NORTHEASTERN UNIVERSITY

Graduate School of Engineering

Thesis Title: Statistical Estimation with $1/f$-Type Prior Models.

Author: Roger Dufour.

Department: Electrical and Computer Engineering.

Approved for Thesis Requirements of the Master of Science Degree:

Professor Eric L. Miller, Advisor	Date
Professor Dana Brooks	Date
Professor Ram Raghavan	Date
Professor J.G. Proakis, Chairman	Date

Graduate School Notified of Acceptance:

Director, Graduate School	Date

NORTHEASTERN UNIVERSITY

Graduate School of Engineering

Thesis Title: Statistical Estimation with $1/f$-Type Prior Models.

Author: Roger Dufour.

Department: Electrical and Computer Engineering.

Approved for Thesis Requirements of the Master of Science Degree:

Professor Eric L. Miller, Advisor	Date
Professor Dana Brooks	Date
Professor Ram Raghavan	Date
Professor J.G. Proakis, Chairman	Date

Graduate School Notified of Acceptance:

Director, Graduate School	Date

Copy Deposited in Library:

Reference Librarian	Date

Statistical Estimation with $1/f$-Type Prior Models

A Thesis Presented

by

Roger Dufour

to

The Electrical and Computer Engineering Department

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in the field of

Electrical Engineering

Northeastern University

Boston, Massachusetts

May 1997

© Copyright 1997 by Roger M. Dufour, Jr.
All Rights Reserved

Abstract

A common problem in signal processing is estimating an object from noise corrupted data which gives an incomplete representation of the unknown. These problems are known as *inverse problems* and are found in many different applications such as geophysical prospecting, satellite and medical imaging, image reconstruction and target and anomaly detection. Because the data is limited and often of low quality, it is impossible to exactly reconstruct the object. We must instead try to construct algorithms for obtaining an accurate approximation of the object. Due to the possible ill-conditioning of the operator matrix, traditional solutions are often unstable. Therefore it is necessary to add a regularizing constraint to the solution.

In this thesis we will pursue a statistical approach to choosing a regularizer. We will assume that the object and noise can be described by stochastic models, specifically that they are zero mean Gaussian random vectors with certain covariance matrices. Since we often will not know the best model to use, it is necessary to estimate the model from statistics of the data. To perform this, we restrict ourselves to a parametric model where we can estimate the parameters of the model from the data. In this thesis we will be examining the $1/f$ fractal model defined in the wavelet transform domain.

This work specifically explores two issues with the use of the $1/f$ models: model mismatch and parameter estimation. We show that the performance is robust to fairly large mismatches between the model and the true process statistics. We also identify

conditions where the sensitivity with respect to a mismatched model is of concern. Finally, the parameter estimation is performed using the Expectation Maximization (EM) algorithm. In this thesis, we present a novel algorithm which takes advantage of the simplicity of the model to greatly reduce the computational burden.

Acknowledgments

Perhaps the greatest pleasure of finishing any work is thanking those who aided in its completion. Indeed, without a great many people this thesis could never have been realized. It would be fitting to thank each in turn, but I must bow to brevity or else this note would possibly outweigh the rest of the thesis.

Foremost, I must thank Professor Eric Miller without whose help and advice this thesis would not be possible. His knowledge and guidance were endless, as was his patience with the actual writing. I enjoyed working with Eric for the past three years and I look forward to working with him towards my Doctorate. I must also thank my committee members, Professors Dana Brooks and Ram Raghavan. They made many helpful suggestions in the final writing of this thesis and it is a better work because of them.

Though the trials of life as a graduate student do not measure up to the endless jokes about it, I must admit that my wits were often at an end as I tried to complete this work. Indeed, the question of my sanity is a popular topic of conversation among my friends. I will leave the verdict to them, but if I did maintain it then they deserve the credit for this. I would like to thank them all, those here in Boston and in Rhode Island and around the world.

Lastly, but with special regard, I would like to thank my family who have always supported me: my brother and sister, my grandparents, and especially my parents. My mother and father gave me the three greatest gifts in my life: a love of learning,

the opportunity to pursue it, and the encouragement to succeed. I could not possibly show my gratitude fully, but perhaps the dedication of this work to them is a start.

Roger Dufour, Jr.
June 5, 1997

Contents

List of Figures

Chapter 1

Introduction

A common problem in signal processing is estimating an object from noise corrupted data which gives an incomplete representation of the unknown. These problems are often known as *inverse problems* and are found in many different applications such as geophysical prospecting, satellite and medical imaging, image reconstruction and target and anomaly detection [1, 2, 3, 4]. Finding the input to a filter given the output, the structure of a radiating source given the fields, the original image given a blurred image and the internal structure given projections such as the Radon Transform are all examples of inverse problems. In these problems, we wish to reconstruct the object of interest from the information contained in the data, but the data elements form a relatively small set, usually one blurred and corrupted image of the object. Because the data is limited and often of low quality, it is impossible to exactly reconstruct the object. We must instead try to synthesize algorithms for obtaining an accurate approximation of the object.

When constructing an inverse solution, we often encounter problems with whether a solution exists and if the solution is unique. Due to noise in the data, it is often the case that no configuration of the object can exactly reproduce the data. Likewise if the distortion operator blocks aspects of the object from appearing in the data (the

operator has a null space), then the solution may not be unique since many objects can produce identical data. A third difficulty, and the one of primary interest in this work, is the stability of the solution to perturbations in the data. Traditional methods for guaranteeing existence and uniqueness, such as the pseudo-inverse solution, are not robust with respect to noise; small perturbations in the data can produce radically different solutions, usually with large high frequency oscillations. We desire our solutions to be stable with respect to small perturbations in the data such as the additive noise, but we often find that the inverse solution changes radically with minor amounts of noise. Therefore methods of regularizing, or stabilizing, the solution have been developed.

Regularization methods impose constraints upon the solution in order to guarantee existence, uniqueness and stability [5, 6, 7]. These constraints are usually in the form of *a priori* assumptions about the structure of the object. Common constraints are a small solution norm, a small derivative of some order, or a combination of several of these. A small solution norm bounds the total energy in the solution while derivatives impose smoothness constraints upon the solution. Since a smoother solution will contain less high frequency components, the constraints impose a lowpass filtering of the solution. The difficulty with using these regularization techniques is construction of the constraint, whether smoothness or some other constraint, and determining how strongly the constraint is applied, the level of regularization. The type of constraint is important in that it determines the type of object that the solution will reconstruct, i.e. whether it will be smooth, step-like, or some other shape. The level of regularization balances the solution between fidelity to the data and satisfying the constraint imposed. This not a trivial issue.

In this thesis we will pursue a statistical approach to choosing a regularizer. We

will assume that the object and noise can be described by stochastic models; specifically that they are zero mean Gaussian random vectors with certain covariance matrices. Instead of using an arbitrary *a priori* constraint in order to regularize the solutions, we will use the stochastic model as the regularizer. We will show in Chapter 2 that the use of the stochastic model in a linear least squares estimator (LLSE) is equivalent to commonly used Tikhonov regularization, and in addition provides the mean square optimal linear estimator if the model accurately describes the object.

In many cases we may not know the proper model. It is therefore necessary to estimate the model from statistics of the data. To perform this, we restrict ourselves to a parametric model where we can estimate the parameters of the model from statistics of the data. In this thesis we will be examining a particular two-parameter family of models which are defined in the wavelet transform domain.

There is much literature exploring certain advantages of operating in the wavelet domain, see [8, 9, 10, 11, 12, 13]. Wavelets are defined as an orthogonal or biorthogonal set of bases which are created by dilation and translation of a single function. This creates a basis which has properties of both scale and spatial localization. The basis set now allows a perspective on a signal where multi-resolution analysis can be used effectively. The wavelet transform has also been shown to sparsify many operator matrices leading to more efficient implementations. In our problem, the wavelet transform will allow us to specify the two parameter model which we will use, and in addition the properties of the wavelet transform will allow us to produce a fast algorithm for finding the parameters of the model.

The models which we will be using are the $1/f$ fractal family of statistical processes which have been developed in [1]. The $1/f$ fractal process displays characteristics of many real low pass processes. They have been used in ocean surface modeling [1] and modeling of Brownian motion. Further, it is shown in Wornell [14] that these processes can be defined extremely simply in the wavelet domain by a diagonal

covariance matrix, with only two parameters controlling the values of the non-zero elements. The use of these models in estimation has been explored for the problem of corrupted data, but not yet examined for more complex systems where the object has undergone a linear transformation. In this thesis, we will explore how these models perform under these more general conditions. We will show that the models perform well under a wide range of noise levels and under differing operators.

1.1 Contributions

As we discussed earlier, the $1/f$ models will be used in situations where the actual covariance matrix is unknown or where we choose not to use it. Thus it is necessary to understand how the models perform when the model is not matched to the true covariance. In Chapter 3, we explore the estimation performance in the case of model mismatch. We show that the performance is robust to fairly large mismatches between the model and the true process statistics. We also identify particular conditions under which performance is sensitive to model mismatch. When we examine the estimation of the models, it is shown that the model estimation performs well in the same situation when performance is most severely degraded by a mismatched model. Conversely, it is shown that in situations where model estimation is poor, model mismatch does not degrade the object estimation performance.

The second contribution of this thesis is a joint estimation algorithm which produces both the model parameters and an estimate of the object. This is performed using the Expectation Maximization (EM) algorithm. The EM algorithm is a two step iterative algorithm for estimation of the model parameters. A byproduct of the Expectation step is a current estimate of the object from the current parameter estimates. The object estimate will continue to improve at each iteration of the algorithm as the parameter estimates converge to the maximum likelihood values.

In this thesis, we will explore a formulation of the EM algorithm which allows for a particularly efficient implementation. By exploiting the properties of the model, it becomes possible to reduce the Maximization step from maximizing over a 2 dimensional function to finding a unique zero of a polynomial. Thus we are able to present a regularization scheme which produces as its output both the model and the reconstructed object, which is stable with respect to model mismatch and which can be efficiently implemented.

1.2 Organization

This thesis is organized into five chapters; the main body of the thesis is Chapters 2, 3 and 4. Chapter 2 is a discussion of the relevant background of inverse problems, regularization, wavelets and the $1/f$ models. Chapter 3 examines the issue of model mismatch. Chapter 4 presents the development of the EM algorithm and the performance bounds upon it, along with a discussion of the algorithm performance. Finally, the conclusions of this work are presented in Chapter 5.

The discussion in Chapter 2 is a brief presentation of the inverse problems and the difficulties encountered when solving them. A discussion of Tikhonov regularization is also presented with a discussion of how it effects the estimation. Next, we briefly discuss the necessary background of wavelets so that we can present the $1/f$ model, which is the last topic of the chapter.

Chapter 3 examines the model mismatch situation. In this chapter we present expressions for calculating the expected error encountered when the model does not match the true statistics of the object process. These expressions are then used to examine the sensitivity issues with respect to model mismatch. We explore the sensitivity across many levels of noise power and blurring function, and determine that the models are fairly robust. We specifically present those situations where

sensitivity may be an issue.

Chapter 4 examines the actual estimation of the model parameters. It begins with an examination of likelihood estimation and presents the necessary likelihood functions for maximum likelihood estimation. Next the Expectation Maximization algorithm is presented since this is the primary way in which the actual estimation is performed. We will present the derivation of the EM algorithm that simplifies the maximization step. In addition we will examine the theoretical bounds upon the estimator performance. Finally, we discuss how the estimation bounds relate to the models mismatch situation. Specifically we found that in situations where model mismatch is a concern, the model parameters can be estimated accurately, and conversely in situations where the model parameters cannot be estimated well, model mismatch is of little concern.

Lastly Chapter 5 will discuss the conclusions of this thesis and some possible ideas for continuation of the work.

Chapter 2

Regularization of Linear Inverse Problems

We are investigating the general linear inverse problem with stochastic models which describe both the object and the additive noise. In this chapter we will briefly examine continuous linear inverse problems and a discretization technique for creating a discrete problem, the solution of which will approximate the continuous solution. We will then examine difficulties associated with solving the discrete problem by using the singular value decomposition of the matrix. After examining these difficulties we will show how several classical regularization techniques can be used to overcome the difficulties. We will also show that the optimal regularizer for linear least squares estimation (LLSE) is to use the covariance matrix of the process as a model term in the LLSE cost function. For a more complete discussion of inverse problems, we direct the reader to [7, 5, 6].

Once we have established the usefulness of an accurate model for the covariance matrix in estimating the object, we will examine a class of statistical models which is defined in the wavelet domain and has been shown to be useful in estimation of many processes. The discussion here will give a brief introduction to the wavelet transform

and will focus on the aspects which make wavelets useful for the basis of the models. The models themselves are $1/f$ type fractal models. As mentioned they are defined in the wavelet transform domain since this allows for a diagonal covariance matrix and a natural scale-space interpretation of the model.

2.1 The Linear Inverse Problem

The general form of the forward problems of interest in this thesis is

$$f(x) = \int_a^b T(x,s)g(s)ds, \quad c \leq x \leq d. \tag{2.1}$$

Here we are given a function $g(s)$ and asked to find the solution $f(x)$. Examples of common forward problems are evaluating a voltage due to a charge distribution, a field due to a radiating source, and a filter response due to an input. The solutions to these problems can be accomplished by solving the integral analytically or by numerous numerical techniques to approximate the integration, such as Riemman summation. Numerical techniques would be necessary if our data were composed of discrete values, that is $\{g(s_n)\}$ where s_n are a collection of points in the interval from a to b, or if the integral was not solvable in closed form due to either the kernel $T(x,s)$ or to the form of $g(s)$.

More difficult to solve is the inverse problem. Here we are given the data as $f(x)$, and we wish to estimate $g(s)$. Analogous problems to those given above are solving for the charge distribution given measurements of the voltage, the source distribution given a field, or the excitation of a filter given a response. This problem may be complicated by the geometry of the problem, or by the sparsity of samples of the data function. We will see that when we have a finite number of samples we must constrain our solution to at most that many degrees of freedom. Thus it will be necessary to solve for a discrete representation of the object, either a sampled object or a projection of the object upon a finite number of basis functions.

2.1.1 Discretization: The Moment Method

Techniques for discretizing a continuous problem have been developed in order to simplify calculation and allow for numerical solutions which can be performed on a computer. An important advantage of the discretized problem is that numerical solutions are often possible in situations where the integral equation is not solvable in closed form. Further, since the measurement of the data in most problems is done discretely, by sampling either in time or space or both, a continuous representation of $f(x)$ is not available except by approximate methods. The approximation technique which we will explore here is the Moment Method (MM) [15, 16].

If we wish to discretize the integral equation in (2.1), we must first observe that we will no longer be working with continuous $g(x)$ and $f(x)$, and must in some way create discrete representations of these. We will first create a discrete $g(x)$ by expanding it as the sum of a finite family of basis functions $\{p_n(x)\}$ with coefficients g_n. This forms an approximation of $g(x)$ with the relation

$$\sum_{n=1}^{N} g_n p_n(x) \approx g(x), \tag{2.2}$$

with equality only if the basis functions span the space of $g(x)$.

The set of basis functions can take many forms. There are two major classes, sub-domain bases and entire-domain bases [15]. Sub-domain bases are those which are non-zero over only a small section of the interval of interest. Examples of these are the set of rectangular pulses, triangular functions, and truncated sinusoids. These bases can always be used without prior knowledge of $g(x)$. Entire domain functions cover the whole segment as the name implies. The Fourier Series, the moments of the object, or a set of orthogonal polynomials are examples of entire domain bases. To properly employ these bases, some knowledge of $g(x)$ may be necessary. For the Fourier Series it may require an unacceptably large number of terms in order to find

a good approximation if $g(x)$ has sharp edges, thus making the Fourier Series impractical. The choice of basis functions is important in describing $g(x)$ as completely as possible, but it can also lead to simpler computation of the integral as we will see shortly [15].

Assuming that the basis can describe $g(x)$ accurately, we will dispense with the approximation for convenience. Inserting our representation of $g(x)$ into the integral equation, we have

$$f(x) = \int_a^b T(x,s) \sum_{n=1}^{N} g_n p_n(s) ds. \tag{2.3}$$

Exchanging the summation and integral gives

$$f(x) = \sum_{n=1}^{N} g_n \int_a^b T(x,s) p_n(s) ds. \tag{2.4}$$

We now have, for a fixed x, a projection of $T(x,s)$ upon the basis functions. Given that $f(x)$ is still continuous we still do not have a discrete problem. We must discretize $f(x)$. In general, we can project $f(x)$ upon a basis of linearly independent functions, $\{w_m(x)\}$ as was done with $g(s)$. In this case $\{w_m(x)\}$ are the weighting functions. For problems with discrete observations, the weighting functions become $w_m(x) = \delta(x - x_m)$ where x_m is the m-th point of observation. The weighting functions $\{w_m(x)\}$ are taken in an inner product with $f(x)$ to form the discrete coefficients f_m as

$$f_m = \langle f(x), w_m(x) \rangle = \int_c^d f(x) w_m(x) dx. \tag{2.5}$$

The coefficients f_m are thus weighted averages of $f(x)$. In some situations weighting functions other than impulses may be practical and necessary. With impulse weighting functions, the behavior of the solution between sample points may be highly ill-behaved. To alleviate this, other weighting functions can be used which constrain the forward behavior to be better.

Now taking the inner product of both sides of (2.4), we have

$$f_m = \sum_{n=1}^{N} g_n \int_c^d \int_a^b T(x,s) p_n(s) w_m(x) ds dx. \tag{2.6}$$

We can see that the weighting functions have complicated the problem by introducing another integration. For this reason, it may be warranted to eliminate the second integral by evaluating only at specific x_m, that is choose $w_m(x) = \delta(x - x_m)$.

We have a representation of $T(x, s)$ which is discrete in both dimensions given as

$$T_{m,n} = \int_c^d \int_a^b T(x, s) p_n(s) w_m(x) ds dx. \tag{2.7}$$

The discrete equation now takes the form

$$f_m = \sum_{n=1}^{N} T_{m,n} g_n. \tag{2.8}$$

For a number of observations, $1 \leq m \leq M$, this equation becomes the matrix equation

$$f = Tg \tag{2.9}$$

where f and g are now vectors of coefficients and T is the matrix which is a discrete approximation to $T(x, s)$.

We are now ready to evaluate the integrals. If $T(x, s)$ is such that a certain choice of basis functions, $\{p_n(x)\}$, allows for a closed form solution, then the integral can be evaluated exactly. Commonly, the integral is approximated. If the set of rectangular pulses was chosen for both the basis and the weights, then the approximation may take the form of the well known Riemmen summation by evaluating $T(x, s)$ at the midpoint of the pulse. Let us apply this discretization technique to the example which is commonly used throughout this thesis.

The problem which is used extensively in this thesis is convolution with a Gaussian kernel. In this case, the kernel of the integral equation is

$$T(x, s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x - s)^2}{2\sigma^2}\right). \tag{2.10}$$

We will discretize this function using delta functions for the weights upon $f(x)$ and rectangular blocks upon $g(s)$ with the support of $p_n(s)$ being s_{n-1} to s_n where $a =$

$s_0 < s_1 < \cdots < s_n = b$. The matrix T is now

$$T_{m,n} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{s_{n-1}}^{s_n} \exp\left(\frac{(x_m - s)^2}{2\sigma^2}\right) ds, \tag{2.11}$$

which is the common error function and can be evaluated explicitly.

2.1.2 Ill-Conditioning

The discrete linear equation corresponding to (2.1), is given in (2.9). In this section, we will be examining how the geometry of the matrix T affects the solution to the problem. In an ideal situation with an invertible matrix T and noise free observations, the solution to the discrete inverse problem would be exact regardless of the matrix T. As noise is unavoidable, we will introduce another term into the equation to represent the noise in the data. The vector n will represent additive noise in the data so the observations take the form,

$$f = Tg + n. \tag{2.12}$$

The problem is to estimate the underlying object g given the data, f. For this we need some way in which to invert the matrix, T. The estimate of g will be signified as $\hat{g}$. The simplest way to estimate g is to use the inverse matrix of T, but this is not possible if T is not square or is less than full rank. In these situations, the pseudoinverse can be defined. In order to introduce the pseudoinverse, we shall define the singular values decomposition (SVD) of the matrix.

The SVD of the matrix T is [17]

$$T = U\Sigma V' \text{ where } \Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \tag{2.13}$$

Here the matrices U and V are the singular vectors and form orthonormal bases for the row and column spaces respectively. The matrix Σ_1 is a diagonal matrix of the

singular values. The matrix Σ is filled with zeros so that its dimensions match those of the original matrix.

Using the SVD, we can regard the operation of the matrix as an orthonormal projection of the vector onto a generalized Fourier space, a filtering operation performed by the Σ matrix and a projection into the data space. To construct the inverse, our projections will remain and the filtering operation will be inverted. Further, to avoid complications created by the null space of T we only invert Σ_1. Thus we can define the pseudoinverse using the SVD as

$$T^\dagger = V\Sigma^\dagger U' \quad \text{where} \quad \Sigma^\dagger = \begin{bmatrix} \Sigma_1^{-1} & 0 \\ 0 & 0 \end{bmatrix}. \tag{2.14}$$

It is now seen that the singular values of the new matrix are the inverse of the singular values of T, except that any values which were zero will remain zero.

The pseudoinverse solves any problems related to existence or uniqueness of the solution, but if we examine the singular values we shall see that stability is still in question. The norm of a matrix bounds the amplification that it can have upon any vector. It is therefore equal to the largest singular value. We now define the condition number as the ratio of the relative change in the object given a relative change in the data. We use δf as a perturbation in f and δg as a perturbation in the solution g, and derive

$$\begin{aligned} K(T) &= \max_{f,\delta f} \frac{\|\delta g\|/\|g\|}{\|\delta f\|/\|f\|} && (2.15)\\ &= \max_{f,\delta f} \frac{\|\delta g\|/\|g\|}{\|T\delta g\|/\|Tg\|} && (2.16)\\ &= \max_{f,\delta f} \frac{\|\delta g\|}{\|T\delta g\|}\frac{\|Tg\|}{\|g\|} && (2.17)\\ &= \frac{\|T\|}{\|T^\dagger\|} = \frac{\sigma_{\max}}{\sigma_{\min}}. && (2.18) \end{aligned}$$

Thus the conditioning is the ratio of the largest singular value to the smallest. It can now be seen that a small condition number relates to a stable problem, since

small perturbations in the data result in only small changes in the solution. A high condition number will instead result in large changes to the solution.

We now apply the pseudoinverse to the problem at hand and expand f, as $Tg+n$.

$$\hat{g} = T^{\dagger}f \tag{2.19}$$

$$= T^{\dagger}Tg + T^{\dagger}n \tag{2.20}$$

Next, we expand g into g_r and g_n, where g_r lies in the range space of the operator T and g_n lies in the null space of T. Therefore,

$$\hat{g} = T^{\dagger}Tg_r + T^{\dagger}Tg_n + T^{\dagger}n \tag{2.21}$$

$$= T^{\dagger}Tg_r + T^{\dagger}n \tag{2.22}$$

$$= g_r + T^{\dagger}n \tag{2.23}$$

Thus pseudoinverse matrix recovers the component of the object which lies in the range space of T and it leaves as 0 that which lies in the null space of T. But the perturbations which were in the data, n, have also been affected by the pseudoinverse matrix. Given the condition number, we can calculate the upper bound upon the change in the estimate given a perturbation in the data and we will see that the maximum change in g can be very large given a relatively small perturbation in f. The change is bounded by the condition number as

$$\frac{\|\delta g\|}{\|g\|} \leq K(T)\frac{\|\delta f\|}{\|f\|}. \tag{2.24}$$

If we consider the perturbation upon f as noise in the data then if the condition number of the matrix T is large so is the upper bound on the noise amplification.

If we now consider the vector spaces associated with the large noise amplification, we see that they are associated with the small singular values in T. Thus we are unlikely to have much object information in these spaces in our data since it would be attenuated in the forward problem. If we consider a low pass filter problem, we see

that the small singular values would be associated with the high frequency singular vectors. Thus in the forward solution a low pass or broadband object would produce data which is primarily low pass. Only in the case of a high pass object would this not be the case. If we add noise to our low pass data, then we will have broadband data. When we produce the solution using the pseudoinverse the high frequencies in the noise will be amplified creating a solution which is primarily composed of high frequency noise. This solution will obviously be worthless and we must therefore seek methods to suppress this behavior of the pseudoinverse. Tikhonov regularization is one such method and will be the method of choice in this thesis since it will be shown to be equivalent to the Linear Least Squares Estimator (LLSE).

2.2 Tikhonov Regularization

Tikhonov regularization [5] uses a cost function which balances fidelity to the data with some norm that imposes an *a priori* constraint upon the solution such as small total energy or smoothness. The constraint also serves to eliminate the instability in the solution by lowering the condition number of the operator. The cost function for this estimation can be written as

$$\hat{g} = \arg\min_{g}(\|Tg - f\|^2 + \lambda^2\|Lg\|^2) \tag{2.25}$$

Here we have the least squares estimator with a cost term added which will serve as the regularizer. The cost term has two factors, a regularization matrix, L, which imposes a constraint upon $\hat{g}$, and the regularization factor, λ, which determines the level of influence the two terms have upon the minimization. Usual choices for L are the identity matrix, which constrains total energy, or a derivative operator, which imposes smoothness. Combining the two terms and solving the normal equations

produces a closed form solution as

$$\hat{g} = \arg\min_{g} \| \begin{bmatrix} T \\ \lambda L \end{bmatrix} g - \begin{bmatrix} f \\ 0 \end{bmatrix} \|^2 \tag{2.26}$$

$$\hat{g} = (T'T + \lambda^2 L'L)^{-1} T'f = Mf \tag{2.27}$$

We can now see the effect of regularization on the condition number by examining the singular values of the estimation matrix M. We will use a regularization matrix $L = I$ and the properties of the SVD that $U'U = I$, $VV' = I$, $U^{-1} = U'$, $V^{-1} = V'$. If we expand T with the SVD given in the previous section, we obtain

$$\begin{aligned} M &= (V\Sigma U' U \Sigma V' + \lambda^2 I)^{-1} V \Sigma U' && (2.28) \\ &= (V\Sigma^2 V' + \lambda^2 VV')^{-1} V \Sigma U' && (2.29) \\ &= V(\Sigma^2 + \lambda^2 I)^{-1} V' V \Sigma U' && (2.30) \\ &= V\left((\Sigma^2 + \lambda^2 I)^{-1}\Sigma\right) U' && (2.31) \end{aligned}$$

Thus the SVD of the matrix M, is now in the same form as that of $T^\dagger$, except that the singular values of M are given by

$$\mu_i = \frac{\sigma_i}{\sigma_i^2 + \lambda^2}, \tag{2.32}$$

whereas the singular values of $T^\dagger$ were given by $\frac{1}{\sigma_i}$. It can now be seen that if a singular value was large in the matrix T, then in both the pseudoinverse and the regularized inverse the corresponding singular value will be small. But if the singular value was very small, in the pseudoinverse it would have been very large, whereas in the regularized inverse the regularization parameter will constrain it. In the regularized inverse the largest possible singular value occurs when $\sigma_i = \lambda$ and therefore the singular values will be bounded by

$$\mu_i \leq \frac{1}{2\lambda}. \tag{2.33}$$

Other cases of the regularization matrix produce similar results, but the singular vector spaces U and V are not identical between both the pseudoinverse and regularized inverse and consequently do not lead to simple expressions. They will, however, produce a regularization provided that T and L do not have a common null space.

There are difficulties with implementing this technique. First, a regularization matrix must be chosen which will adequately handle the condition of the problem. Second, the level of regularization must be set so as to apply the proper amount of regularization to stabilize the estimation without over-regularizing and producing a trivial estimation.

The regularization matrix must be chosen so as to constrain those aspects of the estimation which are unwanted and primarily due to the effects of the high condition number. A common choice is to use the identity matrix, which will apply an equal amount of constraint across all singular values. This will lower the condition number effectively, but will also suppress aspects of the signal which should desirably be allowed. Another choice is a discrete derivative operator. This will also lower the condition number while penalizing against high frequencies in the solution.

2.3 The Optimal Linear Estimator

In this thesis we are concerned with the estimation of an object which can be modeled as a Gaussian random process in additive Gaussian noise. In this section we will show that there is a mean square optimal linear estimator provided that the first and second order statistics of the process are known, and that this estimator is a Tikhonov regularization with the regularization provided by a stochastic model composed of these statistics. This will motivate the subsequent chapters' estimation techniques which attempt to match a model matrix as closely as possible to the true covariance matrix and to use this matrix as the regularization term. Therefore, let us define g

as a Gaussian random process with zero mean and covariance matrix P_0, or

$$g \sim N(0, P_0), \tag{2.34}$$

and the noise statistics as

$$n \sim N(0, R). \tag{2.35}$$

Also, let us use some linear estimator for g, $\hat{g} = My = MTg + Mn$, and compute the mean square error of this estimator as

$$\begin{aligned}
\epsilon^2 &= E\{(g-\hat{g})'(g-\hat{g})\} && (2.36)\\
&= E\left\{\mathrm{tr}\{(g-\hat{g})(g-\hat{g})'\}\right\} && (2.37)\\
&= \mathrm{tr}\left\{E\{(g-\hat{g})(g-\hat{g})'\}\right\} && (2.38)\\
&= \mathrm{tr}\left\{E\{gg' - g\hat{g}' - \hat{g}g' + \hat{g}\hat{g}'\}\right\} && (2.39)
\end{aligned}$$

Now we will expand our estimator $\hat{g}$ and then take the expectation. The expectations are $E\{gg'\} = P_0$, $E\{nn'\} = R$, and $E\{gn'\} = 0$.

$$\begin{aligned}
\epsilon^2 &= \mathrm{tr}\left\{E\{gg'\} - 2E\{g(MTg+Mn)'\}\right. && (2.40)\\
&\quad \left. + E\{(MTg+Mn)(MTg+Mn)'\}\right\} && (2.41)\\
&= \mathrm{tr}\left\{E\{gg'\} - 2E\{gg'\}T'M' + ME\{gn'\}\right. && (2.42)\\
&\quad \left. + MTE\{gg'\}T'M' + 2MTE\{gn'\} + ME\{nn'\}M'\right\} && (2.43)\\
&= \mathrm{tr}\{P_0 + MTP_0T'M' + MRM' - 2P_0T'M'\} && (2.44)\\
&= \mathrm{tr}\{P_0 + M'M(TP_0T' + R) - 2M'P_0T'\} && (2.45)
\end{aligned}$$

We now wish to minimize this across the estimator M. This is a parabolic function in M, and we can find the minimum by finding the zero of the first derivative. Therefore the "optimal" estimator, $\check{M}$, is

$$\check{M} = \arg\min_M(\mathrm{tr}\{P_0 + (TP_0T' + R)M'M - 2P_0T'M'\}) \tag{2.46}$$

$$\frac{\partial \epsilon^2}{\partial m_{i,j}} = \mathrm{tr}\{\frac{\partial}{\partial m_{i,j}}(P_0 + (TP_0T' + R)M'M - 2P_0T'M')\} \quad (2.47)$$

$$= \mathrm{tr}\{\frac{\partial}{\partial m_{i,j}}P_0 + (TP_0T' + R)\frac{\partial}{\partial m_{i,j}}M'M - 2P_0T'\frac{\partial}{\partial m_{i,j}}M'\} \quad (2.48)$$

We define $\Delta_{i,j}$ to be the derivative of M with respect to the element $m_{i,j}$. This is a matrix of zero with one in the ith row, jth column. The relation $\Delta'_{i,j} = \Delta_{j,i}$ is easily verifiable. If we continue with the derivation we have

$$\frac{\partial \epsilon^2}{\partial m_{i,j}} = \mathrm{tr}\{(TP_0T' + R)(\Delta'_{i,j}M + M'\Delta_{i,j}) - 2P_0T'\Delta'_{i,j}\} \quad (2.49)$$

$$= \mathrm{tr}\{(TP_0T' + R)(\Delta_{j,i}M + M'\Delta_{i,j}) - 2P_0T'\Delta_{j,i}\} \quad (2.50)$$

$$= \mathrm{tr}\{(M(TP_0T' + R) - 2P_0T')\Delta_{j,i}\} + \mathrm{tr}\{(TP_0T' + R)M'\Delta_{i,j}\} \quad (2.51)$$

The matrix $\Delta_{i,j}$ extracts only the ith column of the matrix being multiplied and leaves it in the jth column of an otherwise all zero matrix. The trace then sums the diagonal of the matrix. The result from these two operations is that the equation becomes a sum of individual elements of the matrices as

$$\frac{\partial \epsilon^2}{\partial m_{i,j}} = [M(TP_0T' + R) - 2P_0T']_{i,j} + [(TP_0T' + R)M']_{j,i} \quad (2.52)$$

$$= [M(TP_0T' + R) - 2P_0T']_{i,j} + [M(TP_0T' + R)]_{i,j} \quad (2.53)$$

$$(2.54)$$

Now setting this equal to zero we can solve for all i, j by using the matrix equation

$$M(TP_0T' + R) - 2P_0T' + M(T'P_0T + R) = \underline{0} \quad (2.55)$$

$$M(TP_0T' + R) = P_0T' \quad (2.56)$$

$$M = P_0T'(TP_0T' + R)^{-1} \quad (2.57)$$

This can be rearranged by using the following result of the partitioned matrix lemma [18]

$$(A - BD^{-1}C)^{-1}B = -A^{-1}B(D - CA^{-1}B)^{-1}D, \quad (2.58)$$

with the matrices defined as follow:

$$A = P_0^{-1} \tag{2.59}$$

$$B = T'R^{-1} \tag{2.60}$$

$$C = R^{-1}T \tag{2.61}$$

$$D = -R^{-1}. \tag{2.62}$$

Substituting, we obtain

$$M = (T'R^{-1}T + P_0^{-1})^{-1}T'R^{-1}, \tag{2.63}$$

which when compared to Tikhonov regularization shows that the optimal linear regularizer is the inverse of the covariance matrix of g.

2.4 The Wavelet Transform

The wavelet transform has received much attention in the last few years for its application to digital signal processing problems. The wavelet transform is particularly useful since it retains temporal or spatial information while at the same time providing a division of the frequency content. It is able to do this by employing compactly supported basis functions instead of the infinite sinusoidal functions of traditional Fourier analysis. The basis functions for the wavelet transform are also not constrained to have the same length, as they are with the windowed bases of the Short Time Fourier Transform. For fine scale detail, the functions are narrow, giving excellent resolution of short transients in the signal, while for coarse scale structure, they are much longer but have more localized frequency characteristics. Viewed as a division of the frequency domain they are a constant-Q filter bank [19].

The ability of the wavelet transform to localize transient information and resolve frequency content has made it quite useful in the fields of signal and image compression, transient detection, signal smoothing, and numerical signal processing. Indeed,

in many applications of image compression, the wavelet transform performs significantly better than the discrete cosine transform [20]. Its natural tree structure enables transient detection algorithms to use multi-scale search techniques which lower computational cost. In numerical processing, it has been shown [9] that many matrices are naturally sparsified in the wavelet domain and matrix multiplication and inversion can be performed at a much reduced computational cost.

It is helpful to compare how the time frequency plane is divided by the wavelet transform, the Fourier Transform (FT) and the short time Fourier Transform (STFT). Figure 2.1 shows this division for a signal and the three transforms. The first plot (a) is the time representation of a signal, giving complete time resolution, but no frequency resolution. The Fourier Transform which is given by

$$F(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-j\omega t} dt \tag{2.64}$$

is a projection of the signal upon the family of complex sinusoids. It is evident that the complex sinusoids basis is infinite in time, but of a single frequency. Therefore the FT gives exact frequency resolution, but no time resolution. The plot in (b) shows this. Third we have the Short Time Fourier Transform given by

$$F(\omega, \tau) = \int_{-\infty}^{+\infty} f(t) w(t-\tau) e^{-j\omega t} dt. \tag{2.65}$$

Here a windowing function $w(t-\tau)$ upon the Fourier basis functions limits the support to a finite time, allowing for some time resolution controlled by a shift parameter τ. Of course multiplying in the time domain creates a convolution in the frequency domain, blurring the frequency resolution. The limit upon resolution of the time and frequency bins is given by

$$\Delta t \Delta \omega \geq \frac{1}{2} \tag{2.66}$$

where Δt and $\Delta \omega$ are the size of the time-frequency bin [21]. However, it is evident that the frequency bins are of equal support across the entire frequency axis, and

likewise for the time axis. It is however logical that one would want better time resolution with respect to fast changing i.e. high frequency, components of the signal and better frequency resolution with low frequency components. The wavelet transform achieves this by dividing the plane into different sized bins. This gives better resolution in time for high frequency components, and better frequency resolution for low frequency components.

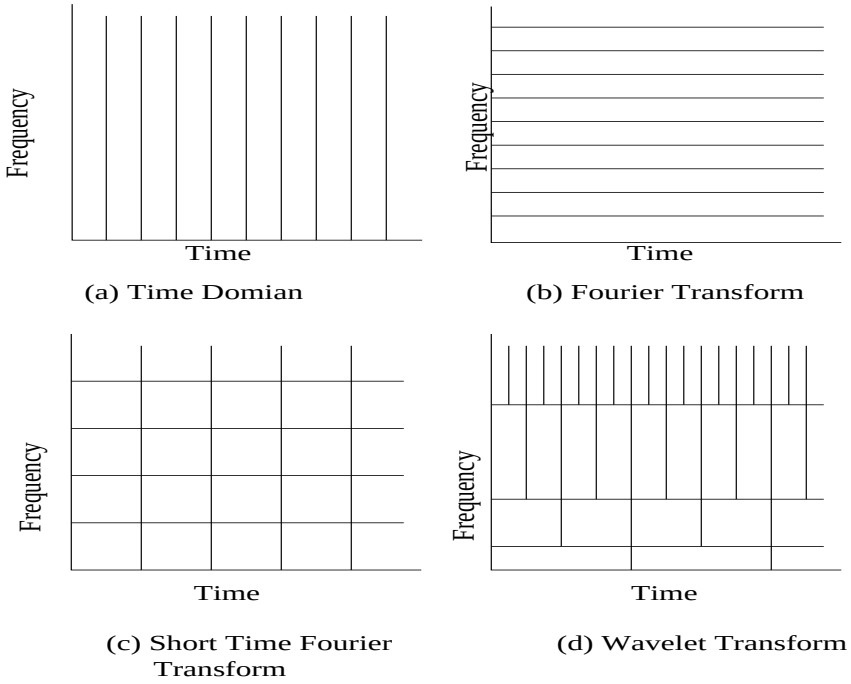

Figure 2.1: The division of the time and frequency domains for several transforms. (a) Time Domain (i.e. no transform) (b) Fourier Transform (c) Short Time Fourier Transform (d) Wavelet Transform.

To develop the wavelet transform, we begin with the complete signal space V_0. We will create a smoother space V_1 which is an approximation to the complete signal space. Signals in the space V_1 are approximations of those in V_0, lacking the fine scale structure which we have removed, therefore we will call these spaces approximation spaces. Continuing this iteration upon each V_j, we have a set of nested approximation

spaces to V_0 with the relationship

$$V_J \subset \cdots \subset V_1 \subset V_0. \tag{2.67}$$

Figure 2.2 shows how the approximation spaces are related.

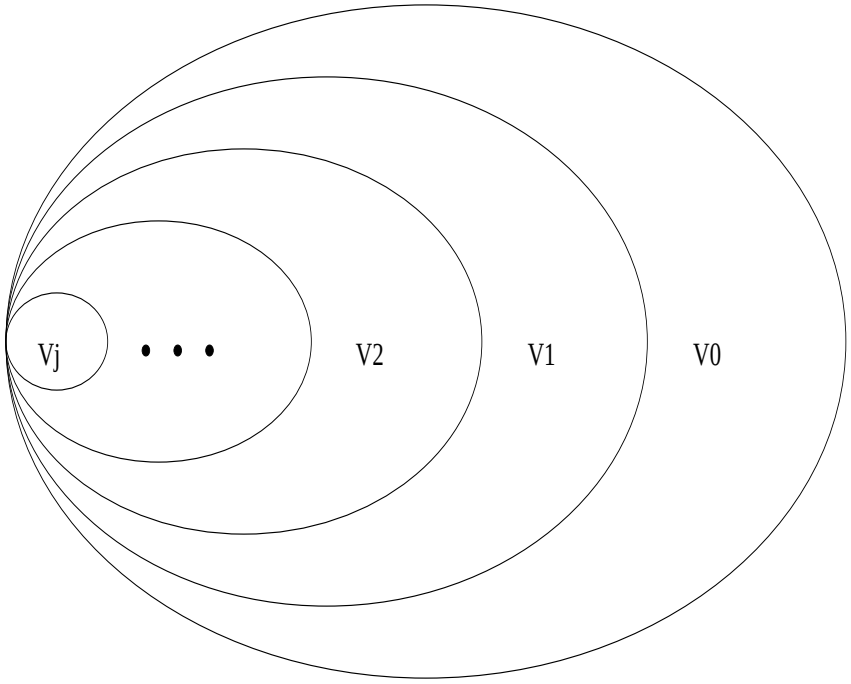

Figure 2.2: The approximation spaces V_j. Each is an embedded subspace of the space above. The detail space W_j is the orthogonal compliment of W_j in V_{j-1}.

Given that each of the V_j are a subspace of the space above, we can now construct a space for the detail which was removed as the orthogonal compliment to the approximation space. We can divide each V_j into the sum of two spaces as

$$V_j = V_{j+1} \oplus W_{j+1}. \tag{2.68}$$

Where the operation $\oplus$ denotes direct addition, and thus the space W_{j+1} is the compliment space to V_{j+1} in the space V_j. It is obvious that each new space W_j is orthogonal to V_j, and since each W_j exists in a different V_{j-1}, all the W_j are

orthogonal. Therefore, if we take the set of spaces W_j and the coarsest approximation space V_J, we will have a set of orthogonal spaces which sum to the space V_0, as in

$$V_J \oplus W_J \oplus \cdots \oplus W_1 \oplus W_0 = V_0. \tag{2.69}$$

The implementation of this breakdown for a discrete signal can be easily seen as a

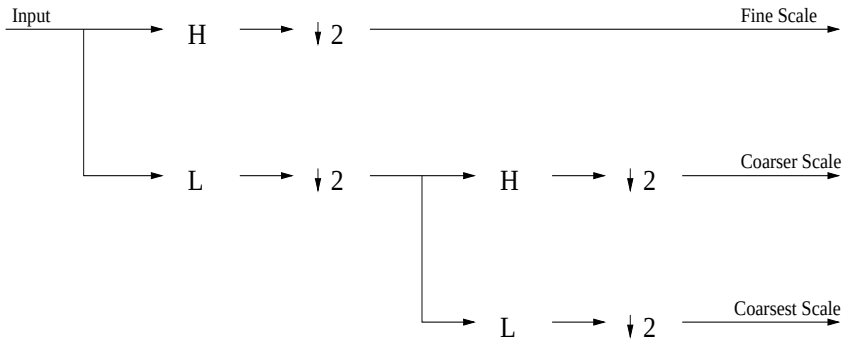

Figure 2.3: Filter bank implementation of the wavelet transform.

filter bank such as Figure 2.3. Here we have two stages of a wavelet decomposition; more levels can be taken by decomposing the coarsest scale at each stage. Thus the fine scale coefficients correspond to the W_0, the next coarser to W_1 and the coarsest to V_1. If more stages are added, this corresponds to the division of the V_j space into a corresponding W_j and V_{j+1}.

Each stage of the filter is performed by specifying a pair of FIR filters, one high pass and one low pass. For an orthogonal wavelet decomposition these filters form a power symmetric pair [12]. Let us specify the filter coefficients at each stage by $h(k)$ for the low pass and $g(k)$ for the highpass. Then we have the relationship

$$g(k) = (-1)^k h(K-k). \tag{2.70}$$

This forms a perfect reconstruction pair, and we can down sample the outputs by 2 without any loss of information. Thus the output of the filter banks has an equal

number of elements as the input. And if we stack the elements into a new vector of the wavelet coefficients, we can represent the action of the filters by a matrix multiplication. If we use W_y as the wavelet decomposition of the data vector, then the vector of wavelet coefficients η is

$$\eta = W_y y. \tag{2.71}$$

Since we are using orthogonal perfect reconstruction filter banks, the transpose matrix applied to the wavelet coefficients will reconstruct the original vector. Thus the matrix W_y is orthonormal and $W_y^T = W_y^{-1}$.

We will also introduce the matrix W_g which is the corresponding wavelet transform for the object space. The transformation of the problem in (2.12) to the wavelet domain can be viewed as the operation

$$W_y y = (W_y T W_g^T) W_g g + W_y n \tag{2.72}$$

$$\eta = \Theta\gamma + \nu. \tag{2.73}$$

We now have each of the vectors represented by a vector of its wavelet coefficients. We will use the corresponding Greek characters η, Θ, and ν to represent the wavelet transforms of y,T and n. The transforms of the covariance matrices are

$$P_\gamma = W_g P_g W_g^T \tag{2.74}$$

$$R_\nu = W_y R_n W_y^T. \tag{2.75}$$

Thus we have the problem of estimating the wavelet coefficients γ from the wavelet coefficients of the data η. Using the results of the previous section, the optimal linear estimator would be the LLSE using as a model matrix the covariance matrix of γ, P_γ. Since in this thesis we are assuming that we do not know this matrix, we will instead in the next section present a useful model which is specified in the wavelet transform domain and has been shown in [1] to accurately describe many real world processes.

2.5 The $1/f$ Model

The $1/f$ model defined in the wavelet domain will be a model for the process covariance. As shown earlier, the optimal linear estimator is the LLSE estimator with the covariance matrix as the regularization term. Since we will often not have the actual covariance or it may be too complicated to use, it is important to have a simple model which in practice will adequately approximate the actual covariance, thereby leading to a low MSE in the solution. In keeping with this idea, we adopt a simple model with only two parameters, defined as

$$[P_\gamma]_{i,j} = \begin{cases} \kappa 2^{-\alpha s} & i = j \\ 0 & i \neq j \end{cases} \tag{2.76}$$

Where s is a non-negative integer representing the level of the wavelet coefficient, the coarsest level being zero and the finest being $S-1$, S being the number of levels. The coarsest level is the approximation space, which for our purposes is assigned a covariance value of κ.

The parameter α is the fractal parameter and controls how quickly the spectrum drops off. A low α is a wide-band process, with zero being white noise, and a high α is a very highly correlated process. For many formulae in this section, we use only the normalized matrix designated as

$$[F(\alpha)]_{i,j} = \begin{cases} 2^{-\alpha s} & i = j \\ 0 & i \neq j \end{cases} \tag{2.77}$$

The parameter κ determines the overall energy of the process and is equivalent to the λ^2 of the classic regularizers while $F(\alpha)$ is equivalent to $L'L$.

It will be useful in our estimations to constrain the model so that the overall energy remains constant while the model changes. Therefore we often wish to use the process energy e_γ instead of κ. In addition to allowing us to use constant energy models, it will be easier to estimate the energy than the regularization parameter. Finally, given

an energy, it is possible to bound κ above and below to simplify the estimation of the model. In other words, with the energy, we can preset the regularization level while we search for a good α.

Given that the energy of the noise is known, it can be shown that the maximum likelihood estimate for the energy of the process is

$$\begin{aligned} e_\eta &= \|\eta\|^2 && (2.78)\\ &= \|\Theta\gamma + \nu\|^2 && (2.79)\\ &= \mathrm{tr}\{\Theta P_0 \Theta' + R\} && (2.80)\\ &= \kappa \mathrm{tr}\{\Theta F(\alpha)\Theta'\} + \mathrm{tr}\{R\} && (2.81)\\ &= \kappa \mathrm{tr}\{\Theta F(\alpha)\Theta'\} + e_\nu && (2.82) \end{aligned}$$

Solving this for κ gives

$$\kappa(\alpha) = \frac{e_\eta - e_\nu}{\mathrm{tr}\{\Theta F(\alpha)\Theta'\}} \qquad (2.83)$$

This can be shown to be a bounded monotonically increasing function in α. The asymptotic lower and upper bounds can be determined by calculating $\kappa(\alpha)$ at $\alpha = 0$ and as $\alpha \to \infty$. For this, we use the relations $F(0) = I$ and $\lim_{\alpha\to\infty} F(\alpha) = I_c$, where I_c is a zero matrix except the upper left hand corner is identity only for the coarsest scale coefficients. The bounds then are

$$\begin{aligned} \kappa_{\min} &= \frac{e_\eta - e_\nu}{\mathrm{tr}\{\Theta\Theta'\}} && (2.84)\\ \kappa_{\max} &= \frac{e_\eta - e_\nu}{\mathrm{tr}\{\Theta I_c \Theta'\}} && (2.85) \end{aligned}$$

The function of κ versus α is shown in Figure 2.4. The dotted lines show the upper and lower bounds upon this function.

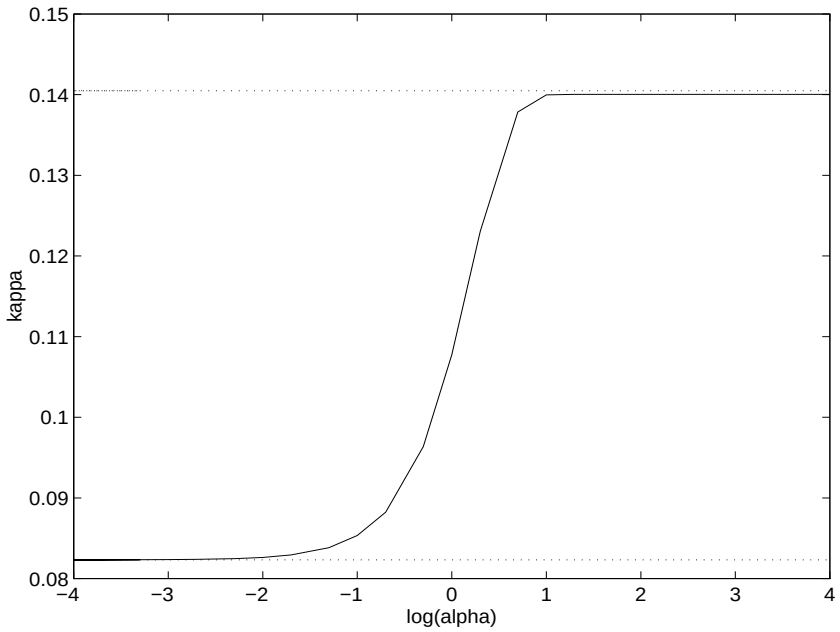

Figure 2.4: The κ-Function versus α for a fixed energy with upper and lower bounds.

Chapter 3

Model Mismatch

Since it is often the case that we do not have the true covariance matrix, we often will be estimating it based upon a parametric model such as the $1/f$ model presented in the previous chapter. Since the estimate will always have a certain amount of error, an important subject to examine regarding our models is model mismatch. Here we are concerned with the performance of the LLSE estimator when the parameters of the model do not match those of the true γ. It will be seen that the models perform well within a wide range of operator specifications and noise conditions. In the most severe case of model mismatch, we will also examine the estimator performance when the actual γ is not a $1/f$ process, but a first order Gauss Markov (FOGM) process which possesses an entirely different covariance matrix than our model. Specifically, whereas in the wavelet domain the models are diagonal matrices, the covariance matrix of the FOGM process is not diagonal, but possesses strong off diagonal elements. We will show that the models still perform well and even when highly mismatched they still act as a classical regularizer to produce a smooth and stable solution.

3.1 MSE Performance

To measure the performance we will use the mean square error in the solution as a function of the true covariance P_0 and the model P_γ

$$\mathrm{MSE}(P_\gamma; P_0) = E\{|\hat{\gamma} - \gamma|\} = \mathrm{tr}\left\{E\{(\hat{\gamma} - \gamma)(\hat{\gamma} - \gamma)'\}\right\}. \tag{3.1}$$

This is a function of the object and model covariance matrices, P_0 and P_γ respectively, the noise covariance matrix R and the matrix Θ, as

$$\mathrm{MSE}(P_\gamma; P_0) = \mathrm{tr}\left\{(\Theta' R^{-1}\Theta + P_\gamma^{-1})^{-1}(\Theta' R^{-1}\Theta + P_\gamma^{-1} P_0 P_\gamma^{-1})(\Theta' R^{-1}\Theta + P_\gamma^{-1})^{-1}\right\}. \tag{3.2}$$

Since it is desirable to describe how the models perform with respect to the optimal mean square error, we define MSE_{opt} as the optimal performance which can be achieved with a linear estimator. As was shown in section 2.5, this is achieved with a LLSE where $P_\gamma = P_0$. The optimal MSE therefore is given by

$$\mathrm{MSE}_{opt} = \mathrm{MSE}(P_0; P_0) = \mathrm{tr}\{(\Theta R^{-1}\Theta + P_0^{-1})^{-1}\}. \tag{3.3}$$

We define the normalized mean square error (NMSE) to be the ratio

$$\mathrm{NMSE}(P_\gamma; P_0) = \frac{\mathrm{MSE}(P_\gamma; P_0)}{\mathrm{MSE}_{opt}}. \tag{3.4}$$

The optimal NMSE is therefore equal to 1, and degradation of performance occurs for values above 1. For example a NMSE of 1.1 is a 10% degradation below the optimal performance of the estimator.

In the following examples, we will examine the performance degradation of the model in the mean square error with respect to model mismatch, under different situations. We will show that the performance is relatively insensitive to mismatch in the model under a wide range of operator specifications and noise conditions. For the first several experiments, we will be using a covariance matrix for the object which

is $1/f$. For these, the optimal performance is achieved when the parameters of the model equal the parameters of the covariance matrix of the object. For the remainder of the examples, we will be estimating a first order Gauss Markov (FOGM) process, for which any $1/f$ model could only be an approximation, and therefore optimal performance will not be achievable, but we will show that even in this severe case of model mismatch, the models still perform well in their estimates. In all cases SNR is measured as the ratio of clean data power $\|\Theta\gamma\|$ to noise power $\|\nu\|$.

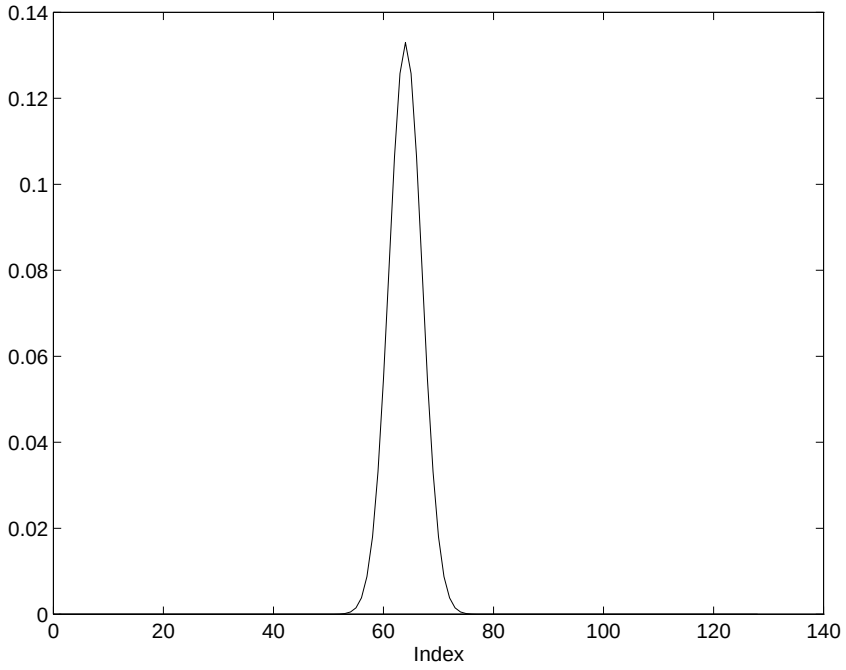

Figure 3.1: The Gaussian kernel with $\sigma = 3$.

For the examples, the matrix Θ is constructed as the wavelet transform of a discrete convolution of a Gaussian kernel, as was shown in section 2.1.1, with delta basis functions for both the data and the object. To aid in the wavlet transform, T is a circulant matrix; that is we assume periodicity of the data. Figure 3.1 shows a kernel for $\sigma = 3$. The vector sizes are 128 elements. An important consideration for the regularization is the condition number of the operator. The condition number for

the Gaussian operator for σ between .25 to 5, which are the values of interest for the examples at hand, ranges from 10^{16} to 10^{21}.

3.2 Estimation of $1/f$ Processes

We will first examine the estimation of $1/f$ processes. In these cases, the model can be matched perfectly to the true covariance if the parameters are correct, thus the estimator performance will achieve the optimal MSE. If the model is not matched perfectly, then the estimation will be degraded, however as is shown in the following examples, such losses are minimal.

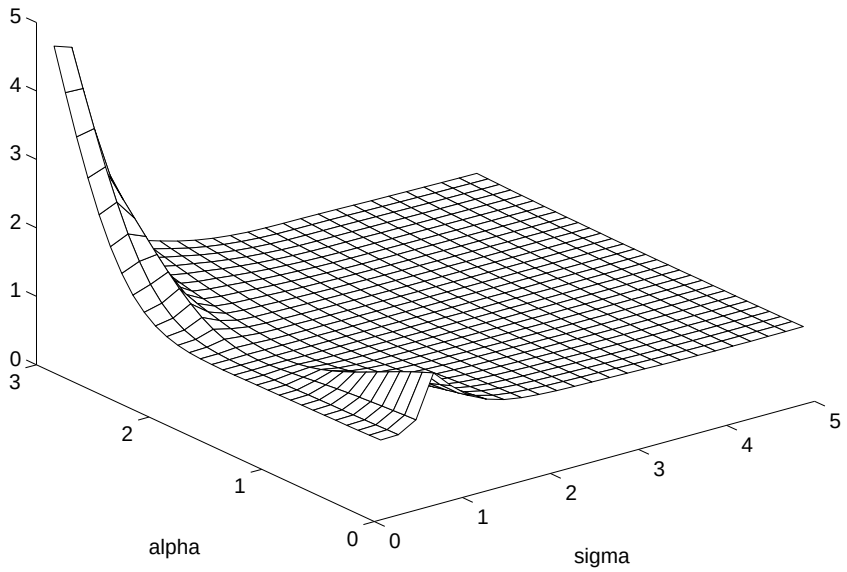

Figure 3.2: The normalized MSE performance of the estimator versus the model parameter and the operator parameter. The process has $\alpha = 1.5$ and the SNR is 20dB.

Figure 3.2 shows the NMSE resulting from estimating a $1/f$ process with $\alpha = 1.5$ with a mismatched model versus σ, the parameter which controls the amount of

blurring. The signal to noise ratio for this example is 20 dB. In this plot, it can be seen that when the model is matched, i.e. α of the model equals 1.5, the optimal NMSE of 1 is achieved. At low values of σ, the degradation is much more severe for the mismatched model than at higher values. At higher values the blurring function essentially eliminates the fine scale detail and little information remains. For low σ, a substantial amount of blurring still occurs, and the condition number of the operator is still very large, but a fairly closely matched model will achieve very acceptable results. Indeed in most cases of model mismatch, the estimator still performs better than regularizing with identity (equivalent to $\alpha = 0$ with our estimator). Particular care must be taken to avoid values of α which are too large, as this can cause overly smooth solutions and a performance which is worse than that of standard regularization.

Figure 3.3 shows the worst case performance across all σ for the mismatched model. At a noise level of 20 dB, a performance within 10% of the optimal can always be achieved with an α which varies from 1.1 to 1.8, more than a 20% deviation above or below the true value of α.

This example demonstrates the robustness of the model with respect to a Gaussian convolution across many degrees of blurring. Of particular interest is the high sensitivity of the model within a certain range of σ, 0.4 to 1.0 for the example shown. This range is dependent upon the signal to noise ratio, and can be seen to move to higher σ as the SNR increases. This sensitivity also increases as the SNR increases as will be seen in the next example. It can also be seen from Figure 3.3 that a model α which is larger than the true α can degrade performance much more than a lower α.

The next example, shown in Figure 3.4 with slices at 4 SNRs shown in Figure 3.5, demonstrates the performance of the mismatched model over a wide range of signal to noise ratios. At very high signal to noise ratios, much of the signal is retained and the degradation from using a mismatched model is shown by the sharp increase

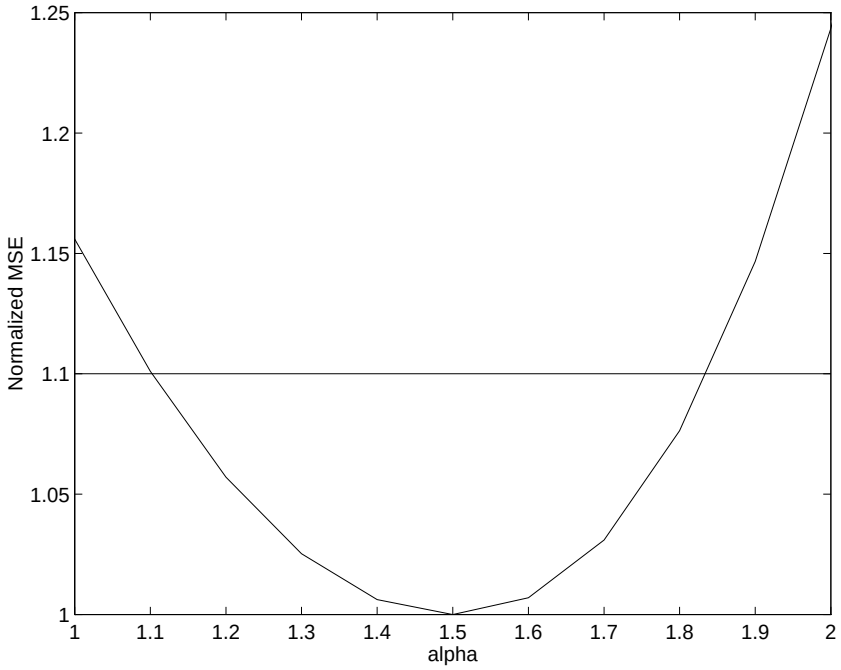

Figure 3.3: Worst case model mismatch performance across all values of σ for SNR=20 dB. The performance is within 10% of optimal over a wide range of α.

in MSE for both high and low values of α. Conversely, as the SNR becomes worse, the acceptable range of α increases, showing that the estimation is less sensitive to model mismatch. At all levels of SNR the matched or closely matched models perform significantly better than the standard identity regularization. It can be seen that at SNR= 0 dB, the range of α within 10% of the optimal is quite large.

3.3 Estimation of FOGM Processes

We will now examine the estimation of first order Gauss Markov (FOGM) processes using $1/f$ models. In this case, the models do not match the true covariance, thus the performance under model mismatch is important in that it is an example of severe model mismatch. The sample path of a FOGM process is described by a covariance

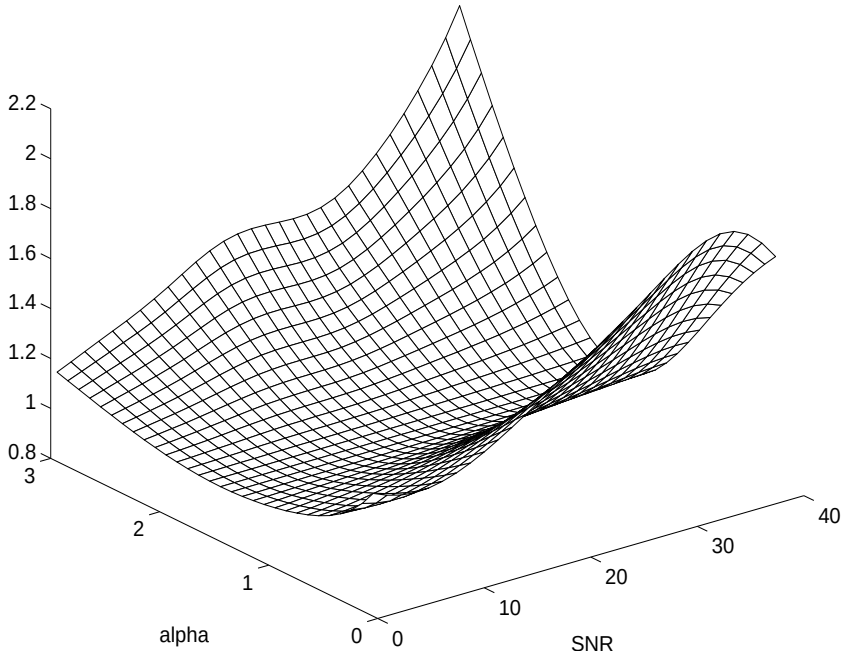

Figure 3.4: The normalized MSE performance of the estimator versus the model parameter and the signal to noise ratio. The process has $\alpha = 1.5$ and $\sigma = 1.0$.

matrix with the following element values

$$[P_g]_{i,j} = \rho^{|i-j|} \tag{3.5}$$

where ρ is the correlation value, and may vary from 0 to 1. At low values the samples are relatively uncorrelated and the process becomes white. At values near 1, the samples are highly correlated and the process becomes increasingly low pass. Viewed in this way, it can be seen that low values of α should correspond to low values of ρ and vice versa. This will be seen in Figure 3.6 where the α of best performance, defined by lowest NMSE, will increase with the value of ρ. The example in Figure 3.8 shows the NMSE performance for a range of α and σ. Here it can be seen that at low values of σ when little blurring occurs, the best performance possible with the models does not meet the optimal. At best it is 20% above optimal. The models however still perform better than identity regularizers ($\alpha = 0$). A similar structure to that found

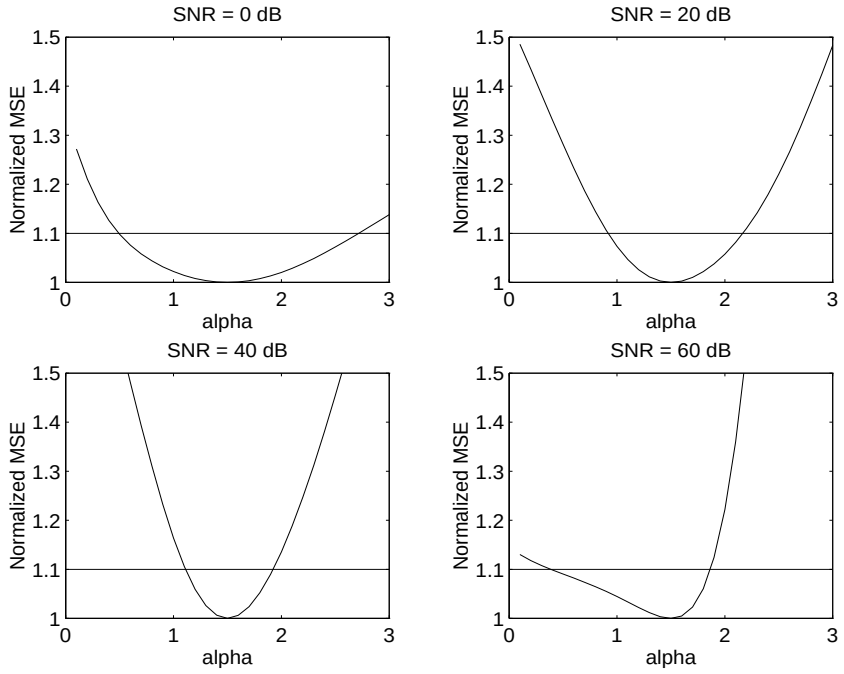

Figure 3.5: The normalized MSE performance for four signal to noise ratios, demonstrating the robustness with respect to model mismatch at low SNRs.

in the estimation of the $1/f$ processes also occurs in this plot. At values of σ in the range of .4 to 1.0, the MSE is more sensitive to α than at other ranges. Likewise, this region of sensitivity can be seen to move to higher values when the SNR increases.

Lastly, in Figure 3.9 and Figure 3.10 we examine the performance for two different correlation values, .25 and .75, versus SNR. We see that as in the $1/f$ case, the highest region of sensitivity occurs at high values of SNR. As the SNR decreases, the degradation due to model mismatch is much less, essentially all estimates degrade at the low SNR values. A contrast between the two plots shows that the more uncorrelated process, $\rho = .25$, is very resilient to model mismatch, but as was seen in the $1/f$ plots, when the model parameter α is large, performance does degrade quickly. It is useful to keep this in mind when working with the estimators for the parameters as we will in the next chapter.

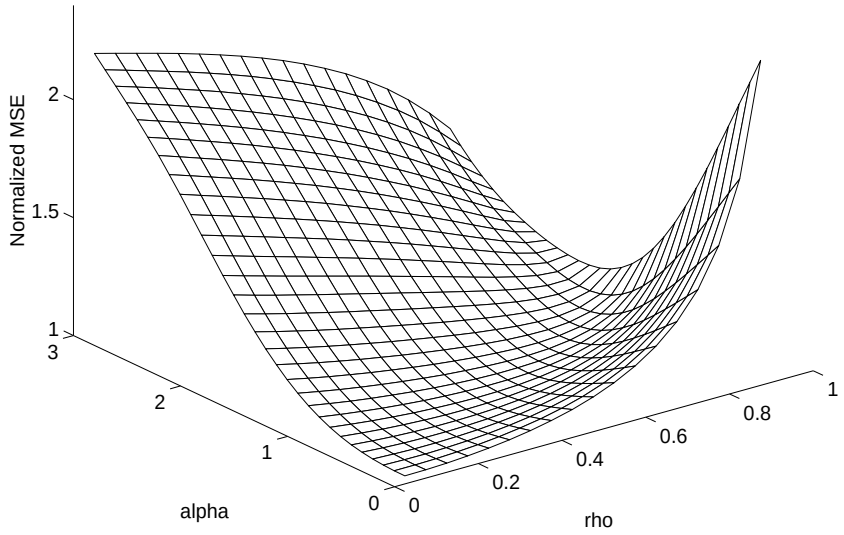

Figure 3.6: Normalized MSE for α versus ρ. The α of best performance can be seen to increase with ρ. The SNR is 20dB and $\sigma = 1.0$.

In this chapter we examined how the $1/f$ type models performed when estimating $1/f$ type processes and also FOGM processes. We saw that the model is fairly robust to model mismatch in both situations. Specifically we noted that the regions of highest sensitivity to model mismatch were in the cases of high SNR, narrow operator, and fairly correlated, i.e. low pass processes. Thus in these cases we should demand a more accurate estimation of the model parameters. The next chapter examines how to estimate the model parameters.

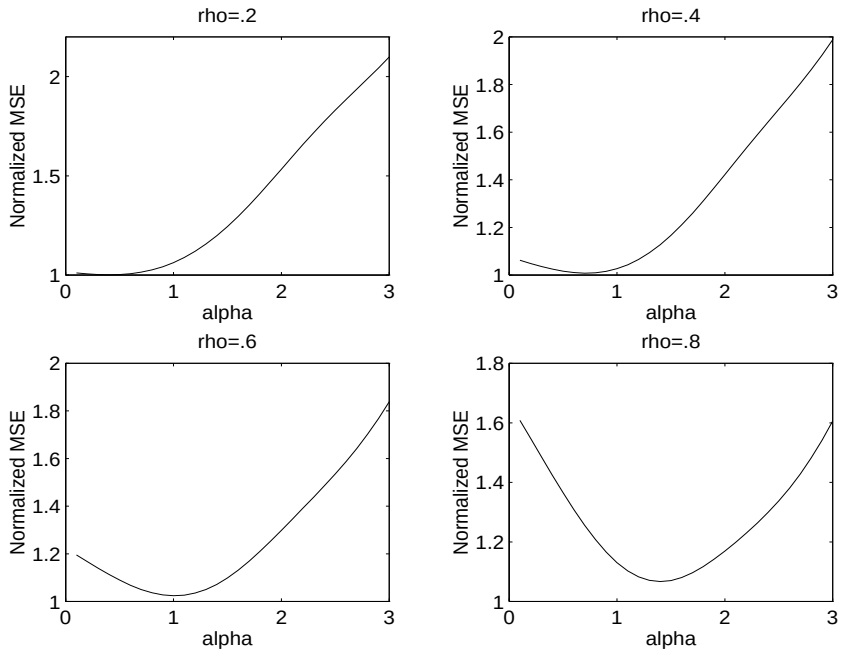

Figure 3.7: The Normalized MSE performance for 4 values of ρ. The value of α for best performance can be seen to increase with ρ. The SNR is 20dB and $\sigma = 1.0$.

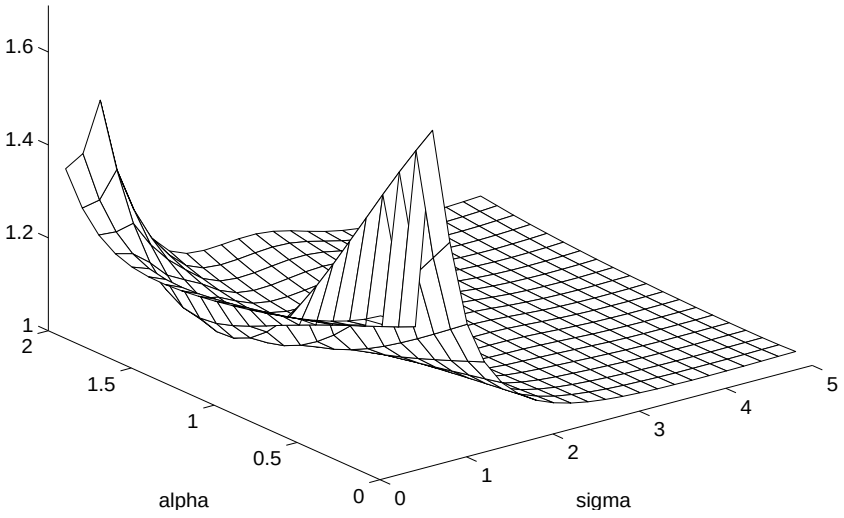

Figure 3.8: Normalized MSE for α versus σ. The SNR is 20dB and $\rho = .75$

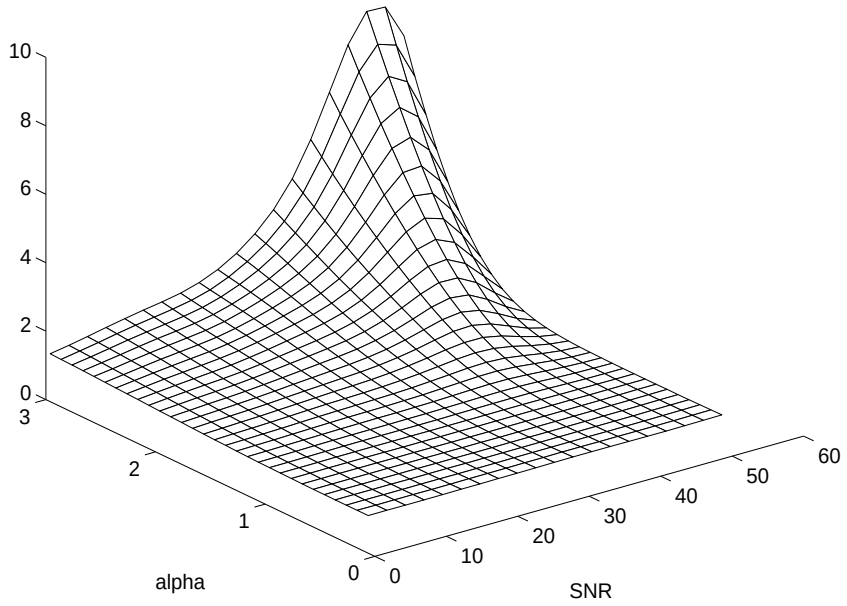

Figure 3.9: Normalized MSE for α versus SNR. $\sigma = 1.0$ and $\rho = .25$

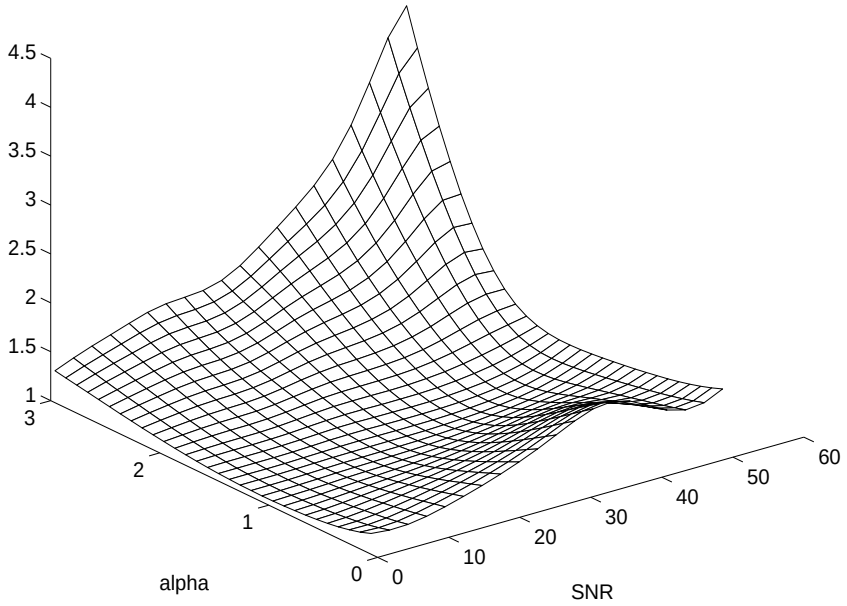

Figure 3.10: Normalized MSE versus SNR for $\rho = .75$.

Chapter 4

Parameter Estimation

In the last chapter we introduced the stochastic models and demonstrated their use in the LLSE estimator as a regularization term. These models were shown to perform well even when mismatched with the process, although of course the optimal performance is achieved when the models match the process. In this chapter we will explore techniques for estimating the parameters of the model accurately so as to minimize the performance degradation due to model mismatch. In order to do this we will employ a maximum likelihood (ML) estimation framework. We will examine the Cramer-Rao bounds upon the estimation as this will show the optimal possible performance. In addition we will look at the Expectation Maximization (EM) algorithm as the primary tool for implementing the ML estimation.

4.1 Likelihood

For the estimation of the object γ, we have been using linear least squares estimation, which requires only knowledge of the mean and variance of the object in order to obtain the estimate [22]. In the absence of this knowledge, it is essential that a good estimate of the mean and variance is made. Estimation of the entire covariance matrix

is not possible with a single realization of the data, but it is possible to fit the two parameter $1/f$ model to the data and achieve an approximate covariance matrix.

In Chapter 3, it was shown that the $1/f$ models are fairly robust to model mismatch even in the estimation of processes other than $1/f$, performing as a standard regularizer in the highly mismatched cases. However in order to produce estimates which approach the optimal MSE we must have a model matrix which closely approximates the actual covariance. We will use maximum likelihood estimation in order to fit the model to the data. Likelihood will allow us to estimate the model parameters accurately and thereby approach the optimal MSE estimation of γ.

The likelihood function is constructed from the probability density function $p(x;t)$, where x is a realization of a random vector and t is a vector of parameters which describes the probability density function, such as the mean and variance of a Gaussian random process. This function is typically regarded as having fixed t and the independent variable is x. In the likelihood function the variable of interest is now the vector of parameters t for a fixed realization of x as in

$$L(t;x) = p(x;t). \tag{4.1}$$

With this likelihood function, we may now question which parameters are the most likely parameters for the family of densities which we have chosen to use, given the data x. Therefore the maximum likelihood estimation of the parameters $\hat{t}$ is given as

$$\hat{t} = \arg\max_t L(t;x). \tag{4.2}$$

Of course the success of the parameter estimation is greatly dependent on the family of densities, or models, which are being used.

4.2 Maximum Likelihood Estimation with γ

In our problem, we are using the family of Gaussian densities for γ with zero mean and a variance specified by P_γ. The probability function for this multivariate Gaussian is therefore

$$p(\gamma; P_\gamma) = \frac{1}{(2\pi)^{N/2}|P_\gamma|^{1/2}} \exp\left(-\gamma' P_\gamma^{-1}\gamma/2\right). \tag{4.3}$$

Where N is the dimension of γ. The parameter of interest for the likelihood function will be the model covariance matrix P_γ. Fixing a realization for γ and taking the elements of the matrix P_γ as the independent variables, we form the likelihood function as

$$L(P_\gamma; \gamma) = \frac{1}{(2\pi)^{N/2}|P_\gamma|^{1/2}} \exp\left(-\gamma' P_\gamma^{-1}\gamma/2\right). \tag{4.4}$$

The likelihood expression will be simpler to use without the exponential. The logarithm being a monotonic transform will eliminate the exponential operation while preserving the maximum. We will also discard the constant terms since they will have no impact upon the evaluation of the maximum. Thus the log-likelihood function for our problem is

$$l(P_\gamma; \gamma) = -\frac{1}{2}\gamma' P_\gamma^{-1}\gamma - \frac{1}{2}\log|P_\gamma|. \tag{4.5}$$

A typical likelihood function is shown in Figure 4.1. Recall that the matrix P_γ can be written as

$$P_\gamma = \kappa F(\alpha) \quad \text{and the inverse} \quad P_\gamma^{-1} = \frac{1}{\kappa}F^{-1}(\alpha), \tag{4.6}$$

where $F(\alpha)$ is a diagonal matrix.

Using the model, (4.5) can be rewritten as

$$\begin{aligned} l(\alpha, \kappa; \gamma) &= -\frac{1}{2\kappa}\gamma' F^{-1}(\alpha)\gamma - \frac{1}{2}\log|\kappa F(\alpha)| & (4.7)\\ &= -\frac{1}{2\kappa}\gamma' F^{-1}(\alpha)\gamma - \frac{N}{2}\log\kappa - \frac{1}{2}\log|F(\alpha)|. & (4.8) \end{aligned}$$

To solve for the maximum of these equations, we will take the first derivative with respect to each parameter, and setting them equal to zero solve for the ML estimates

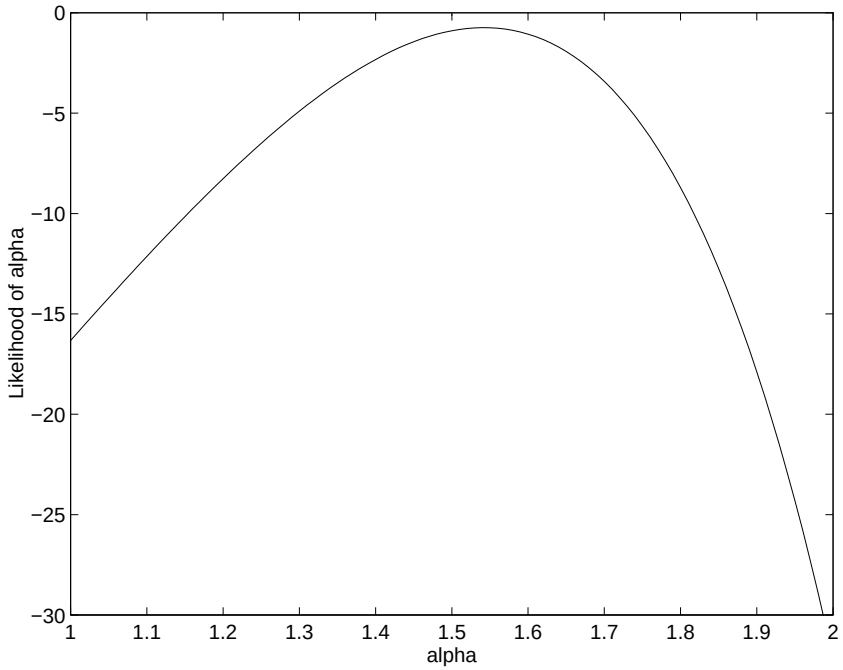

Figure 4.1: The likelihood function of α for a specific γ. The true value for α is 1.5.

of α and κ. For this, it will be useful to have the relationship for the derivative of F. For neater notation, the dependence of F upon α will be assumed but no longer written. Further, the matrix S will be introduced as a diagonal matrix, which has along the diagonal the scale index corresponding to the respective wavelet coefficient. Using this, the derivative of F with respect to α is

$$\dot{F} = \frac{\partial F}{\partial \alpha} = 2^{\alpha} S F \frac{d}{d\alpha}(2^{-\alpha}) = -(\log 2) S F. \tag{4.9}$$

Taking the derivatives of (4.7) with respect to κ and α gives

$$\frac{\partial l}{\partial \kappa} = \frac{1}{2\kappa^2}\gamma' F^{-1}\gamma - \frac{N}{2\kappa} \tag{4.10}$$

$$\begin{aligned} \frac{\partial l}{\partial \alpha} &= \frac{1}{2\kappa}\gamma' F^{-1}\dot{F}F^{-1}\gamma - \frac{1}{2}\mathrm{tr}\{F^{-1}\dot{F}\} \\ &= -\frac{\log 2}{2\kappa}\gamma' F^{-1} S\gamma + \frac{\log 2}{2}\mathrm{tr}\{S\} \end{aligned} \tag{4.11}$$

Setting Equation (4.10) to zero, we can be solve for κ giving

$$\kappa_0 = \frac{1}{N}\gamma' F^{-1}\gamma = \frac{\|\gamma\|^2_{F^{-1}}}{N} \tag{4.12}$$

The solution of (4.11) for α is not as obvious.

$$\begin{aligned} 0 &= \frac{\log 2}{2\kappa}\gamma' F^{-1} S\gamma - \frac{\log 2}{2}\mathrm{tr}\{S\} && (4.13)\\ &= \mathrm{tr}\{\gamma\gamma' F^{-1} S\} - \kappa \mathrm{tr}\{S\} && (4.14)\\ &= \mathrm{tr}\{\gamma\gamma' F^{-1} S\} - \mathrm{tr}\{\gamma\gamma' F^{-1}\}\frac{\mathrm{tr}\{S\}}{N} && (4.15)\\ &= \mathrm{tr}\left\{\gamma\gamma' F^{-1}(S - \frac{\mathrm{tr}\{S\}}{N} I)\right\} && (4.16) \end{aligned}$$

Now since F,S, and I are diagonal matrices, this equation can be rewritten as a summation of the diagonal elements as

$$0 = \sum_i 2^{\alpha s_i}\gamma_i^2 (s_i - \frac{\mathrm{tr}\{S\}}{N}) = P(2^\alpha) \tag{4.17}$$

where the values of s_i are the scale numbers of the wavelet coefficients. Since these form the exponents of 2^α, this can be regarded as a polynomial in 2^α. Since the values of the F matrix are covariances, they must be real and positive, therefore the solution for 2^α must be the real positive root of this polynomial. Since the first two factors of the summation are positive, the sign of each coefficient is determined by the third factor. Since this value is the scale number minus the mean of the scale numbers, and the scale numbers are non-decreasing across i, this number will start negative then become positive. The polynomial has only one sign change, and therefore only one real positive zero. The proof of this is in Appendix A. The ML estimate for α is then computed as the zero of this polynomial, as in

$$\alpha_0 = \arg_\alpha (P(2^\alpha) = 0). \tag{4.18}$$

The likelihood function has a unique maximum on the values of interest. The whole solution would thus take the form of solving the polynomial for the unique positive

root which gives α_0 and then solve (4.12) for κ_0. The pair (α_0,κ_0) are the parameters of the model with the maximum likelihood.

We can now solve for the variance of the estimation using the Fisher information matrix, which is the inverse of the covariance matrix of the estimations. The Fisher Information matrix is composed of the expected values of the second derivatives of the likelihood function [23, 22] evaluated at the true parameters. Taking the second derivatives of the likelihood function gives

$$\frac{\partial^2 l}{\partial \kappa^2} = -\frac{1}{\kappa^3}\gamma' F^{-1}\gamma + \frac{N}{2\kappa^2} \tag{4.19}$$

$$\frac{\partial^2 l}{\partial \alpha^2} = -\frac{\log^2 2}{2\kappa}\gamma'(S^2 F^{-1})\gamma \tag{4.20}$$

$$\frac{\partial^2 l}{\partial \alpha \partial \kappa} = \frac{\log 2}{2\kappa^2}\gamma' S F^{-1}\gamma \tag{4.21}$$

The expectation of these with respect to all γ can be calculated by applying the relation $E\{\gamma\gamma'\} = \kappa F$. The expectations are

$$E\{\frac{\partial^2 l}{\partial \kappa^2}\} = -\frac{N}{2\kappa^2} \tag{4.22}$$

$$E\{\frac{\partial^2 l}{\partial \alpha^2}\} = -\frac{\log^2 2}{2}\mathrm{tr}\{S^2\} \tag{4.23}$$

$$E\{\frac{\partial^2 l}{\partial \kappa \partial \alpha}\} = \frac{\log 2}{2\kappa}\mathrm{tr}\{S\} \tag{4.24}$$

The Fisher information matrix J is given as the negative of the expectations of the second derivatives,

$$J = \begin{bmatrix} -E\{\frac{\partial^2 l}{\partial \kappa^2}\} & -E\{\frac{\partial^2 l}{\partial \kappa \partial \alpha}\} \\ -E\{\frac{\partial^2 l}{\partial \alpha \partial \kappa}\} & -E\{\frac{\partial^2 l}{\partial \alpha^2}\} \end{bmatrix}. \tag{4.25}$$

The variances of the estimators can now be calculated by one element of J and the determinant of J. The determinant of this matrix is

$$|J| = \frac{N\log^2 2}{4\kappa^2}\mathrm{tr}\{S^2\} - \frac{\log^2 2}{4\kappa^2}\mathrm{tr}\{S\}\mathrm{tr}\{S\} \tag{4.26}$$

$$= \frac{\log^2 2}{4\kappa^2}\left(N\mathrm{tr}\{S^2\} - \mathrm{tr}^2\{S\}\right) \tag{4.27}$$

which now allows us to calculate the variance of each of our estimators as

$$\mathrm{var}(\kappa_0) = \frac{J_{2,2}}{|J|} = \frac{2\kappa^2 \mathrm{tr}\{S^2\}}{(\mathrm{tr}^2\{S\} - N\mathrm{tr}\{S^2\})} \tag{4.28}$$

$$\mathrm{var}(\alpha_0) = \frac{J_{1,1}}{|J|} = \frac{2N}{\log^2 2(\mathrm{tr}^2\{S\} - N\mathrm{tr}\{S^2\})} \tag{4.29}$$

It can be seen that the variance of κ is a function of the signal power and the scale numbers while the variance of α is simply a function of the scale numbers. With more scales the performance of the estimators improves.

4.3 Estimation in the Presence of Blurring and Noise

In the problem at hand, we do not have γ, instead we must use a likelihood function with respect to η. In this case we now wish also to estimate the unknown object γ, thus the likelihood function becomes

$$l(\gamma, \alpha, \kappa; \eta) = -\frac{1}{2}\gamma'(\kappa F(\alpha))^{-1}\gamma - \frac{1}{2}(\Theta\gamma - \eta)'R^{-1}(\Theta\gamma - \eta) - \frac{1}{2}\log|\kappa F(\alpha)|. \tag{4.30}$$

Now taking our first derivatives, which above allowed for a closed form solution, we see

$$\begin{aligned} \frac{\partial l}{\partial \kappa} &= \frac{1}{2}\eta'(\kappa\Theta F\Theta' + R)^{-1}\Theta F\Theta(\kappa\Theta F\Theta' + R)^{-1}\eta \\ &\quad -\frac{1}{2}\mathrm{tr}\{(\kappa\Theta F\Theta' + R)^{-1}\Theta F\Theta\} \qquad (4.31) \\ &= \frac{1}{2}\mathrm{tr}\left\{(\kappa\Theta F\Theta' + R)^{-1}\Theta F\Theta'((\kappa\Theta F\Theta' + R)^{-1}\eta\eta' - I)\right\} \qquad (4.32) \end{aligned}$$

$$\begin{aligned} \frac{\partial l}{\partial \alpha} &= \frac{\kappa}{2}\eta'(\kappa\Theta F\Theta' + R)^{-1}\Theta\dot{F}\Theta'(\kappa\Theta F\Theta' + R)^{-1}\eta \\ &\quad -\frac{\kappa}{2}\mathrm{tr}\{(\kappa\Theta F\Theta' + R)^{-1}\Theta\dot{F}\Theta'\} \qquad (4.33) \end{aligned}$$

$$
\begin{aligned}
&= -\frac{\kappa \log 2}{2}\eta'(\kappa\Theta F\Theta' + R)^{-1}\Theta S F\Theta'(\kappa\Theta F\Theta' + R)^{-1}\eta \\
&\quad + \frac{\kappa \log 2}{2}\mathrm{tr}\{(\kappa\Theta F\Theta' + R)^{-1}\Theta S F\Theta'\} \qquad (4.34) \\
&= \frac{\kappa \log 2}{2}\mathrm{tr}\left\{(\kappa\Theta F\Theta' + R)^{-1}\Theta S F\Theta'(I - (\kappa\Theta F\Theta' + R)^{-1}\eta\eta')\right\} \qquad (4.35)
\end{aligned}
$$

There is no simple way in which to solve these two functions for zero. Therefore, when solving for the maximum likelihood, we will use the Expectation Maximization algorithm discussed in the next section.

We will calculate the second derivatives in order to find the Cramer-Rao lower bounds upon the variances of the estimates. As in the previous case, we will first take the second derivatives.

$$
\begin{aligned}
\frac{\partial^2 l}{\partial \kappa^2} &= -\eta'(\kappa\Theta F\Theta' + R)^{-1}\Theta F\Theta(\kappa\Theta F\Theta' + R)^{-1}\Theta F\Theta(\kappa\Theta F\Theta' + R)^{-1}\eta \\
&\quad + \frac{1}{2}\mathrm{tr}\{(\kappa\Theta F\Theta' + R)^{-1}\Theta F\Theta(\kappa\Theta F\Theta' + R)^{-1}\Theta F\Theta\}. \qquad (4.36)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 l}{\partial \alpha^2} &= -\kappa^2 \log^2 2\eta'(\kappa\Theta F\Theta' + R)^{-1}\Theta S F\Theta'(\kappa\Theta F\Theta' + R)^{-1}\Theta S F\Theta'(\kappa\Theta F\Theta' + R)^{-1}\eta \\
&\quad + \frac{\kappa \log^2 2}{2}\eta'(\kappa\Theta F\Theta' + R)^{-1}\Theta S^2 F\Theta'(\kappa\Theta F\Theta' + R)^{-1}\eta \\
&\quad + \frac{\kappa^2 \log^2 2}{2}\mathrm{tr}\{(\kappa\Theta F\Theta' + R)^{-1}\Theta S F\Theta'(\kappa\Theta F\Theta' + R)^{-1}\Theta S F\Theta'\} \\
&\quad - \frac{\kappa \log^2 2}{2}\mathrm{tr}\{(\kappa\Theta F\Theta' + R)^{-1}\Theta S^2 F\Theta'\} \qquad (4.37)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 l}{\partial \alpha \partial \kappa} &= -\frac{\log 2}{2}\mathrm{tr}\{(\kappa\Theta F\Theta' + R)^{-1}\Theta S F\Theta'((\kappa\Theta F\Theta' + R)^{-1}\eta\eta' - I)\} \\
&+\kappa \log 2\mathrm{tr}\{(\kappa\Theta F\Theta' + R)^{-1}\Theta F\Theta(\kappa\Theta F\Theta' + R)^{-1}\Theta S F\Theta'((\kappa\Theta F\Theta' + R)^{-1}\eta\eta' - \frac{1}{2}I)\}
\end{aligned}
$$

These functions simplify greatly under the expectation across η. The relation in this case is

$$E\{\eta\eta'\} = \kappa\Theta F\Theta' + R.$$

$$E\{\frac{\partial^2 l}{\partial \kappa^2}\} = -\frac{1}{2}\mathrm{tr}\{(\kappa\Theta F\Theta' + R)^{-1}\Theta F\Theta'(\kappa\Theta F\Theta' + R)^{-1}\Theta F\Theta'\} \tag{4.38}$$

$$E\{\frac{\partial^2 l}{\partial \alpha^2}\} = -\frac{\kappa^2 \log^2 2}{2}\mathrm{tr}\{(\kappa\Theta F\Theta' + R)^{-1}\Theta SF\Theta'(\kappa\Theta F\Theta' + R)^{-1}\Theta SF\Theta'\} \tag{4.39}$$

$$E\{\frac{\partial^2 l}{\partial \alpha \partial \kappa}\} = -\frac{\kappa \log 2}{2}\mathrm{tr}\{(\kappa\Theta F\Theta' + R)^{-1}\Theta F\Theta'(\kappa\Theta F\Theta' + R)^{-1}\Theta SF\Theta'\} \tag{4.40}$$

And as before, the determinant of J is

$$\begin{aligned} |J| \;=\; & \frac{\kappa^2 \log^2 2}{4}\Big(\mathrm{tr}\{(\kappa\Theta F\Theta' + R)^{-1}\Theta F\Theta'(\kappa\Theta F\Theta' + R)^{-1}\Theta F\Theta'\} \\ & \cdot\mathrm{tr}\{(\kappa\Theta F\Theta' + R)^{-1}\Theta SF\Theta'(\kappa\Theta F\Theta' + R)^{-1}\Theta SF\Theta'\} \\ & -\mathrm{tr}^2\{(\kappa\Theta F\Theta' + R)^{-1}\Theta F\Theta'(\kappa\Theta F\Theta' + R)^{-1}\Theta SF\Theta'\}\Big) \end{aligned} \tag{4.41}$$

The calculations of the variances can now be performed.

$$\mathrm{var}(\alpha_0) \;=\; \frac{J_{2,2}}{|J|} \tag{4.42}$$

$$\mathrm{var}(\kappa_0) \;=\; \frac{J_{1,1}}{|J|} \tag{4.43}$$

4.4 The Expectation Maximization Algorithm

The Expectation Maximization (EM) algorithm is a two step iterative algorithm found in a wide variety of signal processing applications [24, 25]. We will be using it as a tool to jointly estimate the object as well as the model parameters. First the model parameters are fixed at some initial guess. The Expectation step is performed by taking the expectation of the object given the data for the fixed model parameters; this is an MAP estimate of the object with the current model. In the Maximization step, the current estimate of the object is used to computed the likelihood, then we maximize over the likelihood function to compute an update of the model parameters. Since the likelihood function being maximized over is that of the object, it is a much

simpler calculation than maximizing over the likelihood function of the data. In addition, an estimate of the object is produced at each iteration, progressively improving as the model parameters converge. At each step of the algorithm, the likelihood value increases, therefore the solution is guaranteed to converge to some maximum of the likelihood function, but a global maximum is not guaranteed. If local maxima exist on the likelihood function, then the location of the initial guess of the parameters will dictate to which solution the algorithm will converge. This is not a concern for our problem since, as was shown, the likelihood function in our case has a unique maximum.

4.4.1 The Expectation Step

The data which we have available, η, is the *incomplete data* . The vector $[\eta'\gamma']'$ is the *complete data* , which is not observed directly but will be estimated. The algorithm is initialized with a guess at the parameter values κ and α. In subsequent iterations, the values of the parameters will be supplied by the maximization step. In the expectation step, we will compute an estimate of the object as the expectation of γ given the data η as

$$\gamma^{[i]} = E\{\gamma|\eta, \alpha^{[i-1]}, \kappa^{[i-1]}\} \tag{4.44}$$

$$= \int_\Gamma \gamma p(\gamma|\eta, \alpha^{[i-1]}, \kappa^{[i-1]})d\gamma \tag{4.45}$$

Which is the mean of the conditional probability density function (pdf) of γ given η. This can be calculated as the LLSE of γ using the model with the current estimates of α and κ as

$$\gamma^{[i]} = C^{[i]}\Theta' R^{-1}\eta \tag{4.46}$$

Where the matrix $C^{[i]}$ is an estimate of the covariance between the object and data and is given as

$$C^{[i]} = (\Theta' R^{-1}\Theta + \frac{1}{\kappa^{[i-1]}}F^{-1}(\alpha^{[i-1]}))^{-1}. \tag{4.47}$$

This is a regularized estimate of γ with the model term as a regularizer, avoiding any problems of stability in the estimation. As was shown in chapter 3, the performance of this estimate will improve as the model parameters more closely match the true parameters.

4.4.2 The Maximization Step

In the maximization step, we will use the $\gamma^{[i]}$ and $C^{[i]}$ which were estimated in the expectation step as if they were the true γ and C, and we can now maximize the likelihood function [25] to produce $\kappa^{[i+1]}$ and $\alpha^{[i+1]}$. The maximization step then becomes

$$\begin{bmatrix} \alpha^{[i]} \\ \kappa^{[i]} \end{bmatrix} = \arg\max_{\alpha,\kappa} \left(-\frac{1}{2\kappa}(\gamma^{[i]})' F(\alpha)^{-1}(\gamma^{[i]}) - \frac{1}{2\kappa}\mathrm{tr}\{F(\alpha)^{-1}C^{[i]}\} - \frac{N}{2}\log\kappa - \frac{1}{2}\log|F(\alpha)| \right). \tag{4.48}$$

If we combine the first two terms by changing the inner product to a trace we obtain

$$\begin{bmatrix} \alpha^{[i]} \\ \kappa^{[i]} \end{bmatrix} = \arg\max_{\alpha,\kappa} \left(-\frac{1}{2\kappa}\mathrm{tr}\{F(\alpha)^{-1}(\gamma^{[i]}\gamma^{[i]T} + C^{[i]})\} - \frac{N}{2}\log\kappa - \frac{1}{2}\log|F(\alpha)| \right). \tag{4.49}$$

If we observe that the matrix F is diagonal, then the trace operation will be only using diagonal elements of the argument. In other words, the elements of $(\gamma\gamma^T + C)$ are scaled independently by the elements of F and then summed up. Therefore we can combine the estimation of the object and covariance into a single statistic vector

$$t_j = \gamma_j^2 + C_j. \tag{4.50}$$

With this vector we can reformulate our maximization as the same equations for the unblurred exact data case. Thus we can now use the simpler equations without

blurring and noise. The maximum of this function can then be produced using the steps shown in section 4.2. This is a very interesting result, since it allows us to completely eliminate the computationally burdensome step of maximizing across a two dimensional function and instead find the unique positive root of a polynomial.

4.4.3 EM Iteration and Examples

The two steps above can now be performed iteratively by taking a LLSE estimate of the object then performance a maximum likelihood estimation of the parameters using the procedure from section 4.2. Therefore we can now write the complete algorithm in the following equations:

E-Step:

$$C^{[i]} = (\Theta' R^{-1} \Theta + \frac{1}{\kappa^{[i-1]}} F(\alpha^{[i-1]})^{-1})^{-1} \tag{4.51}$$

$$\gamma^{[i]} = C^{[i]} \Theta' R^{-1} \eta \tag{4.52}$$

$$t_j^{[i]} = (\gamma_j^{[i]})^2 + C_{j,j}^{[i]} \quad \text{for} \quad j = 1 \ldots N \tag{4.53}$$

M-Step:

$$P(2^\alpha) = \sum_j 2^{\alpha s_j} t_j^{[i]} (s_j - \frac{\text{tr}\{S\}}{N}) \tag{4.54}$$

$$\alpha^{[i]} = \arg(P(2^\alpha) = 0). \tag{4.55}$$

$$\kappa^{[i]} = \frac{1}{N} \| F(\alpha^{[i]})^{-1} t^{[i]} \|_1 \tag{4.56}$$

Figure 4.2 shows the performance of the algorithm in a situation with 20dB SNR and an operator with $\sigma = 1$. As can be seen from the plot the algorithm converges in 20 steps to an accurate estimate of the parameters. Here the estimate for α is 1.529 for a true α of 1.5. The value of κ converged to 9.358 for a true κ of 10. The estimate of the object using these parameters is shown in Figure 4.3. It can be seen that the estimate is an accurate representation of the original.

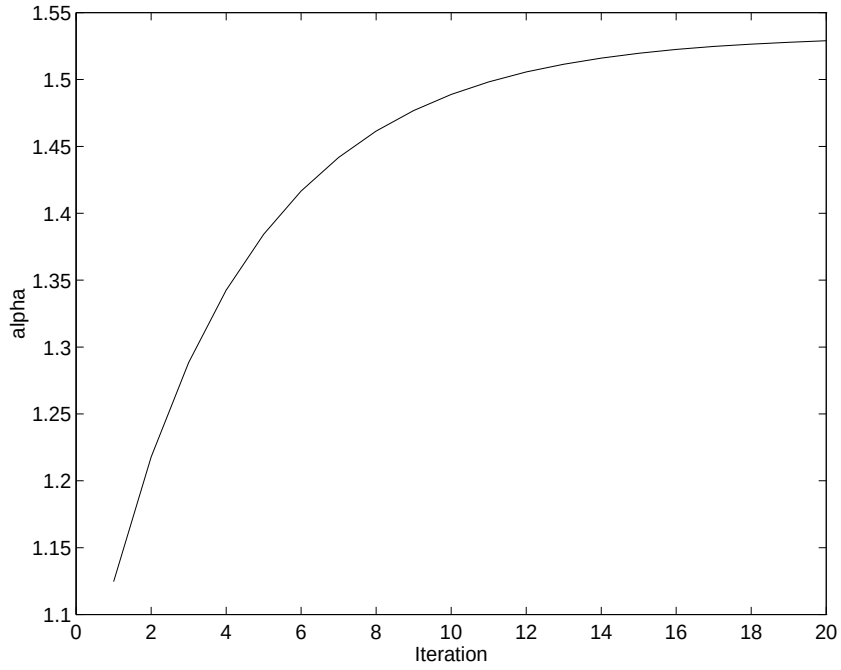

Figure 4.2: The expectation maximization algorithm's convergence for α for 20 iterations. The true $\alpha = 1.5$.

The next plot shows the estimation of a FOGM process using the same algorithm. In this case, the model does not accurately represent the true covariance matrix, but the estimation will produce a decent regularization term for the estimate of γ. Here we have a $\sigma = 1$ and 20dB SNR, with a correlation parameter $\rho = 0.8$. In this case, the algorithm converges in 22 steps to a value of 1.46 for α. The estimation of γ from this is shown in Figure 4.4.

4.5 Variance of the Estimations

In this section, we will examine the performance of the EM algorithm across a variety of situations. We will examine the estimation variances with respect to the Cramer-Rao bounds while varying the noise level, the blurring operator, the length of the estimation vector, and the number of wavelet levels.

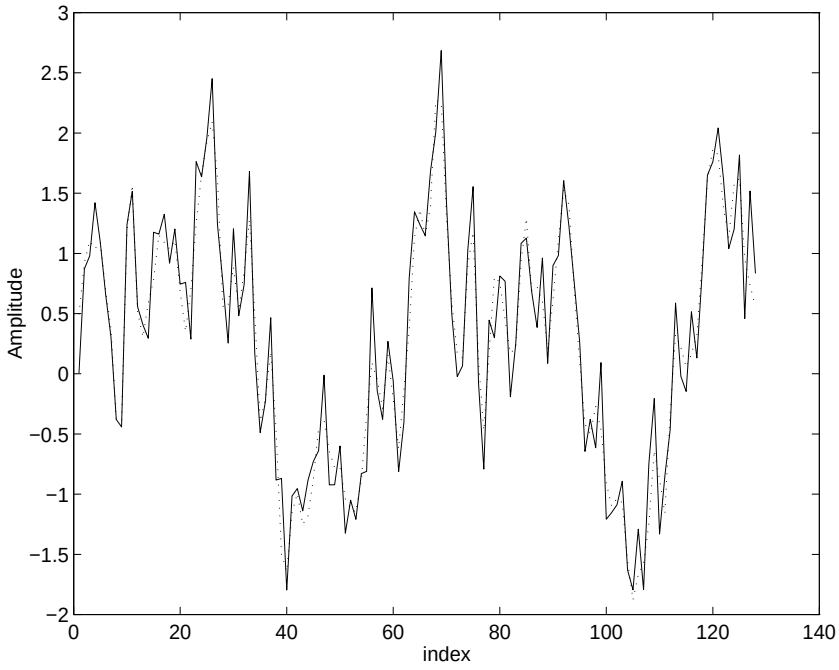

Figure 4.3: The estimate of γ produced with the estimated model parameters. Though the estimate (dotted) is a fairly good representation of the true gamma (solid).

Figure 4.5 shows the variance of the estimate of α using the EM algorithm described in the last section. The circles represent Monte Carlo simulations of the variance in order to confirm that the estimator does achieve the variances given by (4.42) and (4.43). There is quite a large error bound on the Monte Carlo simulation for the wide operator and low SNR case. Longer simulation should result in the circle moving to a location above the bounds. Likewise due to the large error bounds upon the variance estimate of κ, an inordinate number of simulations would have been necessary to confirm the bounds. As can be seen in the plot, the estimation performs very well over a wide range of SNR, with performance not breaking down significantly until the SNR falls below 5 dB. This plot was generated with a blurring kernel of $\sigma = 0.5$. The effect of changing the blurring kernel can be seen as shifting the plot to higher variance as σ increases. The breakdown in estimation variance as

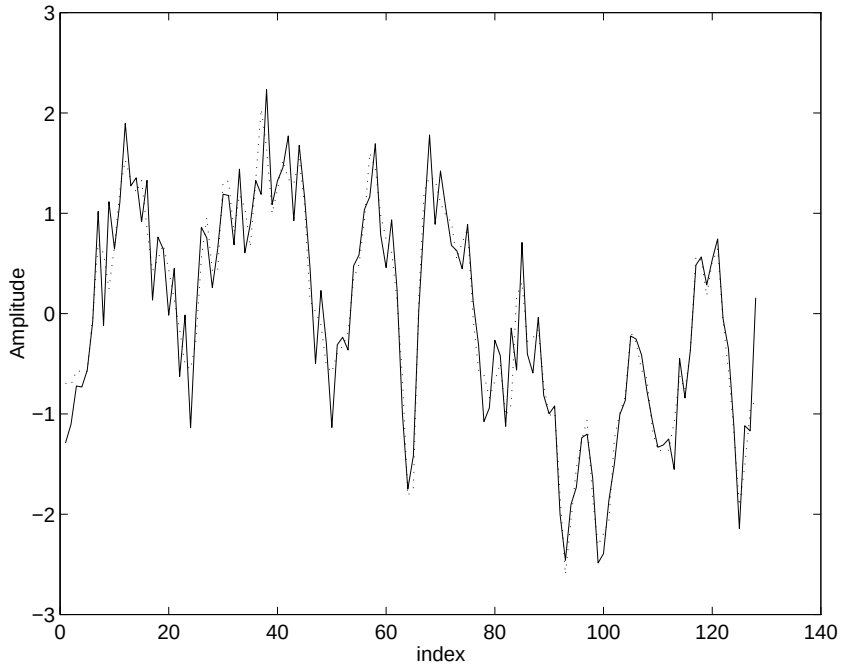

Figure 4.4: The estimate of γ produced with the estimated model parameters for a FOGM process with $\rho = 0.8$.

the SNR becomes worse does not translate into a serious degradation in the estimation of γ as can be seen from the results of Chapter 3. As the SNR decreases, the sensitivity to model mismatch decreases. Thus the estimation performs best, in the higher SNR cases, where the better estimates are needed.

Figure 4.6 shows the variance bounds with respect to the width of the blurring parameter. These variances were generated for a SNR of 20 dB. Similar curves develop for other SNR, with lower SNR shifting the curve to higher variances as was seen in Figure 4.5. It can be seen that at a low σ, i.e. a narrow blurring kernel, the variance levels off. There is little difference between estimation with a kernel of $\sigma = 0.5$ and no blurring at all. This is significant since it allows fairly accurate estimation of the model parameters and thus the object in the range of non-trivial blurring functions. However, as the blurring kernel becomes wider the performance degrades.

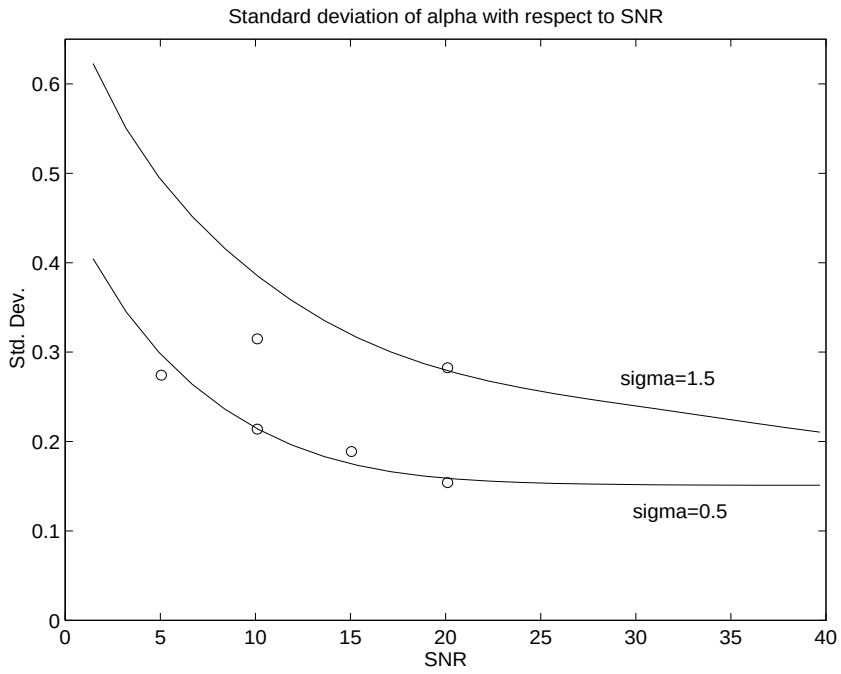

Figure 4.5: The variance of the EM estimate of α versus SNR. Th circles represent the variance from 100 Monte Carlo simulations of estimation.

For blurring operators of $\sigma > 2$, the performance has degraded to a point where the variance is 5 times that of the unblurred case. However, the degradation in estimation of the object is not as severe. The results of chapter 3 showed that as the blurring operator becomes wider, the reconstruction becomes much less sensitive to model mismatch. This means that in the case of a wide operator the reconstruction methods demonstrated will still perform well since the model mismatch from the EM algorithm will have little effect on the reconstruction. Conversely, in the case were the reconstruction is sensitive to model mismatch, that of a narrow kernel, the EM estimation performs very well.

Figure 4.7 shows the effect of the number of samples upon the estimates of the parameters. As expected, the variance decreases as the number of samples increases. This plot is for a constant SNR of 20 dB and a $\sigma = 0.5$. The speed of the decrease is fast, at $1/N$. Thus more samples will quickly achieve a higher degree of accuracy in

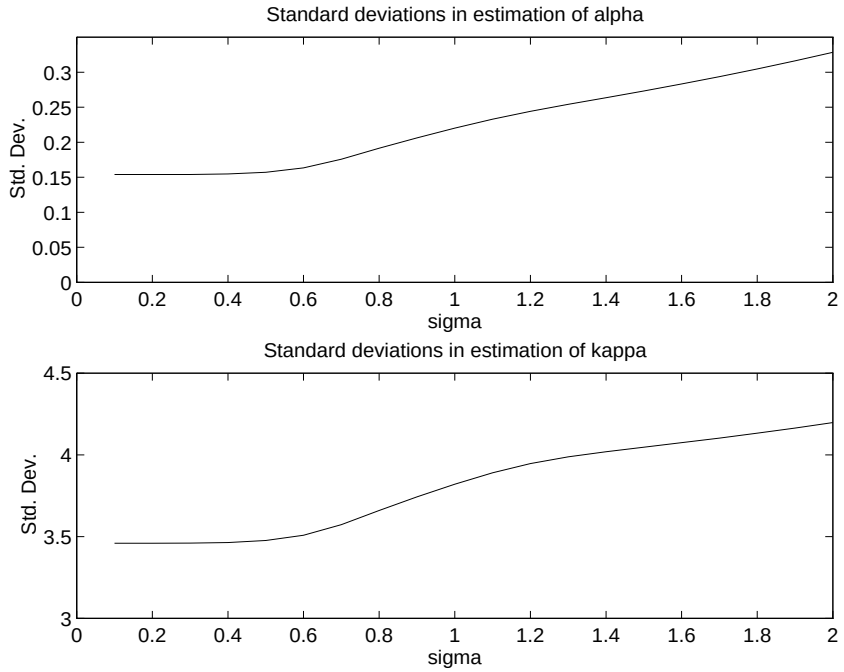

Figure 4.6: The variance of the EM estimates of α and κ versus the blurring parameter σ.

the parameter estimation for a constant SNR and blurring function. Lastly, Figure 4.8 shows the effect of the number of wavelet levels upon the variance of α. Since the estimate of α is an estimate of the sample variances of the data at each scale, it is logical that as the number of scales available for comparison increases, the variance will decrease. The plot shows that the variance decreases sharply as the number of scales increases from 4 to 6. Thus increases estimation accuracy can also be achieved with the use of more wavelet scales, if this is reasonable. Due to the length of the object vector in this problem, $N = 128$, more than 6 scales would not be practical.

We have seen that through the use of the EM algorithm, the model parameters can be estimated with acceptable accuracy in most cases. Where the estimation begins to break down, either through a large blurring operator or a low SNR, the need for high accuracy in the model parameters is alleviated by the relative insensitivity of the object estimation to the mismatch in the models as was shown in chapter 3. The

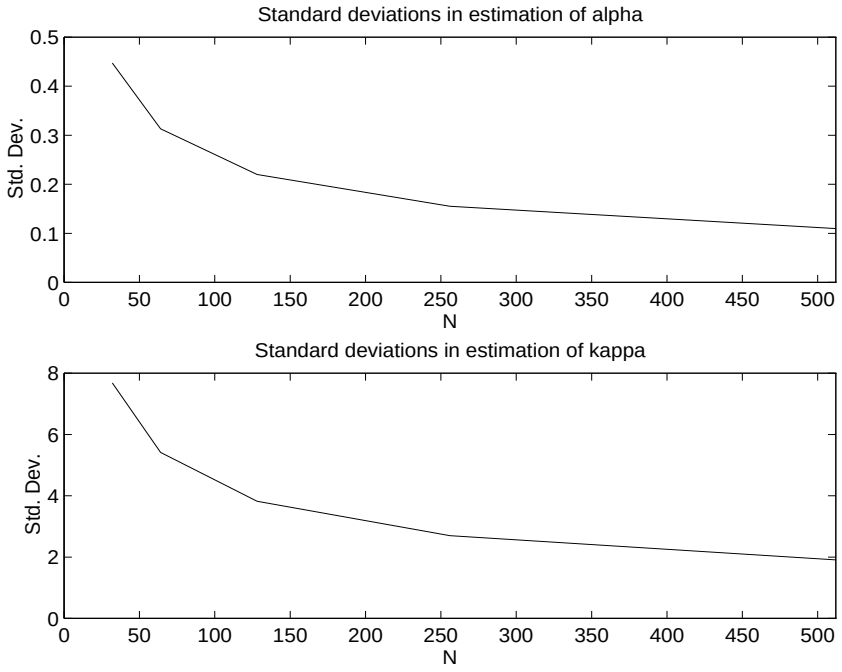

Figure 4.7: The variance of the EM estimates of α and κ versus the number of samples N.

the EM algorithm gives us an estimator which performs well over the class of cases where the higher performance is needed more.

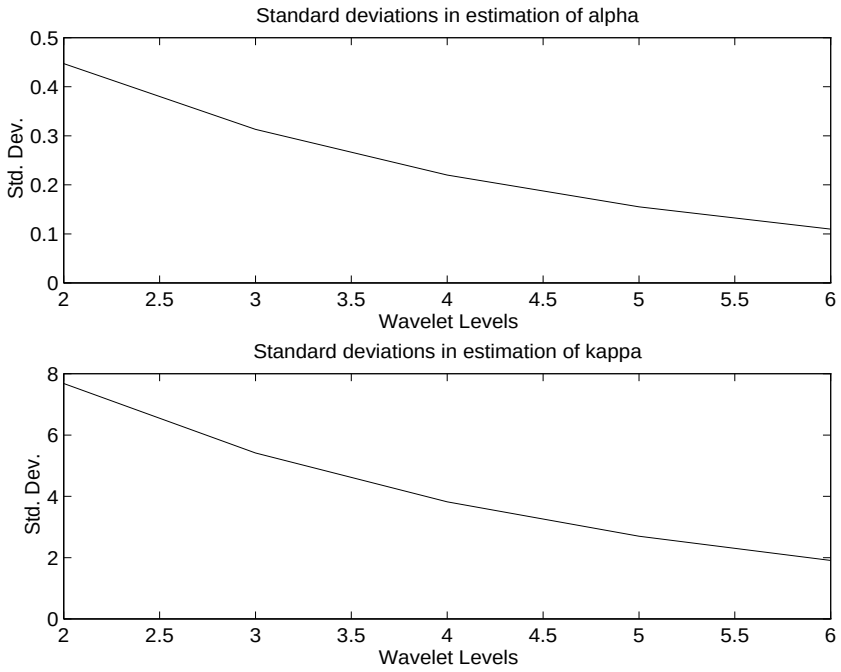

Figure 4.8: The variance of the EM estimates of α and κ versus the number of wavelet levels S for a fixed number of samples.

Chapter 5

Conclusions and Future Work

This thesis examined the use of $1/f$ type fractal prior models for the estimation of a stochastic object in a linear inverse problem. The performance of the $1/f$ models was shown to be insensitive to model mismatch under a range of operators and SNRs. This insensitivity is important since we wish to use the models in situations where we do not know the true statistics of the unknown object and will instead estimate them. Particular conditions were identified where the model does have a higher sensitivity to model mismatch. These conditions (narrow blurring kernel and high SNR) were the same conditions where it was shown that the estimation of the models is most accurate. We also examined a particular formulation of the Expectation Maximization algorithm where exploiting the model parameters allows a more efficient implementation.

Chapter 2 introduced the problem of recovering a degraded signal from a limited number of noise corrupted samples. Problems of this type are often ill-conditioned and proper solution requires the introduction of some prior constraints. The $1/f$ model has been shown to perform well in this role as a regularization term. The flexibility of the model allows differing levels of smoothness to be imposed while retaining the simplicity of only two parameters. A further simplification can be imposed upon the

model by eliminating the parameter κ. This is made possible by estimating the energy in the process and then conditioning κ upon α and the energy. This simplification reduces the model to a single parameter constraining smoothness.

In Chapter 3 the issue of model mismatch was examined under the estimation of a $1/f$ fractal signal and a non-$1/f$ signal (specifically FOGM). These were chosen to show the sensitivity with respect to a mismatched model when a matched model achieves optimal performance, and where the $1/f$ model cannot match the true covariance matrix. The model was shown to be robust. Across a range of SNR, the performance is rather insensitive to model mismatch. The sensitivity increases as the SNR increases. Thus, it is the conditions of high SNR where accurate parameter estimation is necessary. Likewise, across a range of blurring kernels, the sensitivity increases as the kernel becomes narrower. With narrow kernels, the parameter estimation must perform more accurately.

With the specific needs of model mismatch in mind, Chapter 4 examined the estimation of the model parameters. The Expectation Maximization algorithm was used. The EM algorithm solved the maximum likelihood function iteratively to achieve the maximum. In this way, it avoids seeking a closed form solution for the ML estimate where none exists. By first exploring the EM algorithm under the assumption of no blurring and no noise, a more efficient implementation of the algorithm was arrived at which eliminates the maximization step and replaces it with solving for the root of a polynomial. By exploiting the model in the more complex case of blurring and noise, the same maximization step from the simple case can be implemented. This vastly increases the speed of implementation, leading to the ability to solve problems with much larger vectors in less time.

5.1 Future Work

Overall, this thesis has shown that the $1/f$ models are a useful tool for solving a specific class of inverse problems. They are robust to model mismatch, perform a similar function to classical regularizers, and are completely defined by two parameters. In addition, it is possible to eliminate the power parameter by bounding it with an estimate of the energy, further simplifying the model. The estimation of the parameters has been examined and it is shown that it can be performed accurately at a greatly reduced computational burden. There are still however many aspects to explore.

1. The model mismatch issue has not been completely covered. This work examined mismatch under the conditions of a Gaussian blurring function across a range of widths and SNRs. The theoretical work is general enough to incorporate all linear operators and therefore a further exploration could be done with many common operators seen in signal and image processing. More significant, this work did not explore the performance degradation when the object distribution was other than Gaussian. When the Gaussian restriction is relaxed to more general distributions, the performance will most certainly decay and should be understood.

2. Though the work of chapter 4 lead to a more efficient implementation of the EM algorithm for these problems by eliminating the maximization step, the speed can still be increased. The expectation step requires solving for the information matrix and using this to solve for the LLSE of the object. Work has been done in [26] which leads to a fast approximation of the information matrix by operating in the wavelet domain. Inclusion of this in the expectation step would lead to a further reduction in computation and allow for larger signals. Additionally properties of the model might also lead to reduction of this step.

3. This thesis explored the use of the $1/f$ models for one dimensional signals. A natural extension of this is two dimensional images. Some work has been done on this problem. Further work conducted here would certainly benefit from the previous recommendation, since a major constraint upon this work is the size of the images which can be examined reasonably.

4. In this thesis, the model parameters are used globally across the whole signal. In extending this work to images, especially large images, a spatially adaptive implementation may be preferable. This would require fitting the model to the local statistics; possibly by iterating from the top of the wavelet tree to the finer scales. This would lead to much better performance on non-stationary images.

5. Finally, this work did not implement these models in actual problems. An obvious extension to this work is applying the results and the algorithm to a real problem.

Appendix A

Proof of Unique Root

Theorem A polynomial with only one sign change in the coefficients has one positive real root.

Proof.

Let $P(x)$ be a polynomial with

$$P(x) = c_0 + c_1 x + \cdots + c_{m-1}x^{m-1} - c_m x^m - \cdots - c_n x^n$$

where the coefficients $c_i \geq 0$ for all i.

Then $P^{(m)}(x)$, the m^{th} derivative of $P(x)$, is

$$P^{(m)}(x) = -m!c_m - \frac{(m+1)!}{1!}c_{m+1}x - \cdots - \frac{n!}{(n-m)!}x^{n-m}.$$

$P^{(m)}(x) < 0$ for all $x > 0$. The function $P^{(m-1)}$ is the $(m-1)^{\text{th}}$ derivative of $P(x)$, and has $P^{(m)}(x)$ as its first derivative.

$$P^{(m-1)}(x) = (m-1)!c_{m-1} - \frac{m!}{1!}c_m x - \frac{(m+1)!}{2!}c_{m+1}x^2 - \cdots - \frac{n!}{(n-m+1)!}x^{n-m+1}.$$

$P^{(m-1)}(0) > 0$, and $\lim_{x\to\infty} P^{(m-1)}(x) = -\infty$. By the Intermediate Value Theorem, $P^{(m-1)}(x)$ has at least one point where $P^{(m-1)}(x) = 0$. Also, since $P^{(m)}(x) < 0$ for

all values of $x > 0$, $P^{(m-1)}(x)$ is monotonically decreasing for $x > 0$, therefore by the Mean Value Theorem the point where $P^{(m-1)}(x) = 0$ is unique. Let this point be x_0.

Now,

$$P^{(m-2)}(x) = (m-2)!c_{m-2}+(m-1)!c_{m-1}x-\frac{m!}{2!}c_m x^2-\frac{(m+1)!}{3!}c_{m+1}x^3-\cdots-\frac{n!}{(n-m+2)!}x^{n-m+2}.$$

$P^{(m-2)}(0) > 0$ and is monotonically increasing over the interval $0 < x < x_0$, therefore there is no point in this interval where $P^{(m-2)}(x) = 0$. $P^{(m-2)}(x_0) > 0$ and $\lim_{x\to\infty} P^{(m-2)}(x) = -\infty$, thus by the same reasoning as above, there is one point $x > x_0$ for which $P^{(m-2)}(x) = 0$. Let this point be x_1, and repeat the proceedure until $P(x)$ is reached at which point there is a unique point $x > 0$ for which $P(x) = 0$. QED.

Bibliography

[1] A.S. Willsky P.W. Fieguth and W.C. Karl. Multiresolution stochastic imaging of satellite oceanographic altimeter data. In *IEEE International Conference on Image Processing*, volume 2, 1994.

[2] W.C. Karl M.R. Luettgen and A.S. Willsky. Efficient multiscale regularization with application to the computation of optical flow. *IEEE Transactions on Image Processing*, 3(1):41–64, 1994.

[3] E.L. Miller and A.S. Willsky. A multiscale approach to sensor fusion and the solution to linear inverse problems. *Applied and Computational Harmonic Analysis*, 2:127–147, 1995.

[4] J. Zhang G. Wang and G.W. Pan. Solution of inverse problems in image processing by wavelet expansion. *IEEE Transactions on Image Processing*, 4(5):579–593, 1995.

[5] A.N. Tikhonov and V.Y. Arsenin. *Solutions to Ill-Posed Problems*. V.H. Winston & Sons, 1977.

[6] P.C. Hansen. Analysis of discrete ill-posed problems by means of the l-curve. *SIAM Review*, 34(4):561–580, December 1992.

[7] M. Bertero. Linear inverse and ill-posed problems. In Peter W. Hawkes, editor, *Advances in Electronics and Electron Physics*, volume 75. Academic Press, 1989.

[8] I. Daubechies. *Ten Lectures on Wavelets.* SIAM CBMS No. 61, 1992.

[9] R. Coifman G. Beylkin and V. Rokhlin. Fast wavelet transforms and numerical algorithms, i. *Commun. on Pure Appl. Math.*, 44:141–183, 1991.

[10] C.K. Chui. *An Introduction to Wavelets.* Academic Press Inc., 1992.

[11] G. Strang. Wavelets and Dilation Equations: A Brief Introduction. *SIAM Review*, 31(4):614–627, December 1989.

[12] P.P. Vaidyanathan. *Multirate Systems and Filter Banks.* Prentice-Hall Inc., 1993.

[13] M. Vetterli and J. Kovacevic. *Wavelets and Subband Coding.* Prentice-Hall Inc., 1995.

[14] G.W. Wornell. A karhuenen-loeve-like expansion for $1/f$ processes via wavelets. *IEEE Transactions on Information Theory*, 36:859–861, July 1990.

[15] C.A. Balanis. *Advanced Engineering Electromagnetics.* John Wiley and Sons, 1989.

[16] L.L. Tsai. Moment methods in electromagnetis for undergraduates. *IEEE Transactions on Education*, 21:14–22, February 1978.

[17] L.L. Scharf. *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis.* Addison Wesley, 1991.

[18] A.M. Mathai and S.B. Provost. *Quadratic Forms in Random Variables.* Marcel Dekker, I.nc, 1992.

[19] O. Rioul and M. Vetterli. Wavelets and Signal Processing. *IEEE Signal Processing Magazine*, 8:14–38, October 1991.

[20] David Hoag. *Compression techniques using the wavelet transform with applications to underwater image/video data.* PhD thesis, Northeastern University, 1996.

[21] B. Porat. *Digital Processing of Random Signals.* Prentice Hall, Inc., 1994.

[22] S.D. Silvey. *Statistical Inference.* Chapman and Hall, 1975.

[23] A.W.F. Edwards. *Likelihood.* The John Hopkins University Press, 1992.

[24] Todd K. Moon. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, pages 47–60, November 1996.

[25] A. K. Katsaggelos. *Digital Image Restoration.* Spinger-Verlag, 1991.

[26] Eric L. Miller. Efficient statistical inversion in the wavelet transform domain. Technical Report TR-CDSP-97-41, Northeastern University, Boston, MA, January 1997.