to \infty$, is 0, which is less than $q\sigma^2$.

From (A.1.1) we see that

$$C(0) = \sum_{i=1}^{n} \left(\frac{w_i}{d_i}\right)^2$$

We can say that $C(0) > 0$, but nothing else beyond this. Therefore, to ensure that $\hat{\lambda}_t$ (the solution of (A.1.1)) is unique (if it exists), we must assume that $C(0) > q\sigma^2$.

A.1.2 Case Using $\hat{\sigma}_a^2$

Now we will examine $\hat{\lambda}_a$. We can use the arguments which showed (3.2.2) had a unique solution only when

$$\sum_{i=1}^{n} \left(\frac{w_i}{d_i}\right)^2 > q\sigma^2$$

to argue that $\hat{\lambda}_a$ is the unique solution to $\hat{\boldsymbol{\beta}}_\lambda' \boldsymbol{\Sigma}_X \hat{\boldsymbol{\beta}}_\lambda / \hat{\sigma}_a^2 = q$ only when

$$\sum_{i=1}^{n} \left(\frac{w_i}{d_i}\right)^2 > q\hat{\sigma}_a^2.$$

A.1.3 Case Using $\hat{\sigma}_{lik}^2$

We now examine $\hat{\lambda}_{lik}$, which solves

$$\frac{\hat{\boldsymbol{\beta}}_\lambda' \boldsymbol{\Sigma}_X \hat{\boldsymbol{\beta}}_\lambda}{\hat{\sigma}_{lik(\lambda)}^2} = q \ . \qquad \text{(A.1.2)}$$

From (A.1.1) we know that

$$\hat{\boldsymbol{\beta}}_\lambda' \boldsymbol{\Sigma}_X \hat{\boldsymbol{\beta}}_\lambda = \sum_{i=1}^{n} \frac{w_i^2 d_i^2}{(d_i^2 + \lambda)^2} \qquad \text{(A.1.3)}$$

Using the SVD of $\mathbf{X}\boldsymbol{\Sigma}_X^{-1/2}$ we find

$$\begin{aligned}
(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda) &= \mathbf{y}'(\mathbf{I}-\mathbf{UD}(\mathbf{D}'\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}'\mathbf{U}')' \\
&\quad \times (\mathbf{I}-\mathbf{UD}(\mathbf{D}'\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}'\mathbf{U}')\mathbf{U}'\mathbf{y} \\
&= \mathbf{y}'\mathbf{U}\ \mathrm{diag}\left(\frac{\lambda^2}{(\lambda+d_1^2)^2},\dots,\frac{\lambda^2}{(\lambda+d_n^2)^2}\right)\mathbf{U}'\mathbf{y} \\
&= \lambda^2\sum_{i=1}^{n}\frac{w_i^2}{(d_i^2+\lambda)^2} \qquad\qquad (\text{A.1.4})
\end{aligned}$$

We can use (A.1.3) and (A.1.4) to write $\hat{\sigma}^2_{lik(\lambda)}$ as

$$\hat{\sigma}^2_{lik(\lambda)} = \frac{\lambda^2}{n}\sum_{i=1}^{n}\frac{w_i^2}{(d_i^2+\lambda)^2} + \frac{\lambda}{n}\sum_{i=1}^{n}\frac{w_i^2 d_i^2}{(d_i^2+\lambda)^2} \tag{A.1.5}$$

Using (A.1.3) and (A.1.5) we can write the left-hand side of (A.1.2) as

$$\frac{\hat{\boldsymbol{\beta}}_\lambda'\boldsymbol{\Sigma}_X\hat{\boldsymbol{\beta}}_\lambda}{\hat{\sigma}^2_{lik(\lambda)}} = \frac{n\sum_{i=1}^{n} w_i^2 d_i^2/(d_i^2+\lambda)^2}{\lambda\sum_{i=1}^{n} w_i^2/(d_i^2+\lambda)} = \Psi(\lambda)\ ,$$

for convenience. We will show that $\hat{\lambda}_{lik}$ is the unique solution for $\Psi(\lambda) = q$. First,

$$\lim_{\lambda\to 0}\Psi(\lambda) = \infty > q\ ,\ \lim_{\lambda\to\infty}\Psi(\lambda) = 0 < q\ .$$

Finally,

$$\begin{aligned}
\frac{\partial\Psi(\lambda)}{\partial\lambda} &= \frac{-n}{\lambda^2\left[\sum_{i=1}^{n} w_i^2/(d_i^2+\lambda)\right]^2} \\
&\quad\times\left[2\lambda\sum_{i=1}^{n}\frac{w_i^2}{(d_i^2+\lambda)}\sum_{i=1}^{n}\frac{d_i^2 w_i^2}{(d_i^2+\lambda)^3}\right. \\
&\quad\left.+\sum_{i=1}^{n}\frac{d_i^2 w_i^2}{(d_i^2+\lambda)^2}\left(\sum_{i=1}^{n}\frac{w_i^2}{(d_i^2+\lambda)}-\lambda\sum_{i=1}^{n}\frac{w_i^2}{(d_i^2+\lambda)^2}\right)\right]
\end{aligned}$$

Since $\lambda > 0$ and $d_i^2 \geq 0$, the derivative will be negative if

$$\sum_{i=1}^{n}\frac{w_i^2}{(d_i^2+\lambda)} - \lambda\sum_{i=1}^{n}\frac{w_i^2}{(d_i^2+\lambda)^2} > 0\ .$$

But

$$\sum_{i=1}^{n}\frac{w_i^2}{(d_i^2+\lambda)} - \lambda\sum_{i=1}^{n}\frac{w_i^2}{(d_i^2+\lambda)^2} = \sum_{i=1}^{n}\frac{d_i^2 w_i^2}{(d_i^2+\lambda)^2} > 0\ ,$$

so $\Psi(\lambda)$ is monotone decreasing for $\lambda > 0$. Combining this with $\Psi(0) > q$ and $\lim_{\lambda \to \infty} \Psi(\lambda) < q$ tells us that $\hat{\lambda}_{lik}$ is the unique solution to (A.1.2). Therefore, for a given value of q, we will always be able to find $\hat{\lambda}_{lik}$. This is interesting improvement over the case where σ^2 is known. In that situation we had to make an assumption about the behavior of $\mathbf{X}\Sigma^{-1/2}$ at $\lambda = 0$.

Appendix B

Simulation Results with Known q

B.1 Introduction

This appendix contains the results of various simulations to supplement those given in chapter 3. The results in tables B.1 and B.2 summarize results for models in which the rows of $\mathbf{X}$ are generated from the AR(1) process

$$x_{ij} = \phi_1 x_{i,j-1} + \eta_{ij}, \quad j = 1, \ldots, p \tag{B.1.1}$$

where $\eta_{ij} \sim N(0, \sigma_\eta^2)$, and we use $\phi_1 = 0.5$ and $\sigma_\eta^2 = 1$. Table B.3 summarizes results from models in which the rows of $\mathbf{X}$ are generated from the AR(2) process (3.3.1). Each table summarizes results of 1000 simulated data sets, assuming that Σ_X is known.

We have included the results using $\hat{\lambda}_t$ and $\hat{\lambda}_a$ for illustrative purposes in table B.2, but we have omitted them from subsequent tables. This was because they gave estimators that performed worse based on predictive ability, and were less likely to satisfy the constraint equation for large values of q. Further details on this were given in chapter 3.

We follow these results with an examination of the effect that estimating Σ_X has on the smoothing parameters. The rows of $\mathbf{X}$ are generated from the AR(1) process (B.1.1) with $\phi_1 = 0.5$ and $\sigma_\eta^2 = 1$. We assume that the rows are generated using

(B.1.1), and we estimate ϕ_1 and σ^2_η. Table B.4 summarizes the results over 1000 simulated data sets. The $\hat{\lambda}$ should be compared to those in table B.2 to see that the estimation of Σ_X has had little effect on the resulting smoothing parameters. This is especially true for $\hat{\lambda}_{lik}$.

		q			
		15.19	2.16	1.22	0.78
	σ^2	25	36	64	100
$\hat{\sigma}^2_{lik(\hat{\lambda})}$	mean	14.78	21.37	38.55	61.64
	median	14.75	21.1	37.93	60.7
	var	1.90	11.63	46.13	132.1
$\hat{\sigma}^2_a$	mean	23.40	38.71	67.67	104.9
	median	23.41	38.48	67.18	103.6
	var	2.964	32.77	133.9	372.8

Table B.1: Summary of estimates of σ^2 when rows of $\mathbf{X}$ are generated from an AR(1) process. Models use $n = 50$, $p = 80$

		q			
		15.19	2.16	1.22	0.78
$\hat{\lambda}_{EPMSE}$		5.66	37.57	64.54	97.44
$\hat{\lambda}_{EGCV}$		2.32	36.46	59.42	87.15
$\hat{\lambda}_{lik}$	mean	3.033	16.22	25.32	35.24
	median	3.037	16.26	25.36	35.27
	var	1.47×10^{-3}	0.58	2.13	5.167
$\hat{\lambda}_t$	mean	0.00235	3.526	9.26	16.13
	median	0	3.771	8.96	15.58
	var	2.83×10^{-5}	6.95	16.67	31.24
$\hat{\lambda}_a$	mean	0	2.635	7.89	14.67
	median	0	2.611	7.96	14.81
	var	0	2.72	4.63	5.57
$\hat{\lambda}_{GCV}$	mean	3.586	32.24	61.39	6.32×10^7
	median	0.6958	28.01	46.88	69.29
	var	25.96	935.94	5931.79	1.15×10^{18}
PSS_{EPMSE}	mean	46.565	56.560	93.294	137.20
	median	45.314	55.611	91.873	135.27
	var	87.176	136.16	371.13	799.99
PSS_{EGCV}	mean	47.451	56.565	93.343	137.31
	median	46.413	55.628	91.854	135.57
	var	90.382	136.22	372.15	804.06
PSS_{lik}	mean	47.080	58.771	98.166	145.60
	median	45.948	57.618	96.162	143.25
	var	89.049	145.43	408.58	902.49
PSS_t	mean	49.855	65.511	108.10	158.50
	median	49.013	64.113	105.64	155.44
	var	99.761	173.85	480.32	1044.5
PSS_a	mean	49.856	67.012	110.02	160.68
	median	49.013	65.830	108.09	157.55
	var	99.750	183.33	502.95	1082.3
PSS_{GCV}	mean	48.638	61.594	102.80	151.96
	median	47.666	59.779	99.487	146.17
	var	97.001	200.03	615.53	1434.2

Table B.2: Comparison of $\hat{\lambda}$ and PSS values when rows of $\mathbf{X}$ generated from an AR(1) process. Models use $n = 50$, $p = 80$

		q	
		26.47	7.33
$\hat{\sigma}^2_{lik}$	mean	8.995	7.678
	median	8.984	7.664
	var	0.337	0.788
$\hat{\lambda}_{EPMSE}$		4.931	22.542
$\hat{\lambda}_{EGCV}$		1.17×10^{-3}	5.060
$\hat{\lambda}_{lik}$	mean	1.487	5.181
	median	1.487	5.184
	var	1.71×10^{-6}	1.08×10^{-3}
$\hat{\lambda}_{GCV}$	mean	0.369	9.703
	median	0	1.856
	var	3.981	201.85
PSS_{EPMSE}	mean	44.861	48.762
	median	43.933	47.750
	var	125.14	110.43
PSS_{EGCV}	mean	47.631	50.089
	median	46.622	49.183
	var	122.78	134.75
PSS_{lik}	mean	47.587	49.369
	median	46.582	48.386
	var	122.59	130.13
PSS_{GCV}	mean	47.933	49.972
	median	47.277	49.183
	var	133.22	132.60

Table B.3: Results when rows of **X** generated from AR(2) process. Models use $n = 40$, $p = 120$, $\sigma^2 = 25$

		q			
		15.19	2.16	1.22	0.78
	σ^2	25	36	64	100
$\hat{\sigma}^2_{lik(\hat{\lambda})}$	mean	14.97	21.55	38.79	61.93
	median	14.94	21.28	38.16	60.96
	var	1.94	11.83	46.76	133.5
λ_{EPMSE}		5.587	36.921	63.709	96.484
$\hat{\lambda}_{EGCV}$		2.285	36.654	59.678	87.568
$\hat{\lambda}_{lik}$	mean	3.030	16.19	25.28	35.18
	median	3.034	16.24	25.31	35.22
	var	1.49×10^{-3}	0.581	2.12	5.14
$\hat{\lambda}_{GCV}$	mean	3.530	32.519	62.670	6.68×10^{7}
	median	0.717	28.170	46.963	69.221
	var	25.079	961.80	7113.3	1.19×10^{18}

Table B.4: Estimates of σ^2 and $\hat{\lambda}$ values when rows of $\mathbf{X}$ generated from AR(1) process, correctly estimated as having AR(1) form. Models use $n = 50$, $p = 80$

Appendix C

Additional Robust Results

C.1 Introduction

In chapter 4 we presented empirical results that demonstrated our robust procedure appeared to have a bounded influence function in the $q = 7.21$ model, using our robust procedure with weights (4.5.2). Tables C.1 through C.3 illustrate the same results for three other choices of q.

We then present results on the empirical influence function of our second robust procedure, which uses the weights defined by (4.6.1). Tables C.4 to C.6 summarize the results when we increase the value of one element in $\mathbf{y}$ by a factor $s > 0$. We see the robust estimators have bounded influence functions for these values of q, while the nonrobust estimators have unbounded influence functions.

Tables C.7 to C.9 summarize the results when we increase the value of one row in $\mathbf{X}$ by a factor $s > 0$. We see that both procedures appear to have bounded influence functions in these situations, as we saw in chapter 4 for the $q = 7.21$ model.

		s			
		1	20.8	60.4	100
$\hat{\sigma}^2_m$	mean	13.700	15.222	15.953	16.486
	median	13.870	15.235	15.899	15.963
$\hat{\sigma}^2_{lik}$	mean	18.272	350.25	2909.8	7996.7
	median	17.509	234.42	1862.1	5081.6
$\hat{\lambda}_m$	mean	2.699	2.702	2.699	2.700
	median	2.707	2.710	2.709	2.711
$\hat{\lambda}_{lik}$	mean	2.617	2.490	2.460	2.454
	median	2.623	2.483	2.459	2.455
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	244.96	264.12	271.27	276.60
	median	250.85	256.72	264.82	266.26
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	295.69	6856.2	5.84×10^4	1.61×10^5
	median	278.45	4349.7	3.59×10^4	9.87×10^4
c		2	2	2	2

Table C.1: Empirical measure of change in influence function when one value in $\mathbf{y}$ modified, using first robust ridge regression proposal, $q = 17.37$

		s			
		1	20.8	60.4	100
$\hat{\sigma}^2_m$	mean	9.504	10.915	11.180	11.529
	median	9.550	10.655	10.853	10.899
$\hat{\sigma}^2_{lik}$	mean	11.002	231.65	1938.4	5332.4
	median	10.580	150.96	1218.8	3344.6
$\hat{\lambda}_m$	mean	1.766	1.766	1.766	1.766
	median	1.768	1.774	1.777	1.772
$\hat{\lambda}_{lik}$	mean	1.731	1.659	1.644	1.641
	median	1.734	1.656	1.646	1.640
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	234.92	267.74	272.28	279.90
	median	233.63	263.69	264.25	264.82
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	261.45	7100.4	6.10×10^4	1.68×10^5
	median	252.96	4559.1	3.70×10^4	1.02×10^5
c		2	2	2	2

Table C.2: Empirical measure of change in influence function when one value in **y** modified, using first robust ridge regression proposal, $q = 27.14$

		s			
		1	20.8	60.4	100
$\hat{\sigma}^2_m$	mean	4.174	4.692	4.884	4.946
	median	4.206	4.632	4.730	4.774
$\hat{\sigma}^2_{lik}$	mean	4.829	114.72	968.22	2666.7
	median	4.718	73.889	605.19	1665.7
$\hat{\lambda}_m$	mean	0.853	0.853	0.853	0.852
	median	0.854	0.854	0.853	0.853
$\hat{\lambda}_{lik}$	mean	0.846	0.824	0.819	0.818
	median	0.846	0.822	0.818	0.817
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	207.44	228.36	233.97	235.60
	median	207.58	217.68	220.44	219.64
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	222.60	7459.9	6.47×10^4	1.79×10^5
	median	216.97	4856.7	4.09×10^4	1.13×10^5
c		2	2	2	2

Table C.3: Empirical measure of change in influence function when one value in $\mathbf{y}$ modified, using first robust ridge regression proposal, $q = 57.44$

		s			
		1	20.8	60.4	100
$\hat{\sigma}^2_m$	mean	12.542	13.544	13.774	13.940
	median	12.504	13.793	13.848	13.904
$\hat{\sigma}^2_{lik}$	mean	14.113	291.20	2416.8	6637.5
	median	14.334	134.80	1012.9	2740.3
$\hat{\lambda}_m$	mean	2.729	2.734	2.731	2.732
	median	2.736	2.739	2.735	2.734
$\hat{\lambda}_{lik}$	mean	2.650	2.487	2.451	2.444
	median	2.657	2.489	2.452	2.446
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	218.27	231.27	232.23	233.73
	median	216.93	232.03	233.57	233.66
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	222.35	5785.8	5.0×10^4	1.39×10^5
	median	221.88	2396.1	1.95×10^4	5.34×10^5
c		2	2	2	2

Table C.4: Empirical measure of change in influence function when one value in **y** modified, using second robust ridge regression proposal, $q = 17.37$

		s			
		1	20.8	60.4	100
$\hat{\sigma}^2_m$	mean	8.206	8.900	9.084	9.114
	median	8.203	8.896	9.030	8.992
$\hat{\sigma}^2_{lik}$	mean	9.136	202.32	1686.8	4635.4
	median	9.267	86.32	649.31	1757.4
$\hat{\lambda}_m$	mean	1.778	1.779	1.778	1.777
	median	1.781	1.778	1.776	1.776
$\hat{\lambda}_{lik}$	mean	1.740	1.653	1.634	1.630
	median	1.743	1.663	1.639	1.637
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	207.66	221.58	223.36	223.62
	median	207.65	219.50	219.85	219.98
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	211.44	6286.3	5.47×10^4	1.51×10^5
	median	210.00	2384.0	1.94×10^4	5.34×10^4
c		2	2	2	2

Table C.5: Empirical measure of change in influence function when one value in $\mathbf{y}$ modified, using second robust ridge regression proposal, $q = 27.14$

		s			
		1	20.8	60.4	100
$\hat{\sigma}^2_m$	mean	4.018	4.386	4.421	4.439
	median	4.011	4.342	4.356	4.365
$\hat{\sigma}^2_{lik}$	mean	4.386	106.04	888.86	2444.3
	median	4.427	45.528	343.60	928.44
$\hat{\lambda}_m$	mean	0.855	0.855	0.855	0.855
	median	0.856	0.855	0.856	0.856
$\hat{\lambda}_{lik}$	mean	0.846	0.819	0.813	0.812
	median	0.847	0.822	0.815	0.814
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	196.16	211.80	213.41	213.99
	median	196.04	204.79	209.06	209.11
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	199.29	6976.8	6.09×10^4	1.69×10^5
	median	198.91	2883.5	2.36×10^4	6.49×10^4
c		2	2	2	2

Table C.6: Empirical measure of change in influence function when one value in **y** modified, using second robust ridge regression proposal, $q = 57.44$

		s					
		1	20.8	40.6	60.4	80.2	100
$\hat{\sigma}^2_m$	mean	12.542	12.622	12.657	12.672	12.681	12.686
	median	12.504	12.763	12.934	12.934	12.934	12.934
$\hat{\sigma}^2_{lik}$	mean	14.113	14.202	14.204	14.205	14.205	14.205
	median	14.334	14.334	14.327	14.324	14.323	14.322
$\hat{\lambda}_m$	mean	2.729	2.736	2.736	2.735	2.735	2.735
	median	2.736	2.743	2.743	2.743	2.743	2.743
$\hat{\lambda}_{lik}$	mean	2.650	2.655	2.655	2.655	2.655	2.655
	median	2.657	2.671	2.671	2.670	2.670	2.670
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	218.27	216.97	217.02	217.06	217.08	217.09
	median	216.93	213.84	213.80	213.78	213.77	213.77
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	222.35	221.16	221.14	221.14	221.14	221.14
	median	221.88	217.40	217.58	217.64	217.67	217.69
c		2	2	2	2	2	2

Table C.7: Empirical Measure of Change in Influence Function when one row in $\mathbf{X}$ modified, using second robust ridge regression proposal, $q = 17.37$

		s					
		1	20.8	40.6	60.4	80.2	100
$\hat{\sigma}^2_m$	mean	8.206	8.280	8.297	8.302	8.305	8.306
	median	8.203	8.524	8.526	8.526	8.526	8.526
$\hat{\sigma}^2_{lik}$	mean	9.136	9.186	9.187	9.188	9.188	9.188
	median	9.267	9.234	9.230	9.228	9.228	9.227
$\hat{\lambda}_m$	mean	1.778	1.780	1.780	1.780	1.780	1.780
	median	1.781	1.784	1.784	1.784	1.784	1.784
$\hat{\lambda}_{lik}$	mean	1.740	1.743	1.743	1.743	1.743	1.743
	median	1.743	1.750	1.750	1.750	1.750	1.750
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	207.66	206.82	206.89	206.91	206.92	206.93
	median	207.65	204.20	204.17	204.16	204.15	204.15
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	211.44	210.61	210.60	210.59	210.59	210.59
	median	210.00	208.75	208.90	208.95	208.97	208.99
c		2	2	2	2	2	2

Table C.8: Empirical Measure of Change in Influence Function when one row in $\mathbf{X}$ modified, using second robust ridge regression proposal, $q = 27.14$

		s					
		1	20.8	40.6	60.4	80.2	100
$\hat{\sigma}^2_m$	mean	4.018	4.057	4.059	4.060	4.060	4.060
	median	4.011	4.146	4.146	4.145	4.145	4.145
$\hat{\sigma}^2_{lik}$	mean	4.386	4.404	4.405	4.405	4.405	4.405
	median	4.427	4.418	4.421	4.422	4.423	4.423
$\hat{\lambda}_m$	mean	0.855	0.856	0.856	0.856	0.856	0.856
	median	0.856	0.856	0.856	0.856	0.856	0.856
$\hat{\lambda}_{lik}$	mean	0.846	0.846	0.846	0.846	0.846	0.846
	median	0.847	0.848	0.848	0.848	0.848	0.848
$(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$	mean	196.16	195.91	195.93	195.94	195.94	195.94
	median	196.04	194.81	194.75	194.73	194.72	194.72
$(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{lik} - \boldsymbol{\beta})$	mean	199.29	198.86	198.86	198.86	198.86	198.86
	median	198.91	197.87	197.84	197.83	197.83	197.82
c		2	2	2	2	2	2

Table C.9: Empirical Measure of Change in Influence Function when one row in **X** modified, using second robust ridge regression proposal, $q = 57.44$

Bibliography

Altman, N. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association* **85**, 749–759.

Anderson, T. (1971). *The Statistical Analysis of Time Series.* New York: Wiley.

Andrews, D. (1991). Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics* **47**, 359–377.

Askin, R. G. and D. C. Montgomery (1980). Augmented robust estimators. *Technometrics* **22**, 333–341.

Barry, D. (1995). A Bayesian analysis for a class of penalised likelihood estimators. *Communications in Statistics A: Theory and Methods* **24**, 1057–1071.

Bennett, A. (1992). *Inverse Methods in Physical Oceanography.* London: Cambridge University Press.

Björck, Å. (1990). Least squares methods. In P. Ciarlet and J. Lions (Eds.), *Handbook of Numerical Analysis, vol.1: Finite Difference Methods-Solution of Equations in* $\mathbf{R}^n$. New York: Elsevier.

Björkström, A. and R. Sundberg (1996). Continuum regression is not always continuous. *Journal of the Royal Statistical Society, Series B* **58**, 703–710.

Boswell-Purdy, J. E. (1995). Partial least squares and a modification, with an application to predicting glucose concentration in plasma. Master's thesis, Dalhousie University, Halifax NS.

Breiman, L. and J. H. Friedman (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society, Series B* **59**, 3–54.

Bretherton, F. P., R. E. Davis, and C. Fandry (1976). A technique for objective analysis and design of oceanographic experiments applied to MODE-73. *Deep-Sea Research* **23**, 559–582.

Brockwell, P. J. and R. A. Davis (1991). *Time Series: Theory and Methods*, second edition. New York: Springer-Verlag.

Carroll, R., D. Ruppert, and L. Stefanski (1995). *Measurement Error in Nonlinear Models*. New York: Chapman and Hall.

Casella, G. and R.L. Berger (1990). *Statistical Inference*. Belmont, Ca.: Wadsworth.

Chiu, S. (1989). Bandwidth selection for kernel estimation with correlated noise. *Statistics and Probability Letters* **8**, 347–354.

Chu, C. and J. Marron (1991). Comparison of two bandwidth selectors with dependent errors. *Annals of Statistics* **4**, 1906–1918.

Clements, J. C., R. Carroll, and M. Horáček (1996). On regularization parameters for inverse problems in electrocardiography. In D. Ghista (Ed.), *Biomedical and Life Physics*. Wiesbaden: Vieweg Verlag.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 377–403.

Delaney, N. J. and S. Chaterjee (1986). Use of the bootstrap and cross-validation in ridge regression. *Journal of Business and Economic Statistics* **4**, 255–262.

Diggle, P. and M. Hutchinson (1988). On spline smoothing with autocorrelated errors. *Australian Journal of Statistics* **31**, 166–182.

Dowd, M. and K. R. Thompson (1996). Extraction of tidal streams from a shipborne acoustic Doppler current profiler using a statistical-dynamical model. *Journal of Geophysical Research* **101**, 8943–8956.

Field, C. and B. Smith (1994). Robust estimation-a weighted maximum likelihood approach. *International Statistical Review* **62**, 405–424.

Frank, I. and J. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–148.

Fuller, W. (1987). *Measurement Error Models*. New York: Wiley.

Golub, G., M. Heath, and G. Wahba (1979). Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215–223.

Golub, G. H. and C. F. Van Loan (1989). *Matrix Computations*, second edition. Baltimore: Johns Hopkins University Press.

Hall, P. and D. Titterington (1987). Common structure of techniques for choosing smoothing parameters in regression problems. *Journal of the Royal Statistical Society, Series B* **49**, 184–198.

Hampel, F., E. Ronchetti, P. Rousseeuw, and W. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

Haney, R. L., R. A. Hale, and C. A. Collins (1995). Estimating subpycnocline density fluctuations in the California Current region from upper ocean observations. *Journal of Atmospheric and Oceanic Technology* **12**, 550–566.

Hansen, P. (1992). Analysis of discrete ill-posed problems by means of the L-curve. *Siam Review* **34**, 561–580.

Hart, J. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society, Series B* **53**, 173–187.

Harvey, A. C. (1993). *Time Series Models*, second edition. Bodmin, U.K.: Harvester Wheatsheaf.

Hastie, T. and C. Mallows (1993). In discussion of A statistical view of some chemometrics regression tools. *Technometrics* **35**, 140–143.

Helland, I. (1988). On the structure of partial least squares regression. *Communications in Statistics B: Simulation and Computation* **17**, 581–607.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Annals of Statistics* **1**, 799–821.

Kay, J. (1992). Asymptotic comparison factors for smoothing parameter choices in regression problems. *Statistics and Probability Letters* **15**, 329–335.

Kohn, R., C. Ansley, and C. Wong (1992). Nonparametric spline regression with autoregressive moving average errors. *Biometrika* **79**, 335–346.

Krasker, W. S. and R. E. Welsch (1982). Efficient bounded-influence regression estimation. *Journal of the American Statistical Association* **77**, 595–604.

Lawless, J. (1978). Ridge and related estimation procedures: theory and practice. *Communications in Statistics A: Theory and Methods* **7**, 139–164.

Lawrence, K. D. and L. C. Marsh (1984). Robust ridge estimation methods for predicting U.S. coal mining fatalities. *Communications in Statistics A: Theory and Methods* **13**, 130–149.

Li, K. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Annals of Statistics* **14**, 1101–1112.

Lindley, D. and F. Smith (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B* **34**, 1–40.

Manchester, L. and K. Haller (1997). A new approach to undetermined regression. Presented at the Annual Meeting of The Statistical Society of Canada, Fredericton NB.

Marazzi, A. (1993). *Algorithms, Routines, and S Functions for Robust Statistics.* New York: Chapman and Hall.

Nebebe, F. and T. Stroud (1986). Bayes and empirical Bayes shrinkage estimation of regression coefficients. *Canadian Journal of Statistics* **14**, 267–280.

Neter, J., W. Wasserman, and M. Kutner (1985). *Applied Linear Statistical Models*, second edition. Homewood, Il.: Irwin.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science* **1**, 502–527.

O'Sullivan, F. and G. Wahba (1985). A cross-validated Bayesian retrieval algorithm for nonlinear remote sensing experiments. *Journal of Computational Physics* **59**, 441–455.

Pfaffenberger, R. C. and T. E. Dielman (1990). A comparison of regression estimators when both multicollinearity and outliers are present. In K. D. Lawrence and J. L. Arthur (Eds.), *Robust Regression: Analysis and Applications*. New York: Marcel Dekker.

Priestley, M. (1981). *Spectral Analysis and Time Series*. London: Academic Press.

Ronchetti, E., C. Field, and W. Blanchard (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association* **999**, 999–999. To appear.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association* **88**, 486–494.

Silvapulle, M. J. (1991). Robust ridge regression based on an M-estimator. *Australian Journal of Statistics* **33**, 319–333.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B* **36**, 111–147.

Stone, M. and R. Brooks (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society, Series B* **52**, 237–269.

Tarantola, A. (1987). *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*. New York: Elsevier.

Thacker, W. (1988). Fitting models to inadequate data by enforcing spatial and temporal smoothness. *Journal of Geophysical Research* **93**, 10655–10665.

Thacker, W. C. and R. B. Long (1988). Fitting dynamics to data. *Journal of Geophysical Research* **93**, 1227–1240.

Thompson, A., J. Kay, and D. Titterington (1989). A cautionary note about cross-validatory choice. *Journal of Statistical Computation and Simulation* **33**, 199–216.

Tikhonov, A. and V. Arsenin (1977). *Solutions of Ill-Posed Problems.* New York: Wiley.

Titterington, D. (1985). Common structure of smoothing techniques in statistics. *International Statistical Review* **53**, 141–170.

van der Linde, A. (1994). On cross-validation for smoothing splines in the case of dependent errors. *Australian Journal of Statistics* **36**, 67–73.

Van Huffel, S. (Ed.) (1997). *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling.* Philadelphia: SIAM.

Van Huffel, S. and J. Vandewalle (1991). *The Total Least Squares Problem: Computational Aspects and Analysis.* Philadelphia: SIAM.

Van Loan, C. (1976). Generalizing the singular value decomposition. *SIAM Journal on Numerical Analysis* **13**, 76–83.

Victoria-Feser, M.-P. and E. Ronchetti (1994). Robust methods for personal income models. *Canadian Journal of Statistics* **22**, 1247–1258.

Vogel, A., C. O. Ofoegbu, R. Gorenflo, and B. Ursin (Eds.) (1990). *Geophysical Data Inversion Methods and Applications.* Weisbaden-Braunschweig: Vieweg.

Wahba, G. (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. In P. Krishnaiah (Ed.), *Applications of Statistics.* New York: North-Holland.

Wahba, G. (1978). Smoothing and ill-posed problems. In M. A. Goldberg (Ed.), *Solution Methods for Integral Equations-Theory and Applications.* New York: Plenum.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society, Series B* **45**, 133–150.

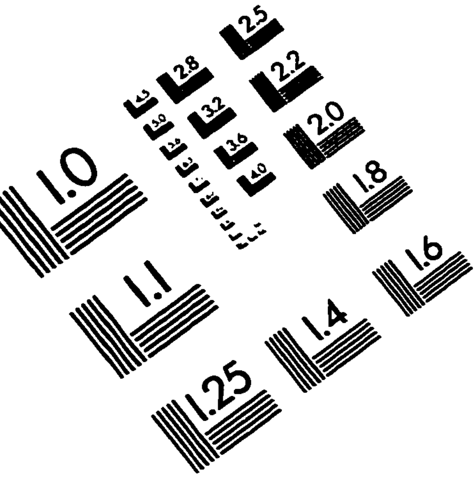
Wahba, G. (1990). *Spline models for Observational Data.* Philadelphia: SIAM. CBMS-NSF Regional Conference series in applied mathematics, volume 59.

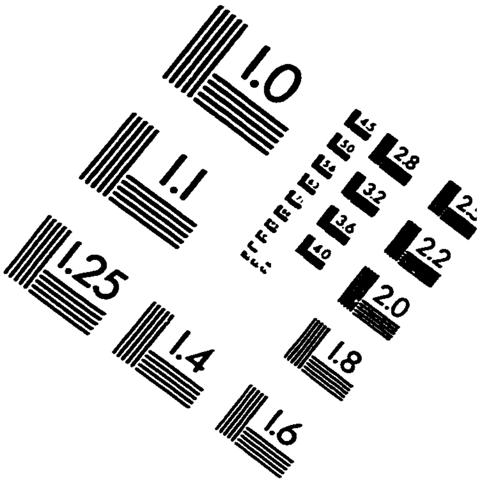
Wahba, G., D. Johnson, F. Gao, and J. Gong (1995). Adaptive tuning of numerical weather prediction models: randomized GCV in three and four dimensional data assimilation. *Monthly Weather Review* **123**, 3358–3369.

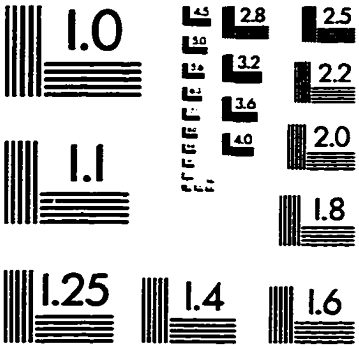
Wahba, G. and J. Wendelberger (1980). Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly Weather Review* **108**, 1122–1143.

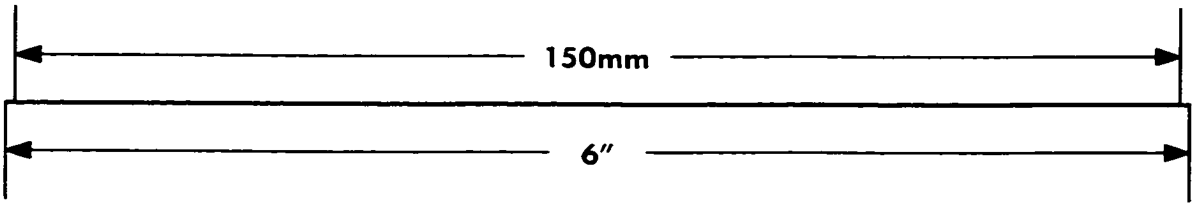
Weisberg, S. (1985). *Applied Linear Regression*, second edition. New York: Wiley.

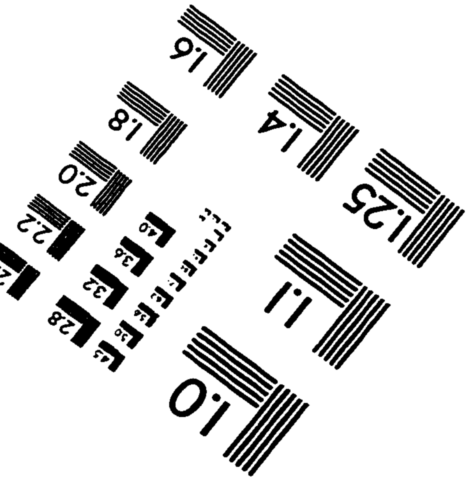
IMAGE EVALUATION TEST TARGET (QA-3)











APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989