IMM

Numerical and Statistical Approaches to Inverse Problems

Master Thesis

Delievered by Robert Owusu - s030509
Supervised by Ole Winther and Per Christian Hansen
Formally Supervised by John Aasted Sørensen of ITU

Lyngby, June 23, 2005

Preface

This Master thesis serves as a documentation for my final assignment which is a requirement for achieving the Master of Science in Multimedia Technology at the IT University of Copenhagen. This work was carried out at the Informatic and Mathematical Modelling Department at the Technical University of Denmark. I extend my sincere thanks to Associate Professor Ole Winther and Professor Per Christian Hansen for the great guidance and supervision they accorded me. I also thank Associate Professor John Aasted Sørensen for inspiring me through out the thesis.

Abstract

The focus of this thesis is on Statistical and Numerical Approaches for solving ill-posed deconvolution (or inverse) problems using the L-Curve for Tikhonov's Regularization, Maximum A Priori (MAP), Maximum Likelihood (both viewed in the context of Statistical Regularization), Evidence Framework for Bayesian Inference and Variational Bayesian Expected Maximization (VBEM) as an alternative method for optimizing the parameters in the Bayesian Inference Framework.

Furthermore, concise treatments of Empirical Bayes, ML Expected Maximization algorithm, Variational Bayes ML and Variational MAP are given.

The main aim and objective is to have a new look at Regularization schemes within the Statistical and Numerical environment. We therefore compare and contrast existing and new methods and based on the formulae given by the methods find their corresponding estimates to see whether they exhibit some consistencies in results.

Contents

List of Figures

Nomenclature

We used the following symbols and abbreviations.

$\tilde{g}$	is the output of dimension n.
K	discretized kernel matrix of dimension $n \times n$.
Σ	covariance matrix.
D	diagonal matrix consisting of the singular values of K
U	matrix consisting of the left singular vectors of K
V	matrix consisting of the right singular vectors of K
f_{ls}	standard Least Squares estimator for true f.
d	vector consisting of the singular values of K.
f	is the n-vector to be found.
$\mathbf{u}_i$	Left Singular Vector at i^{th} column.
$\mathbf{v}_i$	Right Singular Vector at i^{th} column.
λ^2_{rls}	numerical regularization parameter.
λ^2_{eb}	empirical bayes regularization parameter.
λ^2_{ml}	statistical regularization parameter.
α	equivalent to λ^2_{ml}.
β	precision parameter noise variance.
$f_{map_{\lambda_{ml},\sigma}}$	MAP posterior estimate for f.

$f_{ml_{\lambda_{ml},\sigma}}$ ML posterior estimate for f.

$f_{\lambda_{rls}}$ numerical estimate for f.

f_{MP} most probable estimate for f using ML principle.

BIM Bayesian Inference Method

VBEM Variational Inference Expectation Maximization Method

MLM or ML Regularized Maximum Likelihood Method

LCM The L-Curve for Tikhonov's Regularization Method

MAP Maximum A Priori

Introduction

The main purpose of data modelling is to design models that can capture the relevant information from a noisy observed data. This task have always drawn experts from various fields of study into desgning systems or models with the goal of finding an explanation to the underlying structure of the data at hand. However, this is not easy to achieve since we base our decisions on the results of filtering or predictions or inferences about the data we have in hand and possibly on what we expected to observe before the data arrived.

It is often very difficult to know which aspects of the data are relevant for an inference or filtering (or prediction) task and which part should be regarded as noise.

In this thesis, we exploit both Numerical and Statistical approaches to modelling an inverse problem with emphasis on methods and estimates for a particular application. We consider the standard model $\tilde{g} = Kf + \epsilon$ where it is assumed that $K \in \mathbb{R}^{n\times n}$ and of rank n, $f \in \mathbb{R}^n$ and ϵ has mean zero and variance a scalar multiple of the identity matrix I. We focus on the case where the Least Squares do not make sense when put into the context of the Physics, Chemistry and engineering of the process which is generating the data $\tilde{g}$.

The goal is to treat Tikhonov's form of Regularization from both Numerical and Statistical viewpoints by comparing methods Numerically and Statistically and further use the methods to estimate the parameters in the models to see whether consistencies exists among the methods.

In order to enhance consistency in our work, we dealt with a particular problem and maintained the same number of parameters throughout our work. We categorized the whole thesis into the following Chapters:

Chapter 1 is devoted to only the simulation model. It treats the given problem as a Fredholm integral equation of the first kind. Chapter 2 is about Numerical Least Squares and Regularization and Chapter 3 handles the same problem using Stochastic Modelling concepts within the Statististical environment. Chapter 4 follows with a 'Take Home' message about some comparisons between the Numerical and Statistical Framework. Chapter 5 is about Numerical and Statistical Estimation Theory and our contributions with Chapter 6 showing results from the estimates based on the methods.

CHAPTER 1

Simulation Model

The model problem to be used in this thesis is a geomagnetic prospecting problem taken from [1]. We will use it as our simulation model for deconvolution (or the inverse problem). Figure (1.1) assumes a 1-D horizontal mass distribution at a depth h below a given surface. It shows the geometry and the location of the s and t axes.

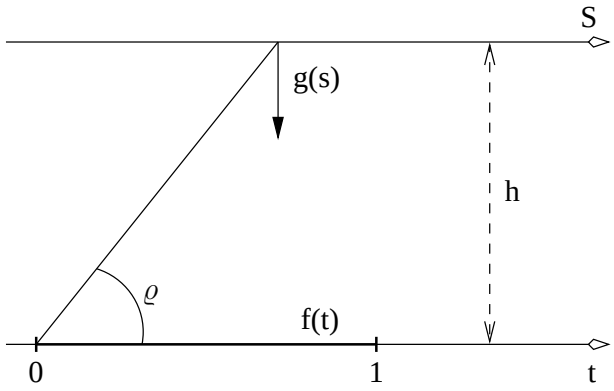

Figure 1.1: *A geometrical illustration of a gravity surveying problem in* $1-$*dimension. The measured signal* $g(s)$ *is the vertical component of the gravity field due to a* $1-$*dimensional mass distribution* $f(t)$ *at a depth* h*.*

1.0.1 Problem Formulation

Our objective is to determine or estimate the input f and at the same time minimize some performance criterion. The formulation of the problem requires that we do the following:
(i) Give a mathematical description of the overall system to be dealt with.
(ii) Give a statement of constraints where necessary.
(iii) Give a specification of a performance criterion.

1.0.2 Mathematical Model

From the measurements of the vertical component of the gravity field, denoted $g(s)$, at the surface, we want to compute the mass distribution, denoted $f(t)$, along the t-axis. In the following, we derive below the necessary equations governing the model to be used in this thesis.

Given a small infinitesimal change dt of the mass distribution $f(t)$, the corresponding small change dg is given by

$$dg \;=\; \frac{\sin(\varrho)}{r^2}\, f(t)dt \tag{1.1}$$

and the distance between the two points on the s and t axes is given by $r = \sqrt{h^2 + (s-t)^2}$. Using that $\sin(\varrho) = h/r$, we get

$$\frac{\sin(\varrho)}{r^2} \;=\; \frac{h}{[h^2 + (s-t)^2]^{\frac{3}{2}}} \tag{1.2}$$

The total value of g for any s is

$$g(s) \;=\; \int_0^1 \frac{h}{[h^2 + (s-t)^2]^{\frac{3}{2}}}\, f(t)dt \tag{1.3}$$

with the limit of integration constrained to lie within the unit line. Equation (1.3) leads to a deconvolution problem for computing the latent variable f with kernel $h\,/\,\{h^2 + (s-t)^2\}^{-\frac{3}{2}}$. The discretization of the continuous integral equation (1.3) together with the measured output g is always contaminated with errors. Furthermore, numerical computations often involve non-negligible rounding errors. Such inaccuracies always lead to inevitably small perturbations which make the direct practical inversion process of f highly unstable. For this problem, we let the quantity f be given by

$$f(t) \;=\; \sin(\pi t) \;+\; 0.5\sin(2\pi t) \tag{1.4}$$

and let $T(s,t)$ represent

$$T(s,t) \;=\; \frac{h}{\{\, h^2 + (s-t)^2 \,\}^{3/2}} \tag{1.5}$$

Equation (1.3), becomes

$$g(s) \quad = \quad \int_0^1 \frac{h}{\{\, h^2 + (s-t)^2 \,\}^{3/2}} \Big(\sin(\pi t) \;+\; 0.5\sin(2\pi t) \Big)\, \mathrm{d}t \tag{1.6}$$

The above continuous integral is then expressed as a quadrature through an appropriate quadrature method based on quadrature rules. This rule is used to sample equation (1.6) at $n - \textit{equally}$ spaced abscissa's $s_1, s_2 ....., s_n$. The quadrature rule for computing an approximation to any arbitrary definite integral (in general) takes the form

$$\int_0^1 \varphi(t)dt \;=\; \sum_{j=1}^{n} w_j\, \varphi(t_j) \tag{1.7}$$

Next, we apply the midpoint rule (or the trapezoidal rule for periodic functions) to the problem using the formulae

$$t_j \;=\; \frac{j \;-\; 0.5}{n} \qquad ; \qquad w_j \;=\; \frac{1}{n}$$

The subsequent approximation to the continuous integral equation (1.6) then becomes

$$\begin{aligned} \int_0^1 \frac{d}{[d^2+(s-t)^2]^{\frac{3}{2}}} \{\sin(\pi t) + 0.5\sin(2\pi t)\} dt &\approx \sum_{j=1}^{n} w_j\, T(s,t_j)\, \tilde{f}(t_j) \quad j = 1,2,..,n \\ &= \psi(s) \end{aligned} \tag{1.8}$$

We let $K_{i,j} = w_j T(s_i,t_j)$, $g(s_i) = \psi(s_i)$ and $f_j = \tilde{f}(t_j)$.

The elements of $\tilde{f}(t_j)$ are the computed samples at discrete abscissa's $t_1, t_2, ..., t_n$. It is straight forward to conclude that the discretized function

$$\psi(s_i) = g(s_i) \quad i = 1,2,....,n$$

For simplicity, we will always assume that the discretization of T is square.

CHAPTER 2

Numerical Least Squares and Regularization

The mathematical description of the simulation model in Figure (1.1) satisfies the definition of a first order Fredholm integral equation of the form

$$g(s) \;=\; \int_{\Omega} T(s,t) f(t)\,\mathrm{d}t \tag{2.1}$$

where Ω defines the limit of integration in $n-$dimensional space and the notations T, f, g are the same as mentioned in Chapter 1. Several methods for solving equations of the first kind numerically have been proposed. One should view equation (2.1) as a linear operator, operating on the function $f(t)$ to produce $g(s)$. The nature of the operator does not often allow it to have a bounded inverse [1]. For instance if we let $f(t)$ be a solution of equation (2.1) and define it as $f(t) \;=\; \sin(2\pi pt) \quad p = 1, 2, ...,$.
Then for any integrable kernel, we have

$$g(s) \;=\; \int_{\Omega} T(s,t)\sin(2\pi pt)\,\mathrm{d}t \longrightarrow 0 \quad \text{as} \quad p \longrightarrow \infty \tag{2.2}$$

Equation (2.2) implies that an infinitesimal small change dg in g can cause a corresponding arbitrarily large change $\mathrm{d}f$ in f. Hence, the ability to solve equation (2.1) successfully depends largely on the accuracy of $g(s)$ and the shape of $T(s,t)$.

2.1 Why Numerical Least Squares and Regularization

If a solution corresponding to equation (2.1) for $g(s)$ exists, a slight perturbation of $g(s)$ may give rise to an arbitrarily large variation in the solution $f(s)$. This results in an equation which may be closely satisfied by a function that bears the same resemblance to the true solution. However, there are some difficulties associated with this instability. This is often due to the fact that in practice the specification of $g(s)$ is usually inexact because of the data at hand. Thus, the "true" or actual data g are corrupted with some noisy samples at certain discrete abscissas $s_1, s_2....., s_n$. We can sometimes be confronted with an ill-conditioned inverse problem in contrast to a well-conditioned inverse problem. In either case, we state the problem as

$$\tilde{g}(s) \;=\; g(s) \;+\; \epsilon(s) \tag{2.3}$$

where ϵ is an arbitrary function referred to as measurement noise and it is measured based on some condition about the size. The problem statement is often related to a functional inequality $|\,\epsilon|$ bounded above such that

$$|\,\epsilon(s)| \leq M \qquad \text{or} \qquad \int_{\Omega} w(s)\,\epsilon^2(s)\,\mathrm{d}s \;\leq\; \tilde{M}; \quad w(s) \;>\; 0$$

[1] Sometimes, the operator may not have an inverse at all. For simplicity, we will assume it has one.

where $w(s)$ are weights. For a moment, let us address the functional with weights $w(s)$.

Instead of having a unique solution to equation (2.3), we obtain a family $\mathbb{F}$ of solutions. Our problem is then to pick out from the family of functions $\mathbb{F}$, the true solution f. This is impossible to find if additional information about the problem represented by equation (2.3) is not given. Here, we have made the assumption that, the functional form of f is unknown, hence our *inability to use a Least Square fit* alone to find the best fit to f. Moreover, we expect the function f to be reasonably smooth (which is often the case). One probably has to choose from an entire family of functions say $f_s \in \mathbb{F}$, the best approximation to f which is smoothest in some sense. This calls for the need of a regularizer. We will assume that, the functions f, g and T are all identically zero outside the unit line (i.e the limit of integration Ω is confined to $\Omega \subset [0,1]$ for a 1-D case).[2] Methods used in discretizing the continuous integral equation (2.1) coupled with the associated ill-posed nature do welcome techniques in Numerical Linear Algebra for solving inverse (or deconvolution) problems.

By applying the quadrature method(s) described in Chapter 1 to equation (2.1) we get

$$\sum_{i=0}^{n} w_j T(s_i, t_j)\tilde{f}(t_j) \;=\; g(s_i) \qquad ; \quad i,j = 1,2,...,n \tag{2.4}$$

where $\tilde{f}_j = \tilde{f}(t_j)$, $\tilde{g}_i = g(s_i)$, $\epsilon_i = \epsilon(s_i)$, $w_j T(s_i, t_j) = K_{i,j}$ and w_j are weighting coefficients whose values depend on the quadrature formula used.

The condition on the magnitude of ϵ is defined by

$$\sum_{j=0}^{n} \epsilon_j^2 \;=\; \epsilon^2$$

where ϵ^2 is a constant.

A convenient way to express equation (2.4) is

$$\tilde{g} \;=\; Kf \;+\; \epsilon \tag{2.5}$$

The naive solution (which we shall denote $\tilde{f}$) of equation (2.5) often gives a poor representation of the true solution and it is when $\epsilon = 0$. The solution have an oscillatory feature which conflicts with our apriori knowledge. Figure (2.1) shows how the naive solution can be very different from the true solution f. The elements of the computed naive vector

$$\tilde{f} \;=\; K^{-1}\,\tilde{g} \tag{2.6}$$

are, in principle, mere approximations to the desired solution. Thus, $Kf = \tilde{g}$ is infact $\tilde{g} = g + \epsilon$ and the vector ϵ also represents perturbation of the exact data. In other words, a good representation of the true solution is only attainable when ϵ is non-zero. To verify this, just introduce the matrix notation

$$K_{ij} \;=\; w_i k_{ij} \qquad \text{and} \quad \text{let} \quad K_{ij}^{-1} \;=\; \nu_{ij}$$

[2]For a square we have $\Omega \subset [0,1] \times [0,1]$ and $\Omega \subset [0,1]_1 \times [0,1]_2, ... \times [0,1]_n$ for unit hypercube in n-dimensions.

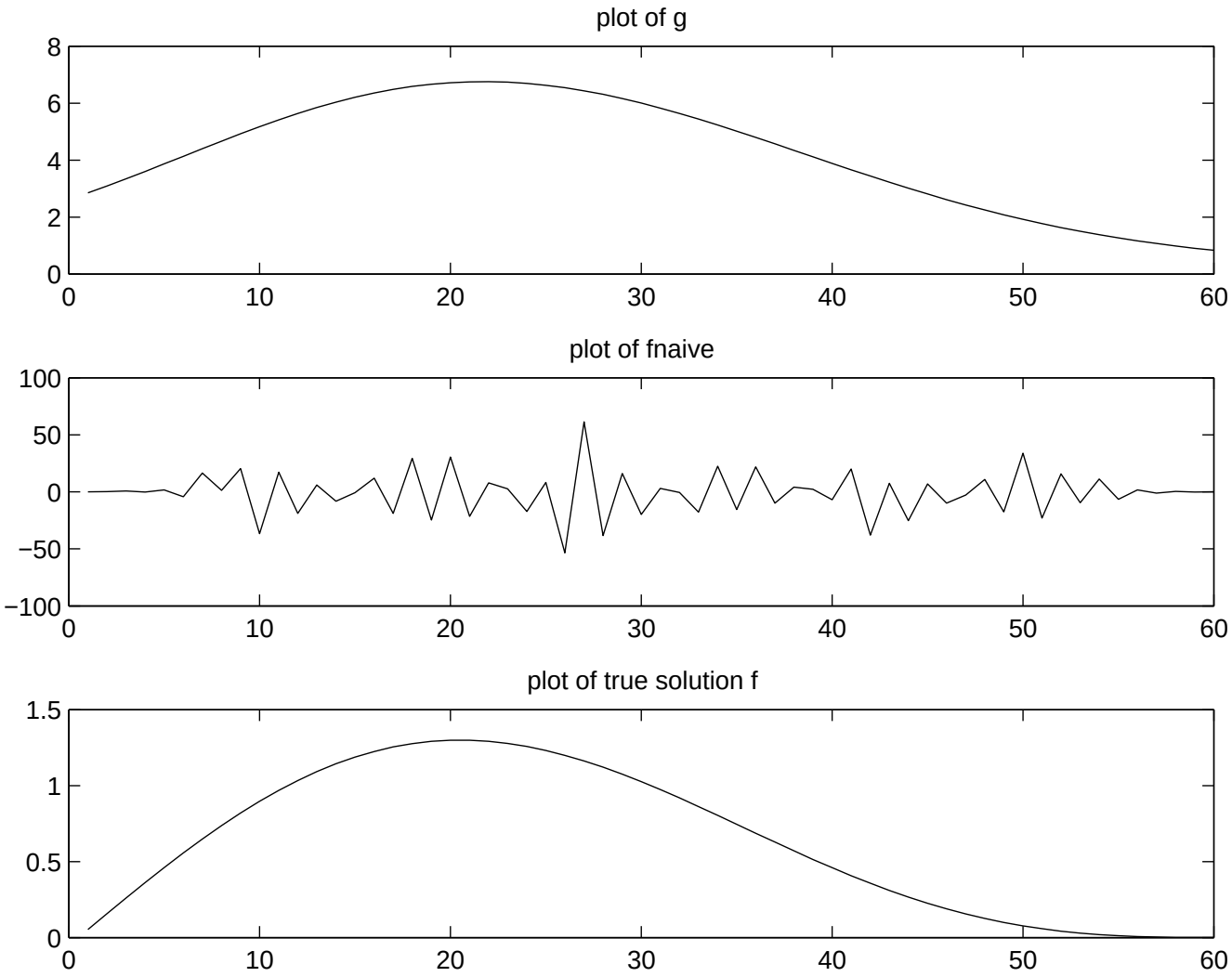

Figure 2.1: plot of output $\tilde{g}$ (top), naive $\tilde{f}$ (middle) and true function f (bottom) versus the number of components i (for $i = 1, 2, ..., 60$). The function $\tilde{f}$ is obtained from the direct inversion, $K^{-1}\tilde{g}$. However it should be noted from above Figure that the 'plot of g' at the top refers to $\tilde{g}$ instead.

Then it easy to see that

$$f = K^{-1}g + K^{-1}\epsilon \tag{2.7}$$

which explains the reason why the behaviour of the weighted kernel K must also be taken into consideration because f in equation (2.7) is a linear function of g and ϵ. Furthermore, by taking partial derivatives of f_i with respect to either g_j or ϵ_j gives the inverse of the weighted kernel ν_{ij};

$$\frac{\partial f_i}{\partial g_j} = \nu_{ij} = \frac{\partial f_i}{\partial \epsilon_j} \quad ; \quad i,j = 1,2,.....n \tag{2.8}$$

In situations where the dimension is high, the matrix K may be rank deficient, hence a stable inverse does not exist. By introducing a regularizer, it is possible to achieve a reasonably smooth function say $f_{\lambda_{rls}}$ which can be accepted as a good (or best) representation of the exact function f.

2.2 Least Squares and Normal Equations

Given the problem of finding the vector $f \in \mathbb{R}^n$ from

$$\tilde{g} = K f \tag{2.9}$$

where $K \in \mathbb{R}^{n\times n}$ is the data matrix and $\tilde{g} \in \mathbb{R}^n$ is the output. Here we assume that both K and $\tilde{g}$ are available. Practically speaking, we do not expect systems of the form of equation (2.9) to have solutions since the output vector $\tilde{g}$ must be an element of the range space of K which is a proper subspace of $\mathbb{R}^n$.

Our objective is to minimize $\|\tilde{g} - Kf\|_p$ for a suitable choice of p . That is

$$\min_f \quad \left\|\tilde{g} - Kf\right\|_p \tag{2.10}$$

In contrast to the p-norm we choose $p = 2$ for two tractable reasons which are as follows:
(i)

$$\phi(f) = \frac{1}{2}\left\| \tilde{g} - Kf \right\|_2^2 \tag{2.11}$$

is a differentiable function of f and so a minimizer of ϕ satisfies $\nabla\phi(f) = 0$. This operation leads to the construction of a symmetric linear system (i.e by forcing any anti-symmetric component of K to vanish) which is positive definite if K has full rank.
(ii) The 2-norm is preserved under orthogonal transformation. That is,
$\|(U^TK)f - U^T\tilde{g}\|_2$ is easy to solve whilst it maintains the equivalent minimizer of $\|\tilde{g} - Kf\|_2^2$. We shall see in subsection (2.2.2) that the length and angle are preserved under an orthogonal transformation.

Differentiating $\phi(f)$ of equation (2.11) with respect to f:

$$K^T(\tilde{g} - Kf_{ls}) = 0 \tag{2.12}$$

shows that the minimum residual denoted ϵ_{ls} is orthogonal to Ran(K) in Figure (2.2).

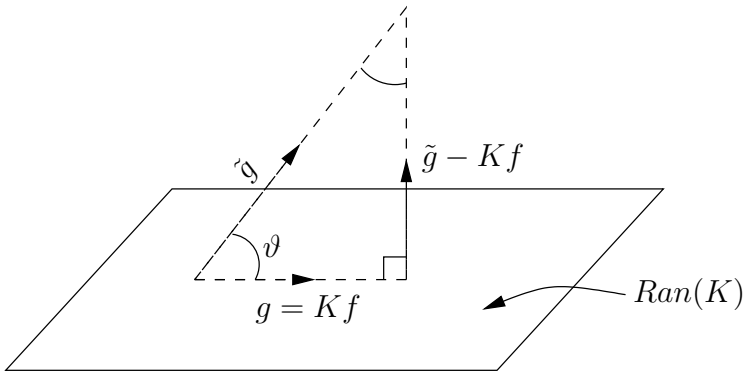

Figure 2.2: *A geometric illustration of Least Squares*

The residual ϵ_{ls};

$$\epsilon_{ls} = \tilde{g} - Kf_{ls} \tag{2.13}$$

is called the *minimum residual vector*. The corresponding size $\|\epsilon_{ls}\|_2$ given by

$$\|\epsilon_{ls}\|_2 = \|\tilde{g} - Kf_{ls}\|_2 \tag{2.14}$$

is also referred to as the *minimum residual* of the Least Squares Problem. Equation (2.12) is called the normal equations since $\nabla\phi(f) = K^T(\tilde{g} - Kf_{ls})$. The solution to the normal equations is tantamount to solving the gradient equation $\nabla\phi(f) = 0$. Furthermore, the 2-norm, $\|\tilde{g} - Kf_{ls}\|_2$ is a non-zero residual as could be seen from Figure (2.2).

In short, we state the Least Squares problem in relation to the Gravity Model of Figure (1.1):
Given

$$\tilde{g} = Kf$$

we seek

$$\min_f \| \tilde{g} - Kf \|_2$$

with solution to the normal equations given by

$$K^T(\tilde{g} - Kf_{ls}) = 0 \tag{2.15}$$

2.2.1 Orthogonality and Orthonormality

Given the set of vectors $\{\mathbf{u}_i\,;\,\mathbf{u}_i \in \mathbb{R}^n\}$ for $i = 1, 2, ..., n$. If $\mathbf{u}_i^T\mathbf{u} = 0$ for $i \neq j$, then the set of vectors is said to be *orthogonal*. If on the other hand, $\mathbf{u}_i^T\mathbf{u}_j = \delta_{ij}$ then the set of vectors is said to be *orthonormal* .

2.2.2 Singular Value Decomposition

If K is a real $n \times n$ matrix, then there exists orthogonal matrices

$$U = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_n] \in \mathbb{R}^{n\times n} \quad \text{and} \quad V = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n] \in \mathbb{R}^{n\times n}$$

such that

$$K = U\,D\,V^T = \sum_{i=1}^{n} \mathbf{u}_i\, d_i\, \mathbf{v}_i^T \tag{2.16}$$

and

$$\mathbf{u}_i^T \mathbf{u}_j = \mathbf{v}_i^T \mathbf{v}_j = \delta_{ij} \quad \text{or} \quad U^T U = V^T V = I$$

where I is the identity matrix and the set of pairs $\{\mathbf{u}_i, \mathbf{v}_j\}$ are respectively called the Left and Right Singular Vectors. Also

$$U^T K V = \text{diag}(d_1, d_2, ..., d_n) \in \mathbb{R}^{n\times n}$$

where $d_1 \geq d_2 \geq, ..., d_n \geq 0$ are the singular values of K. Also $K\mathbf{v}_i = d_i\mathbf{u}_i$. Hence

$$\|K\mathbf{v}_i)\|_2^2 = (K\mathbf{v}_i)^T(K\mathbf{v}_i) = (d_i\mathbf{u}_i)^T(d_i\mathbf{u}_i) = d_i^2\mathbf{u}_i^T\mathbf{u}_i \quad i = 1, 2....n$$

and since the $\mathbf{u}_i's$ form an orthonormal set we have

$$\|K\mathbf{v}_i)\|_2 = d_i$$

The consequence of the orthogonal transformation property preserves the length (or magnitude) of f and the angle between two vectors say f_1 and f_2. To see this, let

$$\grave{f} = U^T f$$

then

$$\|\grave{f}\|_2^2 = f^T U U^T f = \|f\|_2^2$$

and

$$\grave{f_1}^T \grave{f_2} = f_1^T U U^T f_2 = f_1^T f_2$$

That is, the effect of multiplication by an orthogonal matrix U^T is equivalent to a rigid rotation of the coordinate system.

2.2.3 Higher Dimensional Problems

In higher dimensions, it sometimes happen that the matrix K has many singular values of different magnitude close to the origin; thus rendering K to have an ill-determined rank. Therefore a Least Squares fit is not able to capture the relevant information contained in the output $\tilde{g}$. The kernel K smooths out the high frequency components of the signal which results in loss of information at high frequency components of f. Strictly speaking, $\tilde{g} = K\,\tilde{f}$. See equation (2.6). Therefore,

$$\tilde{g} = K\,\tilde{f} = g + \epsilon \tag{2.17}$$

An important consequence is the non-uniqueness of solution to the linear system of equation (2.17). Any solution subjected to the high frequency perturbations will fit the data, $\tilde{g}$ equally well. This makes the deconvolution problem of reconstructing (or recovering) the signal ill-posed. This ill-posedness is accompanied by inevitable effects of instability of the solutions. Small perturbations of the data may result in a completely different solution.

Figure (2.3) is an example that illustrates one of the difficulties that can arise when an inverse operation is performed in the frequency domain. The function K is a low-pass filter designed to handle the smoothing operation in the frequency domain. The function g is a 'clean' speech signal which is free from any form of noise. The clean speech g is then corrupted with additive white noise, ϵ which is normally distributed with zero mean and standard deviation σ_ϵ. The corrupted speech $\tilde{g}$ is

$$\tilde{g} = g + \epsilon$$

The Discrete Fourier Transform, $DFT(\tilde{g})$ is given by

$$\tilde{g}(w) = g(w) + \epsilon(w)$$

where $g(w) = DFT(g)$ and $\epsilon(w) = DFT(\epsilon)$. [3]

Hence,

$$\begin{aligned} \tilde{f}(w) &= \tilde{g}(w) \oslash K(w) \\ &= g(w) \oslash K(w) + \epsilon(w) \oslash K(w) \end{aligned} \tag{2.18}$$

where $f(w) = DFT(f)$, $K(w) = DFT(K)$ in equation (2.18). Here, we have taken cognizance of the fact that, in the Fourier (i.e frequency) domain, inverse operation involving matrix-vector division is possible. Equation (2.18) shows that, the direct division by $K(w)$ unbounds the high frequency components of $\tilde{f}$ due to the division of elements in $\epsilon(w)$ by insignificant (or very small) elements in $K(w)$. See Figure (2.3) below.

The illustrated Figure example consists of a short sequence of 250 samples of a clean speech signal. A low-pass filter with filter coeffiecients $0.5, 1, 1, 1$ and 0.5 is applied to the speech example. A noise vector ϵ is also generated from matlab through the bulit-in m-file randn.m

$$\epsilon = 0.001 \times \text{randn}(250, 1)$$

We cannot solve these problems without making assumptions. In view of that we make the following assumptions without going into the details surrounding the theoretical concepts

1. The matrix K has full rank.
2. K is ill-conditioned with no significant gap in the singular value spectrum. (Problems arise when the singular values d_i are within the range $0 < d_i << d_1$).
3. The true data g is corrupted with noise.
4. The discretization error caused by approximating the continuous operator is much smaller than the noise.
5. The system satisfies the *discrete Picard conditions* which we informally deduce and state in subsection (2.2.4).

[3] The symbol w here is different from the weights given in equation 2.4. Moreover w used in Chapter 1 is also different from both. We should however to note the differences. The same symbol was use due to shortage of notations.

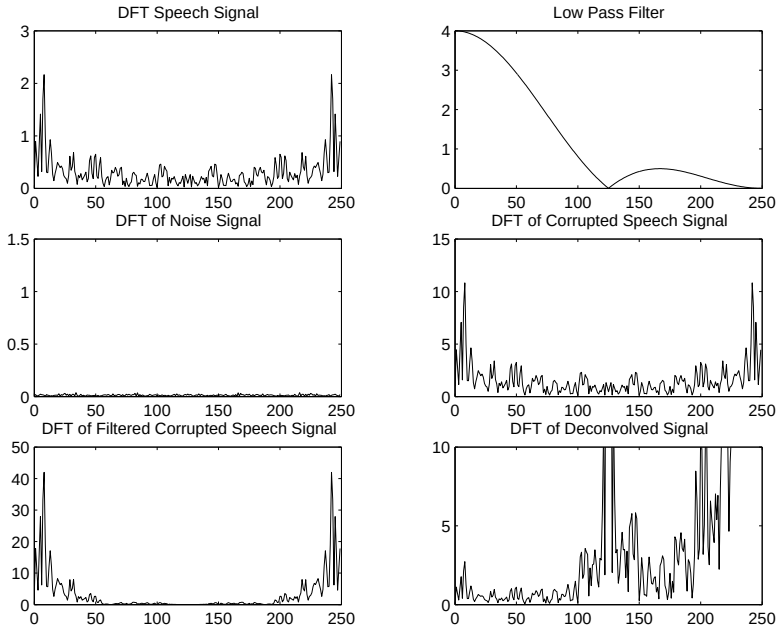

Figure 2.3: Power spectra of the various signals in a low-pass filtering with 5 filter coefficients. It illustrates the power spectra of the clean speech signal g, corrupted signal $\tilde{g}$, filtered signal f and the deconvolved signal $\tilde{f}$ obtained from equation (2.18). As seen from above, the high frequency components of the convolved signal are perturbed greatly especially around the zeros of the low-pass filter. The effect renders inverse operation of DFT meaningless. That is $\tilde{f} = IDFT[\tilde{g}(w) \oslash K(w)]$

2.2.4 Discrete Picard Condition

From (2.3)

$$\tilde{g} \quad = \quad Kf + \epsilon \tag{2.19}$$

we have

$$f \quad = \quad K^{-1}\tilde{g} - K^{-1}\epsilon \tag{2.20}$$

The SVD of the naive solution is

$$\tilde{f} = K^{-1}\tilde{g} = \sum_{i=1}^{n} \Big(\frac{\mathbf{u}_i^T \tilde{g}}{d_i}\Big)\mathbf{v}_i \tag{2.21}$$

The corresponding noise denoted f^ϵ is

$$f^\epsilon = \sum_{i=1}^{n} \Big(\frac{\mathbf{u}_i^T \epsilon}{d_i}\Big)\mathbf{v}_i \tag{2.22}$$

The solution f follows from

$$\begin{aligned} f &= \tilde{f} - f^\epsilon \\ &= \sum_{i=1}^{n} \frac{\mathbf{u}_i^T(\tilde{g} - \epsilon)}{d_i}\mathbf{v}_i \end{aligned} \tag{2.23}$$

The singular values d_i of K in equation (2.23) must neither approach zero nor be zero or else $\| f\|_2^2$ will be large or undefined. A consequence of this leads to loss of much of the information about the system or it can happen that no information will be gained. Especially, for normalized singular values d_i (between 0 and 1) we do not expect d_i to decay faster than either $\mathbf{u}_i^T(\tilde{g} - \epsilon)$ or $\mathbf{u}_i^T\tilde{g}$, otherwise in the neighbourhood where either $d_i \to 0$ (or both $\mathbf{u}_i^T\tilde{g} \to 0$ and $d_i \to 0$), the expression

$$\frac{\mathbf{u}_i^T(\tilde{g} - \epsilon)}{d_i} \to \infty \quad \text{or} \quad \frac{\mathbf{u}_i^T\tilde{g}}{d_i} \to \infty \quad \text{for} \quad i \to \infty$$

Definition
A system is said to satisfy the discrete Picard condition if for large enough values of the discretization parameter n, the sequence of true data values $\{\mathbf{u}_i^T(\tilde{g} - \epsilon)\}$ *goes to zero faster than the sequence of singular values* $\{d_i\}$. *Thus for terms greater than or equal to some parameter k,* $\mathbf{u}_i^T(\tilde{g} - \epsilon) \approx 0$

Figures (2.4), (2.5) and (2.6) are picard plots of the Gravity Surveying Model problem of Figure (1.1) with additive noise $\sigma_\epsilon^2 = 0$, $\sigma_\epsilon^2 = 10^{-6}$ and $\sigma_\epsilon^2 = 10^{-3}$ respectively. They

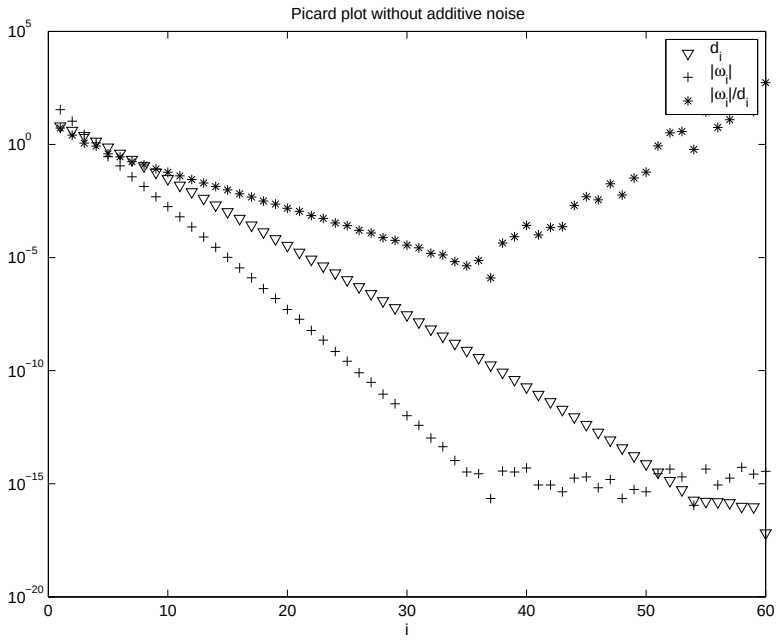

Figure 2.4: Singular values d_i of matrix K and the computed quantities $|\mathbf{u}_i^T\tilde{g}|$ and $|\frac{\mathbf{u}_i^T\tilde{g}}{d_i}|$ of the Gravity Surveying Model problem. The noise contribution comes from only rounding and discretization errors.

show plots of d_i , $[\omega_i = |\mathbf{u}_i^T\tilde{g}|\,]$ and $[\omega_i/d_i = |\frac{\mathbf{u}_i^T\tilde{g}}{d_i}|\,]$ versus i. The quantity $|\mathbf{u}_i^T\tilde{g}|$ in each case decays faster than d_i until it reaches a level set by the machine's precision. At locations where $|\mathbf{u}_i^T\tilde{g}|$ levels off, the quantity $|\frac{\mathbf{u}_i^T\tilde{g}}{d_i}|$ begins to increase steeply and in the neighbourhood where both $|\mathbf{u}_i^T\tilde{g}|$ and d_i approach zero, the ratio $|\frac{\mathbf{u}_i^T g}{d_i}|$ becomes larger and larger. Figure (2.4) illustrates how the naive solution

$$\tilde{f} = \sum_{i=1}^{n} \frac{\mathbf{u}_i^T\tilde{g}}{d_i}\mathbf{v}_i$$

is completely dominated by large values of $|\frac{\mathbf{u}_i^T \tilde{g}}{d_i}|$. They come from components corresponding to the smallest singular values. This explains why the plot "plot of fnaive" $\tilde{f}$ in Figure (2.1) appears as a high oscillatory solution. The norm of $\tilde{f}$ is 6.1×10^{15}.

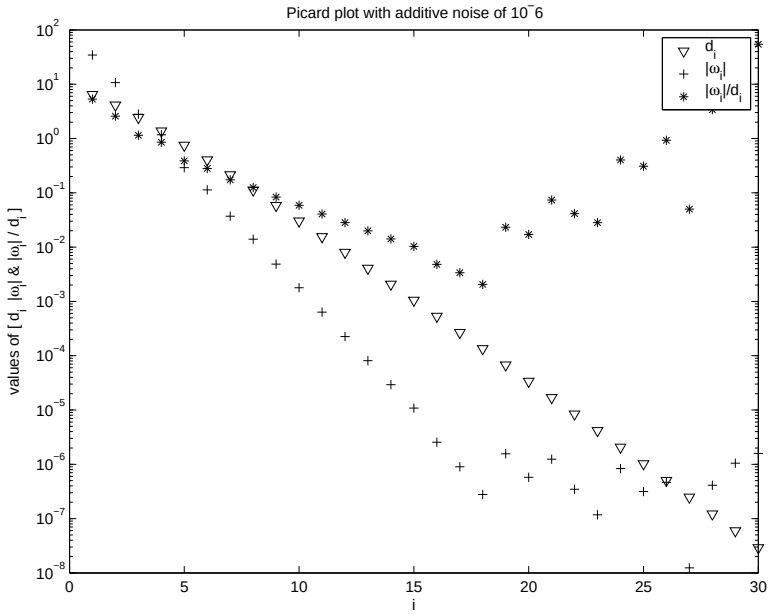

Figure 2.5: Picard plot with an additive noise of 10^{-6}. The acute angle between the decaying regions of $|\mathbf{u}_i^T \tilde{g}|$ and d_i shrinks considerably as compared to Figure (2.4), thereby causing the quantity $|\mathbf{u}_i^T \tilde{g}|$ to level off at lower indices i than without additive noise.

Also, Figures (2.5) and (2.6) illustrate the same problem but with additive noise σ_ϵ^2 of magnitudes 10^{-6} and 10^{-3}. Their respective noise vectors $\epsilon_{10^{-6}}$ and $\epsilon_{10^{-3}}$ are normally distributed with zero mean and variance $\sigma_{10^{-6}}^2$ and $\sigma_{10^{-3}}^2$. It can be seen that the greater the magnitude of the additive noise term the more information we lose. That is, the quantity $|\mathbf{u}_i^T \tilde{g}|$ starts to level off at a much lower indices of the index i. The larger the smoothing effect of the function K, the faster d_i decay. Moreover, small singular values lead to solutions which fit the data well but result in large energy. In an act of trying to find a stable meaningful solution pushes us to employ regularization schemes. We now devote the rest of this Chapter to Numerical Regularization and it is the main subject of this thesis.

2.3 Numerical Approach to Regularization

These are algorithmic techniques which can be used for stabilizing solutions so that they become less sensitive to perturbations. Such algorithms are called Regularization Algorithms. The method encourages smoother functional mappings by adding a penalty term, say Φ to the residual error function r_ϵ to give an implicit form;

$$\beth(f) = r_\epsilon + \lambda^2 \Phi(f) \tag{2.24}$$

where $\beth$ is a functional called the *standard residual error function* and the parameter λ^2 controls the effect of the penalty term Φ on the form of solution. It comes in two

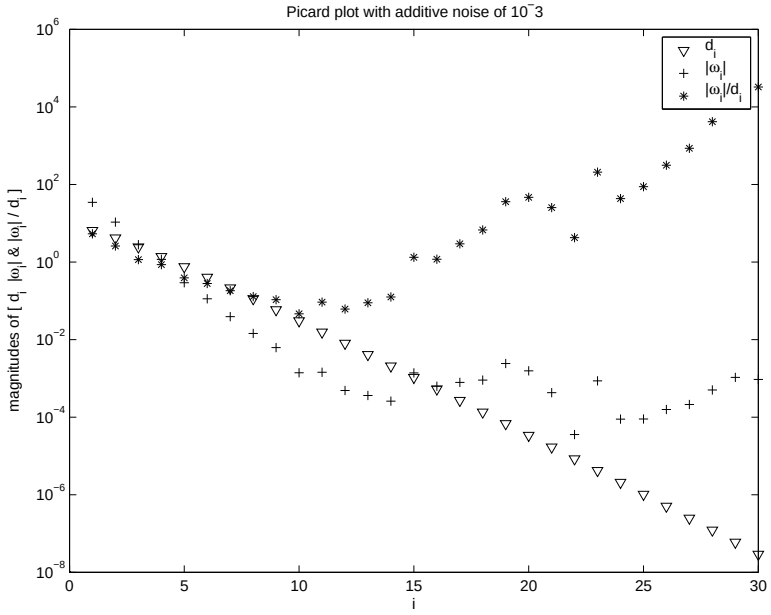

Figure 2.6: Picard plot with an additive noise of 10^{-3}. The acute angle between the decaying regions of $|\mathbf{u}_i^T \tilde{g}|$ and d_i shrinks much more compared to Figures (2.4) and (2.5). The quantity $|\mathbf{u}_i^T \tilde{g}|$ levels off at much lower indices i than the ones shown in Figures (2.4) and (2.4).

flavours; either by (i) truncating the matrix K or (ii) adding a regularizer.
We will first deal with the simplest approach to the smoothness problem called the *Truncated SVD*.

2.3.1 Truncated Singular Value Decomposition

In a much more simple approach, the SVD of the matrix $K \in \mathbb{R}^{n \times n}$ is computed. Thus, we have

$$K \;=\; U\, D\, V^T \;=\; \sum_{i=1}^{n} d_i \mathbf{u}_i \mathbf{v}_i^T \tag{2.25}$$

where $U = U(1:n, 1:n)$ and $n = \text{rank}(K)$. The singular values d_i of the $(n \times n)$ diagonal matrix D are in decreasing order; $d_1 \geq d_2 \geq, ..., d_r \geq, .., d_{n-2} \geq d_{n-1} \geq d_n > 0$.

Given an integer $k \leq r$, we partition the SVD according to

$$K \;=\; (U_k, U_0) \begin{pmatrix} D_k & 0 \\ 0 & D_0 \end{pmatrix} (V_k, V_0)^T$$

where $D_k = \text{diag}(d_1, ..., d_k)$ and $D_0 = \text{diag}(d_{k+1}, ..., d_n)$ are diagonal matrices consisting of the k largest and *(n-k)* smallest singular values respectively. The matrix K_k, defined by

$$K_k \;=\; U_k D_k V_k^T$$

is considered to be an approximation to the original matrix K with a corresponding decrease in rank from n to k. This is the underlying concept of the *truncated SVD*.

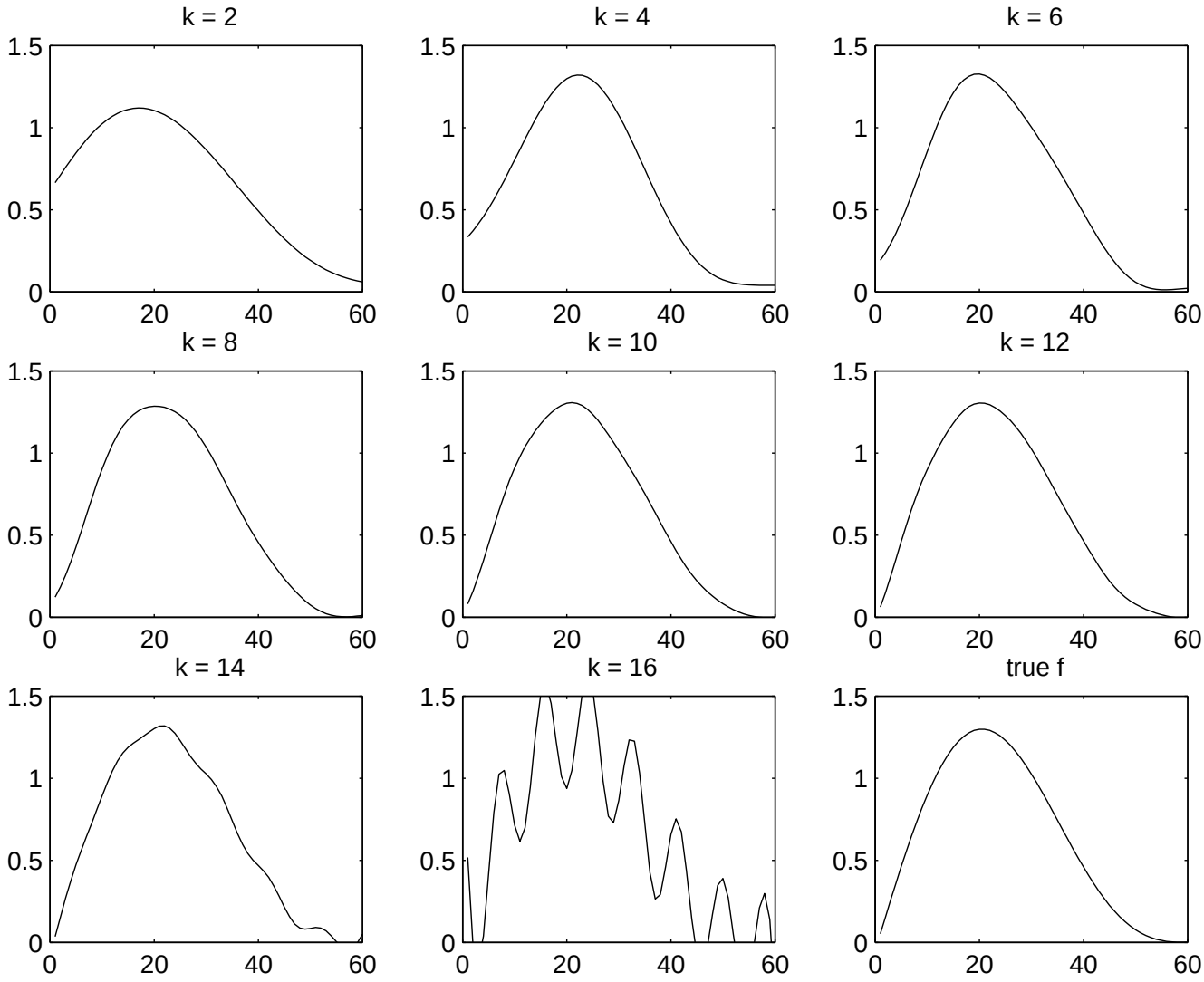

Figure 2.7: The figure shows plots of f_k of the Gravity Survey Model versus i for varying k at constant noise level $\sigma^2 = 10^{-6}$. The exact solution or unpertuurbed solution f is shown at the bottom 'true f '.

Then follows the question anybody would most likely ask about the truncated SVD!

QUESTION : If the parameter k forms the basis in determining a best approximation to our true solution f, How then do we choose an appriopriate k from the triplets $(\mathbf{u}_i, d_i, \mathbf{v}_i)$ in order to capture the most relevant information in K?

The choice of k does not depend on any direct formula(e) in question and therefore not fixed, rather it depends on the particular application. For instance, if in a particular application, the noise level is given in the form of a threshold, say τ then k is chosen so as to make only the first $d_n, d_{n-1}, ..., d_{n-k}$ which are strictly less than τ to be discarded. This leads to a numerical rank deficiency in K (or a singular subspace of K).

Alternatively, we can use the 'Brute force' approach to compute the solution f for each k using the formula of $\tilde{f}$ and changing the summation interval for each choice of k. By running this iteratively at regular steps among all choices of k, we opt for the one that is smoothest in some sense. For each k we have

$$f_k = \sum_{i=1}^{k} \frac{\mathbf{u}_i^T \tilde{g}}{d_i} \mathbf{v}_i \tag{2.26}$$

An illustration of this is demonstrated in Figure (2.7) with the Gravity Example at a noise level $\sigma^2 = 10^{-6}$. It is an alternative procedure which can be used without any restriction on the threshold τ, meanwhile it is equally good enough in abandoning the irrelevant noise components. If K is symmetric, then we have

$$K_k = U_k D_k U_k^T \tag{2.27}$$

The simulation result of the computed solution f_k of the Gravity Model with additive noise, $\sigma^2 = 10^{-6}$ is illustrated in Figure (2.7). The actual plot of the function f is shown at the bottom right corner.

For eack k, we used the formula

$$f_k = \sum_{i=1}^{k} \frac{\mathbf{u}_i^T \tilde{g}}{d_i} \mathbf{v}_i$$

It can be seen that, the smoothness of the solution f_k improves from $k = 2$ to about $k = 12$. At $k \geq 14$, the noise components take over the true solution as a result of large values of their corresponding norm. See how the "sickness" begins to crop up from $k = 14$ and beyond.

2.3.2 Adding a Regularizer

The alternative form of smoothness is governed by how a given function say f, is continuously differentiable with respect to say s. Various forms of Regularizers have been studied in connection with linear models but the one of interest to us here is the class of *Tikhonov Regularizers*, Φ which in general takes the functional form;

$$\Phi(f) = \frac{1}{2} \sum_{k=0}^{R} \int_{\Omega} h_k \Big(\frac{d^k f}{ds^k} \Big)^2 d^k s \tag{2.28}$$

where $s = (s_1, s_2, ..., s_n)$ and $\{h_k \geq 0 \text{ for } k = 0, 1, ..., R-1\}$ are weights such that $h_R > 0$ (Tikhonov and Arsenin, 1977).

The functional Φ could be from a $1-D$ space onto the real line $\mathbb{R}$ (i.e single-input single output) or a higher dimensional space onto $\mathbb{R}$ (i.e multi-input single-output). We view Φ as a functional defined in terms of f with smoothness dependent on the function f. If $R = 1$, then it is obvious that, the derivative operator say L is the identity matrix, I.

Generalized Functional Regularization

From the vector-norm sense, $\tilde{g} - Kf$ and $Lf - f_0$ are vectors which can be of same or different dimensions. The sum of the two residual vectors is given by

$$(\tilde{g} - Kf) + (Lf - f_0)$$

Applying the triangle inequality, we have

$$\|(Lf - f_0) + (\tilde{g} - Kf)\|_2^2 \leq \|Lf - f_0\|_2^2 + \|\tilde{g} - Kf\|_2^2 \tag{2.29}$$

where the Left Hand Side is a lower bound on the right hand side with equality only when the residuals $\tilde{g} - Kf$ and $Lf - f_0$ are at right angles to each other. The norm of either residuals is non-zero, positive and finite;

$$0 < \|Lf - f_0\|_2^2 < \infty \qquad \text{and} \qquad 0 < \|\tilde{g} - Kf\|_2^2 < \infty$$

We try to attain a lower bound on the right hand side of inequality (2.29). This problem can alternatively be identified as a Least Squares minimization problem with quadratic equality constraint which is (more or less) equivalent to the Lagrange multiplier problem of determining a real positive regularization parameter λ_{rls}^2 such that

$$\arg\min \left\| \begin{bmatrix} K \\ \lambda_{rls} L \end{bmatrix} f - \begin{bmatrix} \tilde{g} \\ \lambda_{rls} f_0 \end{bmatrix} \right\|_2 = \arg\min \| \tilde{g} - Kf \|_2^2 + \lambda_{rls}^2 \| Lf - f_0 \|_2^2 \tag{2.30}$$

where the solution to equation (2.30) is the total regularized minimum residual (which is also called the regularized squared error). In a sense, we view λ_{rls}^2 as an indicator of sufficiency of the output $\tilde{g}$ as examples that specify the form of solution $f_{\lambda_{rls}}$. If $\lambda_{rls}^2 = 1$, we have inequality (2.29). We now find the asymptotic behaviour of equation (2.30).

For asymptotes, we write expression (2.30) in the explicit form

$$\min_f \left\{ \lambda_{rls} \left\{ \frac{1}{\lambda_{rls}} \|\tilde{g} - Kf\|_2^2 + \lambda_{rls} \|Lf - f_0\|_2^2 \right\} \right\} \tag{2.31}$$

The limiting case, $\lambda_{rls} \to \infty$;

$$\frac{1}{\lambda_{rls}} \|\tilde{g} - Kf\|_2^2 \to 0 \quad \text{and} \quad \lambda_{rls} \|Lf - f_0\|_2^2 \to \infty$$

implies that the prior smoothness constraint imposed by the differential operator L is by itself sufficient to specify the solution $f_{\lambda_{rls}}$ and it is the same as saying that the output $\tilde{g}$ is unreliable. So

$$\min_f \left\{ \lambda_{rls} \left\{ \frac{1}{\lambda_{rls}} \|\tilde{g} - Kf\|_2^2 + \lambda_{rls} \|Lf - f_0\|_2^2 \right\} \right\} \to \infty \tag{2.32}$$

is said to violate (i) *non-zero residual* of $\frac{1}{\lambda_{rls}} \|\tilde{g} - Kf\|_2^2$ and (ii) a *non-large* value of the total regularized minimum residuals. In this case, the regularized solution $f_{\lambda_{rls}}$ is given by the regularizer alone without taking the actual data into consideration thereby neglecting the information about the data in question. Thus, the solution is said to be independent of the data misfit.

The other limiting case, $\lambda_{rls} \to 0$;

$$\lambda_{rls} \|Lf - f_0\|_2^2 \to 0 \quad \text{and} \quad \frac{1}{\lambda_{rls}} \|\tilde{g} - Kf\|_2^2 \to \infty$$

implies that the problem is unconstrained with the solution $f_{\lambda_{rls}}$ completely determined from the examples. It therefore approaches the Least Squares problem formulation. So, we have

$$\min_f \left\{ \lambda_{rls} \left\{ \frac{1}{\lambda_{rls}} \|\tilde{g} - Kf\|_2^2 + \lambda_{rls} \|Lf - f_0\|_2^2 \right\} \right\} \to \infty \tag{2.33}$$

which also violates a (ii) *non-zero regularized residual* of $\lambda_{rls} \|Lf - f_0\|_2^2$ and (ii) non-large total minimum residual. In this case, the regularized solution $f_{\lambda_{rls}}$ is given by the residuals from the data alone which is the same as saying that the solution is independent of the reason for adding a regularizer.

2.3.3 Tikhonov Functional Regularization

A combination of Tikhonov's Regularizer and S. Twomey's reformulation of Phillip's expression for a regularized f in normal equations settings is referred to as "Regularized Normal Equations" with the purpose stated as follows:

To find the function $f_{\lambda_{rls}}$ that minimizes the Tikhonov functional $\rho(f)$

$$\rho(f) = r_\epsilon(f) + \lambda_{rls}^2 \Phi(f) \tag{2.34}$$

where $r_\epsilon(f)$ is the standard error term, $\Phi(f)$ is the regularizing term and λ_{rls} is the numerical regularization parameter.

The Numerical Framework functional regularizer Φ is of the form

$$\Phi(f) = \frac{1}{2} \| L f \|_2^2 \tag{2.35}$$

with $f_0 = 0$ if no a priori estimate of f is given.

The solution which we denote $f_{\lambda_{rls}}$ that minimizes a weighted combination of the residual norm [4] and the added smoothness constraint is

$$f_{\lambda_{rls}} = \arg\min\left\{ \|\tilde{g} - Kf\|_2^2 + \lambda_{rls}^2 \|Lf\|_2^2 \right\} \tag{2.36}$$

where L is a discrete derivative operator of some order. By taking partial derivatives with respect to f of the expression in the curly brackets of equation (2.36) and setting to zero, we write

$$\nabla_f \left\{ \|\tilde{g} - Kf\|_2^2 + \lambda_{rls}^2 \|Lf\|_2^2 \right\} = 0 \tag{2.37}$$

The solution to equation (2.37) gives the regularized normal equations

$$\left(K^T K + \lambda_{rls}^2 L^T L\right) f_{\lambda_{rls}} = K^T \tilde{g} \tag{2.38}$$

If $L = I$, equation (2.38) is said to be in its standard Tikhonov's form

$$K^T K f_{\lambda_{rls}} + \lambda_{rls}^2 f_{\lambda_{rls}} = K^T \tilde{g} \tag{2.39}$$

In this thesis, we choose our derivative operator L to be the identity matrix I. Hence the standard Tikhonov's solution is

$$f_{\lambda_{rls}} = \left(K^T K + \lambda_{rls}^2 I\right)^{-1} K^T \tilde{g} \tag{2.40}$$

Figure (2.8) is the Tikhonov's solution $f_{\lambda_{rls}}$ to the Gravity model problem of Figure (1.1) at $n = 60$, $d = 0.25$ and for different values of the numerical regularization parameter λ_{rls}^2 with additive noise $\sigma^2 = 10^{-6}$. We can see that the best values of λ_{rls}^2 which give good approximations to the true function is neither too big nor too small. Values of λ_{rls}^2 that are too small tend to overfit whereas values that are too large also give bias estimates. The choice of λ_{rls}^2 is therefore a compromise between the two extremes.

[4] standard error term is the same as the residual norm in most literature.

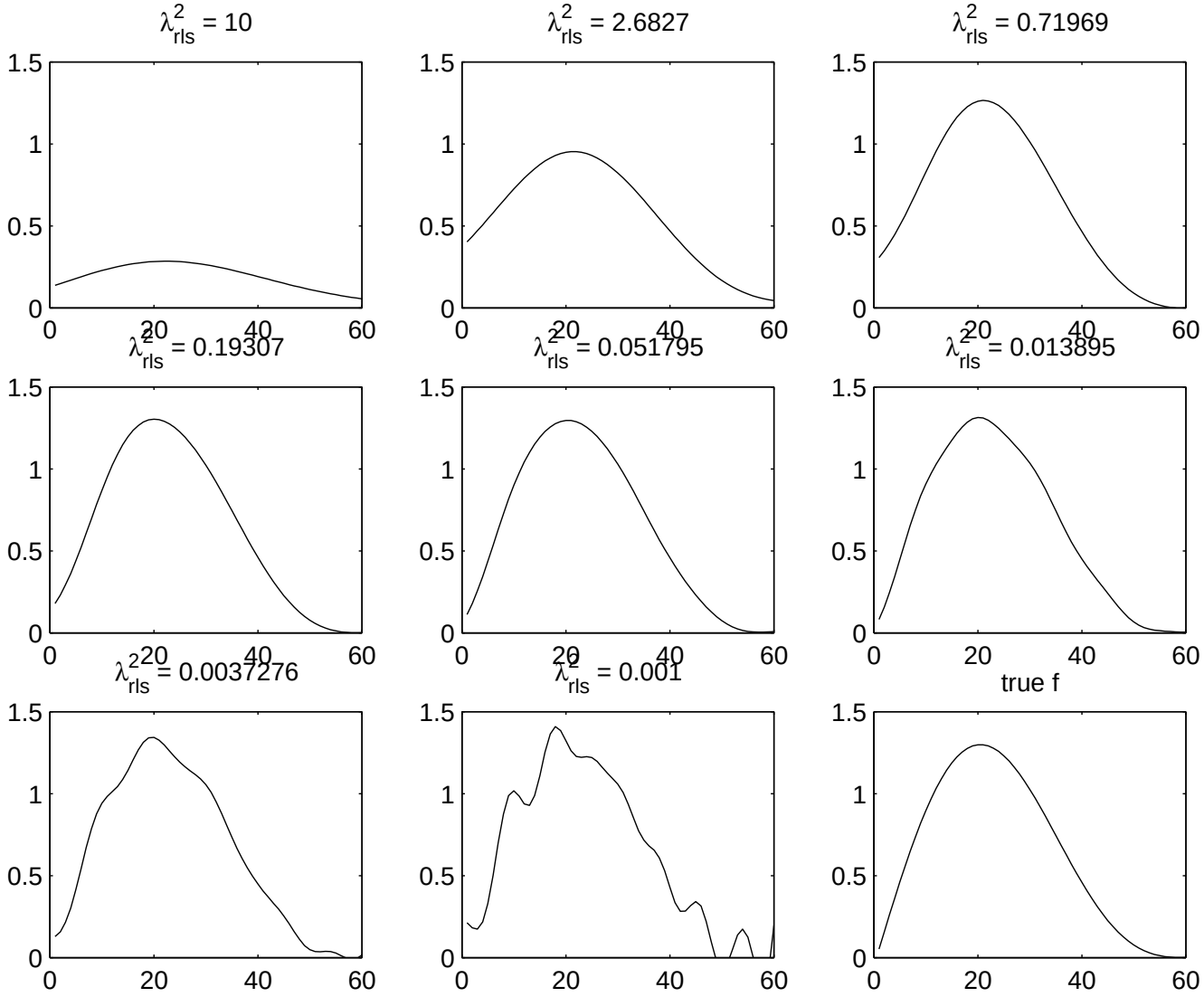

Figure 2.8: Tikhonov's solution $f_{\lambda_{rls}}$ to the Gravity surveying model problem for different choices of the regularization parameter λ_{rls} with noise $\sigma^2 = 10^{-6}$ for λ^2_{rls} in the range $10^{-3} - 10$ plus the exact solution.

CHAPTER 3

Stochastic Modelling

3.1 Filtering and Linear System Theory

Introduction

In this chapter, we develop some results which are required for the solution of the estimation problem under consideration. We will restrict ourselves to some properties of conditional distributions of Gaussian random variables and then give a geometric interpretation.

Like we begun in Chapter 2 with "Why Least Squares and Regularization" in the Numerical Framework, we do the same in this Chapter by looking into Least Squares from a Statistical viewpoint and further move on Statistical Regularization using both the Maximum a Priori (MAP) and Maximum Likelihood (ML) principles to the same Gravity problem. We will appeal to four theorems on multivariable Gaussian distributions. We will then see that these theorems have geometric interpretations with a strong intuitive appeal.

In section (3.2), we formulate the problem of Filtering and Estimation for Discrete Time Systems, state the Deconvolution problem in a Statistical environment and finish up with Least Squares. Section (3.3) deals with the Statistical Approach to Regularization; we explore exact approaches in relation to maximum apriori (MAP) method and marginalization over continuous variables of the Maximum Likelihood principle by performing integration. We follow up with the EM-Algorithmic principle where we will round up with the difference between MAP and ML.

We shall continue to work under normality conditions.

3.2 Formulation of Filtering and Estimation Problems for Discrete-Time Systems

We consider $\{g(s),\, s \in T\}$ and $\{\epsilon(s),\, s \in T\}$ as two real stochastic processes which are signal and noise respectively. We assume that the observation (output) or measurement $\tilde{g}(s)$ are given by

$$\tilde{g}(s) \;=\; g(s) \;+\; \epsilon(s) \tag{3.1}$$

From equation (3.1), we mean that at time s, we have obtained a realization $\{\tilde{g}(\tau), \quad s_k < \tau < s\}$ of the measured variable. Based on this realization, the best estimate of the value of the signal at time s_k can be determined. Here s_k could be one or more of the following:
(a) $s_k < s$ (which leads to a smoothing problem).
(b) $s_k = s$ (which leads to a filtering problem).
(c) $s_k > s$ (which leads to a prediction problem).

Define a notation for a realization, say $\tilde{g}$, by

$$\tilde{g}(s) = (\tilde{g}_{s_1}^T, \tilde{g}_{s_2}^T, \tilde{g}_{s_3}^T, ..., \tilde{g}_{s_n}^T) \tag{3.2}$$

From equation (3.2) we have indicated explicitly that $\tilde{g}$ depends on s. We let $\tilde{g} \in \tilde{\mathbf{g}}$ and $g \in \mathbf{g}$. An estimator (filter, predictor, interpolator) is a function which maps $\tilde{\mathbf{g}}$ *into* $\mathbf{g}$. The value of this function for a particular measurement or observation $\tilde{g}$ is called an *estimate* g. In this Chapter, we will describe filtering problems in relation to (3.1) by specifying the following:
(i) the signal and noise processes.
(ii) the criterion which defines the best estimate.
(iii) the restriction on the admissible estimators.

3.2.1 The Inverse (Deconvolution) Problem

The signal and noise processes are characterized by covariance functions through a linear equation of the form:

$$\tilde{g} = K f + \epsilon \tag{3.3}$$

where $g = Kf$ and ϵ is a sequence of independent Gaussian random variables.

From above, we ask ourselves the question below!

Question : From a realization of the output $\tilde{g}(\tau)$, $s_1 \leq \tau \leq s$. How or by what means can we estimate the input vector f in (3.3)?

This forms an estimation problem. The necessary skills we need to acquire for this problem are discussed in subsections (3.2.2) through (3.3.4).

3.2.2 Statistical Modelling of the Loss Function

The statistical information which the observations give about the stochastic variables $g(s)$ is contained in the conditional distribution

$$p\{g_{s_k} \leq \sigma | \tilde{g}(\tau) = \varphi(\tau) : \quad t_0 \leq \tau \leq t\} = F(\sigma|\varphi) \tag{3.4}$$

In the left (most) hand side of equation (3.4), the parameter σ symbolizes a deviation conditioned on the output value $\tilde{g}(\tau) = \varphi(\tau)$. The corresponding density of the distribution (3.4) is denoted by $p(\sigma|\varphi)$. We define a loss function l, which is a real function with properties $l \geq 0$, $l(\delta) = l(-\delta)$ and l non-decreasing for $\delta \geq 0$. The loss function is then a stochastic variable $l(\tilde{g} - g)$ with the *best estimate* g chosen to be the one that minimizes the average loss $\langle\, l(\tilde{g} - g)\,\rangle$

Theorem 1

This is based on the assumption that the conditional distribution of g given $\tilde{g} = \varphi$ has a density function which is symmetric around the conditional mean $\mu = \int \sigma\, p(\sigma|\varphi)\, \mathrm{d}\sigma$ (where σ is the standard deviation) and non-decreasing for $(\sigma \geq \mu)$. The loss function l is considered to be symmetric and non-decreasing for positive arguments. The *best estimate* is then given by the conditional mean

$$g \;=\; g(\varphi) \;=\; \langle\, \tilde{g} \,|\, \varphi \,\rangle \;=\; \int \sigma\, p(\sigma|\varphi)\, \mathrm{d}\sigma \tag{3.5}$$

The proof is based on an elementary lemma on real function. For proof and more on this see [43].

3.2.3 Multivariate Gaussian Distribution Theorems

The probability density function of a normal n-dimensional variable with mean $\mu_{\tilde{g}}$ and covariance $R_{\tilde{g}}$ is given by

$$p(\tilde{g}) \;=\; (2\pi)^{-n/2}\, \det(\, R_{\tilde{g}}\,)^{-1/2}\, \exp\, -\frac{1}{2}\{(\tilde{g} - \mu_{\tilde{g}})^T R_{\tilde{g}}^{-1} (\tilde{g} - \mu_{\tilde{g}})\} \tag{3.6}$$

where we have made an assumption that the covariance matrix $R_{\tilde{g}}$ is non-singular and

$$\det(\, R_{\tilde{g}}\,) \;=\; \mid R_{\tilde{g}} \mid$$

is the determinant of R.

Theorem 2

If $\tilde{f}$ and $\tilde{g}$ are both $n \times 1$ vectors and we make an assumption that $\begin{bmatrix} \tilde{f} \\ \tilde{g} \end{bmatrix}$ is Gaussian with mean $\begin{bmatrix} \mu_{\tilde{f}} \\ \mu_{\tilde{g}} \end{bmatrix}$ and covariance $R = \begin{bmatrix} R_{\tilde{f}} & R_{\tilde{f}\tilde{g}} \\ R_{\tilde{g}\tilde{f}} & R_{\tilde{g}} \end{bmatrix}$ then the vector ς given by

$$\varsigma \;=\; \tilde{f} \;-\; \mu_{\tilde{f}} \;-\; R_{\tilde{f}\tilde{g}} R_{\tilde{g}}^{-1} (\tilde{g} - \mu_{\tilde{g}}) \tag{3.7}$$

is independent of $\tilde{g}$ has zero mean and covariance

$$R_{\varsigma} \;=\; R_{\tilde{f}} \;-\; R_{\tilde{f}\tilde{g}} R_{\tilde{g}}^{-1} R_{\tilde{g}\tilde{f}} \tag{3.8}$$

For proof and more on this see [43].

Theorem 3

If $\tilde{f}$ and $\tilde{g}$ are two vectors which are jointly Gaussian, then the conditional distribution of $\tilde{f}$ given $\tilde{g}$ is normal with mean

$$\langle\, \tilde{f} \mid \tilde{g} \,\rangle \;=\; \mu_{\tilde{f}} \;+\; R_{\tilde{f}\tilde{g}} R_{\tilde{g}}^{-1} (\tilde{g} - \mu_{\tilde{g}}) \tag{3.9}$$

and covariance

$$\Big\langle\, \Big\{ \tilde{f} \;-\; \langle\, \tilde{f} \,|\, \tilde{g} \,\rangle \Big\} \Big\{ \tilde{f} \;-\; \langle\, \tilde{f} \,|\, \tilde{g} \,\rangle \Big\}^T \;\Big|\; \tilde{g} \,\Big\rangle \;=\; R_{\tilde{f}} \;-\; R_{\tilde{f}\tilde{g}} R_{\tilde{g}}^{-1} R_{\tilde{g}\tilde{f}} \;=\; R_{\varsigma} \tag{3.10}$$

The stochastic variables $\tilde{g}$ and $[\tilde{f} - \langle \tilde{f} \,|\, \tilde{g} \rangle]$ are independent.

For proof and more on this see [43].

Theorem 4

(a) Linear functions (and therefore conditional expectations) on a Gaussian random process are Gaussian random variables.
(b) Orthogonal Gaussian random variables are independent.
(c) Given any random process with means $\langle \tilde{g}(s) \rangle$ and covariances $\langle \tilde{g}(s)\, \tilde{g}(t) \rangle$, there exists a unique Gaussian random process with the same means and covariances.

Interpretation

The state estimation Theorem (1) implies that the best estimate is given by the conditional mean: i.e

$$\langle \tilde{f} \,|\, \tilde{g} \rangle = \mu_{\tilde{f}} + R_{\tilde{f}\tilde{g}} R_{\tilde{g}}^{-1} (\tilde{g} - \mu_{\tilde{g}})$$

and the estimation error has the covariance

$$\langle \varsigma \varsigma^T \,|\, \tilde{g} \rangle = R_{\tilde{f}} - R_{\tilde{f}\tilde{g}} R_{\tilde{g}}^{-1} R_{\tilde{g}\tilde{f}} \tag{3.11}$$

It further implies from Theorem (2) and Theorem (3) that, the estimation error

$$\varsigma = \tilde{f} - \langle \tilde{f} \,|\, \tilde{g} \rangle = \tilde{f} - \mu_{\tilde{f}} - R_{\tilde{f}\tilde{g}} R_{\tilde{g}}^{-1} (\tilde{g} - \mu_{\tilde{g}}) \tag{3.12}$$

is independent of $\tilde{g}$.

3.2.4 Geometric Interpretation

The above multivariable Gaussian distribution theorems gives a strong intuitive appeal when they are illustrated geometrically. See Figure (3.1). For simplicity, we illustrate this by assumming that both variables, $\mu_{\tilde{g}}$ and $\mu_{\tilde{f}}$ have zero mean (i.e $\mu_{\tilde{f}} = \mu_{\tilde{g}} = 0$). We then represent the variables $\tilde{f}$ and $\tilde{g}$ as elements in the *Euclidean Space* with scalar product defined by

$$(\tilde{f}\,,\, \tilde{g}) = \langle \tilde{f}^T \tilde{g} \rangle = \mathrm{cov}(\tilde{f} - 0\,,\, \tilde{g} - 0) = \mathrm{cov}(\tilde{f}\,,\, \tilde{g}) \tag{3.13}$$

The norm is given by

$$\|\tilde{f}\|_2^2 = (\tilde{f}\,,\, \tilde{f}) = \langle \tilde{f}^T \tilde{f} \rangle \tag{3.14}$$

Define the two lines l_1 and l_2 which intersect at the origin. The angle $\breve{a}$ between the lines is given by

$$\cos \breve{a} = \frac{\langle \tilde{f}^T \tilde{g} \rangle}{\|\tilde{f}\|_2 \cdot \|\tilde{g}\|_2} = \frac{\mathrm{cov}(\tilde{f}\,,\, \tilde{g})}{\|\tilde{f}\|_2 \cdot \|\tilde{g}\|_2} \tag{3.15}$$

The stochastic variable $\tilde{f}$ is represented as a vector along l_1 with the length $\|\tilde{f}\|_2 = \sqrt{\langle \tilde{f}^2 \rangle}$ and the stochastic variable $\tilde{g}$ is represented by a vector along l_2 with length $\|\tilde{g}\|_2 = \sqrt{\langle \tilde{g}^2 \rangle}$.

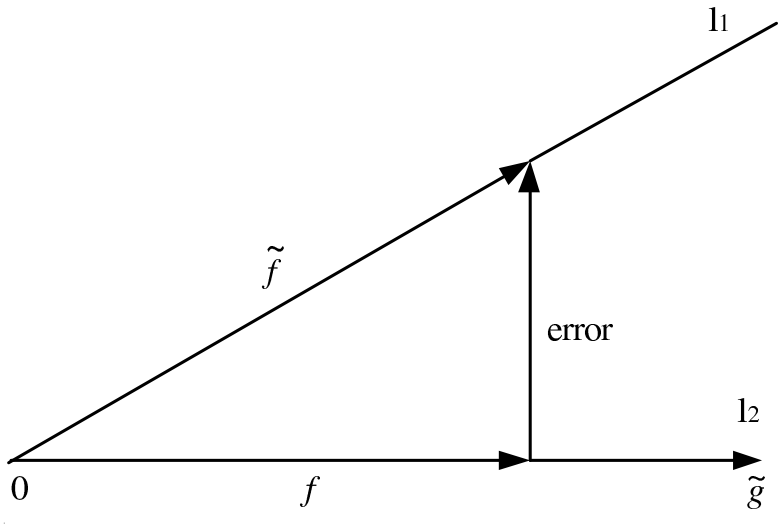

Figure 3.1: *Geometric illustration of the conditional mean values of normal random variables. The conditional mean $f = \langle \tilde{f} \,|\, \tilde{g} \rangle$ is represented by the projection of $\tilde{f}$ on $\tilde{g}$.*

Recalling the assumption that $\tilde{f}$ has zero mean, we find that *Theorem 2* implies that the stochastic variable defined by

$$\varsigma = \tilde{f} - R_{\tilde{f}\tilde{g}} R_{\tilde{g}}^{-1} \tilde{g} \tag{3.16}$$

is independent of $\tilde{g}$. Hence

$$(\varsigma\,,\, \tilde{g}) = \langle \varsigma^T \tilde{g} \rangle = 0 \tag{3.17}$$

Theorem (2) thus implies that ς is orthogonal to $\tilde{g}$ and that the norm of ς is

$$\| \varsigma \|_2^2 = R_{\tilde{f}} - R_{\tilde{f}\tilde{g}} R_{\tilde{g}}^{-1} R_{\tilde{f}\tilde{g}} = \| \tilde{f} \|_2^2 - \frac{(\tilde{f}\,,\, \tilde{g})^2}{\| \tilde{g} \|_2^2}$$

The projection of $\tilde{f}$ on $\tilde{g}$ is

$$\Big(\tilde{f}\,,\, \frac{\tilde{g}}{\| \tilde{g} \|_2} \Big) \frac{\tilde{g}}{\| \tilde{g} \|_2} = \frac{(\tilde{f}\,,\, \tilde{g}) \tilde{g}}{\| \tilde{g} \|_2^2} = R_{\tilde{f}\tilde{g}} R_{\tilde{g}}^{-1} \tilde{g} = \langle \tilde{f} \,|\, \tilde{g} \rangle \tag{3.18}$$

where the equality in equation (3.17) follows from equations (3.13) and (3.14), and the last equality follows from *Theorem 3*.

The variable $\tilde{f} - \varsigma = R_{\tilde{f}\tilde{g}} R_{\tilde{g}}^{-1} \tilde{g} = \langle \tilde{f} \,|\, \tilde{g} \rangle$ equals the best mean estimate of $\tilde{f}$ based on $\tilde{g}$ and should be interpreted geometrically as the projection of $\tilde{f}$ on $\tilde{g}$.
[1]

3.3 Statistical Approach to Regularization

Before we proceed we will use a different notation for the covariance matrix R. It shall be replaced by Σ.

[1] The symbol ς used on this page is equivalent to *error* vector in Figure (3.1).

3.3.1 Maximum Apriori Function and the Regularized Precision Matrix

The linear model is our "old friend":

$$\begin{aligned} \tilde{g} &= g + \epsilon \\ &= Kf + \epsilon \end{aligned} \tag{3.19}$$

where $\tilde{g}$, K, f and ϵ represent the same notations used previously.

We now state our statistical model.

The conditional density (or noise model) is

$$p(\tilde{g}|f,\sigma^2) = C_l(\sigma)\exp - \Big\{\frac{1}{2\sigma^2}\Big\| \tilde{g} - Kf \Big\|_2^2\Big\} \tag{3.20}$$

where $C_l(\sigma)$ is a normalization factor given by

$$\Big\{\int \exp - \frac{1}{2\sigma^2}\Big\| \tilde{g} - Kf \Big\|_2^2 \,\mathrm{d}f\Big\}^{-1}$$

is equivalent to the probability density of the noise contributions ϵ, of zero mean and covariance σ^2:

$$p(\epsilon|\sigma^2) = C_l(\sigma)\exp\Big[-\frac{1}{2\sigma^2}\| \epsilon \|_2^2\Big] \tag{3.21}$$

In Numerical Regularization, we chose a standard quadratic functional Φ with a regularization parameter λ_{rls}^2 by setting the derivative operator $L = I$;

$$\lambda_{rls}^2\,\Phi(f) = \frac{1}{2}\,\lambda_{rls}^2\,\|f\|_2^2$$

We repeat it here by introducing a similar regularizer which we will define as the prior probability in the 'Statistical Regularization' Framework:

$$p(f|\lambda_{ml}^2) = C_p(\lambda_{ml})\exp - \Big[\frac{\lambda_{ml}^2}{2}\| f \|_2^2\Big] \tag{3.22}$$

where $C_p(\lambda_{ml})$ is a normalization factor given by

$$\Big\{\int \exp - \frac{\lambda_{ml}^2}{2}\| f \|_2^2 \,\mathrm{d}f\Big\}^{-1}$$

We wish to compute an estimate for f *given* $\tilde{g}$, σ^2 and the *prior* $p(f|\lambda_{ml}^2)$. This is a standard procedure by applying Bayes' Rule.

From Bayes' Rule, the posterior probability of f is

$$p(f\,|\,\tilde{g},\sigma^2,\lambda_{ml}^2) = \frac{p(\tilde{g}|f,\sigma^2)\,p(f|\lambda_{ml}^2)}{p(\tilde{g}|\sigma^2,\lambda_{ml}^2)} \tag{3.23}$$

where the denominator on the right hand side of equation (3.23) is also a normalization factor defined by

$$p(\tilde{g}|\sigma^2,\lambda_{ml}^2) = \int p(\tilde{g}|f,\sigma^2)\,p(f|\lambda_{ml}^2)\;\mathrm{d}f$$

and the densities $p(\tilde{g}|f,\sigma^2)$ and $p(f|\lambda_{ml}^2)$ are their respective likelihood function and prior probability.

Substituting equations (3.21) and (3.22) into equation (3.23) and taking the logarithm of both sides gives

$$\log p(f \,|\, \tilde{g}, \sigma^2, \lambda_{ml}^2) \;=\; -\frac{1}{2}\left\{ \frac{1}{\sigma^2}\|\tilde{g} - Kf\|^2 \;+\; \lambda_{ml}^2 \| f \|^2 \right\} \;+\; \kappa \tag{3.24}$$

where κ is a constant defined by

$$\kappa \;=\; \log\, C_l(\sigma)\, C_p(\lambda_{ml})\, [p(\tilde{g} \,|\, \sigma^2, \lambda_{ml}^2)]^{-1}$$

We take the first partial derivatives of equation (3.24) with respect to f and solve for the zeros of f to obtain the maximum apriori (MAP) estimate, which we denote by $f_{map_{\lambda,\sigma}}$. Thus,

$$\begin{aligned} \nabla_f \log p(f \,|\, \tilde{g}, \sigma^2, \lambda_{ml}^2) &= \frac{1}{\sigma^2}\, K^T\, (\tilde{g} - Kf) \;-\; \lambda_{ml}^2 f \\ &= 0 \end{aligned} \tag{3.25}$$

[2] It is straight forward to write the zeros of equation (3.25) as

$$(K^T K \;+\; \sigma^2\, \lambda_{ml}^2 I)\, f_{map_{\lambda,\sigma}} \;=\; K^T\, \tilde{g} \tag{3.26}$$

which must be viewed as *Normal Equations in the Statistical Framework sense.* This equation is of the same form as Equation (2.40) of Tikhonov's Regularization in the Numerical Framework. The parameter λ_{ml} of the solution

$$f_{map_{\lambda,\sigma}} \;=\; (K^T K \;+\; \sigma^2 \lambda_{ml}^2 I)^{-1}\, K^T \tilde{g} \tag{3.27}$$

should be considered as provision of a *non-zero* parameter which makes inversion of the matrix $(K^T K \;+\; \sigma^2 \lambda_{ml}^2 I)$ possible.

A comparison of equation (3.26) with equation (2.38) for which $L^T L = I$, shows that there is a relation between the numerical regularization parameter λ_{rls} and the statistical regularization parameter λ_{ml}. The relation is

$$\lambda_{rls}^2 \;=\; (\sigma\, \lambda_{ml})^2 \tag{3.28}$$

We finally take the second order partial derivatives of the log-posterior with respect to f to obtain the curvature information. We denote a negation of the curvature information by $J(\lambda_{ml}, \sigma)$:

$$\begin{aligned} J(\lambda_{ml}, \sigma) &= -\nabla_f^2 \log p(f|\tilde{g}, \sigma^2, \lambda_{ml}^2) \\ &= \frac{K^T K}{\sigma^2} \;+\; \lambda_{ml}^2\, I \end{aligned} \tag{3.29}$$

[2] The maximum $f_{map_{\lambda,\sigma}}$ must lie on the stationary point satisfying:

$$\nabla_f \log p(f|K, \tilde{g}, \sigma^2, \lambda_{ml}^2) \;=\; 0$$

It can be seen from equation (3.29) that, the Matrix J is independent of f. Hence, the expectation of Matrix J with respect to the distribution of f gives

$$\begin{aligned} \langle\, J(\lambda_{ml}, \sigma)\,\rangle &= J(\lambda_{ml}, \sigma) \\ &= \frac{K^T K}{\sigma^2} + \lambda_{ml}^2 I \end{aligned} \tag{3.30}$$

From now through the end of this thesis, we will let the notation Σ_f represent $\langle\, J(\lambda_{ml}, \sigma)\,\rangle$

$$\Sigma_f = \frac{K^T K}{\sigma^2} + \lambda_{ml}^2 I \tag{3.31}$$

The corresponding inverse matrix also called the Precision Matrix is

$$\Sigma_f^{-1} = \Big(\frac{K^T K}{\sigma^2} + \lambda_{ml}^2 I\Big)^{-1} \tag{3.32}$$

The use of Bayes' Rule exploits the capabilities of taking prior information into account. It incorporates and maps 'the event space' of our subjective beliefs onto the space $\mathbb{R}$ (of real numbers) by expressing the 'degree of belief' as 'probability'. This is what we earlier referred to it as *prior probability* in equation (3.22).

Furthermore, the stochastic variable $\tilde{g}$ can be characterized by specifying its finite dimensional distribution $p\,(\tilde{g})$. With the first and second moments of $p\,(\tilde{g})$ in hand, we can (partially) answer all probabilistic questions about the joint probability density function of $\tilde{g}$ and f. This calls for a need to express the two moments in terms of the mean and variance-covariance.

The standard deviation (which is the square root of the variance) is a measure that is used to determine how far we are from our estimate, $f_{map_{\lambda,\sigma}}$. The two moments when put together can enable us construct error bars on our estimate. For a variable say $\hat{\mu}$, the error bars has the property:

$$\hat{\mu} = f_{map_{\lambda,\sigma}} \pm \sqrt{\mathrm{diag}(\Sigma_{f_{\lambda,\sigma}})} \tag{3.33}$$

where $\sqrt{\mathrm{diag}(\Sigma_{f_{\lambda,\sigma}})}$ is the standard deviation and it is obtainable from taking the square root of the diagonal elements of the variance-covariance matrix of $\Sigma_{f_{\lambda,\sigma}}$ which we shall re-visit shortly.

3.3.2 Decomposition of the Regularized Precision Matrix by SVD

In general, the SVD of an inverse matrix of the form

$$\Big(\frac{K^T K}{\sigma^2} + \lambda_{ml}^2 L^T L\Big)^{-1}$$

where L is a derivative operator are as follows:

By beginning with the SVD of K, we have

$$K = UDV^T$$

The corresponding SVD of (K^TK) decomposes into

$$\frac{K^TK}{\sigma^2} = V\,\frac{D^2}{\sigma^2}\,V^T$$

where the matrices U and V consists of the eigenvectors of K such that $V^TV = VV^T = I$ and $U^TU = UU^T = I$. As a result, $V^T = V^{-1}$ and $U^T = U^{-1}$. Moreover, the matrix D^2 consists of the singular values of K^TK and their singular values are equal to the eigenvalues of K^TK since the matrix K^TK is symmetric. Hence,

$$\begin{aligned}\Big(\frac{K^TK}{\sigma^2} + \lambda_{ml}^2 L^TL\Big)^{-1} &= \Big(V\,\frac{D^2}{\sigma^2}\,V^T + \lambda_{ml}^2\,L^TL\Big)^{-1}\\ &= \Big(V\,\frac{D^2}{\sigma^2}\,V^T + \lambda_{ml}^2\,L^T\,V\,V^T\,L\Big)^{-1}\end{aligned} \tag{3.34}$$

We use the above properties of real symmetric matrices to accomplish our objective by replacing the general regularizer operator L with the identity matrix I.

If $L = I$, we have

$$\begin{aligned}\Big(\frac{K^TK}{\sigma^2} + \lambda_{ml}^2 I\Big)^{-1} &= \Big(V\,\frac{D^2}{\sigma^2}\,V^T + \lambda_{ml}^2\,V\,V^T\Big)^{-1}\\ &= V\Big(\frac{D^2}{\sigma^2} + \lambda_{ml}^2\,I\Big)^{-1}V^T\\ &= \sum_{i=1}^{n}\mathbf{v}_i\Big(\frac{\sigma^2}{d_i^2 + \sigma^2\,\lambda_{ml}^2}\Big)\mathbf{v}_i^T\end{aligned} \tag{3.35}$$

Equation (3.35) is a powerful analytical result which serves as a tool for analysis of rank deficient and discrete ill-posed problems. The SVD allows the factorization of the $n \times n$, square symmetric matrix Σ_f^{-1} into orthogonal/orthonormal components. It is also quite computationally efficient for computing inverses and determinants of smaller systems. We will suspend the importance and details of SVD here until we get to Chapter 5 where it shall be used for analysis.

3.3.3 Exact Computation of $\Sigma_{f_{\lambda,\sigma}}$

We deduce the error bars on the estimate $f_{map_{\lambda,\sigma}}$ from the following:

Let $g_{\lambda_{ml}} = Kf_{\lambda_{ml}}$ where $g_{\lambda_{ml}}$ and $f_{\lambda_{ml}}$ represents exact regularized output and input. We write the difference:

$$\begin{aligned}f_{map_{\lambda,\sigma}} - f_{\lambda_{ml}} &= (K^TK + \sigma^2\lambda_{ml}^2 I)^{-1}K^T\tilde{g} - f_{\lambda_{ml}}\\ &= \Big\{(K^TK + \sigma^2\lambda_{ml}^2 I)^{-1}\,K^T(Kf_{\lambda_{ml}} + \epsilon_{\lambda_{ml}}) - (K^TK + \sigma^2\lambda_{ml}^2 I)^{-1}\,K^T(Kf_{\lambda_{ml}})\Big\}\\ &= (K^TK + \sigma^2\lambda_{ml}^2 I)^{-1}\,K^T\epsilon_{\lambda_{ml}}\end{aligned} \tag{3.36}$$

where $f_{\lambda_{ml}} = (K^TK + \sigma^2\lambda_{ml}^2 I)^{-1}\,K^Tg_{\lambda_{ml}}$.

The variance-covariance on the estimate $f_{\lambda_{ml},\sigma}$ is

$$\begin{aligned}\Sigma_{f_{\lambda,\sigma}} &= \Big\langle \Big(K^TK + \sigma^2\lambda_{ml}^2 I\Big)^{-1} K^T \epsilon_{\lambda_{ml}}\, \epsilon_{\lambda_{ml}}^T K\big((K^TK + \sigma^2\lambda_{ml}^2 I\Big)^{-1}\Big\}\Big\rangle \\ &= \Big(K^TK + \sigma^2\lambda_{ml}^2 I\Big)^{-1} K^T \,\langle\, \epsilon_{\lambda_{ml}}\, \epsilon_{\lambda_{ml}}^T \,\rangle\, K \Big(K^TK + \sigma^2\lambda_{ml}^2 I\Big)^{-1} \\ &= \sigma^2\Big(K^TK + \sigma^2\lambda_{ml}^2 I\Big)^{-1} K^TK\Big(K^TK + \sigma^2\lambda_{ml}^2 I)^{-1}\end{aligned} \tag{3.37}$$

where $\langle\, \epsilon_{\lambda_{ml}}\, \epsilon_{\lambda_{ml}}^T \,\rangle = \sigma^2 I$.

Decomposing equation (3.37) by SVD:

$$\begin{aligned}\Sigma_{f_{\lambda,\sigma}} &= \sigma^2\Big\{V\Big(D^2 + \sigma^2\lambda_{ml}^2 I\Big)^{-1} D^2\Big(D^2 + \sigma^2\lambda_{ml}^2 I\Big)^{-1} V^T\Big\} \\ &= \sigma^2 \sum_{i=1}^{n} \mathbf{v}_i\Big(\frac{d_i}{d_i^2 + \lambda_{ml}^2\sigma^2}\Big)^2 \mathbf{v}_i^T \\ &= \sum_{i=1}^{n} \mathbf{v}_i\, \Lambda_i^2\, \mathbf{v}_i^T\end{aligned} \tag{3.38}$$

where

$$\Lambda = \frac{D}{\sigma}\Big(\frac{D^2}{\sigma^2} + \lambda_{ml}^2\, I\Big)^{-1} \quad : \quad \sigma \neq 0 \tag{3.39}$$

is the deviation from the estimate $f_{map_{\lambda,\sigma}}$. Equation (3.39) intuitively shows that the error bars depends on the singular values of the Regularized Precision matrix Σ_f^{-1} and therefore can be approximated by Σ_f^{-1}. In subsection (3.3.4), we deduce how the approximation Σ_f^{-1} can be obtained from Taylor's expansion.

3.3.4 An Approximation to $\Sigma_{f_{\lambda,\sigma}}$

It is common to summmarize the posterior distribution by $f_{map_{\lambda,\sigma}}$ and construct approximate error bars on the fit for that particular values of λ_{ml}^2 and σ^2. By Taylor expanding the log-posterior up to the second order (via Laplace approximation), we have

$$\log p\,(f \,|\, \tilde{g}, \sigma^2, \lambda_{ml}^2) \;=\; \log p\,\Big(f_{map_{\lambda,\sigma}} \,|\, \tilde{g}, \sigma^2, \lambda_{ml}^2\Big) \;+\; \frac{1}{2}\Big(f - f_{map_{\lambda,\sigma}}\Big)^T \Sigma_f \Big(f - f_{map_{\lambda,\sigma}}\Big) \tag{3.40}$$

where

$$\Sigma_f \;=\; -\nabla_f^2 \log p\,(f \,|\, \tilde{g}, \sigma^2, \lambda_{ml}^2)$$

Hence

$$p\Big(f \,|\, \tilde{g}, \sigma^2, \lambda_{ml}^2\Big) \;=\; p\Big(f_{map_{\lambda,\sigma}} \,|\, \tilde{g}, \sigma^2, \lambda_{ml}^2\Big)\, \exp\, -\frac{1}{2}\Big\{\Big(f - f_{map_{\lambda,\sigma}}\Big)^T \Sigma_f \Big(f - f_{map_{\lambda,\sigma}}\Big)\Big\} \tag{3.41}$$

From equation (3.41), the posterior can be locally approximated as a Gaussian with precision matrix (or error bars) Σ_f^{-1}:

$$\Sigma_f^{-1} \;=\; \Big(\frac{K^TK}{\sigma^2} + \lambda_{ml}^2 I\Big)^{-1} \tag{3.42}$$

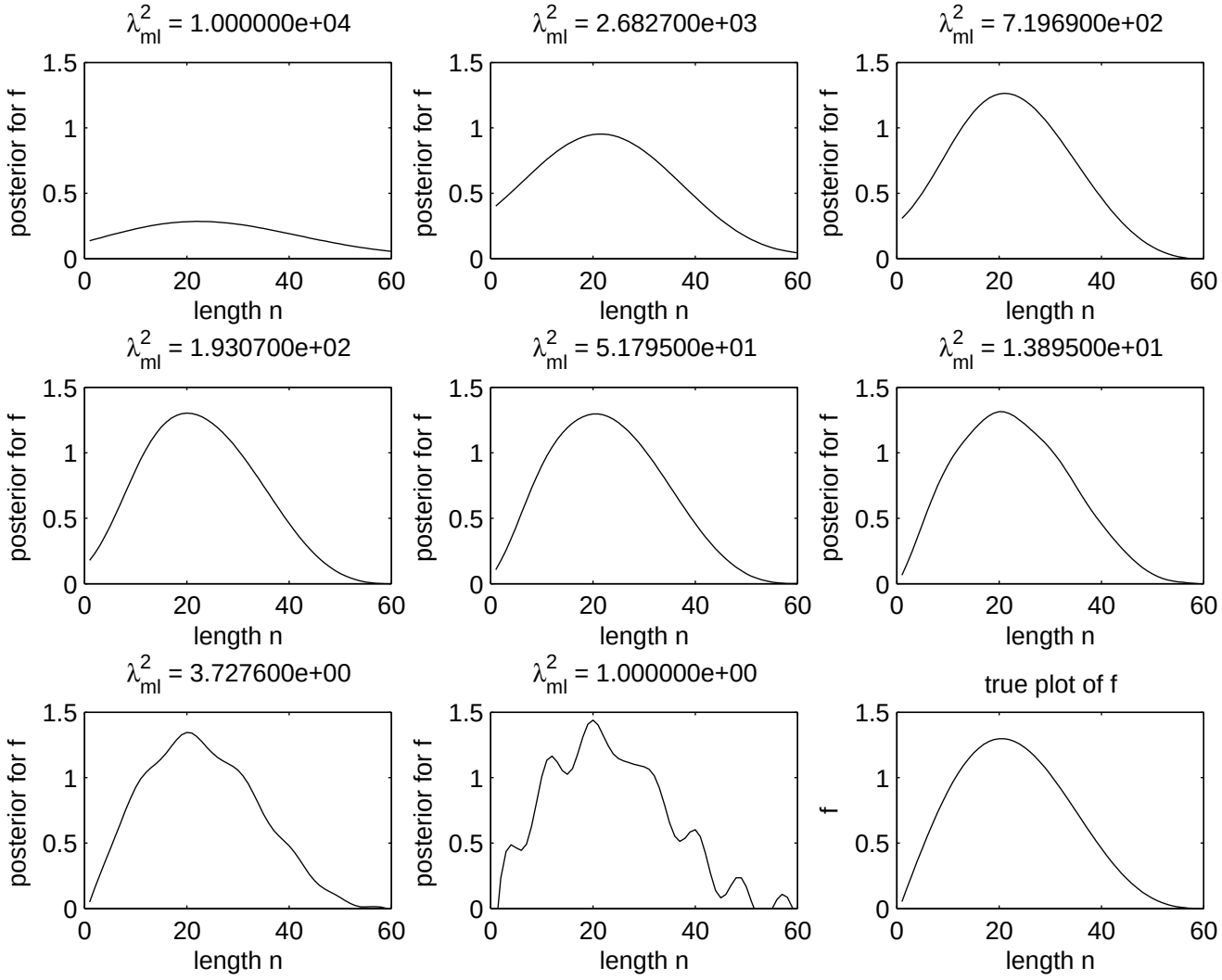

Figure 3.2: Estimates of the MAP posterior at $\sigma^2 = 10^{-6}$. The value of λ^2_{ml} for each posterior f satisfies the relation we established in of equation (3.28); $\lambda^2_{ml} = \lambda^2_{rls}/\sigma^2$. The shape of the subplots are replica of the subplots in Figure (2.8) except for a difference in the values of λ^2_{ml} and λ^2_{rls} as a result of the factor σ^2 which is factored into λ^2_{rls}.s This tells us that, if the noise level σ^2 and the regularization parameter λ^2_{rls} are given, the parameter λ^2_{ml} can be obtained and vice-versa.

Analysis

Equation (3.42) is a good approximation to $\Sigma_{f_{\lambda,\sigma}}$ of equation (3.38) for any given pair of hyperparameters λ_{ml} and σ. This approximation holds for all quadratic functional regularizers (priors) and noise models that are assumed to be Gaussian. Therefore, analysis of $\Sigma_{f_{\lambda,\sigma}}$ is tantamount to doing analysis on the approximation,

$$\sum_{i=1}^{n} \mathbf{v}_i \left(\frac{d_i^2}{\sigma^2} + \lambda_{ml}^2 \right)^{-1} \mathbf{v}_i^T \tag{3.43}$$

for any given K and σ^2 since the ratio D/σ can viewed as some sort of scaling. However, we note that $(d_i/\sigma\,;\,\forall\, i)$ of equation (3.39) is known since d_i comes from the singular values of the matrix K at the noise level σ^2. For *non-zero* λ_{ml}^2 and σ^2, we have $\Lambda_i \to 0$ if and only if $d_i/\sigma \to 0$. In this case, the error bars depends on the SVD components (or singular values) of K having numerical values roughly equal to (or approaching) zero. Also from equation (3.38), we can easily see that when $\lambda_{ml} = 0$, we have

$$\Sigma_{f_{\lambda_{ml},\sigma}} \quad = \quad \sigma^2 \sum_{i=1}^{n} \mathbf{v}_i \Big(\frac{1}{d_i^2} \Big) \mathbf{v}_i^T \tag{3.44}$$

which corresponds to the variance-covariance of the Least Squares estimates.

From above, it is enough to focus on the analysis of Σ_f^{-1} :

$$\Sigma_f^{-1} \quad = \quad \sum_{i=1}^{n} \mathbf{v}_i \left(\frac{d_i^2}{\sigma^2} + \lambda_{ml}^2 \right)^{-1} \mathbf{v}_i^T \tag{3.45}$$

For fixed σ^2, $(\sigma^2 > 0)$:

(a) **if $\lambda_{ml} << d_i$ for all i, we have**

$$\Sigma_f^{-1} \simeq \sum_{i=1}^{n} \mathbf{v}_i \Big(\frac{\sigma^2}{d_i^2} \Big) \mathbf{v}_i^T$$

(b) **if $\lambda_{ml} >> d_i$ for all i, we have**

$$\Sigma_f^{-1} \simeq \sum_{i=1}^{n} \mathbf{v}_i \left(\frac{\sigma^2}{\lambda_{ml}^2 \, \sigma^2} \right) \mathbf{v}_i^T = \sum_{i=1}^{n} \mathbf{v}_i \left(\frac{1}{\lambda_{ml}^2} \right) \mathbf{v}_i^T$$

Figures (3.2), (3.3) and (3.4) are the estimates, exact error bars using equation (3.38) and approximate error bars using equation (3.42) on the Gravity Problem for $n = 60$, $d = 0.25$ and at a noise level of 10^{-3}. In the case of the exact error bars, we can see the effect of small d_i/σ at the middle to the tail-ends of each subplot for $\sigma^2 > 0$.

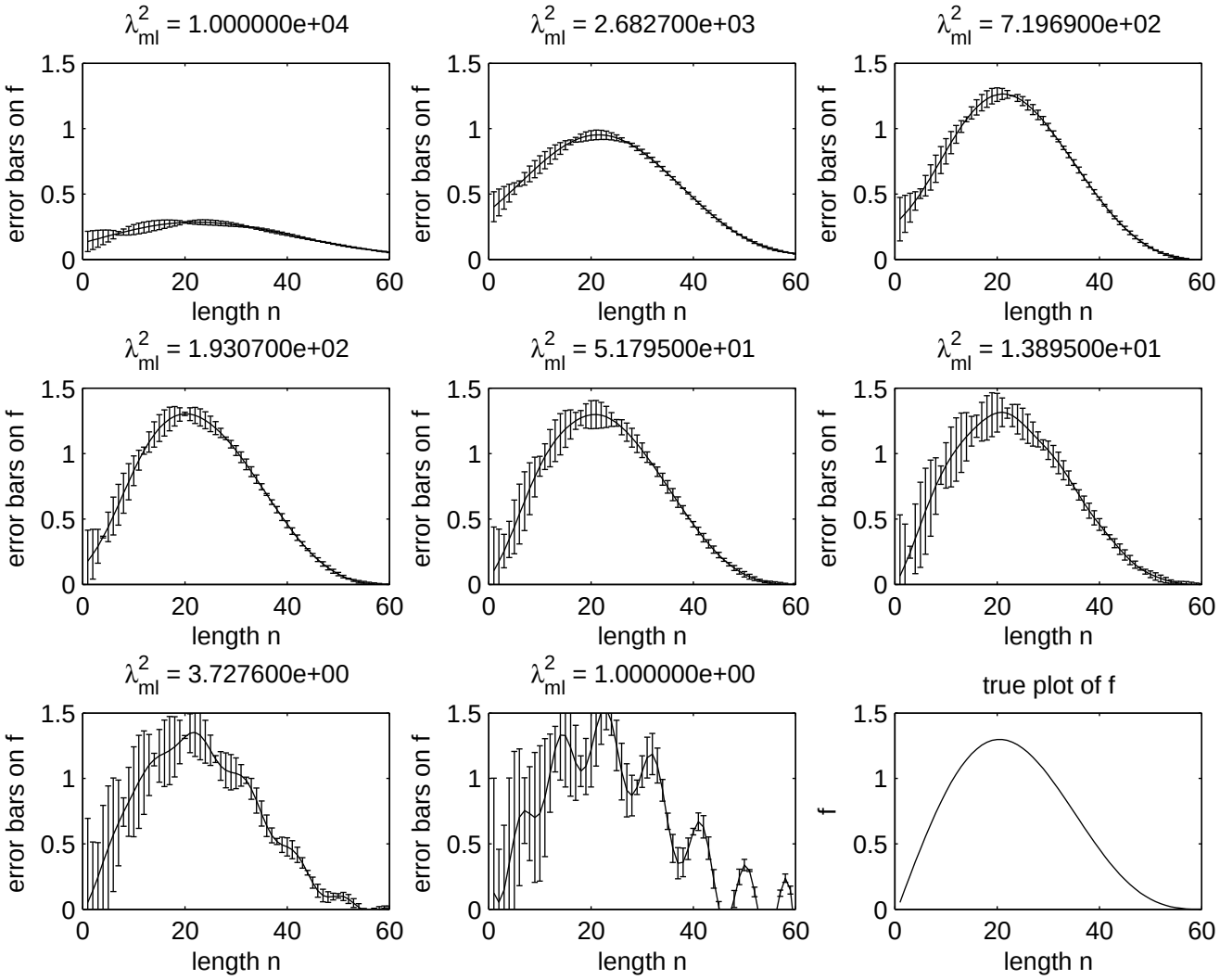

Figure 3.3: The exact Error bars on the estimates of the posterior for f of Figure (3.2) using the expression for $\Sigma_{f_{\lambda,\sigma}}$ at the same $\sigma^2 = 10^{-6}$.

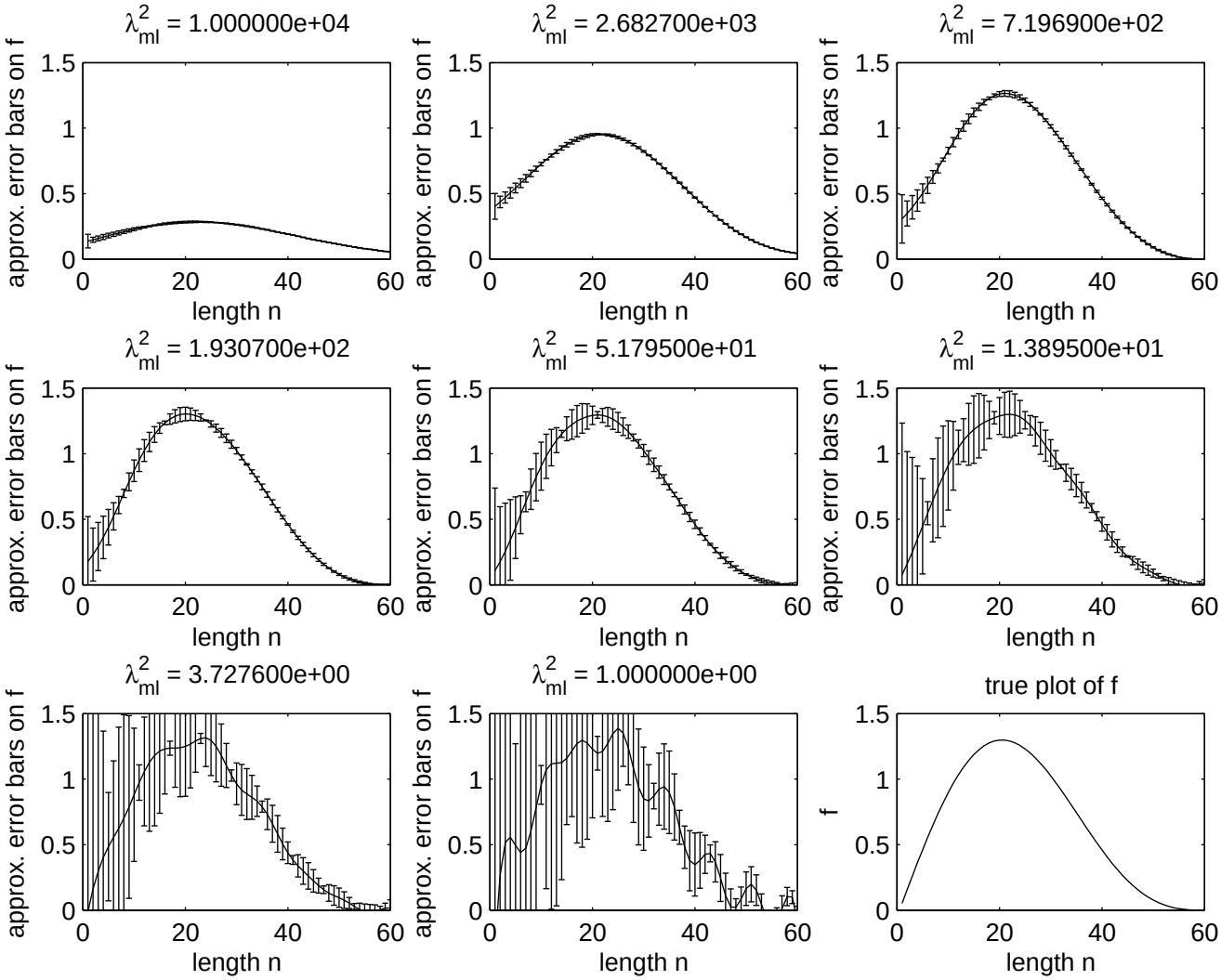

Figure 3.4: The error bars approximated by Σ_f^{-1}.

3.3.5 Application of the Maximum Likelihood Principle to the Problem

The stochastic process $\tilde{g}$ is characterized by specifying the finite dimensional conditional distribution

$$p\{\tilde{g}(s)\,|\sigma^2, \lambda_{ml}^2\} \;=\; p\{(\tilde{g}_{s_1}^T, \tilde{g}_{s_2}^T, \tilde{g}_{s_3}^T, ..., \tilde{g}_{s_n}^T)\,|\sigma^2, \lambda_{ml}^2\}$$

Bayes rule further allows us to re-characterize the above finite dimensional conditional distribution by specifying the joint conditional density function through the marginal distribution of $\tilde{g}$ given σ^2 and λ_{ml}^2. [3]

In general, the re-characterization of the marginal distribution of $\tilde{g}$ given the parameter pair $(\sigma^2, \lambda_{ml}^2)$ is defined by:

$$p(\tilde{g}|\lambda_{ml}^2, \sigma^2) \quad = \quad \int p(\tilde{g}\,|\,f, \sigma^2, \lambda_{ml}^2)\, p(f\,|\,\lambda_{ml}^2, \sigma^2)\,\mathrm{d}f \tag{3.46}$$

where the stochastic input variable f and the parameter σ are assumed to be independent and the integrand is called the joint conditional density. A functional say l defined by

$$l(f, \lambda_{ml}^2, \sigma^2) \;=\; p(\tilde{g}\,|\,f, \sigma^2, \lambda_{ml}^2)$$

is called the *conditional likelihood-function* (conditioned on f).

From Bayes' rule we have

$$p(\tilde{g}\,|\,f, \sigma^2)\, p(f\,|\,\lambda_{ml}^2) \;=\; p(\tilde{g}, f\,|\,\lambda_{ml}^2, \sigma^2) \tag{3.47}$$

where $p(\tilde{g}\,|\,f, \sigma^2)$ and $p(f\,|\,\lambda_{ml}^2)$ are the same as given in equations (3.21) and (3.22) and the equivalence of their corresponding normalizing constants $C_l(\sigma)$ and $C_p(\lambda_{ml})$ for this problem are

$$\left\{ \int \exp -\frac{1}{2\sigma^2}\|\tilde{g} \,-\, Kf\|^2 \,\mathrm{d}f \right\}^{-1} \;=\; (2\,\pi\,\sigma^2)^{-n/2} \tag{3.48}$$

and

$$\left\{ \int \exp[-\frac{\lambda_{ml}^2}{2}\| \,f\,\|^2 \,\mathrm{d}f \right\}^{-1} \;=\; \Big(\frac{\lambda_{ml}^2}{2\,\pi}\Big)^{n/2} \tag{3.49}$$

The integrand of the conditional density $p(\tilde{g}\,|\sigma^2, \lambda_{ml}^2)$ is equivalent to the joint conditional density of equation (3.47). Hence by substitution of equations (3.47), (3.48) and (3.49)

$$\begin{aligned} p(\tilde{g}, f\,|\,\sigma^2, \lambda_{ml}^2) \quad &= \quad \Big(\frac{\lambda_{ml}^2}{2\,\pi}\Big)^{n/2} \Big(2\,\pi\,\sigma^2\Big)^{-n/2} \exp\Big\{ \,-\frac{1}{2\,\sigma^2}\,\Big\|\,\tilde{g} \,-\, Kf\,\Big\|_2^2 \,+ \\ &\qquad -\frac{\lambda_{ml}^2}{2}\,\Big\|\,f\,\Big\|_2^2\Big\} \\ &= \quad C(\lambda_{ml}, \sigma) \exp\Big\{ -\frac{1}{2\,\sigma^2}\Big[\,\Big\|\,\tilde{g} \,-\, Kf\,\Big\|_2^2 + \sigma^2\lambda_{ml}^2\,\Big\|\,f\,\Big\|_2^2\Big]\Big\} \end{aligned} \tag{3.50}$$

[3] The parameter λ_{ml}^2 is usually referred to as a *hyperparameter*. It controls the distribution of other parameters in equation (3.46).

where $C(\lambda_{ml}, \sigma) = C_p(\lambda_{ml})\, C_l(\sigma)$. Also, the functional $p(\tilde{g}\,|\,\sigma^2, \lambda_{ml}^2)$ is referred to as the likelihood function and we shall denote it by $L(\theta)$ with $\theta = (\lambda_{ml}, \sigma)$.

Our objective is then to find the pair say $(\lambda_{ml}^{2\,\star}, \sigma^{2\,\star})$ from the set of parameters $\{\lambda_{ml}, \sigma\}$ for which L is maximum. If the above distributions are defined in terms of the exponential family of distributions, then it is usually advisable to take the logarithm of L before we proceed to finding the maximum. From the two *'hoods'*, likelihoods multiply and log-likelihoods add.

3.3.6 Explicit Result for the Gaussian Model

Here we will make a general assumption about the mean and variance-covariance of the stochastic variables $\tilde{g}$ and f. We are considering the case where the mean of $\tilde{g}$ is different from zero. We state our assumptions as follows:

$$p(\epsilon|\sigma^2) \sim \mathbb{N}(0, \sigma^2 I) \quad , \quad p(f|\lambda_{ml}^2) \sim \mathbb{N}(0, \lambda_{ml}^{-2} I) \quad \text{and} \quad \tilde{g} \sim \mathbb{N}(Kf, \Sigma_f)$$

where the vectors ϵ and f are assumed to be independent.

Since the output $\tilde{g}$ and input f of the process are Gaussian, the definitions of their conditional probability densities are equivalent to the previous ones. A replica of equation (3.46) to the model gives the following:

$$\begin{aligned} p(\tilde{g}\,|\,\lambda_{ml}^2, \sigma^2) &= C(\lambda_{ml}, \sigma) \int_{-\infty}^{\infty} \exp\Big\{ -\frac{1}{2\sigma^2} \Big\{ f^T(K^TK + \lambda_{ml}^2\sigma^2 I)f \,-\, 2\tilde{g}^T K f \\ &\quad + \tilde{g}^T\tilde{g} \Big\} \Big\}\, \mathrm{d}f \\ &= C(\lambda_{ml}, \sigma)\, (2\pi)^{n/2} \,\Big|\, \Sigma_f \,\Big|^{-1/2} \frac{\exp\, \frac{1}{2}\Big(\nu^T\, \Sigma_f^{-1} \nu \Big)}{\exp \frac{1}{2\sigma^2} \,\|\, \tilde{g} \,\|^2} \end{aligned} \tag{3.51}$$

where

$$C(\lambda_{ml}, \sigma) = \Big(\frac{\lambda_{ml}^2}{2\pi}\Big)^{n/2} \Big(\frac{1}{2\pi\,\sigma^2}\Big)^{n/2} \quad \text{and} \quad \nu^T \,=\, \frac{\tilde{g}^T K}{\sigma^2}$$

A decomposition of the determinant $\Big|\, \Sigma_f \,\Big|$ by SVD gives

$$\begin{aligned} \Big|\, \Sigma_f \,\Big| &= \Big| V\Big(\frac{D^2}{\sigma^2} \,+\, \lambda_{ml}^2 I\Big) V^T \Big| \\ &= \Big| \Big(\frac{D^2}{\sigma^2} \,+\, \lambda_{ml}^2 I\Big) \Big| \\ &= \prod_{i=1}^{n} \Big(\frac{d_i^2}{\sigma^2} \,+\, \lambda_{ml}^2\Big) \end{aligned} \tag{3.52}$$

By substitution of equation (3.52) into equation (3.51):

$$p(\tilde{g}|\lambda_{ml}^2, \sigma^2) \;=\; \Big(\frac{\lambda_{ml}^2}{2\pi\sigma^2}\Big)^{n/2} \;\Big[\prod_{i=1}^{n} \Big(\frac{d_i^2}{\sigma^2} \,+\, \lambda_{ml}^2\Big)\Big]^{-1/2} \frac{\exp\, \frac{1}{2}\Big[\, \upsilon_i^T \Big(\frac{d_i^2}{\sigma^2} \,+\, \lambda_{ml}^2\Big)^{-1} \upsilon_i\Big]}{\exp \frac{1}{2\sigma^2} \,\|\, \tilde{g} \,\|^2} \tag{3.53}$$

where $\upsilon_i^T = \nu^T \mathbf{v}_i = \frac{\tilde{g}^T K}{\sigma^2} \mathbf{v}_i$.
Taking logarithms of both sides of equation (3.53) gives

$$\begin{aligned} \log\{p(\tilde{g} \mid \lambda_{ml}^2, \sigma^2)\} &= \frac{n}{2}\log\Big(\frac{\lambda_{ml}^2}{2\pi\sigma^2}\Big) - \frac{1}{2}\sum_{i=1}^{n}\log\Big(\frac{d_i^2}{\sigma^2} + \lambda_{ml}^2\Big) \\ &\quad - \frac{1}{2\,\sigma^2}\sum_{i=1}^{n}\tilde{g}_i{}^T\tilde{g}_i + \frac{1}{2}\Big\{\sum_{i=1}^{n}\upsilon_i^T\Big(\frac{d_i^2}{\sigma^2} + \lambda_{ml}^2\Big)^{-1}\upsilon_i\Big\} \end{aligned} \tag{3.54}$$

In general, equation (3.51) would have been a very difficult integral to perform if the prior was not Gaussian. Secondly, we can easily see from either equation (3.53) or (3.54) that the marginal distribution is non-linear in λ_{ml}^2 and σ^2. Hence we cannot just differentiate the likelihood function with respect to the hyperparameters σ^2 and λ_{ml}^2 (which should have been the case). We end here with the exact integration procedure above. However, we will refer to some of the equations in this subsection when we get to Bayesian Inference methods. The above problem shall then be addressed to encompass Non-Gaussian priors that assumes an approximation to Gaussian distributions as well.

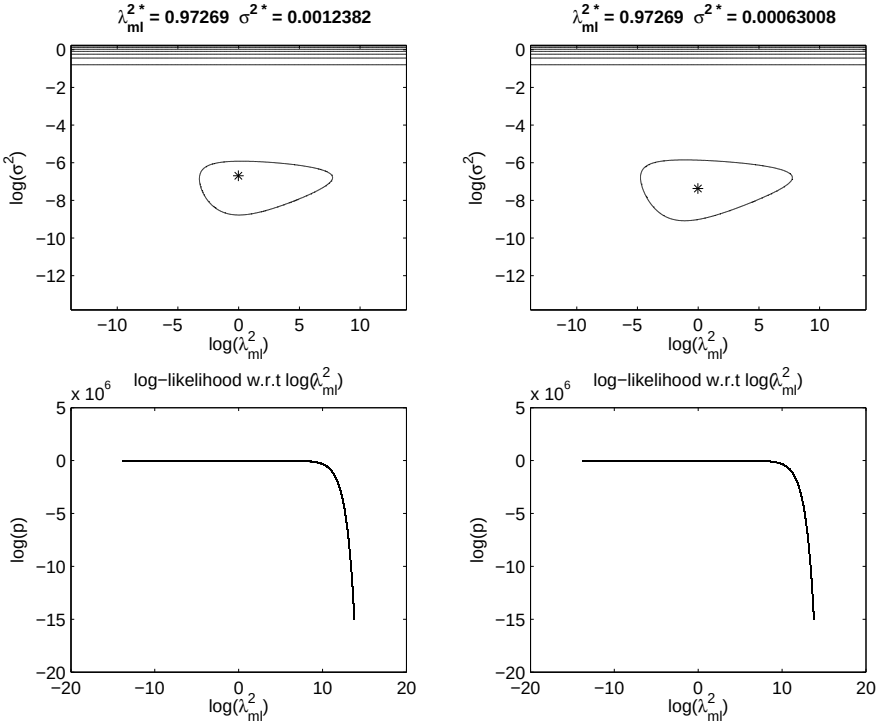

Figure 3.5: log-likelihood contours (top-row) and log-likelihood plots (bottom-row) for $n = 30$, $d = 10$ (first-column) and $d = 1000$ (second column) of The Gravity Problem Model.

Figures (3.5) and (3.6) are log-likelihood contour plots, log-likelihoods plots and surfaces of the log-likelihood. Figures (3.7) and (3.8) also shows likelihood contour plots, likelihood plots and surfaces of the likelihood. The set of values of λ_{ml}^2 and σ^2 for all the plots were respectively generated from matlab using logspace$(-6, 6, 500)$ and logspace$(-6, 0, 500)$ for $n = 30$ and at different values of d; $d = 10$ for the first columns

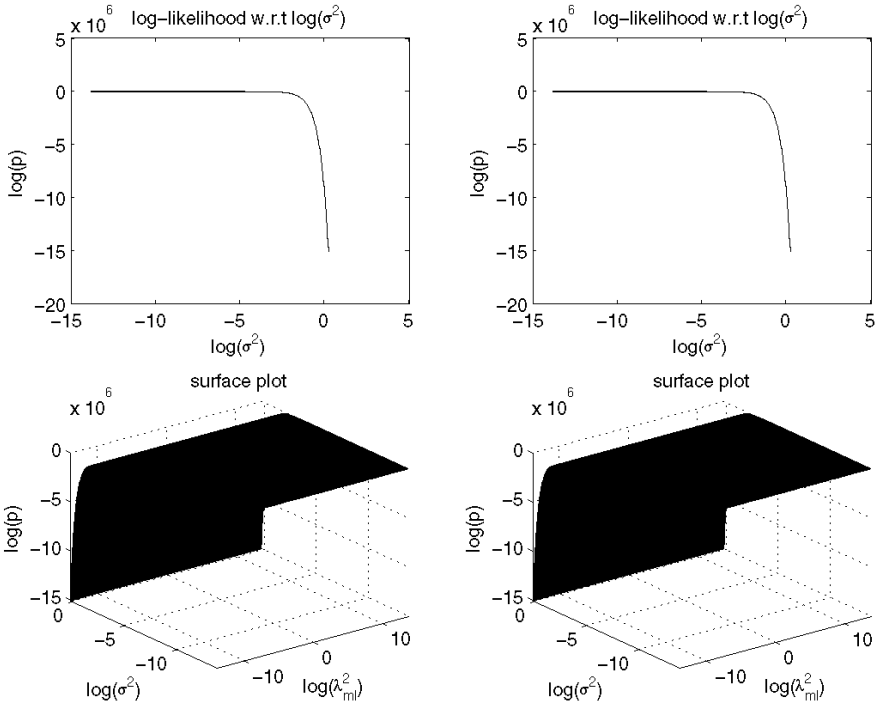

Figure 3.6: log-likelihood w.r.t $\log(\sigma^2)$ (top-row) and surface of the log-likelihood (bottom-row) for $n = 30,\ d = 10$ (first-column) and $d = 1000$ (second column) of the Gravity Model.

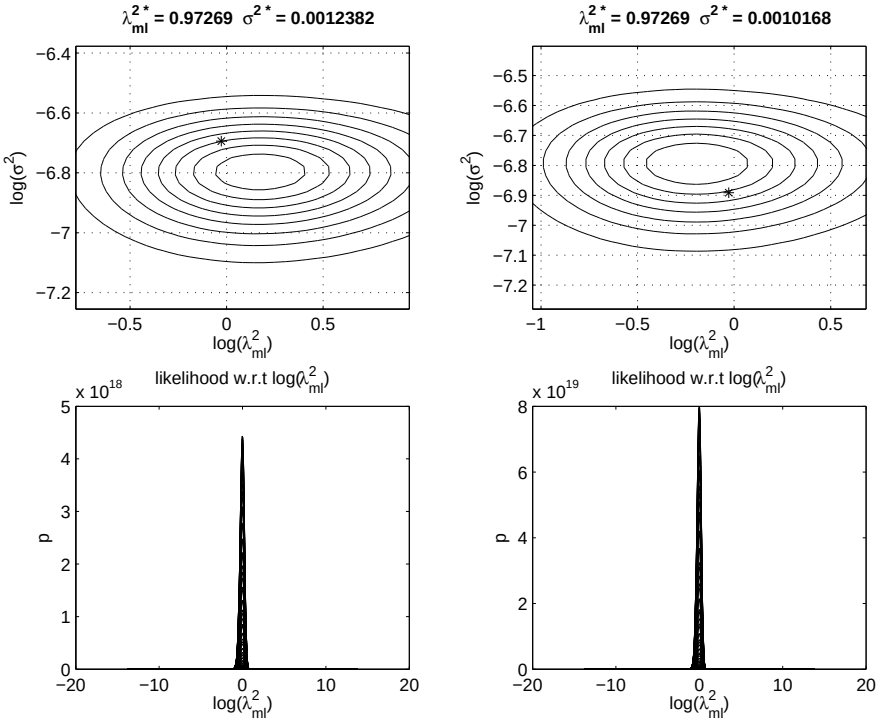

Figure 3.7: likelihood contours (top-row) and likelihood plots (bottom-row) for $n = 30,\ d = 10$ (first-column) and $d = 1000$ (second column) of the Gravity Model.

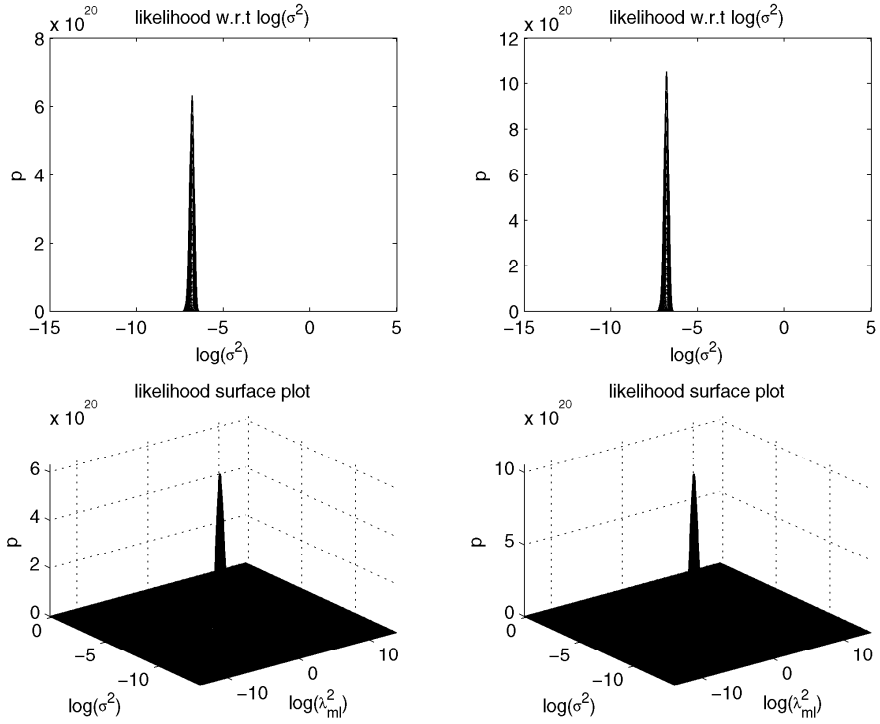

Figure 3.8: likelihood w.r.t $\log(\sigma^2)$ (top-row) and surface of the likelihood (bottom-row) for $n = 30$, $d = 10$ (first-column) and $d = 1000$ (second column) of the Gravity Model.

and $d = 1000$ for the second columns of each of the Figure.

The most probable value of the pair of parameters $(\lambda^{2\star}_{ml}, \sigma^{2\star})$ for each column is marked $\star$ in each contour plot and they have their respective values shown on top.

In subsection (3.3.7), we appeal to two different set of assumptions for a simple case where the mean of $\tilde{g}$ is assumed to be zero and another case where the mean of $\tilde{g}$ is different from zero and use the definition of probability density function of a normal multi-dimensional variable of equation (3.6) on the same problem.

3.3.7 Multivariate Gaussian Distribution Approach to the Model

The alternative method to solving the same problem is to make a *tacit* assumption about the output $\tilde{g}$ which can be either; $\tilde{g}$ comes from the sum of two Gaussians with mean zero or otherwise. We state our assumptions for the two cases as follows:

$$\text{(a)} \quad p(f|\lambda^2_{ml}) \sim \mathbb{N}(0, \frac{1}{\lambda^2_{ml}} I) \quad , \quad p(\epsilon|\sigma^2) \sim \mathbb{N}(0, \sigma^2 I) \quad \text{and} \quad \tilde{g} \sim \mathbb{N}(0, \Sigma_{\tilde{g}})$$

$$\text{(b)} \quad \tilde{g} \sim \mathbb{N}(Kf, \sigma^2 I) \quad , \quad p(f|\lambda^2_{ml}) \sim \mathbb{N}(0, \frac{1}{\lambda^2_{ml}} I) \quad \text{and} \quad p(\epsilon|\sigma^2) \sim \mathbb{N}(0, \sigma^2 I)$$

(a) Zero Mean

Using the assumption in (a), the covariance of marginal for $\tilde{g}$ is obtainable from the following.

$$\begin{aligned} \Sigma_{\tilde{g}} &= \langle\, \tilde{g}\tilde{g}^T \,\rangle \\ &= \langle\, (Kf + \epsilon)(Kf + \epsilon)^T \,\rangle \\ &= \langle\, Kff^TK^T \,\rangle + \langle\, \epsilon\epsilon^T \,\rangle \end{aligned} \tag{3.55}$$

where

$$\langle\, Kf\epsilon^T \,\rangle = \langle\, \epsilon f^T K^T \,\rangle = 0$$

because the vectors ϵ and f are assumed to be independent.
A decomposition of equation (3.55) by SVD results in

$$\begin{aligned} \Sigma_{\tilde{g}} &= [K\frac{1}{\lambda_{ml}^2}K^T] + \sigma^2 I \\ &= U\Big\{\frac{D^2}{\lambda_{ml}^2} + \sigma^2 I\Big\}U^T \end{aligned} \tag{3.56}$$

Since $\tilde{g}$ is asumed to be Gaussian, its normalized joint conditional probability density $p(\tilde{g}|\sigma^2, \lambda_{ml}^2)$ is

$$p(\tilde{g}|\sigma^2, \lambda_{ml}^2) = \Big(\frac{1}{2\pi}\Big)^{n/2} |\, \Sigma_{\tilde{g}} \,|^{-1/2} \exp - \Big\{ \frac{1}{2}\tilde{g}^T\Sigma_{\tilde{g}}^{-1}\tilde{g}\Big\} \tag{3.57}$$

Substituting the equivalent expressions of $\Sigma_{\tilde{g}}^{-1}$ and $|\, \Sigma_{\tilde{g}} \,|$ into equation (3.58) above gives

$$p(\tilde{g}|\sigma^2, \lambda_{ml}^2) = \Big(\frac{1}{2\pi}\Big)^{n/2} \Big[\prod_{i=1}^{n}\Big(\frac{d_i^2}{\lambda_{ml}^2} + \sigma^2\Big)\Big]^{-1/2} \exp - \frac{1}{2}\left\{ \sum_{i=1}^{n}\omega_i^T\Big(\frac{d_i^2}{\lambda_{ml}^2} + \sigma^2\Big)^{-1}\omega_i \right\} \tag{3.58}$$

where $\omega_i = \mathbf{u}_i^T\tilde{g}$ and we have made use of the initial assumption that the mean $\langle\, \tilde{g} \,\rangle = 0$.

Taking the logarithm of both sides of equation (3.58) gives us another important equation

$$\begin{aligned} \log p\,(\tilde{g}|\sigma^2, \lambda_{ml}^2)\} &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log\Big\{\Big(\frac{d_i^2}{\lambda_{ml}^2} + \sigma^2\Big)\Big\} \\ &\quad \frac{1}{2}\left\{ \sum_{i=1}^{n}\omega_i^T\Big(\frac{d_i^2}{\lambda_{ml}^2} + \sigma^2\Big)^{-1}\omega_i \right\} \end{aligned} \tag{3.59}$$

c We already know that the marginal distribution is non-linear in the parameters λ_{ml}^2 and σ^2 in the zero neighbourhood. So, a closed form ML principle is intractable. Therefore, a maximization of the log-likelihood through differentiation and solving for the zeros directly is impossible. Interestingly, it is possible to find the best possible pair of hyperparameters say $(\lambda_{ml}^{2\star}, \sigma^{2\star})$ in equations (3.57) from a given set of hyperparameters $\{\lambda_{ml}^2, \sigma^2\}$ but it is impossible to do the same for non-zero mean due to the presence of

the latent variable f and/or the assumptions made in (b).

We present an iterative scheme for estimating $(\lambda^{2\,\star}_{ml}\,,\,\sigma^{2\,\star})$ using the ML EM Algorithm to the problem. This procedure is (more or less) a showcase or test-bed for understanding the Variational Bayesian EM Algorthm which we will treat later. It will also make it easy for us to see the difference between Regularization in MAP and ML.

Application of EM Algorithm (Tikhonov EM Regularization)

Before we stoop to this iterative algorithm, we shall let the symbols $\beta = 1/\sigma^2$ and $\alpha = \lambda^2_{ml}$. We are also dealing with a single data point in n-dimensions so there is no need for a summation sign here.

In this EM application, instead of maximizing the likelihood $p(\tilde{g}\,|\,\beta,\alpha)$, [4] we rather seek to maximize the joint likelihood $p(\tilde{g}, f\,|\beta,\alpha)$ of the unobserved random variables in the model which is a function of the latent variable f. [5] The quantity $p(\tilde{g}, f\,|\,\beta,\alpha)$ then becomes a function of the unobserved random variables f. Hence we have

$$L_c(\beta,\alpha) \quad = \quad p(\tilde{g},\, f\,|\beta,\alpha) \tag{3.60}$$

In using Bayes' rule, followed by taking logarithm of both sides and making use of the fact that f is independent on ϵ, we have

$$\log L_c(\beta,\alpha) \,=\, \log p(\tilde{g}\,|\,f,\,\beta) \,+\, \log p\,(f\,|\,\alpha) \tag{3.61}$$

[6] The complete log-likelihood function, $(\log L_c\,)$ of equation (3.61) reduces to

$$\begin{aligned}\log L_c(\alpha,\beta) \quad &= \quad -\,\mathrm{n}\log(2\,\pi) \,+\, \frac{n}{2}\log\beta \,+\, \frac{n}{2}\log\alpha \,- \\ &\qquad \frac{\beta}{2}\Big(\tilde{g} \,-\, K\,f\Big)^T\Big(\,\tilde{g} \,-\, K\,f\,\Big) \,-\, \frac{\alpha}{2}\,f^T f\end{aligned} \tag{3.62}$$

The M Step of EM

The M step involves taking the expectation of the complete log-likelihood and maximizing it with respect to β.

Differentiating with respect to β:

$$\frac{\mathrm{d}\,\langle\,\log L_c(\alpha,\beta)\,\rangle}{\mathrm{d}\,\beta} \quad = \quad \frac{n}{2\,\beta} \,-\, \frac{1}{2}\Big\{\Big\langle\Big(\tilde{g} - K\,f\Big)^T\Big(\,\tilde{g} - K\,f\Big)\Big\rangle\Big\} \tag{3.63}$$

[4] $p(\tilde{g}\,|\beta,\alpha)$ is also referred to as the Incomplete Data Likelihood

[5] $p(\tilde{g}, f\,|\beta,\alpha)$ of the unobserved random variables in the model is also known as Complete Data Likelihood.

[6] The second term on the right hand side of equation (3.61) is independent on β. Therefore, finding the most probable β simply means that we only need the first term on the right hand side of equation (3.61). On the otherhand, the first term on the right hand side of equation (3.61) is independent on α. Therefore, finding the most probable α simply means that we only need the second term on the right hand side of equation (3.61).

Setting the derivative to zero and solving for β gives

$$\begin{aligned} \beta &= \frac{n}{\Big\langle \Big(\tilde{g} - K\,f\Big)^T \Big(\tilde{g} - K\,f\Big) \Big\rangle} \\ &= \frac{n}{\tilde{g}^T\,\tilde{g} \;-\; \Big(\tilde{g}\,\langle\, f^T \rangle\Big) K^T \;+\; \mathrm{Tr}[K^T K\,\langle f\, f^T \rangle]\Big\}} \end{aligned} \tag{3.64}$$

where $f^T(K^T K)\,f \;=\; \mathrm{Tr}[K^T K\,f\,f^T]$ and we have used the relation $x^T A x = \mathrm{Tr}[A x x^T]$.

At this point we are still left with the problem of determining the actual values of $\langle f \rangle$ and $\langle f\, f^T \rangle$. This is where the E-step comes in.

The E Step of EM

The E-step involves the maximization of the log-posterior $p(f \,|\, \tilde{g}, \beta, \alpha)$. The analytical form of the expectation for $p(f \,|\, \tilde{g}, \beta, \alpha)$ can be obtained from Bayes' rule

$$p\,(f \,|\, \tilde{g}, \beta, \alpha) \quad \propto \quad p(\tilde{g} \,|\, f, \beta)\; p\,(f \,|\, \alpha) \tag{3.65}$$

Taking logarithm of both sides and expanding gives

$$\begin{aligned} \log p(f \,|\, \tilde{g}, \beta, \alpha) &= -\frac{n}{2}\log 2\pi \;+\; \frac{n}{2}\log \beta \;-\; \frac{\beta}{2}\Big\{ \Big(\tilde{g} - K\,f\Big)^T \Big(\tilde{g} - K\,f\Big) \Big\} \\ &\quad -\; \frac{n}{2}\log 2\pi \;-\; \frac{\alpha}{2} f^T f \;+ \frac{n}{2}\log \alpha + \;\Upsilon \\ &= -\frac{1}{2}\Big\{ \tilde{g}^T \beta \tilde{g} \;-\; 2 f^T K^T\, \beta \tilde{g} \;+\; f^T(\alpha\, I + K^T \beta K\,) f \Big\} \;+\; \Upsilon' \end{aligned} \tag{3.66}$$

where Υ and Υ' are independent of f and we have assumed that the prior is a quadratic functional given by $p\,(f \,|\, \alpha) = (\alpha/2\,\pi)^{n/2} \exp\, -\,\frac{\alpha}{2} f^T f$. From equation (3.66), it is easy to infer that

$$p\,(f \,|\, \tilde{g}, \beta, \alpha) \sim \mathbb{N}\Big(\langle f \rangle, \Sigma_f^{-1}\Big)$$

where Σ_f^{-1}:

$$\Sigma_f^{-1} \quad = \quad (\alpha\, I \;+\; \beta\, K^T K)^{-1} \tag{3.67}$$

and $\langle\, f\,\rangle$ is

$$\begin{aligned} \langle f \rangle &= \Sigma_f^{-1} K^T \beta\, \tilde{g} \\ &= f_{ml_{\lambda,\sigma}} \end{aligned} \tag{3.68}$$

where Σ_f^{-1} and $\langle\, f\,\rangle$ are the variance-covariance and mean with a prior on f.

The corresponding $\langle f\, f^T \rangle$ is

$$\langle f\, f^T \rangle \quad = \quad \Sigma_f^{-1} + \langle\, f\,\rangle \langle\, f\,\rangle^T \tag{3.69}$$

If there is no prior on f or ($\alpha = 0$) then we assume non-informative priors. In this case

$$\langle f \rangle = (\beta K^T K)^{-1} K^T \beta \tilde{g} = (K^T K)^{-1} K^T \tilde{g} \tag{3.70}$$

which gives the updates equation for the standard EM-Algorithm (and it is equal to the Least Squares solution). The difference between equations (3.68) and (3.70) comes from the addition of a prior. For quadratic functional priors, we view the conditional prior $p(f \mid \lambda_{ml}^2)$ as equivalent to an unconditional prior $p(f)$ when $\lambda_{ml}^2 = 1$.

3.3.8 Summary of Equations for the EM Algorithm

E Step

$$\begin{aligned} \Sigma_f^{-1} &= (\alpha I + \beta K^T K)^{-1} \\ \langle f \rangle &= \Sigma_f^{-1} K^T \beta \tilde{g} = (\alpha I + \beta K^T K)^{-1} \beta \tilde{g} \\ \langle f f^T \rangle &= \Sigma_f^{-1} + \langle f \rangle \langle f \rangle^T \end{aligned} \tag{3.71}$$

where $\beta = \sigma^{-2}$ in the above.

M Step

$$\begin{aligned} \beta &= \frac{n}{\left\langle \left(\tilde{g} - K f\right)^T \left(\tilde{g} - K f\right) \right\rangle} \\ &= \frac{n}{\tilde{g}\tilde{g}^T - \left(\tilde{g} \langle f^T \rangle\right) K^T + \mathrm{Tr}[K^T K \langle f f^T \rangle]\Big\}} \end{aligned} \tag{3.72}$$

The solutions to the equations above is sometimes called the solution to the *maximum penalized likelihood* (MPL). The penalty term that assessses the physical plausibility of the solution is

$$\frac{1}{\sigma^2} \| f \|_2^2$$

The MPL solution is the same as the Tikhonov's Regularized solution

$$f_{ml_{\lambda,\sigma}} = (K^T K + \sigma^2 \lambda_{ml}^2 I)^{-1} K^T g \tag{3.73}$$

3.3.9 The Difference between MAP and Maximum Likelihood

We have so far seen that, differences in ML and MAP do not lie in the equation governing the Bayesian posterior for the stochastic variable f because the same Bayes' rule is used for estimating f in either case. We will use the concept of exact marginalization over continuous variables to find out whether difference(s) really exists.

From Bayes' Rule we have

$$p(\tilde{g}, f, \lambda_{ml}^2, \sigma^2) = p(\tilde{g}, f \mid \lambda_{ml}^2, \sigma^2)\, p(\lambda_{ml}^2, \sigma^2) \tag{3.74}$$

Integrating out f

$$\begin{aligned} p(\lambda_{ml}^2, \sigma^2 \,|\, \tilde{g})\, p\,(\tilde{g}) &= p\,(\lambda_{ml}^2, \sigma^2) \int p\,(\tilde{g}, f \,|\, \lambda_{ml}^2, \sigma^2)\, \mathrm{d}f \\ &= p\,(\lambda_{ml}^2, \sigma^2)\; p\,(\tilde{g} \,|\, \lambda_{ml}^2, \sigma^2) \end{aligned} \tag{3.75}$$

Hence

$$\begin{aligned} p(\lambda_{ml}^2, \sigma^2 \,|\, \tilde{g}) &= \frac{p\,(\lambda_{ml}^2, \sigma^2)}{p\,(\tilde{g})}\; p\,(\tilde{g} \,|\, \lambda_{ml}^2, \sigma^2) \\ &\propto p\,(\lambda_{ml}^2, \sigma^2)\, p\,(\tilde{g} \,|\, \lambda_{ml}^2, \sigma^2) \end{aligned} \tag{3.76}$$

where $p(\lambda_{ml}^2, \sigma^2 \,|\, \tilde{g})$ is the MAP posterior of the parameters given the data $\tilde{g}$ and $p\,(\tilde{g} \,|\, \lambda_{ml}^2, \sigma^2)$ is the likehood of the parameters.

With regards to equation (3.76), we view MAP as a substitute to maximizing the likelihood $p\,(\tilde{g} \,|\, \lambda_{ml}^2, \sigma^2)$ by maximizing the Bayesian posterior probability density of the set of parameters with the only difference arising from an introduction of a prior over the parameters we want to infer due to the knowledge we knew (or initial assumption we made) about the distribution over the parameters. If the prior distribution of λ_{ml}^2 and σ^2 are assumed to be independent then $p\,(\lambda_{ml}^2, \sigma^2) = p\,(\lambda_{ml}^2)\, p\,(\sigma^2)$. In addition to the independency, if both priors are assumed to be non-informative then estimates obtained from ML and MAP should coincide otherwise we expect MAP to out-perform ML due to the inclusion of priors.

CHAPTER 4

Take Home on Numerical and Statistical Regularization

Similarity in Equations Between Numerical and Statistical Regularization (for Gaussian random variables)

In chapter (2), the measurables were of the form

$$\tilde{g}(s) \;=\; \int_{\Omega} K(s,t) f(t)\,\mathrm{d}t \;=\; g(s) \;+\; \epsilon(s)$$

and we explained that the equation above is often related to a functional inequality $|\,\epsilon\,|$ bounded above such that

$$|\,\epsilon(s)| \;\leq \mathbb{M} \qquad \text{or} \qquad \int_{\Omega} \epsilon^2(s)\, w(s)\,\mathrm{d}s \;\leq\; \tilde{\mathbb{M}}: \quad w(s) \;>\; 0 \tag{4.1}$$

Also from equation (3.46), the functional equation is given by the marginal distribution of $\tilde{g}$ given (or conditional on) σ^2 and λ^2_{ml}. This is of the form

$$p(\tilde{g}|\lambda^2_{ml}, \sigma^2) \;=\; \int\; p(\tilde{g}\,|\,f, \sigma^2)\, p(f\,|\,\lambda^2_{ml})\,\mathrm{d}f \tag{4.2}$$

Both $\epsilon^2(s)$ and $p(\tilde{g}\,|\,f, \sigma^2)$ have the same quadratic functional forms and the corresponding weights $w(s)$ and $p(f\,|\,\lambda^2)$ are also quadratic functional regularizers. Without loss of generality, $|\,p(\tilde{g}\,|\,f, \sigma^2)\,|$ also satisfies

$$|\,p(\tilde{g}\,|\,f, \sigma^2)\,| \leq \hat{\mathbb{M}} \qquad \text{or} \qquad \int\; p(\tilde{g}\,|\,f, \sigma^2)\, p(f\,|\,\lambda^2)\,\mathrm{d}f \leq \hat{\tilde{\mathbb{M}}}$$

Sub-conclusion on Numerical and Statistical Framework

(a) The conditional mean in the Statistical framework/settings (i.e MAP) is the same as that of Regularized Normal Equations in the Numerical Methods framework if we are dealing with Multivariate Gaussian Distribution. We can consider the Regularized Normal Equations in the Numerical Methods framework settings as a special case of Stochastic Modelling theory when we are dealing with Gaussian Random Variables.

(b) The tuning parameter λ_{rls} in the Numerical Methods framework settings is a product of the Statistical parameter λ_{ml} and the noise level σ.

(c) Setting the regularization parameter λ_{ml} to zero for $\sigma < \infty$ is the same as finding a solution to a Least Squares problem.

CHAPTER 5

Numerical and Statistical Estimation Theory

The aim of this Chapter is to compare and contrast equations and expressions which leads to understanding the features and analysis of
(i) the L-Curve for Tikhonov Regularization in Numerical Ridge Regression.
(ii) the Emprical Bayes (Regularization) in Statistical (Bayesian) Ridge Regression

We already know from Chapters 2 and 3 that the standard Least Squares (LS) estimate f_{ls} is equivalent to the standard Maximum Likelihood estimate f_{ml}. Also, due to ill-posedness which is beyond both the standard LS and ML estimates, we extended the estimation procedure to be based on adding the small positive constant $\alpha = \lambda_{ml}^2$ to the singular values of the symmetric matrix $K^T K/\sigma^2$ or $K^T K$ so that the inverse matrix associated with $f_{\lambda_{rls}}$ or $f_{map_{\lambda_{ml},\sigma}}$ or $f_{ml_{\lambda_{ml},\sigma}}$ becomes non-singular.

We now turn our attention to the problem of how optimal estimates for λ_{rls}^2, λ_{ml}^2 and σ^2 can efficiently be determined. We shall continue to work under normality conditions.

5.1 The L-Curve for Tikhonov's Numerical Regularization

5.1.1 SVD for Tikhonov's Regularization

A decomposition of the standard regularized solution of equation (2.40) by SVD gives

$$f_{\lambda_{rls}} \quad = \quad \sum_{i=1}^{n} \Big(\frac{d_i^2}{d_i^2 \, + \, \lambda_{rls}^2} \Big) \frac{\mathbf{u}_i^T \tilde{g}}{d_i} \, \mathbf{v}_i \tag{5.1}$$

which is of the form

$$f_{\lambda_{rls}} \quad = \quad \sum_{i=1}^{n} x_i \, \frac{\mathbf{u}_i^T \tilde{g}}{d_i} \, \mathbf{v}_i \tag{5.2}$$

where

$$x_i = \frac{d_i^2}{d_i^2 \, + \, \lambda_{rls}^2} \quad ; \quad \forall \text{ i} \tag{5.3}$$

are called the Tikhonov's filter factors and they satisfy the inequality $0 < x_i < 1$. Another way of writing equation (5.3) is

$$\frac{d_i^2}{d_i^2 \, + \, \lambda_{rls}^2} = 1 \, - \, \frac{\lambda_{rls}^2}{d_i^2 \, + \, \lambda_{rls}^2} \tag{5.4}$$

(a) If $d_i >> \lambda_{rls}$, then $\lambda_{rls}^2/(d_i^2 \, + \, \lambda_{rls}^2) \, \rightarrow \, 0$. Hence

$$\frac{d_i^2}{d_i^2 \, + \, \lambda_{rls}^2} \approx 1 \tag{5.5}$$

(*b*) If $d_i << \lambda_{rls}$, then $\lambda_{rls}^2/(d_i^2 + \lambda_{rls}^2) \rightarrow 1$. Hence

$$\frac{d_i^2}{d_i^2 + \lambda_{rls}^2} \approx 1 - 1 = 0 \tag{5.6}$$

(*c*) For a δ-neighbourhood (δ-small s.t $\delta > 0$), if $d_i = \lambda_{rls} \pm \delta$, then the filter factor x_i are in transition between the two extreme regions of (*a*) and (*b*) above. Hence,

$$\frac{d_i^2}{d_i^2 + \lambda_{rls}^2} \approx 1 - 1/2 = 1/2 \tag{5.7}$$

The naive solution $\tilde{f}$ is obtained when $\lambda_{rls} = 0$. The SVD components corresponding to $d_i > \lambda_{rls}$ contributes strength that is more than half of the naive case $\tilde{f}$. For case(s) where $d_i >> \lambda_{rls}$, it contributes with almost full strength to the solution $f_{\lambda_{rls}}$. On the otherhand, the SVD components corresponding to singular values $d_i < \lambda_{rls}$ are damped considerably and contribute very little to the solution $f_{\lambda_{rls}}$. Hence, the truncation parameter k has a relation with λ_{rls}^2 given by $d_k \approx \lambda_{rls}$. For more on this see [2].

5.1.2 Analysis of L-Curve for Tikhonov Regularization

The analysis to be presented here is due Hansen. We will not dwell much into the surrounding details. For a thorough discussion on this, kindly see [2]. The analysis starts by writing $\tilde{g}$ as the sum of an exact unperturbed data $\bar{g}$ and noise ϵ;

$$\tilde{g} \quad = \quad \bar{g} + \epsilon \quad \text{and} \quad \bar{g} = K\bar{f} \tag{5.8}$$

where $\bar{f} = K^{\dagger}\bar{g}$. The Tikhonov solution is expressed as

$$f_{\lambda_{rls}} \quad = \quad \bar{f}_{\lambda_{rls}} + f^{\epsilon}_{\lambda_{rls}} \tag{5.9}$$

where $\bar{f}_{\lambda_{rls}}$ is the regularized version of the exact solution $\bar{f}$ and it is also given by

$$\bar{f}_{\lambda_{rls}} \quad = \quad (K^T K + \lambda_{rls}^2 I)^{-1} K^T \bar{g} \tag{5.10}$$

The Least Squares solution $\bar{f} = K^{\dagger}\bar{g}$ to the unperturbed problem satisfies the discrete picard condition and for that matter $|\,\mathbf{v}_i^T \bar{f}\,| = |\,\mathbf{u}_i^T \bar{g}/d_i\,|$ also decay. The residual norm which is characterized by data misfit $\tilde{g}^{\epsilon}_{\lambda_{rls}}$ is given by

$$\| \tilde{g}^{\epsilon}_{\lambda_{rls}} \|_2^2 = \| \tilde{g} - K f_{\lambda_{rls}} \|_2^2 \quad = \quad \sum_{i=1}^{n} \left((1 - x_i)\, \mathbf{u}_i^T \tilde{g} \right)^2 \tag{5.11}$$

The norm of the deviation $f^{\epsilon}_{\lambda_{rls}}$ from the estimates is

$$\begin{aligned} \| f^{\epsilon}_{\lambda_{rls}} \|_2^2 = \| \bar{f}_{\lambda_{rls}} - f_{\lambda_{rls}} \|_2^2 \quad &= \quad \sum_{i=1}^{n} \left(\frac{d_i \sigma}{d_i^2 + \lambda_{rls}^2} \right)^2 \\ &\approx \quad \sigma^2 \left\{ \sum_{i=1}^{k} \left(\frac{1}{d_i} \right)^2 + \sum_{i=k+1}^{n} \left(\frac{d_i}{\lambda_{rls}^2} \right)^2 \right\} \end{aligned} \tag{5.12}$$

It follows from the solution norm $\| f_{\lambda_{rls}} \|_2^2 = \sum_{i=1}^{m} (x_i \frac{\mathbf{u}_i^T g}{d_i})^2$ that the norm of the regularized version of the exact solution can be given by the approximation

$$\| \bar{f}_{\lambda_{rls}} \|_2^2 \approx \sum_{i=1}^{k} \left(\mathbf{v}_i^T \bar{f} \right)^2 \approx \sum_{i=1}^{n} \left(\mathbf{v}_i^T \bar{f} \right)^2 = \| \bar{f} \|_2^2 \tag{5.13}$$

where the last $n-k$ terms contribute very little to the sum. Hence
(i) if $\lambda_{rls} \to \infty$, and $k \to 0$ we have $\bar{f}_{\lambda_{rls}} \to 0$ which implies that $\| \bar{f}_{\lambda_{rls}} \|_2 \to 0$.
(ii) if $\lambda_{rls} \to 0$, then $\| \bar{f}_{\lambda_{rls}} \|_2 \to \| \bar{f} \|_2$

The residual corresponding to $\bar{f}_{\lambda_{rls}}$ then satisfies

$$\| \bar{g} - K \bar{f}_{\lambda_{rls}} \|_2^2 = \sum_{i=k}^{n} \left(\mathbf{u}_i^T \bar{g} \right)^2 \tag{5.14}$$

Therefore, the L-Curve for the unperturbed problem is a flat curve at $\| \bar{f}_{\lambda_{rls}} \|_2 \approx \bar{f}$ except for large values of the residual norm $\| \bar{g} - K \bar{f}_{\lambda_{rls}} \|_2$ where the curve approaches the abscissa axis.

Finally, the first and second sum of $\| f_{\lambda_{rls}}^{\epsilon} \|_2^2$ in equation (5.12) are respectively dominated by $d_k^{-2} \approx \lambda_{rls}^{-2}$ and $d_{k+1}^2 \approx \lambda_{rls}^2$ and can be approximated by

$$\| f_{\lambda_{rls}}^{\epsilon} \|_2^2 \approx \varpi_{\lambda_{rls}} \sigma / \lambda_{rls} \tag{5.15}$$

where $\varpi_{\lambda_{rls}}$ is a quantity that varies slowly with λ_{rls}. Hence, $f_{\lambda_{rls}}^{\epsilon}$ increases monotonoically from 0 as λ_{rls} decreases until

$$\| K^{\dagger} \epsilon \|_2 \approx \sigma \| K^{\dagger} \|_F \quad \text{for} \quad \lambda_{rls} \to 0 \quad \text{is attained.}$$

The corresponding residuals satisfies

$$\| K f_{\lambda_{rls}}^{\epsilon} - \tilde{g} \|_2^2 \approx \sum_{i=k}^{n} \sigma^2 = (n-k)\sigma^2 \tag{5.16}$$

Hence, $\| K f_{\lambda_{rls}}^{\epsilon} - \epsilon \|_2 \approx \sigma \sqrt{n-k}$ is a slowly varying function of λ_{rls} which lies in the range from 0 to $\| \epsilon \|_2 \approx \sigma \sqrt{n}$. Therefore the L-Curve for ϵ is an overall very steep curve located slightly to the left of $\| K f_{\lambda_{rls}}^{\epsilon} - \epsilon \|_2 \approx \| \epsilon \|_2$, except for small values of λ_{rls} where it approaches the ordinate axis.

It is emphasized that the analysis is valid only when the L-Curve is plotted in log-log scale and that it is a plot of $\frac{1}{2} \ln \| f_{\lambda_{rls}} \|_2^2$ versus $\frac{1}{2} \ln \| K f_{\lambda_{rls}} - \tilde{g} \|_2^2$. It is further assumed that the noise is a scalar multiple of the identity matrix I. So the expected values of the SVD coefficients of $\mathbf{u}_i^T \epsilon$ are independent of i;

$$\langle (\mathbf{u}_i^T \epsilon)^T (\mathbf{u}_i^T \epsilon) \rangle = \sigma^2 \quad ; \quad i = 1, 2, ..., n \tag{5.17}$$

5.1.3 The L-Curve Method for Estimating the Parameter λ^2_{rls}

The L-Curve is a log-log plot of the norm of the regularized solution $\| f_{\lambda_{rls}} \|_2$ versus the norm of the corresponding residual norm $\| K f_{\lambda_{rls}} - \tilde{g} \|_2$. Thus, from the set of parameter values $\{ \lambda^2_{rls} \}$, if we let η and ρ be represented respectively by

$$\eta = \| f_{\lambda_{rls}} \|_2^2 \quad ; \quad \rho = \| K f_{\lambda_{rls}} - \tilde{g} \|_2^2 \tag{5.18}$$

and let

$$\hat{\eta} = \ln \eta \quad ; \quad \hat{\rho} = \ln \rho \tag{5.19}$$

Then

$$\eta = \exp \hat{\eta} = \| f_{\lambda_{rls}} \|_2^2 \quad \text{and} \quad \rho = \exp \hat{\rho} = \| K f_{\lambda_{rls}} - \tilde{g} \|_2^2 \tag{5.20}$$

Hence

$$\hat{\eta} = \ln \| f_{\lambda_{rls}} \|_2^2 \quad \text{and} \quad \hat{\rho} = \ln \| K f_{\lambda_{rls}} - \tilde{g} \|_2^2 \tag{5.21}$$

Hansen in [1] derived an expression for the curvature $\varkappa$ of the L-Curve as a function of λ_{rls} and had

$$\varkappa = \frac{2 \eta \rho}{\eta'} \frac{\lambda^2 \eta' \rho + 2 \lambda \eta \rho + \lambda^4 \eta \eta'}{(\lambda^2 \eta^2 + \rho^2)^{3/2}} \tag{5.22}$$

where

$$\lambda = \lambda_{rls} \quad , \quad \eta' = \frac{-4}{\lambda} \sum_{i=1}^{n} (1 - x_i) x_i^2 \frac{\omega_i^2}{d_i^2} \quad \text{and} \quad \omega_i = \mathbf{u}_i^T \tilde{g}$$

The strategy of choosing the best estimate for the regularization parameter $\hat{\lambda}^2_{rls}$ lies at the corner. The corner separates the flat and the vertical parts of the curve where the solution is dominated by regularization and perturbation errors.

5.2 Empirical Bayes for Statistical Ridge Regression

Hoerl and Kennard (1970), were the first to propose the Ridge Regression estimator of f_{ls};

$$f_{\lambda_{eb}} \quad = \quad (K^T K + \lambda^2_{eb} I)^{-1} K^T \tilde{g} \quad \lambda_{eb} > 0 \tag{5.23}$$

where $f_{\lambda_{eb}}$ is the regularized solution and λ^2_{eb} is the regularization constant. [1] They further used the term instability to signify that $\langle f_{ls}^T f_{ls} \rangle$ is too large or much larger than $\| f \|_2^2$. We view the Ridge Estimate $f_{\lambda_{rls}}$ in the above context as an extension of the standard Least Squares or standard ML estimator when $K^T K$ have at least one singular value to be small. A small singular value d_i ($d_i \to 0$) of K tends to make the Least Squares estimator unstable in the sense that small changes in $\tilde{g}$ may produce large changes in f_{ls}.

We now present the first of two types of Statistical Bayes Ridge Regression namely Truncated SVD for Empirical Bayes Ridge Regression.

[1] the subcript eb stands for Empirical Bayes and λ_{eb} is the empirical bayes regularization parameter.

5.2.1 SVD for Empirical Bayes

We can re-write equation (5.23) in the following alternative forms.

$$f_{\lambda_{eb}} = WK^T\tilde{g} \tag{5.24}$$

where $W = (K^TK + \lambda_{eb}^2 I)^{-1}$. The alternative form of equation (5.24) is

$$\begin{aligned} f_{\lambda_{eb}} &= \left[I + \lambda_{eb}^2 (K^TK)^{-1}\right]^{-1} (K^TK)^{-1}K^T\tilde{g} \\ &= Y\, f_{ls} \end{aligned} \tag{5.25}$$

where $Y = \left[I + \lambda_{eb}^2 (K^TK)^{-1}\right]^{-1}$

By further manipulating Y, we get

$$\begin{aligned} Y &= \left[K^TK + \lambda_{eb}^2 I\right]^{-1} K^TK \\ &= I - \lambda_{eb}^2 [K^TK + \lambda_{eb}^2 I]^{-1} \quad \text{by simple long division arithmetic} \\ &= I - \lambda_{eb}^2\, W \end{aligned} \tag{5.26}$$

We also let $\xi(W)$ and $\xi(Y)$ be singular values of W and Z such that

$$\xi(W) = \sum_{i=1}^{n} \frac{1}{d_i^2 + \lambda_{eb}^2} \tag{5.27}$$

$$\xi(Y) = \sum_{i=1}^{n} \frac{d_i^2}{d^2 + \lambda_{eb}^2} \tag{5.28}$$

The SVD of $f_{\lambda_{eb}}$ of equation (5.24) is

$$f_{\lambda_{eb}} = \sum_{i=1}^{n} \mathbf{v}_i \frac{d_i^2}{d^2 + \lambda_{eb}^2} \frac{\mathbf{u}_i^T\tilde{g}}{d_i} \tag{5.29}$$

and it is comparatively of the same form as equation (5.1). The estimator $f_{\lambda_{eb}}$ depends on the choice of the corresponding precision parameter λ_{eb}^2 and it is generally not guaranteed to be better than f_{ls} in terms of risk under any quadratic loss. In view of this, we seek to produce *minimax adaptive ridge-regression* [2] that are uniformly better than the Least Squares estimator.

5.2.2 Truncated SVD for Empirical Bayes

The inverse of the matrix K^TK is of concern to us here since the solution depends on it. The singular values of $(K^TK)^{-1}$ is obtainable from

$$V^T(K^TK)^{-1}V = D^{-2} \tag{5.30}$$

[2] The term minimax is used to refer to an estimator that is uniformly better than the Least Squares estimator and the word adaptive indicates that the ridge constant is estimated from data.

where $D^{-2} = \text{diag}(d_n^{-2}, d_{n-1}^{-2}, d_{n-2}^{-2}, ..., d_1^{-2})$ and $d_n^{-2} > d_{n-1}^{-2} > d_{n-2}^{-2} >, ..., > d_1^{-2}$. We partition V such that

$$V = (V_1, V_2) \quad \text{where} \quad V_1 \in \mathbb{R}^{n \times k} \quad \text{and} \quad V_2 \in \mathbb{R}^{n \times (n-k)} \quad \text{for some k.} \tag{5.31}$$

We write

$$\begin{aligned} f = VV^T f &= V_1 V_1^T(f) + V_2 V_2^T f \\ &= V_1 \varsigma_{eb} + V_2 \gamma \end{aligned} \tag{5.32}$$

where ς_{eb} corresponds to the smaller singular values of $K^T K$.

It is desirable to impose the constraint

$$f = V_2 \gamma \tag{5.33}$$

We construct ridge-type regression estimators using the information about which singular values are smaller in some sense. This we do by

$$\begin{aligned} V^T (K^T K) V = D^2 &= \text{diag}(d_1^2, d_2^2, d_3^2, ..., d_n^2) \\ &= \begin{pmatrix} D_1^2 & 0 \\ 0 & D_2^2 \end{pmatrix} \end{aligned} \tag{5.34}$$

where $D_1^2 = \text{diag}(d_1^2, d_2^2, d_3^2, ..., d_{n-k}^2)$ and D_2^2 is an $(n-k+1) \times (n-k+1)$ diagonal matrix consisting of the smaller singular values. Since ς_{eb} of equation (5.32) corresponds to small singular values and for this reason must not be included in the model, we shrink f_{ls} towards the linear constraints

$$H_0 : f = V_2 \gamma \quad \gamma \in \mathbb{R}^{n-k} \tag{5.35}$$

Using the decompositions given in equations (5.31) and (5.34), the estimate of γ is

$$\hat{\gamma} = V_2^T f_{ls} \tag{5.36}$$

The truncated SVD (or principal component) regression estimator of f is given by

$$f^{PC} = V_2 \underbrace{V_2^T f_{ls}}_{\hat{\gamma}} \tag{5.37}$$

Also from equations (5.36) and (5.37) we have

$$V_2 \hat{\gamma} = f^{PC} \tag{5.38}$$

By conveniently treating them in canonical form, we can choose to let $z = V^T f_{ls}$. Then

$$z \sim \mathbb{N}(V^T f, \sigma^2 D^{-2}) \tag{5.39}$$

5.2.3 Analysis of Empirical Bayes Estimators

Let $\xi_i(Y)$ be the $i^{th}-$diagonal singular value of equation (5.28). For $\lambda_{eb} > 0$, we have

$$\max_i \xi_i(Y) \quad = \quad \frac{{d_1}^2}{{d_1}^2 + \lambda_{eb}^2} < 1 \tag{5.40}$$

where ${d_1}^2$ is the largest singular value of $K^T K$ and that equation (5.40) equals 1 if and only if $\lambda_{eb} = 0$. Hence

$$\| f_{\lambda_{eb}} \|_2^2 \quad < \quad \| f_{ls} \|_2^2 \tag{5.41}$$

Also from equations (5.28) and (5.26)

$$\begin{aligned} \lim_{\lambda_{eb}\to\infty} \frac{{d_i}^2}{{d_i}^2 + \lambda_{eb}^2} &= \lim_{\lambda_{eb}\to\infty} 1 - \frac{\lambda_{eb}^2}{{d_i}^2 + \lambda_{eb}^2} \quad ; \quad \forall\, i \\ &= 0 \end{aligned} \tag{5.42}$$

That is, when $\lambda_{eb} \to \infty$, the largest singular value ${d_1}^2$ becomes insignificant and the solution $f_{\lambda_{eb}}$ is independent on i. Hence the truncation parameter k also approaches zero ($k \to 0$) and the corresponding norm of the regularized solution is also zero.

Method of Estimating the Parameter λ_{eb}^2

This estimator shrinks the Least Squares estimator towards the principal components. The Bayes estimator $f_{\lambda_{eb}}$ is

$$f_{\lambda_{eb}} \quad = \quad f_{ls} - \Big(I + \frac{1}{\lambda_{eb}^2} K^T K\Big)^{-1} \Big(f_{ls} - V_2 \gamma\Big) \tag{5.43}$$

with the estimate of γ given by the weighted Least Squares estimator

$$\hat{\gamma} = \Big(V_2^T K^T K\, V_2\Big)^{-1} V_2^T K^T K\, f_{ls} \tag{5.44}$$

where $\hat{\gamma}$ is obtainable from the minimization of the weighted squared loss

$$\Big(f_{ls} - V_2\gamma\Big)^T K^T K \Big(f_{ls} - V_2\gamma\Big) \tag{5.45}$$

The best variable estimate which we denote by $\hat{f}_{\lambda_{eb}}$ is

$$\hat{f}_{\lambda_{eb}} \quad = \quad f_{ls} - \Big(I + \frac{1}{\lambda_{eb}^2} K^T K\Big)^{-1} \Big(f_{ls} - f_{ls}^{PC}\Big) \tag{5.46}$$

and

$$1/\hat{\lambda}_{eb}^2 \quad = \quad \max(1/\lambda_{eb}^{2\star}, 1/\lambda_0^2) \tag{5.47}$$

where $\lambda_{eb}^{2\star}$ is given by the root of the equation

$$\Big(f_{ls} - f_{ls}^{PC}\Big)^T \Big((K^T K)^{-1} + \frac{1}{\lambda_{eb}^{2\star}} I\Big)^{-1} \Big(f_{ls} - f_{ls}^{PC}\Big) \quad = \quad \frac{(n-k-2)}{n+2} \Big\{ \Big(\tilde{g} - K f_{ls}\Big)^T \Big(\tilde{g} - K f_{ls}\Big) \Big\} \tag{5.48}$$

where $f_{ls}^{PC} = f^{PC}$ and λ_0 is also the root of the equation

$$\sum_{i=1}^{k+1} \frac{(d_{n-i})^{-2} - (d_{k+1})^{-2}}{(d_{n-i})^{-2} + (1/\lambda_0)^2} = (k-1)/2 \tag{5.49}$$

The above estimator $f_{\lambda_{eb}}$ incorporates both methods of Ridge Regression and truncated SVD. This should however be viewed as an extension of truncated SVD. For more on this see an example in [38].

5.3 Bayesian Inference for Statistical Bayes Ridge Regression

In general, the statistical information envisaged in parameter estimation (like we have $\tilde{g}$) gives some evidence concerning some hypothesis say $H_1, H_2, ,,.$ (since H might be the statement that its parameter(s) lies within an interval) and we make inferences about them solely from what we observe. The very act of choosing a model by sampling distribution conditional on H is considered as a means of expressing some kind of prior knowledge about the existence and nature of H and its observable effects.

In effect, we see it as a rule for constructing informative priors when we have partial prior information that restricts the possibility significantly but not completely. In contrast to Bayesian Inference, D. C Mackay in his book "Information Theory, Inference and Learning Algorithms" [6] argues

"Once we have made explicit all our assumptions about the model and the data, our inferences are mechanical. Whatever question we wish to pose, the rules of probability theory give a unique answer which consistently takes into account all the given information"

Nevertheless, Bayesian Inference tends to imitate both Sampling theory and even Numerical Methods in that it incorporates little or no prior information beyond the choice of the model and so seeks "non-informative" priors, otherwise it is expected to out-perform Sampling Methods only when the latter faces a problem like insufficient (or small) data.

5.3.1 The Evidence Framework and Occam Razor

The Framework (due techniques developed by Gull and Skilling), integrates over the precision model parameters $\alpha = \lambda_{ml}^2$ and $\beta = 1/\sigma^2$ and the resulting evidence maximized over the hyperparameters. The hyperparameters are then used to define a Gaussian approximation to the posterior distribution. The Bayesian Adaptive learning begins with the probability of everything;

$$p(\tilde{g}\,,\, f\,,\, \alpha\,,\, \beta) = p(\tilde{g}\,,\, f\,,\, H_i\,) \tag{5.50}$$

where $H_i = \{\alpha, \beta\}$ is a sub-model of the hypothetical space H. Two levels of inference are involved in the Ridge Regression task;

(*i*) Model fitting where we infer f by obtaining a compact representation for model H_i

$$\begin{aligned} p(f \,|\, \tilde{g}\,, \alpha\,, \beta) &= \frac{\overbrace{p\,(\tilde{g} \,|\, f\,, \beta)}^{likelihood}\ \overbrace{p\,(f \,|\, \alpha)}^{prior}}{\underbrace{p\,(\tilde{g} \,|\, \alpha, \beta)}_{evidence}} \\ &= \frac{p\,(\tilde{g} \,|\, f\,, H_i)\, p\,(f \,|\, H_i)}{p\,(\tilde{g} \,|\, H_i)} \end{aligned} \tag{5.51}$$

The error bars are obtainable from Taylor expanding the log-posterior about the most probable f_{MP}:

$$p(f \,|\, \tilde{g}, H_i) \approx p(f_{MP} \,|\, \tilde{g}, H_i) \exp -\frac{1}{2}(f - f_{MP})^T \Sigma_f \,(f - f_{MP}) \tag{5.52}$$

(*ii*) Given a collection of models of H_i, we wish to find our initial beliefs about the relative plausibilities in terms of a list of quatified $p(H_i)$ such that

$$\sum_i p(H_i) = 1 \tag{5.53}$$

We use Bayes' rule to update our belief in the models in the light of $\tilde{g}$. We do model comparison using the relation

$$\begin{aligned} p(H_i \,|\, \tilde{g}) \,=\, p(\alpha, \beta \,|\, \tilde{g}) \,=\, \frac{p(\tilde{g} \,|\, H_i)\, p(H_i)}{p(\tilde{g})} &\propto p(\tilde{g} \,|\, H_i)\, p(H_i) \\ &= p(\tilde{g} \,|\, \alpha, \beta)\, p(\alpha, \beta) \quad ; \quad \forall \mathrm{i} \end{aligned} \tag{5.54}$$

The denominator

$$p(\tilde{g}) = \sum_i p(\tilde{g} \,|\, H_i)\, p(H_i) \tag{5.55}$$

makes our final beliefs $p(H_i \,|\, \tilde{g})$ adds up to 1. In the light of $\tilde{g}$, the relative plausibility of any two alternatives say H_1 and H_2 is obtainable from

$$\frac{p(H_1 \,|\, \tilde{g})}{p(H_2 \,|\, \tilde{g})} = \frac{p(\tilde{g} \,|\, H_1)\, p(H_1)}{p(\tilde{g} \,|\, H_2)\, p(H_2)} \tag{5.56}$$

Their normalizing constants are the same so they cancel out. The ratio $p(H_1 \,|\, \tilde{g})/p(H_2 \,|\, \tilde{g})$ measures how our initial beliefs favour H_1 over H_2. The ratio $p\,(H_1)/p\,(H_2)$ will also cancel out if we have no reason to assign different priors for $p(H_1)$ and $p(H_2)$. Finally, the ratio $p(\tilde{g} \,|\, H_1)/p(\tilde{g} \,|\, H_2)$ expresses how well $\tilde{g}$ is predicted by H_1 compared to H_2.

Figure (5.1) is a schematic diagram of the marginal likelihoods for a complex, too simple and "just ok" models. The more complex models are able to describe a greater range of a given data set. However, for a given data $\tilde{g}$, the "just ok" model has a greater evidence than either the too simple model or the too complex model. Thus, model complexity is governed by Occam Razor which tends to favour neither too simple nor too complex models.

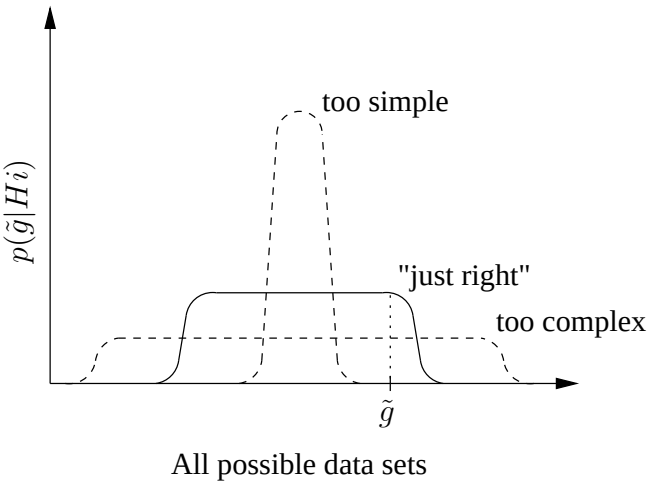

Figure 5.1: A schematic diagram taken from [Zoubin, Mackay, Bishop] with an adjustment to suit the explanation to the evidence Framework descibed above. It is a plot of the marginal likelihood versus $\tilde{g}$. It shows for a given $\tilde{g}$, the corresponding marginal likelihood $p(\tilde{g}\,|\,H_i)$ for a too simple model, too complex model and 'just right'. The more complex model is able to describe a greater range of data set and vice-versa for a too simple one.

5.3.2 Evaluation of the Evidence and Occam Factor

The evidence [3] is the normalizing constant at the first level of inference and it is given by

$$p(\tilde{g}\,|\,\alpha,\beta) \quad = \quad \int p(\tilde{g}\,|\,f,\beta)\,p(f\,|\,\alpha)\mathrm{d}f \tag{5.57}$$

The posterior $p(f\,|\,\tilde{g},\alpha,\beta)$ is proportional to the integrand $p(\tilde{g}\,|\,f,\beta)\,p(f\,|\,\alpha)$ of equation (5.57). In the ML principle, the distribution of $p(\tilde{g}\,|\,\alpha,\beta)$ is sharply peaked around the most probable variable f_{MP}. Hence the evidence can be approximated by

$$p(\tilde{g}\,|\,\alpha,\beta) \quad \approx \quad p(\tilde{g}\,|\,f_{MP},\beta)\,p(f_{MP}\,|\,\alpha)\,\Delta f \tag{5.58}$$

where Δf is the width and $p(f_{MP}\,|\,\alpha)$ is the prior which can be imagined to be uniform on some large interval. Therefore,

$$p(f_{MP}\,|\,\alpha) \quad = \quad \frac{1}{\Delta^0 f} \tag{5.59}$$

with the *Occam factor* given by

$$\Delta f/\Delta^0 \tag{5.60}$$

The n-dimensional posterior distribution is well approximated by a Gaussian with the corresponding Occam factor obtained from the determinant of the Gaussian covariance matrix:

$$p(\tilde{g}\,|\,\alpha,\beta) \quad \approx \quad \underbrace{p(\tilde{g}\,|\,f_{MP},\beta)}_{best\ fit\ likelihood}\ \underbrace{p(f_{MP}\,|\,\alpha)\,(2\,\pi)^{n/2}\,|\,\Sigma_f\,|^{-1/2}}_{Occam\ factor} \tag{5.61}$$

[3] The evidence is the same as the marginal likelihood of equation (3.51) except that the parameters have been defined in terms of the precision parameters α and β.

The *evidence* is approximately evaluated from the following: Let

$$Z_{\tilde{g}}(\beta) \,=\, (2\,\pi/\beta)^{n/2}\;;\quad Z_f(\alpha) \,=\, (2\,\pi/\alpha)^{n/2}\;;\quad E_{\tilde{g}} \,=\, \frac{1}{2}\,\|\,\tilde{g} - K f\,\|_2^2\;;\quad E_f \,=\, \frac{1}{2}\,\|\,f\,\|_2^2 \tag{5.62}$$

and also let

$$M(f,\alpha,\beta) \,=\, \beta\,E_{\tilde{g}} \,+\, \alpha\,E_f \tag{5.63}$$

Then we can write equation (5.57) as

$$p(\tilde{g}\,|\,\alpha,\beta) \quad = \quad \frac{Z_M(f,\alpha,\beta)}{Z_f(\alpha)\,Z_{\tilde{g}}(\beta)} \tag{5.64}$$

where

$$Z_M(f,\alpha,\beta) = \int \exp\,-\Big\{M(f,\alpha,\beta)\Big\} \tag{5.65}$$

Taylor expanding M about f_{MP} to second order gives

$$M \quad = \quad M(f_{MP}) \,+\, \frac{1}{2}\,(f - f_{MP})^T \Sigma_f (f - f_{MP}) \tag{5.66}$$

Substituting into equation (5.65) and solving gives

$$Z_M(f,\alpha,\beta) \quad = \quad (2\,\pi)^{n/2}\,|\,\Sigma_f\,|^{-1/2} \exp\,-M(f_{MP}) \tag{5.67}$$

The general form of writing the log-marginal likelihood of equation (5.64) to embody non-quadratic regularizer functional using a Gaussian approximation is

$$\begin{aligned} \ln p(\tilde{g}\,|\,\alpha,\beta) \quad &= \quad -\ln\,Z_f(\alpha) \,-\, \alpha\,E_f^{MP} \,-\, \frac{1}{2}\,\ln|\,\Sigma_f\,| + \frac{n}{2}\ln(2\,\pi) \\ &\quad\; -\,\ln\,Z_{\tilde{g}}(\beta) \,-\, \beta\,E_{\tilde{g}}^{MP} \end{aligned} \tag{5.68}$$

with $\beta E_{\tilde{g}}^{MP}$ representing the misfit of the interpolant (or filter) to $\tilde{g}$ and $\alpha\,E_f^{MP}$ measuring how far f_{MP} is from its null value. The Occam factor is

$$\frac{\Delta f}{\Delta^0 f} = \frac{(2\,\pi)^{n/2}\,|\,\Sigma_f\,|^{-1/2}}{Z_f(\alpha)} \tag{5.69}$$

5.3.3 Analysis and Method of Estimating α and β

From the evidence approximation of equation (5.68), we can now differentiate the log-evidence to get optimal estimates for α and β .

Differentiating with respect to α

$$\begin{aligned} \frac{d}{d\alpha}\,\ln p\,(\tilde{g}\,|\,\alpha,\beta) \quad &= \quad -E_f^{MP} \,-\, \frac{1}{2}\,\mathrm{Tr}\left(\Sigma_f^{-1}\,\frac{d\Sigma_f}{d\alpha}\right) + \frac{n}{2\,\alpha} \\ &= \quad -E_f^{MP} \,-\, \frac{1}{2}\,\mathrm{Tr}\left(\Sigma_f^{-1}\right) + \frac{n}{2\,\alpha} \end{aligned} \tag{5.70}$$

Setting the derivatives to zero, it is straight forward that the maximum satisfies

$$\Psi \,=\, 2\,\alpha\,E_f^{MP} \quad = \quad n \,-\, \alpha\,\mathrm{Tr}\,\Sigma_f^{-1} \tag{5.71}$$

Solving for α from equation (5.71) gives

$$\begin{aligned} \alpha &= \frac{\mathrm{n}/2}{\frac{1}{2}\,\mathrm{Tr}\,(\Sigma_f^{-1}) + E_f^{MP}} \\ &= \frac{\mathrm{n}}{\mathrm{Tr}\,(\Sigma_f^{-1}) + \| f_{MP} \|_2^2} \end{aligned} \tag{5.72}$$

with E_f^{MP} equivalent to its corresponding expression of equation (5.62). The quantity Ψ is a dimensionless measure which can be interpreted as some sort of χ_f [4] for f since it can be written in the form;

$$\chi_f = \frac{1}{\sigma_f^2} \| f_{MP} \|_2^2 = 2\,\alpha\, E_f^{MP} \tag{5.73}$$

where $\sigma_f^2 = 1/\alpha$ is the variance of f from the null value of the fitted parameters. Another way of expressing Ψ of equation (5.71) is

$$\begin{aligned} \Psi &= \mathrm{n} - \sum_{i=1}^{n} \frac{\alpha}{\beta\, {d_i}^2 + \alpha} \\ &= \sum_{i=1}^{n} \frac{\beta\, {d_i}^2}{\beta\, {d_i}^2 + \alpha} \end{aligned} \tag{5.74}$$

with $\lambda_{rls}^2 = \alpha/\beta = \lambda_{ml}^2\,\sigma^2$.
Differentiating with respect to β.

$$\begin{aligned} \frac{d}{d\beta} \ln p\,(\tilde{g}\,|\,\alpha,\beta) &= -E_{\tilde{g}}^{MP} - \frac{1}{2}\,\mathrm{Tr}\left(\Sigma_f^{-1}\,\frac{d\,\Sigma_f}{d\,\beta}\right) + \frac{n}{2\,\beta} \\ &= -E_{\tilde{g}}^{MP} - \frac{1}{2}\,\mathrm{Tr}\left(K\,\Sigma_f^{-1}\,K^T\right) + \frac{n}{2\,\beta} \end{aligned} \tag{5.75}$$

Setting equation (5.75) to zero and manipulating gives

$$\begin{aligned} \Xi = 2\,\beta\, E_{\tilde{g}}^{MP} &= n - \beta \sum_{i=1}^{n} \frac{{d_i}^2}{\beta\, {d_i}^2 + \alpha} \\ &= \sum_{i=1}^{n} 1 - \frac{\beta\, {d_i}^2}{\beta\, {d_i}^2 + \alpha} \end{aligned} \tag{5.76}$$

Solving for β from equation (5.75) gives

$$\begin{aligned} \beta &= \frac{n/2}{E_{\tilde{g}}^{MP} + \frac{1}{2}\,\mathrm{Tr}\,(K\,\Sigma_f^{-1}\,K^T)} \\ &= \frac{n}{\| \tilde{g} - K f_{MP} \|_2^2 + \mathrm{Tr}\,(K\,\Sigma_f^{-1}\,K^T)} \end{aligned} \tag{5.77}$$

where we have substituted the equivalence of $E_{\tilde{g}}^{MP}$ in equation (5.62).

[4] For N independent Gaussian variables with mean μ and standard deviation σ, the statistic $\chi = \sum \frac{(x-\mu)^2}{\sigma^2}$ is a measure of misfit.

5.3.4 Analytical Interpretation

The quantity $\Psi = n - \alpha \operatorname{Tr} \Sigma_f^{-1}$ of equation (5.71) is the number of good parameter measurements and has value between 0 and n. The quantity α measures how strongly the parameters are determined by the prior. Thus, $(\forall i \text{ and } \alpha > 0) \quad \beta\, d_i^2/(\beta\, d_i^2 + \alpha)$ is a number between 0 and 1 which measures the strength of the data relative to the prior in the i direction. A direction in parameter space for which $d_i{}^2\beta$ (or d_i^2/σ^2) [5] is small compared to α does not contribute to the number of good parameter measurements. As $\alpha/\beta \to 0$, χ_f increases from 0 to n.

5.4 Variational Inference Methods

This is a technique one can employ whenever a complex or complicated distribution is encountered in a statistical data modelling task. It evolves around Gibbs inequality method and is often associated with the Kullback and Leibler divergence theorem $[\, D_{KL}(Q \,\|\, P)\,]$ between two probability distribution say $Q(X)$ and $P(X)$. Mathematically, $D_{KL}(Q \,\|\, P)$ (also called Relative Entropy) over the same alphabet say $\mathbb{A}_X$is defined as

$$\begin{aligned} D_{KL}(Q \,\|\, P) &= \int Q(X) \ln \frac{Q(X)}{P(X)}\, \mathrm{d}X \\ &= -\int Q(X) \ln \frac{P(X)}{Q(X)}\, \mathrm{d}X \\ &= -\mathcal{F}(Q) \end{aligned} \tag{5.78}$$

where

$$\mathcal{F} = \int Q(X) \ln \frac{P(X)}{Q(X)}\, \mathrm{d}X \tag{5.79}$$

Equation (5.78) satisfies

$$D_{KL}(Q \,\|\, P) \geq 0$$

with equality if and only if $Q = P$. [6]

Variational Inference in its own world (from Statistical Physics) attempts to approximate an integrand until the integral becomes tractable. The idea is to either bound the integrand from above or below so that the integral can be reduced to an optimization problem. No parameter estimation is required and the quantity of the integral is optimized directly. It further allows flexibility in specifying the prior and makes provision for attaining bounds on the value of the evidence. See Figure (5.2) for an illustration.

The methods to be presented in this section is due Hinton and van Camp.

[5] We recall that $d_i{}^2\beta = d_i{}^2/\sigma^2$ is the scaling factor of the exact expression for $\Sigma_{f_{\lambda,\sigma}}$ in equation (3.38)

[6] In general $D_{KL}(Q \,\|\, P) \neq D_{KL}(P \,\|\, Q)$

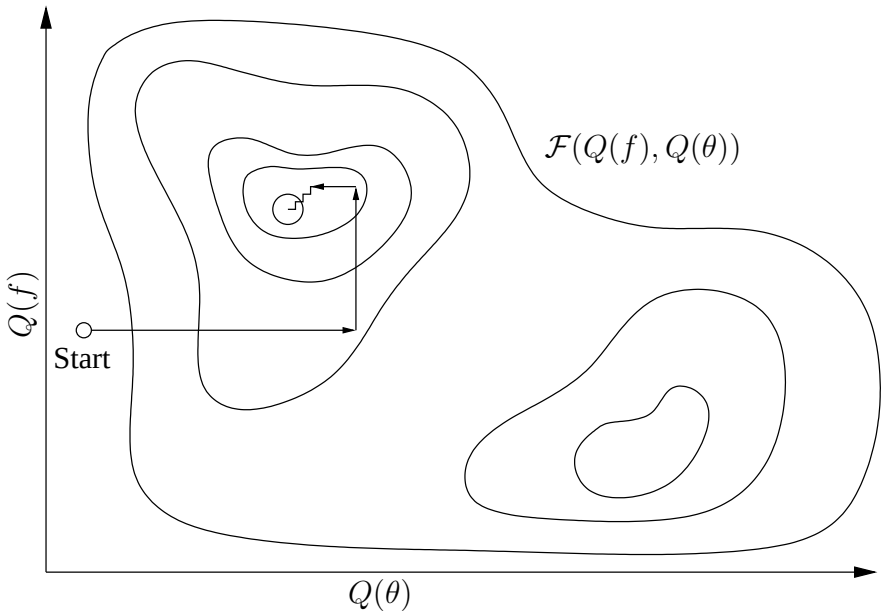

Figure 5.2: An illustration to show that Variational Methods is a coordinate ascent algorithm in $\mathcal{F}$

5.4.1 Variational ML and MAP

The incomplete log-likelihood function of a given parameter say ϕ for $\tilde{g}$ is

$$\ln L(\phi) \quad = \quad \ln \int p\,(\tilde{g}, f \,|\, \phi)\, \mathrm{d}f \tag{5.80}$$

By introducing the simpler dsitribution $Q\,(f)$ and maximizing $L\,(\phi)$ with respect to ϕ we have

$$\begin{aligned} \ln L(\phi) &= \ln \int Q(f)\, \frac{p\,(\tilde{g}, f \,|\, \phi)}{Q(f)}\, \mathrm{d}f \\ &\geq \int Q(f) \ln \frac{p\,(\tilde{g}, f \,|\, \phi)}{Q(f)}\, \mathrm{d}f \\ &= \langle \ln p\,(\tilde{g}, f \,|\, \phi) \rangle_{Q(f)} + \Upsilon_{Q(f)} \\ &= \mathcal{F}\,(Q(f), \phi) \end{aligned} \tag{5.81}$$

where $\Upsilon_{Q(f)}$ is the entropy of the distribution $Q(f)$ and we have made use of Jensen's inequality which makes use of the fact that the logarithmic function is concave.

Exact Optimization of Variational ML using the EM principle

The *E-step* involves optimizing the posterior $Q(f)$;

$$\mathcal{F}(Q(f), \phi) \quad = \quad \int Q(f) \ln \frac{p\,(\tilde{g}, f \,|\, \phi)}{Q(f)}\, \mathrm{d}f \tag{5.82}$$

If $Q(f) = p\,(f \,|\, \tilde{g}, \phi)$ in equation (5.82) we arrive at the following:

$$\begin{aligned} \mathcal{F}(Q(f), \phi) &= \int p\,(f \,|\, \tilde{g}, \phi) \ln \frac{p\,(f \,|\, \tilde{g}, \phi)\, p\,(\tilde{g} \,|\, \phi)}{p\,(f \,|\, \tilde{g}, \phi)}\, \mathrm{d}f \\ &= p\,(\tilde{g} \,|\, \phi) \int p\,(f \,|\, \tilde{g}, \phi) \mathrm{d}f \\ &= p\,(\tilde{g} \,|\, \phi) \end{aligned} \tag{5.83}$$

subject to a normalizing equality constraint

$$\int Q\,(f)\ \mathrm{d}f \quad = \quad 1 \tag{5.84}$$

Thus, the functional $\mathcal{F}$ of equation (5.83) becomes independent of f whenever $Q\,(f)$ equals $p\,(f \,|\, \tilde{g}, \phi)$. This signifies some kind of tight bounds at the E-step. By introducing a Lagrange multiplier α, $(\alpha > 0)$, the new functional say $\mathcal{F}_{\alpha,\phi}$ becomes

$$\begin{aligned} \mathcal{F}_{\alpha,\phi}(Q(f), \phi) &= \mathcal{F}(Q(f), \phi) + \alpha\Big[-1 + \int Q(f)\, \mathrm{d}f \Big] \\ &= \int p\,(f \,|\, \tilde{g}, \phi) \ln \frac{p\,(f \,|\, \tilde{g}, \phi)\, p\,(\tilde{g} \,|\, \phi)}{p\,(f \,|\, \tilde{g}, \phi)}\, \mathrm{d}f + \alpha\Big[-1 + \int Q(f)\, \mathrm{d}f \Big] \end{aligned} \tag{5.85}$$

By re-substituting $p\,(f \,|\, \tilde{g}, \phi)$ with its original simpler distribution $Q(f)$ into equation (5.85) and taking functional derivatives of $\mathcal{F}_{\alpha,\phi}$ with respect to $Q(f)$ gives

$$\frac{\partial}{\partial\, Q(f)} \mathcal{F}_{\alpha,\phi}(Q(f), \phi) \quad = \quad \ln p(g, f \,|\phi) \ - \ 1 \ - \ \ln Q\,(f) \ + \ \alpha \tag{5.86}$$

Setting the functional derivative to zero and solving for $Q(f)$ gives the updates equation for the posterior:

$$\begin{aligned} Q^{t+1}\,(f) &\longleftarrow \exp\,(\alpha - 1)\, p\,(f, \tilde{g} \,|\, \phi^t) \\ &= p\,(f \,|\, \tilde{g}, \phi^t) \end{aligned} \tag{5.87}$$

The parameter α can also be expressed in terms of the normalization constant to give

$$\alpha \ = \ 1 \ - \ \ln \int p\,(\tilde{g}, f \,|\, \phi^t)\ \mathrm{d}f \tag{5.88}$$

The *M-step* involves the optimization of $\mathcal{F}$ of equation (5.80) with respect to ϕ;

$$\begin{aligned} \mathcal{F}(Q(f), \phi) &= \int \mathrm{d}f\, Q(f) \ln \frac{p\,(\tilde{g}, f \,|\, \phi)}{Q(f)} \\ &= \int \mathrm{d}f\, Q(f) \ln p\,(\tilde{g}, f \,|\, \phi) \ - \int \mathrm{d}f\, Q(f) \ln Q(f) \end{aligned} \tag{5.89}$$

The entropy $Q(f)$ is independent of ϕ. Hence optimizing $\mathcal{F}$ with respect to ϕ is restricted to the first integral on the right hand side of equation (5.89). This is because

the parameter ϕ^t associated with $Q(f)$ in equation (5.87) is/are the previous estimates obtained from the t^{th}-iterate of an M-step and this was used in computing for the current $(t+1)^{th}$-iterate of Q. Therefore the hyperparameter ϕ^t associated with Q in equation (5.89) is held fixed whilst optimizing $\mathcal{F}$ at the M-step.

By taking functional derivatives of $\mathcal{F}$ of equation (5.89) with respect to ϕ and solving for the zero of ϕ gives the updates

$$\phi^{t+1} \longleftarrow \arg_\phi \max \int \mathrm{d}f \, p\left(f \,|\, \tilde{g}, \phi^t\right) \ln p\left(\tilde{g}, f | \, \phi\right) \tag{5.90}$$

where the optimization is over the second ϕ.

Exact Optimization of Variational MAP using the EM principle

From similar lines, the M-step of Variational MAP have updates given by

$$\phi^{t+1} \longleftarrow \arg_\phi \max \left\{ \ln p(\phi) + \int \mathrm{d}f \, p\left(f \,|\, \tilde{g}, \phi^t\right) \ln p\left(\tilde{g}, f | \, \phi\right) \right\} \tag{5.91}$$

If in equation (5.91), the prior $p\left(\phi\right)$ is non-informative, then the expression for the updates is approximately given by equation (5.90). [7]

5.4.2 Summary of Update Equations for Variational ML and MAP EM-Algorthm

For ML we have the following updates
E-Step:

$$Q^{t+1}\left(f\right) \quad \longleftarrow \quad \exp\left(\alpha - 1\right) p\left(f, \tilde{g} \,|\, \phi^t\right) \tag{5.92}$$

M-Step:

$$\phi^{t+1} \longleftarrow \arg_\phi \max \int \mathrm{d}f \, p\left(f \,|\, \tilde{g}, \phi^t\right) \ln p\left(\tilde{g}, f | \, \phi\right) \tag{5.93}$$

For MAP, we have the following updates
E-Step:

$$Q^{t+1}\left(f\right) \quad \longleftarrow \quad \exp\left(\alpha - 1\right) p\left(f, \tilde{g} \,|\, \phi^t\right) \tag{5.94}$$

M-Step:

$$\phi^{t+1} \longleftarrow \arg_\phi \max \left\{ \ln p(\phi) + \int \mathrm{d}f \, p\left(f \,|\, \tilde{g}, \phi^t\right) \ln p\left(\tilde{g}, f | \, \phi\right) \right\} \tag{5.95}$$

[7] Recall, that the posterior for f in MAP and ML have the same equation but difference exist in the estimation of parameters. The difference is due to the introduction of a prior distribution over the parameters for the MAP. So that if the prior is non-informative the two becomes the same.

What we failed to do and why?

If anything at all, we have been able to extend Variational Learning methods to ML and MAP. However, we did not apply the method to the Gravity problem since our main objective here is to use MAP and ML as platforms for understanding the underlying theory and concepts within the Variational Methods Framework. Despite the fact that we have not implemented the algorithm to obtain α and β, we have been able to at least capitalize on the free form optimization and EM-like algorithmic technique associated with the objective function $\mathcal{F}$. Therefore, we end ML and MAP here and focus on Variational Learning Algorithms for Bayesian Methods.

5.4.3 Variational Learning for Bayesian Methods

For convenience, we let the Variational Bayesian model be

$$p(\tilde{g}, f, \alpha, \beta) = p(\tilde{g}, f, \theta) \tag{5.96}$$

where the hyperparameter (or parameters) α and β still maintains $\alpha = \lambda_{ml}^2$, $\beta = 1/\sigma^2$ and $\theta = \{\alpha, \beta\}$. Recall; the precision hyperparameter β defines a noise variance $\sigma^2 = 1/\beta$ and the precision hyperparameter α is the regularization constant. In following the footsteps of the Bayesian Inference paradigm, we begin with the two levels of inference;
(i) Model fitting where we infer f by obtaining a compact representation of $p(f \,|\, \tilde{g}, \theta)$ for a given value of θ;

$$p(f \,|\, \tilde{g}, \theta) = \frac{p(\tilde{g} \,|\, f, \theta)\, p(f \,|\, \theta)}{p(\tilde{g} \,|\, \theta)} \tag{5.97}$$

(ii) infer θ by maximizing the evidence $p(\tilde{g} \,|\, \theta)$ of equation (5.97) in (i);

$$p(\theta \,|\, \tilde{g}) = \frac{p(\tilde{g} \,|\, \theta)\, p(\theta)}{p(\tilde{g})} \tag{5.98}$$

For α and β;
We assume here that we have no knowledge about α and β so we wish to construct an appropriate prior that embodies our ignorance. This is where the concept of *conjugate priors* is really needed. [8] We shall therefore not assign random values to α and β like we did previously by generating values using logspace. Rather we assume in addition to the likelihood of equation (5.98) a Gamma prior distribution over α and a Gamma prior distribution over β. To be realistic, we cannot place a Gaussian distribution over α and β since they are both non-negative. The Gamma distribution for α and β are respectively defined by

$$\begin{aligned} p(\alpha \,|\, a_\alpha, b_\alpha) &= \Gamma(\alpha; a_\alpha, b_\alpha) \\ &= \frac{b_\alpha^{a_\alpha}}{\Gamma(a_\alpha)}\, \alpha^{a_\alpha - 1} \exp -(b_\alpha\, \alpha) \quad ; \quad 0 \le \alpha < \infty \end{aligned} \tag{5.99}$$

and

$$\begin{aligned} p(\beta \,|\, c_\beta, d_\beta) &= \Gamma(\beta; c_\beta, d_\beta) \\ &= \frac{d_\beta^{c_\beta}}{\Gamma(c_\beta)}\, \beta^{c_\beta - 1} \exp -(d_\beta\, \beta) \quad ; \quad 0 \le \beta < \infty \end{aligned} \tag{5.100}$$

[8] Conjugate priors are priors whose functional forms belongs to the same family of distribution as the likelihood.

where $\Gamma(a_\alpha)/d_\beta^{c_\beta}$ and $\Gamma(c_\beta)/d_\beta^{c_\beta}$ are normalizing factors defined by $\int_0^\infty \alpha^{a_\alpha - 1} \exp -(b_\alpha\,\alpha)\,\mathrm{d}\alpha$ and $\int_0^\infty \beta^{c_\beta - 1} \exp -(d_\beta\,\beta)\,\mathrm{d}\beta$ respectively and the constants a_α, b_α, c_β and d_β are called hyper-hyperparameters. Their respective mean and variance are a_α/b_α and a_α/b_α^2 for α and c_β/d_β and c_β/d_β^2 for β.

5.4.4 Bounds for the Marginal Likelihood

We lower bound the log-marginal likelihood of $\tilde{g}$ for model H by introducing a distribution Q over both f and θ. Thus,

$$\begin{aligned} \ln p\,(\tilde{g}) &= \ln \int p\,(\tilde{g}, f, \theta\,)\,\mathrm{d}\theta\;\mathrm{d}f \\ &\geq \int Q(f,\theta) \ln \frac{p\,(\tilde{g}, f, \theta)}{Q(f,\theta)}\,\mathrm{d}\theta\,\mathrm{d}f \\ &= \langle \ln p\,(\tilde{g}, f, \theta\,) \rangle_{Q(f,\theta)} + \Upsilon_{Q(f,\theta)} \\ &= \mathcal{F}[\,Q(\Theta)\,] \end{aligned} \tag{5.101}$$

where $\Theta = \{f, \alpha, \beta\}$ and $\Upsilon_{Q(f,\theta)}$ is the entropy for Q and the inequality was possible through the usual appeal to Jensen's inequality.

Exact Optimization of Variational Bayesian using the EM principle

The learning rules also follows the Bayesian paradigm by integrating out nuisance parameters/variables.

We derive the E-step and M-step for any arbitrary distribution Q. First, we let

$$Q(\Theta) \,=\, Q(f,\theta) \,=\, Q(f,\alpha,\beta)$$

From inequality (5.101), we have

$$\mathcal{F}[\,Q(\Theta)\,] \;=\; \int Q(f,\alpha,\beta) \ln \frac{p\,(\tilde{g}, f, \alpha, \beta)}{Q(f,\alpha,\beta)}\;\mathrm{d}f\,\mathrm{d}\alpha\,\mathrm{d}\beta \tag{5.102}$$

The distributions of α and β are assumed to be independent, so their joint distribution $Q(\alpha,\beta)$ assumes the separable form

$$Q(\alpha,\beta) \,=\, Q(\theta) \,=\, Q_\alpha(\alpha)\; Q_\beta(\beta) \tag{5.103}$$

However, there is a problem with the separable form of the joint distribution of $Q(f,\alpha,\beta)$ since there exists some stochastic dependencies between f and θ (i.e between f and α to be precise). That is,

$$Q(f,\theta) \,=\, Q(f,\alpha,\beta) \,=\, Q(f\,|\,\alpha)\,Q(\alpha)\,Q(\beta) \tag{5.104}$$

Maximizing the lower bound of inequality (5.101) with respect to $Q(\Theta)$ to attain equality demands that the free distribution:

$$Q(f,\theta) \,=\, p(f,\theta\,|\,\tilde{g})$$

To achieve tight bounds requires that we know the normalizing constant, the marginal likelihood if an exact posterior is to be evaluated. We can go around this problem provided the distribution of $Q(f, \alpha, \beta)$ is separable. At this point, the best we can do is to assume and accept that we can (no matter what ever circumstance pertains to the stochastic dependency between f and α), constrain the posterior to a simple factorized form for the distribution Q to an approximation

$$Q(f, \alpha, \beta) \approx Q(f)\, Q(\alpha)\, Q(\beta) \tag{5.105}$$

Inserting approximation (5.105) into equation (5.102) gives

$$\begin{aligned} \mathcal{F}[\, Q(\Theta)\,] &= \int Q(f)\, Q(\alpha)\, Q(\beta) \ln \frac{p\,(\tilde{g}, f, \alpha, \beta \,|\, H)}{Q(f)\, Q(\alpha)\, Q(\beta)}\, \mathrm{d}f\, \mathrm{d}\alpha\, \mathrm{d}\beta \\ &= \langle \ln p\,(\tilde{g}, \Theta)\,\rangle_{Q(\Theta)} + \Upsilon_{Q(\Theta)} \end{aligned} \tag{5.106}$$

Using calculus of variation (shown in appendix A), the solution for maximizing the functional $\mathcal{F}[\, Q(\Theta)\,]$ with respect to each of the individual Q distribution is of the form

$$Q_i\,(\,\Theta\,) = \frac{\exp \langle \ln p\,(\tilde{g}, \Theta)\rangle_{Q_{k\neq i}}}{\int \exp \langle \ln p\,(\tilde{g}, \Theta)\,\rangle_{Q_{k\neq i}} \mathrm{d}\Theta_i} \tag{5.107}$$

or

$$\ln Q_i\,(\,\Theta\,) = \exp \langle p\,(\tilde{g}, \Theta)\rangle_{Q_{k\neq i}} + \text{constant} \tag{5.108}$$

where $\langle \bullet \rangle_{k\neq i}$ denotes the expectation with respect to every distribution other than $Q_i(\Theta_i)$.

Equation (5.108) embodies both the *E-step* and *M-step* of the Variational Bayesian EM-Algorithm.

5.4.5 Application of Variational Bayesian EM (VBEM) to the Gravity Problem

We now apply the above equations to the Gravity Problem. Figure (5.3) is a schematic diagram of the Graphical model for VBEM. From probability of everything,

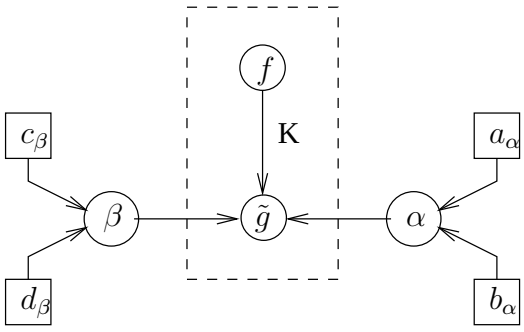

Figure 5.3: A graphical model of the Variational Bayesian approach

$$p\,(\tilde{g}, f, \alpha, \beta \,|\, H) = p\,(\tilde{g} \,|\, f, \beta, H)\, p\,(f \,|\, \alpha, H)\, p\,(\alpha \,|\, H)\, p(\beta \,|\, H) \tag{5.109}$$

The log-probabilities of each of the expressions on the right hand side of equation (5.109) are as follows:

$$\begin{aligned} \ln p\,(\alpha) &= \ln p(\alpha \,|\, a_\alpha, b_\alpha) \\ &= (a_\alpha - 1)\ln\alpha \,-\, b_\alpha\,\alpha \,+\, \wp' \end{aligned} \tag{5.110}$$

$$\begin{aligned} \ln p\,(\beta) &= \ln p(\beta \,|\, c_\beta, d_\beta) \\ &= (c_\beta - 1)\ln\beta \,-\, d_\beta\,\beta \,+\, \wp'' \end{aligned} \tag{5.111}$$

where the $\wp'$ and $\wp''$ are constants given by the log of their normalization factors $\Gamma(a_\alpha)/b_\alpha^{a_\alpha}$ and $\Gamma(c_\beta)/d_\beta^{c_\beta}$ respectively.

$$\ln p\,(f \,|\, \alpha) \;=\; -\frac{\alpha}{2}\,f^T f \,+\, \frac{n}{2}\ln\alpha \,+\, \wp''' \tag{5.112}$$

$$\ln p\,(\tilde{g} \,|\, f, \beta) \;=\; \frac{n}{2}\ln\beta \,-\, \frac{\beta}{2}(\tilde{g} - Kf)^T(\tilde{g} - Kf) \,+\, \wp'''' \tag{5.113}$$

where $\wp''' = \wp'''' = -\frac{n}{2}\ln(2\pi)$.

Substituting equations (5.110), (5.111), (5.112) and (5.113) into the logarithm of $p\,(\tilde{g}, f, \alpha, \beta \,|\, H)$ in equation (5.109) gives us

$$\begin{aligned} \ln p\,(\tilde{g}, f, \alpha, \beta \,|\, H) \;=\; & \frac{n}{2}\ln\beta \,-\, \frac{\beta}{2}(\tilde{g} - Kf)^T(\tilde{g} - Kf) \,+ \\ & -\,\frac{\alpha}{2}\,f^T f \,+\, \frac{n}{2}\ln\alpha \\ & +\,(a_\alpha - 1)\ln\alpha \,-\, b_\alpha\,\alpha \\ & +\,(c_\beta - 1)\ln\beta \,-\, d_\beta\,\beta \,+\, \mathrm{C} \end{aligned} \tag{5.114}$$

where $C = \wp' + \wp'' + \wp''' + \wp''''$ is the overall constant and we have made the hyper-hyperparameters a_α, b_α, c_β and d_β explicit with respect to α and β.

Optimization of $Q_\alpha(\alpha)$

As a distribution of $Q_\alpha(\alpha)$, we take expectations of equation (5.114) with respect to the distribution of $Q_f(f)\,Q_\beta(\beta)$ with all other terms indepedent of α put together and considered as a normalizing constant (of the distribution of α).

Thus, using equations (5.108) and (5.114), the logarithm of $Q_\alpha(\alpha)$ gives

$$\begin{aligned} \ln Q_\alpha(\alpha) &= \langle \ln p\,(\tilde{g}, \Theta) \rangle_{Q_f(f)\,Q_\beta(\beta)} \\ &= \langle \ln p\,(\tilde{g}, f, \alpha, \beta) \rangle_{Q_f(f)\,Q_\beta(\beta)} \\ &= -\frac{\alpha}{2}\langle\, f^T f \,\rangle \,+\, \frac{n}{2}\ln\alpha \,+\, (a_\alpha - 1)\ln\alpha \,-\, b_\alpha\,\alpha \,+\, C(f, \beta) \\ &= \left\{ \left(\frac{n}{2} + a_\alpha\right) - 1 \right\} \ln\alpha \,-\, \left\{ \frac{1}{2}\langle\, f^T f \,\rangle \,+\, b_\alpha \right\} \alpha \,+\, C(f, \beta) \end{aligned} \tag{5.115}$$

where $C(f, \beta)$ is a constant expression given by all terms on the right hand side of equation (5.114) not containing α.

Comparing coefficients of $\ln \alpha$ and α of equations (5.115) and (5.110), we can easily infer that the optimal for $Q_\alpha(\alpha)$ denoted $Q_\alpha^{opt}(\alpha)$ satisfies the Gamma distribution

$$Q_\alpha^{opt}(\alpha) \;=\; \Gamma(\alpha; \hat{a}, \hat{b}) \tag{5.116}$$

with update equations

$$\begin{aligned} \hat{a} &= \frac{n}{2} + a_\alpha \\ \\ \hat{b} &= \frac{1}{2} \langle\, f^T f \,\rangle + b_\alpha \end{aligned} \tag{5.117}$$

The mean and variance of $Q_\alpha^{opt}(\alpha)$ are given by $\hat{a}/\hat{b}$ and $\hat{a}/\hat{b^2}$ respectively. We are still left with an expression for $\langle\, f^T f \,\rangle$ of equation (5.117) which can be obtained from $Q_f(f)$.

Optimization of $Q_f(f)$

In the optimization of $Q_f(f)$, we take expectations of equation (5.114) with respect to the distribution of $Q_\alpha(\alpha)\, Q_\beta(\beta)$:

$$\begin{aligned} \ln Q_f(f) &= \langle \ln p\,(\tilde{g}, \Theta) \rangle_{Q_\alpha(\alpha)\, Q_\beta(\beta)} \\ &= \langle \ln p\,(\tilde{g}, f, \alpha, \beta) \rangle_{Q_\alpha(\alpha)\, Q_\beta(\beta)} \\ &= -\frac{1}{2} \Big\{ \tilde{g}^T \langle\, \beta \,\rangle \tilde{g} \;-\; 2\, f^T K^T \langle\, \beta \,\rangle\, \tilde{g} \\ &\quad + f^T \Big(K^T \langle\, \beta \,\rangle\, K \;+\; \langle\, \alpha \,\rangle\, I \Big)\, f \Big\} + C(\alpha, \beta) \end{aligned} \tag{5.118}$$

where $C(\alpha, \beta)$ is a constant given by all other expressions independent of f. The optimizing distribution $Q_f^{opt}(f)$ is a Gaussian identical to the posterior distribution for particular values of $\alpha = \hat{\alpha} = \langle \alpha \rangle$ and $\beta = \hat{\beta} = \langle \beta \rangle$. Thus

$$Q_f^{opt}(f) \;=\; p\,(f \,|\, \tilde{g}, \hat{\alpha}, \hat{\beta}) \;\sim\; \mathbb{N}\Big(\hat{f}_{MP_{\hat{\alpha},\hat{\beta}}}\,,\, \hat{\Sigma}_f^{-1} \Big) \tag{5.119}$$

with update equations

$$\hat{\Sigma}_f^{-1} \;=\; \Big(K^T \hat{\beta}\, K \;+\; \hat{\alpha} I \Big)^{-1}$$

$$\begin{aligned} \hat{f}_{MP_{\hat{\alpha},\hat{\beta}}} = \langle\, f \,\rangle &= \hat{\Sigma}_f^{-1}\, K^T \hat{\beta}\, \tilde{g} \\ &= \Big(K^T \hat{\beta}\, K \;+\; \hat{\alpha} I \Big)^{-1} K^T \hat{\beta}\, \tilde{g} \end{aligned}$$

$$\begin{aligned} \langle\, f^T f \,\rangle &= \mathrm{Tr}\,(\hat{\Sigma}_f^{-1}) \;+\; \hat{f}^T_{MP_{\hat{\alpha},\hat{\beta}}}\, \hat{f}_{MP_{\hat{\alpha},\hat{\beta}}} \\ &= \mathrm{Tr}\,(\hat{\Sigma}_f^{-1}) \;+\; \| \hat{f}_{MP_{\hat{\alpha},\hat{\beta}}} \|_2^2 \end{aligned}$$

Optimization of $Q_\beta(\beta)$

By following the previous steps, we take expectations with respect to the distribution $Q_f(f)$ and $Q_\alpha(\alpha)$.

$$\begin{aligned}
\ln Q_\beta(\beta) &= \langle \ln p\,(\tilde{g}, \Theta) \rangle_{Q_\alpha(\alpha)\,Q_f(f)} \\
&= \langle \ln p\,(\tilde{g}, f, \alpha, \beta) \rangle_{Q_\alpha(\alpha)\,Q_f(f)} \\
&= \frac{n}{2} \ln \beta \,-\, \left\langle \frac{\beta}{2} (\tilde{g} - Kf)^T (\tilde{g} - Kf) \right\rangle_{Q_\alpha(\alpha)\,Q_f(f)} \\
&\quad + (c_\beta - 1) \ln \beta \,-\, d_\beta\, \beta \,+\, \mathrm{C}(f, \alpha) \\
&= \left\{ \left(\frac{n}{2} + c_\beta \right) \,-\, 1 \right\} \ln \beta \,-\, \left\{ d_\beta \,+\, \frac{1}{2} \left\langle\, (\tilde{g} - Kf)^T (\tilde{g} - Kf) \,\right\rangle \right\} \beta \\
&\quad + C(f, \alpha)
\end{aligned} \tag{5.120}$$

where $C(\alpha, f)$ is a constant given by all other expressions independent of β.

Comparing the coefficients of $(\ln \beta)$ and β in equations (5.120) and (5.111), we can easily infer that the optimal for $Q_\beta(\beta)$ denoted $Q_\beta^{opt}(\beta)$ satisfies the Gamma distribution

$$Q_\beta^{opt}(\beta) \,=\, \Gamma(\beta; \hat{c}, \hat{d}) \tag{5.121}$$

with update equations

$$\begin{aligned}
\hat{c} &= \frac{n}{2} + c_\beta \\
\\
\hat{d} &= d_\beta \,+\, \frac{1}{2} \left\langle\, (\tilde{g} - Kf)^T (\tilde{g} - Kf) \,\right\rangle
\end{aligned} \tag{5.122}$$

The mean and variance of $Q_\beta^{opt}(\beta)$ are given by $\hat{c}/\hat{d}$ and $\hat{c}/\hat{d}^2$ respectively. We are still left with an expression for $\left\langle\, (\tilde{g} - Kf)^T (\tilde{g} - Kf) \,\right\rangle$ of equation (5.122) and it can be obtained as follows:

$$\begin{aligned}
\left\langle\, (\tilde{g} - Kf)^T (\tilde{g} - Kf) \,\right\rangle &= \| \tilde{g} \|_2^2 \,-\, 2\, \tilde{g}^T K \langle f \rangle \,+\, \langle\, (Kf)^T (Kf) \,\rangle \\
&= \| \tilde{g} \|_2^2 \,-\, 2\, \tilde{g}^T K \hat{f}_{MP} \,+\, \left((K \hat{f}_{MP})^T (K \hat{f}_{MP}) \right) \,+\, \mathrm{Tr}\,(K \hat{\Sigma}_f^{-1} K^T) \\
&= \| \tilde{g} \,-\, K \hat{f}_{MP} \|_2^2 \,+\, \mathrm{Tr}\,(K \hat{\Sigma}_f^{-1} K^T)
\end{aligned} \tag{5.123}$$

[9] where $\hat{f}_{MP}$ and Σ_f^{-1} are from equation (5.120).

Finally, from equations (5.117), (5.119), (5.120) and (5.120), the mean $\hat{\alpha}$ for the optimized Gamma distribution becomes

$$\hat{\alpha} \;=\; \frac{n/2 \,+\, a_\alpha}{\frac{1}{2} \mathrm{Tr}\,(\hat{\Sigma}_f^{-1}) \,+\, \frac{1}{2} \| \hat{f}_{MP_{\hat{\alpha},\hat{\beta}}} \|_2^2 \,+\, b_\alpha} \tag{5.124}$$

[9] For a stochastic vector x with mean m, covariance M and central moments $\langle (x-m)^r \rangle$
$\langle (Ax)^T (Ax) \rangle \,=\, \mathrm{Tr}\,(A\,M\,A^T) \,+\, (Am)^T (Am)$

A special case of equation (5.124) is when the prior on α becomes non-informative (that is $a_\alpha \to 0$ and $b_\alpha \to 0$). We obtain

$$\hat{\alpha} = \frac{n}{\mathrm{Tr}\,(\hat{\Sigma}_f^{-1}) + \| \hat{f}_{MP_{\hat{\alpha},\hat{\beta}}} \|_2^2} \tag{5.125}$$

Also from equations (5.119),(5.121) and (5.122), we have

$$\hat{\beta} = \frac{n/2 + c_\beta}{\frac{1}{2}\| \tilde{g} - K\,\hat{f}_{MP} \|_2^2 + \frac{1}{2}\mathrm{Tr}\,(K\,\hat{\Sigma}_f^{-1}K^T) + d_\beta} \tag{5.126}$$

A special case of equation (5.126) is when the prior on β becomes non-informative (that is $c_\beta \to 0$ and $d_\beta \to 0$). We obtain

$$\hat{\beta} = \frac{n}{\| \tilde{g} - K\,\hat{f}_{MP} \|_2^2 + \mathrm{Tr}\,(K\,\hat{\Sigma}_f^{-1}K^T)} \tag{5.127}$$

The optimal $\hat{\alpha}$ and $\hat{\beta}$ of equations (5.125) and (5.127) are the same as the optimal obtained in the Evidence Framework for α and β. Hence any optimum of the evidence approximation also correspond to the optimum of Variational Bayes.

5.5 Comparison between L-Curve for Tikhonov and Bayesian Inference

The main analysis tool of the L-Curve is the Truncated SVD. The flat regions (regions almost parallel to the abscissa and ordinate axes) are residuals. The solution lies within the truncated region with the best estimate (optimal parameter) given by the value of λ_{rls}^2 at utmost corner. The flat regions gives no information since they are made up of residuals and therefore does not contribute to the Regularized solution.

Analysis in the Bayesian Inference Framework, is governed by an expression which is equivalent to the filter factors in the Numerical Framework if the relation $\lambda_{rls}^2 = \alpha/\beta = \lambda_{ml}^2\,\sigma^2$ is substituted into the expressions for Ψ and Ξ. It is straight forward to see this from equations (5.74) and (5.76). Thus;

$$\begin{aligned}\Psi &= \sum_{i=1}^{n} \frac{{d_i}^2}{{d_i}^2 + \alpha\,\sigma^2} \\ &= \sum_{i=1}^{n} \frac{{d_i}^2}{{d_i}^2 + \lambda_{rls}^2}\end{aligned}$$

and

$$\begin{aligned}\Xi = 2\,\beta\,E_{\tilde{g}}^{MP} &= n - \beta \sum_{i=1}^{n} \frac{{d_i}^2}{\beta\,{d_i}^2 + \alpha} \\ &= \sum_{i=1}^{n} \frac{\lambda_{ml}^2\sigma^2}{{d_i}^2 + \lambda_{ml}^2\sigma^2} \\ &= \sum_{i=1}^{n} \frac{\lambda_{rls}^2}{{d_i}^2 + \lambda_{rls}^2}\end{aligned}$$

The expression for Ψ_i is equal to x_i (i.e $\Psi_i = x_i$) of equation (5.3). The difference $\Xi_i = 1 - \Psi_i$ are the corresponding SVD components of the residuals which characterizes data misfit. Moreover, analysis in Emprical Bayes of equations (5.40), (5.41) and (5.42) stands the same as the analysis of the L-Curve of equations (5.12) and (5.13).

We attribute differences to come from the fact that, the Evidence Framework (Bayesian) has a well defined functional approximation for evaluating α and β whereas the L-Curve for Tikhonov's Regularization is heuristic.

Finally, a search in parameter space of $\{\lambda^2_{rls}\}$ for which the functional $\int_\Omega \epsilon^2(s)\, w(s)\, \mathrm{d}s$ as a function of λ^2_{rls} is optimal is comparable to the ML principle discussed if we are dealing with Gaussianity. Hence, $\int_\Omega \epsilon^2(s)\, w(s)\, \mathrm{d}s$ can be viewed as a likelihood function of λ^2_{rls} such that $\lambda^2_{rls} = \sigma^2\, \lambda^2_{ml}$. The most probable parameter in this case is also $\lambda^{2\,\star}_{rls}$.

CHAPTER 6

Simulation Results

6.1 The Ill-Posed Inverse Problem Competition using the Gravity Model Example

In this Chapter, we compare estimates obtained for the inverse problem using the methods described in the previous Chapters namely; standard ML (or LS), Bayesian Inference Method (BIM), Variational Bayes EM (VBEM), Regularized Maximum Likelihood Method(MLM) and the L-Curve Method (LCM). We denote the estimated parameters of BIM by $(\alpha_{bim}, \beta_{bim})$, VBEM by $(\alpha_{vbm}, \beta_{vbm})$, MLM by $(\alpha_{mlm}, \beta_{mlm})$ and LCM by $\alpha_{lcm} = \lambda^2_{rls}$. The kernel K and f are the same as used previously. We shall first consider the case with additive noise $\sigma^2 = 10^{-6}$ and $\sigma^2 = 10^{-3}$ using the matlab built-in function

$$\epsilon_{10^{-6}} = \sqrt{(10^{-6})} \times \text{randn}(n, 1) \quad \text{and} \quad \epsilon_{10^{-3}} = \sqrt{(10^{-6})} \times \text{randn}(n, 1)$$

respectively. The depth h is still maintained at 0.25.

Figures (6.1) through (6.6) has values of n set to 50 in their first columns and 100 in their second columns. The higher the value of n, the more the system's matrix K has a lot of its singular values to be very small. Figure (6.1) are subplots of the true function f in the first row and corrupted output $\tilde{g}$ at the second and third rows with $\sigma^2 = 10^{-6}$ and $\sigma^2 = 10^{-3}$ respectively. Figures (6.2) through (6.6) have $\tilde{g}$ corrupted at a noise level $\sigma^2 = 10^{-6}$ in their first rows and $\sigma^2 = 10^{-3}$ for the second rows.

With respect to Figures (6.3) through (6.6), the optimal values, $\alpha_{method}, \beta_{method}$ for each method is located on top of the Figure with the ratio $\alpha_{est}/\beta_{est} = \lambda^2_{est} \times \sigma^2_{est}$[1] . Our interest here is to find whether the ratios show some consistency for different values of n and σ^2. Each subplot of Figures (6.3) through (6.6) consists of the true input f and the estimated (or reconstructed) input.

In each of Figures (6.7) through (6.12), we have all the reconstructed (estimated) input f_{est} for $n = 50$, $n = 100$ and $\sigma^2 = 10^{-6}$, $\sigma^2 = 10^{-3}$. The value of n tally with the last value of i at the abscissa axis. The norm of the difference in f_{est} and true f (i.e $\| f - f_{est} \|_2$) for each method is also calculated for each n and σ^2. In the remaining Figures we plotted $\| f - f_{est} \|_2$ versus n in steps of 10 to 200 for noise levels $\sigma^2 = 10^{-6}$, $\sigma^2 = 10^{-5}$, $\sigma^2 = 10^{-4}$, $\sigma^2 = 10^{-3}$, $\sigma^2 = 10^{-2}$ and $\sigma^2 = 10^{-1}$.

[1] σ^2_{est} is different from noise level σ^2. Here, σ^2_{est} is the reciprocal of β_{est}.

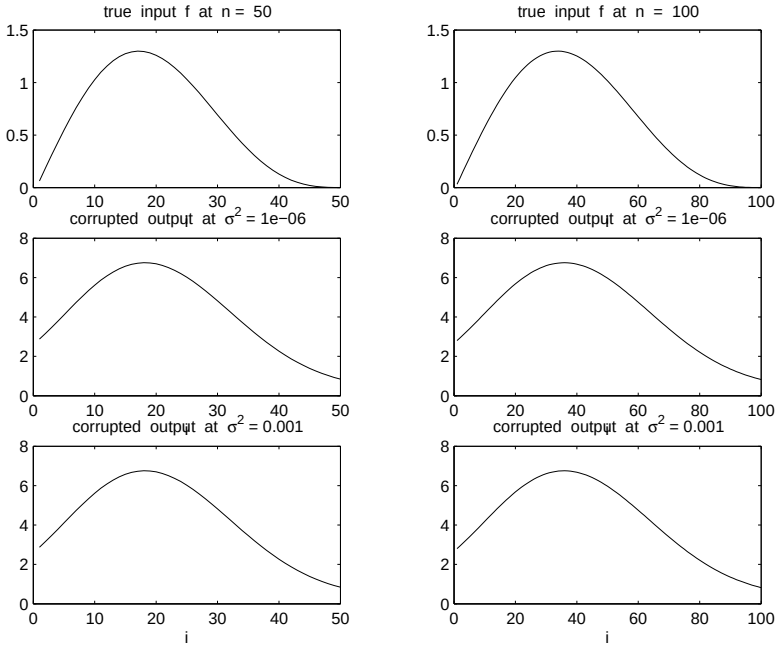

Figure 6.1: First Row : *The true input f when $n = 50$ (left) and $n = 100$ (right).*
Second Row : *Output $\tilde{g}$ with an additive noise of 10^{-6} for $n = 50$ (left) and $n = 100$ (right).*
Third Row : *Output $\tilde{g}$ with an additive noise of 10^{-3} for $n = 50$ (left) and $n = 100$ (right).*

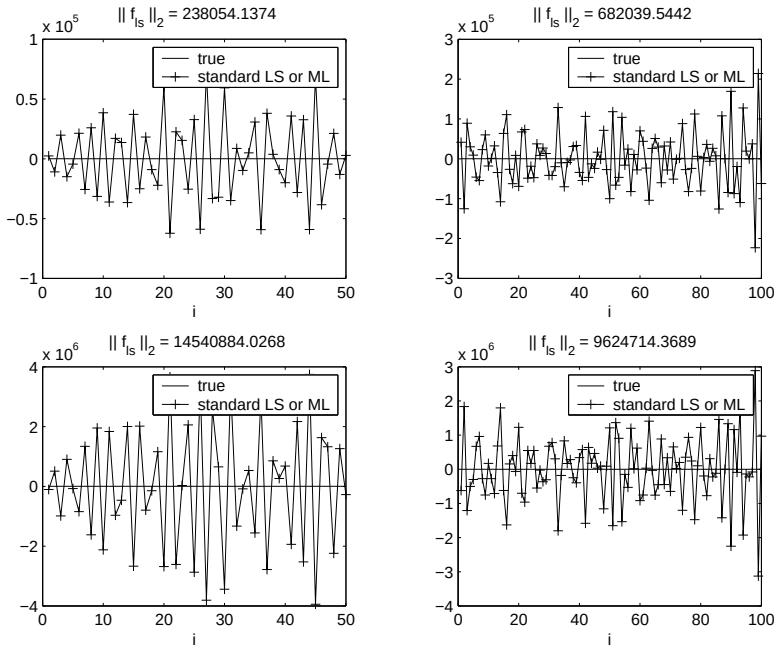

Figure 6.2: *Least Squares Estimates illustrating large norm $\| f_{ls} \|_2$. The true curve is seen to be flat due to large values on the ordinate axis. $\| f_{ls} \|_2$ is large even at low noise level.*

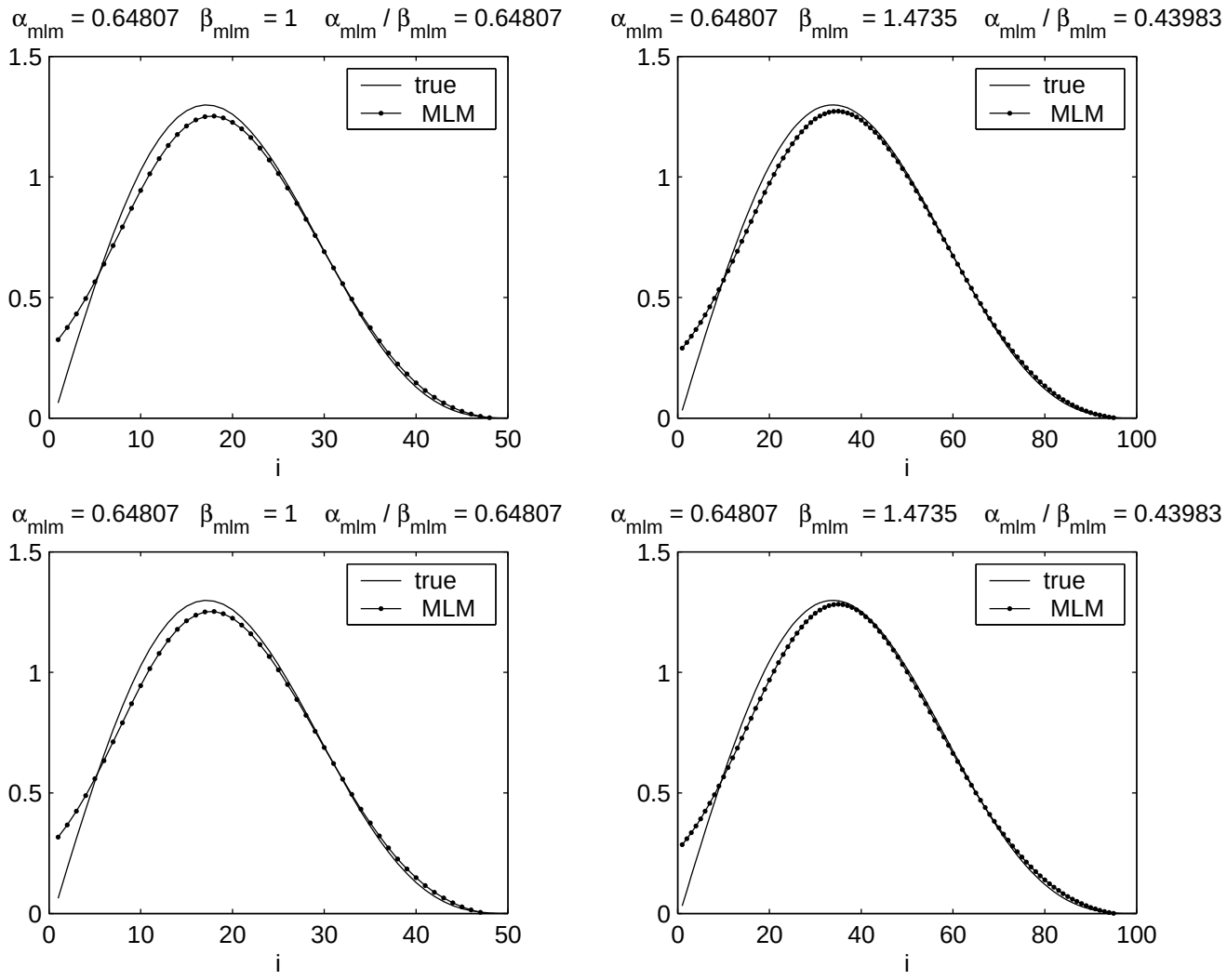

Figure 6.3: First Row : *Estimates of α_{mlm} and β_{mlm} and the ratio α_{mlm}/β_{mlm} using the Maximum Likelihood method at $\sigma^2 = 10^{-6}$ for $n = 50$ (left) and $n = 100$ (right).*
Second Row : *Estimates of α_{mlm} and β_{mlm} and the ratio α_{mlm}/β_{mlm} at $\sigma^2 = 10^{-3}$ for $n = 50$ (left) and $n = 100$ (right).*

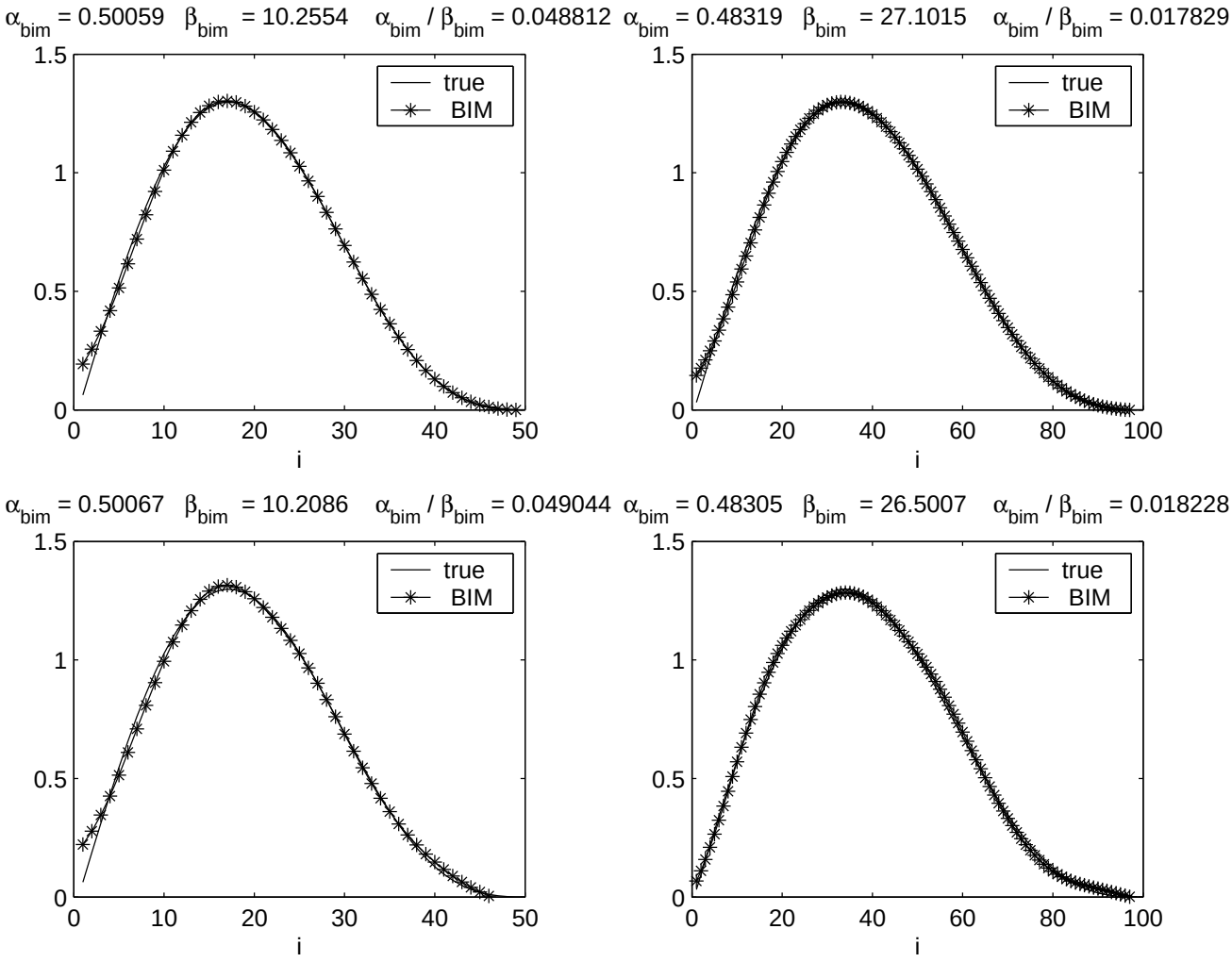

Figure 6.4: First Row : *Estimates of α_{bim} and β_{bim} and the ratio α_{bim}/β_{bim} using the Bayesian Inference method at $\sigma^2 = 10^{-6}$ for $n = 50$ (left) and $n = 100$ (right).*
Second Row : *Estimates of α_{bim} and β_{bim} and the ratio α_{bim}/β_{bim} at $\sigma^2 = 10^{-3}$ for $n = 50$ (left) and $n = 100$ (right).*

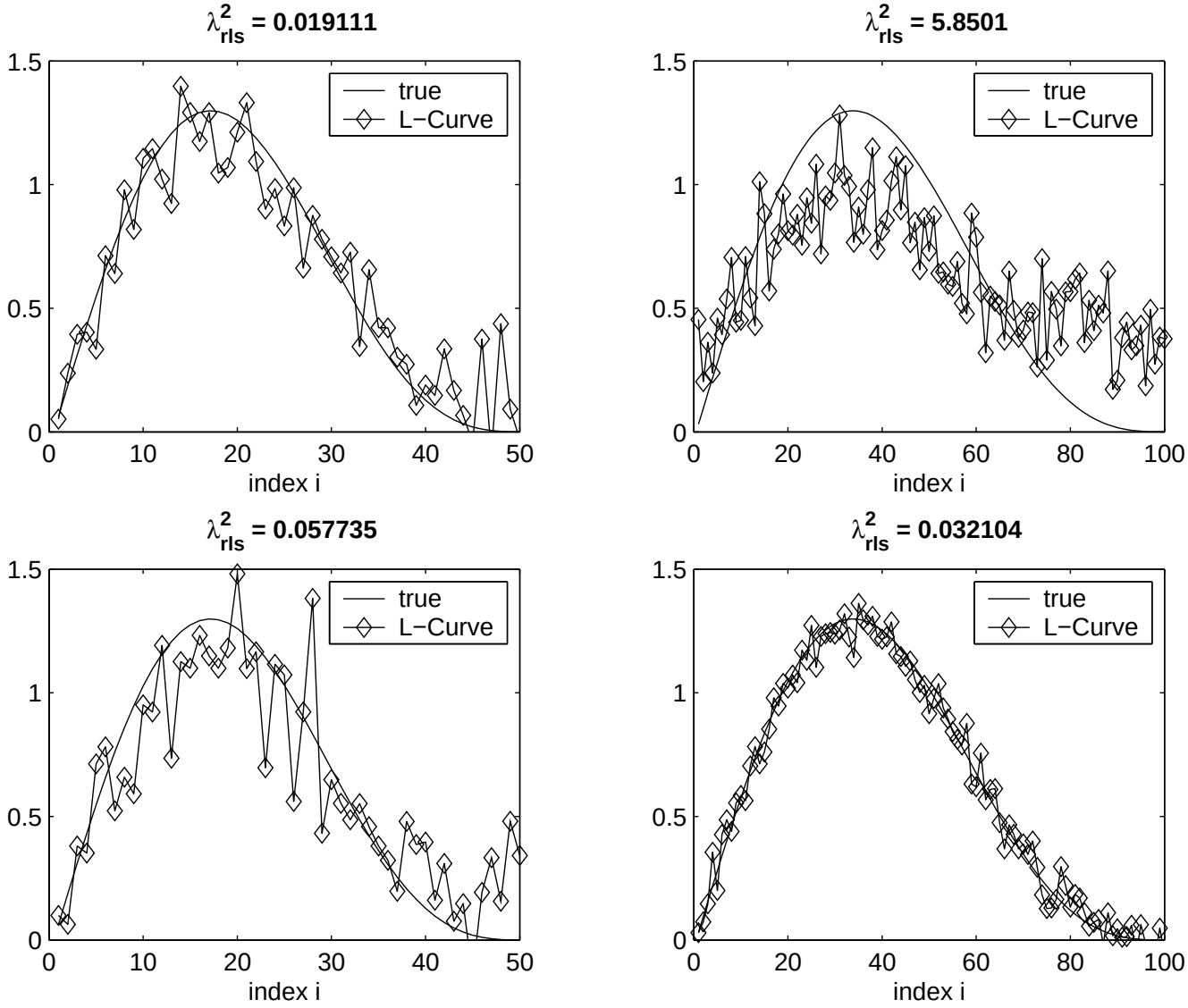

Figure 6.5: First Row : *Estimates of* $\alpha_{lcm} = \lambda^2_{rls}$ *using the L-Curve method for Tikhonov's Regularization at* $\sigma^2 = 10^{-6}$ *for* $n = 50$ *(left) and* $n = 100$ *(right).*
Second Row : *Estimates of* $\alpha_{lcm} = \lambda^2_{rls}$ *at* $\sigma^2 of = 10^{-3}$ *for* $n = 50$ *(left) and* $n = 100$ *(right).*

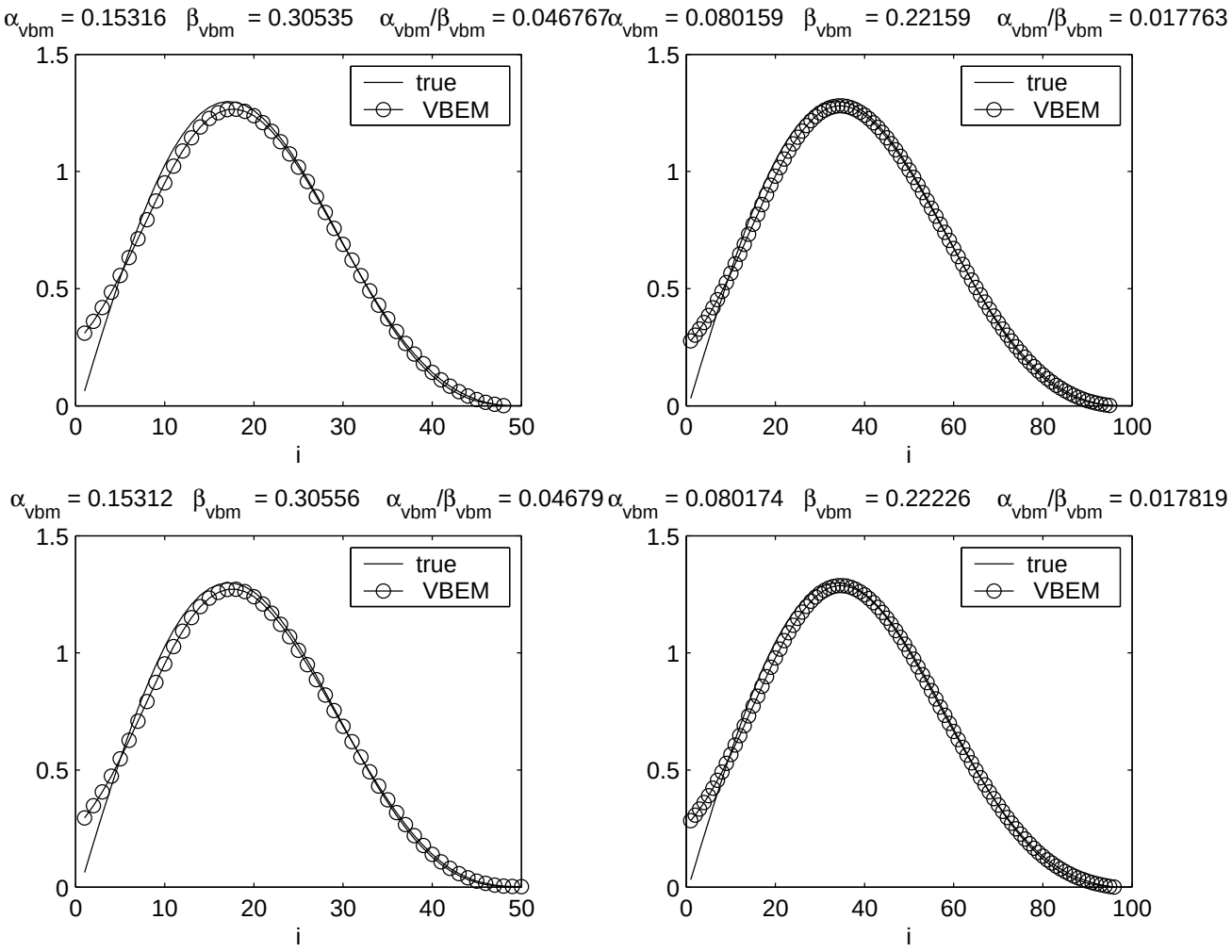

Figure 6.6: First Row : *Estimates of α_{vbm} and β_{vbm} and the ratio α_{vbm}/β_{vbm} using the Variational Bayesian EM algorithm at $\sigma^2 = 10^{-6}$ for $n = 50$ (left) and $n = 100$ (right) for 500 iterations.*

Second Row : *Estimates of α_{vbm} and β_{vbm} and the ratio α_{vbm}/β_{vbm} at $\sigma^2 = 10^{-3}$ for $n = 50$ (left) and $n = 100$ (right) at the same number of iterations.*

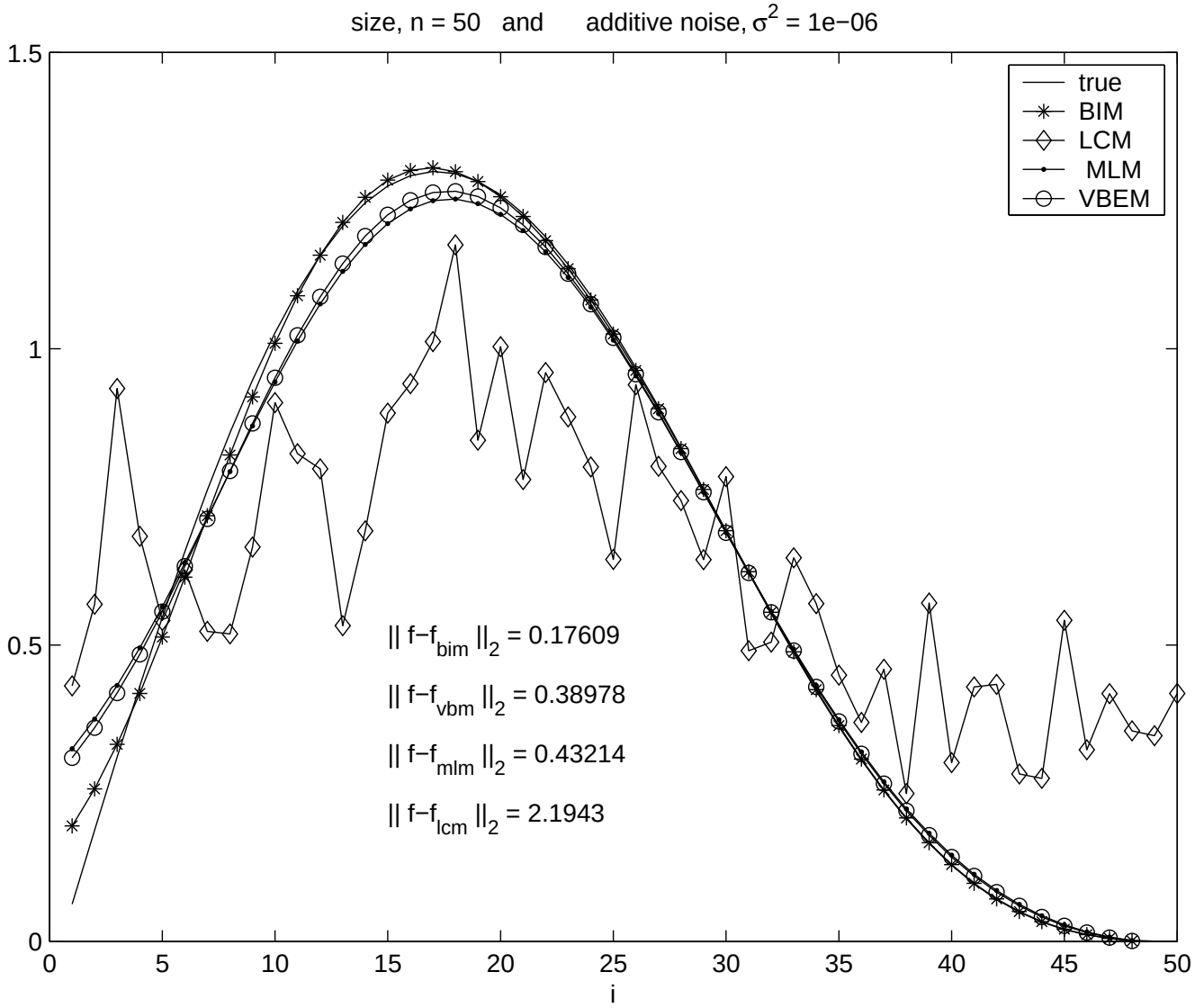

Figure 6.7: *The estimated f_{est} using the methods described and root of squared deviations of f_{est} from f at $\sigma^2 = 10^{-6}$ for $n = 50$. BIM gives the smallest $\| f - f_{est} \|_2$ followed by VBEM followed by MLM and LCM.*

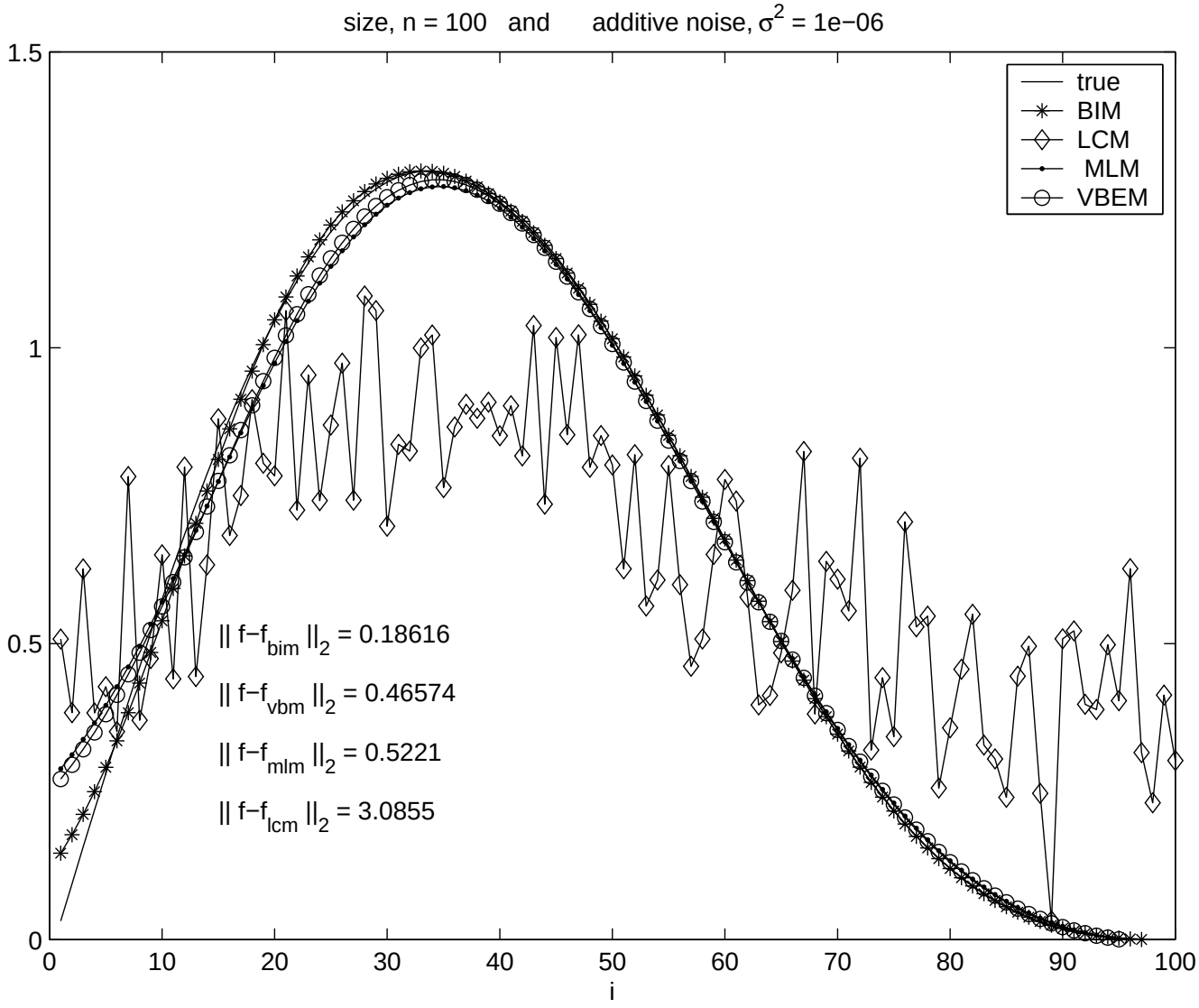

Figure 6.8: *The estimated f_{est} using the methods described and root of squared deviations of f_{est} from f at $\sigma^2 = 10^{-6}$ for $n = 100$. BIM gives the smallest $\| f - f_{est} \|_2$ followed by VBEM followed by MLM and LCM.*

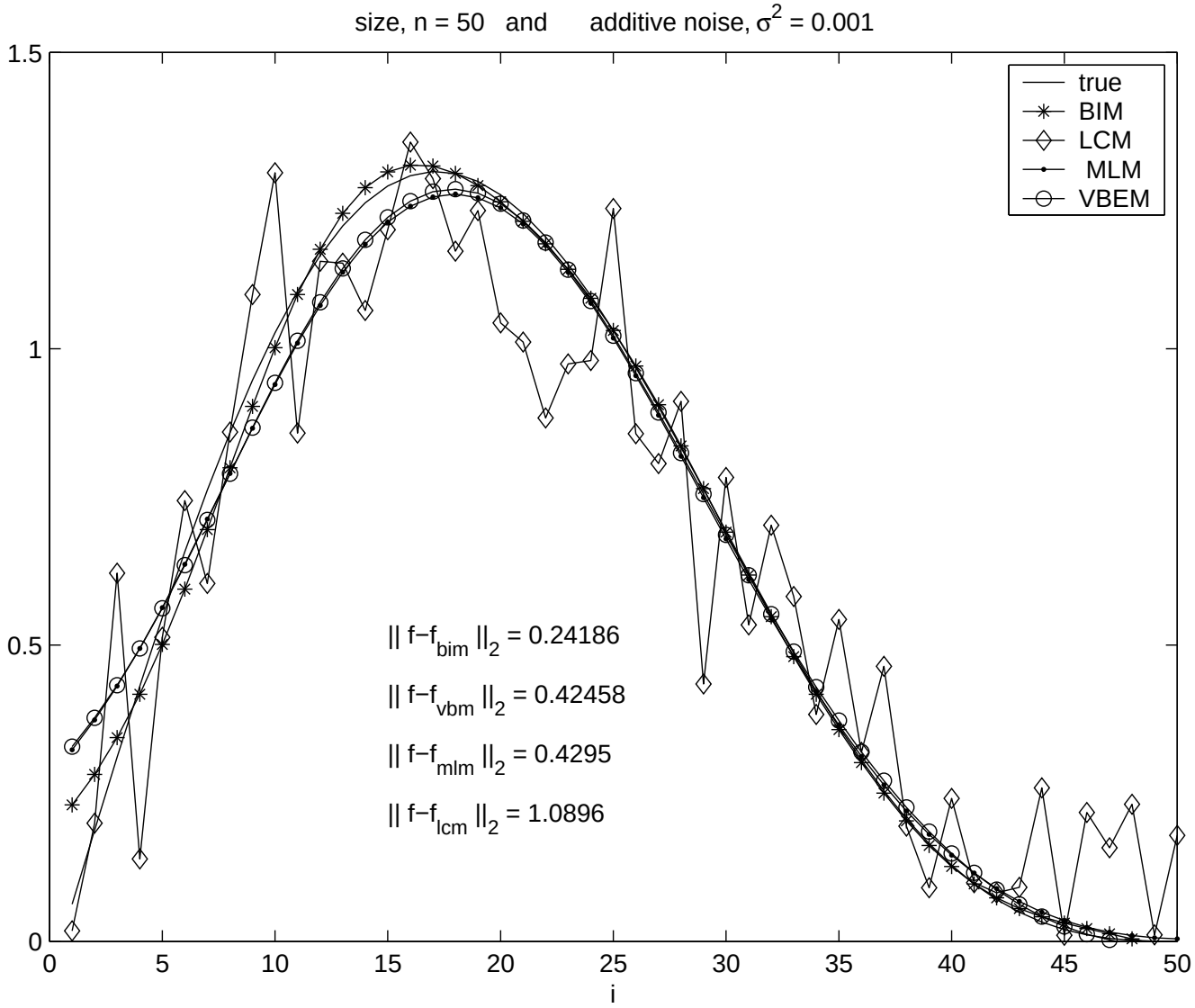

Figure 6.9: *The estimated f_{est} using the methods described and root of squared deviations of f_{est} from f at $\sigma^2 = 10^{-3}$ for $n = 50$. BIM gives the smallest $\| f - f_{est} \|_2$ followed by VBEM followed by MLM and LCM. But estimates from LCM has improved considerably.*

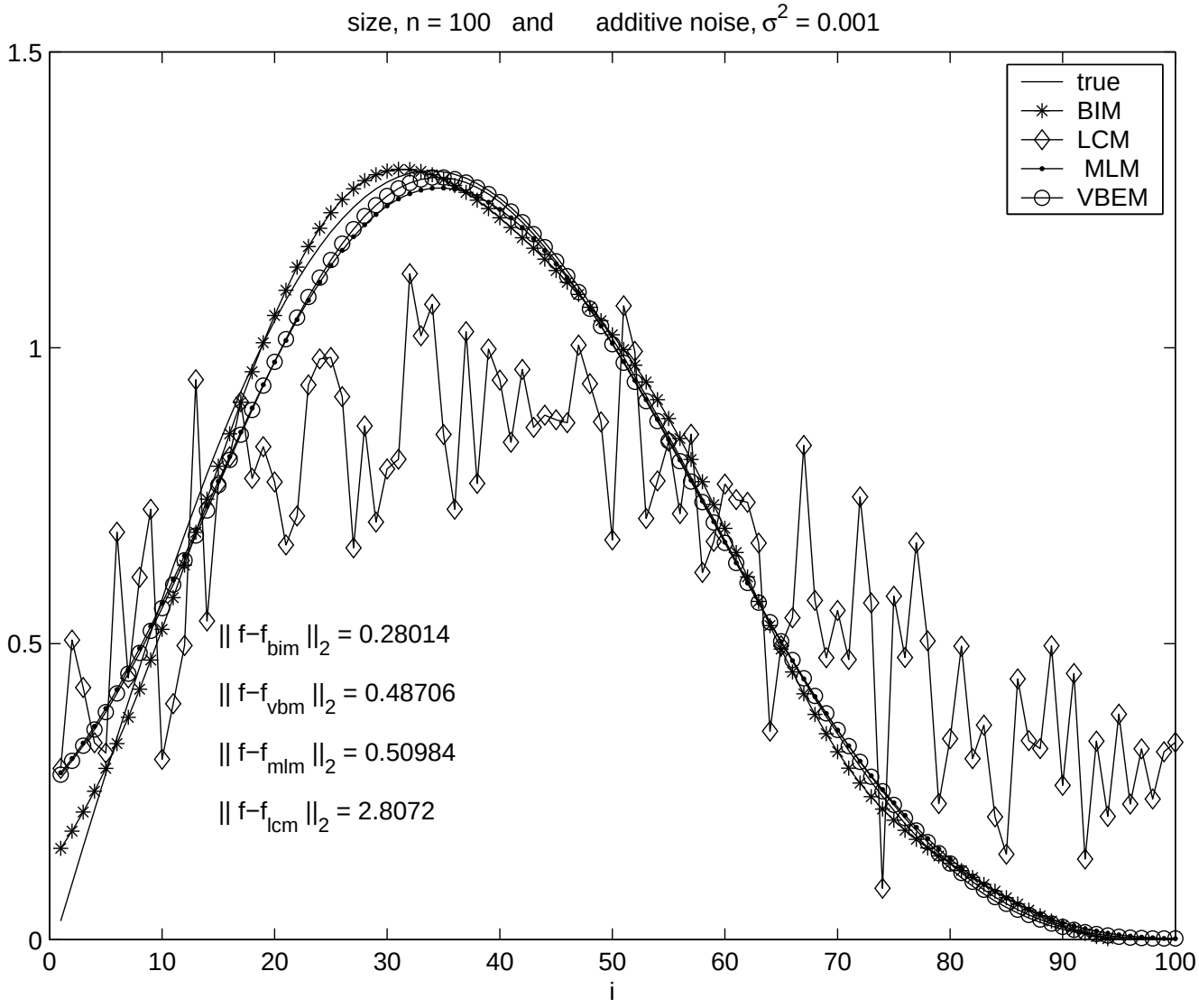

Figure 6.10: *The estimated f_{est} using the methods described and root of squared deviations of f_{est} from f at $\sigma^2 = 10^{-3}$ for $n = 100$. BIM gives the smallest $\| f - f_{est} \|_2$ followed by VBEM followed by MLM and LCM.*

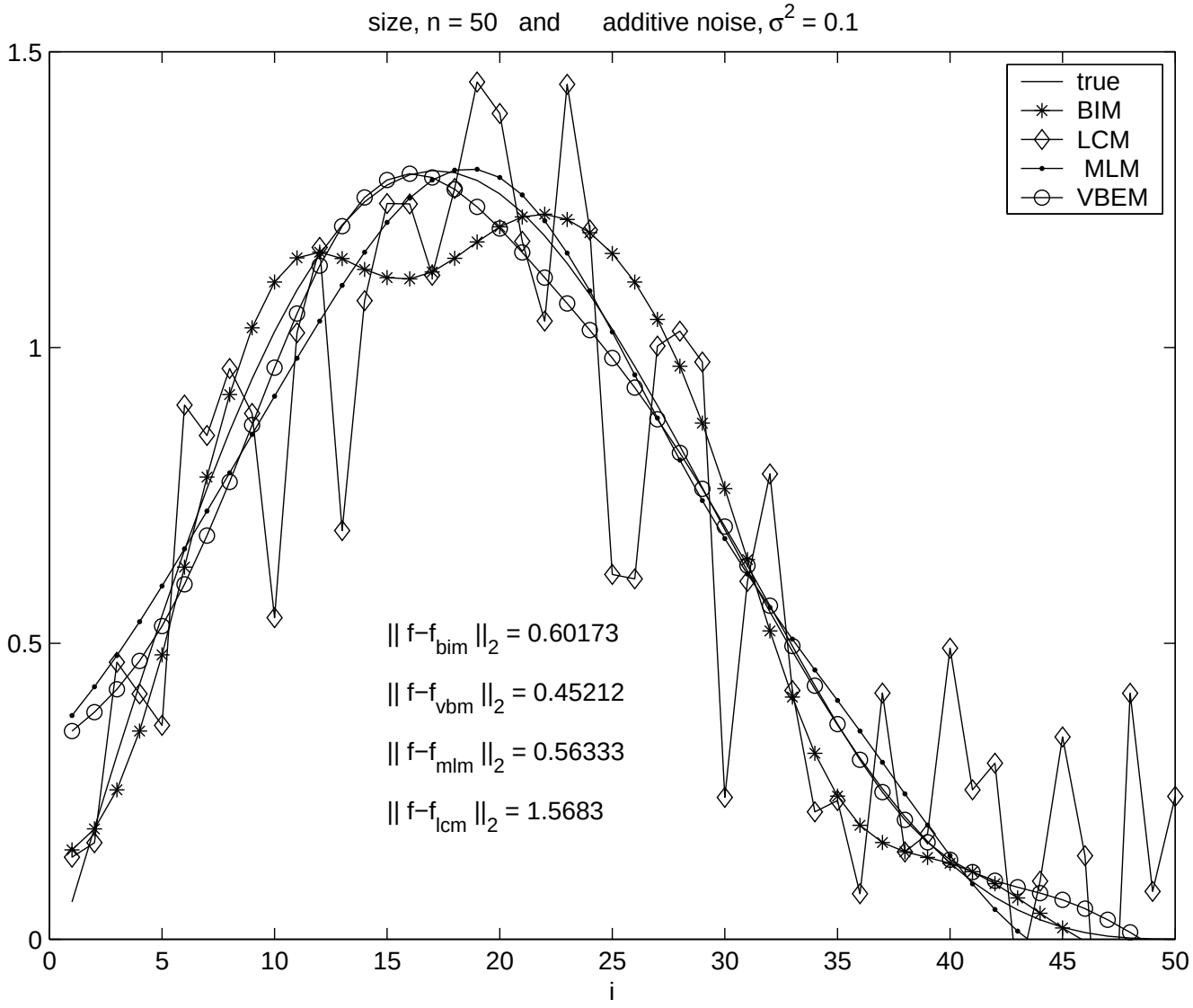

Figure 6.11: *VBEM gives the smallest* $\| f - f_{est} \|_2$ *followed by MLM followed by BIM and LCM. at* $\sigma^2 = 10^{-1}$ *for* $n = 50$.

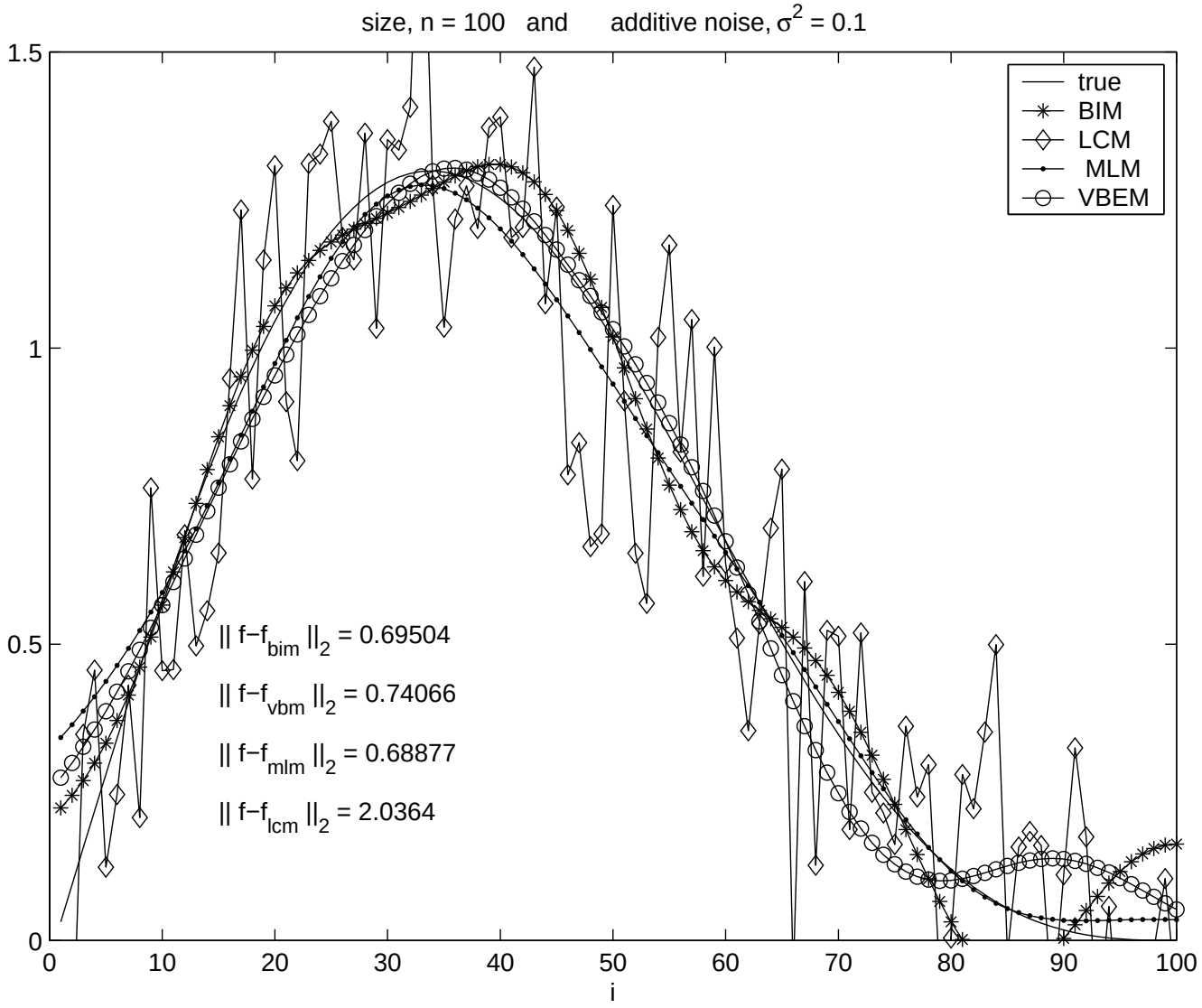

Figure 6.12: *MLM gives the smallest* $\| f - f_{est} \|_2$ *followed by BIM followed by VBEM and LCM at* $\sigma^2 = 10^{-1}$ *for* $n = 100$.

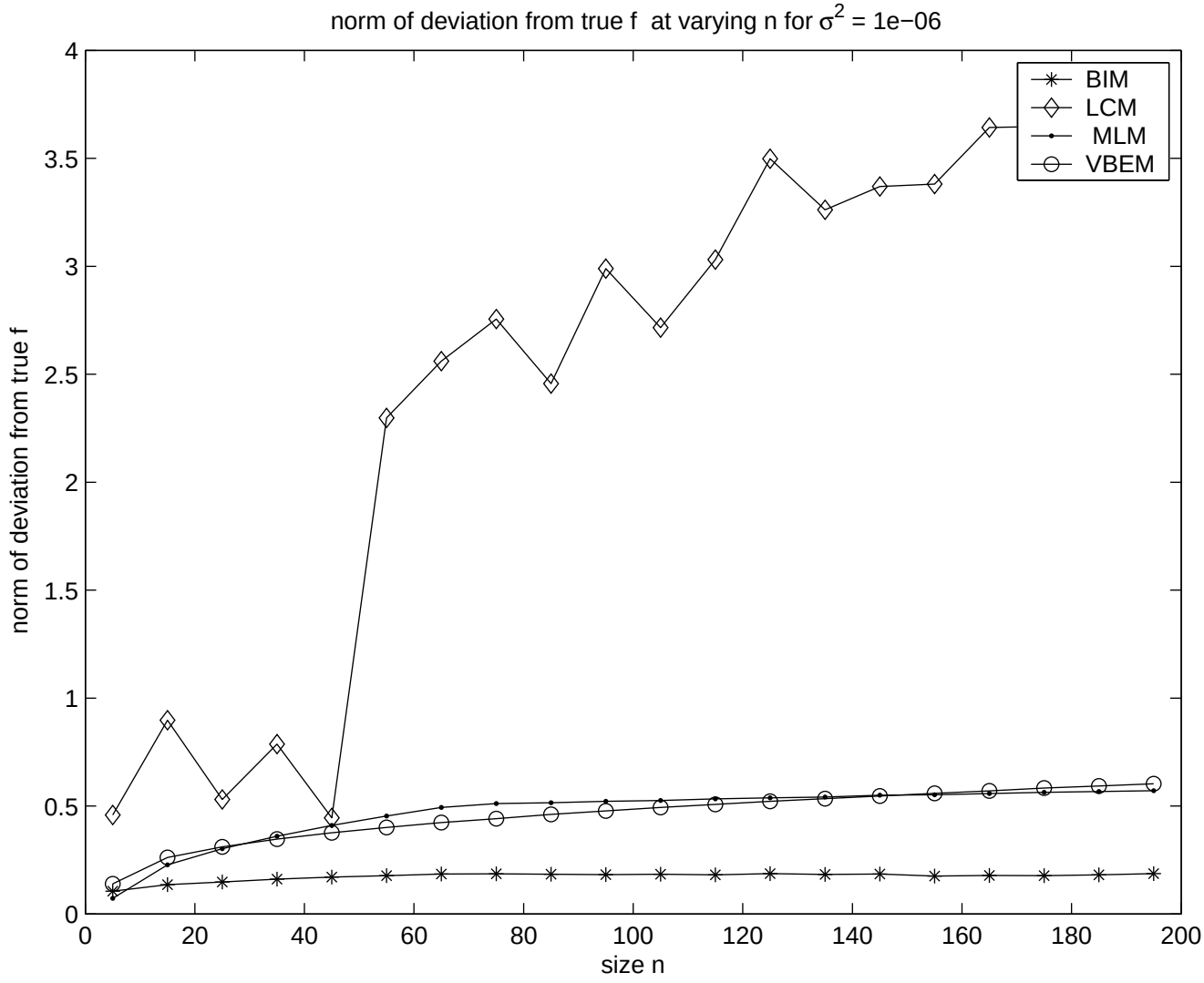

Figure 6.13: $\| f - f_{est} \|_2$ *at varying* n *at* $\sigma^2 = 10^{-6}$.

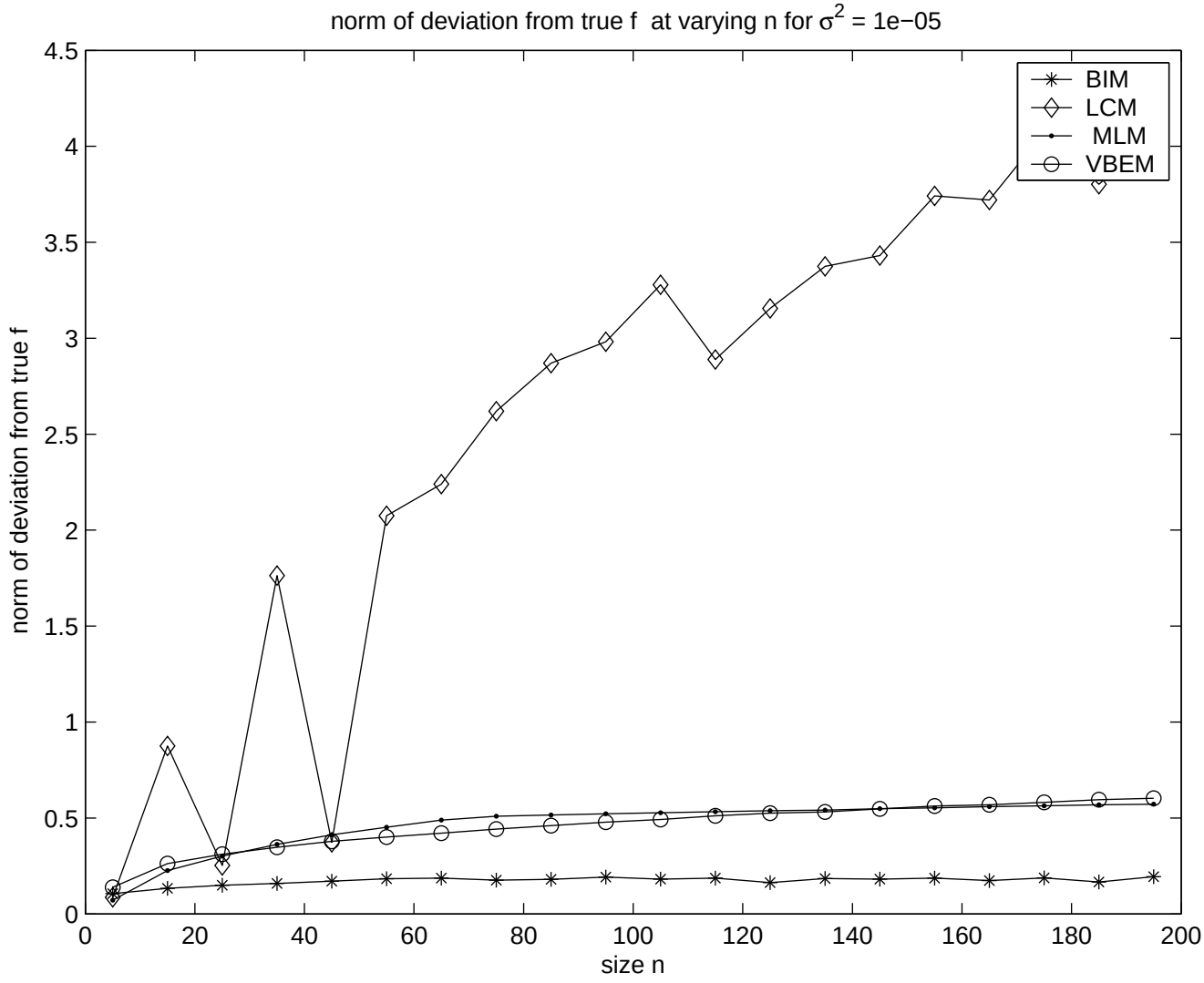

Figure 6.14: $\| f - f_{est} \|_2$ *at varying* n *at* $\sigma^2 = 10^{-5}$.

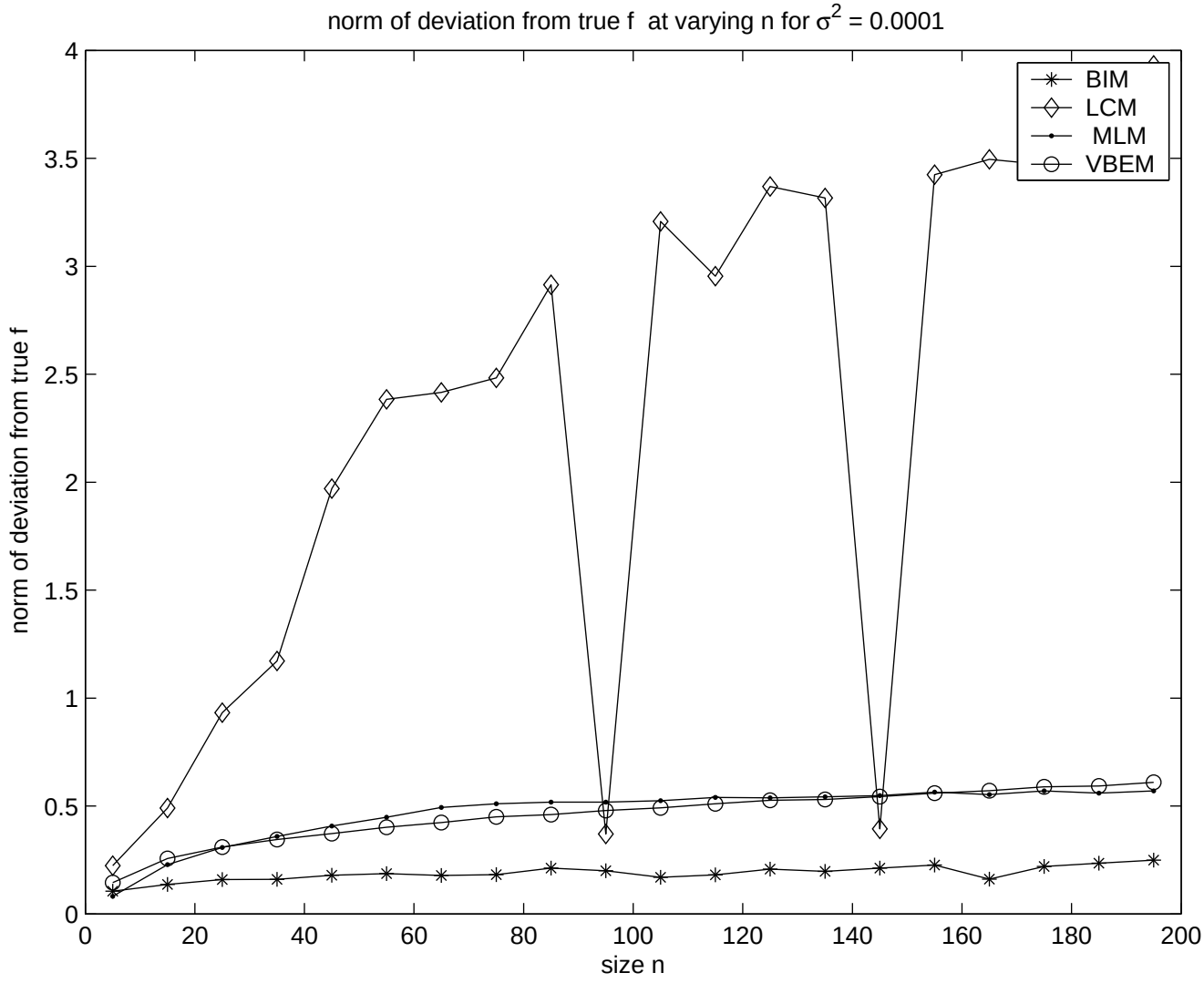

Figure 6.15: $\| f - f_{est} \|_2$ *at varying* n *at* $\sigma^2 = 10^{-4}$.

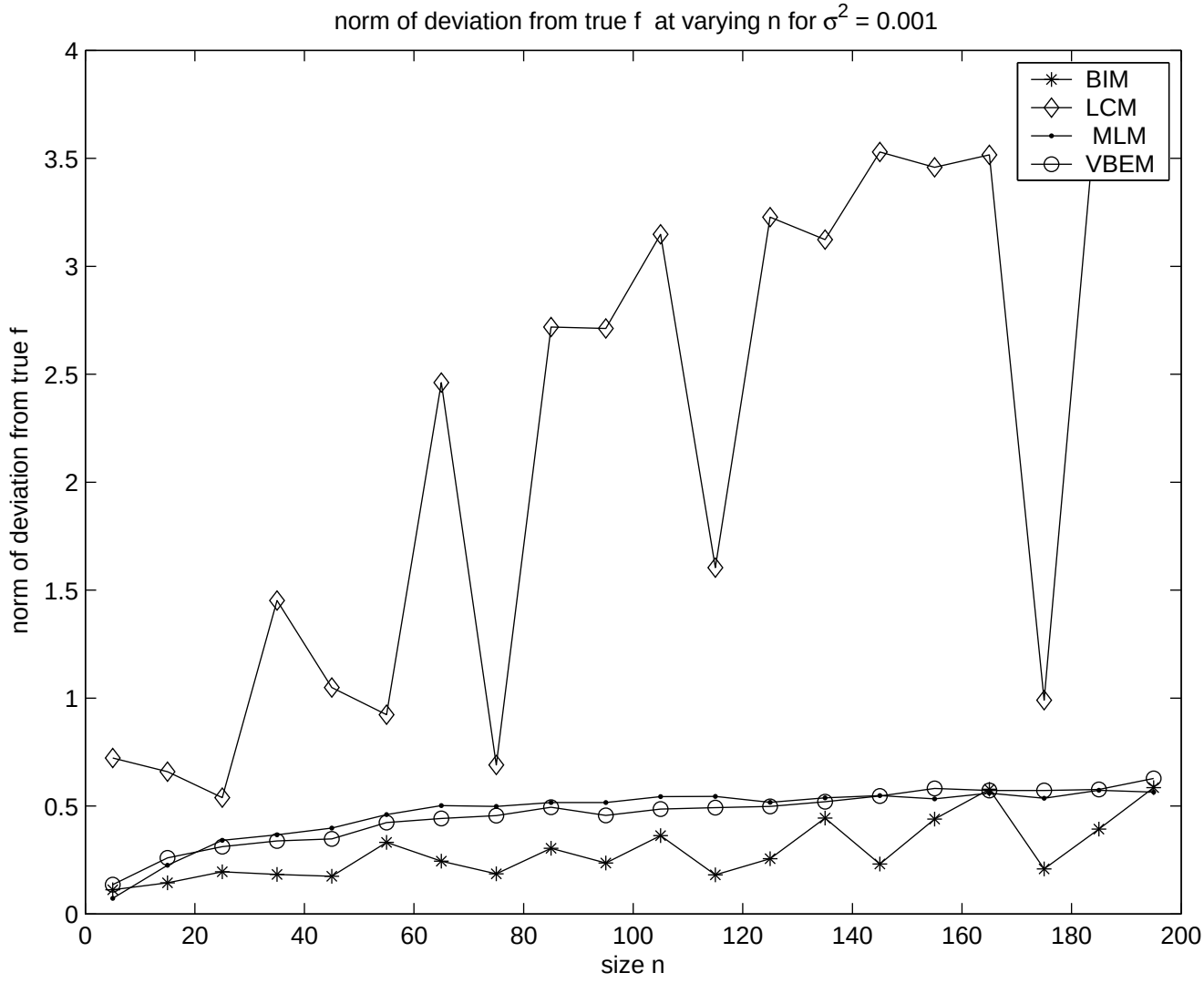

Figure 6.16: $\| f - f_{est} \|_2$ *at varying* n *at* $\sigma^2 = 10^{-3}$.

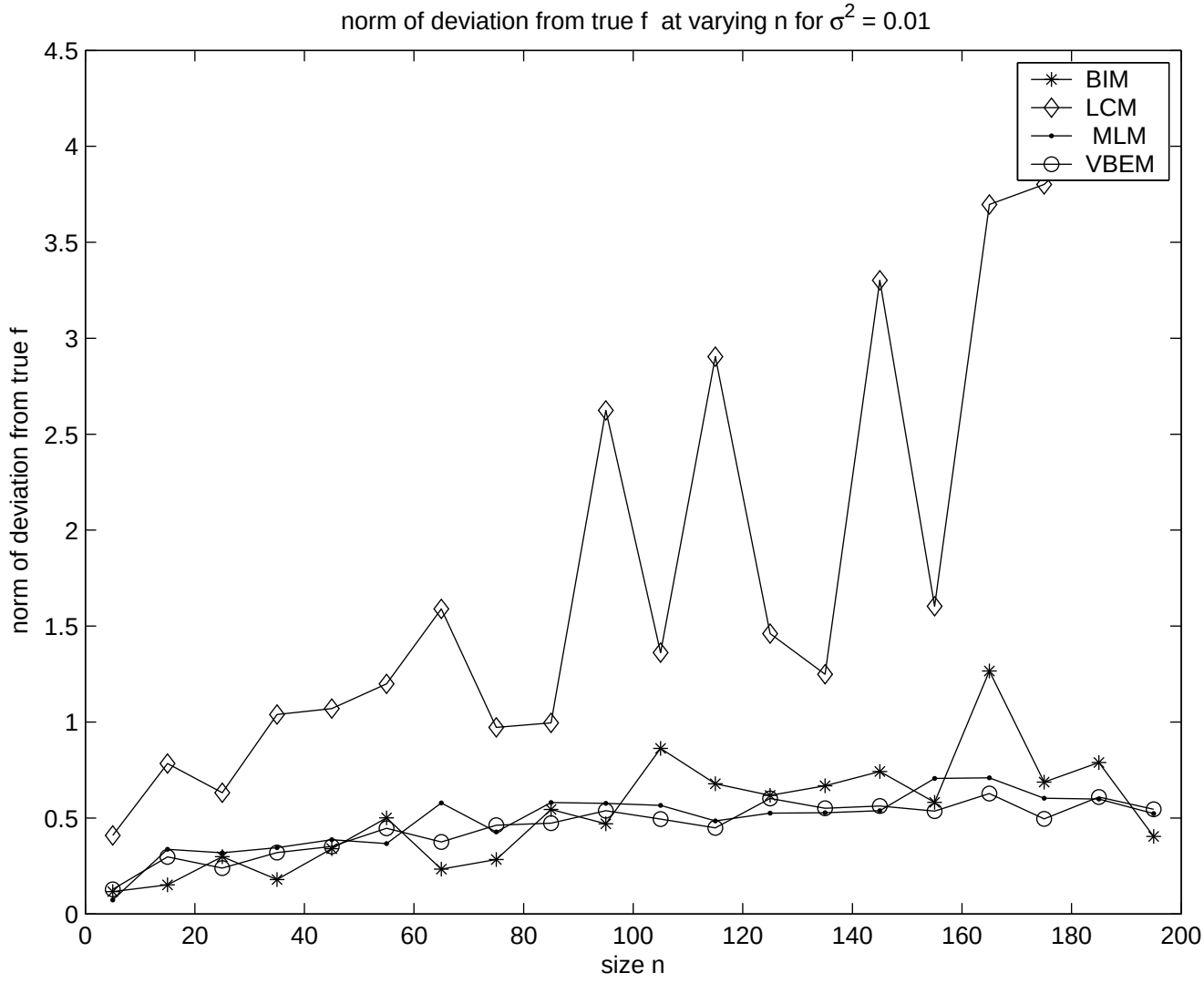

Figure 6.17: $\| f - f_{est} \|_2$ *at varying* n *at* $\sigma^2 = 10^{-2}$.

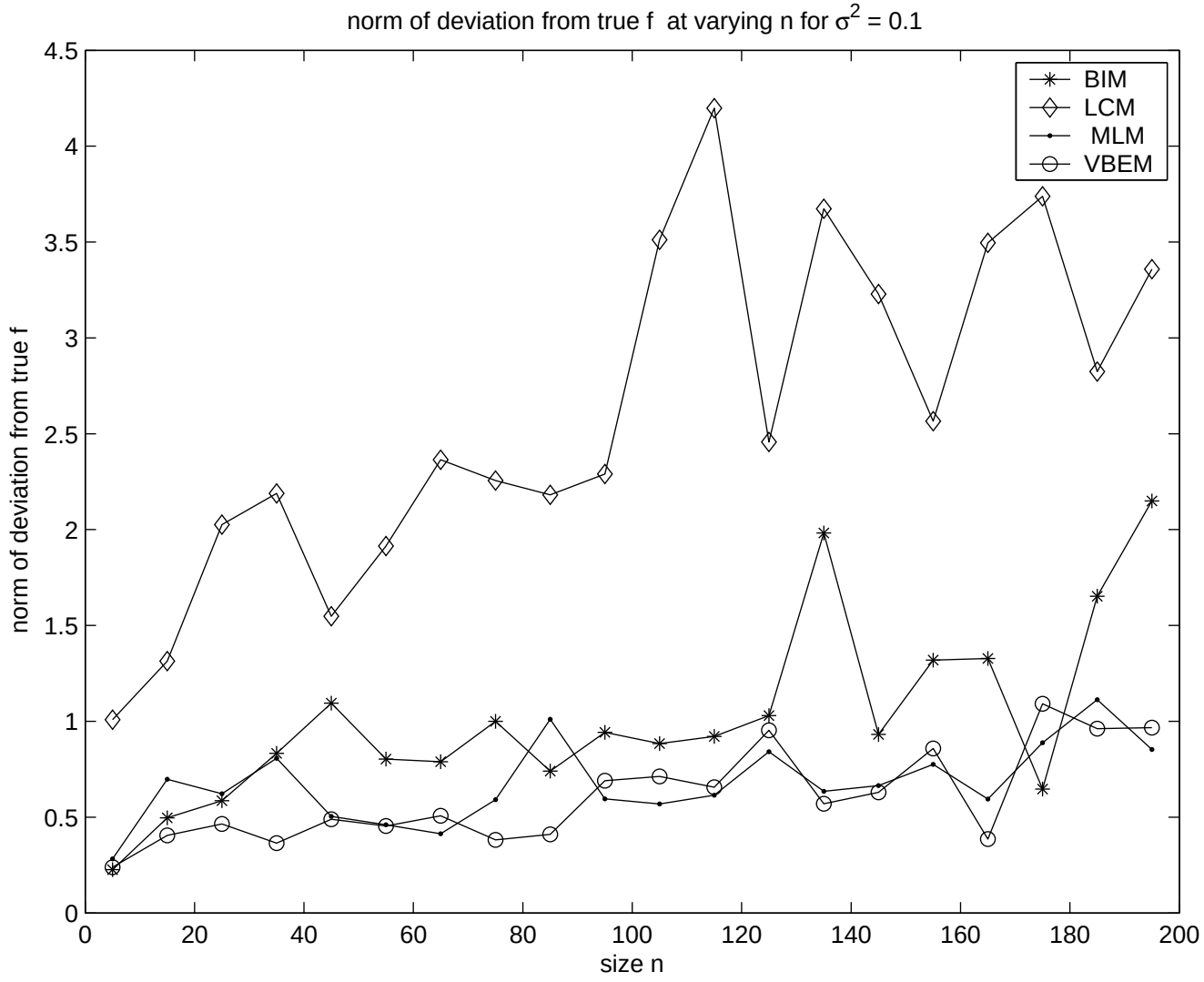

Figure 6.18: $\| f - f_{est} \|_2$ *at varying* n *at* $\sigma^2 = 10^{-1}$.

From the simulation results, BIM looks more sensitive to high noise than ML, VBEM and LCM. The results from the plots show that for large values of the discretization parameter n, $\| f - f_{est} \|_2$ increases as well. At higher noise level LCM is more robust followed by VBEM and MLM with BIM as worst. At a given noise level of say $\sigma^2 \leq 10^{-3}$ BIM is better. On the average, I will go in for VBEM since it is quite robust, computes quite faster and it is able to handle intractable problems which the Bayesian Inference finds difficult to handle.

Contributions

1. Came out with two important sub-conclusions.
2. VBEM optimization algorithm have been derived for the inverse (deconvolution) problem.
3. Shown that any optimum of the Bayesian also corresponds to an optimum of Variational Bayes.
4. Justified that Bayesian can also be viewed as an extension of ML from a Regularization viewpoint (for an acceptable noise-level, σ^2).
5. Shown that for non-informative priors; differences which arise between ML and Bayesian Inference is due to the correction term $\mathrm{Tr}(K\Sigma_f^{-1}K^T)$.
6. Shown from the Gravity Example that estimates obtained using Statistical Methods are better than the L-Curve for Tikhonov Regularization.
7. On heuristic grounds; any optimum of the L-Curve approximation in the Numerical Framework can be used as an estimate in the Statistical Framework.
8. Seen that $\int_\Omega \epsilon^2(s)w(s)\,\mathrm{d}s$ can be viewed as a likelihood function of λ_{rls}^2 with its most probable parameter also given by $\lambda_{rls}^{2\star}$.
9. Proposed a method which incorporated the method of truncated SVD into Ridge Regression.

CHAPTER 7

Conclusion

The work in this thesis focussed on methods within the Statistical and Numerical Framework by using the Gravity Model as our 'yard stick'. We viewed Tikhonov's Regularization from a Statistical viewpoint using ML and MAP. We extended the Tikhonov's functional ML using the Evidence Framework for Bayesian Inference and also estimated the parameters using Variational Bayesian EM Algorithm and we arrived at the same equations which the Evidence Framework gave for the parameters α and β if non-informative priors are assigned. We also saw from Chapter 4 that the Regularization parameter $\lambda^2_{rls} = \lambda^2_{ml}\,\sigma^2$. All the methods followed some similar analytical path. Their main individual differences lie on the criterion for estimating the parameters.

In using the Gravity Model example at varying dimensions, we found that Statistical Methods consistently out-performed the L-Curve estimates at all noise levels considered. The Evidence Framework was better in terms of Root of Square Deviations from true f.

At $\sigma^2 = 0.1$, VBEM, ML were little better than the Evidence Framework with the L-Curve as the worst in terms of Root of Squared Deviations from the true f. In all, the L-Curve was not as consistent as the Evidence Framework, neither was it as consistent as ML (Regularized) nor VBEM. This may be due to the fact that the L-Curve estimation procedure is not built on any well-defined functional form like that given by Statistical methods. This may account for its robustness and ability to handle perturbations consisting of correlated noise.

With respect to the Gravity problem example, we will prefer VBEM to all since it is able to manage well and compute faster too. Moreover, estimates obtained using Statistical Methods have all proven to be better than the L-Curve in terms of smoothness in relation to the reconstruction of input and the norm of Squared Deviations of the reconstructed input and the true input f.

CHAPTER A

Appendix

A.0.1 SVD Asymptotic Analysis of the Explicit Likelihood Function

We have at the back of our minds that the matrix K has some of its singular values d_i to be very small.

$$p(\tilde{g}|\lambda_{ml}^2,\sigma^2) \propto \left(\frac{\lambda_{ml}^2}{2\pi\sigma^2}\right)^{n/2}\left|\left(\frac{D^2}{\sigma^2}+\lambda_{ml}^2 I\right)\right|^{-1/2}\exp\frac{1}{2}\left\{\frac{\tilde{g}^T K}{\sigma^2}\left\{V\left(\frac{D^2}{\sigma^2}+\lambda_{ml}^2 I\right)^{-1}V^T\right\}\frac{K^T\tilde{g}}{\sigma^2}\right\}$$

Asymptotes of the Likelihood Function

(a) asymptotes of $p(\tilde{g}|\lambda_{ml}^2,\sigma^2)$ when $\lambda_{ml}\to\infty$ and $0<\sigma<\infty$

$$\begin{aligned} p(\tilde{g}|\lambda_{ml}^2,\sigma^2) &\simeq \lambda_{ml}^n\left(\frac{1}{\lambda_{ml}}\right)^n\exp\frac{1}{\lambda_{ml}^2} \\ &\simeq \lambda_{ml}^n\left(\frac{1}{\lambda_{ml}}\right)^n\exp\ 0 \\ &\to 1 \end{aligned} \tag{A.1}$$

(b) behaviour of $p(\tilde{g}|\lambda_{ml}^2,\sigma^2)$ when $\lambda_{ml}\to 0$ and $0<\sigma<\infty$

In this case, we have the constraint;

Say, $D_{n,n}>0$ [1] fixed small, such that $\exists\ \lambda_{ml}<D_{n,n}$ for which

$$\begin{aligned} p(\tilde{g}|\lambda_{ml}^2,\sigma^2) &\simeq \lambda_{ml}^n\left\{\Pi_{i=1}^n\left(\frac{\sigma}{d_i}\right)\right\}\exp\sum_{i=1}^n\frac{\sigma^2}{d_i^2} \\ &\to 0 \end{aligned} \tag{A.2}$$

if and only if two of the three expressions on the right hand side of the approximation in (A.2) satifies;

$$\left\{\Pi_{i=1}^n\left(\frac{\sigma}{d_i}\right)\right\}<\infty \quad\text{and}\quad \exp\sum_{i=1}^n\frac{\sigma^2}{d_i^2}<\infty$$

[1] $D_{n,n}$ is the last or smallest singular value.

otherwise

we are kind of doubtful about the likelihood function. Thus

$$p(\tilde{g}|\lambda_{ml}^2, \sigma^2) \quad \rightarrow \quad 0 \times \text{large} \times \text{large} \quad \text{for some} \quad d_i \rightarrow 0 \tag{A.3}$$

In the case of (A.3), the exponential term increases faster than the determinant term. Hence, taking logarithm of the likelihood function will not make the problem go away since it is unlikely, it can nullify the effect of the positive sign on the exponent.

(c) asymptotes of $p(\tilde{g}|\lambda_{ml}^2, \sigma^2)$ when $\sigma \rightarrow \infty$ and $0 < \lambda_{ml} < \infty$

$$\begin{aligned} p(\tilde{g}|\lambda_{ml}^2, \sigma^2) &\simeq \left(\frac{1}{\sigma}\right)^n \left(\frac{1}{\lambda_{ml}}\right)^n \exp\ 0 \\ &\rightarrow 0 \end{aligned} \tag{A.4}$$

(d) behaviour of $p(\tilde{g}|\lambda_{ml}^2, \sigma^2)$ when $\sigma \rightarrow 0$ and $0 < \lambda_{ml} < \infty$

$$p(\tilde{g}|\lambda_{ml}^2, \sigma^2) \quad \rightarrow \quad \infty \quad \text{ie become undefined} \tag{A.5}$$

A.0.2 Variational Bayesian Factorial Approximation

We consider the lower bound of the log evidence defined in terms of $\varnothing$;

$$\mathcal{F} = \int Q(\varnothing) \ln \frac{p(\tilde{g}, \varnothing)}{Q(\varnothing)} \, \mathrm{d}\varnothing \tag{A.6}$$

for say some $\varnothing = \{\varnothing_1, \varnothing_2, \varnothing_3\}$.

Maximizing the functional $\mathcal{F}(\varnothing)$ over the space of probability distribution $Q(\varnothing)$ begins with an assumption about the factorization of $Q(\varnothing)$ which can be written in a separable form:

$$Q(\varnothing) = \Pi_{i=1}^3 Q_i(\varnothing_i) = Q_1(\varnothing_1)\, Q_2(\varnothing_2)\, Q_3(\varnothing_3)$$

Hence

$$\mathcal{F}(Q) = \int Q_1(\varnothing_1)\, Q_2(\varnothing_2)\, Q_3(\varnothing_3) \ln \frac{p(\tilde{g}, \varnothing)}{Q_1(\varnothing_1)\, Q_2(\varnothing_2)\, Q_3(\varnothing_3)} \, \mathrm{d}\varnothing_1 \, \mathrm{d}\varnothing_2 \, \mathrm{d}\varnothing_3 \tag{A.7}$$

From calculus of variation, a maximization of the functional $\mathcal{F}(Q)$ with respect to $Q(\varnothing)$ constrains equation (A.7) to satisfy

$$\int Q(\varnothing) \, \mathrm{d}\varnothing = 1 \tag{A.8}$$

The integrand of equation (A.8) was 'dribbled around' the differential equation

$$\begin{aligned} \dot{z} &= Q(\varnothing) \\ \Rightarrow \quad \mathrm{d}z &= Q(\varnothing)\,\mathrm{d}\varnothing \end{aligned}$$

before initial conditions were applied to obtain an idefinite integral on the left hand side to give the value, 1 on the right hand side. With this notion in mind, we therefore define a new function $z(\varnothing)$ as follows:

$$z(\varnothing) \;=\; \int_{-\infty}^{\varnothing} Q_1(\varnothing_1')\,Q_2(\varnothing_2')\,Q_3(\varnothing_3')\;\mathrm{d}\varnothing_1'\,\mathrm{d}\varnothing_2'\,\mathrm{d}\varnothing_3' \tag{A.9}$$

which gives rise to the differential constraint:

$$\dot{z} \;-\; Q_1\,Q_2\,Q_3 \;=\; 0 \tag{A.10}$$

with the initial conditions (end points) constrained to be $z(-\infty) = 0$ and $z(\infty) = 1$.

Let the integrand of equation (A.7) be defined by

$$h(Q_1, Q_2, Q_3, \varnothing) \;=\; Q_1(\varnothing_1)\,Q_2(\varnothing_2)\,Q_3(\varnothing_3)\ln\frac{p(\tilde{g},\varnothing)}{Q_1(\varnothing_1)\,Q_2(\varnothing_2)\,Q_3(\varnothing_3)} \tag{A.11}$$

Introducing a Lagrange multiplier say α to the constraint of equation (A.10) and adding to equation (A.11) gives

$$\begin{aligned} h_L(Q_1, Q_2, Q_3, \varnothing) \quad &= \quad Q_1(\varnothing_1)\,Q_2(\varnothing_2)\,Q_3(\varnothing_3)\ln\frac{p(\tilde{g},\varnothing)}{Q_1(\varnothing_1)\,Q_2(\varnothing_2)\,Q_3(\varnothing_3)} \\ &\quad + \alpha\,(\dot{z} \;-\; Q_1\,Q_2\,Q_3\,) \end{aligned} \tag{A.12}$$

The integrand of the integral equation (A.7) then takes the form of equation (A.12). Maximizing the functional $\mathcal{F}(Q)$ with respect to each distribution Q_i is tantamount to solving the set of Euler-Lagrange equations;

$$\frac{\partial\, h_L}{\partial Q_i} \;-\; \frac{d}{d\varnothing}\Big(\frac{\partial\, h_L}{\partial \dot{Q}_i}\Big) \;=\; 0 \tag{A.13}$$

$$\frac{\partial\, h_L}{\partial z} \;-\; \frac{d}{d\varnothing}\Big(\frac{\partial\, h_L}{\partial \dot{z}}\Big) \;=\; 0 \tag{A.14}$$

where $\dot{q} = dQ/d\varnothing$.

From equation (A.12), $\partial h_L/\partial\dot{z} = \alpha$. Substituting into (A.14), we get

$$\frac{d\alpha}{d\varnothing} \;=\; 0 \tag{A.15}$$

which shows that α is independent of $\varnothing$. Similarly, differentiating h_L of equation (A.12) with respect to say Q_1 and solving for the zero results

$$Q_2\,Q_3\Big\{\ln p\,(\tilde{g},\varnothing) \;-\; \ln Q_1 \;-\; \ln Q_2\,Q_3\Big\} \;-\; Q_2\,Q_3 \;-\; \alpha\,Q_2\,Q_3 \;=\; 0 \tag{A.16}$$

Integrating the above with respect to $\varnothing_2$ and $\varnothing_3$ is equivalent to the mean (or expectation) under the distributions of $Q(\varnothing_2)$ and $Q(\varnothing_3)$. Thus we obtain

$$\langle \ln p\,(\tilde{g}, \varnothing)\rangle_{Q_2 Q_3} \;-\; \ln Q_1 \;-\; \int Q_2\, Q_3 \ln Q_2 Q_3 \;\mathrm{d}\varnothing_2\, \mathrm{d}\varnothing_3 \;-\; 1 \;-\; \alpha \;=\; 0 \tag{A.17}$$

and solving for Q_1, we get

$$Q_1 \;=\; \frac{\exp \langle \ln p\,(\tilde{g}, \varnothing)\rangle_{Q_2 Q_3}}{\exp\,(1 + \alpha + \int Q_2\, Q_3 \ln Q_2 Q_3 \;\mathrm{d}\varnothing_2\, \mathrm{d}\varnothing_3)} \tag{A.18}$$

From equation (A.15) and our assumption about the factorized form;

$$Q(\varnothing) \;=\; \Pi_i Q_i(\varnothing_i)$$

we can see that the denominator of equation (A.18) is independent of $\varnothing_1$ and so it can be considered as a normalization constant. Hence, we can generally express the solution for the individual Q_i that maximizes the functional $\mathcal{F}(Q)$ under the assumed factorization for each i as

$$Q_i \;=\; \frac{\exp \langle \ln p\,(\tilde{g}, \varnothing)\;\rangle_{Q_{k\neq i}}}{\exp \langle \ln p\,(\tilde{g}, \varnothing)\;\rangle_{Q_{k\neq i}}\, \mathrm{d}\varnothing_i} \tag{A.19}$$

Bibliography

[1] Hansen, P.C., *Deconvolution and regularization with Toeplitz matrices*, Journal, Numerical Algorithms, Department of Mathematical Modelling, Technical University of Denmark. (2002) vol.29 .

[2] Hansen, P.C., *The L-Curve and its use in the numerical treatment of inverse problems*, Department of Mathematical Modelling, Technical University of Denmark.

[3] Hansen, P.C., *Regularization Tools: a Matlab package for analysis and solution of discrete ill-posed problems*, Numerical Algorithms 6. (1994)

[4] Hansen, P.C., *Rank-Deficient and Discrete Ill-Posed Problems*, Society for Industrial and Applied Mathematics, SIAM, Philadelphia, 1988.

[5] Fierro, R.D., Hansen, P.C. and Hansen, P.S.K., *UTV TOOLS: Matlab Templates for Rank-Revealing UTV Decompositions*, Version $1 \cdot 0$ 0 for Matlab $5 \cdot 2$, Technical Report, IMM-REP-1999-2

[6] Mackay, D.J.C., *Information Theory, Inference and Learning Algorithms*, Cambridge University Press.

[7] Mackay, D.J.C., *Bayesian Methods for Neural Networks: Theory and Application*, University of Cambridge Programme for Industry, Cavendish Laboratory, Cambridge.

[8] Mackay, D.J.C., *Bayesian Non-Linear Modelling for the Prediction Competition*, Cavendish Laboratory, Cambridge.

[9] Mackay, D.J.C., *The Evidence Framework applied to Classification Networks*, Computation and Neural Systems, Carlifornia Institute of Technology 139-74, Pasadena CA 91125

[10] Mackay, D.J.C., *Comparison of Approximate Methods for Handling Hyperparameters*, Cavendish Laboratory, Cambridge.

[11] Lartey, G.O., *Lecture notes on Integral Equations*, Integral Equations, Kwame Nkrumah University of Science and Technology, Ghana.

[12] Proakis, J.G. and Dimitris, G. M., *Digital Signal Processing*, Principles, Algorithms and Applications, Prentice Hall, INC. Third Edition

[13] Jazwinski, A.H., *Stochastic Processes and Filtering Theory*, Analytical Mechanics Associates, Inc. Seabrook, Maryland.

[14] Hansen, P.S.K., , Ph.D Thesis, Department of Mathematical Modelling, Technical University of Denmark. 2000.

[15] Urmanov, A.M, Gribok, A.V., Bozdogan, H., Hines, J.W. and Uhrig, R.E., *Information complexity-based regularization parameter selection for solution of ill conditioned inverse problems*, The University of tenenessee, Nuclear Engineering and Statistics Departments, Knoxville.

[16] Bjorck, Ake, *NUMERICAL METHODS FOR Least Squares PROBLEMS*, Society for Industrial and Applied Mathematics, (SIAM), Philadelphia.

[17] Graham, A.,Kronecker Products and Matrix Calculus with Applications, Halsted Press, John Wiley and Sons, NY. 1981.

[18] O'leary, D.P., *NEAR-OPTIMALITY PARAMETERS FOR TIKHONOV AND OTHER REGULARIZATION METHODS**, Society for Industrial and Applied Mathematics, SIAM.

[19] Golub, Gene H. and Van Loan C.F., *Matrix Computation*, John Hopkins university Press, Baltimore and London. Third Edition

[20] Phillips, D.L., *A technique for the Numerical Solution of Certain Integral Equations of the First Kind**, Argonne National Laboratory, Argonne, Illinois.

[21] Twomey, S., *On the Numerical Solution of Fredholm Integral Equations of the First Kind by the Inversion of Linear System Produced by Quadrature**, U.S Weather Bureau, Washington D.C.

[22] Bishop, C.M., *Neural Networks for Pattern Recognition*, Institute of Adaptive and Neural Computation, Division of Informatics, Edinburgh University, OXFORD UNIVERSITY PRESS.

[23] Vio, R., Nagy, J.G., Tenorio, L., Andreani, P., Baccigalupi, C and Wamsteker W., *Digital deblurring of CMB maps: Performance and efficient implementation*, Astronomy and Astrophysics.

[24] Calvetti, D., Lewis, B. and Reichel L. *A hybrid GMRES and TV-norm based method for image restoration*, Department of Mathematics, Kent State University, Kent.

[25] Haykin, Simon, *NEURAL NETWORKS*, A COMPREHENSIVE FOUNDATION, SECOND EDITION (INTERNATIONAL EDITION), Prentice Hall.

[26] Kalman, R.E.,*A New Approach to Linear Filtering and Prediction Problems*, Research Institute for Advanced Study, Baltimore. Md.

[27] Davis, M.H.D. and Vinter R.B., *Stochastic Modelling and Control*, Department of Electrical Engineering, Imperial College, London. CHAPMAN AND HALL.

[28] Thyregod P and Madsen H., *An Introduction to General and Generalized Linear Models*, $0 \cdot 7$ Edition, IMM

[29] Ghahramani Z and Beal J.M., *Variational Inference for Bayesian Mixtures of Factor Analysers*, Gatsby Computational Neuroscience Unit, University College London, England

[30] Ghahramani Z and Hinton G.E., *The EM Algorithm for Mixtures of Factor Analyzers*, Department of Computer Science, University of Toronto, Canada.

[31] Ghahramani Z and Beal J.M., *Propagation Algorithms for Variational Bayesian Learning*, Gatsby Computational Neuroscience Unit, University College London, England

[32] Ghahramani Z and Beal J.M., *The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structure*, Gatsby Computational Neuroscience Unit, University College London, England

[33] Ghahramani Z and Beal J.M., *Graphical models and variational methods*, Gatsby Computational Neuroscience Unit, University College London, England

[34] Attias H., *A Variational Bayesian Framework for Graphical Models*, Gatsby Computational Neuroscience Unit, University College London, England

[35] Ghahramani Z., *On Structured variational Approximations*, Gatsby Computational Neuroscience Unit, University College London, England

[36] Wang X and George E.I., *GA Hierarchical Bayes Approach to variable Selection for Generalized Linear Models*, Department of Statistical Science, Southern Methodist University, Dallas, Texas.

[37] Strawderman W.E., *Minimax Generalized Ridge Regression Estimators*, Journal of the American Statistical Association.

[38] Kubokawa T and Srivasta M.S., *Minimax Empirical Bayes Ridge-Principal Component Regression Estimators*, University of Tokyo.

[39] Sundberg R., *Continum Regression and Ridge Regression*, Stockholm University, Sweden. ,

[40] Kubokawa T and Srivasta M.S., *Improved Empirical Bayes Ridge Regression Estimators under Multicollinearity*, University of Tokyo.

[41] Gonzalez R.C and Woods R.E., *Digital Image Processing*, Review Material, prentice Hall. Second Edition

[42] Beal J.M., *VARIATIONAL ALGORITHMS FOR APPROXIMATE BAYESIAN INFERENCE*, Gatsby Computational Neuroscience Unit, University College London, England

[43] Åstrom K.J., *INTRODUCTION TO STOCHASTIC CONTROL THEORY*, Division of Automatic Control, Lund Institute of Technology, Lund, Sweden