

**ROBUST ESTIMATION OF UNKNOWN~~S~~ IN A  
LINEAR SYSTEM OF EQUATIONS WITH MODELING  
UNCERTAINTIES**

**A THESIS  
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND  
ELECTRONICS ENGINEERING  
AND THE INSTITUTE OF ENGINEERING AND SCIENCES  
OF BILKENT UNIVERSITY  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
MASTER OF SCIENCE**

**By  
Fehmi CHEBIL**

**July 1997**

QA  
276.8  
.C44  
1997

ROBUST ESTIMATION OF UNKNOWN IN A  
LINEAR SYSTEM OF EQUATIONS WITH MODELING  
UNCERTAINTIES

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND

ELECTRONICS ENGINEERING

AND THE INSTITUTE OF ENGINEERING AND SCIENCES

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Fehmi Chebil  
*da'afodan baji'la' masta*

By

Fehmi Chebil

July 1997

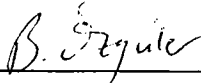
QA  
276.8  
.C44  
1997

B- 038255

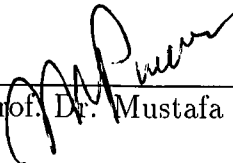
I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

  
Assist. Prof. Dr. Orhan Arıkan(Supervisor)

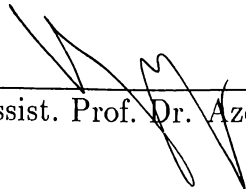
I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

  
Prof. Dr. A. Bülent Özgüler

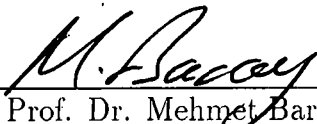
I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

  
Assist. Prof. Dr. Mustafa Çelebi Pınar.

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

  
Assist. Prof. Dr. Azer Kerimov

Approved for the Institute of Engineering and Sciences:

  
Prof. Dr. Mehmet Baray  
Director of Institute of Engineering and Sciences

## ABSTRACT

### ROBUST ESTIMATION OF UNKNOWNNS IN A LINEAR SYSTEM OF EQUATIONS WITH MODELING UNCERTAINTIES

Fehmi Chebil

M.S. in Electrical and Electronics Engineering

Supervisor: Assist. Prof. Dr. Orhan Arikan

July 1997

Robust methods of estimation of unknowns in a linear system of equations with modeling uncertainties are proposed. Specifically, when the uncertainty in the model is limited to the statistics of the additive noise, algorithms based on adaptive regularized techniques are introduced and compared with commonly used estimators. It is observed that significant improvements can be achieved at low signal-to-noise ratios. Then, we investigated the case of a parametric uncertainty in the model matrix and proposed algorithms based on non-linear ridge regression, maximum likelihood and Bayesian estimation that can be used depending

on the amount of prior information. Based on a detailed comparison study between the proposed and available methods, it is shown that the new approaches provide significantly better estimates for the unknowns in the presence of model uncertainties.

*Keywords:* Robust Estimation, Parametric measurement uncertainties, Ridge Regression, Wavelet based reconstruction, Mean Square Error.

## ÖZET

### BENZETİM BELİRSİZLİKLERİ OLAN DOĞRUSAL DENKLEM SİSTEMLERİNDE BİLİNMEYENLERİN GÜRBÜZ KESTİRİMİ

Fehmi Chebil

Elektrik ve Elektronik Mühendisliği Bölümü Yüksek Lisans

Tez Yöneticisi: Yardımcı Doçent Orhan Arıkan

Temmuz 1997

Doğrusal denklem sistemlerinde bilinmeyenlerin kestiriminde kullanılmak üzere pekçok yöntem önerilmiştir. Sistemin belirsizlikler içermesi durumunda kestirim başarımı yüksek gürbüz yöntemlere duyulan ihtiyaç nedeniyle, tez kapsamında yeni yöntemler önerilmektedir. Sistem belirsizliğinin ölçüm gürültüsünün istatistiksel tanımlanması üzerinde olduğu durumlarda kullanılmak üzere önerdiğimiz yöntemler kullanılmakta olan yöntemler ile kıyaslanmış ve oldukça daha iyi kestirimler elde edilebildiği gösterilmiştir. Özellikle sinyal-gürültü oranının düşük olduğu durumlarda yeni yöntemler çok daha iyi kestirimler verebilmektedir. Parametrik yapıya sahip sistem matrislerinde belirsizlikler olması durumunda

kullanılabilecek yeni kestirim yöntemleri de önerilmektedir. Bu yöntemlerin kullanılmakta olan diğer yöntemlerde olan detaylı kıyaslamasında yeni yöntemlerin daha gürbüz ve yüksek başarımlı kestirim sonuçları verebildiği gösterilmiştir.

*Anahtar Kelimeler:* Gürbüz Kestirim, Parametrik Ölçüm Belirsizlikleri, Diyagonal Düzenleştirme, Dalgacık tabanlı oluşturma, Hata Karesinin Ortalaması.



## ACKNOWLEDGEMENT

I gratefully thank my supervisor Assist. Prof. Dr. Orhan Arıkan for his suggestions, supervision and guidance throughout the development of this thesis. Moreover, I would like to acknowledge his positive personality that turned working with him to a pleasure.

I would also like to thank Prof. Dr. A. Bülent Özgüler, Assist. Prof. Dr. M. Ç. Pınar and Assist. Prof. Dr. Azer Kerimov, the members of my jury, for reading and commenting on the thesis.

An exhaustive list of those for whom I want to dedicate this work is by no means possible. However, I would like to thank all my friends for their friendship and my officemates for the favourable environment they have provided throughout the achievement of this work.

Last, but not least, special thanks to my parents for their never ending support.

# Contents

<b>1</b>	<b>INTRODUCTION</b>	<b>1</b>
<b>2</b>	<b>Commonly Used Estimation Approaches</b>	<b>4</b>
2.1	Introduction . . . . .	4
2.2	Known Measurement Kernel . . . . .	5
2.2.1	Least Squares Fitting to the Measurements . . . . .	5
2.2.2	Ridge Regression . . . . .	10
2.2.3	Simulation Results . . . . .	14
2.3	Uncertain Model	21
2.3.1	Total Least Squares . . . . .	22
2.3.2	Simulation Results . . . . .	25
2.3.3	Nonlinear Least Squares Modeling . . . . .	26

2.3.4	Simulation Results . . . . .	29
<b>3</b>	<b>Proposed Estimation Methods</b>	<b>33</b>
3.1	Introduction . . . . .	33
3.2	Known Measurement Kernel . . . . .	34
3.2.1	Error Dependent Ridge Regression Constant . . . . .	34
3.2.2	Simulation Results . . . . .	35
3.2.3	Gauss-Markov Estimate with recursive updates	36
3.2.4	Simulation Results . . . . .	40
3.2.5	A Wavelet Based Recursive Reconstruction Algorithm . . .	41
3.2.6	Simulation Results . . . . .	46
3.2.7	Comparing performances . . . . .	47
3.3	Uncertain Model	50
3.3.1	Nonlinear Ridge Regression Modeling . . . . .	50
3.3.2	Maximum Likelihood and Least Squares Bayesian Inversion Approaches . . . . .	54
3.3.3	Simulation Results and Comparing Performances	56

<i>CONTENTS</i>	x
<b>4 Conclusions</b>	<b>63</b>
<b>APPENDIX</b>	<b>65</b>
<b>A Computation of the Jacobian matrices.</b>	<b>66</b>

# List of Figures

2.1	Maximum Likelihood principle:Typical density functions. . . . .	7
2.2	Least Squares and Ridge Regression Estimators: Bias and Covariance. . . . .	13
2.3	Least Squares Estimator, % error= 4.44.	17
2.4	Theobald Estimator, % error= 4.44. . . . .	17
2.5	Schmidt Estimator, % error= 4.44.	17
2.6	Swamy Mehta and Rappoport Estimator, % error= 4.45. . . . .	18
2.7	Goldstein Estimator, % error= 11.46. . . . .	18
2.8	Least Squares Estimator, % error= 113. . . . .	18
2.9	Theobald Estimator, % error= 19.3. . . . .	19
2.10	Schmidt Estimator, % error= 23.7.	19
2.11	Swamy Mehta and Rappoport Estimator, % error= 16.2. . . . .	19

2.12 Goldstein Estimator, % error= 26.8. . . . .	20
2.13 Estimation error versus SNR for Least Squares(LS), Theobald, Schmidt, Swamy-Mehta-Rappoport(SMR) and Goldstein-Smith(GS). 20	
2.14 Estimation error versus kernel matrix condition number for Least Squares(LS), Theobald, Schmidt, Swamy-Mehta-Rappoport(SMR) and Goldstein-Smith(GS). . . . .	21
2.15 TLS and OLS % $e_{TLS}$ = 31.2, % $e_{OLS}$ = 47.6 .	26
2.16 Effect of $\theta$ on the $R(A)$ .	30
2.17 Application of Cadzow's algorithm with SNR=80dB and kernel of low condition number, %error= 3.2 . . . . .	31
2.18 Non linear least squares modeling algorithm with SNR=28dB, %error = 64.2 . . . . .	31
2.19 Comparing TLS and Nonlinear Least Squares Modeling, % $error_{TLS}$ = 71.5 and % $error_{Cad}$ = 24.1	32
3.1 Choice of Ridge Regression: increasing curve is the sample variance of the fit error vector as a function of $\mu$ , horizontal line is $\sigma^2$ . . . .	35
3.2 Error Dependent Ridge Regression(EDRR) with % error = 12.4 and Swamy-Mehta-Rappoport (SMR) with % error = 21.5 Esti- mates . . . . .	36

3.3	Application of Gauss-Markov with recursive updates algorithm. SNR=45dB and low kernel matrix condition number, %error= 2. .	41
3.4	Application of Gauss-Markov with recursive updates algorithm. SNR=45dB and high kernel matrix condition number, %error= 8.	41
3.5	Fit Error and Magnitude of the estimate versus the number of basis components. . . . .	47
3.6	Reconstructed estimate by using 10 components of the basis. . . .	47
3.7	WBRR, EDRR, GM, SMR estimates, with %errors: $e_{WBRR} =$ $18.37$ , $e_{EDRR} = 15.23$ , $e_{GM} = 13.82$ , $e_{SMR} = 24.54$ . . . . .	48
3.8	Estimation Error vs. SNR for the WBRR, EDRR, GM and SMR estimates. . . . .	49
3.9	Estimation Error vs. SNR for the WBRR and EDRR estimates.	49
3.10	Nonlinear Ridge Regression algorithm with non linear minimiza- tion technique. SNR=65dB, %error = 3.11 . . . . .	57
3.11	Nonlinear Ridge Regression algorithm with non linear minimiza- tion technique. SNR=46dB and $\kappa = 10^4$ , %error = 7.42	57
3.12	Nonlinear Ridge Regression algorithm with Gradient technique. SNR=46dB, $\kappa = 10^4$ . %error = 3.24.	58
3.13	Maximum Likelihood-Bayesian approach, %error = 13.17. . . . .	59
3.14	Estimation error versus the bound vector norm. . . . .	59

3.15 Least Squares-Bayesian approach, %error = 12.75. . . . .	59
3.16 Nonlinear Least Squares Modeling, %error = 24.36. . . . .	60
3.17 Nonlinear Ridge Regression Modeling with non linear optimization, %error = 16.7. . . . .	61
3.18 Nonlinear Ridge Regression Modeling with gradient descent minimization, %error = 15.9. . . . .	61
3.19 Likelihood Bayesian approach, %error = 21.8. . . . .	61
3.20 Estimation Error versus SNR for the presented algorithms: Cadzow(Cad), Nonlinear Ridge Regression(RR), Nonlinear Ridge Regression with Gradient descent algorithm(Gradient), Bayesian-Likelihood(BLik) and Bayesian-Least Squares(BLeast). . . . .	62



**To my family ...**

# Chapter 1

## INTRODUCTION

The basic job of an experimenter is to describe what he or she sees, try to explain what is observed and use this knowledge to help answer questions encountered in the future. The explanation often takes the form of a physical model, which is a theoretical explanation of the physical phenomenon under study. Models make it possible to explore situations which in the actual system would be hazardous or demanding. Aircraft and space vehicle simulators are well known examples. A model is usually expressed verbally first then formalized into one or more equations giving rise to the mathematical model. A characteristic of science is its use of mathematical models to extract the essentials from complicated evidence and to quantify the implications.

An important reason behind modeling is to provide the required framework for the estimation of the unknowns. Experience has shown that no measurement, however carefully made, can be completely free of errors. In science the word

“error” does not carry the usual connotations of mistake. Error in a scientific measurement means the inevitable uncertainty that attends measurements. Uncertainty is not the ignorance of outcomes. As a matter of fact when a coin is tossed, we are certain that one of two outcomes will occur. What is not known is heads or tails. Again, when a die is tossed, it is certain that 1, 2, 3, 4, 5 or 6 will turn up. What is not known is which of these numbers. The future outcome of a coin toss or a die toss is not only unknown but also not knowable in advance. Thus uncertainty is the certainty that one of several outcomes will occur; but which specific outcome will prevail is unknown and unknowable.

A basic problem that arises in a broad class of scientific disciplines is to perform estimation of certain parameters from a model within uncertainties. In this thesis, we treat this problem when the model is a linear statistical one, which is described by:

$$\mathbf{A} \mathbf{x} = \mathbf{y} \quad , \quad (1.1)$$

where  $\mathbf{x}$  is the unknown vector,  $\mathbf{y}$  is the measurement vector and  $\mathbf{A}$  is the measurement kernel. As mentioned previously, there are no measurements free of error, the obtained data presented in the vector  $\mathbf{y}$  are considered to be erroneous. An additive noise vector  $\mathbf{n}$  is added to the observation to stress that fact. The uncertainty could come from the kernel matrix  $\mathbf{A}$ , the entries of this matrix are also subject to sampling errors, measurement errors, modeling errors and instrument errors. Again the matrix  $\mathbf{A}$  could depend on an unknown real valued set of parameters  $\boldsymbol{\theta}$  belonging to a set  $S$ . This is the case of array signal processing applications where  $\boldsymbol{\theta}$  refers to direction of arrivals of signals. Thus the problem we are dealing with is estimating an  $M$  dimensional vector  $\mathbf{x}$  from an  $N$  dimensional

data vector  $\mathbf{y}$  with:

$$\mathbf{y} = \mathbf{A}(\boldsymbol{\theta})\mathbf{x} + \mathbf{n} \quad . \quad (1.2)$$

In chapter 2, some of the commonly used approaches to the estimation of the unknowns in the presence of measurement uncertainties will be presented. In chapter 3, we will introduce the proposed approaches to the estimation problem. In order to compare the estimation performance of the old and new approaches, extensive simulations are provided throughout the thesis.

## Chapter 2

# Commonly Used Estimation Approaches

### 2.1 Introduction

The commonly used approaches to estimate the unknown parameters from a model under uncertainties are presented in this chapter. Over synthetically generated examples, these approaches are compared with each other in terms of their performances. In the following, measurement relationship is modeled as:

$$\mathbf{y} = \mathbf{A}(\boldsymbol{\theta})\mathbf{x} + \mathbf{n} \quad , \quad (2.1)$$

where  $\mathbf{y}$  is the  $N$ -dimensional vector of available measurement data,  $\mathbf{A}$  is the measurement kernel or operator,  $\mathbf{x}$  is the  $M$ -dimensional unknown vector,  $\mathbf{n}$  is the additive measurement noise and  $\boldsymbol{\theta}$  is  $P$ -dimensional vector parameterizing the uncertainty in the model.

We shall start by providing the approaches used for a fixed  $\boldsymbol{\theta}$ , that is to solve the overdetermined set of equations:

$$\mathbf{y} = \mathbf{A} \mathbf{x} + \mathbf{n} \quad . \quad (2.2)$$

We will investigate the Least Squares approach, then the ridge regression estimate. For the model uncertainty problem which is characterized by equation 2.1, we will consider the Total Least Squares estimate and the nonlinear least squares modeling algorithm.

## 2.2 Known Measurement Kernel

### 2.2.1 Least Squares Fitting to the Measurements

The least squares method of estimation is extensively utilized in a wide variety of applications such as communications, control, signal processing and numerical analysis, since it requires no information on the statistics of the data, and it is usually simple to implement. As we will see, it provides reasonably good estimates when the condition number of  $\mathbf{A}$  is relatively small and the signal to noise ratio (SNR) of the measurements is high.

In the method of least squares, we want to find an estimate  $\hat{\mathbf{x}}$  such that the norm of the fit error

$$\mathbf{e} = \mathbf{y} - \mathbf{A} \hat{\mathbf{x}} \quad (2.3)$$

is minimized.

The least squares estimate satisfies the well known normal equations:

$$(\mathbf{A}^H \mathbf{A}) \hat{\mathbf{x}}_{LS} = \mathbf{A}^H \mathbf{y} \quad (2.4)$$

If  $(\mathbf{A}^H \mathbf{A})$  is full rank then the least squares solution can be found as:

$$\hat{\mathbf{x}}_{LS} = (\mathbf{A}^H \mathbf{A})^{-1} \mathbf{A}^H \mathbf{y} \quad (2.5)$$

When  $(\mathbf{A}^H \mathbf{A})$  is rank deficient, the least squares estimator is given by:

$$\hat{\mathbf{x}}_{LS} = \mathbf{A}^\dagger \mathbf{y} \quad , \quad (2.6)$$

where  $\mathbf{A}^\dagger$  is called the pseudo-inverse or the Moore Penrose generalized inverse of  $\mathbf{A}$ , which can be obtained from the singular value decomposition (SVD) of  $\mathbf{A}$ .

When the measurement noise vector has independent identically distributed normal entries, the least squares estimator also corresponds to the maximum likelihood estimator. The maximum likelihood theory is widely applied to a number of important applications in signal processing such as system identification, array signal processing and signal decomposition. It is also applied to find an estimate to uncertain model parameters. The principle of maximum likelihood is illustrated by the following example [1].

Let  $y$  be a random variable for which the probability density function  $f_x(y)$  is parameterized by an unknown parameter  $x$ . A typical density function is given in figure 2.1. In this figure two densities are illustrated, one for parameter  $x_1$  and one for parameter  $x_2$ . Suppose that the value  $\hat{y}$  is observed. Based on the prior model  $f_x(y)$  shown in 2.1 we can say that  $\hat{y}$  is more probably observed when  $x = x_2$  than when  $x = x_1$ . More generally there may be a unique value of  $x$  for which  $\hat{y}$  is more probably observed than for any other. We call this value of

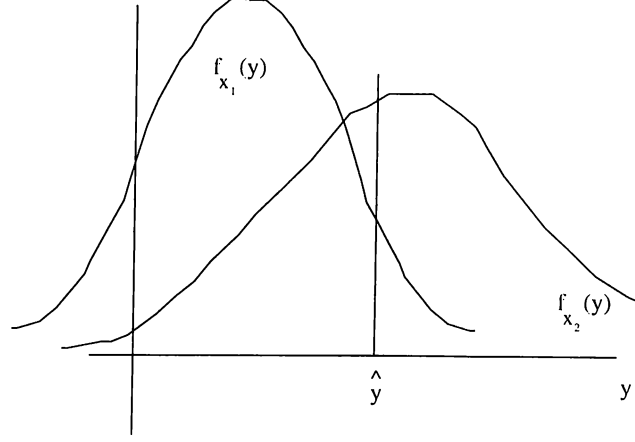


Figure 2.1: Maximum Likelihood principle: Typical density functions.

$x$  that makes  $\hat{y}$  most probable, or most likely, the maximum likelihood estimate  $\hat{x}_{ML}$ :

$$\hat{x}_{ML} = \arg \max_x f_x(\hat{y}) . \quad (2.7)$$

We obtain the maximum likelihood estimate by evaluating the conditional density  $f_{Y|X}(y|x)$  at the value of observation  $\hat{y}$  and then searching for the value of  $x$  that maximizes  $f_{Y|X}(\hat{y}|x)$ . The function  $l(x, \hat{y}) = f_{Y|X}(\hat{y}|x)$  is called the likelihood function and its logarithm  $L(x, \hat{y}) = \ln f_{Y|X}(\hat{y}|x)$  is called the log likelihood function.

In our problem we have  $N$  observations summarized in the vector  $\mathbf{y}$  obtained by this relation:

$$\mathbf{y} = \mathbf{A} \mathbf{x} + \mathbf{n} , \quad (2.8)$$

with  $\mathbf{n}$  normally distributed having zero mean and covariance matrix  $\mathbf{R}$ , the conditional probability density function of  $\mathbf{y}$  given  $\mathbf{x}$  is:

$$f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^{N/2} |\mathbf{R}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{A} \mathbf{x})^H \mathbf{R}^{-1} (\mathbf{y} - \mathbf{A} \mathbf{x}) \right\} \quad (2.9)$$



where  $|\mathbf{R}|$  denotes the determinant of  $\mathbf{R}$ . The corresponding log-likelihood function is:

$$\begin{aligned} L(\mathbf{x}, \mathbf{y}) &= \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \\ &= -\frac{N}{2} \ln 2\pi - \frac{1}{2} \ln |\mathbf{R}| - \frac{1}{2} (\mathbf{y} - \mathbf{A}\mathbf{x})^H \mathbf{R}^{-1} (\mathbf{y} - \mathbf{A}\mathbf{x}) . \end{aligned}$$

The maximum likelihood estimator is obtained by differentiating the log-likelihood function with respect to  $\mathbf{x}$  and setting it to 0, yielding:

$$\hat{\mathbf{x}}_{ML} = (\mathbf{A}^H \mathbf{R}^{-1} \mathbf{A})^{-1} \mathbf{A}^H \mathbf{R}^{-1} \mathbf{y} \quad (2.10)$$

As pointed earlier, the maximum likelihood estimator coincides with the least squares estimator when  $\mathbf{R} = \sigma_n^2 \mathbf{I}$ .

In the remaining of this subsection we will investigate the mean and covariance of the least squares and maximum likelihood estimators. Let  $\mathcal{R}(\mathbf{A})$  be the subspace spanned by the columns of  $\mathbf{A}$ . If we call  $\mathbf{y}_1$  the projection of  $\mathbf{y}$  onto  $\mathcal{R}(\mathbf{A})$  and  $\mathbf{y}_2$  the projection of  $\mathbf{y}$  onto the orthogonal complement of  $\mathcal{R}(\mathbf{A})$ , then  $\mathbf{y}_1 - \mathbf{A}\mathbf{x}$  belongs to  $\mathcal{R}(\mathbf{A})$  and is orthogonal to  $\mathbf{y}_2$ . Hence we can write :

$$C = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 = \|\mathbf{y}_1 - \mathbf{A}\mathbf{x}\|^2 + \|\mathbf{y}_2\|^2 , \quad (2.11)$$

which attains its minimum when  $\|\mathbf{y}_1 - \mathbf{A}\mathbf{x}\|^2$  is minimized with respect to  $\mathbf{x}$ . Since we can always find an  $\mathbf{x}$  satisfying  $\mathbf{A}\mathbf{x} = \mathbf{y}_1$  and it is unique if and only if

$$\text{null}(\mathbf{A}) = \{\mathbf{x}; \mathbf{A}\mathbf{x} = \mathbf{0}\} = \mathbf{0} , \quad (2.12)$$

then, the least squares estimate always exists, and it is unique if and only if  $\mathbf{A}$  is full column rank. The statistical behavior of an estimator can be investigated by finding its mean and covariance. Assuming that the additive measurement

noise vector is zero mean, the expected value of the least squares estimator can be found as:

$$\begin{aligned}
 E\{\hat{\mathbf{x}}_{LS}\} &= E\{(\mathbf{A}^H \mathbf{A})^{-1} \mathbf{A}^H \mathbf{y}\} \\
 &= E\{(\mathbf{A}^H \mathbf{A})^{-1} \mathbf{A}^H (\mathbf{A} \mathbf{x} + \mathbf{n})\} \\
 &= (\mathbf{A}^H \mathbf{A})^{-1} (\mathbf{A}^H \mathbf{A}) \mathbf{x} + (\mathbf{A}^H \mathbf{A})^{-1} \mathbf{A}^H E\{\mathbf{n}\} \\
 &= \mathbf{x} \ ,
 \end{aligned}$$

which implies that the least squares estimator is unbiased. The covariance of the least squares estimator is given by:

$$Cov\{\hat{\mathbf{x}}_{LS}\} = E\{(\hat{\mathbf{x}}_{LS} - E\{\hat{\mathbf{x}}_{LS}\})(\hat{\mathbf{x}}_{LS} - E\{\hat{\mathbf{x}}_{LS}\})^H\} \ . \quad (2.13)$$

With the assumption that the noise vector  $\mathbf{n}$  is normally distributed having zero mean and covariance matrix  $\mathbf{R}_{nn} = \sigma_n^2 \mathbf{I}$ , the required computation can be performed easily, yielding:

$$Cov\{\hat{\mathbf{x}}_{LS}\} = \sigma_n^2 (\mathbf{A}^H \mathbf{A})^{-1} \quad (2.14)$$

How much an estimator could deviate on the average from the actual parameters is given by the Mean Square Error (MSE) criterion. This is obtained by:

$$\begin{aligned}
 MSE(\mathbf{x}_{LS}) &= \text{trace}(Cov\{\hat{\mathbf{x}}_{LS}\}) \\
 &= \sigma_n^2 \sum_{i=1}^M \frac{1}{\lambda_i} \ ,
 \end{aligned} \quad (2.15)$$

where  $\sqrt{\lambda_i}$  is the  $i^{th}$  singular value of  $\mathbf{A}$ . Hence, if the matrix kernel has a high condition number then the MSE will be large.

Another criterion to quantify statistical performance of estimators, is comparing their error covariance matrix with the Cramer-Rao lower bound, which

establishes a lower bound on the covariance matrix for any estimator of a parameter. The Cramer-Rao theorem states that if  $\mathbf{y}$  is an  $N$ -dimensional vector with probability density function  $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$  and the estimator  $\hat{\mathbf{x}}$  is an unbiased estimator of  $\mathbf{x}$ , then the error covariance matrix of  $\hat{\mathbf{x}}$  is bounded as [1].

$$\mathbf{C} = E\{(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^H\} \geq \mathbf{J}^{-1}, \quad (2.16)$$

where  $\mathbf{J}$  is called the Fisher information matrix and it is given by:

$$\mathbf{J}(\mathbf{x}) = E\left\{\left[\frac{\partial}{\partial \mathbf{x}} \ln f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\right]\left[\frac{\partial}{\partial \mathbf{x}} \ln f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\right]^H\right\}. \quad (2.17)$$

For the Least Squares estimator, the Fisher information matrix becomes:

$$\mathbf{J} = \frac{\mathbf{A}^H \mathbf{A}}{\sigma_n^2}. \quad (2.18)$$

Thus, the covariance matrix of the least squares estimator given in equation 2.14 meets the Cramer-Rao lower bound. Hence, when the measurement noise is identically independently distributed normal, the least squares estimator is the best linear unbiased efficient (BLUE) estimator.

### 2.2.2 Ridge Regression

In 1970, Hoerl and Kennard showed that based on the Mean Square Error criterion, a biased estimation procedure could yield better parameter estimates of a linear model than the analogous estimates obtained via classical least squares [2]. This procedure is introduced initially to avoid the ill effects of quasi-collinearity in ordinary least squares estimators. In order to avoid widely oscillating estimates of least squares, obtained in the case of measurement kernels with a large

condition number, a penalty term on the weighted magnitude of the estimated variables is incorporated to the ridge regression cost function:

$$C = \mathbf{e}^H \mathbf{e} + \mathbf{x}^H \mathbf{D} \mathbf{x} \quad , \quad (2.19)$$

with  $\mathbf{e} = \mathbf{y} - \mathbf{A} \mathbf{x}$  and  $\mathbf{D} = \text{diag}(k_i), i = 1, \dots, N$ , where the weights  $k_i > 0$  are known as the ridge regression constants. The Ridge Regression estimator,  $\hat{\mathbf{x}}_{RR}$ , can be found as the unique minimizer of the above cost function resulting in:

$$\hat{\mathbf{x}}_{RR} = (\mathbf{A}^H \mathbf{A} + \mathbf{D})^{-1} \mathbf{A}^H \mathbf{y} \quad (2.20)$$

Similar estimator was obtained by Levenberg(1944) and Marquardt(1963) in developing an algorithm for nonlinear least squares minimizations [3]. In the presence of little or no prior information, the choice of the ridge regression constants becomes a difficult task. Therefore in many applications the weights are all chosen to be the same, reducing the search space for the right set of parameters to one. This case of uniform weighting is known as ordinary ridge regression and its corresponding estimator is:

$$\hat{\mathbf{x}}_{ORR} = (\mathbf{A}^H \mathbf{A} + k \mathbf{I})^{-1} \mathbf{A}^H \mathbf{y} \quad (2.21)$$

The expected value of the ridge regression estimator can be found as:

$$\begin{aligned} E\{\hat{\mathbf{x}}_{RR}\} &= E\{(\mathbf{A}^H \mathbf{A} + \mathbf{D})^{-1} \mathbf{A}^H \mathbf{y}\} \\ &= (\mathbf{A}^H \mathbf{A} + \mathbf{D})^{-1} \mathbf{A}^H \mathbf{A} \mathbf{x} \quad , \end{aligned}$$

which has a bias of:

$$E\{\mathbf{x} - \hat{\mathbf{x}}_{RR}\} = \mathbf{V} \text{diag}\left(\frac{k_i}{\lambda_i + k_i}\right) \mathbf{V}^H \mathbf{x} \quad , \quad (2.22)$$

where  $\mathbf{V}$  is the right singular matrix and  $\sqrt{\lambda_i}$ 's are the singular values of the  $\mathbf{A}$  matrix. Likewise, the covariance of the Ridge Regression estimator can be found as:

$$Cov\{\hat{\mathbf{x}}_{RR}\} = \sigma_n^2 \mathbf{V} \text{diag}\left(\frac{\lambda_i^2}{(\lambda_i + k_i)^2}\right) \mathbf{V}^H, \quad (2.23)$$

where  $\sigma_n^2$  is the noise variance.

Since the Generalized Ridge Regression is a biased estimator, we use the Cramer Rao lower bound for biased estimators :

$$Cov\{\hat{\mathbf{x}}_{RR}\} \geq \left[\frac{\partial}{\partial \mathbf{x}} E\{\hat{\mathbf{x}}_{RR}\}\right]^H \mathbf{J}^{-1} \left[\frac{\partial}{\partial \mathbf{x}} E\{\hat{\mathbf{x}}_{RR}\}\right], \quad (2.24)$$

where  $\mathbf{J}$  is the Fisher information matrix for  $\mathbf{x}$ . since

$$Cov\{\hat{\mathbf{x}}_{RR}\} = \sigma_n^2 \mathbf{V} \text{diag}\left(\frac{\lambda_i^2}{(\lambda_i + k_i)^2}\right) \mathbf{V}^H, \quad (2.25)$$

$$\frac{\partial}{\partial \mathbf{x}} E\{\hat{\mathbf{x}}_{RR}\} = \mathbf{V} \text{diag}\left(\frac{\lambda_i}{\lambda_i + k_i}\right) \mathbf{V}^H, \quad (2.26)$$

$$\mathbf{J} = \frac{\mathbf{A}^H \mathbf{A}}{\sigma_n^2}, \quad (2.27)$$

it can be shown that:

$$Cov\{\hat{\mathbf{x}}_{RR}\} = \left[\frac{\partial}{\partial \mathbf{x}} E\{\hat{\mathbf{x}}_{RR}\}\right]^H \mathbf{J}^{-1} \left[\frac{\partial}{\partial \mathbf{x}} E\{\hat{\mathbf{x}}_{RR}\}\right]. \quad (2.28)$$

Hence, the Ridge Regression estimator meets the Cramer-Rao bound for the biased case.

The main task in Ridge Regression estimators is how to choose the Ridge Regression constants. The criterion that we are using to judge estimators is the Mean Square Error criterion, so the Ridge Regression estimator would outperform the Least Squares estimator if :

$$MSE(\hat{\mathbf{x}}_{RR}) \leq MSE(\hat{\mathbf{x}}_{LS}). \quad (2.29)$$

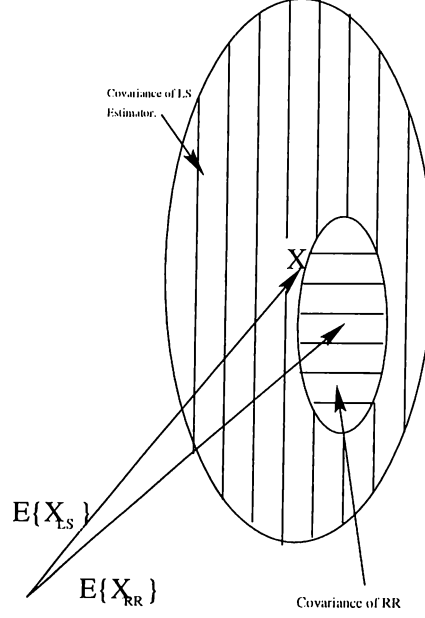


Figure 2.2: Least Squares and Ridge Regression Estimators: Bias and Covariance.

In other words we would like to get the situation illustrated by figure 2.2. The corresponding Mean Square Error of the Ridge Regression estimator is:

$$\begin{aligned} MSE(\hat{\mathbf{x}}_{RR}) &= E\{(\hat{\mathbf{x}}_{RR} - \mathbf{x})^H (\hat{\mathbf{x}}_{RR} - \mathbf{x})\} \\ &= \sum_{i=1}^N \frac{k_i^2 \|\mathbf{V}_i^H \mathbf{x}\|^2}{(\lambda_i + k_i)^2} + \frac{\lambda_i^2 \sigma_n^2}{(\lambda_i + k_i)^2} . \end{aligned} \quad (2.30)$$

Theobald proved that to provide the condition in equation 2.29 we should have :

$$\frac{\mathbf{V}^H \mathbf{x}^H \mathbf{x} \mathbf{V}}{\sigma_n^2} \ll \text{diag}\left(\frac{2}{k} + \frac{1}{\lambda}\right) , \quad (2.31)$$

which is satisfied for  $k_i > 0$  or  $k_i < -2\lambda_i$  for  $i = 1, \dots, M$ . When the  $k_i$ 's are fixed, the domain of parameters where the generalized ridge regression estimator is better than the least squares one are given by:

$$\mathbf{x}^H \mathbf{V} \text{diag}\left(\frac{k\lambda}{2\lambda + k}\right) \mathbf{V}^H \mathbf{x} < \sigma_n^2 , \quad (2.32)$$

which is an ellipsoid. Several suggestions were proposed for the choice of the ridge regression constant: Goldstein and Smith (1974) proposed to take  $k_i =$

$2\sigma_n^2(\gamma_i^2 - \sigma_n^2\lambda_i^{-1})^{-1}$  where  $\boldsymbol{\gamma} = \mathbf{V}^H \mathbf{x}$  [4]. Schmidt (1976) suggested that  $k$  could be taken as  $k = \frac{\sigma_n^2}{\max(\gamma_i)}$ . Swamy, Mehta and Rappoport (1978) showed that if a priori information about the norm of the parameter vector  $\mathbf{x}$  is provided then we can get better estimates. For instance, if we suppose that  $\mathbf{x}$  lies in a hyper-space of radius  $r$ , that is

$$\mathbf{x}^H \mathbf{x} \leq r^2 < \infty, \quad (2.33)$$

the value of  $\mathbf{x}$  that minimizes  $\frac{\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2}{\sigma_n^2}$  subject to equation 2.33 is:

$$\hat{\mathbf{x}}_{SMR}(k) = (\mathbf{A}^H \mathbf{A} + \sigma_n^2 k \mathbf{I})^{-1} \mathbf{A}^H \mathbf{y}, \quad (2.34)$$

where :

$$k = \frac{c\lambda_{max}}{\mathbf{y}^H \mathbf{Q} \mathbf{y}}, \quad (2.35)$$

and

$$\mathbf{Q} = \lambda_{max} \mathbf{A} (\mathbf{A}^H \mathbf{A})^{-2} \mathbf{A}^H + (N - M)^{-1} \mathbf{A} (\mathbf{A}^H \mathbf{A})^{-1} \mathbf{A}^H \quad (2.36)$$

with  $c$  a positive constant and  $\sigma_n^2$  the noise variance.

### 2.2.3 Simulation Results

In the simulations we generate randomly a matrix  $\mathbf{A}$ , a vector  $\mathbf{x}$  and a Gaussian random vector  $\mathbf{n}$  then we find the observation vector  $\mathbf{y}$  by

$$\mathbf{y} = \mathbf{A} \mathbf{x} + \mathbf{n} \quad (2.37)$$

Then, based on  $\mathbf{A}$  and  $\mathbf{y}$  we apply the algorithms described in this chapter to find an estimate for  $\mathbf{x}$ . All through the simulations we will give the estimation error values for an estimate  $\hat{\mathbf{x}}$  of  $\mathbf{x}$  by error percentage:  $\%error = \frac{100\|\hat{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|}$ .

In figures 2.3- 2.7, the estimated and actual  $\mathbf{x}$  obtained using the method of Least Squares and also methods proposed by Theobald's, Schmidt's, Swamy-Mehta & Rappoport and Goldstein are shown. In this simulation the kernel matrix has a condition number  $\kappa \leq 10$  and the signal to noise ratio,  $\text{SNR} = -20 \log(\frac{\sigma_x}{\sigma_n})$ , is 80dB. As expected for such a case the least squares estimator is performing well, the estimate is very close to the true unknown variables. The estimates obtained via the proposed ridge regression procedures provide very close results to the theoretical values. In such cases, one would prefer to use the Least Squares estimator since it does not need any prior knowledge on the noise or data statistics, and the inversion of the system matrix  $\mathbf{A}\mathbf{A}^H$  can be performed without any trouble.

However, as shown in figures 2.8- 2.12. when the signal to noise ratio decreases below 40dB a Ridge Regression estimators provide far more accurate results. This is because of the fact that the least square estimator is more sensitive to the measurement noise. The least squares estimate is more noisy along the right singular vectors corresponding to the smaller singular values. Since, typically smaller singular values are associated with oscillatory singular vectors, the estimates obtained at low SNR have widely oscillatory behavior as shown in figure 2.8.

In order to obtain statistically more significant comparison results, we repeated the above comparisons for various realizations of  $\mathbf{y}$  at different SNR values, and plotted the average errors in the obtained estimates in figure 2.13. At each SNR value 25 different realizations have been used. As it can be seen, performance of the least squares estimator degrades badly at low SNR values



compared with the results obtained by the ridge regression family of estimators.

In order to test the performance of these estimators in the case of measurement kernels with high condition number, we compared the performances of the estimators of various condition numbers. In figure 2.14, for each estimator, we plotted the average error norm as a function of the kernel condition number. As seen from this figure, the performance of the least squares estimator degrades drastically as the condition number gets large.

The superiority of the ridge regression estimators over least squares is due to the utilization of available prior information. The methods presented by Theobald, Schmidt, Swamy Mehta and Rappoport, and Goldstein outperform least squares when the noise standard deviation and the magnitude of the unknown vector are available. Unless a priori knowledge about the signal and the noise statistics are provided, the performance of the suggested ridge regression estimators deteriorates. From the performance of the Ridge Regression estimators we can also conclude that Swamy, Mehta and Rappoport's give better results than the other estimators.

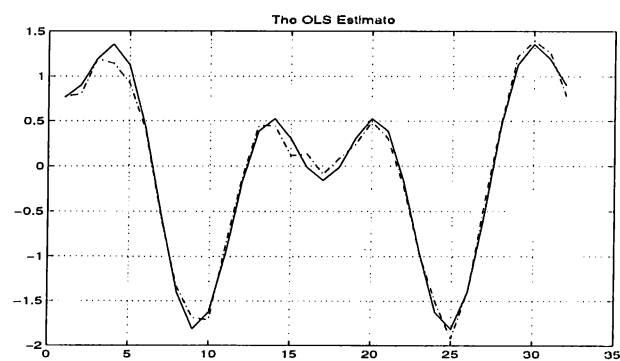


Figure 2.3: Least Squares Estimator, % error= 4.44.

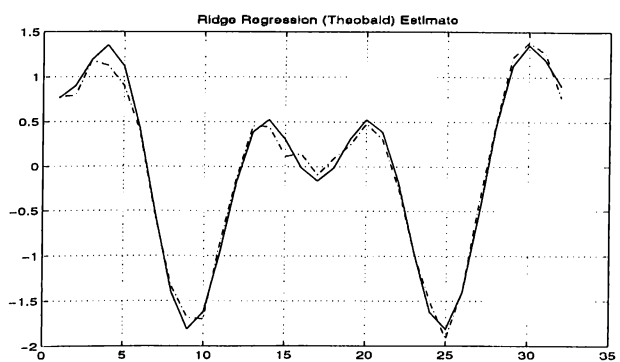


Figure 2.4: Theobald Estimator, % error= 4.44.

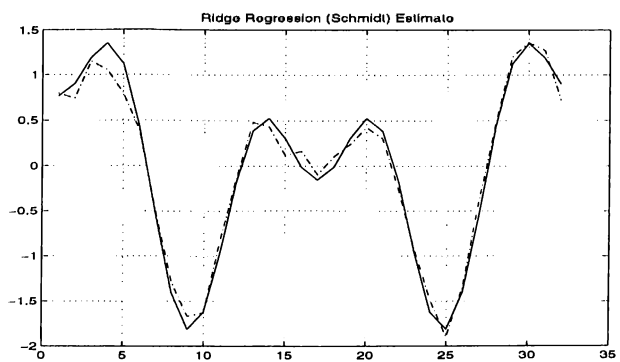


Figure 2.5: Schmidt Estimator, % error= 4.44.

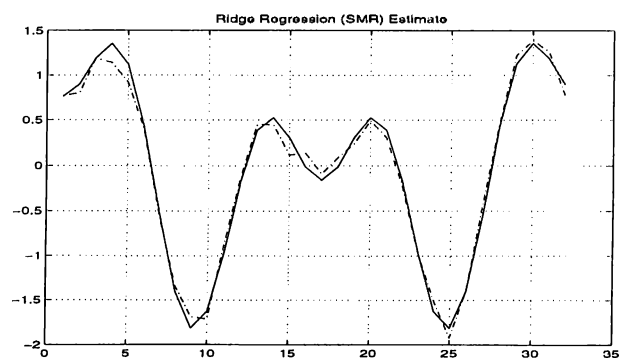


Figure 2.6: Swamy Mehta and Rappoport Estimator, % error= 4.45.

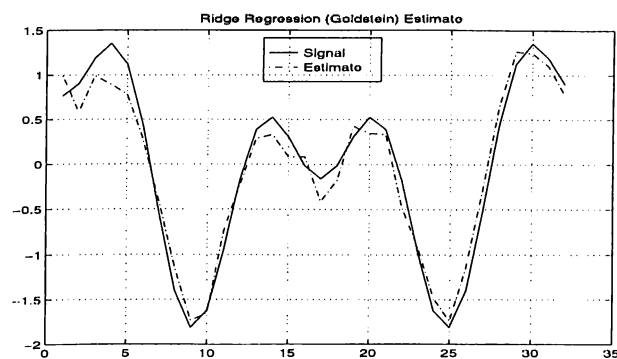


Figure 2.7: Goldstein Estimator, % error= 11.46.

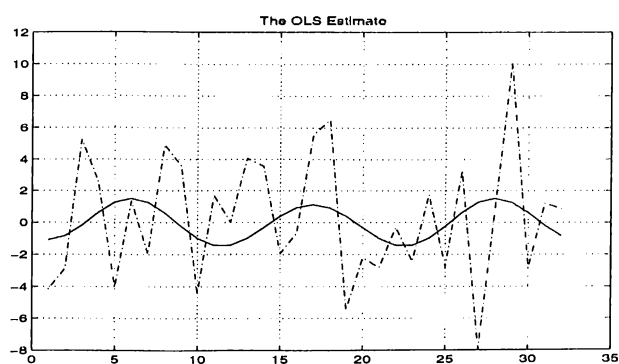


Figure 2.8: Least Squares Estimator, % error= 113.

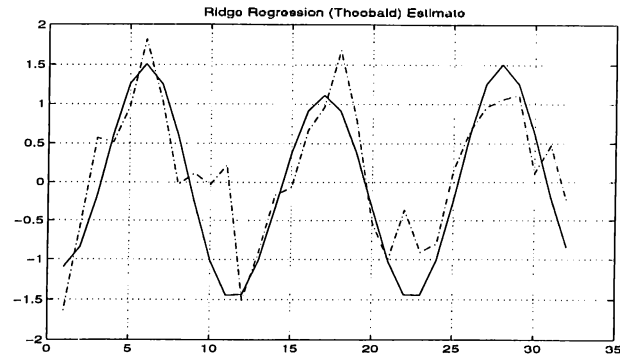


Figure 2.9: Theobald Estimator, % error= 19.3.

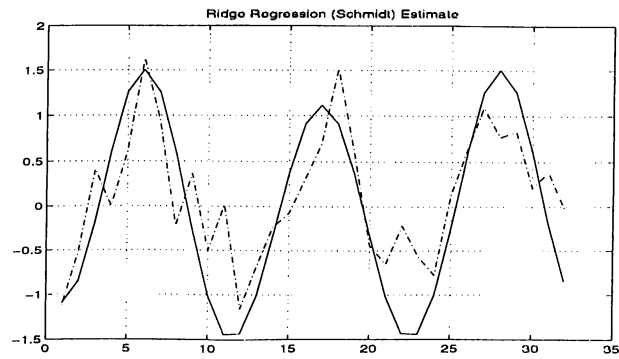


Figure 2.10: Schmidt Estimator, % error= 23.7.

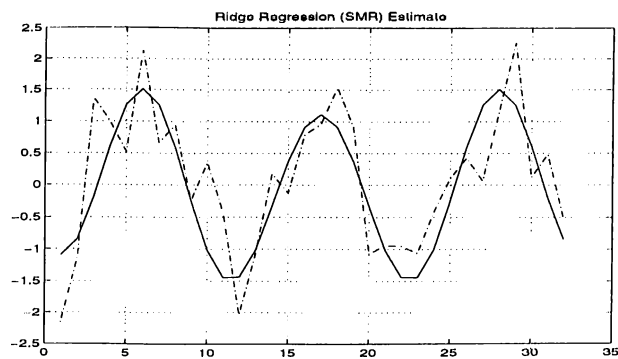


Figure 2.11: Swamy Mehta and Rappoport Estimator, % error= 16.2.

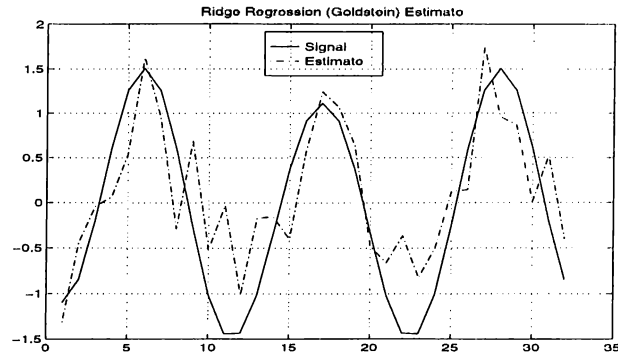


Figure 2.12: Goldstein Estimator, % error= 26.8.

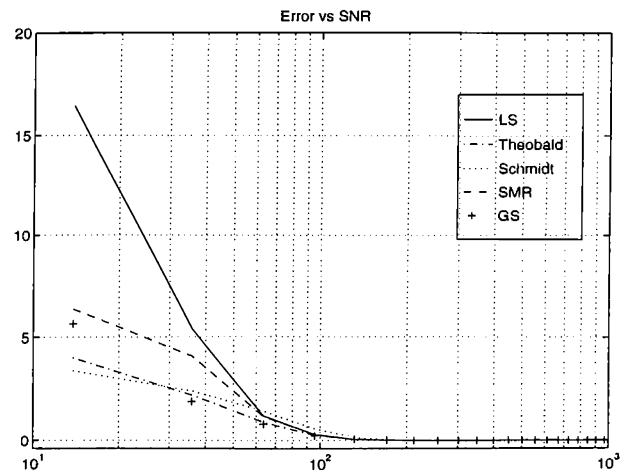


Figure 2.13: Estimation error versus SNR for Least Squares(LS), Theobald, Schmidt, Swamy-Mehta-Rappoport(SMR) and Goldstein-Smith(GS).

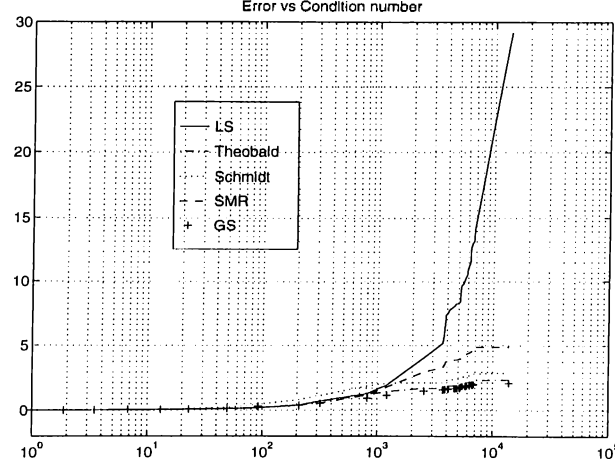


Figure 2.14: Estimation error versus kernel matrix condition number for Least Squares(LS), Theobald, Schmidt, Swamy-Mehta-Rappoport(SMR) and Goldstein-Smith(GS).

## 2.3 Uncertain Model

In the previous section, the measurement matrix entries are assumed to be known exactly, hence, the only source of uncertainty in the observation vector  $\mathbf{y}$  is the additive noise vector  $\mathbf{n}$ . However this assumption is often unrealistic. In practice, we seldom face an exactly known measurement kernel. Errors that do take place during modeling and sampling may imply inaccuracies on the measurement matrix  $\mathbf{A}$  as well. The inaccuracies in  $\mathbf{A}$  can be due to uncertainties in a few parameters which define  $\mathbf{A}$ , or to the uncertainties in each individual entry in  $\mathbf{A}$  which do not fit to a low order parametric description. In the latter case the Total Least Squares (TLS) is one of the commonly used methods of obtaining estimates when there are errors in both the observation vector  $\mathbf{y}$  and the data matrix  $\mathbf{A}$ . Although computationally more intensive and limited in terms of its application areas, nonlinear least squares modeling is the method of choice if there is a parametric description of the measurement matrix.

### 2.3.1 Total Least Squares

The Total Least Squares approach has been introduced in recent years in the numerical analysis literature as an alternative for the least squares in the case that both  $\mathbf{A}$  and  $\mathbf{y}$  are affected by errors. A good way to introduce the Total Least Squares method is to recast the Ordinary Least Squares problem.

In the Least Squares estimation, the unknown  $\mathbf{x}$  is obtained as the minimizer of the following optimization problem:

$$\begin{aligned} \min_{\mathbf{y}' \in \mathbb{R}^M} \quad & \|\mathbf{y} - \mathbf{y}'\|^2 \\ \text{Subject to} \quad & \mathbf{y}' \in \mathcal{R}(\mathbf{A}) . \end{aligned}$$

Once  $\mathbf{y}'$  is found, the minimum norm  $\mathbf{x}$  satisfying  $\mathbf{A}\mathbf{x} = \mathbf{y}'$  is called the Least Squares solution. The underlying assumption here is that errors only occur in the vector  $\mathbf{y}$  and that the matrix  $\mathbf{A}$  is exactly known, which is often far from reality. The least squares estimator is obtained by solving the smallest perturbation on the measurements so that the perturbed measurement will lie in the range space of  $\mathbf{A}$ . When there are errors in both  $\mathbf{A}$  and  $\mathbf{y}$ , the same idea of perturbation can be applied to both  $\mathbf{A}$  and  $\mathbf{y}$  such that the perturbed measurements will lie in the range space of the perturbed  $\mathbf{A}$  matrix. Again we want to find the minimal perturbation on both  $\mathbf{A}$  and  $\mathbf{y}$ . In the TLS, this is achieved by finding the solution to the following optimization problem:

$$\begin{aligned} \min_{[\hat{\mathbf{A}}, \hat{\mathbf{y}}] \in \mathbb{R}^{N \times (M+1)}} \quad & \|[\mathbf{A}, \mathbf{y}] - [\hat{\mathbf{A}}, \hat{\mathbf{y}}]\|_F \\ \text{Subject to} \quad & \hat{\mathbf{y}} \in \mathcal{R}(\hat{\mathbf{A}}) , \end{aligned}$$

where  $\|\cdot\|_F$  denotes the frobenius norm. Once a minimizing  $[\hat{\mathbf{A}}, \hat{\mathbf{y}}]$  is found,  $\mathbf{x}$  satisfying  $\hat{\mathbf{A}}\mathbf{x} = \hat{\mathbf{y}}$  is called the Total Least Squares solution.

To solve this problem, we bring  $\mathbf{A} \mathbf{x} \simeq \mathbf{y}$  into the following form

$$[\mathbf{A}, \mathbf{y}] \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} \simeq \mathbf{0} . \quad (2.38)$$

Let  $[\mathbf{A}, \mathbf{y}] = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^H$  be the singular value decomposition of  $[\mathbf{A}, \mathbf{y}]$ , with

$$\begin{aligned} \mathbf{U} &= [\mathbf{u}_1, \dots, \mathbf{u}_{M+1}] , \\ \mathbf{\Sigma} &= \text{diag}(\sigma_1, \dots, \sigma_{M+1}) , \\ \mathbf{V} &= [\mathbf{v}_1, \dots, \mathbf{v}_{M+1}] . \end{aligned}$$

If  $\sigma_{M+1} \neq 0$  then  $[\mathbf{A}, \mathbf{y}]$  is of rank  $M + 1$  and the subspace  $\mathcal{S}$  generated by the rows of  $[\mathbf{A}, \mathbf{y}]$  coincides with  $R^{M+1}$  and there is no nonzero vector in the orthogonal complement of  $\mathcal{S}$ , hence equation 2.38 is incompatible. To obtain a solution the rank of  $[\mathbf{A}, \mathbf{y}]$  must be reduced to  $M$ .

Eckart-Young-Mirsky matrix approximation theorem says: Let the singular value decomposition of  $\mathbf{C} \in R^{N \times M}$  be given by:  $\mathbf{C} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T$  with  $r = \text{rank}(\mathbf{C})$ . If  $k < r$  and  $\mathbf{C}_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T$  then

$$\min_{\text{rank}(\mathbf{D})=k} \|\mathbf{C} - \mathbf{D}\|^2 = \|\mathbf{C} - \mathbf{C}_k\|^2 = \sigma_{k+1}^2 \quad (2.39)$$

and

$$\min_{\text{rank}(\mathbf{D})=k} \|\mathbf{C} - \mathbf{D}\|_F = \|\mathbf{C} - \mathbf{C}_k\|_F = \sqrt{\sum_{i=k+1}^p \sigma_i^2} \quad (2.40)$$

with  $p = \min(M, N)$ . Using this theorem, the best rank  $M$  Total Least Squares approximation  $[\hat{\mathbf{A}}, \hat{\mathbf{y}}]$  of  $[\mathbf{A}, \mathbf{y}]$  which minimizes the deviation in variance is given by:

$$[\hat{\mathbf{A}}, \hat{\mathbf{y}}] = \mathbf{U} \hat{\mathbf{\Sigma}} \mathbf{V}^H , \quad (2.41)$$



where  $\hat{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_M, 0)$ . It is clear that the approximate set

$$[\hat{\mathbf{A}}, \hat{\mathbf{y}}] \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} \simeq \mathbf{0} \quad (2.42)$$

is compatible and its solution is given by the vector  $\mathbf{v}_{M+1}$  the last column of  $\mathbf{V}$ .

Thus the total least squares solution is :

$$\hat{\mathbf{x}}_{TLS} = \frac{-1}{\mathbf{V}_{M+1,M+1}} [\mathbf{V}_{1,M+1}, \dots, \mathbf{V}_{M,M+1}]^T \quad (2.43)$$

$$= (\mathbf{A}^H \mathbf{A} - \sigma_{M+1}^2 \mathbf{I})^{-1} \mathbf{A}^H \mathbf{y} \quad (2.44)$$

exists and is unique solution to

$$\hat{\mathbf{A}} \mathbf{x} = \hat{\mathbf{y}} \quad (2.45)$$

Whenever  $\mathbf{V}_{M+1,M+1} \neq 0$ , the Total Least Squares solution is solvable and is therefore called generic. Problems may occur if  $\sigma_p > \sigma_{p+1} = \dots = \sigma_{M+1}$  for  $p \leq M$  and if all  $\mathbf{V}_{M+1,i} = 0$  for  $i = p+1, \dots, M+1$  these problems are called non generic.

For the generic case when  $\sigma_p > \sigma_{p+1} = \dots = \sigma_{M+1}$  for  $p \leq M$ , if not all  $\mathbf{V}_{M+1,i} = 0$  for  $i = p+1, \dots, M+1$  then the minimum norm Total Least Squares solution is given by:

$$\hat{\mathbf{x}}_{TLS} = \frac{-1}{\sum_{i=p+1}^{M+1} \mathbf{V}_{M+1,i}^2} \sum_{i=p+1}^{M+1} \mathbf{V}_{M+1,i} [\mathbf{V}_{1,i}, \dots, \mathbf{V}_{M,i}]^T \quad (2.46)$$

For the non generic case when  $\mathbf{V}_{M+1,j} = 0$  for  $j = p+1, \dots, M+1$ . If  $\sigma_{p-1} > \sigma_p$  and  $\mathbf{V}_{M+1,p} \neq 0$  then the Total Least Squares Solution is given by:

$$\hat{\mathbf{x}}_{TLS} = \frac{-1}{\mathbf{V}_{M+1,p}} [\mathbf{V}_{1,p}, \dots, \mathbf{V}_{M,p}]^T \quad (2.47)$$

### 2.3.2 Simulation Results

To test the performance of the Total Least Squares Estimator, we generated a matrix  $\mathbf{A}_n$  of independent identically distributed random variables, with zero mean. This matrix is added to the kernel matrix  $\mathbf{A}$ , then we generate the data vector  $\mathbf{y}$  in the same way we did for testing Ridge Regression and Least Squares estimators.

Figure 2.15 shows the Total Least Squares (TLS) and the Ordinary Least Squares (OLS) estimates when applied to a case where the Total Least Squares solution is generic. The Total Least Squares outperforms the Least Squares for several reasons. Ordinary Least Squares takes into account only errors in the observed data  $\mathbf{y}$ . However, Total Least Squares considers that both the data vector  $\mathbf{y}$  and the kernel matrix  $\mathbf{A}$  are erroneous, and it searches for the smallest perturbation on both  $\mathbf{A}$  and  $\mathbf{y}$  to reach a compatible set of equations.

The main problem in using the TLS approach is how to determine the rank of the augmented matrix  $[\mathbf{A}, \mathbf{y}]$  and how to choose  $p$  for which  $\sigma_p \neq 0$ . The performance of the TLS estimator deteriorates drastically when the rank is chosen inaccurately.

Despite this drawback, the Total Least Squares estimate remains the only way to solve the problem of linear parameter estimation under model uncertainties that are treated as independently distributed random variables.

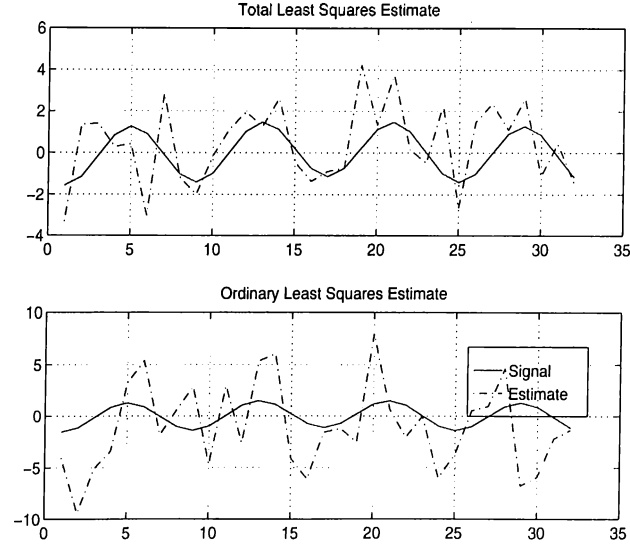


Figure 2.15: TLS and OLS  $\%e_{TLS} = 31.2$ ,  $\%e_{OLS} = 47.6$  .

### 2.3.3 Nonlinear Least Squares Modeling

In many applications of interest the phenomenon under investigation can be represented by a system of linear equations in which the elements of the system matrix are known functions of a set of parameters. For instance, in array signal processing the parameters correspond to direction of arrivals of the received signals, or in inverse problems, the parameters correspond to the measurement device geometry. For these cases measurements relation is modeled as:

$$\mathbf{y} = \mathbf{A}(\boldsymbol{\theta})\mathbf{x} + \mathbf{n} \quad , \quad (2.48)$$

where  $\boldsymbol{\theta} \in \mathcal{R}^P$  is a vector containing  $P$  parameters characterizing the uncertainty in the model. To solve this problem the Nonlinear Least Squares Modeling technique has been applied [5]. In this approach, which was presented by Cadzow, a selection of the parameter vector  $\boldsymbol{\theta}$  and the unobserved vector  $\mathbf{x}$  are tried to be found so that  $\mathbf{A}(\boldsymbol{\theta})\mathbf{x}$  best approximates  $\mathbf{y}$  in the Euclidean norm sense.

More precisely,  $\boldsymbol{\theta}$  and  $\boldsymbol{x}$  are found by solving the following squared  $l_2$  norm optimization problem:

$$\min_{\boldsymbol{x} \in \mathbb{C}^M} \min_{\boldsymbol{\theta} \in \mathbb{R}^P} \|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{x}\|^2. \quad (2.49)$$

Due to the nonlinear fashion in which  $\boldsymbol{x}$  and  $\boldsymbol{\theta}$  appear, generally there is no closed form expression for the solution to this optimization problem. So it is necessary to apply nonlinear optimization techniques to search for a solution. The search space dimension can be reduced to the dimension of  $\boldsymbol{\theta}$  by observing that the optimal value of  $\boldsymbol{x}$  given  $\boldsymbol{\theta}$  is:

$$\hat{\boldsymbol{x}}_{LS} = \boldsymbol{A}^\dagger(\boldsymbol{\theta})\boldsymbol{y} \quad (2.50)$$

Hence, the overall optimization can be recast in terms of  $\boldsymbol{\theta}$  alone as:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^P} \|\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{\theta})\|\boldsymbol{y}\|^2, \quad (2.51)$$

where  $\boldsymbol{P}(\boldsymbol{\theta}) = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{A}^\dagger(\boldsymbol{\theta})$  is the projection operator onto the range of  $\boldsymbol{A}(\boldsymbol{\theta})$ . Since  $\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{\theta})$  is the projection operator onto the perpendicular space of range  $\boldsymbol{A}(\boldsymbol{\theta})$ , optimal value of  $\boldsymbol{\theta}$  can be found as the maximizer of

$$g(\boldsymbol{\theta}) = \|\boldsymbol{P}(\boldsymbol{\theta})\boldsymbol{y}\|^2. \quad (2.52)$$

Thus we would chose  $\boldsymbol{\theta}$  so that the projection of  $\boldsymbol{y}$  onto the subspace spanned by the columns of  $\boldsymbol{A}(\boldsymbol{\theta})$  has the maximum squared norm.

The maximization of  $g(\boldsymbol{\theta})$  necessitates a nonlinear programming algorithm to approximate an optimal solution. Cadzow suggested the method of descent in which the present estimator  $\boldsymbol{\theta}$  is additively perturbed to  $\boldsymbol{\theta} + \boldsymbol{\delta}$ , where  $\boldsymbol{\delta}$  is referred to as perturbation vector. The basic task becomes to select the perturbation

vector so that the improving condition

$$g(\boldsymbol{\theta} + \boldsymbol{\delta}) > g(\boldsymbol{\theta}) \quad (2.53)$$

is satisfied. For a sufficiently small, in size, perturbation vector, a Taylor series expansion, of the perturbed criterion can be made, in which only the first two terms are retained:

$$\begin{aligned} g(\boldsymbol{\theta} + \boldsymbol{\delta}) &= \|\mathbf{P}(\boldsymbol{\theta} + \boldsymbol{\delta})\mathbf{y}\|^2 \\ &= \left\| \left[ \mathbf{P}(\boldsymbol{\theta}) + \sum_{k=1}^P \frac{\partial \mathbf{P}(\boldsymbol{\theta})}{\partial \theta_k} \delta_k \right] \mathbf{y} \right\|^2 \\ &= \left\| \left[ \mathbf{P}(\boldsymbol{\theta}) \right] \mathbf{y} + \sum_{k=1}^P \left[ \frac{\partial \mathbf{P}(\boldsymbol{\theta})}{\partial \theta_k} \mathbf{y} \right] \delta_k \right\|^2 \\ &= \left\| \left[ \mathbf{P}(\boldsymbol{\theta}) \right] \mathbf{y} + \mathbf{L}(\boldsymbol{\theta}) \boldsymbol{\delta} \right\|^2, \end{aligned} \quad (2.54)$$

where:

$$\mathbf{L}(\boldsymbol{\theta}) = \left[ \frac{\partial \mathbf{P}(\boldsymbol{\theta})}{\partial \theta_1} \mathbf{y} : \frac{\partial \mathbf{P}(\boldsymbol{\theta})}{\partial \theta_2} \mathbf{y} : \dots : \frac{\partial \mathbf{P}(\boldsymbol{\theta})}{\partial \theta_P} \mathbf{y} \right], \quad (2.55)$$

and  $\delta_j$  or  $\theta_j$  are the  $j^{th}$  entry of  $\boldsymbol{\delta}$  or  $\boldsymbol{\theta}$ , respectively. A logical choice of the perturbation vector would be one that maximizes the Euclidean norm criterion given by equation 2.54.

$$\begin{aligned} g(\boldsymbol{\theta} + \boldsymbol{\delta}) &= g(\boldsymbol{\theta}) + \boldsymbol{\delta}^H \mathbf{L}(\boldsymbol{\theta})^H \mathbf{P}(\boldsymbol{\theta}) \mathbf{y} \\ &\quad + \mathbf{y}^H \mathbf{P}(\boldsymbol{\theta}) \mathbf{L}(\boldsymbol{\theta}) \boldsymbol{\delta} + \boldsymbol{\delta}^H \mathbf{L}(\boldsymbol{\theta})^H \mathbf{L}(\boldsymbol{\theta}) \boldsymbol{\delta}. \end{aligned}$$

By setting the gradient of this expression, with respect to  $\boldsymbol{\delta}$ , to the zero vector the optimal selection can be found as :

$$\boldsymbol{\delta}^* = -[\Re\{\mathbf{L}(\boldsymbol{\theta})^H \mathbf{L}(\boldsymbol{\theta})\}]^\dagger \Re\{\mathbf{L}(\boldsymbol{\theta})^H \mathbf{P}(\boldsymbol{\theta}) \mathbf{y}\}. \quad (2.56)$$

To ensure a sufficiently small perturbation, a scaled perturbation vector  $\alpha \boldsymbol{\delta}^*$  is instead used. The nonlinear programming algorithm is given in table 2.1. The stopping condition to be evaluated is the fit error norm.

Step	Description
1	Start by an initial $\boldsymbol{\theta}$
2	Evaluate $\ \mathbf{P}(\boldsymbol{\theta})\mathbf{y}\ ^2$
3	Determine $\mathbf{L}(\boldsymbol{\theta})$
4	Compute the optimum perturbation vector $\boldsymbol{\delta}^*$
5	Evaluate $\ \mathbf{P}(\boldsymbol{\theta} + \alpha\boldsymbol{\delta}^*)\mathbf{y}\ ^2$ for $\alpha = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots$ until improvement.
6	Evaluate stopping conditions, if not satisfied, set $\boldsymbol{\theta} = \boldsymbol{\theta} + \alpha\boldsymbol{\delta}^*$ and go to step 2.

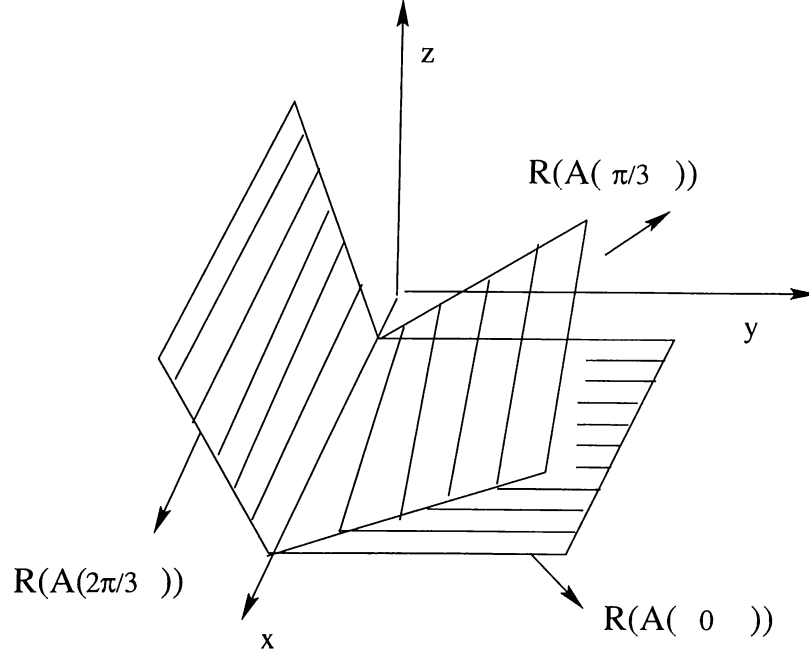
Table 2.1: Nonlinear Programming Algorithm.

### 2.3.4 Simulation Results

Consider the following matrix:

$$\mathbf{A}(\theta) = \begin{bmatrix} 1 & 0 \\ 0 & \cos(\theta) \\ 0 & \sin(\theta) \end{bmatrix}. \quad (2.57)$$

The range of this matrix is a plane in the three dimensional space. The parameter  $\theta$  determines the slant angle of the plane as shown in figure 2.16. Based on the above parametric representation we investigate the the performance of the nonlinear least squares modeling estimator. So, we choose a random  $\boldsymbol{\theta}^0$  then generate  $\mathbf{x}$  and  $\mathbf{n}$  in the previously described fashion, and we obtain  $\mathbf{y}$  by  $\mathbf{A}(\boldsymbol{\theta}^0)\mathbf{x} + \mathbf{n}$ . Then, based on  $\mathbf{y}$  and the known parametric description of  $\mathbf{A}$ , we estimate  $\mathbf{x}$ . Figure 2.17 shows the result obtained in the case when the system is nearly noise free  $\text{SNR} = 80\text{dB}$ . The nonlinear least squares modeling algorithm converges in few steps to the optimal vector of unknowns  $\boldsymbol{\theta}$  and the estimate  $\hat{\mathbf{x}}$  is very close to the true unknown vector.

Figure 2.16: Effect of  $\theta$  on the  $R(A)$ .

However in the case when we increase the signal to noise ratio the quality of the estimates deteriorates and the estimated parameters deviate significantly from the true values. This is because in the search for optimal  $\theta$  that results in the largest projection of  $\mathbf{y}$  onto the subspace spanned by  $\mathbf{A}(\theta)$ , the additive noise vector is also projected. Hence, if the condition number of  $\mathbf{A}(\theta)$  is large, the noise component of the projection may result in significant estimation errors along the singular vectors corresponding to the smaller singular values, as illustrated by figure 2.18, where the SNR = 28dB. This is due to the fact that this algorithm is based on the least square estimator, and whatever the least squares suffer from will be inherited in this procedure.

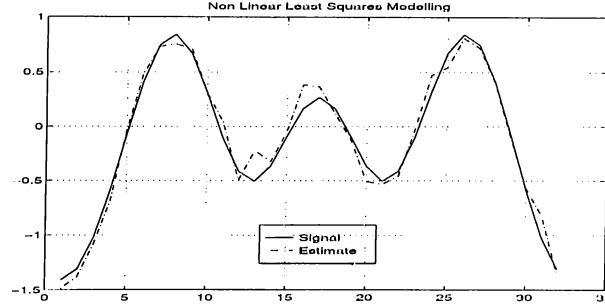


Figure 2.17: Application of Cadzow's algorithm with SNR=80dB and kernel of low condition number, %error= 3.2

A comparison between the performance of Total Least Squares and Nonlinear Least Squares algorithms is held in figure 2.19, both estimates were applied for the case of an SNR 47dB, the results show the superiority of the Nonlinear Least Squares Modeling. Thus modeling the uncertainty of the kernel matrix as a function of nuisance parameters will improve the quality of the estimate.

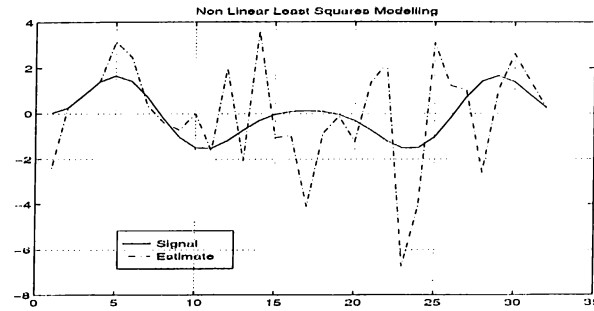


Figure 2.18: Non linear least squares modeling algorithm with SNR=28dB, %error = 64.2 .



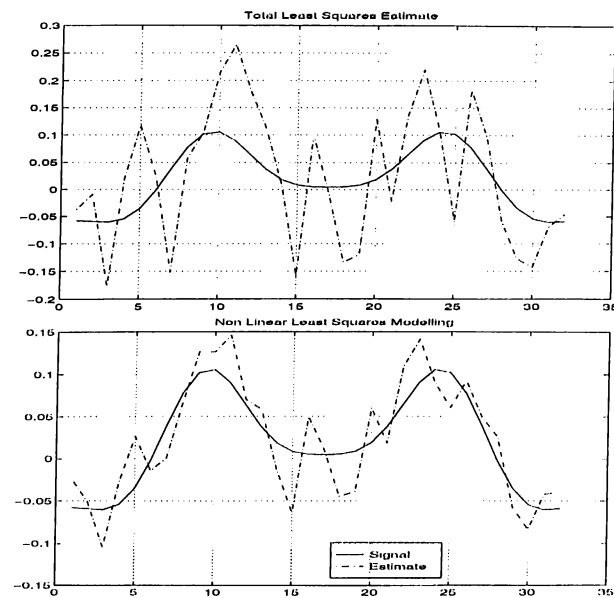


Figure 2.19: Comparing TLS and Nonlinear Least Squares Modeling,  $\%error_{TLS} = 71.5$  and  $\%error_{Cad} = 24.1$

# Chapter 3

## Proposed Estimation Methods

### 3.1 Introduction

In the previous chapter we presented the commonly used estimation approaches in the presence of measurement uncertainty. We investigated the Least Squares and Ridge Regression estimators for the fixed kernel case. The drawbacks of these approaches were observed in the simulations, due to noise standard deviation and the structure of the data matrix for the former; and to the necessity of prior information for the choice of the ridge regression constant for the latter. For the uncertain kernel case we examined the Total Least Squares and the Non-linear Least Squares Modeling approaches. In this chapter, we will present new approaches to provide more reliable estimates of the unknowns.

Following the same plan as the previous chapter, we start by proposing the methods that we can apply when we fix the parameter  $\theta$  and we assume that

it is known. Then we present algorithms when this parameter is unknown and need to be estimated as well.

## 3.2 Known Measurement Kernel

In this section, we will present a way to choose the Ridge Regression constant by constraining the estimate to lead to a fit error having the same statistics as that of the additive noise. Then, we introduce a method that alleviates the necessity of a priori information by iteratively estimating the noise and the signal variance. Then, we propose an algorithm to solve large linear system of equations, the algorithm recursively updates the solution in an increasingly larger dimensional subspace whose basis vectors are a subset of a complete wavelet basis.

### 3.2.1 Error Dependent Ridge Regression Constant

As presented in the previous chapter, Ridge Regression methods provide a family of solutions depending on the regression parameter. Some of the commonly used ways of choosing the ridge regression parameter have been presented in the previous chapter. However no firm recommendation for optimal Ridge Regression parameter seems to emerge. The simplest single parameter family of ridge regression estimates are in the following form:

$$\hat{\mathbf{x}}_{RR}(\mu) = (\mathbf{A}^H \mathbf{A} + \mu \mathbf{I})^{-1} \mathbf{A}^H \mathbf{y} \quad (3.1)$$

where  $\mu$  is the regression parameter.

Different methods of choosing  $\mu$  will lead to different fit error vectors  $\hat{\mathbf{e}}_{RR}(\mu) = \mathbf{y} - \mathbf{A} \hat{\mathbf{x}}_{RR}(\mu)$ . If there is prior information on the statistics of the random noise vector, one can try to choose  $\mu$  such that  $\hat{\mathbf{e}}_{RR}(\mu)$  will look like a realization of the random noise vector. The similarity can be measured based on the deviation of sample moments of  $\hat{\mathbf{e}}_{RR}(\mu)$  from the known moments of the noise vector. In practice, only the first few moments can be used for this purpose. Here we suggest to choose  $\mu$  such that the sample variance of  $\hat{\mathbf{e}}_{RR}(\mu)$  is the same as the noise variance.

### 3.2.2 Simulation Results

Over a synthetically generated example, this practical approach is compared with the Swamy-Mehta and Rappoport approach, and the obtained results are shown in figure 3.2. As seen from this figure, the performance of the proposed approach is better. This is a typical case over moderate sized problems. When the dimension of  $\mathbf{y}$  gets larger, the performance difference gets smaller.

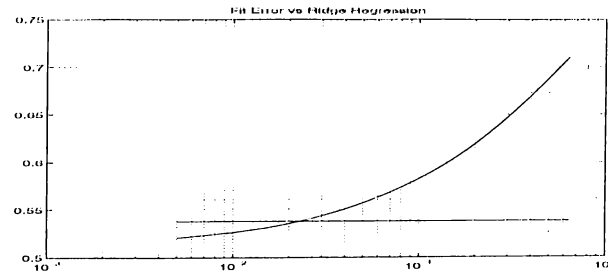


Figure 3.1: Choice of Ridge Regression: increasing curve is the sample variance of the fit error vector as a function of  $\mu$ , horizontal line is  $\sigma^2$ .

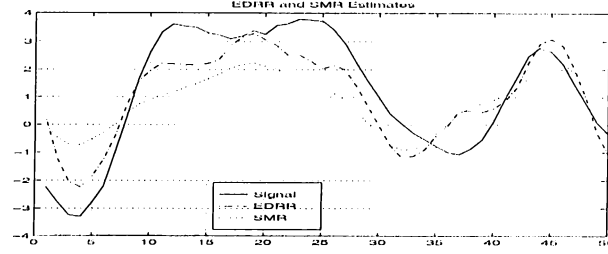


Figure 3.2: Error Dependent Ridge Regression(EDRR) with % error = 12.4 and Swamy-Mehta-Rappoport (SMR) with % error = 21.5 Estimates .

### 3.2.3 Gauss-Markov Estimate with recursive updates

There are two fundamentally different ways of solving statistical problems: The classical and the Bayesian approaches. In the classical approach, a set of data generated in accordance with some unknown probability law will be used without making any assumption about the unknown law. In the Bayesian approach, the use of any reasonable prior knowledge about the unknown is recommended.

In deriving the maximum likelihood estimator we have inferred the value of the unknown parameter  $\mathbf{x}$  by chasing  $\hat{\mathbf{x}}$  to be the parameter that maximizes the likelihood of the observed data  $\mathbf{y}$ , this is a classical view of the problem. In the following we will treat the unknown parameter  $\mathbf{x}$  as a realization of a random experiment from which the unknowns are endowed with prior distribution. This is the Bayesian approach where the information available prior to and carried by the measurements are optimally combined to obtain an estimate for  $\mathbf{x}$ .

If we define  $P(\mathbf{x}|\mathbf{y})$  to be the conditional probability that  $\mathbf{x}$  is true given  $\mathbf{y}$ , then Bayes theorem gives the desired  $P(\mathbf{x}|\mathbf{y})$  from the computable probability  $P(\mathbf{y}|\mathbf{x})$  and from the probabilities  $P(\mathbf{x})$  and  $P(\mathbf{y})$ .  $P(\mathbf{x})$  is called the prior

probability because it is known in advance, somehow, to obtain  $\mathbf{y}$ , and  $P(\mathbf{x}|\mathbf{y})$  is called the a posteriori probability because it is what we aim to obtain after considering the above facts.

$$P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{P(\mathbf{y})} \quad (3.2)$$

Let  $\hat{\mathbf{x}}_B$  be the Bayes estimator. The quality of the estimator  $\hat{\mathbf{x}}_B$  is measured by a real-valued function with some specific properties, known as the loss function, denoted by  $L[\mathbf{x}, \hat{\mathbf{x}}_B]$ . A typical loss function would be the quadratic one:

$$L[\mathbf{x}, \hat{\mathbf{x}}_B] = [\mathbf{x} - \hat{\mathbf{x}}_B]^H [\mathbf{x} - \hat{\mathbf{x}}_B] \quad , \quad (3.3)$$

which assigns a loss equal to the Euclidean distance between the actual value of  $\mathbf{x}$  and the estimated value  $\hat{\mathbf{x}}_B$ . The Bayes estimator under quadratic loss is given by:

$$\hat{\mathbf{x}}_B = \arg \min_{\hat{\mathbf{x}}} \int L[\mathbf{x}, \hat{\mathbf{x}}_B] f(\mathbf{x}|\mathbf{y}) d\mathbf{x} \quad , \quad (3.4)$$

which is the conditional mean of  $\mathbf{x}$  given  $\mathbf{y}$ , that is

$$\hat{\mathbf{x}}_B = E\{\mathbf{x}|\mathbf{y}\} \quad . \quad (3.5)$$

Our problem is to estimate  $\mathbf{x}$  from the overdetermined set of equations:

$$\mathbf{y} = \mathbf{A} \mathbf{x} + \mathbf{n} \quad . \quad (3.6)$$

Assuming that  $\mathbf{x}$  and  $\mathbf{n}$  are independent, zero-mean Gaussian random vectors, with autocorrelation matrices  $\mathbf{R}_{xx}$  and  $\mathbf{R}_{nn}$ , respectively, we get the following jointly Gaussian density for  $\mathbf{x}$  and  $\mathbf{y}$ :

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim N \left[ \mathbf{0}; \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{R}_{xx} \mathbf{A}^H \\ \mathbf{A} \mathbf{R}_{xx} & \mathbf{A} \mathbf{R}_{xx} \mathbf{A}^H + \mathbf{R}_{nn} \end{bmatrix} \right] \quad (3.7)$$

To find the Bayesian estimator  $\hat{\mathbf{x}}_B$  that minimizes the mean squared error, we have to find the expectation of the random variable  $\mathbf{z} = \mathbf{x}|\mathbf{y}$ . Gauss-Markov theorem states that if  $\mathbf{x}$  and  $\mathbf{y}$  are random vectors that are distributed according to the multivariable distribution

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim N \left[ \begin{bmatrix} \mathbf{m}_x \\ \mathbf{m}_y \end{bmatrix}; \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{R}_{xy} \\ \mathbf{R}_{yx} & \mathbf{R}_{yy} \end{bmatrix} \right] \quad (3.8)$$

Then the conditional distribution of  $\mathbf{x}$  given  $\mathbf{y}$  is multivariate normal:

$$P(\mathbf{x}|\mathbf{y}) \sim N(\hat{\mathbf{x}}, \mathbf{P}) \quad , \quad (3.9)$$

where the mean  $\hat{\mathbf{x}}$  and the covariance  $\mathbf{P}$  are given by:

$$\begin{aligned} \hat{\mathbf{x}} &= \mathbf{m}_x + \mathbf{R}_{xy} \mathbf{R}_{yy}^{-1} (\mathbf{y} - \mathbf{m}_y) \\ \mathbf{P} &= \mathbf{R}_{xx} - \mathbf{R}_{xy} \mathbf{R}_{yy}^{-1} \mathbf{R}_{yx} \quad . \end{aligned}$$

Thus the Bayesian estimator for  $\mathbf{x}$  is given by:

$$\hat{\mathbf{x}}_B = (\mathbf{A}^H \mathbf{R}_{nn}^{-1} \mathbf{A} + \mathbf{R}_{xx}^{-1})^{-1} \mathbf{A}^H \mathbf{R}_{nn}^{-1} \mathbf{y} \quad (3.10)$$

In the case of  $\mathbf{x}$  and  $\mathbf{n}$  are composed of independent identically distributed random variables, i.e.,  $\mathbf{R}_{xx} = \sigma_x^2 \mathbf{I}$  and  $\mathbf{R}_{nn} = \sigma_n^2 \mathbf{I}$ , we get

$$\hat{\mathbf{x}}_B = (\mathbf{A}^H \mathbf{A} + \frac{\sigma_n^2}{\sigma_x^2} \mathbf{I})^{-1} \mathbf{A}^H \mathbf{y} \quad , \quad (3.11)$$

$$\mathbf{P} = \sigma_n^2 (\mathbf{A}^H \mathbf{A} + \frac{\sigma_n^2}{\sigma_x^2} \mathbf{I})^{-1} \quad . \quad (3.12)$$

Note that  $\hat{\mathbf{x}}_B$  is a ridge regression estimator with ridge regression constant  $\mu = \frac{\sigma_n^2}{\sigma_x^2}$  and the corresponding mean square error is:

$$\begin{aligned} MSE(\hat{\mathbf{x}}_B) &= \text{trace}(\sigma_n^2 \mathbf{V} \text{diag}(\frac{\sigma_x^2}{\lambda \sigma_x^2 + \sigma_n^2}) \mathbf{V}^H) \\ &= \sum_{i=1}^M \frac{\sigma_n^4}{\lambda_i \sigma_x^2 + \sigma_n^2} \end{aligned} \quad (3.13)$$

which is smaller than the MSE of the maximum likelihood estimator due to the optimal use of the prior information on  $\mathbf{x}$ .

In the case of unknown  $\sigma_x^2$  and  $\sigma_n^2$ , we can first obtain their maximum likelihood estimates and then we use them in the Bayesian estimator. The approach in developing the Bayesian estimator was to find :

$$\hat{\mathbf{x}}_B = \max_{\mathbf{x}} P(\mathbf{x} | \mathbf{y}) . \quad (3.14)$$

The maximum likelihood estimator for  $\mathbf{z} = \begin{bmatrix} \sigma_x^2 & \sigma_n^2 \end{bmatrix}$  can be obtained as:

$$\hat{\mathbf{z}} = \arg \max_{\mathbf{z}} P_{\mathbf{Y}}(\mathbf{Y} = \mathbf{y}) . \quad (3.15)$$

If  $\mathbf{U}$  is the left singular matrix of  $\mathbf{A}$ , then  $\mathbf{y}_m = \mathbf{U}^H \mathbf{y}$  has a normal distribution with zero mean and diagonal covariance matrix  $\sigma_x^2 \mathbf{A} + \sigma_n^2 \mathbf{I}$  where  $\mathbf{A}$  is the diagonal matrix with entries which are the square of the singular values of  $\mathbf{A}$ . Hence, the probability density function of  $\mathbf{y}_m$  is:

$$f_{\mathbf{y}_m}(\mathbf{Y}_m) = \prod_{i=1}^N \frac{\exp(\frac{-y_{mi}^2}{\lambda_i \sigma_x^2 + \sigma_n^2})}{\sqrt{2\pi(\lambda_i \sigma_x^2 + \sigma_n^2)}} . \quad (3.16)$$

Maximizing  $f_{\mathbf{y}_m}(\mathbf{Y}_m)$  with respect to  $\mathbf{z}$  is equivalent to maximizing

$$\begin{aligned} J(\sigma_x^2, \sigma_n^2) &= \log f_{\mathbf{y}_m}(\mathbf{Y}_m) \\ &= \sum_{i=1}^N \frac{-y_{mi}^2}{\lambda_i \sigma_x^2 + \sigma_n^2} - \log \sqrt{2\pi(\lambda_i \sigma_x^2 + \sigma_n^2)} . \end{aligned} \quad (3.17)$$

Taking partial derivatives with respect to  $\sigma_x^2$  and  $\sigma_n^2$  we obtain:

$$\frac{\partial J}{\partial \sigma_x^2} = \sum_{i=1}^N \lambda_i \frac{y_{mi}^2 - \lambda_i \sigma_x^2 - \sigma_n^2}{(\sigma_x^2 \lambda_i + \sigma_n^2)^2} \quad (3.18)$$

$$\frac{\partial J}{\partial \sigma_n^2} = \sum_{i=1}^N \frac{y_{mi}^2 - \lambda_i \sigma_x^2 - \sigma_n^2}{(\sigma_x^2 \lambda_i + \sigma_n^2)^2} . \quad (3.19)$$



To find the solutions that annihilate these quantities we may use the successive substitution method, to get:

$$\sigma_{x(k+1)}^2 = \frac{\sum_{i=1}^N \frac{\frac{1}{2}\lambda_i - \lambda_i y_{mi}^2 \sigma_{n(k)}^2}{(\sigma_{x(k)}^2 \lambda_i + \sigma_{n(k)}^2)^2}}{\sum_{i=1}^N \frac{\lambda_i^2 y_{mi}^2}{(\sigma_{x(k)}^2 \lambda_i + \sigma_{n(k)}^2)^2}}, \quad (3.20)$$

$$\sigma_{n(k+1)}^2 = \frac{\sum_{i=1}^N \frac{\frac{1}{2} - y_{mi}^2 \sigma_{n(k)}^2}{(\sigma_{x(k)}^2 \lambda_i + \sigma_{n(k)}^2)^2}}{\sum_{i=1}^N \frac{y_{mi}^2 \lambda_i}{(\sigma_{x(k)}^2 \lambda_i + \sigma_{n(k)}^2)^2}}, \quad (3.21)$$

where  $\sigma_{x(k)}^2$  and  $\sigma_{n(k)}^2$  stand for the values of  $\sigma_x^2$  and  $\sigma_n^2$  at step  $k$  of the iterations. We could also use a gradient descent method to successively converge to the solutions for  $\sigma_x^2$  and  $\sigma_n^2$ .

### 3.2.4 Simulation Results

To test the performance of the above proposed algorithm we make use of the same synthetic example used in the previous sections. The results are shown in Figure 3.3 and 3.4. As seen from these figures, the estimator we suggested do not suffer from the multi-collinearity problem in the kernel matrix  $\mathbf{A}$  because it belongs to the class of ridge regression estimators. The signal and noise variance estimation process in the Gauss-Markov with recursive updates algorithm ends up by converging to values that are within 10% of the actual values (  $\sigma_x^2 = 1$  and  $\hat{\sigma}_x^2 = 0.92$ ,  $\sigma_n^2 = 0.02$  and  $\hat{\sigma}_n^2 = 0.018$  ). These results are plugged into equation 3.11 and the estimate that we obtain shows good performance which is robust to the noise standard deviation or kernel matrix condition number.

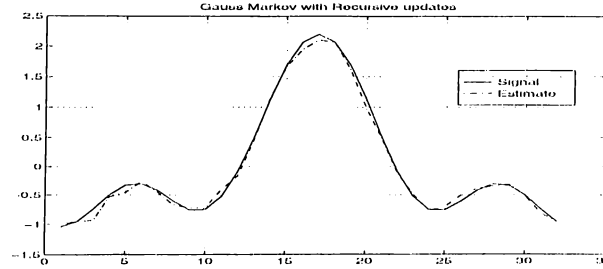


Figure 3.3: Application of Gauss-Markov with recursive updates algorithm. SNR=45dB and low kernel matrix condition number, %error= 2.

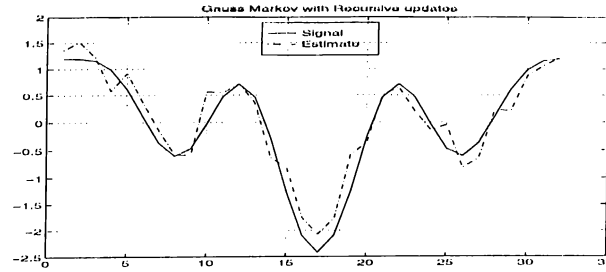


Figure 3.4: Application of Gauss-Markov with recursive updates algorithm. SNR=45dB and high kernel matrix condition number, %error= 8.

### 3.2.5 A Wavelet Based Recursive Reconstruction Algorithm

Reconstruction of the unknowns from the data has been the subject matter of many inverse problems arising in a vast class of applications as geophysical signal processing and speech processing. A very important first step of the inverse problems is the parameterization of the unknowns. In many applications, where the sensitivity of the measurements varies across the space of the unknowns, the space of the unknowns is partitioned into cells of non-uniform sizes. The dimensions of cells becomes larger when the sensitivity of the measurements to those

cells becomes weaker. In order to keep the computational complexity at a low level, usually data independent partitions are used. In this way the reconstruction performance ameliorate with respect to the case when uniform partitions are used. However this result could be further improved when the partitions are chosen adaptively based on the available data.

A new data dependent recursive reconstruction algorithm has been proposed for robust and efficient estimation of the unknowns [6]. In this algorithm, the parameterization of the space of unknowns are performed by using an appropriate wavelet basis for the application at hand. The algorithm recursively updates the solution in an increasingly larger dimensional subspace whose basis vectors are chosen as a subset of the wavelet basis. Robust criteria on how to choose the basis vectors at each iteration, and when to stop the iterations are given in [6].

In this approach, an optimal subspace of the domain of  $\mathbf{x}$  will be searched such that least-squares inversion within this subspace provides a satisfactory reconstruction. For this purpose, a properly chosen wavelet basis can be used.

Wavelets are relatively recent development in applied mathematics. Their name itself was coined in 1982 [7]. But interest in them has grown at an explosive rate. There are several reasons for their wide spread use: Wavelets have been successfully used in subband coding, signal analysis and numerical analysis. The wavelet transform is a tool that cuts up data or functions or operators into different frequency components and then studies each component with a resolution matched to its scale, i.e., the localization in space and scaling are the hallmarks of the wavelet expansion.

The search for the appropriate dimensional subspace of the unknown  $\mathbf{x}$  will be performed in steps of increasing dimensions with the addition of new basis components to the existing ones in the previously formed subspace. The order in which the basis components should be used must be determined efficiently. A close approximation to the set of basis can be obtained by using the matching-pursuit algorithm where the first basis component  $\phi_1$  is chosen as the one which maximizes  $\|\mathbf{y}^H \mathbf{A} \phi_i\|^2$ , and then at step  $n$  of the recursions the optimal set of basis components is updated by adding the basis vector which has the largest absolute inner product with the residual measurement vector, i.e.,

$$\mathbf{b}_{n+1} = \arg \max_{\mathbf{b}_i} |(\mathbf{y} - \mathbf{y}_n)^H \mathbf{b}_i| \quad (3.22)$$

where  $\mathbf{y}_n$  is the estimate of the measurement by using  $n$  basis components.

Define the decomposition of  $\mathbf{x}$  onto the first  $n$  basis components to be:

$$\mathbf{x}_n = \sum_{i=1}^n \phi_i \alpha_i \quad (3.23)$$

Our aim is to determine  $\hat{\mathbf{x}}_n$  such that

$$\hat{\mathbf{x}}_n = \arg \min_{\mathbf{x}_n} \|\mathbf{y} - \mathbf{A} \mathbf{x}_n\|^2 + \mu \|\mathbf{x}_n\|^2 \quad (3.24)$$

where  $\mu$  is the ridge regression constant which when set to 0 yields the Least Squares estimator. The above minimization problem is equivalent to finding:

$$\hat{\boldsymbol{\alpha}}_n = \arg \min_{\boldsymbol{\alpha}_n} \|\mathbf{y} - \mathbf{B}_n \boldsymbol{\alpha}_n\|^2 + \mu \|\boldsymbol{\alpha}_n\|^2 \quad (3.25)$$

with  $\boldsymbol{\alpha}_n = [\alpha_1 \dots \alpha_n]^T$ , and  $\mathbf{B}_n = [\mathbf{b}_1 \dots \mathbf{b}_n]$ . As given previously the solution to this optimization problem is:

$$\hat{\boldsymbol{\alpha}}_n = (\mathbf{B}_n^H \mathbf{B}_n + \mu \mathbf{I})^{-1} \mathbf{B}_n^H \mathbf{y} \quad (3.26)$$

Since the optimal number of basis components to be used is not known a priori, estimates for  $\hat{\alpha}_n$  for various values  $n$  should be obtained in the search for the right number of basis components. If equation 3.26 is used for the estimates, then the order of computations is high because of the matrix inversion that should be performed at each step. Fortunately, there is an efficient way of updating the matrix  $(\mathbf{B}_n^H \mathbf{B}_n + \mu \mathbf{I})^{-1}$  for two consecutive values of  $n$ . At step 1 compute:

$$\mathbf{H}_1 = (\mathbf{B}_1^H \mathbf{B}_1 + \mu \mathbf{I})^{-1} \quad (3.27)$$

$$\mathbf{h}_1 = \mathbf{b}_1^H \mathbf{y} \quad (3.28)$$

$$\hat{\alpha}_1 = \mathbf{H}_1 \mathbf{h}_1 \quad (3.29)$$

The general step of the algorithm which updates  $\mathbf{H}_i$  makes use of the following matrix inversion result:

$$\begin{bmatrix} \mathbf{R} & \mathbf{r} \\ \mathbf{r}^H & \rho \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{Q} & \mathbf{q} \\ \mathbf{q}^H & \kappa \end{bmatrix} \quad (3.30)$$

where  $\mathbf{R}$  is an invertible matrix,  $\mathbf{r}$  is a vector and  $\rho$  is a scalar and:

$$\kappa = \frac{1}{\rho - \mathbf{r}^H \mathbf{R}^{-1} \mathbf{r}} \quad (3.31)$$

$$\mathbf{q} = -\kappa \mathbf{R}^{-1} \mathbf{r} \quad (3.32)$$

$$\mathbf{Q} = \mathbf{R}^{-1} - \mathbf{R}^{-1} \mathbf{r} \mathbf{q}^H \quad (3.33)$$

Based on the above formula, the general step of the recursion in the update from  $n$  to  $n+1$  is given by:

$$\boldsymbol{\theta}_{n+1} = \mathbf{B}_n^H \mathbf{b}_{n+1} \quad (3.34)$$

$$\gamma_{n+1} = \mathbf{H}_n \boldsymbol{\theta}_{n+1} \quad (3.35)$$

$$\beta_{n+1} = \frac{1}{\mathbf{b}_{n+1}^H \mathbf{b}_{n+1} + \mu - \boldsymbol{\theta}_{n+1}^H \gamma_{n+1}} \quad (3.36)$$

$$\boldsymbol{\eta}_{n+1} = -\gamma_{n+1}\beta_{n+1} \quad (3.37)$$

$$\mathbf{H}_{n+1} = \begin{bmatrix} \mathbf{H}_n - \gamma_{n+1}\boldsymbol{\eta}_{n+1}^H & \boldsymbol{\eta}_{n+1} \\ \boldsymbol{\eta}_{n+1}^H & \beta_{n+1} \end{bmatrix} \quad (3.38)$$

$$\epsilon_{n+1} = \beta_{n+1} (\mathbf{b}_{n+1}^H \mathbf{y}) \quad (3.39)$$

$$\nu_{n+1} = (\boldsymbol{\theta}_{n+1}^H \hat{\boldsymbol{\alpha}}_n) \beta_{n+1} - \epsilon_{n+1} \quad (3.40)$$

$$\hat{\boldsymbol{\alpha}}_{n+1} = \begin{bmatrix} \hat{\boldsymbol{\alpha}}_n + \nu_{n+1} \gamma_{n+1} \\ -\nu_{n+1} \end{bmatrix} \quad (3.41)$$

For an  $N$  dimensional  $\mathbf{y}$  we have  $2n^2 + n(N+4) + 2N + 2$  multiplications at each step of the recursion, the total number of multiplications required to compute  $\hat{\boldsymbol{\alpha}}_n$  is  $O(Nn^2)$  for  $N > n$ , whereas the direct use of equation 3.26 requires  $O(Nn^2)$  multiplications at each step. Therefore, the computational saving of the recursive algorithm over the direct solution is significant. Also, the recursive algorithm provides estimates  $\hat{\boldsymbol{\alpha}}_n$  at each step of the recursion making it possible to easily implement criteria to stop the iteration. One important quantity that is helpful in the decision to stop the iterations is the measurement fit error:

$$e(n) = \|\mathbf{y} - \mathbf{B}_n \hat{\boldsymbol{\alpha}}_n\|^2, \quad (3.42)$$

which is a decreasing function of  $n$ . One commonly used criterion stops the iterations when  $e(n)$  is either small enough or reaches a plateau region following a fast decrease. Another stop criterion makes also the use of the norm of the estimate at each step.

### 3.2.6 Simulation Results

To test the performance of this algorithm we generated a measurement kernel such that the norm its columns decreases rapidly as the column index gets larger. This type of rapid decrease is a common case in remote sensing applications where the domain of unknowns is partitioned with a uniform grid. The SNR in this simulation is 35dB. The Haar basis is chosen to be the wavelet basis for the domain of unknowns in this example. The ridge regression parameter,  $\mu$ , is set to  $\sigma_n^2/\sigma_x^2$ .

The criterion we suggest for the number of the basis components used in the estimation is based on the magnitude of the reconstructed vector  $\mathbf{x}_r$ . Since the fit error will decrease rapidly till it reaches a plateau where almost no improvement is observed, the magnitude of the reconstructed vector would indicate for us when the noise fitting process starts, this occurs when an abrupt change in the magnitude takes place. Thus we will avoid that by choosing the appropriate number of basis components. Figure 3.5 shows the plot of the fit error and the estimated vector magnitudes versus the number of basis components used for the reconstruction, figure 3.6 shows the reconstructed estimator by using 10 components of the basis since we stop when the norm of the estimate increases rapidly.

The obtained estimate is very close to the true values of the unknown vector, this result shows that the presented algorithm provides satisfactory results with a highly reduced cost of computations.

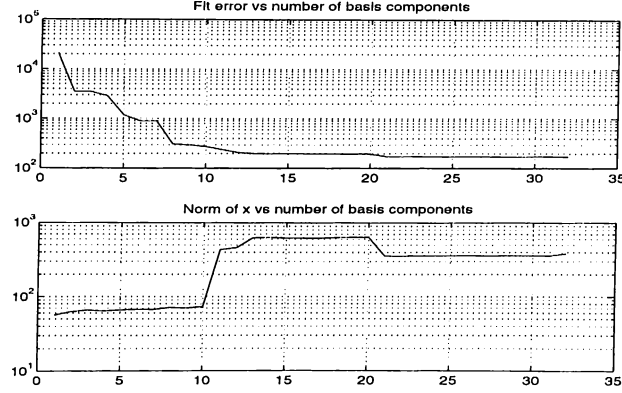


Figure 3.5: Fit Error and Magnitude of the estimate versus the number of basis components.

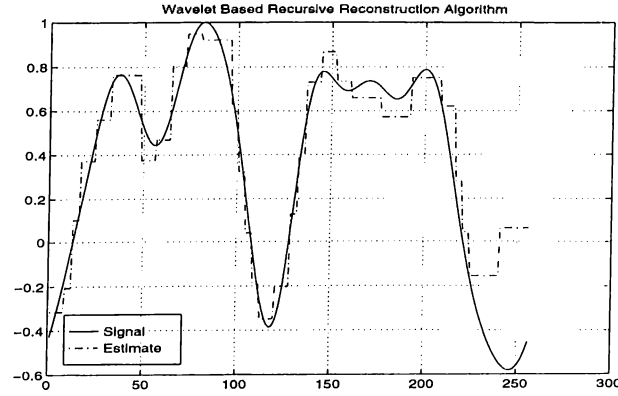


Figure 3.6: Reconstructed estimate by using 10 components of the basis.

### 3.2.7 Comparing performances

Figure 3.7 shows the performance of the Error Dependent Ridge Regression, Gauss Markov with recursive updates, Wavelet based recursive reconstruction algorithms together with the Ridge regression estimator suggested by Swamy, Mehta and Rappoport when applied to a case where the sensitivity of the measurements varies across the space of the unknowns. The algorithms that we



presented outperform the Swamy, Mehta and Rappoport Ridge Regression estimate. Figure 3.8 shows a plot of the estimation error  $\|\mathbf{x} - \hat{\mathbf{x}}\|$  versus the signal to noise ratio for the above mentioned methods. The Wavelet Based Recursive Reconstruction and the Error Dependent Ridge Regression algorithms beat the Gauss-Markov and the Swamy-Mehta and Rappoport's algorithms. This is because the Wavelet based algorithm takes into consideration the sensitivity of the measurements across the space of the unknowns by searching for the optimal subspace of the unknown  $\mathbf{x}$  and the Error Dependent Ridge Regression method imposes on the estimate to yield to an error having the same statistics as the noise, thus for low signal to noise ratio the emphasis on the noise is stressed more than in other approaches. The Error Dependent Ridge Regression and the Wavelet based Recursive reconstruction algorithms performances are shown on figure 3.9.

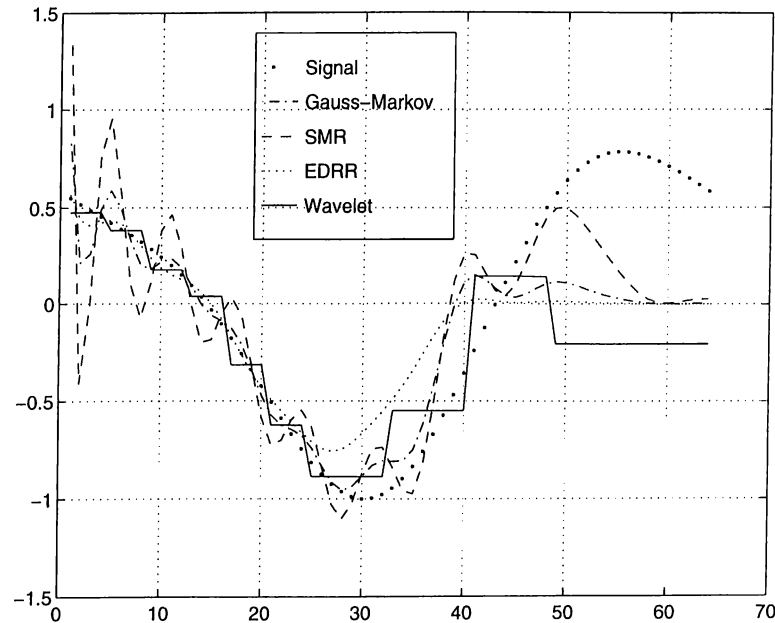


Figure 3.7: WBRR, EDRR, GM, SMR estimates, with %errors:  $e_{WBRR} = 18.37$ ,  $e_{EDRR} = 15.23$ ,  $e_{GM} = 13.82$ ,  $e_{SMR} = 24.54$

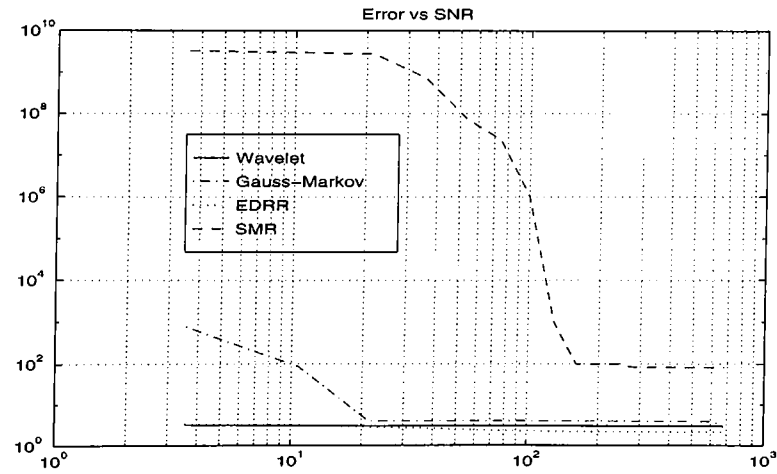


Figure 3.8: Estimation Error vs. SNR for the WBRR, EDRR, GM and SMR estimates.

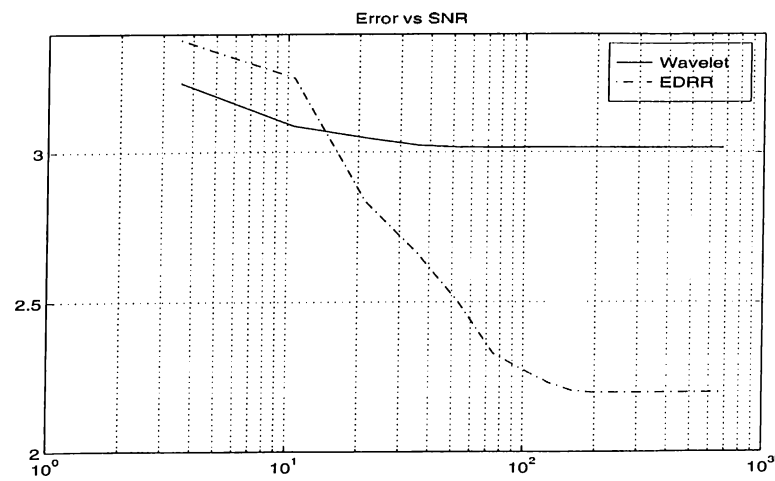


Figure 3.9: Estimation Error vs. SNR for the WBRR and EDRR estimates.

### 3.3 Uncertain Model

#### 3.3.1 Nonlinear Ridge Regression Modeling

In the section of the previously suggested algorithms we presented the Nonlinear Least Squares Modeling method proposed by Cadzow to solve the system of equations when the system matrix elements are known functions of a set of parameters, and through a synthetic example, we saw that its performance deteriorates in the case of high signal to noise ratio or high kernel matrix condition number. In a way these were the same problems we faced when considering the Least Squares problem. Therefore, by introducing a penalty term on the squared norm of the estimate, we can hope to obtain a similar improvement in the performance of the non linear least squares estimator. In the following we will call the penalized approach as the non linear ridge regression modeling which provides estimates for  $\mathbf{x}$  and  $\boldsymbol{\theta}$  as the solution to:

$$\min_{\mathbf{x} \in \mathbb{C}^M} \min_{\boldsymbol{\theta} \in \mathbb{R}^P} \|\mathbf{y} - \mathbf{A}(\boldsymbol{\theta})\mathbf{x}\|^2 + \mu \|\mathbf{x}\|^2, \quad (3.43)$$

where  $\mu$  is the ridge regression constant. For the general case there is no direct form solution to this problem. One way to find the optimal  $\hat{\mathbf{x}}$  and  $\hat{\boldsymbol{\theta}}$  is to use the same non linear optimization technique applied previously in non linear least squares modeling. For any value of  $\boldsymbol{\theta}$  the optimal estimator that minimizes the above cost is :

$$\hat{\mathbf{x}}_{RR} = (\mathbf{A}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})^H + \mu\mathbf{I})^{-1} \mathbf{A}(\boldsymbol{\theta})^H \mathbf{y} \quad (3.44)$$

Substituting equation 3.44 into equation 3.43 the dimension of the optimization problem can be reduced to the dimension of  $\boldsymbol{\theta}$ .

$$\begin{aligned}
& \min_{\mathbf{x} \in \mathbb{C}^M} \min_{\boldsymbol{\theta} \in \mathbb{R}^P} \|\mathbf{y} - \mathbf{A}(\boldsymbol{\theta})\mathbf{x}\|^2 + \mu\|\mathbf{x}\|^2 \\
&= \min_{\boldsymbol{\theta} \in \mathbb{R}^P} \|\mathbf{y} - \mathbf{A}(\boldsymbol{\theta})(\mathbf{A}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})^H + \mu\mathbf{I})^{-1}\mathbf{A}(\boldsymbol{\theta})^H\mathbf{y}\|^2 \\
&\quad + \mu\|(\mathbf{A}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})^H + \mu\mathbf{I})^{-1}\mathbf{A}(\boldsymbol{\theta})^H\mathbf{y}\|^2 \\
&= \min_{\boldsymbol{\theta} \in \mathbb{R}^P} \|[\mathbf{I} - \mathbf{D}(\boldsymbol{\theta})]\mathbf{y}\|^2 + \mu\|\mathbf{B}(\boldsymbol{\theta})\mathbf{y}\|^2, \tag{3.45}
\end{aligned}$$

where  $\mathbf{D}(\boldsymbol{\theta})$  and  $\mathbf{B}(\boldsymbol{\theta})$  are:

$$\mathbf{D}(\boldsymbol{\theta}) = \mathbf{A}(\boldsymbol{\theta})(\mathbf{A}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})^H + \mu\mathbf{I})^{-1}\mathbf{A}(\boldsymbol{\theta})^H \tag{3.46}$$

$$\mathbf{B}(\boldsymbol{\theta}) = (\mathbf{A}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})^H + \mu\mathbf{I})^{-1}\mathbf{A}(\boldsymbol{\theta})^H \tag{3.47}$$

If we define the function to be minimized by:

$$f(\boldsymbol{\theta}) = \|[\mathbf{I} - \mathbf{D}(\boldsymbol{\theta})]\mathbf{y}\|^2 + \mu\|\mathbf{B}(\boldsymbol{\theta})\mathbf{y}\|^2, \tag{3.48}$$

then the problem reduces to finding a perturbation vector  $\boldsymbol{\delta}$  such that:

$$f(\boldsymbol{\theta} + \boldsymbol{\delta}) < f(\boldsymbol{\theta}). \tag{3.49}$$

Assuming that the perturbation vector is sufficiently small in size, we can use Taylor series expansion for  $\mathbf{D}(\boldsymbol{\theta} + \boldsymbol{\delta})$ , giving:

$$\mathbf{D}(\boldsymbol{\theta} + \boldsymbol{\delta}) \simeq \mathbf{D}(\boldsymbol{\theta}) + \sum_{k=1}^P \frac{\partial \mathbf{D}(\boldsymbol{\theta})}{\partial \theta_k} \delta_k. \tag{3.50}$$

Similarly using the Taylor series expansion for  $\mathbf{B}(\boldsymbol{\theta} + \boldsymbol{\delta})$ , we obtain:

$$\mathbf{B}(\boldsymbol{\theta} + \boldsymbol{\delta}) \simeq \mathbf{B}(\boldsymbol{\theta}) + \sum_{k=1}^P \frac{\partial \mathbf{B}(\boldsymbol{\theta})}{\partial \theta_k} \delta_k. \tag{3.51}$$

Using the above expansions

$$\begin{aligned}
f(\boldsymbol{\theta} + \boldsymbol{\delta}) &\simeq \|[\mathbf{I} - \mathbf{D}(\boldsymbol{\theta} + \boldsymbol{\delta})]\mathbf{y}\|^2 + \mu\|\mathbf{B}(\boldsymbol{\theta})\mathbf{y}\|^2 \\
&= \|[\mathbf{I} - \mathbf{D}(\boldsymbol{\theta}) - \sum_{k=1}^P \frac{\partial \mathbf{D}(\boldsymbol{\theta})}{\partial \theta_k} \delta_k]\mathbf{y}\|^2 + \mu\|[\mathbf{B}(\boldsymbol{\theta}) - \sum_{k=1}^P \frac{\partial \mathbf{B}(\boldsymbol{\theta})}{\partial \theta_k} \delta_k]\mathbf{y}\|^2 \\
&= \|[\mathbf{I} - \mathbf{D}(\boldsymbol{\theta})]\mathbf{y} - \mathbf{L}_1(\boldsymbol{\theta})\boldsymbol{\delta}\|^2 + \mu\|\mathbf{B}(\boldsymbol{\theta})\mathbf{y} - \mathbf{L}_2(\boldsymbol{\theta})\boldsymbol{\delta}\|^2, \quad (3.52)
\end{aligned}$$

where the Jacobian matrices  $\mathbf{L}_1(\boldsymbol{\theta})$  and  $\mathbf{L}_2(\boldsymbol{\theta})$  are obtained by:

$$\mathbf{L}_1(\boldsymbol{\theta}) = \left[ \frac{\partial \mathbf{D}(\boldsymbol{\theta})}{\partial \theta_1} \mathbf{y} : \frac{\partial \mathbf{D}(\boldsymbol{\theta})}{\partial \theta_2} \mathbf{y} : \dots : \frac{\partial \mathbf{D}(\boldsymbol{\theta})}{\partial \theta_P} \mathbf{y} \right] \quad (3.53)$$

$$\mathbf{L}_2(\boldsymbol{\theta}) = \left[ \frac{\partial \mathbf{B}(\boldsymbol{\theta})}{\partial \theta_1} \mathbf{y} : \frac{\partial \mathbf{B}(\boldsymbol{\theta})}{\partial \theta_2} \mathbf{y} : \dots : \frac{\partial \mathbf{B}(\boldsymbol{\theta})}{\partial \theta_P} \mathbf{y} \right], \quad (3.54)$$

and  $\delta_j$  or  $\theta_j$  are the  $j^{th}$  entry of  $\boldsymbol{\delta}$  or  $\boldsymbol{\theta}$ , respectively. A logical choice of the perturbation vector would be one that minimizes the Euclidean norm criterion given by equation 3.52. By setting the gradient of this expression, with respect to  $\boldsymbol{\delta}$ , to the zero vector the optimal selection is :

$$\begin{aligned}
\boldsymbol{\delta}^* &= \Re\{(\mathbf{L}_1(\boldsymbol{\theta})^H \mathbf{L}_1(\boldsymbol{\theta}) + \mu \mathbf{L}_2(\boldsymbol{\theta})^H \mathbf{L}_2(\boldsymbol{\theta}))^{-1}\} \\
&\quad \Re\{\mathbf{L}_1(\boldsymbol{\theta})^H [\mathbf{I} - \mathbf{D}(\boldsymbol{\theta})]\mathbf{y} + \mu \mathbf{L}_2(\boldsymbol{\theta})^H \mathbf{B}(\boldsymbol{\theta})\mathbf{y}\}. \quad (3.55)
\end{aligned}$$

To ensure a sufficiently small perturbation, a scaled perturbation vector  $\alpha \boldsymbol{\delta}^*$  is used instead. The stopping conditions could be the fit error norm. The steps of non linear optimization algorithm are given in table 3.1.

An other way to solve the minimization problem equation 3.45 is to apply gradient descent method to reach the optimal  $\hat{\boldsymbol{\theta}}$  that minimizes  $J(\boldsymbol{\theta}) = \|[\mathbf{I} - \mathbf{D}(\boldsymbol{\theta})]\mathbf{y}\|^2 + \mu\|\mathbf{B}(\boldsymbol{\theta})\mathbf{y}\|^2$ . Starting with an initial  $\hat{\boldsymbol{\theta}}_0$  we perform updating by:

$$\hat{\boldsymbol{\theta}}_k = \hat{\boldsymbol{\theta}}_{k-1} + \mu_{\theta} \frac{\partial J}{\partial \boldsymbol{\theta}} \big|_{\hat{\boldsymbol{\theta}}_{k-1}}, \quad (3.56)$$

Step	Description
1	Start by an initial $\boldsymbol{\theta}$
2	Evaluate $\ [\mathbf{I} - \mathbf{D}(\boldsymbol{\theta})]\mathbf{y}\ ^2 + \mu\ \mathbf{B}(\boldsymbol{\theta})\mathbf{y}\ ^2$
3	Determine $\mathbf{L}_1(\boldsymbol{\theta})$ and $\mathbf{L}_2(\boldsymbol{\theta})$
4	Compute the optimum perturbation vector $\boldsymbol{\delta}^*$
5	Evaluate $\ [\mathbf{I} - \mathbf{D}(\boldsymbol{\theta} + \alpha\boldsymbol{\delta}^*)]\mathbf{y}\ ^2 + \mu\ \mathbf{B}(\boldsymbol{\theta} + \alpha\boldsymbol{\delta}^*)\mathbf{y}\ ^2$ for $\alpha = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots$ until improvement.
6	Evaluate stopping conditions, if not satisfied, set $\boldsymbol{\theta} = \boldsymbol{\theta} + \alpha\boldsymbol{\delta}^*$ and go to step 2.

Table 3.1: Nonlinear Ridge Regression Algorithm.

where  $\frac{\partial J}{\partial \boldsymbol{\theta}}$ , the derivative of the cost function with respect to  $\boldsymbol{\theta}$ , is given by:

$$\begin{aligned}
\frac{\partial J}{\partial \theta_j} = & -\mathbf{y}^H \frac{\partial \mathbf{D}(\boldsymbol{\theta})}{\partial \theta_j} \mathbf{y} - \mathbf{y}^H \frac{\partial \mathbf{D}(\boldsymbol{\theta})}{\partial \theta_j} \mathbf{y} \\
& + \mathbf{y}^H \frac{\partial \mathbf{D}(\boldsymbol{\theta})}{\partial \theta_j} \mathbf{D}(\boldsymbol{\theta}) \mathbf{y} + \mathbf{y}^H \mathbf{D}(\boldsymbol{\theta})^H \frac{\partial \mathbf{D}(\boldsymbol{\theta})}{\partial \theta_j} \mathbf{D}(\boldsymbol{\theta}) \mathbf{y} \\
& + \mu \mathbf{y}^H \frac{\partial \mathbf{B}(\boldsymbol{\theta})}{\partial \theta_j} \mathbf{B}(\boldsymbol{\theta}) \mathbf{y} + \mu \mathbf{y}^H \mathbf{B}(\boldsymbol{\theta})^H \frac{\partial \mathbf{B}(\boldsymbol{\theta})}{\partial \theta_j} \mathbf{y} ,
\end{aligned}$$

for  $j = 1..P$ , and  $\mu_\theta$  is the step size which can be taken small enough to ensure the convergence of the algorithm. Once we reach a minimizing  $\hat{\boldsymbol{\theta}}$ , we plug it into equation 3.44 to find  $\hat{\mathbf{x}}_{RR}$  then check that this point  $Q(\hat{\mathbf{x}}_{RR}, \hat{\boldsymbol{\theta}})$  is a global minimum for both parameters by finding the eigenvalues of the Hessian matrix  $\mathbf{H}$  at  $Q$ , where

$$\mathbf{H}(\mathbf{x}, \boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial^2 J}{\partial \mathbf{x}^2} & \frac{\partial^2 J}{\partial \mathbf{x} \partial \boldsymbol{\theta}} \\ \frac{\partial^2 J}{\partial \boldsymbol{\theta} \partial \mathbf{x}} & \frac{\partial^2 J}{\partial \boldsymbol{\theta}^2} \end{bmatrix} \quad (3.57)$$

At the global minimum point the Hessian matrix is positive definite. The speed of convergence of this method is highly related to the starting point of the algorithm and to the step size used. Taking a large step size may stick the iterations at a local minimum. One way to overcome this problem is to apply the algorithm

by initializing several starting points, so that we can avoid being clung at local minima points.

The derivation of the Jacobian matrices  $\mathbf{L1}(\boldsymbol{\theta})$  and  $\mathbf{L2}(\boldsymbol{\theta})$  are provided in the Appendix.

### 3.3.2 Maximum Likelihood and Least Squares Bayesian Inversion Approaches

Both non linear least squares and ridge regression estimators of  $\mathbf{x}$  in the presence of a parametric uncertainty in  $\mathbf{A}(\boldsymbol{\theta})$  provide estimates in the absence of any prior information. In this section, we will present two approaches to the estimation of unknowns  $\mathbf{x}$  when there is available prior information on the set of parameters  $\boldsymbol{\theta}$ . Typically this prior information on  $\boldsymbol{\theta}$  can be a constrained set, such as:

$$\boldsymbol{\theta}_L \leq \boldsymbol{\theta} \leq \boldsymbol{\theta}_U, \quad (3.58)$$

or it can be a density function  $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ . One way of incorporating this type of prior information on  $\boldsymbol{\theta}$  to the estimation of  $\mathbf{x}$  is given in the following estimator which is based on maximum likelihood principle:

$$\hat{\mathbf{x}}_M = \arg \max_{\mathbf{x}} \mathcal{E}_{\boldsymbol{\theta}} \{ \mathcal{L}(\mathbf{y} - \mathbf{A}(\boldsymbol{\theta})\mathbf{x}) \} , \quad (3.59)$$

where  $\mathcal{E}_{\boldsymbol{\theta}} \{ \mathcal{L}(\mathbf{y} - \mathbf{A}(\boldsymbol{\theta})\mathbf{x}) \}$  is the expected likelihood with respect to  $\boldsymbol{\theta}$ .

$$\mathcal{E}_{\boldsymbol{\theta}} \{ \mathcal{L}(\mathbf{y} - \mathbf{A}(\boldsymbol{\theta})\mathbf{x}) \} = \int \mathcal{L}(\mathbf{y} - \mathbf{A}(\boldsymbol{\theta})\mathbf{x}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta} , \quad (3.60)$$

where  $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$  is the prior density on  $\boldsymbol{\theta}$ , For the commonly used zero mean Gaussian noise model with autocorrelation  $\mathbf{R}_{nn}$ , the likelihood function is:

$$\mathcal{L}(\mathbf{y} - \mathbf{A}(\boldsymbol{\theta})\mathbf{x}) = \frac{1}{2\pi^{N/2}|\mathbf{R}_{nn}|^{1/2}} \exp\{-(\mathbf{y} - \mathbf{A}(\boldsymbol{\theta})\mathbf{x})^H \mathbf{R}_{nn}^{-1}(\mathbf{y} - \mathbf{A}(\boldsymbol{\theta})\mathbf{x})\} . \quad (3.61)$$

Hence,  $\hat{\mathbf{x}}_M$  in equation 3.59 can be obtained as the maximizer of:

$$J(\mathbf{x}, \boldsymbol{\theta}) = \int \frac{1}{(2\pi\sigma_n)^{N/2}} \exp\{-(\mathbf{y} - \mathbf{A}(\boldsymbol{\theta})\mathbf{x})^H \mathbf{R}_{nn}^{-1}(\mathbf{y} - \mathbf{A}(\boldsymbol{\theta})\mathbf{x})\} f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta} \quad (3.62)$$

There is no closed form expression for the maximizer of this cost. However, numerical optimization tools can be utilized to compute the maximizer  $\hat{\mathbf{x}}_M$ . For instance, assuming that  $\mathbf{R}_{nn} = \sigma_n^2 \mathbf{I}$ , then  $\hat{\mathbf{x}}_M$  that maximizes  $J$  is the vector that annihilates:

$$\frac{\partial J}{\partial \mathbf{x}^H} = \int \frac{1}{(2\pi\sigma_n)^{N/2}} \exp\left\{-\frac{\|\mathbf{y} - \mathbf{A}(\boldsymbol{\theta})\mathbf{x}\|^2}{2\sigma_n^2}\right\} \frac{\mathbf{A}(\boldsymbol{\theta})^H(\mathbf{y} - \mathbf{A}(\boldsymbol{\theta})\mathbf{x})}{2\sigma_n^2} f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta} \quad (3.63)$$

which, by discretizing the subspace of  $\boldsymbol{\theta}$  onto  $K$  possible values of  $\boldsymbol{\theta}_i, i = 1, \dots, K$  equation 3.63 can be approximated by:

$$\frac{\partial J}{\partial \mathbf{x}^H} = \sum_{i=1}^K \frac{1}{(2\pi\sigma_n)^{N/2}} \exp\left\{-\frac{\|\mathbf{y} - \mathbf{A}(\boldsymbol{\theta}_i)\mathbf{x}\|^2}{2\sigma_n^2}\right\} \frac{\mathbf{A}(\boldsymbol{\theta}_i)^H(\mathbf{y} - \mathbf{A}(\boldsymbol{\theta}_i)\mathbf{x})}{2\sigma_n^2} f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i). \quad (3.64)$$

For the solution of equation 3.64, we may use the method of successive substitutions. At iteration  $k+1$   $\mathbf{x}$  is updated by:

$$\mathbf{x}^{(k+1)} = \left( \sum_{i=1}^K f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i) \exp\left\{-\frac{\|\mathbf{y} - \mathbf{A}(\boldsymbol{\theta}_i)\mathbf{x}^{(k)}\|^2}{2\sigma_n^2}\right\} \mathbf{A}(\boldsymbol{\theta}_i)^H \mathbf{A}(\boldsymbol{\theta}_i) \right)^{-1} \left( \sum_{i=1}^K f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i) \exp\left\{-\frac{\|\mathbf{y} - \mathbf{A}(\boldsymbol{\theta}_i)\mathbf{x}^{(k)}\|^2}{2\sigma_n^2}\right\} \mathbf{A}(\boldsymbol{\theta}_i)^H \mathbf{y} \right). \quad (3.65)$$



A second way of incorporating prior information on  $\boldsymbol{\theta}$  is based on least squares principle, where the estimate is obtained as the minimizer of the expected squared norm of the fit error:

$$\hat{\mathbf{x}}_L = \arg \min_{\mathbf{x}} \mathcal{E}_{\boldsymbol{\theta}} \{ \|\mathbf{y} - \mathbf{A}(\boldsymbol{\theta})\mathbf{x}\|^2 \} . \quad (3.66)$$

By using the prior distribution on  $\boldsymbol{\theta}$ , we can rewrite the estimator in the following form:

$$\hat{\mathbf{x}}_L = \arg \min_{\mathbf{x}} \int \|\mathbf{y} - \mathbf{A}(\boldsymbol{\theta})\mathbf{x}\|^2 f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta} \quad (3.67)$$

This quadratic form of the cost function has the following closed form solution for  $\hat{\mathbf{x}}_L$ :

$$\hat{\mathbf{x}}_L = \left[ \int \mathbf{A}^H(\boldsymbol{\theta}) \mathbf{A}(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta} \right]^{-1} \left[ \int \mathbf{A}^H(\boldsymbol{\theta}) \mathbf{y} f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta} \right] . \quad (3.68)$$

In order to avoid potential problems of the required matrix inversion, we can use the following regularized form:

$$\hat{\mathbf{x}}_L = \left[ \int (\mathbf{A}^H(\boldsymbol{\theta}) \mathbf{A}(\boldsymbol{\theta}) + \mu \mathbf{I}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta} \right]^{-1} \left[ \int \mathbf{A}^H(\boldsymbol{\theta}) \mathbf{y} f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta} \right] , \quad (3.69)$$

where  $\mu$  is the regularization parameter.

### 3.3.3 Simulation Results and Comparing Performances

To test the performance of the non linear ridge regression modeling algorithm when applied to estimate an unknown vector  $\mathbf{x}$  under model uncertainties, we apply the same synthetic example used to test the non linear least squares modeling estimator. Figure 3.10 shows the result when the algorithm is used for the case of  $\text{SNR} = 65\text{dB}$  and the kernel matrix condition  $\kappa = 40$ . The algorithm

converges in few steps to the actual values of  $\theta$  and the estimated  $\hat{x}$  is very close to the unknown parameter  $x$ . For kernel matrix  $A(\theta)$  having a condition number  $\kappa = 10^4$  and  $\text{SNR} = 46\text{dB}$ , figure 3.11 displays the result obtained when non linear ridge regression algorithm is applied with use of non linear minimization procedure to solve for  $\theta$  and figure 3.12 shows the result when we apply the gradient descent technique to search for optimal  $\theta$ . The ridge regression constant used is the one provided by Swamy, Mehta and Rappoport.

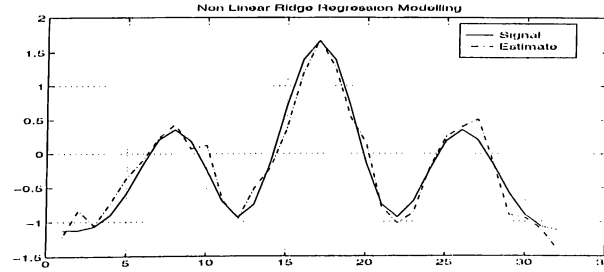


Figure 3.10: Nonlinear Ridge Regression algorithm with non linear minimization technique.  $\text{SNR}=65\text{dB}$ ,  $\% \text{error} = 3.11$

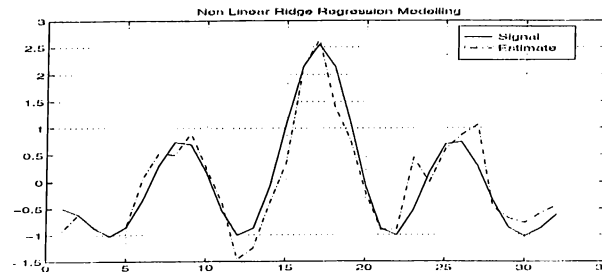


Figure 3.11: Nonlinear Ridge Regression algorithm with non linear minimization technique.  $\text{SNR}=46\text{dB}$  and  $\kappa = 10^4$ ,  $\% \text{error} = 7.42$

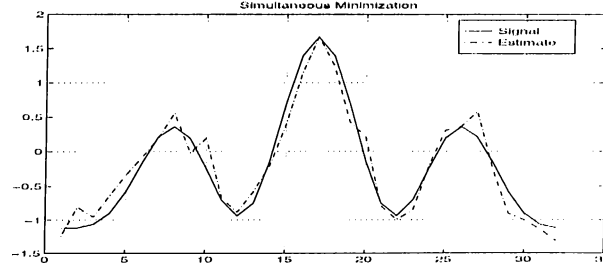


Figure 3.12: Nonlinear Ridge Regression algorithm with Gradient technique. SNR=46dB,  $\kappa = 10^4$ . %error = 3.24.

Figure 3.13 shows the estimated and the actual parameter of  $\mathbf{x}$  by the use of Maximum likelihood-Bayesian approach. The four dimensional space of the unknown parameters  $\boldsymbol{\theta}$  in this example has been sampled to five steps for each dimension. Thus we used 625 different  $\boldsymbol{\theta}$  values. The prior distribution on the  $\boldsymbol{\theta}_i$  space is taken to be proportional to the projection norm of the data vector  $\mathbf{y}$  onto the range space of  $\mathbf{A}(\boldsymbol{\theta}_i)$ . The SNR in this application is 45dB. The performance of this algorithm would increase when the sampling values are taken tighter as shown in figure 3.14, where the estimation error is plotted versus the norm of,  $\boldsymbol{\theta}$ , the bound vector for the parameter  $\boldsymbol{\theta}$ , that is we assume

$$\boldsymbol{\theta}_0 - \boldsymbol{\delta} \leq \boldsymbol{\theta} \leq \boldsymbol{\theta}_0 + \boldsymbol{\delta} \quad (3.70)$$

and we keep the sampling rate described above while varying the bound vector.

Figure 3.14 displays the result when the least squares Bayesian approach is used. The conditions under which this algorithm is applied are the same as in the application of the maximum likelihood-Bayesian approach. Again the performance of this algorithm is highly related to the sampling density applied to discretize the estimate given by equation 3.69.

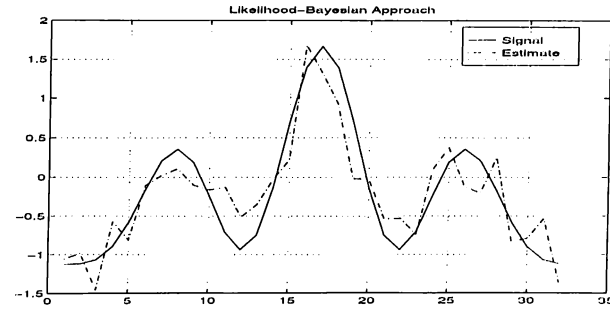


Figure 3.13: Maximum Likelihood-Bayesian approach, %error = 13.17.

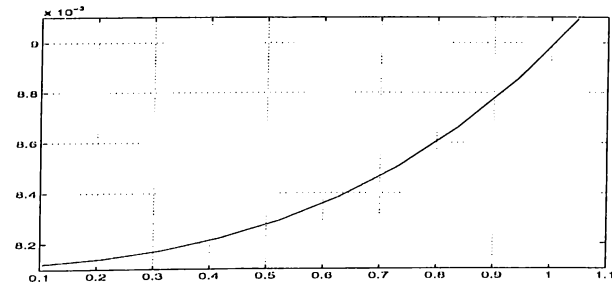


Figure 3.14: Estimation error versus the bound vector norm.

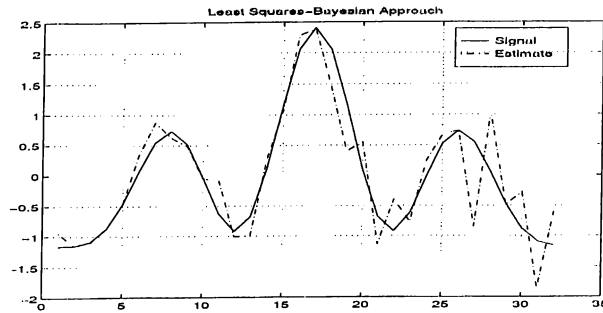


Figure 3.15: Least Squares-Bayesian approach, %error = 12.75.

To end up, we applied the same synthetic example for comparing the non linear least squares modeling, the non linear ridge regression modeling, and the maximum likelihood-bayesian inversion estimators. Figures 3.16- 3.19 display the results obtained for an SNR of 80dB. The non linear least squares modeling approach is based on the least squares estimators, thus its drawbacks that are caused by the noise vector standard deviation and the kernel matrix condition number are inherited. Whereas the non linear ridge regression estimator with its two ways for determining the optimal  $\hat{\theta}$  overcomes the above mentioned problems. On the other hand if prior information on the distribution of the unknown parameter is available, we can use the maximum likelihood-Bayesian inversion or the least squares-Bayesian inversion approaches, these algorithms provide us smooth estimates, and avoid drastically deteriorating results. The disadvantage of those two methods is the need of a fine sampling of the  $\theta$  space that leads to increase in the computational cost. Figure 3.20 gives the estimation error versus the SNR for the presented approaches. We notice that for low SNR the algorithms we suggested outperform the non linear least squares algorithm, however the latter yields better results in the mean square error sense when the SNR is high, this is due to the fact that non linear least squares algorithm gives an unbiased estimator.

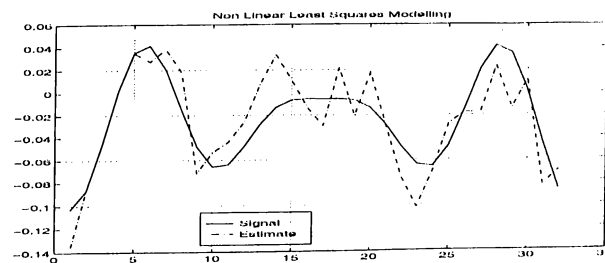


Figure 3.16: Nonlinear Least Squares Modeling, %error = 24.36.

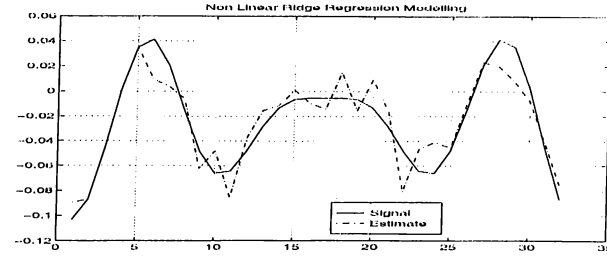


Figure 3.17: Nonlinear Ridge Regression Modeling with non linear optimization, %error = 16.7.

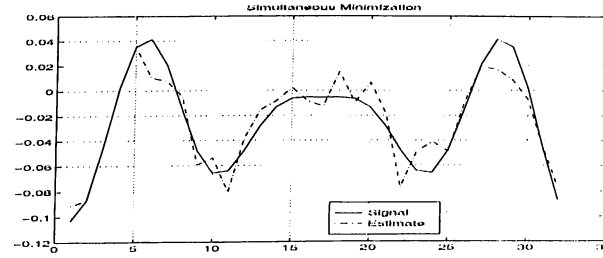


Figure 3.18: Nonlinear Ridge Regression Modeling with gradient descent minimization, %error = 15.9.

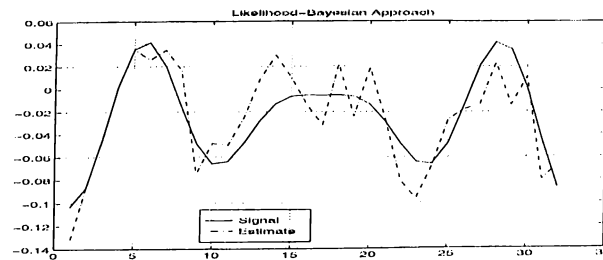


Figure 3.19: Likelihood Bayesian approach, %error = 21.8.

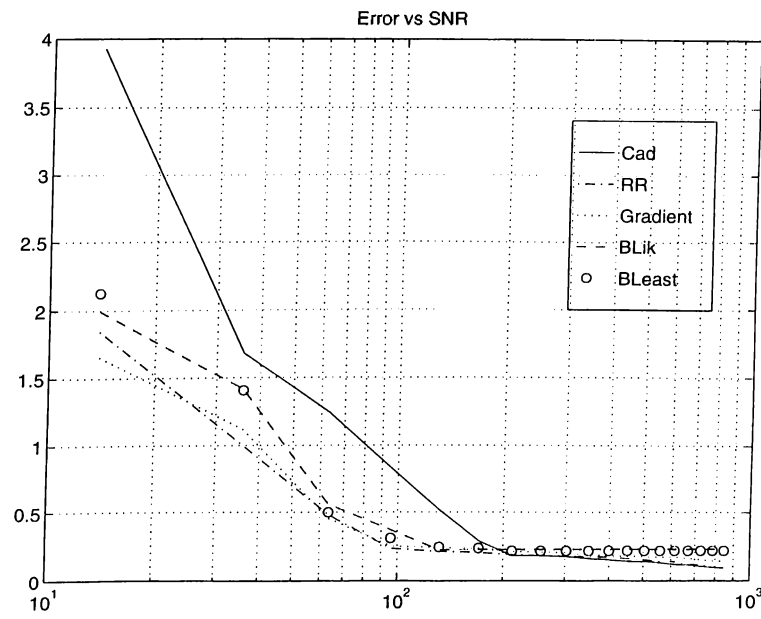


Figure 3.20: Estimation Error versus SNR for the presented algorithms: Cad-zow(Cad), Nonlinear Ridge Regression(RR), Nonlinear Ridge Regression with Gradient descent algorithm(Gradient), Bayesian-Likelihood(BLik) and Bayesian-Least Squares(BLeast).

# Chapter 4

## Conclusions

In this thesis, we have considered the problem of estimation of unknowns in a linear statistical model with uncertainties. We started by reviewing some of the widely used methods, then we introduced our approaches. For both parts we treated the question in two separate cases. When the kernel matrix is known and errors are considered only at the measured data, we investigated the Least Squares and the Ridge Regression estimators. We provided results showing that the mean square error of the Least Squares estimator increases drastically for kernel matrices with high condition numbers and additive noise vectors with large variances. On the other hand, the Ridge Regression estimator overcome such problems of multicollinearity and of low signal to noise ratio at the expense of a required prior information about the unknown parameter vector or about the noise statistics, which are used to determine the ridge regression parameter. In order to avoid such a prior information on the unknowns, we proposed the Error Dependent Ridge Regression approach which chooses a ridge regression



constant that leads to an error with the same second order statistics as the noise vector. The prior information about the noise vector variance can also be avoided by applying the Gauss Markov algorithm with recursive updates. These two algorithms lead to better estimates in the mean square error sense than the Least Squares and the commonly used Ridge Regression estimators. In addition to those methods, a data dependent recursive reconstruction algorithm is proposed for robust and efficient estimation of the unknowns. The algorithm recursively updates the solution in an increasingly larger dimensional subspace whose basis vectors are chosen as a subset of the wavelet basis. Robust criteria on how to choose the basis vectors at each iteration, and when to stop the iterations are provided.

Then we looked into the methods used to solve the problem of uncertainties in the kernel matrix, we first examined the Total Least Squares and the Non Linear Least Squares modeling. Being based on Least Squares, those two methods inherit its deficiencies. To remedy these drawbacks we introduced the Non Linear Ridge Regression Modeling algorithm which is based on the ridge regression estimator. This method reduces the minimization problem with respect to two unknown vectors to a minimization problem with respect to one vector for which a non linear programming algorithm or a gradient descent type algorithm can be used to reach the optimal solution. Finally, to deal with the problem when prior information on the parameter that models the uncertainty in the kernel matrix, is provided we suggested two similar approaches based on maximization of the expected likelihood and minimization of expected least squares cost. Simulation results obtained through synthetic examples demonstrated that the proposed algorithms outperform the commonly used methods providing robust estimates

with smaller mean square errors.

# Appendix A

## Computation of the Jacobian matrices.

In the Non Linear Least Squares Modeling we had:

$$\mathbf{L}(\boldsymbol{\theta}) = [\frac{\partial \mathbf{P}(\boldsymbol{\theta})}{\partial \theta_1} \mathbf{y} ; \frac{\partial \mathbf{P}(\boldsymbol{\theta})}{\partial \theta_2} \mathbf{y} ; \dots ; \frac{\partial \mathbf{P}(\boldsymbol{\theta})}{\partial \theta_P} \mathbf{y}] \quad . \quad (\text{A.1})$$

In the computation of the Jacobian matrix  $\mathbf{L}(\boldsymbol{\theta})$  the problem is to find the derivative of the projection matrix  $\mathbf{P}(\boldsymbol{\theta})$  with respect to the  $P$  dimensional vector  $\boldsymbol{\theta}$ . Knowing that:

$$\mathbf{P}(\boldsymbol{\theta}) = \mathbf{P}(\boldsymbol{\theta})^2 \quad (\text{A.2})$$

and

$$\mathbf{P}(\boldsymbol{\theta}) = \mathbf{P}(\boldsymbol{\theta})^H \quad (\text{A.3})$$

$$\frac{\partial \mathbf{P}(\boldsymbol{\theta})}{\partial \theta_k} = \frac{\partial \mathbf{P}(\boldsymbol{\theta})}{\partial \theta_k} \mathbf{P}(\boldsymbol{\theta}) + \mathbf{P}(\boldsymbol{\theta}) \frac{\partial \mathbf{P}(\boldsymbol{\theta})}{\partial \theta_k} \quad (\text{A.4})$$

$$= \frac{\partial \mathbf{P}(\boldsymbol{\theta})}{\partial \theta_k} \mathbf{P}(\boldsymbol{\theta}) + [\frac{\partial \mathbf{P}(\boldsymbol{\theta})}{\partial \theta_k} \mathbf{P}(\boldsymbol{\theta})]^H \quad (\text{A.5})$$

To compute the terms on the right-hand side of equation A.4, each side of the matrix identity  $\mathbf{P}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta}) = \mathbf{A}(\boldsymbol{\theta})$  is differentiated with respect to  $\theta_k$ .

$$\frac{\partial \mathbf{P}(\boldsymbol{\theta})}{\partial \theta_k} \mathbf{A}(\boldsymbol{\theta}) = [\mathbf{I} - \mathbf{P}(\boldsymbol{\theta})] \frac{\partial \mathbf{A}(\boldsymbol{\theta})}{\partial \theta_k} . \quad (\text{A.6})$$

Right multiplying each side of this relationship by the Moore-Penrose generalized inverse  $\mathbf{A}^\dagger(\boldsymbol{\theta})$ :

$$\frac{\partial \mathbf{P}(\boldsymbol{\theta})}{\partial \theta_k} \mathbf{P}(\boldsymbol{\theta}) = [\mathbf{I} - \mathbf{P}(\boldsymbol{\theta})] \frac{\partial \mathbf{A}(\boldsymbol{\theta})}{\partial \theta_k} \mathbf{A}^\dagger(\boldsymbol{\theta}) . \quad (\text{A.7})$$

This expression is then substituted in equation A.4 to get:

$$\frac{\partial \mathbf{P}(\boldsymbol{\theta})}{\partial \theta_k} = [[\mathbf{I} - \mathbf{P}(\boldsymbol{\theta})] \frac{\partial \mathbf{A}(\boldsymbol{\theta})}{\partial \theta_k} \mathbf{A}^\dagger(\boldsymbol{\theta})] + [[\mathbf{I} - \mathbf{P}(\boldsymbol{\theta})] \frac{\partial \mathbf{A}(\boldsymbol{\theta})}{\partial \theta_k} \mathbf{A}^\dagger(\boldsymbol{\theta})]^H . \quad (\text{A.8})$$

For  $1 \leq k \leq P$ .

In the Non Linear Ridge Regression Modeling algorithm we had:

$$\mathbf{L}_1(\boldsymbol{\theta}) = \left[ \frac{\partial \mathbf{D}(\boldsymbol{\theta})}{\partial \theta_1} \mathbf{y} : \frac{\partial \mathbf{D}(\boldsymbol{\theta})}{\partial \theta_2} \mathbf{y} : \dots : \frac{\partial \mathbf{D}(\boldsymbol{\theta})}{\partial \theta_P} \mathbf{y} \right] \quad (\text{A.9})$$

$$\mathbf{L}_2(\boldsymbol{\theta}) = \left[ \frac{\partial \mathbf{B}(\boldsymbol{\theta})}{\partial \theta_1} \mathbf{y} : \frac{\partial \mathbf{B}(\boldsymbol{\theta})}{\partial \theta_2} \mathbf{y} : \dots : \frac{\partial \mathbf{B}(\boldsymbol{\theta})}{\partial \theta_P} \mathbf{y} \right] , \quad (\text{A.10})$$

where  $\mathbf{D}(\boldsymbol{\theta})$  and  $\mathbf{B}(\boldsymbol{\theta})$  are given by:

$$\mathbf{D}(\boldsymbol{\theta}) = \mathbf{A}(\boldsymbol{\theta})(\mathbf{A}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})^H + \mu \mathbf{I})^{-1} \mathbf{A}(\boldsymbol{\theta})^H$$

$$\mathbf{B}(\boldsymbol{\theta}) = (\mathbf{A}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})^H + \mu \mathbf{I})^{-1} \mathbf{A}(\boldsymbol{\theta})^H$$

The main problem is to compute  $\frac{\partial \mathbf{D}(\boldsymbol{\theta})}{\partial \theta_k}$  and  $\frac{\partial \mathbf{B}(\boldsymbol{\theta})}{\partial \theta_k}$ .

Call  $(\mathbf{A}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})^H + \mu \mathbf{I}) = \mathbf{T}(\boldsymbol{\theta})$ . Then, since  $\mathbf{T}(\boldsymbol{\theta})$  is invertible

$$\mathbf{T}(\boldsymbol{\theta})\mathbf{T}^{-1}(\boldsymbol{\theta}) = \mathbf{I} \quad (\text{A.11})$$

Using this identity and differentiating with respect to  $\theta_k$

$$\frac{\partial(\mathbf{T}(\boldsymbol{\theta})\mathbf{T}^{-1}(\boldsymbol{\theta}))}{\partial\theta_k} = 0 \quad , \quad (\text{A.12})$$

taking partial derivatives

$$\frac{\partial\mathbf{T}(\boldsymbol{\theta})}{\partial\theta_k}\mathbf{T}^{-1}(\boldsymbol{\theta}) + \mathbf{T}(\boldsymbol{\theta})\frac{\partial\mathbf{T}^{-1}(\boldsymbol{\theta})}{\partial\theta_k} = 0 \quad (\text{A.13})$$

therefore the derivative of  $\mathbf{T}^{-1}(\boldsymbol{\theta})$  with respect to  $\boldsymbol{\theta}$  is

$$\frac{\partial\mathbf{T}^{-1}(\boldsymbol{\theta})}{\partial\theta_k} = -\mathbf{T}^{-1}(\boldsymbol{\theta})\frac{\partial\mathbf{T}(\boldsymbol{\theta})}{\partial\theta_k}\mathbf{T}^{-1}(\boldsymbol{\theta}) \quad . \quad (\text{A.14})$$

Thus the derivative of  $\mathbf{D}(\boldsymbol{\theta})$  with respect to  $\theta_k$  is:

$$\begin{aligned} \frac{\partial\mathbf{D}(\boldsymbol{\theta})}{\partial\theta_k} &= \frac{\partial\mathbf{A}(\boldsymbol{\theta})}{\partial\theta_k}(\mathbf{A}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})^H + \mu\mathbf{I})^{-1}\mathbf{A}(\boldsymbol{\theta})^H \\ &\quad - \mathbf{A}(\boldsymbol{\theta})\mathbf{T}^{-1}(\boldsymbol{\theta})\frac{\partial\mathbf{T}(\boldsymbol{\theta})}{\partial\theta_k}\mathbf{T}^{-1}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})^H \\ &\quad + \mathbf{A}(\boldsymbol{\theta})\mathbf{T}^{-1}(\boldsymbol{\theta})\frac{\partial\mathbf{A}(\boldsymbol{\theta})^H}{\partial\theta_k} \quad , \end{aligned} \quad (\text{A.15})$$

and of  $\mathbf{B}(\boldsymbol{\theta})$  with respect to  $\theta_k$  is:

$$\begin{aligned} \frac{\partial\mathbf{B}(\boldsymbol{\theta})}{\partial\theta_k} &= -\mathbf{T}^{-1}(\boldsymbol{\theta})\frac{\partial\mathbf{T}(\boldsymbol{\theta})}{\partial\theta_k}\mathbf{T}^{-1}(\boldsymbol{\theta})\mathbf{A}(\boldsymbol{\theta})^H \\ &\quad + \mathbf{T}^{-1}(\boldsymbol{\theta})\frac{\partial\mathbf{A}(\boldsymbol{\theta})^H}{\partial\theta_k} \quad . \end{aligned} \quad (\text{A.16})$$

# References

- [1] Louis L Scharf. *Statistical Signal Processing*. Addison-Wesley, Boulder, 1991.
- [2] J.S. Mehta P.A.V.B. Swamy and P.N. Rappoport “Two Methods Of Evaluating Hoerl and Kennard’s Ridge Regression.,” *Commun. Statistics.-theory. Meth.*, vol. A7, pp. 1133–1155, 1978.
- [3] R. Fletcher. *Practical Methods of Optimization*. John Willey and Sons, Dundee, Scotland,UK, 1987.
- [4] Irishickesh D. Vinod “A Survey Of Ridge Regression and Related Techniques for Improvement Over Ordinary Least Squares,” *The Review Of Economics and Statistics*, pp. 121–131, 1977.
- [5] James A. Cadzow “Least Squares, Modelling, and Signal Processing.,” *Digital Signal Processing*, vol. 4, 1994.
- [6] Orhan Arikan “Regularized Solution of Linear System of Equations by Using Wavelets.,” in *Proc. IEEE Int. Symp. on Time-Frequency and Time-Scale Analysis*.
- [7] Ingrid Daubechies. *Ten Lectures On Wavelets*. Capital City Press, Montpelier, Vermont, 1992.

- [8] Elisabeth Gassiat “Déconvolution aveugle de systèmes linéaires discrêts bruités,” *Academie des Sciences*, vol. 319, pp. 489–492, 1994.
- [9] Carmen Sanchez-Avilla “An Adaptive Regularized Method for Deconvolution Of Signal With Edges by Convex Projections,” *IEEE Trans. Signal Process.*, vol. 42, pp. 1849–1851, 1994.
- [10] C. Gourieroux C. Fourgeaud and J. Pradel “Some Theoretical Results For Generalized Ridge REgression Estimators,” *Journal of Econometrics*, vol. 2, pp. 191–203, 1984.
- [11] Edwin T. Jaynes “On The Rationale Of Maximum-Entropy Methods,” *Proc. IEEE*, vol. 70, pp. 939–952, 1982.
- [12] J. Nunez and J. Llacer “A General Bayesian Image Reconstruction Algorithm With Entropy Prior. Preliminary Application to IIST Data,” *Astronomical Society Of the Pacific*, vol. 105, pp. 1192–1208, 1993.
- [13] Orhan Arikan “Regularized inversion of a two-dimensional integral equation with applications in borehole induction measurements,” *Radio Science*, vol. 29, pp. 519–538, 1994.
- [14] Sabine Van Huffel and Joos Vanderwalle. *The Total Least Squares Problem*. Society For Industrial and Applied Mathematics, SIAM,, Philadelphia, Pennsylvania, 1991.
- [15] Assem Deif. *Sensitivity Analysis in Linear Systems*. Springer Verlag, Berlin Heidelberg Germany., 1986.

- [16] George K. Chacko. *Decision-Making Under Uncertainty*. Praeger, New York, 1991.
- [17] John R. Taylor. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. Oxford University Press, Colorado, 1982.
- [18] Fehmi Chebil and Orhan Arikan “Robust Estimation Under Model Uncertainties,” in *Sinyal Isleme ve Uygulamalari Kurultayi Bildiriler Kitabi*, volume 2, p. 844, 1997.