UNIVERSITY OF CAPE COAST

LEAST SQUARES OPTIMIZATION WITH L_1-NORM REGULARIZATION

HENRIETTA NKANSAH

2017

UNIVERSITY OF CAPE COAST

LEAST SQUARES OPTIMIZATION WITH L_1-NORM REGULARIZATION

BY

HENRIETTA NKANSAH

Thesis Submitted to the Department of Mathematics and Statistics of the School of Physical Sciences, College of Agriculture and Natural Sciences, University of Cape Coast, in partial fulfilment of the requirements for the award of Doctor of Philosophy degree in Mathematics

OCTOBER 2017

DECLARATION

Candidate's Declaration

I hereby declare that this thesis is the result of my own original research and that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature ..... Date .....

Name: Henrietta Nkansah

Supervisors' Declaration

We hereby declare that the preparation and presentation of the thesis were supervised in accordance with the guidelines on supervision of thesis laid down by the University of Cape Coast.

Principal Supervisor's Signature ..... Date .....

Name: Prof. Francis Benyah

Co-Supervisor's Signature ..... Date .....

Name: Dr. Henry Amankwah

ABSTRACT

The non-differentiable L_1-norm penalty in the L_1-norm regularized least squares problem poses a major challenge to obtaining an analytic solution. The study thus explores smoothing and non-smoothing approximations that yields differentiable loss functional that ensures a close-form solution in over-determined systems. Three smoothing approximations to the L_1-norm penalty have been examined. These include the Quadratic, Sigmoid and Cubic Hermite. Tikhonov regularization is then applied to the resulting loss function. The approximations are a modification of the Lee's approximation to the L_1-norm term. The regularized solution using this approximation has been presented in various forms. Using the Hilbert 12×12 matrix, it is found that for all three methods, a good approximation to the exact solution converges at a regularization parameter $\mu = 10^{-30}$. The solutions show an accuracy to nine digits. In each approximation, a suitable value of the parameter is obtained for which the absolute value function approximates the L_1-norm penalty. The results of the Modified Newton's method based on the Lee's approximation however shows an accuracy of at most two digits. The solution of the smoothing methods also compares favourably with $l1\_ls$ method. Analytic solution of the L_1-norm problem is also obtained by means of the sub-gradient method, after casting the constrained formulation as unconstrained. Attempt at achieving sparsity of the Least Absolute Shrinkage and Selection Operator (LASSO) solution has been made in two ways. The initial solution is expressed in terms of the singular value decomposition so that by truncating smaller singular values, the desired sparsity is achieved using suitable regularization parameter obtained by the K-fold cross-validation of the fit. In another way, the LASSO solution itself has been induced to ensure sparsity. The results show that the LASSO formulation and solution must be appropriately designed for certain type of datasets, particularly those that are severely ill-conditioned and those with monotone trends.

KEY WORDS

Least Squares Optimization

L_1-Norm Regularization

L_2-Norm Regularization

Non-Smoothing Approximations

Smoothing Approximations

Sparsity Induced Systems

ACKNOWLEDGEMENTS

I am most grateful to my Principal Supervisor, Professor Francis Benyah of the Department of Mathematics and Statistics, University of Cape Coast (UCC) for his invaluable suggestions, comments and guidance throughout the stages of this work. I am also grateful to my Co-Supervisor, Dr. Henry Amankwah, also of the Department of Mathematics and Statistics, UCC, for his contribution towards completion of this thesis.

Thanks are also extended to Dr. Nathaniel Howard, Head, Department of Mathematics and Statistics, UCC, for his support towards completion of my Ph.D programme. I am also thankful to Professor Emmanuel K. Essel for his encouragement and support towards the completion of my programme. I acknowledge with gratitude all lecturers and staff of the Department of Mathematics and Statistics, UCC, for their diverse support and contribution to the success of this work.

I am extremely grateful to Professor F. K. A. Allotey (late), President of AIMS-Ghana and Director of the Institute of Mathematical Sciences, Accra, Ghana for providing me financial support throughout the programme.

Finally, I would like to thank my husband, Dr. Bismark K. Nkansah, and my entire family for their unflinching support throughout my programme.

DEDICATION

To my family

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

In this introductory chapter, the motivation for this study has been specified in the background to the study. This section explains the problem associated with the L_1-norm regularized least squares. In particular, the concepts of least squares method for solving over-determined systems have been explained as well as the Tikhonov (L_2-norm) regularization and the L_1-norm regularization (LASSO). These explanations enable us to make a statement of the problem for the study, and hence identify the main objectives that would guide the thesis. The chapter also introduces a number of datasets that would be needed to illustrate the main ideas developed in Chapters Four and Five. The rationale for the choice of these datasets is clearly highlighted by the description of the features of these datasets. The last section specifies the organisation of the thesis.

Background to the Study

Many optimization problems arising in many applications require minimization of an objective cost function that is convex but not differentiable. Such a minimization arises, for example, in signal reconstruction, image processing, data fitting and general allocation problems. To solve convex but non-differentiable problems, we have to employ special methods that can work in the absence of differentiability, while taking advantage of convexity and possibly other special structures that the minimization problem may possess. In this thesis, we study ways of solving convex problems that are not differentiable.

The method of least squares is a standard approach to the approximate solution of over-determined systems of the form

$$\mathbf{X}\alpha = \mathbf{y}, \tag{1.1}$$

where $\mathbf{X} \in \Re^{m \times p}$, $m > p$, $\mathbf{y} \in \Re^m$ and $\alpha \in \Re^p$. Thus, in such sets of equations, there are more equations than unknowns. By least squares method, the overall solution minimizes $\|\mathbf{X}\alpha - \mathbf{y}\|_2^2$, the sum of squares of errors made in the results of every single equation.

The most important application of this method is in data fitting. The best fit in the least squares sense minimizes the sum of squared residuals, a residual being the difference between an observed value and the fitted value provided by a model. When the problem has substantial uncertainties in the independent variable (the x variable), then least squares regression is not suitable; in such a case, the methodology required for fitting errors-in-variables models may be considered instead of least squares.

Least squares problems fall into two categories: linear or ordinary least squares and non-linear least squares, depending on whether or not the residuals are linear in all unknowns. The linear least squares problem occurs in statistical regression analysis which has a closed-form solution that can be evaluated in a finite number of standard operations. The non-linear problem has no closed-form solution and is usually solved by iterative refinement; at each iteration, the system is approximated by a linear one, and thus the core calculation is similar in both cases.

This thesis focuses on optimizing the least squares objective function with an L_1-penalty on the parameters. There is currently significant interest in this and related problems in a wide variety of fields, due to the appealing idea of creating accurate predictive models that also have interpretable or parsimonious representations. Rather than focus on applications or properties of these models, the main contribution of this thesis is an examination of a variety of the approaches that have been proposed for parameter estimation in these models. It then proposes other ways to some of the approaches using regularization. Least square problems often have their origin in fitting models to observations. In

its simplest form, we know this from the problem of fitting a regression line, $y = ax + b$, through a set of data points $(x_i,\ y_i),\ i = 1,\ldots,m$. When $m > 2$, it is in general impossible to put the line through all points. However, we try to determine an optimal line, for example, by determining the pair $\{a^*,\ b^*\}$ which minimizes the objective function

$$f(a,\ b) = \sum_{i=1}^{m} (ax_i + b - y_i)^2.$$

In this simple case, it is easy to derive the solution analytically, but in general the solution has to be found numerically. Since these problems are so common, there has been a lot of work involved in adapting the general algorithms for unconstrained optimization to this special case. It is more effective to use specially adapted algorithms instead of the more general ones. We first consider the linear least squares problem and the normal equations.

Linear Least Squares Problem and the Normal Equations

Least square problems arise in statistical and geometric applications that require fitting a polynomial or a curve to an experimental data. Methods for numerically solving the least squares problem invariably lead to solving a linear system. Linear least squares problems occur when solving over-determined linear systems. In general, over-determined system has no solution, but we may find a meaningful approximate solution by minimizing some norm of the residual vector.

Given the matrix $\mathbf{X}$ in Equation (1.1), we find a vector $\alpha \in \Re^p$ such that the norm of the residual vector, $\mathbf{r} = \mathbf{X}\alpha - \mathbf{y}$ is minimized. That is, we solve the problem

$$\min_{\alpha \in \Re^p} \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 . \tag{1.2}$$

The calculations are simplest when we choose the norm-2. Thus, we will minimize the square of the length of the residual vector

$$\|\mathbf{r}\|_2^2 = r_1^2 + r_2^2 + \cdots + r_m^2.$$

To see that this minimum exists and is attained by some $\alpha \in \Re^p$, we note that $\mathbf{E} = \{\mathbf{X}\alpha - \mathbf{y} |\ \alpha \in \Re^p\}$ is a non-empty, closed and convex subset of $\Re^m$. Since $\Re^m$ equipped with the Euclidean inner product is a Hilbert space, E contains a unique element of smallest norm, so there exists an $\alpha \in \Re^p$ (not necessarily unique) such that $\|\mathbf{X}\alpha - \mathbf{y}\|_2$ is minimum. The minimization problem in Equation (1.2) is the Least Squares Method. We characterise the least squares solution by the following theorem.

Theorem 1: Least Squares Solution

Let

$$\mathbf{S} = \left\{ \alpha \in \Re^p : \min_{\alpha} \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 \right\}$$

be the set of solutions of $\mathbf{X}\alpha = \mathbf{y}$ and let $\mathbf{r}_\alpha = \mathbf{X}\alpha - \mathbf{y}$ denote the residual for a specific α. Then

$$\alpha \in \mathbf{S} \Longleftrightarrow \mathbf{X}^T\mathbf{r}_\alpha = 0 \Longleftrightarrow \mathbf{r}_\alpha \perp R(\mathbf{X}), \tag{1.3}$$

where $R(\mathbf{X})$ denotes the subspace spanned by the columns of $\mathbf{X}$.

proof

We prove the first equivalence, from which the second one follows easily.

($\Leftarrow$:) Let $\mathbf{X}^T\mathbf{r}_\alpha = 0$ and $\mathbf{z} \in \Re^p$ be an arbitrary vector. It follows that

$$\mathbf{r}_\mathbf{z} = \mathbf{X}\mathbf{z} - \mathbf{y} = \mathbf{X}\alpha - \mathbf{y} + \mathbf{X}(\mathbf{z} - \alpha),$$

thus $\mathbf{r}_\mathbf{z} = \mathbf{r}_\alpha + \mathbf{X}(\mathbf{z} - \alpha)$. Now

$$\|\mathbf{r}_\mathbf{z}\|_2^2 = \|\mathbf{r}_\alpha\|_2^2 + 2(\mathbf{z} - \alpha)^T\mathbf{X}^T\mathbf{r}_\alpha + \|\mathbf{X}(\alpha - \mathbf{z})\|_2^2.$$

But $\mathbf{X}^T\mathbf{r}_\alpha = 0$ and therefore $\|\mathbf{r}_\mathbf{z}\|_2 \geq \|\mathbf{r}_\alpha\|_2$. Since this holds for every $\mathbf{z}$, then $\alpha \in \mathbf{S}$.

($\Rightarrow$:) We show this by contradiction: assume $\mathbf{X}^T\mathbf{r}_\alpha = \mathbf{z} \neq 0$. We consider $\mathbf{u} = \alpha + \varepsilon\mathbf{z}$ with $\varepsilon > 0$:

$\mathbf{r}_\mathbf{u} = \mathbf{X}\mathbf{u} - \mathbf{y} = \mathbf{X}\alpha - \mathbf{y} - \varepsilon\mathbf{X}\mathbf{z} = \mathbf{r}_\alpha - \varepsilon\mathbf{X}\mathbf{z}$. Now

$$\|\mathbf{r}_\mathbf{u}\|_2^2 = \|\mathbf{r}_\alpha\|_2^2 - 2\varepsilon\mathbf{z}^T\mathbf{X}^T\mathbf{r}_\alpha + \varepsilon^2\|\mathbf{X}\mathbf{z}\|_2^2.$$

Since $\mathbf{X}^T\mathbf{r}_\alpha = \mathbf{z}$, we obtain

$$\|\mathbf{r}_\mathbf{u}\|_2^2 = \|\mathbf{r}_\alpha\|_2^2 - 2\varepsilon\|\mathbf{z}\|_2^2 + \varepsilon^2\|\mathbf{X}\mathbf{z}\|_2^2.$$

We conclude that, for sufficient small ε, we can obtain

$$\|\mathbf{r}_\mathbf{u}\|_2^2 < \|\mathbf{r}_\alpha\|_2^2.$$

This is a contradiction, since α cannot be in the set of solutions in this case. Thus, the assumption is wrong, that is, we must have $\mathbf{X}^T\mathbf{r}_\alpha = 0$, which proves the first equivalence in Equation (1.3). This ends the proof.

The least squares solution has an important statistical property which is expressed in the following Gauss-Markoff Theorem. Let the vector $\mathbf{y}$ of observations be related to an unknown parameter vector α by the linear relation

$$\mathbf{X}\alpha = \mathbf{y} + \varepsilon, \tag{1.4}$$

where $\mathbf{X} \in \Re^{m\times p}$ is a known matrix and ε is a vector of random errors. In this standard linear model, it is assumed that the random variables ε_j are uncorrelated and all have zero mean and the same variance.

Theorem 2: Gauss-Markoff

Consider the standard linear model in Equation (1.4). Then the best linear unbiased estimator of any linear function $\mathbf{c}^T\alpha$ is the least squares solution of

$$\min_\alpha \|\mathbf{X}\alpha - \mathbf{y}\|_2^2.$$

Equation (1.3) can be used to determine the least squares solution. From $\mathbf{X}^T\mathbf{r}_\alpha = 0$, it follows that $\mathbf{X}^T(\mathbf{X}\alpha - \mathbf{y}) = 0$, and we obtain the Normal Equations of Gauss-Markoff:

$$\mathbf{X}^T\mathbf{X}\alpha = \mathbf{X}^T\mathbf{y}. \tag{1.5}$$

We continue this section with typical examples (Gander, Gander & Kwok, 2014) that lead to least squares problems.

Example 1: Measuring a line segment

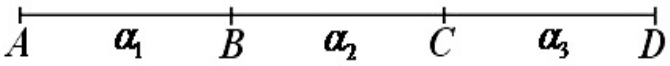

Figure 1: Graph of a Line Segment.

Consider Figure 1 and assume that we have performed five measurements. Let AD = 89 mm, AC= 67 mm, BD = 53 mm, AB = 35 mm and CD = 20 mm, and we want to determine the estimates of the length of the segments $\alpha_1 = AB$, $\alpha_2 = BC$ and $\alpha_3 = CD$. According to the observations, we obtain a linear system with more equations than unknowns. The system is given by

$$\begin{aligned} \alpha_1 + \alpha_2 + \alpha_3 &= 89 \\ \alpha_1 + \alpha_2 &= 67 \\ \alpha_2 + \alpha_3 &= 53 \\ \alpha_1 &= 35 \\ \alpha_3 &= 20 \end{aligned}$$

This can be written in the form

$$\mathbf{X}\alpha = \mathbf{y}, \quad \mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 89 \\ 67 \\ 53 \\ 35 \\ 20 \end{bmatrix}.$$

It can be noticed that if we use the last three equations, then we obtain the least squares solution $\alpha_1 = 35,\ \alpha_2 = 33$ and $\alpha_3 = 20$.

However, if we check the first two equations by inserting this solution we obtain

$$\begin{aligned} \alpha_1 + \alpha_2 + \alpha_3 - 89 &= -1 \\ \alpha_1 + \alpha_2 - 67 &= 1. \end{aligned}$$

So, the equations contradict each other because of measurement errors, and the over-determined system has no solution. A remedy is to find an approximate solution that satisfies the equations as well as possible. For that purpose, one introduces the residual vector $\mathbf{r} = \mathbf{y} - \mathbf{X}\alpha$. One then looks for a vector α that minimizes in some sense this residual vector.

Example 2

The amount h of a component in a chemical reaction decreases exponentially with time t according to the relation

$$h(t) = a_0 + a_1 e^{-bt}.$$

If the material is weighed at different times, we obtain measured values given by $t = t_1, \ldots, t_m$ and $y = y_1, \ldots, y_m$.

The problem now is to estimate the model parameters a_0, a_1 and b from these observations. Each measurement point $(t_i,\ y_i)$ yields an equation:

$$h(t_i) = a_0 + a_1 e^{-bt_i} \approx y_i, \quad i = 1, \ldots, m. \tag{1.6}$$

If there were no measurement errors, then we could replace the approximate symbol in Equation (1.6) by an equality and use three equations from the set to determine the parameters. However, in practice, measurement errors are inevitable. Furthermore, the model equations are often not quite correct and only model the physical behaviour approximately. The equations will therefore in general contradict each other and we need some mechanism to balance the measurement errors, for example, by requiring that Equation (1.6) be satisfied as well as possible.

The above examples are illustrations of different classes of approximation problems. For instance, in Example 1, the equations are linear and in Example 2 (chemical reactions), the system of equations is non-linear.

Example 3

We return to Example 1 and solve it using the Normal Equations.

$$\mathbf{X}^T\mathbf{X}\alpha = \mathbf{X}^T\mathbf{y} \Rightarrow \begin{bmatrix} 3 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} 191 \\ 209 \\ 162 \end{bmatrix}.$$

The solution of this 3×3 system is

$$\alpha = \begin{bmatrix} 35.125 \\ 32.500 \\ 20.625 \end{bmatrix}.$$

The residual for this solution becomes

$$\mathbf{r} = \mathbf{X}\alpha - \mathbf{y} = \begin{bmatrix} 0.7500 \\ -0.6250 \\ -0.1250 \\ -0.1250 \\ -0.6250 \end{bmatrix},$$

with $\|\mathbf{r}\|_2 = 1.1726$.

We notice that for the solution $\alpha = (35,\ 33,\ 20)^T$ obtained by solving the last three equations in Example 1 has a larger residual $\|\mathbf{r}\|_2 = 1.4142$.

There is also a way to understand the normal equations geometrically from Equation (1.3). We want to find a linear combination of columns of the matrix $\mathbf{X}$ to approximate the vector $\mathbf{y}$. The space spanned by the columns of $\mathbf{X}$ is the range of $\mathbf{X}$, $R(\mathbf{X})$, which is a hyperplane in $\Re^m$, and the vector $\mathbf{y}$ in general does not lie in this hyperplane, as shown in Figure 2. Thus, minimizing $\|\mathbf{X}\alpha - \mathbf{y}\|_2$ is equivalent to minimizing the length of the residual vector $\mathbf{r}$, and thus has to be orthogonal to $R(\mathbf{X})$, as shown in Figure 2.

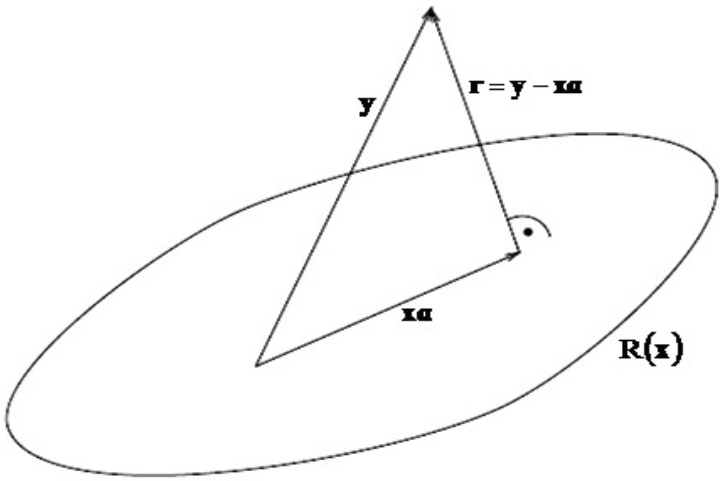

Figure 2: Graph Showing $\mathbf{r}$ Orthogonal to $R(\mathbf{X})$.

The normal equations in Equation (1.5) is concentrated on the matrix $\mathbf{B} = \mathbf{X}^T\mathbf{X}$ which is $p \times p$, and $\mathbf{X}$ is $m \times p$. The matrix $\mathbf{B}$ is symmetric, and if $\text{rank}(\mathbf{X}) = p$, then it is also positive definite. Thus, the natural way to solve the normal equations is by means of the Cholesky decomposition. We notice that when solving linear systems $\mathbf{X}\alpha = \mathbf{y}$ with p equations and p unknowns by Gaussian elimination, reducing the system to triangular form, we make use of the fact that equivalent systems have the same solutions:

$$\mathbf{X}\alpha = \mathbf{y} \Longleftrightarrow \mathbf{B}\mathbf{X}\alpha = \mathbf{B}\mathbf{y} \text{ if } \mathbf{B} \text{ is non-singular.}$$

For a system of equations $\mathbf{X}\alpha \approx \mathbf{y}$ to be solved in the least squares sense, it no longer holds that multiplying by a nonsingular matrix $\mathbf{B}$ leads to an equivalent system. This is because the transformed residual $\mathbf{Br}$ may not have the same norm as $\mathbf{r}$ itself. However, if we restrict ourselves to the class of orthogonal matrices,

$$\mathbf{X}\alpha \approx \mathbf{y} \Longleftrightarrow \mathbf{B}\mathbf{X}\alpha \approx \mathbf{B}\mathbf{y} \text{ if } \mathbf{B} \text{ is orthogonal.}$$

Then, the least squares problems remain equivalent, since $\mathbf{r} = \mathbf{X}\alpha - \mathbf{y}$ and $\mathbf{Br} = \mathbf{By} - \mathbf{BX}\alpha$ have the same length,

$$\|\mathbf{Br}\|_2^2 = (\mathbf{Br})^T(\mathbf{Br}) = \mathbf{r}^T\mathbf{B}^T\mathbf{Br} = \mathbf{r}^T\mathbf{r} = \|\mathbf{r}\|_2^2.$$

Orthogonal matrices and the matrix decompositions containing orthogonal factors therefore play an important role in algorithms for the solution of linear least squares problems. Often it is possible to simplify the equations by pre-multiplying the system by a suitable orthogonal matrix. In this regard, the singular value decomposition (SVD) is an important tool for analysing the linear problem. The SVD solution enables us to identify the smaller singular values which is believed to contribute to the distortions in the solution and hence produces a large residual norm. Methods that make use of SVD isolate these small singular values in an attempt to improve on the solution. The study will explore

the SVD in attempts at improving upon results that would be obtained by the methods developed in the thesis. We will briefly describe here some important aspects of the SVD and provide a more extensive discussion in Chapter Three on review of methods.

Singular Value Decomposition

The singular value decomposition (SVD) of a matrix $\mathbf{X}$ is a very useful tool in the context of least squares problems. It is also very helpful for analysing properties of a matrix.

Theorem 3: Singular Value Decomposition

Let $\mathbf{X} \in \Re^{m\times p}$ with $m \geq p$. Then there exists orthogonal matrices $\mathbf{U} \in \Re^{m\times m}$ and $\mathbf{V} \in \Re^{p\times p}$ and a diagonal matrix $\sum = \text{ diag } (\sigma_1, \ldots, \sigma_p) \in \Re^{m\times p}$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$, such that

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$$

holds. The column vectors of $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$ are called the left singular vectors and similarly, $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ are the right singular vectors. The values σ_i are called the singular values of $\mathbf{X}$. If $\sigma_r > 0$ is the smallest nonzero singular value, then the matrix $\mathbf{X}$ has rank *r*.

proof

The norm-2 of $\mathbf{X}$ is defined by

$$\|\mathbf{X}\|_2 = \max_{\|\alpha\|_2=1} \|\mathbf{X}\alpha\|_2 .$$

Thus, there exists a vector α with $\|\alpha\|_2 = 1$ such that $\mathbf{z} = \mathbf{X}\alpha$, $\|\mathbf{z}\|_2 = \|\mathbf{X}\|_2 =: \sigma$. Let $\mathbf{y} := \mathbf{z}/\|\mathbf{z}\|_2$. This yields $\mathbf{X}\alpha = \sigma\mathbf{y}$ with $\|\alpha\|_2 = \|\mathbf{y}\|_2 = 1$. Next we extend α into an orthonormal basis of $\Re^p$. If $\mathbf{V} \in \Re^{p\times p}$ is the matrix containing the

basis vectors as columns, then $\mathbf{V}$ is an orthogonal matrix that can be written as $\mathbf{V} = [\alpha,\ \mathbf{V}_1]$, where $\mathbf{V}_1^T\alpha = 0$. Similarly, we can construct an orthogonal matrix $\mathbf{U} \in \Re^{m\times m}$ satisfying $\mathbf{U} = [\mathbf{y},\ \mathbf{U}_1]$, $\mathbf{U}_1^T\mathbf{y} = 0$. Now,

$$\mathbf{X}_1 = \mathbf{U}^T\mathbf{X}\mathbf{V} = \begin{bmatrix} \mathbf{y}^T \\ \mathbf{U}_1^T \end{bmatrix} \mathbf{X}[\alpha,\ \mathbf{V}_1] = \begin{bmatrix} \mathbf{y}^T\mathbf{X}\alpha & \mathbf{y}^T\mathbf{X}\mathbf{V}_1 \\ \mathbf{U}_1^T\mathbf{X}\alpha & \mathbf{U}_1^T\mathbf{X}\mathbf{V}_1 \end{bmatrix} = \begin{bmatrix} \sigma & \mathbf{w}^T \\ 0 & \mathbf{B} \end{bmatrix},$$

because $\mathbf{y}^T\mathbf{X}\alpha = \mathbf{y}^T\sigma\mathbf{y} = \sigma\mathbf{y}^T\mathbf{y} = \sigma$ and $\mathbf{U}_1^T\mathbf{X}\alpha = \sigma\mathbf{U}_1^T\mathbf{y} = 0$ since $\mathbf{U}_1 \perp \mathbf{y}$.

We claim that $\mathbf{w}^T := \mathbf{y}^T\mathbf{X}\mathbf{V}_1 = 0$. In order to prove this, we compute

$$\mathbf{X}_1 \begin{bmatrix} \sigma \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} \sigma^2 + \|\mathbf{w}\|_2^2 \\ \mathbf{B}\mathbf{w} \end{bmatrix}$$

and conclude from that equation that

$$\left\| \mathbf{X}_1 \begin{bmatrix} \sigma \\ \mathbf{w} \end{bmatrix} \right\|_2^2 = \left(\sigma^2 + \|\mathbf{w}\|_2^2\right)^2 + \|\mathbf{B}\mathbf{w}\|_2^2 \geq \left(\sigma^2 + \|\mathbf{w}\|_2^2\right)^2.$$

Now, since $\mathbf{V}$ and $\mathbf{U}$ are orthogonal, $\|\mathbf{X}_1\|_2 = \left\|\mathbf{U}^T\mathbf{X}\mathbf{V}\right\|_2 = \|\mathbf{X}\|_2 = \sigma$ holds and

$$\sigma^2 = \|\mathbf{X}_1\|_2^2 = \max_{\|\alpha\|_2\neq 0} \frac{\|\mathbf{X}_1\alpha\|_2^2}{\|\alpha\|_2^2} \geq \frac{\left\| \mathbf{X}_1 \begin{pmatrix} \sigma \\ \mathbf{w} \end{pmatrix} \right\|_2^2}{\left\| \begin{pmatrix} \sigma \\ \mathbf{w} \end{pmatrix} \right\|_2^2} \geq \frac{\left(\sigma^2 + \|\mathbf{w}\|_2^2\right)^2}{\sigma^2 + \|\mathbf{w}\|_2^2}$$

The last equation reads

$$\sigma^2 \geq \sigma^2 + \|\mathbf{w}\|_2^2,$$

and we conclude that $\mathbf{w} = 0$. Thus, we have obtained

$$\mathbf{X}_1 = \mathbf{U}^T\mathbf{X}\mathbf{V} = \begin{bmatrix} \sigma & 0 \\ 0 & \mathbf{B} \end{bmatrix}.$$

This ends the proof.

We can now apply the same construction to the sub-matrix **B** and thus finally end up with a diagonal matrix. Although the proof is constructive, the singular value decomposition is not usually computed in this way. An efficient numerical algorithm was designed by Golub and Reinsch (1970). They first transformed the matrix by orthogonal householder transformations to bidiagonal form. Then the bidiagonal matrix is further diagonalised in an iterative process by a variant of the QR Algorithm.

If we write the equation $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$ in partitioned form, in which $\Sigma_{\mathbf{r}}$ contains only the non-zero singular values, we get

$$\begin{aligned}\mathbf{X} &= [\mathbf{U}_1,\ \mathbf{U}_2]\begin{bmatrix}\Sigma_{\mathbf{r}} & 0\\ 0 & 0\end{bmatrix}[\mathbf{V}_1,\ \mathbf{V}_2]^T\\ &= \mathbf{U}_1\Sigma_{\mathbf{r}}\mathbf{V}_1^T\\ &= \sum_{i=1}^{\mathbf{r}}\sigma_i\mathbf{U}_i\mathbf{V}_i^T.\end{aligned}$$

Given the SVD of **X**, the unique solution with the smallest norm is given as

$$\alpha^* = \mathbf{V}\Sigma^{-1}\mathbf{U}^T\mathbf{y} = \sum_{i=1}^{\mathbf{r}}\frac{\mathbf{U}_i^T\mathbf{y}}{\sigma_i}\mathbf{V}_i.$$

Discrete Ill-Posed Problems

We say that the algebraic problems $\mathbf{X}\alpha = \mathbf{y}$ and $\min_{\alpha\in\Re^p}\|\mathbf{X}\alpha - \mathbf{y}\|_2^2$ are discrete ill-posed problems if the matrix **X** is ill-conditioned and all its singular values decay to zero in such a way that there is no particular gap in the singular value spectrum.

When discrete ill-posed problems are analysed and solved by various numerical regularization techniques, a very convenient way to display information about the regularized solution is to plot the norm or seminorm of the solution versus the norm of the residual vector. In particular, the graph associated with Tikhonov regularization plays a central role. We will illustrate the use of this

graph in the numerical treatment of discrete ill-posed problems. The graph is characterised quantitatively, and several important relations between regularized solutions and the graph are derived. It is also demonstrated that several methods for choosing the regularization parameter are related to locating a characteristic L-shaped "corner" of the graph.

Most numerical methods for treating discrete ill-posed problems seek to overcome the problems associated with the large condition number of $\mathbf{X}$ by replacing the problem with a "nearby" well-conditioned problem whose solution approximates the required solution and, in addition, is a more satisfactory solution than the ordinary (least squares) solution. The latter goal is achieved by incorporating additional information about the sought solution, often that the computed solution should be smooth. Such methods are called regularization methods, and they always include a so-called regularization parameter λ which controls the degree of smoothing or regularization applied to the problem. As the regularization parameter λ varies, we obtain a regularized solutions α_λ having properties that vary with λ. A convenient way to display and understand these properties is to plot the norm or, more generally, a seminorm of the regularized solution, $\|\mathbf{L}\alpha_\lambda\|$, versus the norm of the corresponding residual vector, $\|\mathbf{X}\alpha_\lambda - \mathbf{y}\|$. This was originally suggested in the classic book by Lawson and Hanson (1974).

Ill-Conditioning in Linear Systems

Let $\mathbf{X}$ be a $p \times p$ non-singular matrix and let $\hat{\alpha}$ be the computed solution to the linear system in Equation (1.1). The error vector is given by $\mathbf{e} = \alpha - \hat{\alpha}$. If $\|.\|$ is the norm on $\Re^p$, then $\|\mathbf{e}\|$ is a measure of the absolute error and $\dfrac{\|\mathbf{e}\|}{\|\alpha\|}$ is a measure of the relative error. Generally, we have no way of determining the exact value of $\|\mathbf{e}\|$ and $\dfrac{\|\mathbf{e}\|}{\|\alpha\|}$, since in most practical problems, the exact solution is not known. One way of testing the accuracy of the computed solution $\hat{\alpha}$ is to

compute the residual vector, $\mathbf{r} = \mathbf{y} - \mathbf{x}\hat{\alpha}$ and see how small the relative residual

$$\frac{\|\mathbf{r}\|}{\|\mathbf{y}\|} = \frac{\|\mathbf{y} - \mathbf{X}\hat{\alpha}\|}{\|\mathbf{y}\|}$$

is. Unfortunately, a small residual does not always guarantee the accuracy of the solution, as the following example shows.

Example 4

The linear system $\mathbf{X}\alpha = \mathbf{y}$ given by

$$\begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix}$$

has the exact solution $\alpha = (1,1)^T$. Now, the vector $\hat{\alpha} = (3,0)^T$ produces the residual

$$\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\alpha} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.0002 \end{bmatrix}$$

The relative residual $\dfrac{\|\mathbf{r}\|}{\|\mathbf{y}\|} = 0.000066664$, which is small, even though the solution $\hat{\alpha} = (3,0)^T$ is no where near the exact solution $\alpha = (1,1)^T$. Example 4 phenomena can be explained by the following Theorem.

Theorem 4: The Residual Theorem

Let $\hat{\alpha}$ be the computed solution to the linear system $\mathbf{X}\alpha = \mathbf{y}$. Then

$$\frac{\|\alpha - \hat{\alpha}\|}{\|\alpha\|} \leq \|\mathbf{X}\| \left\|\mathbf{X}^{-1}\right\| \frac{\|\mathbf{r}\|}{\|\mathbf{y}\|}.$$

Proof

From $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\alpha} = \mathbf{X}\alpha - \mathbf{X}\hat{\alpha} = \mathbf{X}(\alpha - \hat{\alpha})$, we have $\alpha - \hat{\alpha} = \mathbf{X}^{-1}\mathbf{r}$ where $\mathbf{X}$ is nonsingular. Taking norms gives

$$\|\alpha - \hat{\alpha}\| \leq \left\|\mathbf{X}^{-1}\right\| \|\mathbf{r}\|. \tag{1.7}$$

Also, from $\mathbf{y} = \mathbf{X}\alpha$, we have $\|\mathbf{y}\| \leq \|\mathbf{X}\| \|\alpha\|$. That is,

$$\frac{1}{\|\alpha\|} \leq \frac{\|\mathbf{X}\|}{\|\mathbf{y}\|}. \tag{1.8}$$

Combining Equations (1.7) and (1.8) gives

$$\frac{\|\alpha - \hat{\alpha}\|}{\|\alpha\|} \leq \|\mathbf{X}\| \left\|\mathbf{X}^{-1}\right\| \frac{\|\mathbf{r}\|}{\|\mathbf{y}\|}.$$

Theorem 4 shows that the relative error in the computed solution $\hat{\alpha}$ depends not only on the relative residual, but also on the quantity $\|\mathbf{X}\| \left\|\mathbf{X}^{-1}\right\|$. A computed solution can be guaranteed to be accurate only when the product $\|\mathbf{X}\| \left\|\mathbf{X}^{-1}\right\| \frac{\|\mathbf{r}\|}{\|\mathbf{y}\|}$ is small. This ends the proof.

The Condition Number of a Matrix

Let $\mathbf{X}$ be $p \times p$ non-singular matrix. The condition number of $\mathbf{X}$, denoted by $\kappa(\mathbf{X})$ is defined as

$$\kappa(\mathbf{X}) = \|\mathbf{X}\| \left\|\mathbf{X}^{-1}\right\|.$$

Example 5

Let

$$\mathbf{X} = \begin{bmatrix} 3 & 3 \\ 4 & 5 \end{bmatrix}. \quad \text{Then} \quad \mathbf{X}^{-1} = \frac{1}{3}\begin{bmatrix} 5 & -3 \\ -4 & 3 \end{bmatrix}$$

$$\|\mathbf{X}\|_\infty = 9, \quad \left\|\mathbf{X}^{-1}\right\|_\infty = \frac{8}{3}. \quad \text{This implies that,} \quad \kappa(\mathbf{X}) = 9.\frac{8}{3} = 24.$$

For any $p \times p$ non-singular matrix $\mathbf{X}$ and natural norm $\|.\|$,

$$1 = \left\|\mathbf{I}_p\right\| = \left\|\mathbf{X}\,\mathbf{X}^{-1}\right\| \leq \|\mathbf{X}\| \left\|\mathbf{X}^{-1}\right\| = \kappa(\mathbf{X}).$$

If $\kappa(\mathbf{X})$ is small (close to 1), then the matrix is said to be well-conditioned. On the other hand, if $\kappa(\mathbf{X})$ is large, that is, if it is significantly larger than 1, then it is said to be ill-conditioned.

Convex Optimization

A convex optimization problem is one of the form

$$\begin{aligned} &\text{minimize } f_0(x) \\ &\text{subject to } f_i(x) \leq b_i,\ i = 1,\ \ldots,\ m, \end{aligned} \tag{1.9}$$

where the functions $f_0,\ \ldots,\ f_m : \Re^n \to \Re$ are convex, and satisfy the inequality

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y)$$

for all $x,\ y \in \Re^n$ and all $\alpha,\ \beta \in \Re$ with $\alpha + \beta = 1,\ \alpha \geq 0,\ \beta \geq 0$. The least squares problem in Equation (1.1) is a special cases of the general convex optimization problem. Using convex optimization is, at least conceptually, very much like using least squares or linear programming. If we can formulate a problem as a convex optimization problem, then we can solve it efficiently, just as we can solve a least-squares problem efficiently. With only a bit of exaggeration, we can say that, if you formulate a practical problem as a convex optimization problem, then you have solved the original problem. There are also some important differences. Recognising a least squares problem is straightforward, but recognising a convex function can be difficult. In addition, there are many more tricks for transforming convex problems than for transforming linear programmes. Recognising convex optimization problems, or those that can be transformed to convex optimization problems, can therefore be challenging.

Regularization Methods

The primary difficulty with the discrete ill-posed problems is that they are essentially under-determined due to the cluster of smaller singular values of the matrix $\mathbf{X}$. Hence, it is necessary to incorporate further information about the desired solution in order to stabilise the problem and to single out a useful and

stable solution. This is the purpose of regularization. Although many types of additional information about the solution α is possible in principle, the dominant approach to regularization of discrete ill-posed problems is to require that the norm-2 or an appropriate semi-norm of the solution be small. An initial estimate $\Lambda\alpha$ of the solution may also be included in the side constraint. Hence, the side constraint involves minimization of the quantity

$$\Omega(\alpha) = \|\mathbf{L}\alpha\|_2 . \tag{1.10}$$

Here, the matrix $\mathbf{L}$ is typically either the identity matrix $\mathbf{I}_p$ or a $p \times n$ discrete approximation of the $(n-p)$-th derivative operator, in which case $\mathbf{L}$ is a banded matrix with full row rank. When the side constraint $\Omega(\alpha)$ is introduced, one must give up the requirement that $\mathbf{X}\alpha$ equals $\mathbf{y}$ in the linear system in Equation (1.1) and instead seek a solution that provides a fair balance between minimizing $\Omega(\alpha)$ and minimizing the residual norm $\|\mathbf{X}\alpha - \mathbf{y}\|_2$. The underlying idea is that a regularized solution with small semi-norm and a suitably small residual norm is not too far from the desired unknown solution to the unperturbed problem underlying the given problem. The same idea also applies to the least squares problem in Theorem 1.

Undoubtedly, the most common form of regularization is the one known as Tikhonov regularization. Here, the idea is to define the regularized solution α_λ as the minimizer of the weighted combination of the residual norm and the side constraint given as

$$\alpha_\lambda = \min_\alpha \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda \|\mathbf{L}\alpha\|_2^2, \tag{1.11}$$

where the regularization parameter λ, controls the weight given to minimization of the side constraint relative to minimization of the residual norm. Clearly, a large λ (equivalent to a large amount of regularization) favours a small solution semi-norm at the cost of a large residual norm, while a small λ (that is, a small amount of regularization) has the opposite effect. λ also controls the sensitivity

of the regularized solution α_λ to perturbations in α and $\mathbf{y}$, and the perturbation bound is proportional to λ^{-1}. Thus, the regularization parameter λ is an important quantity which controls the properties of the regularized solution, and λ should therefore be chosen with care. We remark that an underlying assumption for the use of Tikhonov regularization in the form of Equation (1.11) is that the errors in the right-hand side are unbiased and that their covariance matrix is proportional to the identity matrix. Besides Tikhonov regularization, there are many other regularization methods with properties that make them better suited to certain problems or certain computers.

L_2-Norm Regularization (Ridge Regression)

In the case of over-determined systems, we minimize $\|\mathbf{X}\alpha - \mathbf{y}\|_2^2$. Tikhonov regularization addresses the numerical instability of the matrix inversion and subsequently produces lower variance models. This method adds a positive constant to the diagonals of $\mathbf{X}^T\mathbf{X}$, to make the matrix non-singular (Boyd & Vandenberghe, 2004). The analytic solution then becomes

$$\alpha = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}.$$

proof

A common approach to obtain an approximate solution to a linear system is to minimize the objective function:

$$\phi(\alpha) = \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda\|\alpha\|_2^2$$

where $\lambda > 0$. The function can be restated as

$$\phi(\alpha) = (\mathbf{X}\alpha - \mathbf{y})^T(\mathbf{X}\alpha - \mathbf{y}) + \lambda\alpha^T\alpha.$$

Simplifying gives

$$\phi(\alpha) = \alpha^T\mathbf{X}^T\mathbf{X}\alpha - 2(\mathbf{X}^T\mathbf{y})^T\alpha + \mathbf{y}^T\mathbf{y} + \lambda\alpha^T\alpha.$$

Now, taking the derivative with respect to α, we obtain

$$\begin{aligned}\frac{\partial}{\partial\alpha}\phi(\alpha) &= \mathbf{X}^T\mathbf{X}\alpha+(\mathbf{X}^T\mathbf{X})^T\alpha-2\mathbf{X}^T\mathbf{y}+2\lambda\alpha\\ &= \mathbf{X}^T\mathbf{X}\alpha+\mathbf{X}^T\mathbf{X}\alpha-2\mathbf{X}^T\mathbf{y}+2\lambda\alpha\\ &= 2\mathbf{X}^T\mathbf{X}\alpha-2\mathbf{X}^T\mathbf{y}+2\lambda\alpha\\ &= 2\mathbf{X}^T(\mathbf{X}\alpha-\mathbf{y})+2\lambda\alpha.\end{aligned}$$

Setting the derivative to zero gives

$$\begin{aligned}\frac{\partial}{\partial\alpha}\phi(\alpha)=0 &\Rightarrow \mathbf{X}^T\mathbf{X}\alpha+\lambda\alpha=\mathbf{X}^T\mathbf{y}\\ &\Rightarrow (\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})\alpha=\mathbf{X}^T\mathbf{y}.\end{aligned}$$

Therefore, the minimum norm solution is given by

$$\alpha=(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}.$$

Thus,

$$\min_{\alpha}\|\mathbf{X}\alpha-\mathbf{y}\|_2^2+\lambda\|\alpha\|_2^2$$

implies that

$$\alpha=(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}. \tag{1.12}$$

This ends the proof.

L_1-Norm Regularization (LASSO Regression)

While L_2-norm regularization is an effective means of achieving numerical stability and increasing predictive performance, it does not address other problems with least squares estimate: parsimony of the model and interpretability of the coefficient values. While the size of the coefficient values is bounded, minimizing the sum of squared residual with a penalty on the L_2-norm does not encourage sparsity and the resulting models typically have non-zero values associated with all coefficients. It has been proposed that, rather than simply achieving the goal of 'shrinking' the coefficients, higher λ values for the L_2-penalty

force the coefficients to be more similar to each other in order to minimize their joint L_2-norm (Tibshirani, 1994). It will be noticed in the review of the literature that this problem arises because the L_2-norm method does not take into consideration the sparsity of the solution. In other words, the solution produced presupposes that all the independent variables in the model are significant. A recent trend has been to replace the L_2-norm with an L_1-norm. This L_1-norm regularization has many of the beneficial properties of L_2-norm regularization, but yields sparse models that are more easily interpreted (Hastie, Tibshirani & Friedman, 2001).

An additional advantage of L_1-penalties is that the models produced often outperform those produced with an L_2-penalty, when irrelevant features are present in α. This property provides an alternative motivation for the use of an L_1-penalty. It provides a regularized feature selection method, and thus can give low variance feature selection, compared to the high variance performance of typical subset selection techniques (Hastie et al., 2001). Furthermore, this does not come with a large disadvantage over subset selection methods, since it has been shown that least squares with an L_1-penalty comes as close as subset selection techniques do to an ideal subset selector (Tibshirani, 1994).

Unconstrained Formulation of the L_1-Norm Regularization

The L_1-norm regularization may be formulated as

$$\begin{aligned} g(\alpha) &= \min_{\alpha} \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 \\ &\quad \text{such that } \|\mathbf{L}\alpha\|_1 \leq t. \end{aligned}$$

This is the constrained formulation of the minimization problem. It can be shown (see Chapter Five) that the problem may also be formulated as

$$g(\alpha) = \min_{\alpha} \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda \|\mathbf{L}\alpha\|_1 . \tag{1.13}$$

It is clear that this remains an unconstrained convex optimization problem in terms of α. However, this problem is now non-differentiable when $\alpha_i = 0$ for any α_i. Thus, we cannot obtain a closed form solution for the global minimum in the same way that is done with the L_2-penalty. This drawback has led to the recent introduction of a multitude of techniques for determining the optimal parameters. Several of these algorithms directly use the unconstrained optimization problem (Perkins, Lacker & Theiler, 2003), while other techniques (Sardy, Bruce & Tseng, 1998) use equivalent constrained formulations.

Statement of the Problem

Studies into least squares optimization with an L_1- and L_2-norm regularization are well documented in the literature. Regularization of least squares solution is mainly due to the fact that the solution may not be unique in general. In particular, it is possible to obtain two least squares solutions that have a variable with different signs on the two solutions. This creates problems for interpretation of results. Recent studies on the subject focus particularly on correcting this anomaly. Many of these researchers have proposed various modifications to some of the optimization techniques to go round the problems that have been encountered in the implementation of the procedures and algorithms of these techniques.

Invariably, attempts at obtaining better approaches to the problem also end up with some inherent difficulties. Usually such difficulties are as a result of the ill-conditioning of the data matrix. For others, the weakness is as a result of a sole focus on achieving a model that produces a minimal residual norm without equal consideration for the structure of the data matrix. For example, Lee, Lee, Abbeel and Ng (2006) obtained a smoothing approximation to the L_1-norm regularization functional and later used Modified Newton's method which is an

unconstrained minimization method. They later found the convergence rate of the solution to be slow which was as a result of ill-conditioning of the data matrix. Yet another challenge is that the L_1-norm problem does not generally have a close form solution and hence difficult to obtain an analytic solution. Solutions to such a problem therefore relies on numerical approach.

An attempt at obtaining an analytic solution relies on the use of smoothing approximation that eventually becomes an L_2-norm problem. Success at obtaining a good approximation to the L_1-norm regularization thus requires a sound understanding of the L_2-norm problem. It appears that efforts at obtaining an optimal solution to the L_1-norm problem cannot be said to be over. This thesis is also concerned with the study of an L_1-norm regularization functional.

Objectives of the Study

The task of this thesis is to carry out a general study of smoothing and non-smoothing methods of Least Squares Optimization problems using tools from regularization theory and the theory of ill-posed problems. In particular, we study some methods in approximating the non-differentiable component in the loss function. Specifically, we aim to

1. carry out a general study of Tikhonov Regularization;
2. determine techniques that help to overcome the non-differentiability of the L_1-norm regularization in order to improve on the solution to the least squares problem;
3. investigate the effect of various regularization parameters on the degree of disparity in the solution; and
4. apply both smoothing and non-smoothing approximations to data fitting.

Illustrative Datasets

A number of datasets have been selected to illustrate the concepts developed in this study. These datasets are specifically chosen to enable us highlight our results using their special features. One of these datasets is the Hilbert matrix which is well known for its ill-conditioned nature and widely used as a hypothetical test data for showing accuracy of solutions of optimization techniques. We explore with other types of data that are real but are generally ill-conditioned. These are datasets that exhibit time series components and may require polynomial fitting. These involve Population Growth and Temperature Variation over time. The other datasets cover Ozone concentration, Boston Housing, Crime data and the Prostate Cancer data.

In each of these datasets, we will use our derived method in Chapter Five to obtain a model that minimizes the sum of squared residual and to determine how well the model behaves in relation to the other standard methods. In what follows, we provide detailed description of each of the datasets.

Hilbert Matrices

There are several examples of ill-conditioned matrices, but the class of matrices that are most studied and that arise in applications are those named after David Hilbert. Hilbert matrices are square matrices composed of fractional entries with the largest values located for small values of i and j. In addition, the smallest entries are located for larger values of i and j.

Definition 1

The $n \times n$ matrix $\mathbf{H}_n$, with entries $h_{(ij)} = \dfrac{1}{i+j-1}, \quad 1 \leq i \leq n, \quad 1 \leq j \leq n$ is called the Hilbert matrix of order n.

The ill-conditioning nature of the Hilbert matrices can be traced back to the approximation problem. On the interval, the functions are very nearly linearly dependent. This means that the rows of the Hilbert matrix are very linearly dependent which makes the matrix very nearly singular. In such cases, a small perturbations in the data can result in large perturbations in the answers. In the original problem, small errors in the function or rounding errors in its calculation can result in large changes in the coefficients. Furthermore, these matrices become rapidly ill-conditioned as their size increases. However, this is hard to demonstrate numerically. For example, a 12×12 Hilbert matrix, $\left\|\mathbf{H}^{-1}\right\|$ is about 10^{16}. Hence, even with a small residual it is quite possible that the approximation is not very good. Table 1 demonstrate the relationship between the condition number and the increase in the size of the Hilbert matrix $\mathbf{H}_n$.

Table 1: Condition Number of $n \times n$ Hilbert Matrix

n	Cond($\mathbf{H}_n$)
4	$1.5514e+004$
5	$4.7661e+005$
6	$1.4951e+007$
7	$4.7537e+008$
8	$1.5258e+010$
9	$4.9315e+011$
10	$1.6025e+013$
11	$5.2232e+014$
12	$1.7408e+016$

It is very difficult to compute the inverse of an $n \times n$ Hilbert matrix numerically even with implementation of OCTAVE 3.8.2. For instance, when multiplying a Hilbert matrix and its inverse together, one should expect the identity

matrix. However, the product generates a matrix with diagonal entries of ones, but many other entries are $0(\varepsilon_m)$.

For want of space, we generate the first three columns of a 12×12 Hilbert matrix, $\mathbf{H}_{12\times 3}$, as shown. It can be verified that each entry of the matrix follows Definition 1. Since rounding of the elements could generate considerable differences in the coefficients of the intended model, we will adopt throughout the work the long format of the matrix in which elements are given to 16 decimal places.

$$\begin{pmatrix}
1.0000000000000000 & 0.5000000000000000 & 0.3333333333333333 \\
0.5000000000000000 & 0.3333333333333333 & 0.2500000000000000 \\
0.3333333333333333 & 0.2500000000000000 & 0.2000000000000000 \\
0.2500000000000000 & 0.2000000000000000 & 0.1666666666666667 \\
0.2000000000000000 & 0.1666666666666667 & 0.1428571428571428 \\
0.1666666666666667 & 0.1428571428571428 & 0.1250000000000000 \\
0.1428571428571428 & 0.1250000000000000 & 0.1111111111111111 \\
0.1250000000000000 & 0.1111111111111111 & 0.1000000000000000 \\
0.1111111111111111 & 0.1000000000000000 & 0.0909090909090909 \\
0.1000000000000000 & 0.0909090909090909 & 0.0833333333333333 \\
0.0909090909090909 & 0.0833333333333333 & 0.0769230769230769 \\
0.0833333333333333 & 0.0769230769230769 & 0.0714285714285714
\end{pmatrix}.$$

Notice that by Definition 1, the $(4,\ 1)$ element, for example, is given for $i = 4$ and $j = 1$ which gives

$$h_{(4,\ 1)} = \frac{1}{4+1-1} = \frac{1}{4} = 0.25 = h_{(1,\ 4)}.$$

Generally, $h_{(i,j)} = h_{(j,i)}$ which makes the matrix essentially symmetric. In addition, all the diagonal elements $h_{(j,j)}$ are equal to $h_{(i,j)}$ for $i+j = j+j$. Thus, there are ${}^nC_2 + n - 2$ repeating elements. The two elements that are not repeated are the first and the last entries $h_{(1,1)}$ and $h_{(n,n)}$. Furthermore, the values are

quiet close in magnitude. As a result of these, the columns of the matrix appear singular especially when the size of the matrix gets large.

Example 6

We create an over-determined system by considering the first 7 columns of a 12×12 Hilbert matrix. By considering the linear system of the form $\mathbf{X}\alpha = \mathbf{y}$, $\mathbf{y}$ is chosen such that the exact solution is $\alpha = [1,\ 1,\ 1,\ 1,\ 1, 1,\ 1]^T$. Solving the linear system by the least squares approach, the approximate solution is displayed in Table 2 with the exact solution.

Table 2: Least Squares Solution

α_{exact}	$\hat{\alpha}$
1.00000000000000	1.000088594853878
1.00000000000000	0.997624278068542
1.00000000000000	1.020212173461914
1.00000000000000	0.975856781005859
1.00000000000000	1.089424133300781
1.00000000000000	0.917808532714844
1.00000000000000	1.017202377319336

From Table 2, the error in the computed solution is 0.0894241333007812 and the condition number of the coefficient matrix is $2.31648078701200e+015$ implying that the system is ill-conditioned thereby affecting the accuracy of the solution. If d is the digits of accuracy of the data and k is the power of the condition number expressed to base 10, then the number of digits expected to be lost in the computed solution is $d-k$. Hence, with the data accurate to 16 digits, the accuracy of the computed solution is given by $d-k = 16-15 = 1$ which implies that the least squares solution is accurate to only one digit. The

solution therefore cannot be a good approximation to the least squares problem. Since the system is ill-conditioned, an attempt is made to improve upon the computed solution by using Tikhonov Regularization.

A better idea can be gained by looking at the singular values of a Hilbert matrix. The computed singular values of a 12×12 Hilbert matrix are shown in Table 3. Notice that there is not a clear distinction of the number of nonzero singular values. If we use an estimate of rounding errors as a way to set a singular value to zero, then one might guess that the rank is either 10 or 11. When computed using mathematical software such as Octave, $\text{rank}(\mathbf{H}) = 11$.

Table 3: Singular Values of 12×12 Hilbert Matrix

i	σ_i
1	$1.79537205956200e+000$
2	$3.80275245955037e-001$
3	$4.47385487521811e-002$
4	$3.72231223789118e-003$
5	$2.33089089021776e-004$
6	$1.11633574832269e-005$
7	$4.08237611034721e-007$
8	$1.12286106661227e-008$
9	$2.25196455445750e-010$
10	$3.11134682183084e-012$
11	$2.64997367869193e-014$
12	$1.03133344361456e-016$

The gradual decay of the singular values is an indication that we are not dealing with a rank-deficient matrix, but with a situation that is inherently ill-conditioned. However, it is possible to create some quite complicated schemes

to try to decide between a very small number and a zero, but not all of them always work. Thus, it is definitely possible to compute the rank of a matrix theoretically, however computing the rank is impossible numerically without the risk of an occasional failure. We can use a parameter along with the singular value decomposition as a cut-off value to determine which singular values are retained and others treated as zeros of the rounding error.

Ozone Concentration Data

The data covers the level of atmospheric ozone concentration from eight daily meteorological measurements made in the Los Angeles basin in 1976. The response, referred to as ozone, is actually the logarithm of the daily maximum of the hourly-average ozone concentrations in Upland, California. It involves 330 complete cases of measurements that were made every day that year. Thus, the data covers 330 observations on a total of ten variables which are described as follows:

Ozone : Upland Maximum Ozone

VH : Vandenberg 500 mb Height

Wind : Wind Speed (mph)

Humidity : Humidity

Temp : Sandburg AFB Temperature

IBH : Inversion Base Height

DPG : Daggot Pressure Gradient

IBT : Inversion Base Temperature

Vis : Visibility (miles)

DOY : Day of the Year

For this data, we intend to determine how well the method studied provides a good fit for the Ozone concentration in terms of the nine variables.

Boston Housing Data

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. It is created by Harrison and Rubinfeld (1978) on 'Hedonic prices and the demand for clean air'. The data consists of 506 observations on 14 variables. It seeks to explain the crime rate in Boston using 13 explanatory housing variables. The description of the variables are given as follows:

CRIM: per capita crime rate by town

ZN: proportion of residential land zoned for lots over 25,000 sq.ft.

INDUS: proportion of non-retail business acres per town

CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

NOX: nitric oxides concentration (parts per 10 million)

RM: average number of rooms per dwelling

AGE: proportion of owner-occupied units built prior to 1940

DIS: weighted distances to five Boston employment centres

RAD: index of accessibility to radial highways

TAX : full-value property-tax rate per $10,000$

PTRATIO: pupil-teacher ratio by town

B: $1000(Bk-0.63)^2$ where Bk is the proportion of blacks by town

LSTAT: lower status of the population

MEDV: Median value of owner-occupied homes in $1000's$

The data has been studied in Belsley, Kuh and Welsch (1980) and in Quinlan (1993).

Data for Polynomial Fitting

The layout of data for polynomial fitting is described in the methods in Chapter Three. As will be demonstrated later, such datasets are inherently ill-

conditioned, and are therefore suitable for illustrating the features of the techniques studied in the Thesis. One of such datasets is on population growth. It covers a period of 19 years from an initial period. Figures 3 and 4 show a plot of the data with polynomial fits of various degrees. It is observed in Figure 4 that the 4th-order polynomial almost coincides with the 3rd-order polynomial. Table 4 gives the corresponding coefficient estimates for up to degree 4 for the data.

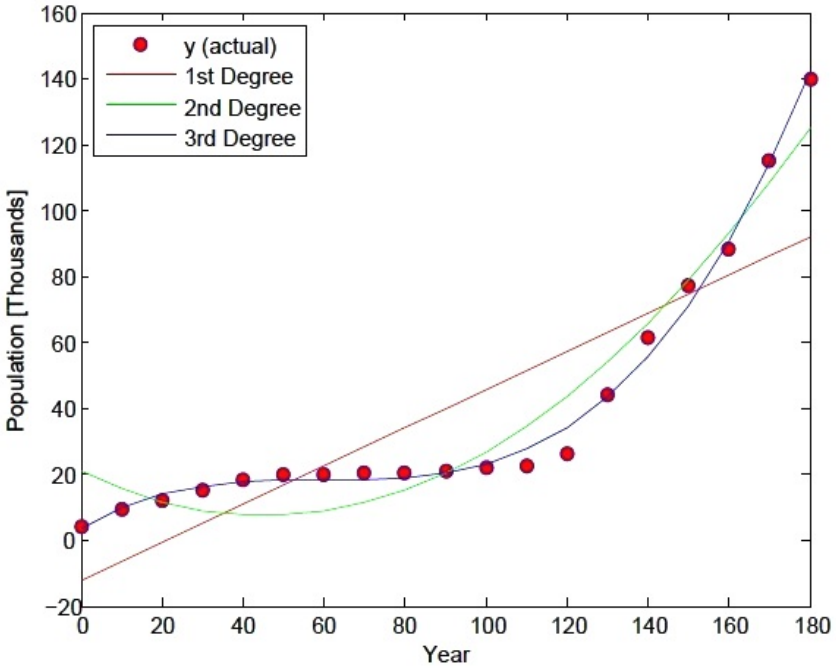

Figure 3: Graphs of Polynomial Fits up to Degree 3 of Population Data.

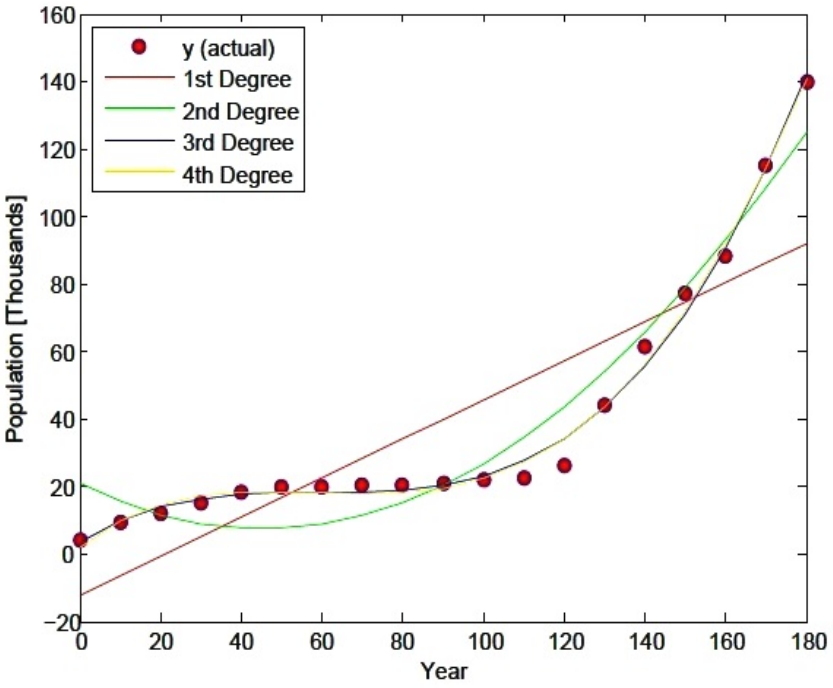

Figure 4: Graphs of Polynomial Fits up to Degree 4 of Population Data.

Table 4: Polynomial Fits of Population Data

Coefficient	Degree 1	Degree 2	Degree 3	Degree 4
α_0	0.5807	0.0065	0.0001	−0.0000
α_1	−12.5513	−0.5898	−0.0125	0.0001
α_2		20.6132	0.7431	−0.0180
α_3			3.3675	0.9497
α_4				1.9602

Table 5: Residual Norms of Polynomial Fits in Table 4

Degree	Residual Norms
1	86.324297014518891
2	41.415541335732172
3	13.846617984349306
4	13.402597620075092

The residual norms of the respective models are given in Table 5. From the table, it appears that the residuals are almost the same for degree 3 and higher-degree polynomials, an indication that the best fit may not exceed degree 3. Using the study methods we will determine the optimum degree for this data.

Since the population data contains few data points, it would be appropriate to consider another data of similar characteristics but with much more points. The other data covers annual global temperature anomalies. It is provided by Jones, Parker, Osborn and Briffa (2016) and also studied in Parker (2016). The data covers a period of 166 years from 1850 to 2015. Figure 5 shows a plot of the data with polynomial fit of order 5, 10 and 20.

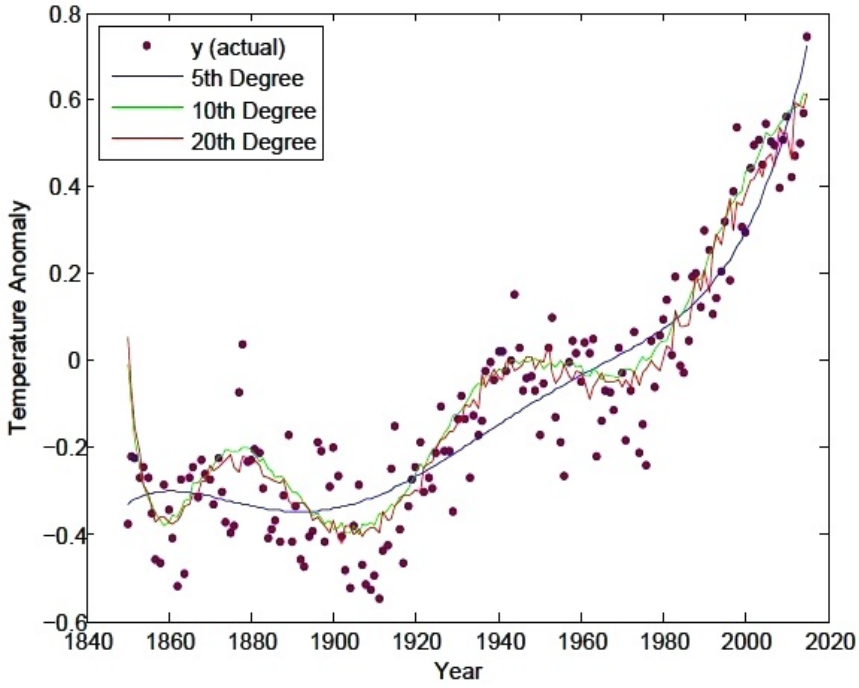

Figure 5: Graphs of Polynomial Fits of Various Degrees of Global Temperature Anomaly Data.

In Figure 5, the fitted polynomials are obtained by least squares method. The figure shows that higher degree polynomial would provide better fit of the

data. In Table 6, we have the residual norms of the fitted polynomial degrees in Figure 5.

Table 6: Residual Norms of Polynomial Fits of Global Temperature Anomaly Data for Specified Degrees

Degree	Residual Norms
5	1.479440012542864
10	1.351162363969778
20	1.329395095141014

The table shows that there are some differences in error explained by fits of degree 10 and higher. This buttresses the fact that higher degree polynomial would be required to minimize the residuals in the data. We provide a study of this in Chapter Five.

Organisation of the Thesis

This section outlines the contents within each of the six chapters of the thesis, and gives a brief description of these contents.

The Introduction is the first chapter of the study. It looks at the background to the study which brings out the history of stages of applications of least squares Optimization with an L_1-norm regularization functional and the need for more innovative efforts in the field. This is followed by the objectives of the study. It gives the Mathematical background and notations that are used in the study of Unconstrained optimization.

Chapter Two is the review of relevant literature. It discusses research made in the field of least squares optimization in relation to both L_2-norm and L_1-norm regularization functional.

In Chapter Three, we review methods used in related works being done with respect to L_1-norm. In this chapter, an ill-posed problem is used to demonstrate how a solution can be retrieved.

Chapter Four deals with the results that have emerged from the study with respect to smoothing approximations of the least squares minimization with an L_1-norm regularization functional. In this chapter, three smoothing approximations are considered which makes use of Tikhonov regularization approach.

Chapter Five also considers non-smoothing methods for solving the same problem by the use of sub-gradient. The last part of this chapter deals with applications in data fitting. We will then discuss some observations from the applications.

In Chapter Six, we summarise the various findings that have emerged from the study. Finally, we draw appropriate conclusions and make recommendations based on the results of the study.

Chapter Summary

The chapter has brought out the main motivation for the study. It explains the concept of L_1-regularized least squares and the general non-differentiable convex problem. Important concepts connected with the problem under study have been briefly introduced. These include singular value decomposition, ill-conditioning in linear systems, general regularization methods and unconstrained formulation of the L_1-norm regularization. It then specifies the statement of the problem and the objectives that will guide the study.

The chapter also describes in detail relevant datasets that will be used to illustrate the concepts developed in the study. It finally provides the organisation of the thesis which covers six chapters in all.

CHAPTER TWO

LITERATURE REVIEW

Introduction

This chapter reviews related work on several aspects of least squares optimization with L_1-norm regularization. We consider L_1-regularized optimization problems, where the L_1-norm is used to obtain sparsity of the solution. Many such problems are versions of the LASSO (Least Absolute Shrinkage and Selection Operator) introduced in Tibshirani (1996). We first review the various formulations of the regularization functional. We then consider properties of the L_1-regularized solution obtained by various methods. The review will also cover related applications of the L_1-regularized optimization. Other related works are also reviewed. It looks at methods that have been proposed in the case where the design matrix $\mathbf{X}$ is orthogonal. We then turn attention to methods that have been proposed for optimizing the LASSO that do not achieve the optimal solution. We will subsequently review other loss functions to which L_1-penalty has been applied, and some of the notable optimization methods used in these works. Also covered are methods that have been proposed for finding an appropriate value of the regularization parameter, and we find that this ties back to our discussion of optimization methods.

The literature review here are in line with the focus of this thesis, which is to find a smoothing approximation of the solution to the L_1-norm functional by the method of regularization instead of the Modified Newton's method which is usually used in the literature.

Formulation of the Regularization Functional

Elastic Net regularization involves a penalty term on the L_q-norm of the parameters, with $q \geq 0$ (and not necessarily integral). L_2-penalties correspond to the case where q is 2, while L_1-penalties correspond to the case where q is 1 (subset selection is defined as the case where q is 0). An interesting note is that L_1-penalties are the smallest Bridge penalties that yield a convex set for the constraint region (Tibshirani, 1994). This view also leads to a geometric interpretation of the sparseness properties of the LASSO (Hastie et al., 2001). Fu (1998) gives a more general form of the corresponding statistical priors associated with Bridge penalties.

An interesting observation about the sparsity of the LASSO coefficients was made in Efron, Hastie, Johnstone, and Tibshirani (2002). They state that the number of non-zero coefficients is bounded by $n-1$. Interestingly, they state that the $n-1$ coefficients for a given value of λ will not necessarily include those selected for the maximum value of λ. A related observation made by Osborne, Presnell and Turlach (2000) is that the search for t can be restricted to the range $[0, \left|\mathbf{X}^T\mathbf{y}\right|_\infty]$.

A final interesting and important property of L_1 regularization is the recent work in Ng (2004) on the effective sample complexity of using L_1 regularization compared to L_2 regularization. This work shows that the sample complexity grows at least linearly in the number of irrelevant features for many loss functions when using L_2 regularization (this includes if L_2-based preprocessing such as Principal Component Analysis was used), while the sample complexity grows only logarithmically when L_1 regularization is used.

Tibshirani (2013) stated the LASSO fit and solution in a more explicit form that satisfies the Karush-Kuhn Tucker (KKT) optimality conditions which is sta-

ted as

$$\mathbf{X}_i^T(\mathbf{y}-\mathbf{X}\hat{\alpha}) = \lambda_i\gamma_i, \tag{2.1}$$

where

$$\gamma_i = \begin{cases} \text{sign}(\alpha_i), & \alpha_i \neq 0 \\ \\ [-1,\ 1], & \alpha_i = 0. \end{cases} \qquad i = 1,\ 2,\ \ldots,\ n$$

Tibshirani defines an Equicorrelation set ε as

$$\varepsilon = \{i \in \{1,\ 2,\ \ldots,\ n\} : \left|\mathbf{X}_i^T(\mathbf{y}-\mathbf{X}\hat{\alpha})\right| = \lambda\} \tag{2.2}$$

and Equicorrelation signs $\mathbf{s} = \mathbf{X}_i^T(\mathbf{y}-\mathbf{X}\hat{\alpha})$. Equation (2.2) may therefore be written as

$$\varepsilon = \{i \in \{1,\ 2,\ \ldots,\ n\} : |\gamma_i| = 1\}.$$

Thus, when $\mathbf{y}$, $\mathbf{X}$ are standardised, ε contains the variables that have equal and maximal absolute correlation with the residual $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\alpha}$. By definition of sub-gradient, $\alpha_i = 0,\ i \notin \varepsilon$, writing the ε portion in Equation (5.14), we have

$$\mathbf{X}_\varepsilon^T(\mathbf{y}-\mathbf{X}_\varepsilon\hat{\alpha_\varepsilon}) = \lambda\mathbf{s}. \tag{2.3}$$

Noting that $\lambda\mathbf{s} \in \text{row}(\mathbf{X}_\varepsilon)$, $\lambda\mathbf{s} = \mathbf{X}_\varepsilon^T(\mathbf{X}_\varepsilon^T)^+\lambda\mathbf{s}$. Making substitutions into Equation (2.3) and simplifying, the unique LASSO fit was obtained as

$$\mathbf{X}_\varepsilon\hat{\alpha}_\varepsilon = \mathbf{X}_\varepsilon(\mathbf{X}_\varepsilon)^+\left[\mathbf{y} - (\mathbf{X}_\varepsilon^T)^+\lambda\mathbf{s}\right]. \tag{2.4}$$

Thus, any LASSO solution is of the form

$$\alpha_{-\varepsilon} = 0 \quad \text{and} \quad \alpha_\varepsilon = (\mathbf{X}_\varepsilon)^+\left[\mathbf{y} - (\mathbf{X}_\varepsilon^T)^+\lambda\mathbf{s}\right] + \mathbf{y}, \tag{2.5}$$

where $\mathbf{y} \in \text{null}(\mathbf{X}_\varepsilon)$. Since any $\mathbf{y} \in \text{null}(\mathbf{X}_\varepsilon)$ produces a LASSO solution α in Equation (2.5) provided $\hat{\alpha}$ has correct signs over non-zero coefficients, that is,

$\mathbf{s}_i = \text{sign}(\alpha_i)$, Equation (2.5) is generally written as

$$\mathbf{y} \in \text{null}(\mathbf{X}_\varepsilon) \quad \text{and} \quad \mathbf{s}_i \cdot \left\{ \alpha_\varepsilon = (\mathbf{X}_\varepsilon)^+ \left[\mathbf{y} - (\mathbf{X}_\varepsilon^T)^+ \lambda \mathbf{s} \right] + \mathbf{y} \right\} \geq 0, \quad i \in \varepsilon. \quad (2.6)$$

Following Equation (2.5), it is easy to identify a sufficient condition for uniqueness of the LASSO solution. It is observed that if $\text{null}(\mathbf{X}_\varepsilon) = \{\mathbf{0}\}$, then $\mathbf{y} = \mathbf{0}$ and the LASSO solution in Equation (2.5) is unique.

A number of algorithms have been proposed for paths of L_1-norm problems. It provides a tool for understanding the behaviour of LASSO solutions. The algorithm by Tibshirani (2013), for example, which is very similar to that of Efron, Johnstone and Tibshirani (2004), computes a LARS (Least Angle Regression Shrinkage) path described as follows:

1. It begins at $\lambda = \infty$ where the solution is trivial.

2. Then as λ decreases, it computes a solution path $\hat{\alpha}^{\text{LARS}}(\lambda)$ which is piecewise linear and continuous as a function of λ.

3. Each knot in the path corresponds to an iteration of the algorithm in which the path's linear trajectory is altered in order to satisfy the KKT conditions.

4. If $\mathbf{X}_\varepsilon^T \mathbf{X}_\varepsilon$ is singular, then the KKT conditions over all $x_i,\ i \in \varepsilon$ no longer have a unique solution. The algorithm then uses the solution with the minimum L_2-norm (as in Rosset, Zhu and Hastie (2004)). It is claimed that this provides the basis for the algorithm correctness in the general case.

A basic assumption underlying most of these algorithms (Osborne et al. (2000)) is that rank $\mathbf{X}_\varepsilon = |\varepsilon|$ throughout the path. This assumption has been identified by Tibshirani (2013) to be incorrect and can lead to errors in LASSO solutions. His algorithm described above therefore makes provision for this

error. Since many algorithms specify a search direction, it is important to make explicit the nature of the direction used in the description of algorithm provided. Equation (2.5) which computes the coefficients for x_i, $i \in \varepsilon$ may be written as

$$\alpha_\varepsilon = (\mathbf{X}_\varepsilon)^+ \left[\mathbf{y} - (\mathbf{X}_\varepsilon^T)^+ \lambda \mathbf{s}\right] = \mathbf{c} - \lambda_k \mathbf{d} \tag{2.7}$$

for the kth iterate where $\lambda = \lambda_k$. In Equation (2.7), $\mathbf{c} = (\mathbf{X}_\varepsilon)^+ \mathbf{y}$ and $\mathbf{d} = (\mathbf{X}_\varepsilon)^+ (\mathbf{X}_\varepsilon^T)^+ \mathbf{s} = (\mathbf{X}_\varepsilon^T \mathbf{X}_\varepsilon)^+ \mathbf{s}$. Equation (2.7) shows that the solution is a linear function of the regularization parameter.

Recent studies (James, Paulson, & Rusmevichientong, 2013) have sort to enhance the performance of methods for L_1-regularized least squares problems by imposing further constraints on the LASSO formulation. This has lead to what is called the Constrained LASSO. This approach augment the standard LASSO with linear equality and inequality constraints. The Constrained LASSO may be stated as

$$\min_\alpha = \|\mathbf{y} - \mathbf{X}\alpha\|_2^2 + \lambda \|\alpha\|_1 \tag{2.8}$$

$$\text{subject to } \mathbf{A}\alpha = \mathbf{b} \text{ and } \mathbf{C}\alpha \leq \mathbf{d}.$$

where $\mathbf{y} \in \Re^m$, $\mathbf{X} \in \Re^{m \times p}$ and $\alpha \in \Re^p$. As indicated, L_1-penalty enables the imposition of prior knowledge in coefficient estimates in terms of sparsity. By augmenting the LASSO, the constraints provide additional tool for prior knowledge about the data to be incorporated into the solution. A typical problem that need such tools involve data that have time series components. Such datasets appear to have monotone trend and the knowledge of this could be and has been incorporated into methods for estimating the trend. This is done by using isotonic regression (Wu, Woodroofe, & Mentz, 2001). In the formation in Equation (2.8), isotonic regression is a special case of the constrained LASSO with $\lambda = 0$. In this case, there is an ordering principle among the parameters as $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_p$. It would be deduced therefore that methods for L_1-regularized least squares problem are not expected to perform well on datasets

with monotone trend. It is thus possible for the least squares and L_2-norm regularization to perform better than L_1-norm regularization.

Properties of Solution to L_1-Norm Problem

It is well-known that optimizing parameters under an L_2-penalty is equivalent to finding the mean (and mode) of the posterior distribution of the parameters subject to Gaussian prior probabilities on the parameters (Tibshirani, 1994). It is shown (Tibshirani, 1994) that an L_1-penalty is equivalent to finding the mode (but not necessarily mean) of the posterior distribution of the parameters under a double exponential prior (whose heavy tailed nature yields further insights into the sparsity of the solutions). Finally, Roth (2004) presented a method to compute posterior predictive probabilities over the predictions made using a model that was estimated subject to an L_1-penalty, by using exponential hyperpriors and analytically integrating them out. This leads to an efficient method for estimating the variance of predictions made by the model. The solution in terms of variables contained in ε in Equation (2.2) have been identified to have important properties. It is found (Tibshirani & Taylor, 2011; Tibshirani, 2013) that for any fixed data matrix $\mathbf{X}$ and $\lambda > 0$, and for every $\mathbf{y} \in \Re^m$, the LASSO solution has what is called an Active Set A, equal to ε and therefore achieves the largest active set of any LASSO solution. Similarly, it has been established (Osborne et al., 2000; Rosset et al., 2004) that for any $\mathbf{y}$, $\mathbf{X}$ and $\lambda > 0$, there exists a LASSO solution whose set of active variables is linearly independent. This means in particular that there exists a solution whose active set A has size $|A| \leq \min(m,\ p)$. This result has been highlighted by Tibshirani (2013). A key point in the proof of this result is to note what is referred to as the support $A = \text{supp}(\hat{\alpha})$, which is the active set, and assume that $\text{rank}(A) < |A|$. Then for

some $i \in A$ and the set of signs $s = \{-1, 1\} \in \Re^{|\varepsilon|}$ defined in Equation (2.2),

$$s_i x_i = \sum_{j \in A \setminus \{i\}} a_j s_j x_j,$$

where

$$\sum_{j \in A \setminus \{i\}} a_j = 1.$$

That is, $s_i x_i$ lies in the affine span of $s_j x_j$, $j \in A \setminus \{i\}$. By this result, it is possible to obtain a smaller set of coefficients $\tilde{\alpha}$ in A such that $\mathbf{X}\tilde{\alpha} = \mathbf{X}\hat{\alpha}$ and $\|\tilde{\alpha}\|_1 = \|\hat{\alpha}\|_1$, and thus, provides a LASSO solution with fewer coefficients. If one proceeds this way, one could obtain a solution with active set that satisfies $\text{rank}(A) = |A|$. It is further pointed out that for the smallest active set, the subspace $\text{col}(\mathbf{X}_A)$ is invariant under all choices of active set A for almost every $\mathbf{y} \in \Re^m$. That is, one cannot find a solution whose active set has size less than $|A|$, as this would necessarily change the span of the active variables.

Tibshirani has described variables in a LASSO solution as either dispensable or indispensable for a given regularization parameter. A dispensable variable has a non-zero coefficient at one solution but a zero coefficient at another. On the other hand, an indispensable variable has a nonzero coefficient on all LASSO solutions. Tibshirani illustrates further relationship between linear dependence and indispensability.

Other Related Works

Under approximation methods, computing the optimal LASSO parameters is a convex optimization problem, and thus any local minimum found is guaranteed to be a global minimum. In addition, we have surveyed in this thesis several highly efficient algorithms for computing the optimal LASSO parameters. Nevertheless, several works (in prominent Machine Learning venues) have presented highly efficient but sub-optimal algorithms. Grandvalet and Canu

(1998) observed that the LASSO is equivalent to a technique called the 'adaptive ridge', that places a separate non-negative penalty on the absolute value of each coefficient (similar to a Relevance Vector Machine). This equivalence simply requires an obvious constraint on the sum of the values of these penalties. Using this equivalence, they propose to use a Fixed Point algorithm for computing the adaptive ridge, in order to compute LASSO parameters.

Beginning from an L_2-penalised solution, the Fixed Point algorithm iterates between estimating these values, and estimating the coefficient vector (similar to the Expectation Maximization algorithm). However, the authors say that the method is likely not to be globally convergent. This counter-intuitive result appears to be due to the slightly different constraints that the adaptive ridge uses, and that the constraint enforcing the equivalence with the LASSO is not properly taken advantage of during the Fixed Point iterations. The authors implementation of this approach was included in several experiments. Based on these experiments, Schmidt (2005) confirmed that this method indeed does not find the optimal solution, the models generated are too sparse for small value of the regularization parameter λ (as in Relevance Vector Machines) and not sparse enough for large values of λ, and also the method finds its sub-optimal solution highly efficiently, but there is no significant saving over some of the optimal methods.

Yongdai and Jinseog (2004) presented a highly efficient but sub-optimal approach for the LASSO problem. This approach used a gradient descent-based approach related to L_1 boosting. The motivation in that work was to avoid Quadratic Programming and approximately estimate the parameters for very large problem sizes.

Although the LASSO formulation has become very popular, L_1-norm regularization has become an even more popular topic recently in the context of classification. Here, the target variable takes one of several classes, and least squares

generally gives poor predictions compared to techniques such as Support Vector Machines, Boosting, and Logistic Regression. Presented in Tibshirani (1994) was a strategy for estimating L_1-penalized parameters of loss functions that can yield a quadratic approximation. This simply involves using a loop that uses a (weighted) LASSO solver. Roth (2004) extended the Active Set method of Osborne et al. (2000) to models (under the name 'Generalised LASSO'), focusing specifically on Logistic Regression. The 'Grafting' method of Perkins et al. (2003) and the 'Gauss-Seidel' method of Shevade and Keerthi (2003) were also presented for the case of Logistic Regression. Several techniques have also been presented exclusively for the case of Logistic Regression with an L_1-penalty, although it is clear that many of these techniques would also apply in the LASSO scenario. Zhang et al. (2002) used a strategy for Logistic Regression with an L_1-penalty called stochastic sub-gradient descent. The key idea behind this method is to 'jitter' away from the point of non-differentiability by taking a small step along a sub-gradient. The weights do not become exactly zero in this model, and are thresholded when the algorithm terminates. Genkin, Lewis and Madigan (2007) presented an approach based on cyclic coordinate descent (with a minor modification to allow re-introduction of variables currently at 0). Three different approaches based on iterative scaling for Logistic Regression with an L_1-norm loss were suggested in Goodman (2004).

Although least squares and Logistic Regression are applicable in a wide variety of scenarios, L_1-penalties have been extended to even wider array of problems. This includes Multi-layer Perceptrons (Neural Networks trained by back propagation) (Perkins et al., 2003), Support Vector Machines and Generalized Additive Models (Grandvalet & Canu, 1998), Probit Regression (Krishnapuram & Hartemink, 2005), the L_1 and the Huber loss functions (Sardy, Tseng, & Bruce, 2001). Finally, as discussed in Rosset et al. (2004), the Boosting algorithm optimizes a criteria that is approximated by an L_1-penalty on the ap-

propriate loss function.

Under orthogonal design matrices, there have been several approaches proposed for computing the LASSO estimate in the special case where $\mathbf{X}^T\mathbf{X} = \mathbf{I}$. In the Basis Pursuit Denoising literature (where Orthogonal Design matrices can be constructed), Sardy et al. (1998) introduced a method based on the Block Coordinate Relaxation (BCR) strategy. Specifically, it minimizes the objective function with respect to a block of variables, keeping the others fixed. Two methods are proposed for selecting these blocks (systematic cycling and optimal descent). Convergence of the algorithm is proved, and an empirical test with the Interior Point method of Chen, Donoho and Saunders (1999) is performed. It is shown that the BCR strategy is at least as fast as this Interior Point method, and that the BCR strategy can yield approximate solutions much more efficiently. Tibshirani also briefly discussed the orthogonal design case, defining the optimal solution and showing that the LASSO gives the same estimate as the Garotte function in this case (Tibshirani, 1994). Later, an efficient and trivial algorithm that takes advantage of this definition was presented in Osborne et al. (2000).

Next on regularization parameter estimation, a parallel issue to optimizing the LASSO parameters given a fixed value of λ is selecting a good value of the regularization parameter λ. Tibshirani (1994) proposed three methods for computing an appropriate value of λ. The first was the simple but computationally expensive cross-validation procedure (since it involves solving a large number of similar problems). The second was a computationally simple but less accurate unbiased estimate of risk (here only one problem is solved). Finally, the third method (and recommended by the author) is a generalised cross-validation scheme that is less computationally expensive than cross-validation but provides a better estimate than the unbiased estimate of risk. Zhang et al. (2002) later proposed a randomized variant of Generalised Approximate Cross Validation for regularization parameter estimation. This work also proposed to use

a strategy called 'slice modeling' to solve the similar optimization problems more effectively. Another contribution of the influential work of Osborne et al. (2000) was an extremely useful tool for hyper-parameter estimation, that also vastly increases the interpretability of the models. They observed that the coefficient values follow a piecewise-linear path as t changes. Combined with their active set method that allows 'warm-starts' from lower t values, they presented a homotopy-based algorithm that computes the LASSO coefficients for all values of t. Efron et al. (2002) later termed the phrase 'regularization path' for this idea, and showed that this method allows computation of all possible values of t using the same asymptotic complexity of solving an ordinary least squares problem. Although it is not explored in detail in this thesis, several of the other LASSO optimization methods discussed would be used for efficient computation of the regularization path.

Applications of L_1-Norm Regularization

Many of the works related to the LASSO have focused exclusively on publicly available (small) benchmark datasets. Among the more ambitious and diverse applications, Sardy et al. (1998) applied the method to detection of incoming radar signatures, Zhang et al. (2002) applied Basis Pursuit to epidemiological studies, and Zheng, Jordan, Liblit and Aiken (2004) applied Logistic Regression with an L_1-penalty for identifying features associated with program crashes. Among the most recent works, two of the areas where the LASSO is showing significant potential are the analysis of microarray and other forms of genetic data (Roth, 2004; Shevade & Keerthi, 2003), and in Natural Language Processing applications (Goodman, 2004). This data usually has an extremely large number of features and relatively few observations. Thus, sparse interpretable models are highly desirable. Another area where the author of this work sees a need for sparse regularization is Computer Vision, where computationally

expensive and redundant filter banks are currently used. Increasing the sparsity of these filter banks would be highly desirable, and could vastly increase the speed at which data can be analysed.

An interesting application of L_1 regularization is total variation denoising (Rudin, Osher & Fatemi, 1992). This has served as inspiration for the L_1 mean and covariance filtering considered by Annergren (2012). Mean and covariance filtering is an important problem in several fields of research, for example economics and biology. It is used for processing data and to discover trends. Through knowledge of the trends one can, for instance, simplify the task of identifying parametric models describing the data. In Kim and Kim (2004), they perform L_1 trend filtering. They assume that the mean is piecewise linear and they use an interior-point method of Boyd and Vandenberghe (2004), to solve the optimization problem.

In Banerjee, Ghaoui and Aspremont (2008), they do model selection based on multivariate Gaussian data, that is, they find the sparsity pattern (elements equal to zero) of the inverse covariance matrix of the data. To promote sparsity, an L_1-regularized maximum likelihood estimator is used and to obtain a convex optimization problem, the decision variable is chosen as the inverse covariance matrix. Annergren (2012) also used an L_1-regularized maximum likelihood estimator and the inverse covariance matrix as the decision variable. However, they seek to find a piecewise constant covariance matrix instead of its sparsity pattern, and they used ADMM.

Many Model Predictive Control (MPC) implementations boil down to solving a quadratic programme (Boyd & Vandenberghe, 2004) at each sampling instance of the system. This requires methods than can find and implement the optimal solution at the same rate as the sampling time. The available methods can be divided into two main groups: explicit MPC and on-line MPC. In explicit MPC, the solutions to all possible quadratic programmes are calculated off-line

and then stored in a look-up table to be used on-line. Unfortunately, the size of the table grows exponentially with respect to the time horizon, number of states and input dimensions used in the MPC (Wang & Boyd, 2010). Therefore, explicit MPC is not suitable for medium - to large scale problems. For an example of explicit MPC, see Bemporad, Morari, Dua and Pistikopoulos (2002). In on-line MPC, the quadratic programme is solved in real-time at each sampling instance. Depending on the system to be controlled and the size of the over-all quadratic programme, this may require a very fast and efficient algorithm. Three commonly used algorithms are the interior-point method, active set method and fast gradient method. For examples of an interior-point method used for MPC, see Wang and Boyd (2010) and Rao, Wright and Rawlings (1998), an active set method, see Ferreau, Bock and Diehl (2008), and a fast gradient method, see Richter, Jones and Morari (2009).

Several tricks exist for improving the speed of on-line MPC when using an interior point method. In Wang and Boyd (2010), they emphasise two already known ideas, exploitation of problem structure and warm-start of algorithm, and a new idea consisting of early termination of the algorithm. They show in a simulation study that, even though the optimal solution obtained in each sample is less accurate, the control performance remains acceptable. In addition, a Riccati recursion is commonly used in both interior-point and active set methods to efficiently solve the set of linear equations that occur when finding the optimal solution.

The basic total least squares problem and its solution by the singular value decomposition was introduced by Golub and Van Loan (1980). Van Huffel and Vandewalle (1989) considered multi-variable and non-generic cases, when the problem has no solution and generalised the algorithm of Golub and Van Loan (1980) to produce a solution in these cases. Statistical properties of the total least squares method were studied by Gleser (1981), who proved that the

method yields a consistent estimator for the true parameter value in the errors in variables setting (Fuller, 1987; Carroll, Ruppert & Stefanski, 1995). The noise assumptions that ensure consistency of the basic total least squares method imply that all elements of the data matrix are measured with equal precision, an assumption that may not be satisfied in practice.

A variation of the total least squares problem is the data least squares problem (Degroat & Dowling, 1991), where the matrix $\mathbf{X}$ is noisy and the vector $\mathbf{y}$ is exact. When the errors are row-wise independent with equal row covariance matrix (which is known up to a scaling factor), the generalised total least squares problem formulation (Van Huffel & Vandewalle, 1989) extends the consistency of the basic total least squares estimator. In addition to the work on least squares temporal difference methods as well as L_1 regularization methods, there has been some recent work on regularization and feature selection in Reinforcement Learning. For instance, Farahmand, Ghavamzadeh, Szepesvari and Mannor (2009) consider a regularized version of Temporal Difference (TD)-based policy iteration algorithms, but only specifically considered L_2 regularization. However, they focused mainly on showing how such regularization can guarantee theoretical convergence properties for policy iteration, which is mainly an orthogonal issue to the L_1 regularization.

The L_1-norm is a matrix norm that penalises the sum of maximum absolute values of each row. This regularizer encourages row sparsity, that is, it encourages the entire rows of the matrix to have zero elements. In essence, this type of regularization aims at extending the L_1 framework for learning sparse models to a setting where the goal is to learn a set of sparse models. Learning algorithms based on L_1-regularized loss functions have had a relatively long history in machine learning, covering a wide range of applications such as sparse sensing, Donoho (2004), L_1-logistic regression, Ng (2004) and structure learning of Markov networks Lee, Battle, Raina and Ng (2007). A well

known property of L_1-regularized models is their ability to recover sparse solutions. Because of this, they are suitable for applications where discovering significant features is of value and where computing features is expensive. In addition, it has been shown that in some cases, L_1 regularization can lead to sample complexity bounds that are logarithmic in the number of input dimensions, making it suitable for learning in high dimensional spaces (Ng, 2004). Turlach, Venables and Wright (2005) developed an interior point algorithm for optimizing a twice differentiable objective regularized with an L_1-norm. One of the limitations of this approach is that it requires the exact computation of the Hessian of the objective function. This might be computationally expensive for some applications both in terms of memory and time. An alternative approach was proposed by Schmidt, Murphy, Fung and Rosale (2008), who combined a gradient-descent method with independent L_∞ projections. For the special case of a linear objective function, the regularization problem can be expressed as a linear programme (Quattoni, Carreras, Collins, & Darrell, 2009). While this is feasible for small problems, it does not scale to problems with large number of variables. Duchi, Shalev-Shwartz, Singer and Chandra (2008) also proposed an L_1 projection algorithm which is a special case of the algorithm where $m = 1$. The derivation of the general case for L_1 regularization is significantly more involved as it requires reducing a set of L_∞ regularization problems tied together through a common L_1-norm to a problem that can be solved efficiently. Similar to the L_1-norm, the L_2-norm has also been proposed for sparse approximation. This norm penalises the sum of the L_2-norms of each row (Yuan & Lin, 2006; Meier, Van de Geer & Buhlmann, 2006; Simila & Tikka, 2007; Park & Hastie, 2006; Obozinski, Taskar & Jordan, 2006; Argyriou, Evgeniou & Pontil, 2007; Schmidt, Van den Berg, Friedlander & Murphy, 2009). The principle of parsimony is central to many areas of science: the simplest explanation to a given phenomenon should be preferred over more complicated ones. In the

context of machine learning, it takes the form of variable or feature selection and it is commonly used in two situations. First, to make the model or the prediction more interpretable or computationally cheaper to use, that is, even if the underlying problem is not sparse, one looks for the best sparse approximation. Second, sparsity can also be used given prior knowledge that the model should be sparse. For variable selection in lineal models, parsimony may be directly achieved by penalisation of the empirical risk or the log-likelihood by the cardinality of the support of the weight vector. However, this leads to hard combinatorial problems. In particular, when the sparse model is assumed to be well-specified, regularization of the L_1-norm is adapted to high-dimensional problems, where the number of variables to learn from may be exponential in the number of observations.

The LASSO (Tibshirani, 1996) is a popular method for regression that uses an L_1 penalty to achieve a sparse solution. In the signal processing literature, the LASSO is also known as basis pursuit (Chen, Donoho & Saunders, 1998). This idea has been broadly applied, for example, to generalised linear models (Tibshirani, 1996) and Cox's proportional hazard models for survival data (Tibshirani, 1997). There has been an enormous amount of research activity devoted to related regularization methods, that is, the grouped LASSO by Yuan and Lin (2007); Meier, Van De Geer and Buhlmann (2008), where variables are included or excluded in groups; the Dantzig selector by Candes and Tao (2007), on a slightly modified version of the LASSO; the Elastic Net by Zou and Hastie (2005) for correlated variables, which uses a penalty that is part L_1 and part L_2; L_1 regularization paths for generalised linear models by Park and Hastie (2007); regularization paths for the support-vector machine by Hastie, Rosset, Tibshirani and Zhu (2004); the graphical LASSO by Friedman, Hastie, Hoefling and Tibshirani (2008) for sparse covariance estimation and undirected graphs.

Efron et al. (2004) developed an efficient algorithm for computing the en-

tire regularization path for the LASSO. Their algorithm exploits the fact that the coefficient profiles are piecewise linear, which leads to an algorithm with the same computational cost as the full least squares fit on the data (see also Osborne et al. (2000)). Rosset and Zhu (2007) characterise the class of problems where piecewise-linearity exists - both the loss function and the penalty have to be quadratic or piecewise linear. Van der Kooij (2007) independently used coordinate descent for solving Elastic-Net penalised regression models. Recent rediscoveries include Friedman, Hastie, Hoefling and Tibshirani (2007) and Wu and Lange (2008a). The first paper recognised the value of solving the problem along an entire path of values for the regularization parameters, using the current estimates as warm starts. This strategy turns out to be remarkably efficient for this problem. Several other researchers have also re-discovered coordinate descent, many for solving the same problems notably Shevade and Keerthi (2003), Krishnapuram and Hartemink (2005), Genkin et al. (2007) and Wu, Chen, Hastie, Sobel and Lange (2009).

Friedman, Hastie and Tibshirani (2010) extended the work of Friedman et al. (2007) and developed fast algorithms for fitting generalised linear models with Elastic-Net penalties. In particular, their models include regression, two-class logistic regression, and multinomial regression problems. Their algorithms can work on very large datasets, and can take advantage of sparsity in the feature set.

Adding an L_1-norm constraint or an L_1-norm regularization term to an optimization problem, a sparse solution can be achieved in some applications. A sparse solution usually benefits us in some aspects: good interpretation, Efron et al. (2004) and memory savings. The LASSO, Efron et al. (2004), a representative L_1- regularized least squares problem, has attracted more and more attentions from the field of artificial intelligence. It has a wide range of applications, such as image deblurring, Beck and Teboulle (2009), sparse coding, Lee

et al. (2006), curve-fitting and classification, Bishop (2006).

In these applications, how to efficiently solve the L_1-regularized least squares problem becomes a critical issue. Most existing optimization methods for the L_1-regularized least squares problem can be broadly classified into three categories. First, some algorithms are designed by transforming L_1-regularized least squares as a constrained quadratic programming problem. This is achieved by either introducing an auxiliary variable, or splitting the variable into the positive and negative parts. Representative algorithms include interior method by Kim, Koh, Lustig, Boyd and Gorinevsky (2007) and Figueiredo, Nowak and Wright (2007).

However, these methods double the variable size, making the optimization more costly. Second, several algorithms are developed in the fixed-point-type framework: A gradient descent operation is first done, and then a soft-thresholding operation is performed. The most representative algorithms is the Fast Iterative Shrinkage thresholding Algorithm (FISTA) by Beck and Teboulle (2009).

One other fixed-point-type algorithms is the Forward-Backward Splitting (FOBOS) by Duchi et al. (2008), etc. However, these algorithms are first-order methods, not utilising the second-order information. It is a further development based on Least Angle Regression (LARS) by Efron et al. (2004) and Feature-Sign (FS) search algorithm by Lee et al. (2007).

Lagrangian relaxation and duality have been effective tools for solving large-scale convex optimization problems and for systematically providing lower bounds on the optimal value of non-convex (continuous and discrete) optimization problems. Sub-gradient methods have played a key role in this framework providing computationally efficient means to obtain near-optimal dual solutions and bounds on the optimal value of the original optimization problem. Most remarkably, in networking applications, over the last few years,

sub-gradient methods have been used with great success in developing decentralised cross-layer resource allocation mechanisms (see Low & Lapsley, 1999); Srikant (2004) for more on this subject). The sub-gradient methods for solving dual problems have been extensively studied starting with Despite widespread use of the sub-gradient methods for solving dual (non-differentiable) problems, there are some aspects of sub-gradient methods that have not been fully studied. In particular, practical applications, the main interest is in solving the primal problem. In this case, the question arises whether we can use the sub-gradient method in dual space and exploit the sub-gradient information to produce primal near-feasible and near-optimal solutions.

Application of the LASSO to logistic regression was proposed in Tibshirani (1996); coordinate descent methods for logistic, multinomial, and Poisson regression were developed in Friedman et al. (2007), Friedman et al. (2010), Wu and Lange (2008a), and Wu et al. (2009).

In the past two decades, regularization methods based on the L_1-norm, have become immensely popular. This led to the question whether L_1-based techniques should replace the simpler, faster and better known L_2-based alternatives as the default approach to regularization techniques. The tremendous advantages of L_1-based techniques are not in doubt. However, such techniques also have their limitations. This thesis explores advantages and disadvantages compared to L_2-based techniques using several practical case studies. Taking into account the considerable added hardship in calculating solutions of the resulting computational problems, L_1-based techniques must offer substantial advantages to be worthwhile.

Ill-posed problems typically require some regularization in order to compute a credible approximate solution in a stable, well-defined manner. In this thesis, we consider such problems where the objective function is convex. Convex non-differentiable, also known as convex non-smooth optimization looks at

problems where the functions involved are not continuously differentiable. The gradient does not exist, implying that the function may have corner points and thus cannot be approximated locally by a tangent hyperplane or by a quadratic approximation. Directional derivatives still exist because of the convexity property.

Chapter Summary

The view of the literature has focused on obtaining the right formulation of the L_1-norm regularization functional that produces an optimal solution. The aim has been to obtain a sparse solution that is better than the Tikhonov regularization in particular. A number of important techniques have been employed in determining suitable methods. These include the sub-gradient methods and truncated Newton's interior point method. The methods also have their accompanying algorithms. The literature also shows widespread applications of the L_1-norm regularization. This thesis will also consider regularization methods that ensure sparsity. It will make use of the Sub-gradient and Singular Value Decomposition as some of the main techniques in the study of a LASSO solution. The focus of application will be data fitting.

CHAPTER THREE

RESEARCH METHODS

Introduction

Regularization methods are often used to obtain a stable and smooth solutions to ill-posed problems. To obtain a meaningful solution, it is necessary to incorporate some additional qualitative information about the desired solution α. Such additional information is then used as a side constraint to control the smoothness of the solution. The first part of this chapter reviews the least squares minimization with an L_2-norm regularization functional and the second part is about a non-smoothing method known as the Sub-gradient method for minimizing a least squares objective function.

Tikhonov Regularization (Ridge Regression)

When $\mathbf{X}$ is rank-deficient or very nearly singular, standard algorithms often give solutions that vary rapidly, with large positive and negative values. To stabilise the computation, we add a term that penalises the large components and thereby reduces them as follows

$$\phi_\lambda(\alpha) = \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda \|\mathbf{L}\alpha\|_2^2, \tag{3.1}$$

where λ is the regularization parameter which controls the weight given to the minimization of the square of the constraint $\|\mathbf{L}\alpha\|_2$, relative to the minimization of the square of the residual norm $\|\mathbf{X}\alpha - \mathbf{y}\|_2$. The matrix $\mathbf{L}$ is the $(n-p)\times n$ discrete approximation of the pth derivative operator. A large λ favours a small solution norm at the expense of a large residual norm, while a small λ favours a large solution norm at the expense of a small residual norm.

The minimizing solution α_λ is given by the non-singular linear system as

$$\alpha_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{L})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3.2}$$

For $p = 0$, $\mathbf{L} = \mathbf{I}$, is the $n \times n$ identity matrix which is known as Order Zero Regularization. Thus, the regularized solution in this case is given as

$$\alpha_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3.3}$$

Regularization by Equation (3.3) dampens components that are large in magnitude, but it may not inhibit components that oscillate with moderate amplitudes. If those components are undesirable, then a stronger regularization is needed. Thus, we introduce another penalty term that is large for rapid changes in the solution. For this we consider $p = 1$, and $\mathbf{L} = (n-1) \times n$, discrete approximation of the first derivative operator is known as Order One Regularization.

Thus, the regularized solution is given as

$$\alpha_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{D}_1^T\mathbf{D}_1)^{-1}\mathbf{X}^T\mathbf{y}, \tag{3.4}$$

where $\mathbf{D}_1$ is the $(n-1) \times n$ matrix with elements defined as

$$(\mathbf{D}_1)_{ij} = \begin{cases} 1, & \text{if } i = j \\ -1, & \text{if } i = j-1 \text{ or } j = i+1 \\ 0, & \text{otherwise.} \end{cases}$$

That is,

$$\mathbf{D}_1 = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & 1 & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{pmatrix}.$$

A stronger regularization is based on another minimization of the form where $p = 2$, which implies $L = (n-2) \times n$ is the discrete approximation of the second derivative operator which is known as Order Two Regularization.

Thus, the regularized solution is given as

$$\alpha_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{D}_2^T\mathbf{D}_2)^{-1}\mathbf{X}^T\mathbf{y}, \tag{3.5}$$

where $\mathbf{D}_2$ is the $(n-2)\times n$ matrix with elements defined as

$$(\mathbf{D}_2)_{ij} = \begin{cases} -2, & \text{if } i = j \\ 1, & \text{if } i = j+1 \text{ or } j = i-1 \\ 0, & \text{otherwise.} \end{cases}$$

That is,

$$\mathbf{D}_2 = \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \vdots \\ 0 & 1 & -2 & \ddots & 0 \\ \vdots & 0 & \ddots & \ddots & 1 \\ 0 & 0 & 0 & 1 & -2 \end{pmatrix}.$$

Next, we consider the computation of the generalised solution to the linear system of the form

$$\mathbf{X}\alpha = \mathbf{y}. \tag{3.6}$$

Computing Generalised Solutions

To avoid the non-existence and non-uniqueness issues in solving Equation (3.6), we need to compute the generalised solution α_0^* of the system. By the well-known rules concerning solution of a linear system and the rank of a matrix, every system has a generalised solution for all $\mathbf{X}$ and $\mathbf{y}$.

Theorem 5

A unique generalised solution for the over-determined system exists for all $\mathbf{X}$ and $\mathbf{y}$ from Equation (3.6) of the form

$$\alpha_0^* = \sum_{i=1}^{p} \frac{\mathbf{U}_i^T \mathbf{y}}{\sigma_i} \mathbf{V}_i \tag{3.7}$$

where $\sigma_1, \sigma_2, \ldots, \sigma_p$ are all the nonzero singular values, with $\mathbf{U}_i$ and $\mathbf{V}_i$ the corresponding right and left singular vectors of $\mathbf{X}$.

Proof

We consider singular value decomposition to solve linear systems, in the exact or the least squares sense. Assume that the singular values are ordered such that $\sigma_1, \sigma_2, \ldots, \sigma_p$. Furthermore, we will also take that $m \geq p$ and also assume that $\sigma_i > 0, \forall i = 1, 2, \cdots, p$; that is rank$(\mathbf{X}) = p$. Then, we develop the generalised solution by using a solution of the least squares approximation, $\mathbf{X}^T\mathbf{X}\alpha_0 = \mathbf{X}^T\mathbf{y}$. Suppose we write

$$\mathbf{y} = \sum_{i=1}^{m} \mathbf{y}_i \mathbf{U}_i$$

where $\mathbf{U}_i$ are the orthonormal eigenvectors of $\mathbf{X}^T\mathbf{X}$. Then, because of the orthonormality of the $\mathbf{U}_i$'s then by a simple substitution $\mathbf{y}_i = \mathbf{U}_i^T\mathbf{y}$. Expanding the result gives

$$\alpha_0 = \sum_{i=1}^{p} \mathbf{a}_i \mathbf{V}_i.$$

Thus, from substitution,

$$\mathbf{X}^T\mathbf{X}\alpha_0 = \mathbf{X}^T\mathbf{X}\sum_{i=1}^{p} \mathbf{a}_i\mathbf{V}_i = \sum_{i=1}^{p} \mathbf{a}_i\sigma_i^2\mathbf{V}_i.$$

In addition,

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\sum_{i=1}^{m} \mathbf{y}_i\mathbf{U}_i = \sum_{i=1}^{p} \mathbf{y}_i\sigma_i\mathbf{V}_i,$$

where the last step comes from $\mathbf{X}^T\mathbf{V}_i = 0$, for $i = p+1,\ p+2, \ldots, m$. It follows then from comparing the expansions for $\mathbf{X}^T\mathbf{y}$ and $\mathbf{X}^T\mathbf{X}\alpha_0$, that

$$\mathbf{a}_i = \frac{\mathbf{U}_i^T\mathbf{y}}{\sigma_i}$$

and therefore we have

$$\alpha_0 = \sum_{i=1}^{p} \frac{\mathbf{U}_i^T\mathbf{y}}{\sigma_i}\mathbf{V}_i.$$

This ends the proof.

In this case, we can change the problem to finding the least squares solution of minimum norm. Thus, we have the following equality

$$\|\alpha_0\|^2 = \sum_{i=1}^{p} \mathbf{a}_i^2$$

in which we obtain the smallest norm by setting the in-determinant coefficients to zero. Therefore, the minimum norm least squares solution is given by Equation (3.7).

Definition 2

Singular value decomposition with damping is the method where a good choice of λ can get a relatively smooth solution that is still a reasonably good approximation to the true solution.

We show that the solution corresponding with the singular value decomposition of Equation (3.3) can be expressed in component form as

$$\alpha_\lambda = \sum_{i=1}^{p} \frac{\sigma_i}{\sigma_i^2 + \lambda} \mathbf{U}_i^T \mathbf{y} \mathbf{V}_i.$$

Proof

Using the singular value decomposition of the matrix $\mathbf{X} = \mathbf{USV}^T$ and then substituting into Equation (3.3) gives

$$\begin{aligned}
\alpha_\lambda &= \left[(\mathbf{USV}^T)^T(\mathbf{USV}^T) + \lambda \mathbf{I}\right]^{-1} (\mathbf{USV}^T)^T \mathbf{y} \\
&= \left(\mathbf{VSU}^T \mathbf{USV}^T + \lambda \mathbf{I}\right)^{-1} \mathbf{VSU}^T \mathbf{y} \\
&= \left(\mathbf{VS}^2 \mathbf{V}^T + \lambda \mathbf{VIV}^T\right)^{-1} \mathbf{VSU}^T \mathbf{y} \\
&= \left[\mathbf{V}(\mathbf{S}^2 + \lambda \mathbf{I})\mathbf{V}^T\right]^{-1} \mathbf{VSU}^T \mathbf{y} \\
&= \mathbf{V}(\mathbf{S}^2 + \lambda \mathbf{I})^{-1} \mathbf{V}^T \mathbf{VSU}^T \mathbf{y} \\
&= \mathbf{V}(\mathbf{S}^2 + \lambda \mathbf{I})^{-1} \mathbf{SU}^T \mathbf{y},
\end{aligned}$$

where $(\mathbf{S}^2 + \lambda \mathbf{I})^{-1}$ is the diagonal matrix given by

$$(\mathbf{S}^2 + \lambda \mathbf{I})^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2 + \lambda} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2 + \lambda} & 0 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2 + \lambda} \end{pmatrix}.$$

Therefore, in component form, we have

$$\alpha_\lambda = \sum_{i=1}^{p} \frac{\sigma_i}{\sigma_i^2 + \lambda} \mathbf{U}_i^T \mathbf{y} \mathbf{V}_i.$$

This ends the proof.

Singular Value Decomposition of $\mathbf{X}$ in component form can be written as

$$\alpha_{\text{reg}} = \sum_{i=1}^{p} f_i \frac{\mathbf{U}_i^T \mathbf{y}}{\sigma_i} \mathbf{V}_i, \tag{3.8}$$

where f_i is the filter factors and is given by

$$f_i = \frac{\sigma_i^2}{\lambda + \sigma_i^2}.$$

The effect of the addition in Equation (3.3) is to dampen the contributions of the terms involving smaller singular values, so that instead of cutting them off altogether, they are modified so that they reduce their impact. Therefore, if

$$\lambda << \sigma_i^2, \ \frac{\sigma_i^2}{\lambda + \sigma_i^2} \approx 1$$

which indicates that the filter factors has no effect on the solution. Thus, Equation (3.2) is without regularization.

On the other hand, if

$$\lambda >> \sigma_i^2, \ \frac{\sigma_i^2}{\lambda + \sigma_i^2} \approx \frac{\sigma_i^2}{\lambda}.$$

We note that $\frac{\sigma_i^2}{\lambda} \to 0$. This indicates that the filter factors has effect on the solution. In this case, it reduces the effect of the smaller singular values. Thus, Equation (3.2) gives the solution with regularization.

Analysis of Error Produced by Tikhonov Regularization Method

We examine how the behaviour of the right hand side vector in Equation (3.6) affects the error in the regularized solutions. For this purpose, we examine

the error in the solution due to the regularization process itself. We define

$$\gamma_i = \frac{\mathbf{U}_i^T \mathbf{y}}{\sigma_i},$$

as a measure of the energy of $\mathbf{y}$ in a singular subspace of $\mathbf{X}$ relative to the power of that subspace. We show that the space of the function γ is the determining factor in regularization error. We also define

$$\gamma_i = \max_{1<i<p} \gamma_i.$$

The error due to regularization is given in the following theorem.

Theorem 6

Let α and α_λ be the exact and the Tikhonov regularized solutions given by Equation (3.7) and Equation (3.8), respectively. The following can be shown to hold:

$$\|\alpha - \alpha_\lambda\|_2 = \left(\sum_{i=1}^{p} \left(\frac{\sigma_i^2}{\lambda + \sigma_i^2} \gamma_i \right)^2 \right)^{\frac{1}{2}} = \begin{cases} \leq \sqrt{p} \, \dfrac{\sigma_p^2}{\lambda + \sigma_p^2} \gamma_p \\ \geq \dfrac{\sigma_p^2}{\lambda + \sigma_p^2} \gamma_p \end{cases}, \tag{3.9}$$

where

$$\frac{\sigma_p^2}{\lambda + \sigma_p^2} \gamma_p = \max_{1<i<p} \frac{\sigma_i^2}{\lambda + \sigma_i^2} \gamma_i$$

and

$$\frac{\|\alpha - \alpha_\lambda\|_2}{\|\alpha\|_2} = \begin{cases} \leq \sqrt{p} \, \dfrac{\sigma_p^2}{\lambda + \sigma_p^2} \dfrac{\gamma_p}{\gamma_l} \\ \geq \dfrac{\sigma_p^2}{\lambda + \sigma_p^2} \dfrac{\gamma_p}{\gamma_l} \end{cases}, \tag{3.10}$$

where $l \leq p, \quad \gamma_p \leq \gamma_l$.

Proof

Rewriting Equation (3.7), the norm of the true solution α is

$$\|\alpha\|_2 = \left\|\mathbf{V}_p[\gamma_1, \cdots, \gamma_p]^T\right\|_2 = \begin{cases} \leq \sqrt{p}\,(\max_{1<i<p}\gamma_i) \\ \geq \max_{1<i<p}\gamma_i \end{cases}, \tag{3.11}$$

where $\mathbf{V}_p$ is the section of $\mathbf{V}$ consisting of p columns. In the over-determined case, $\mathbf{V}_p = \mathbf{V}$. Using Equation (3.8) and Equation (3.11),

$$\|\alpha - \alpha_\lambda\|_2 = \left\|\mathbf{V}_p\left[\frac{\sigma_1^2}{\lambda+\sigma_1^2}\gamma_1, \cdots, \frac{\sigma_p^2}{\lambda+\sigma_p^2}\gamma_p\right]^T\right\|_2 = \begin{cases} \leq \sqrt{p}\,\dfrac{\sigma_p^2}{\lambda+\sigma_p^2}\gamma_p \\ \geq \dfrac{\sigma_p^2}{\lambda+\sigma_p^2}\gamma_p \end{cases},$$

which give Equation (3.9). Combining Equation (3.9) and Equation (3.11) gives the desired result in Equation (3.10). This ends the proof.

Thus regularization produces a disproportionately large error for a component of the data $\mathbf{y}$ in the ill-conditioned subspaces, compared to an equal sized component in the well-conditioned subspaces. The more energy of $\mathbf{y}$ that shifts to the ill-conditioned subspaces, the larger the error due to regularization. This is independent of how close to the optimum the regularization parameter may be.

Convex Optimality Conditions

An important class of optimization problems involves convex cost functions and convex constraints. A set $C \subseteq \Re^p$ is convex if for all $\alpha, \alpha' \in C$ and all scalars, $s \in [0,\ 1]$, all vectors of the form $\alpha(s) = s\alpha + (1-s)\alpha'$ also belong to C. A function $f : \Re^p \to \Re$ is convex means that for any two vectors α, α' in the domain of f and any scalar $s \in (0,\ 1)$, we have

$$f(\alpha(s)) = f(s\alpha + (1-s)\alpha') \leq sf(\alpha) + (1-s)f(\alpha'). \tag{3.12}$$

In geometric terms, this inequality implies that the chord joining the $f(\alpha)$ and $f(\alpha')$ lies above the graph of f, as illustrated in Figure 6 (a). This inequality guarantees that a convex function cannot have any local minima that are not also globally minimal, as illustrated in Figure 6 (b).

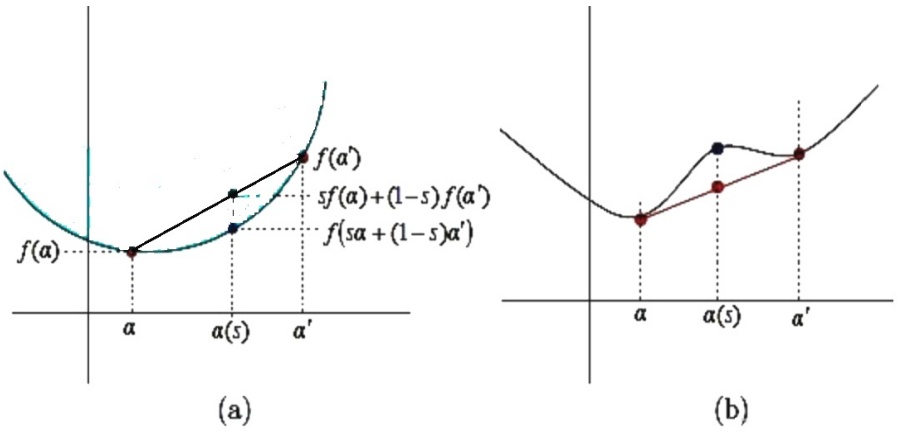

Figure 6: Graphs Showing (a) Convex Function and (b) Non-Convex Function.

In Figure 6 (a) for a convex function, the line $sf(\alpha)+(1-s)f(\alpha)$ always lies above the function value $f(s\alpha+(1-s)\alpha)$. Figure 6 (b) is a non-convex function that violates the inequality in Equation (3.12). Without convexity, there may be local minima that are not globally minima, as shown by the point α'.

Optimality for Differentiable Problems

Consider the constrained optimization problem

$$\underset{\alpha\in\Re^p}{\text{minimize}}\, f(\alpha) \qquad \text{such that} \quad \alpha\in\mathbf{C}, \tag{3.13}$$

where $f:\Re^p\to\Re$ is a convex objective function to be minimized, and $\mathbf{C}\subset\Re^p$ is a convex constraint set. When the cost function f is differentiable, then a necessary and sufficient condition for a vector $\alpha^*\in\mathbf{C}$ to be a global optimum is that

$$\langle\nabla f(\alpha^*),\alpha-\alpha^*\rangle\geq 0 \tag{3.14}$$

for all $\alpha \in C$. The sufficiency of this condition is easy to see; for any $\alpha \in C$, we have

$$f(\alpha) \geq f(\alpha^*) + \langle \nabla f(\alpha^*), \alpha - \alpha^* \rangle \geq f(\alpha^*), \tag{3.15}$$

where the first inequality follows from the convexity of f, and the second inequality follows from the optimality condition in Equation (3.14). As a special case, when $C = \Re^p$ so that the problem in Equation (3.13) is actually unconstrained, then the first-order condition in Equation (3.14) reduces to the classical zero-gradient condition $\nabla f(\alpha^*) = 0$. Frequently, it is the case that the constraint set C can be described in terms of the sub-level sets of some convex constraint functions. For any convex function $g : \Re^p \to \Re$, it follows from the definition in Equation (3.12) that the sub-level set $\{\alpha \in \Re^p | g(\alpha) \leq 0\}$ is a convex set. On this basis, the convex optimization problem

$$\underset{\alpha \in \Re^p}{\text{minimize}} \; f(\alpha) \quad \text{such that} \quad g_j(\alpha) \leq 0 \quad \text{for } j = 1, \ldots, m, \tag{3.16}$$

where g_j, $j = 1, \ldots, m$ are convex functions that express constraints to be satisfied, is an instance of the general programme in Equation (3.13). We let f^* denote the optimal value of the optimization problem in Equation (3.16). An important function associated with the problem in Equation (3.16) is the Lagrangian $L : \Re^p \times \Re^m_+ \to \Re$, defined by

$$L(\alpha; \lambda) = f(\alpha) + \sum_{j=1}^{m} \lambda_j g_j(\alpha). \tag{3.17}$$

The non-negative weights $\lambda \geq 0$ are known as Lagrange multipliers; the purpose of the multiplier λ_j is to impose a penalty whenever the constraint $g_j(\alpha) \leq 0$ is violated. Indeed, if we allow the multipliers to be chosen optimally, then we recover the original programme in Equation (3.16), since

$$\sup_{\lambda \geq 0} L(\alpha; \lambda) = \begin{cases} f(\alpha) & \text{if } g_j(\alpha) \leq 0 \\ +\infty & \text{otherwise.} \end{cases} \tag{3.18}$$

and thus $f^* = \inf_{\alpha \in \Re^p} \sup_{\lambda \geq 0} L(\alpha; \lambda)$. For convex programmes, the Lagrangian allows for the constrained problem in Equation (3.16) to be solved by reduction to an equivalent unconstrained problem. More specifically, under some technical conditions on f and $\{g_j\}$, the theory of Lagrange duality guarantees that there exists an optimal vector $\lambda^* \geq 0$ of Lagrange multipliers such that $f^* = \min_{\alpha \in \Re^p} L(\alpha; \lambda^*)$. As a result, any optimum α^* of the problem in Equation (3.16), in addition to satisfying the feasibility constraints $g_j(\alpha^*) \leq 0$, must also be a zero-gradient point of the Lagrangian, and hence satisfy the equation

$$\nabla_\alpha L(\alpha^*; \lambda^*) = \nabla f(\alpha^*) + \sum_{j=1}^{m} \lambda_j^* \nabla g_j(\alpha^*) = 0. \tag{3.19}$$

When there is only a single constraint function g, this condition reduces to $\nabla f(\alpha^*) = -\lambda^* \nabla g(\alpha^*)$, and has an intuitive geometric interpretation, as shown in Figure 5. In particular, at the optimal solution α^*, the normal vector $\nabla f(\alpha^*)$ to the contour line of f points in the opposite direction to the normal vector to the constraint curve $g(\alpha) = 0$. Equivalently, the normal vector to the contour f lies at right angles to the tangent vector of the constraint. Consequently, if we start at the optimum α^* and travel along the tangent at $g(\alpha) = 0$, we cannot decrease the value of $f(\alpha)$ up to first order.

In general, the Karush-Kuhn-Tucker (KKT) conditions relate the optimal Lagrange multiplier vector $\lambda^* \geq 0$, also known as the dual vector, to the optimal primal vector $\alpha^* \in \Re^p$.

Lagrangian Condition

The pair (α^*, λ^*) satisfies the condition in Equation (3.19). These KKT conditions are necessary and sufficient for α^* to be a global optimum whenever the optimization problem satisfies a regularity condition known as strong duality. The complementary slackness condition asserts that the multiplier λ_j^* must be zero if the constraint $g_j(\alpha) \leq 0$ is inactive at the optimum, that is, if

$g_j(\alpha^*) < 0$.

Consequently, under complementary slackness, the Lagrangian gradient condition in Equation (3.19) guarantees that the normal vector $\nabla f(\alpha^*)$ lies in the positive linear span of the gradient vectors $\{\nabla g_j(\alpha^*) | \lambda_j^* > 0\}$.

Non-differentiable Functions and Sub-gradients

In practice, many optimization problems arising in statistics involve convex but non-differentiable cost functions. For instance, the L_1-norm $g(\alpha) = \sum_{j=1}^{p} |\alpha_j|$ is a convex function, but it fails to be differentiable at any point where at least one coordinate α_j is equal to zero. For such problems, the optimality conditions that have been discussed, in particular the first-order condition in Equation (3.14) and the Lagrangian condition in Equation (3.19), are not directly applicable, since they involve gradients of the cost and constraint functions.

Nonetheless, for convex functions, there is a natural generalisation of the notion of gradient that allows for a more general optimality theory. A basic property of differentiable convex functions is that the first-order tangent approximation always provides a lower bound. The notion of sub-gradient is based on a natural generalisation of this idea. In particular, given a convex function $f : \Re^p \to \Re$, a vector $\mathbf{z} \in \Re^p$ is said to be a sub-gradient of f at α if

$$f(\alpha') \geq f(\alpha) + \langle \mathbf{z},\ \alpha' - \alpha \rangle \quad \text{for all } \alpha' \in \Re^p. \tag{3.20}$$

In geometric terms, the sub-gradient vector $\mathbf{z}$ is the normal to a (non-vertical) hyperplane that supports the epigraph of f. The set of all sub-gradients of f at α is called the sub-differential, denoted by $\partial f(\alpha)$. Whenever f is differentiable at α, then the sub-differential reduces to a single vector namely, $\partial f(\alpha) = \{\nabla f(\alpha)\}$. At points of non-differentiability, the sub-differential is a convex set containing all possible sub-gradients.

For example, for the absolute value function $f(\alpha) = |\alpha|$, we have

$$\partial f(\alpha) = \begin{cases} \{+1\} & \text{if } \alpha > 0 \\ \{-1\} & \text{if } \alpha < 0 \\ [-1,\ +1] & \text{if } \alpha = 0. \end{cases} \tag{3.21}$$

We sometimes write $\mathbf{z} \in \text{sign}(\alpha)$ to mean that $\mathbf{z}$ belongs to sub-differential of the absolute value function at α.

Figure 7 shows a function $f : \Re \to \Re$ and some examples of sub-gradients at the two points x_1 and x_2.

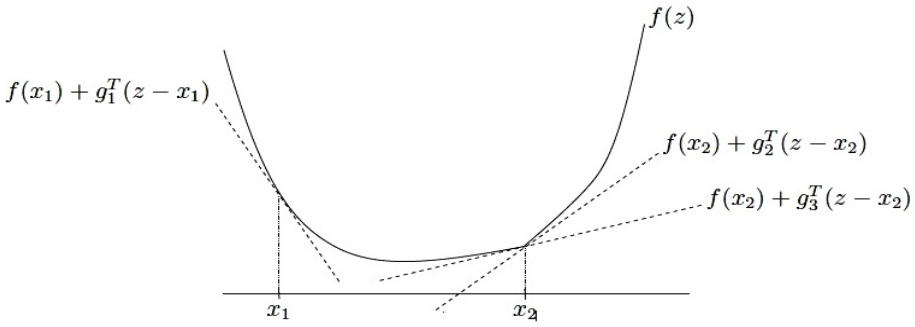

Figure 7: Sub-gradient of f at x_1 and x_2.

At the point x_1, the function is differentiable and hence there is only one sub-gradient namely, $f'(x_1)$. At the point x_2, it is not differentiable, and there are multiple sub-gradients; each one specifies a tangent plane that provides a lower bound on f. From the convex optimization problem in Equation (3.16), this is very useful. Assume that one or more of the functions $\{f,\ g_j\}$ are convex but non-differentiable, then in this case, the zero-gradient Lagrangian condition in Equation (3.19) no longer makes sense. Nonetheless, again under mild conditions on the functions, the generalised KKT theory can still be applied using

the modified condition

$$0 \in \partial f(\alpha^*) + \sum_{j=1}^{m} \lambda_j^* \partial g_j(\alpha^*), \tag{3.22}$$

in which we replace the gradients in the KKT condition in Equation (3.19) with sub-differentials. Since the sub-differential is a set, Equation (3.22) means that the all-zeros vector belongs to the sum of the sub-differentials.

Example 7: LASSO and Sub-gradients

As an example, suppose that we want to solve a minimization problem of the form as Equation (3.16) with a convex and differentiable cost function f, and a single constraint specified by $g(\alpha) = \sum_{j=1}^{p} |\alpha_j| - R$ for some positive constant R. Thus, the constraint $g(\alpha) \leq 0$ is equivalent to requiring that it belongs to an L_1-ball of radius R. Recalling the form of the sub-differential in Equation (3.21) for the absolute value function, the condition in Equation (3.22) becomes

$$\nabla f(\alpha^*) + \lambda^* z^* = 0, \tag{3.23}$$

where the sub-gradient vector satisfies $z_j^* \in \text{sign}(\alpha_j^*)$ for each $j = 1, \cdots, p$. When the cost function f is the squared error $f(\alpha) = \|\mathbf{y} - \mathbf{X}\alpha\|_2^2$.

Example 8

Consider $f(x) = |x|$. For $x < 0$, the sub-gradient is unique: $\partial f(x) = \{-1\}$. Similarly, for $x > 0$ we have $\partial f(x) = \{1\}$. At $x = 0$, the sub-differential is defined by the inequality $|x| \geq \mathbf{g}x$ for all x, which is satisfied if and only if $\mathbf{g} \in [-1, 1]$. Therefore we have $\partial f(0) = [-1, 1]$. This is illustrated in Figure 8.

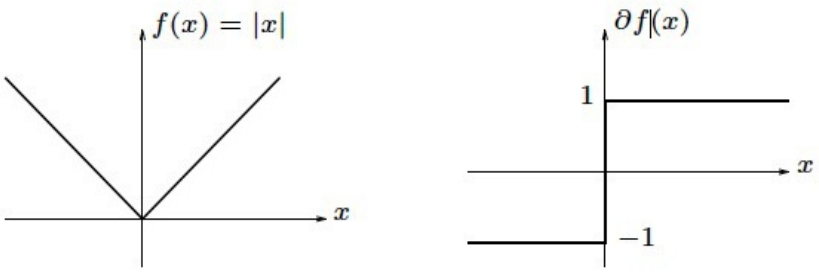

Figure 8: The Absolute Value Function (left), and its Sub-differential $\partial f(x)$ (right).

Basic Properties of Sub-Gradient

The sub-differential $\partial f(x)$ is always a closed convex set, even if f is not convex. This follows from the fact that it is the intersection of an infinite set of halfspaces:

$$\partial f(x) = \bigcap_{z \in \text{dom} f} \{g | f(z) \geq f(x) + \mathbf{g}^T (z - x)\}.$$

Figure 9 shows an epigraph of a function f at a point.

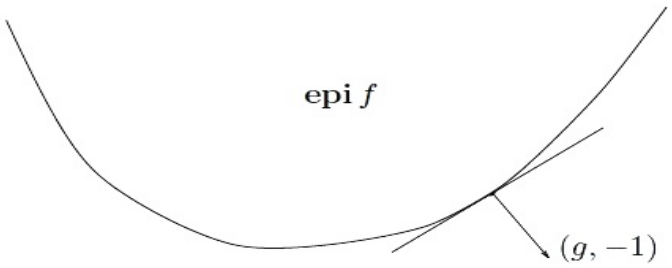

Figure 9: Epigraph of f at a Point.

A vector $\mathbf{g} \in \Re^p$ is a sub-gradient of f at x if and only if $(\mathbf{g}, -1)$ defines a supporting hyperplane to epi f at $(x, f(x))$.

Gram-Schmidt Orthogonalisation

We have seen that it can be very convenient to have an orthonormal basis for a given vector space, in order to compute expansions of arbitrary vectors within that space. Therefore, given a non-orthonormal basis, it is desirable to have a process for obtaining an orthonormal basis from it. Suppose we have a set of functions $\{\mathbf{x}_1, \mathbf{x}_2, \dots\}$ which are not linearly independent. To construct an orthonormal set from this set, we proceed as follows. Given a basis $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p\}$ for a subspace W of V, the method involves using orthogonal projections to construct an orthogonal basis $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p\}$ for V. The $\mathbf{v}_i$'s are constructed so that $\text{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\} = \text{span}\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}, \quad k = 1, \dots, p$. We define the Gram-Schmidt process as follows:

$$
\begin{aligned}
\mathbf{v}_1 &= \mathbf{x}_1 \\
\mathbf{v}_2 &= \mathbf{x}_2 - \frac{\langle \mathbf{x}_2, \mathbf{v}_1 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} \mathbf{v}_1 \\
\mathbf{v}_3 &= \mathbf{x}_3 - \frac{\langle \mathbf{x}_3, \mathbf{v}_1 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} \mathbf{v}_1 - \frac{\langle \mathbf{x}_3, \mathbf{v}_2 \rangle}{\langle \mathbf{v}_2, \mathbf{v}_2 \rangle} \mathbf{v}_2 \\
&\vdots \\
\mathbf{v}_p &= \mathbf{x}_p - \frac{\langle \mathbf{x}_p, \mathbf{v}_1 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} \mathbf{v}_1 - \frac{\langle \mathbf{x}_p, \mathbf{v}_2 \rangle}{\langle \mathbf{v}_2, \mathbf{v}_2 \rangle} \mathbf{v}_2 - \cdots - \frac{\langle \mathbf{x}_p, \mathbf{v}_{p-1} \rangle}{\langle \mathbf{v}_{p-1}, \mathbf{v}_{p-1} \rangle} \mathbf{v}_{p-1}.
\end{aligned}
$$

Then, $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p\}$ is an orthogonal basis for W. Moreover, $\text{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p\} = \text{span}\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p\}, \quad \text{for} \quad 1 \leq k \leq p$.

Relationship Between L_1 and L_2-Norms Regularization

The connection between L_1- and L_2-norms may be expressed through what is referred to as the Elastic Net (Zou & Hastie, 2005) for a β, $0 < \beta < 1$, and a non-negative λ. The Elastic Net problem may be expressed as

$$
\min_{\alpha_0,\, \alpha} \left[\frac{1}{2m} \sum_{i=1}^{m} (y_i - \alpha_0 - x_i^T \alpha)^2 + \lambda E_\beta(\alpha) \right],
$$

where

$$\begin{aligned} E_\beta(\alpha) &= \frac{(1-\beta)}{2}\|\alpha\|_2^2 + \beta\|\alpha\|_1 \\ &= \sum_{j=1}^{p}\left(\frac{1-\beta}{2}\alpha_j^2 + \beta|\alpha_j|\right). \end{aligned}$$

We notice that when $\beta = 1$, Elastic Net is the same as the LASSO. However, as β shrinks to 0, Elastic Net approaches the Ridge regression. For other values of β, the penalty $E_\beta(\alpha)$ is an interpolation between the L_1-norm and L_2-norm of α.

Now since the L_2 regularization least squares problem has the analytic solution

$$\alpha_{L_2} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y},$$

as $\lambda \to 0$, α_{L_2} converges to the Moore-Penrose solution $\mathbf{X}^+\mathbf{y}$ where $\mathbf{X}^+$ is the Moore-Penrose pseudo-inverse of $\mathbf{X}$. In the expression, α_{L_2} is a linear function of $\mathbf{y}$. Equation (5.11) in Chapter Five shows that L_1 regularization gives a solution α_{L_1} which is not linear in $\mathbf{y}$.

There is a finite convergence to zero as λ get to ∞ for L_1 regularization. For L_2 regularization, as λ get to ∞, the optimal solution tends to 0. However, for L_1 regularized LSP, convergence occurs for finite value of λ :

$$\lambda \geq \lambda_{\max} = \left\|2\mathbf{X}^T\mathbf{y}\right\|_\infty,$$

where $\|\mathbf{u}\|_\infty = \max|u_i|$ and denotes the L_∞-norm of vector $\mathbf{u}$. This criterion is clearly deduced from the optimality condition in Equation (5.14). In that equation, the condition that 0 is an optimal solution is that

$$(2\mathbf{X}^T\mathbf{y})_i \in \{-\lambda,\ \lambda\} \quad i = 1,\ 2, \ldots, n.$$

Thus, $\left\|2\mathbf{X}^T\mathbf{y}\right\|_\infty \leq \lambda$. This condition remains crucial in most LASSO solutions, as will be seen in our derived LASSO in Chapter Five. The behaviour of the regularization paths of the two methods are quite different. The solution

α_{L_2} to the Tikhonov regularization problem varies smoothly as the regularization parameter λ varies over $[0, \infty)$. By contrast, (the regularization path of L_1 regularized LSP) the family of solution as λ varies over $(0, \infty)$ has the piece-wise linear solution path property. Thus, there are values $\lambda_1, \lambda_2, \ldots, \lambda_k$ with $0 = \lambda_k < \cdots < \lambda = \lambda_{\max}$ such that the regularized path is a piece-wise linear on $\Re^p$:

$$\alpha_{L_1} = \frac{1}{\lambda_i - \lambda_{i+1}} \left[(\lambda_i - \lambda)\alpha^{(i+1)} + (\lambda - \lambda_{i+1})\alpha^{(i)} \right], \qquad (3.24)$$
$$\lambda_{i-1} < \lambda < \lambda_i, \quad i = 1, 2, \cdots, k-1$$

where $\alpha^{(i)}$ is a solution to Equation (3.24) with $\lambda = \lambda_i$. Thus, $\alpha^1 = 0$ and $\alpha_{L_1} = 0$ for $\lambda \geq \lambda_1$. What this means is that L_1 regularized LS typically yields sparse solution with few non-zero components. As λ decreases, α_{L_1} tends to be sparse but not necessarily (Hastie et al., 2001; Tibshirani, 1996). In contrast, solution α_{LS} to the Tikhonov regularization problem typically has all coefficients non-zero. The knowledge of this would be crucial in determining optimal values of λ particularly in the case where a sparse solution is rather not desired. This is illustrated in a kind of data fitting which is more related to solution retrieved.

Optimality Conditions and Dual Problems

In this section, we review methods associated with optimality conditions and dual problems in regularization.

Optimality Conditions

The Karush-Kuhn-Tucker (KKT) optimality conditions may be stated as

$$\mathbf{X}_i^T(\mathbf{y} - \mathbf{X}\hat{\alpha}) = \lambda\gamma, \qquad (3.25)$$

where

$$\gamma_i = \begin{cases} \text{sign}(\alpha_i), & \alpha_i \neq 0 \\ \\ [-1,\ 1], & \alpha_i = 0 \end{cases} \qquad i = 1,\ 2,\ \ldots,\ p.$$

Here, $\gamma_i \in \Re^p$ is the sub-gradient of $\|\alpha\|_1$. An important property of a LASSO solution is that it satisfies the KKT conditions. As it does, it will have some basic features which is summarised in the following Lemma.

Lemma 1

For any $\mathbf{y}, \mathbf{X}$ and $\lambda \geq 0$, the LASSO problem has the following properties:

(i) There is either a unique LASSO solution or an uncountably infinite solutions.

(ii) Every LASSO solution $\hat{\alpha}$ gives the same fitted value $\mathbf{X}\hat{\alpha}$.

(iii) If $\lambda > 0$, then every LASSO solution $\hat{\alpha}$ has the same L_1-norm.

A sketch of the proof of this Lemma can be seen in Tibshirani (2013).

Duality Problems and Sub-optimality Bound

The L_1-regularized LSP may be presented as a Lagrangian dual. By introducing a new variable $\mathbf{z} \in \Re^m$ and new equality constraint $\mathbf{z} = \mathbf{X}\alpha - \mathbf{y}$, we construct an equivalent problem

$$\min_{\alpha} \ \mathbf{z}^T\mathbf{z} + \lambda_i \|\alpha\|_1 \qquad (3.26)$$

$$\text{subject to } \mathbf{z} = \mathbf{X}\alpha - \mathbf{y}.$$

Let $v_i \in \Re,\ i = 1,\ 2,\ldots,m$ be dual variables associated with the equality constraints $\mathbf{z}_i = (\mathbf{X}\alpha - \mathbf{y})_i$, the Lagrangian is

$$\mathbf{L}(\alpha, \mathbf{z}, \mathbf{v}) = \mathbf{z}^T\mathbf{z} + \lambda \|\alpha\|_1 + \mathbf{v}^T(\mathbf{X}\alpha - \mathbf{y} - \mathbf{z}).$$

The dual form is given as

$$\inf_{\alpha,\mathbf{z}} \mathbf{L}(\alpha, \mathbf{z}, \mathbf{v}) = \begin{cases} -(\frac{1}{4})\mathbf{v}^T\mathbf{v} - \mathbf{v}^T\mathbf{y}, & \left|(\mathbf{X}^T\mathbf{v})_i\right| \leq \lambda_i, \quad i = 1,2,\ldots,m \\ -\infty, & \text{otherwise} \end{cases}$$

The Lagrangian dual of Equation (3.26) is

$$\max\ \mathbf{G}(\mathbf{v}) \tag{3.27}$$

$$\text{subject to} \left|(\mathbf{X}^T\mathbf{v})_i\right| \leq \lambda_i, \quad i = 1,2,\ldots,m,$$

where the dual objective $\mathbf{G}(\mathbf{v})$ is $\mathbf{G}(\mathbf{v}) = -(\frac{1}{4})\mathbf{v}^T\mathbf{v} - \mathbf{v}^T\mathbf{y}$. Equation (3.27) is a convex optimization problem with $\mathbf{v} \in \Re^m$ and is dual feasible if it satisfies the constraints of Equation (3.27). (as described in Boyd & Vandenberghe, 2004).

Now any dual feasible point $\mathbf{v}$ gives a lower bound on the optimal value p^* of the primal problem in Equation (3.27), that is, $\mathbf{G}(\mathbf{v}) \leq p^*$, called weak duality. Furthermore, the optimal values of the primal and dual are dual since the primal problem in Equation (3.27) satisfies Slater's condition, the strong duality (Boyd & Vandenberghe, 2004). A property of the L_1-regularized LSP is that for an arbitrary α, we can derive a bound on the sub-optimality of α, by constructing a dual feasible point $\mathbf{v} = 2\mathbf{w}(\mathbf{X}\alpha - \mathbf{y})$, where

$$\mathbf{w} = \min\left\{\frac{\lambda}{|2(\mathbf{X}^T\mathbf{X}\alpha)_i - 2\mathbf{y}_i|}\right\}, \quad i = 1,2,\ldots,m.$$

The point $\mathbf{v}$ is dual feasible and so $\mathbf{G}(\mathbf{v})$ is a lower bound on p^*, the optimal value of Equation (3.27). The difference

$$\eta = \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda\|\alpha\|_1 - \mathbf{G}(\mathbf{v})$$

between the primal objective value of α and the associated lower bound $\mathbf{G}(\mathbf{v})$ gives the duality gap. By weak duality $\eta \geq 0$ and equality holds at an optimal point (that is, strong duality).

Truncated Newton Interior-Point Method

The L_1-regularized LSP in Equation (1.13) may be transformed to a convex quadratic problem (QP) with linear inequality constraints

$$\begin{aligned} \min_{\alpha} \quad & \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda \sum_{i=1}^{p} u_i \\ \text{subject to} \quad & -u_i \leq \alpha_i \leq u_i, \quad i = 1, 2, \cdots, p \end{aligned}$$

where the variables are $\alpha \in \Re^p$ and $\mathbf{u} \in \Re^p$. An interior-point method for solving this QP is a custom interior point method. Define the logarithm barrier for the bound constraints $-u_i \leq \alpha_i \leq u_i$ in Equation (3.28) as

$$\Phi(\alpha, \mathbf{u}) = -\sum_{i=1}^{p} \log(u_i + \alpha_i) - \sum_{i=1}^{p} \log(u_i - \alpha_i) \tag{3.28}$$

defined over the domain

$$\text{dom}\Phi = \left\{ (\alpha, u) \in \Re^p \times \Re^p \mid |\alpha_i| < u_i, \quad i = 1, 2, \ldots, p \right\}.$$

The central path (CP) consists of the unique minimizer $(\alpha^*(t), \mathbf{u}^*(t))$ of the convex form

$$\phi_t(\alpha, \mathbf{u}) = t \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + t \sum_{i=1}^{p} \lambda u_i + \Phi(\alpha, \mathbf{u})$$

as t varies over $(0, \infty)$. We associate with each $(\alpha^*(t), \mathbf{u}^*(t))$, $\mathbf{v}^*(t) = 2(\mathbf{X}\alpha^*(t) - \mathbf{y})$ which coincides with the dual feasible point v constructed from $\alpha^*(t)$. Hence, $\mathbf{v}^*(t)$ is dual feasible. Thus the path tends to an optimal solution since $(\alpha^*(t), \mathbf{u}^*(t))$ is no more than $\frac{2p}{t}$-suboptimal. We compute a sequence of points on the LP for increasing t. The method can be terminated when $\frac{2p}{t} \leq \varepsilon$, the target duality

gap (see Boyd & Vandenberghe, 2004). In such barrier methods, Newton's method is used to minimize ϕ_t. That is, the exact solution to the Newton's system

$$\mathbf{H}\begin{bmatrix}\Delta\alpha\\ \Delta\mathbf{u}\end{bmatrix} = -\mathbf{g},$$

gives the search direction, where $\mathbf{H} = \nabla^2\phi_t(\alpha,\mathbf{u}) \in \Re^{2p\times 2p}$ is the Hessian and $\mathbf{g} = \nabla\phi_t(\alpha,\mathbf{u}) \in \Re^{2p}$ is the gradient at the current iterate. Kim et al. (2007) specifies the algorithm for the Truncated Newton's Interior Point Method for L_1-regularized least squares problems which we used to verify our solutions.

Generalised Linear Models

Linear models are suitable when the response variable is quantitative, and ideally when the error distribution is Gaussian. However, other types of response arise in practice. For instance, binary variables can be used to indicate the presence or absence of some attribute (for example, "cancerous" versus "normal" cells in a biological assay, or "clicked" versus "not clicked" in web browsing analysis); here the binomial distribution is more appropriate. Sometimes the response occurs as counts (for example, number of arrivals in a queue, or number of photons detected); here the Poisson distribution might be called for.

In this chapter, we discuss generalisations of simple linear models and the LASSO that are suitable for such applications. With a binary response coded in the form $y \in \{0,1\}$, the linear logistic model is often used: it models the log-likelihood ratio as the linear combination

$$\log\frac{Pr(y=1|\mathbf{X}=x)}{Pr(y=0|\mathbf{X}=x)} = \alpha_0 + \alpha^T x, \tag{3.29}$$

where $\mathbf{x} = (x_1,\ x_2,\ \ldots,\ x_p)$ is a vector of predictors,

$\alpha_0 \in \Re$ is an intercept term, and $\alpha \in \Re^p$ is a vector of regression coefficients. Inverting this transformation yields an expression for the conditional probability

$$Pr\Big(y = 1|\mathbf{X} = x\Big) = \frac{e^{\alpha_0+\alpha^T x}}{1+e^{\alpha_0+\alpha^T x}}. \tag{3.30}$$

By inspection, without any restriction on the parameters $(\alpha_0,\ \alpha)$, the model specifies probabilities lying in $(0,\ 1)$. We typically fit logistic models by maximizing the binomial log-likelihood of the data. The logit transformation in Equation (3.30) of the conditional probabilities is an example of a link function. In general, a link function is a transformation of the conditional mean $E[y|\mathbf{X} = x]$in this case, the conditional probability that $y = 1$ to a more natural scale on which the parameters can be fit without constraints. As another example, if the response y represents counts, taking values in $\{0,\ 1,\ 2,\ \ldots\}$, then we need to ensure that the conditional mean is positive. A natural choice is the log-linear model

$$\log E\Big[y|\mathbf{X} = x\Big] = \alpha_0 + \alpha^T x, \tag{3.31}$$

with its log link function. Here, we fit the parameters by maximizing the Poisson log-likelihood of the data. The models in Equation (3.30) and Equation (3.31) are both special cases of generalised linear models. These models describe the response variable using a member of the exponential family, which includes the Bernoulli, Poisson, and Gaussian as particular cases. A transformed version of the response mean $E[Y|\mathbf{X} = x]$ is then approximated by a linear model. In detail, if we use $\mu(x) = E[y|\mathbf{X} = x]$ to denote the conditional mean of y given $\mathbf{X} = x$, then a GLM is based on a model of the form

$$g\Big[\mu(x)\Big] = \underbrace{\alpha_0 + \alpha^T x}_{\eta(x)}, \tag{3.32}$$

where $g : \Re \to \Re$ is a strictly monotonic link function. For example, for a binary response $y \in \{0,\ 1\}$, the logistic regression model is based on the choices

$$\mu(x) = Pr[y = 1|X = x]$$

and

$$g(\mu) = \text{logit}(\mu) = \log(\mu/(1-\mu)).$$

When the response variable is modeled as a Gaussian, the choices

$$\mu(x) = \alpha_0 + \alpha^T x$$

and $g(\mu) = \mu$ recover the standard linear model. Generalised linear models can also be used to model the multicategory responses that occur in many problems, including handwritten digit classification, speech-recognition, document classification and cancer classification. The multinomial replaces the binomial distribution discussed under this section, and we use a symmetric log-linear representation:

$$Pr\Big[y = k|\mathbf{X} = x\Big] = \frac{e^{\alpha_{0k}+\alpha_k^T x}}{\sum_{l=1}^{k} e^{\alpha_{0l}+\alpha_l^T x}}. \tag{3.33}$$

Here, there are K coefficients for each variable (one per class). In this chapter, we discuss approaches to fitting generalised linear models that are based on maximizing the likelihood, or equivalently minimizing the negative log-likelihood along with an L_1-penalty

$$\underset{\alpha_0,\ \alpha}{\text{minimize}} \left\{ -L(\alpha_0,\ \alpha; \mathbf{y},\ \mathbf{X}) + \lambda \|\alpha\|_1 \right\}. \tag{3.34}$$

Here $\mathbf{y}$ is the m-vector of outcomes and $\mathbf{X}$ is the $m \times p$ matrix of predictors, and the specific form the log-likelihood ℓ varies according to the GLM. In the special case of Gaussian responses and the standard linear model, we have

$$-L(\alpha_0,\ \alpha; \mathbf{y},\ \mathbf{X}) = \frac{1}{2\sigma^2} \|\mathbf{y} - \alpha_0\mathbf{1} - \mathbf{X}\alpha\|_2^2 + c,$$

where c is a constant independent of $(\alpha_0,\ \alpha)$, so that the optimization problem in Equation (3.34) corresponds to the ordinary linear least squares LASSO.

Least Squares Minimization for Logistic Regression

We consider the hypothesis

$$h_\alpha(x) = g(\alpha^T x) = \frac{1}{1 + e^{-\alpha^T x}},$$

where

$$g(z) = \frac{1}{1+e^{-z}}$$

is called the logistic function or the sigmoid function. Figure 10 shows a plot of $g(z)$.

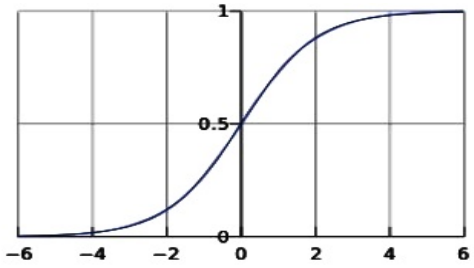

Figure 10: Graph of Sigmoid Function.

We notice that $g(z)$ tends towards 1 as z tends to infinity and $g(z)$ tends towards 0 as $z$tends to negative infinity hence, $h(x)$ and also $g(z)$ are always bounded between 0 and 1. Let us assume that

$$\begin{aligned} Pr(y=1|x;\alpha) &= h_\alpha(x), \text{ and so} \\ Pr(y=0|x;\alpha) &= 1-h_\alpha(x). \end{aligned}$$

This can be written more compactly as

$$Pr(y|x;\alpha) = (h_\alpha(x))^y(1-h_\alpha(x))^{1-y}.$$

Assuming that the m training examples $\{x_i,\ y_i\}_i$ where $y_i \in \{0,\ 1\}$ were generated independently, we can then compute the likelihood of the parameters as

$$\begin{aligned} L(\alpha) &= Pr(\vec{y}|X;\alpha) \\ &= \prod_{i=1}^{m} P(y_i|x_i;\alpha) \\ &= \prod_{i=1}^{m} (h_\alpha(x_i))^{y_i}(1-h_\alpha(x_i))^{1-y_i}. \end{aligned}$$

Before we maximize the log likelihood of the parameters, we first find the derivative of the sigmoid function $g(z)$.

$$\begin{aligned} g'(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\ &= \frac{1}{(1+e^{-z})^2}\,(e^{-z}) \\ &= \frac{1}{(1+e^{-z})}.\left(1-\frac{1}{(1+e^{-z})}\right) \\ &= g(z)(1-g(z)). \end{aligned}$$

Now, maximizing the log likelihood of the parameters gives

$$\begin{aligned} \ell(\alpha) &= \log L(\alpha) \\ \log L(\alpha) &= \log\left(\prod_{i=1}^{m} P(y_i|x_i;\alpha)\right) = \sum_{i=1}^{m}\log P(y_i|x_i;\alpha) \\ &= \sum_{i=1}^{m}\log\left((h_\alpha(x_i))^{y_i}(1-h_\alpha(x_i))^{1-y_i}\right) \\ &= \sum_{i=1}^{m}\left(y_i\log h_\alpha(x_i)+(1-y_i)\log(1-h_\alpha(x_i))\right) \end{aligned}$$

Computing partial derivatives

$$\begin{aligned} \frac{\partial \log L(\alpha)}{\partial \alpha_j} &= \sum_i \frac{\partial y_i \log h_\alpha(x_i)}{\partial \alpha_j} + \frac{(1-y_i)\log(1-h_\alpha(x_i))}{\partial \alpha_j} \\ &= \sum_i y_i \frac{\partial \log g(x_i\alpha)}{\partial \alpha_j} + (1-y_i)\frac{\log(1-g(x_i\alpha))}{\partial \alpha_j} \\ &= \sum_i \frac{y_i}{g(x_i\alpha)}.\frac{\partial g(x_i\alpha)}{\partial \alpha_j} - \frac{1-y_i}{1-g(x_i\alpha)}\frac{\partial g(x_i\alpha)}{\partial \alpha_j} \\ &= \sum_i \left(\frac{y_i}{g(x_i\alpha)} - \frac{1-y_i}{1-g(x_i\alpha)}\right)\frac{\partial g(x_i\alpha)}{\partial \alpha_j} \\ &= \sum_i \left(\frac{y_i}{g(x_i\alpha)} - \frac{1-y_i}{1-g(x_i\alpha)}\right) g(x_i\alpha)(1-g(x_i\alpha))\frac{\partial x_i\alpha)}{\partial \alpha_j} \\ &= \sum_i \left(\frac{y_i}{g(x_i\alpha)} - \frac{1-y_i}{1-g(x_i\alpha)}\right) g(x_i\alpha)(1-g(x_i\alpha))x_{ij} \\ &= (y_i - g(x_i\alpha))x_{ij} \\ &= (y_i - h_\alpha(x_i))x_{ij}. \end{aligned}$$

Logistic regression has been popular in biomedical research for half a century, and has recently gained popularity for modeling a wider range of data. In the high-dimensional setting, in which the number of features p is larger than the sample size, it cannot be used without modification. When $p > m$, any linear model is over-parametrised, and regularization is needed to achieve a stable fit. Such high-dimensional models arise in various applications.

Linear Fitting

Given a set of data points $(x_i,\ y_i),\ i = 1,\ 2,\ \ldots,\ m$, where $x_i \in \Re^p$ and $y_i \in \Re$. We assume that an approximate linear relation holds: $y_i \approx X_i^T \alpha,\ i = 1,\ 2,\ \ldots,\ m$. The corresponding least squares problem is given as

$$\min_{\alpha \in \Re^p} \sum_{i=1}^{m} (X_i^T \alpha - y_i)^2,$$

which has an equivalent formulation as

$$\min_{\alpha \in \Re^p} \|\mathbf{X}\alpha - \mathbf{y}\|^2,$$

where

$$\mathbf{X} = \begin{bmatrix} -\mathbf{x}_1^T - \\ -\mathbf{x}_2^T - \\ \vdots \\ -\mathbf{x}_m^T - \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

Polynomial Fitting

Given a set of data points in $\Re^2 : (u_i,\ y_i),\ i = 1,\ 2,\ \ldots,\ m$ for which the following approximate relation holds for some $a_0,\ \ldots,\ a_d$:

$$\sum_{j=0}^{d} a_j u_i^j \approx y_i,\ i = 1, \ldots,\ m.$$

The system is

$$\begin{pmatrix} 1 & u_1 & u_1^2 & \cdots & u_1^d \\ 1 & u_2 & u_2^2 & \cdots & u_2^d \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & u_m & u_m^2 & \cdots & u_m^d \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_d \end{pmatrix}$$

The least squares solution is well defined if the $m \times (d+1)$ matrix is of full column rank.

Cross-Validation

Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. In typical cross-validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against. The basic form of cross-validation is k-fold cross-validation. Other forms of cross-validation are special cases of k-fold cross-validation or involve repeated rounds of k-fold cross-validation. In k-fold cross-validation, the data is first partitioned into k equally (or nearly equally) sized segments or folds. Subsequently k iterations of training and validation are performed such that within each iteration a different fold of the data is held-out for validation while the remaining $k-1$ folds are used for learning. In data mining and machine learning 10-fold cross-validation (k=10) is the most common. Cross-validation is used to evaluate or compare learning algorithms as follows: in each iteration, one or more learning algorithms use $k-1$ folds of data to learn one or more models, and subsequently the learned models are asked to make predictions about the data in the validation fold. The performance of each learning algorithm on each fold can be tracked using some pre-determined performance metric like accuracy. Upon completion, k samples

of the performance metric will be available for each algorithm. Different methodologies such as averaging can be used to obtain an aggregate measure from these sample, or these samples can be used in a statistical hypothesis test to show that one algorithm is superior to another.

In statistics or data mining, a typical task is to learn a model from available data. Such a model may be a regression model or a classifier. The problem with evaluating such a model is that it may demonstrate adequate prediction capability on the training data, but might fail to predict future unseen data. cross-validation is a procedure for estimating the generalisation performance in this context.

Principal Component Analysis

In practice, there is the tendency to take measurements along dimensions that are more in number that we actually need. In addition, datasets also come with a lot of noise which contaminates the data. In Principal Component Analysis (PCA), the general aim is to re-construct the dataset along a new set of dimensions (of the same number as the original dimensions) so that the most meaningful dimensions can be determined. PCA is closely related to the technique of singular value decomposition (SVD). PCA makes one assumption: linearity. Thus, given data on $\mathbf{X}_{m\times p} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ on p-dimensions, the ith PCA y_i is the linear combination of the x_j given by

$$y_i = \sum_{j}^{p} a_{ij} x_j; \quad i = 1, 2, \ldots, p$$

or

$$y_i = \mathbf{a}_i^T \mathbf{x}; \quad i = 1, 2, \ldots, p. \tag{3.35}$$

Thus, the vector of principal components is given by

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots a_{1j} & \cdots a_{1p} \\ a_{21} & a_{22} & a_{23} & \cdots a_{2j} & \cdots a_{2p} \\ & & & \vdots & \\ a_{i1} & a_{i2} & a_{i3} & \cdots a_{ij} & \cdots a_{ip} \\ & & & \vdots & \\ a_{p1} & a_{p2} & a_{p3} & \cdots a_{pj} & \cdots a_{pp} \end{pmatrix}$$

This representation may be given as

$$\mathbf{y} = \mathbf{A}\mathbf{x},$$

where $\mathbf{A}$ is $p \times p$ matrix whose ith row is the coefficients of the ith PC. Using the matrix $\mathbf{A}$, the data $\mathbf{X}$ may be reconstructed as

$$\mathbf{Y} = \mathbf{A}\mathbf{X}^T, \tag{3.36}$$

where $\mathbf{Y}$ is $p \times m$ matrix of reconstructed data and $\mathbf{X}$ is as defined.

Three main conditions are imposed on the linear combinations in Equation (3.35). The first is that the inner product of any two PCs is such that

$$\mathbf{a}_i\mathbf{a}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases}$$

This means that in Equation (3.36), $\mathbf{A}^T\mathbf{A} = \mathbf{I}$. This condition means that any two PCs are independent and orthonormal.

The variance of y_i is given by $\text{Var}(y_i) = \mathbf{a}_i^T \Sigma_{\mathbf{X}} \mathbf{a}_i$, where $\Sigma_{\mathbf{X}}$ is the variance-covariance matrix of the original variables $\mathbf{x}$. A second condition is that $\text{Var}(y_1) > \text{Var}(y_2) > \cdots > \text{Var}(y_p)$. Since $\text{Var}(y_i)$ represents the amount of variation in the data accounted for by y_i, this condition helps to identify those components that may be considered as redundant. $\text{Var}(y_i)$ is given by the eigenvalue λ_i corresponding to y_i which is the ith eigenvector of the matrix $\Sigma_{\mathbf{X}}$.

The goal of PC is to obtain an orthonormal transformation of the dataset $\mathbf{X}$ given in Equation (3.36). In this case, the variance-covariance matrix of the transformed data $\mathbf{Y}$ is given by

$$\Sigma_{\mathbf{Y}} = \frac{1}{n-1}\mathbf{Y}\mathbf{Y}^T.$$

Making a substitution for $\mathbf{Y}$ from Equation (3.36), we obtain

$$\begin{aligned}
\Sigma_{\mathbf{Y}} &= \frac{1}{n-1}(\mathbf{AX})(\mathbf{AX})^T \\
&= \frac{1}{n-1}\mathbf{A}\mathbf{X}\mathbf{X}^T\mathbf{A}^T \\
&= \frac{1}{n-1}\mathbf{A}(\mathbf{X}\mathbf{X}^T)\mathbf{A}^T \\
&= \frac{1}{n-1}\mathbf{A}\Sigma_{\mathbf{X}}\mathbf{A}^T
\end{aligned}$$

By the first condition on a principal component, the matrix $\Sigma_{\mathbf{Y}}$ must necessarily be diagonal. To see this, we recall that for a symmetric matrix, $\Sigma_{\mathbf{X}}$, there exists a matrix $\mathbf{P}$ such that $\Sigma_{\mathbf{X}} = \mathbf{P}\mathbf{D}\mathbf{P}^T$, where $\mathbf{D}$ is a diagonal matrix and $\mathbf{P}$ is a matrix whose columns are the eigenvectors of $\Sigma_{\mathbf{X}}$. That is, $\mathbf{A} = \mathbf{P}^T$. Thus, $\Sigma_{\mathbf{X}} = \mathbf{A}^T\mathbf{D}\mathbf{A}$, and noting that $\mathbf{A}^{-1} = \mathbf{A}^T$ we have

$$\begin{aligned}
\Sigma_{\mathbf{Y}} &= \frac{1}{n-1}(\mathbf{A}\Sigma_{\mathbf{X}}\mathbf{A}^T) \\
&= \frac{1}{n-1}\mathbf{A}(\mathbf{A}^T\mathbf{D}\mathbf{A})\mathbf{A}^T \\
&= \frac{1}{n-1}(\mathbf{A}\mathbf{A}^T)\mathbf{D}(\mathbf{A}\mathbf{A}^T) \\
&= \frac{1}{n-1}\mathbf{D}.
\end{aligned}$$

Therefore, $\mathrm{Var}(y_i) = \frac{1}{n-1}\mathbf{D}_{ii}$. That is, the variance of the ith principal component is the ith diagonal elements of $\Sigma_{\mathbf{Y}}$.

Singular Value Decomposition

Let $\mathbf{X}$ be an arbitrary $m \times p$ matrix and $\mathbf{X}^T\mathbf{X}$ be of rank r, square, symmetric $p \times p$ matrix. We define the following quantities.

1. The vectors $\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \ldots, \hat{\mathbf{v}}_r$ is the set of r orthonormal $p \times 1$ eigenvectors with associated eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_r$ of the symmetric matrix $\mathbf{X}^T\mathbf{X}$.

$$(\mathbf{X}^T\mathbf{X})\hat{\mathbf{v}}_i = \lambda_i \hat{\mathbf{v}}_i.$$

2. The value $\sigma_i \equiv \sqrt{\lambda_i}$ are positive real and called the singular values.

3. The vectors $\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \ldots, \hat{\mathbf{u}}_r$ is the set of r orthonormal $m \times 1$ vectors defined by the projection

$$\hat{\mathbf{u}}_i \equiv \frac{1}{\sigma_i}\mathbf{X}\hat{\mathbf{v}}_i.$$

4. The inner product $\hat{\mathbf{u}}_i\hat{\mathbf{u}}_j = \delta_{ij}$, since $\mathbf{v}_i$s are orthonormal.

5. The norm of the product $\|\mathbf{X}\hat{\mathbf{v}}_i\| = \sigma_i$.

From the third definition,

$$\mathbf{X}\hat{\mathbf{v}}_i = \sigma_i\hat{\mathbf{u}}_i. \tag{3.37}$$

Let $\Sigma = \text{diag}[\sigma_1,\ \sigma_2,\ \cdots,\ \sigma_r,\ 0,\ 0,\ \cdots,\ 0]$ be a $p \times p$ diagonal matrix with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. The corresponding augmented orthogonal matrices $\mathbf{V}$ and $\mathbf{U}$ are

$$\mathbf{V} = [\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \cdots, \hat{\mathbf{v}}_r, \hat{\mathbf{v}}_{r+1}, \cdots, \hat{\mathbf{v}}_p]$$

$$\mathbf{U} = [\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \cdots, \hat{\mathbf{u}}_r, \hat{\mathbf{u}}_{r+1}, \cdots, \hat{\mathbf{u}}_m].$$

Equation (3.37) may be generalised for all p components as

$$\mathbf{XV} = \mathbf{U}\Sigma. \tag{3.38}$$

Noting that $\mathbf{V}^{-1} = \mathbf{V}^T$, the matrix $\mathbf{X}$ in Equation (3.38) may be decomposed as

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T. \tag{3.39}$$

Thus, in this decomposition, we can identify the principal components of the matrix $\mathbf{X}^T\mathbf{X}$ as well as their corresponding singular values. The result in Equation (3.39) may best be interpreted using Equation (3.37) and by noting that

$$\mathbf{U}^T\mathbf{X} = \mathbf{S},$$

where $\mathbf{S} = \Sigma\mathbf{V}^T$. This represent a span of the columns of $\mathbf{X}$ by the basis $\mathbf{U}^T$. Similarly, we can deduce from Equation (3.39) that

$$\mathbf{V}^T\mathbf{X} = \mathbf{S},$$

where in this case $\mathbf{S} = \mathbf{U}^T\Sigma$. This represent a span of the rows of $\mathbf{X}$ by the basis $\mathbf{v}^T$.

SVD and PCA

With some computations, it can be shown that the two methods are intimately related. We return to the original $m \times p$ data matrix $\mathbf{X}$. We can define a new matrix $\mathbf{Y}$ as a $p \times m$ matrix, where

$$\mathbf{Y} \equiv \frac{1}{\sqrt{n-1}}\mathbf{X}^T.$$

Each column of $\mathbf{Y}$ has zero mean. The definition of $\mathbf{Y}$ becomes clear by analyzing $\mathbf{Y}^T\mathbf{Y}$.

$$\begin{aligned}
\mathbf{Y}^T\mathbf{Y} &= \left(\frac{1}{\sqrt{n-1}}\mathbf{X}^T\right)^T\left(\frac{1}{\sqrt{n-1}}\mathbf{X}^T\right) \\
&= \frac{1}{n-1}\mathbf{X}^{TT}\mathbf{X}^T \\
&= \frac{1}{n-1}\mathbf{X}\mathbf{X}^T \\
\mathbf{Y}^T\mathbf{Y} &= \mathbf{S}_\mathbf{X}.
\end{aligned}$$

By construction $\mathbf{Y}^T\mathbf{Y}$ equals the covariance matrix of $\mathbf{X}$. The principal components of $\mathbf{X}$ are the eigenvectors of $\mathbf{S}_\mathbf{X}$. If we calculate the SVD of $\mathbf{Y}$, the columns of matrix $\mathbf{V}$ contain the eigenvectors of $\mathbf{Y}^T\mathbf{Y} = \mathbf{S}_\mathbf{X}$. Therefore, the columns of $\mathbf{V}$ are the principal components of $\mathbf{X}$.

Chapter Summary

This chapter has outlined various concepts and techniques that would be needed in the study. The techniques cover those that are designed for variable transformation, These include the Gram-Schmidt orthogonalisation process,

principal component and singular value decomposition. Another class of techniques that are covered are those concerned with data fitting procedures. These include the least squares method, polynomial fitting and generalised linear modelling. Techniques of regularization processes have also been reviewed. These are the L_1-norm and L_2-norm regularization. A number of procedures used in regularization have also been covered. Of interest are the sub-gradient calculus, various optimality conditions and the Truncated Newton's Interior Point method.

CHAPTER FOUR

SMOOTHING APPROXIMATIONS FOR THE L_1-NORM REGULARIZATION FUNCTIONAL

Introduction

In this chapter, we consider smoothing approximations for the L_1-norm penalty in the regularized least squares problem given by

$$\min_{\alpha} g(\alpha) = \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda \|\alpha\|_1 , \tag{4.1}$$

which is already introduced in previous chapters. Three main smoothing approximations will be explored. These include a Quadratic approximation of the Lee et al. (2006) approximation, Sigmoid Function approximation (Chen & Mangasarian, 1996) and Cubic Hermite approximation. In each case, we will apply the Tikhonov regularization to the resulting smooth least squares minimization problem which we represent by

$$\min_{\alpha} g(\alpha) = \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \mu J(\alpha), \tag{4.2}$$

where μ is a function of λ in Equation (4.1) and a parameter of the approximating functional. It is possible to apply Tikhonov regularization to the resulting problem in Equation (4.2) since $J(\alpha)$ is now differentiable. The regularized solution to Equation (4.2) will be derived and then specified in terms of singular value decomposition. The performance of the approximation will be assessed by using the Hilbert sub-matrix of dimension 12×7. Subsequently, we will compare the solutions obtained from the regularization method with that of the Modified Newton's method, which is the usual practice in the literature.

Quadratic Approximation

Lee et al. (2006) proposed a method for transforming the non-differentiable L_1-norm function into a differentiable function by replacing it with a differen-

tiable approximation. For a one-dimensional case, the approximation to the absolute value function is given by

$$|x| \approx \sqrt{x^2 + \varepsilon}.$$

To determine the best approximate solution, we first examine the nature of the plot for various values of ε. Approximations of the absolute value function for different values $1,\ 0.01$ and 0.0001, of ε are given in Figure 11.

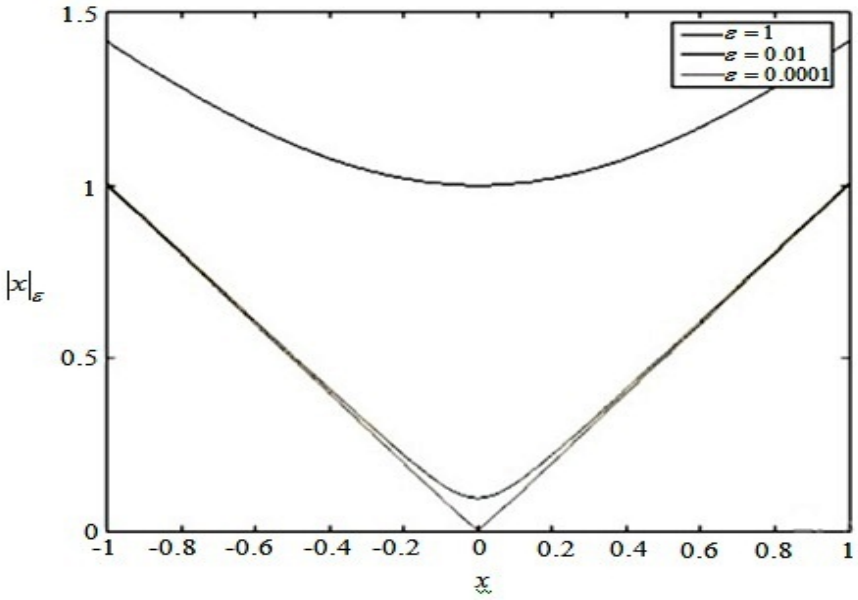

Figure 11: Quadratic Approximation of $|x|_\varepsilon$ for Various Values of Approximating Parameter, ε.

From Figure 11, $|x|_\varepsilon \to |x|$ as $\varepsilon \to 0$. That is,

$$\lim_{\varepsilon \to 0} |x|_\varepsilon = |x|\,.$$

Thus, it would be suitable to choose $\varepsilon = 0.0001$ for the subsequent implementation. The gradient $\nabla(|x|_\varepsilon)$ and the Hessian $\nabla^2(|x|_\varepsilon)$ of the smoothing approximation of the absolute value function given in single variable form are derived as follows:

$$\nabla(|x|_\varepsilon) = \frac{x}{\sqrt{x^2+\varepsilon}} \quad \text{and} \quad \nabla^2(|x|_\varepsilon) = \frac{\varepsilon}{\left(\sqrt{x^2+\varepsilon}\right)^3}.$$

For $\mathbf{x} \in \Re^p$,

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{p} |x_i| \approx \sum_{i=1}^{p} |x_i|_\varepsilon .$$

The loss function given in Equation (4.2) therefore becomes

$$g(\alpha) \approx \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \mu \sum_{i}^{p} \sqrt{\alpha_i^2 + \varepsilon}. \tag{4.3}$$

The regularized solution of Equation (4.3) is given as

$$\alpha_\mu = (\mathbf{X}^T\mathbf{X} + \mu\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}, \tag{4.4}$$

where $\mu = \frac{1}{2}\lambda\varepsilon^{-\frac{1}{2}}$. The regularized solution α_μ written in terms of singular value decomposition (SVD) is given in component form as

$$\alpha_\mu = \sum_{i=1}^{p} \frac{\sigma_i}{\sigma_i^2 + \mu}(\mathbf{U}_i^T\mathbf{y})\mathbf{V}_i,$$

where σ are the singular values of the matrix $\mathbf{X}$.

Proof

Let

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T.$$

Now, from Equation (4.4),

$$\begin{aligned}
\alpha_\mu &= \left[(\mathbf{USV}^T)^T\mathbf{USV}^T+\mu\mathbf{I}\right]^{-1}(\mathbf{USV}^T)^T\mathbf{y} \\
&= \left[(\mathbf{VSU}^T)(\mathbf{USV}^T)+\mu\mathbf{I}\right]^{-1}(\mathbf{VSU}^T)\mathbf{y} \\
&= \left[\mathbf{VS}^2\mathbf{V}^T+\mu\mathbf{I}\right]^{-1}(\mathbf{VSU}^T)\mathbf{y} \\
&= \left[\mathbf{VS}^2\mathbf{V}^T+\mu\mathbf{VIV}^T\right]^{-1}(\mathbf{VSU}^T)\mathbf{y} \\
&= \left[\mathbf{V}\left(\mathbf{S}^2+\mu\mathbf{I}\right)\mathbf{V}^T\right]^{-1}(\mathbf{VSU}^T)\mathbf{y} \\
&= \mathbf{V}\left(\mathbf{S}^2+\mu\mathbf{I}\right)^{-1}\mathbf{V}^{-1}(\mathbf{VSU}^T)\mathbf{y} \\
&= \left(\mathbf{S}^2+\mu\mathbf{I}\right)^{-1}(\mathbf{VSU}^T)\mathbf{y}
\end{aligned}$$

Therefore, the singular value decomposition solution in component form is given as

$$\alpha_\mu = \sum_{i=1}^{p}\frac{\sigma_i}{\sigma_i^2+\mu}(\mathbf{U}_i^T\mathbf{y})\mathbf{V}_i$$

which ends the proof.

Derivation of Analytic Solution using Lee-Quadratic Approximation

Let

$$k(\mathbf{x}) = \sum_{i}^{p}\sqrt{x_i^2+\varepsilon} = \|\mathbf{x}\|_\varepsilon \approx \|\mathbf{x}\|_1 .$$

Since $k(\mathbf{x})$ is differentiable at $x=0$, a Taylor's expansion about $\mathbf{x}=0$ is given as

$$\begin{aligned}
k(\mathbf{x}) &\approx k(\mathbf{x}_0)+\nabla k(\mathbf{x}_0)^T(\mathbf{x}-\mathbf{x}_0)+\frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T\nabla^2 k(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)+\cdots \\
&\approx k(0)+\nabla k(0)^T\mathbf{x}+\frac{1}{2}\mathbf{x}^T\nabla^2 k(0)\mathbf{x}+\cdots,
\end{aligned}$$

where

$$k(0) = p\varepsilon^{\frac{1}{2}}, \quad \nabla k(0) = \nabla k(\mathbf{x}) = \left.\frac{x_i}{\sqrt{x_i^2+\varepsilon}}\right|_{x=0} = 0,$$

and

$$\nabla^2 k(0) = \nabla^2 k(\mathbf{x}) = \frac{\varepsilon}{\left(\sqrt{x_i^2+\varepsilon}\right)^3}\bigg|_{x=0} = \begin{pmatrix} \varepsilon^{-\frac{1}{2}} & 0 & \cdots & 0 \\ 0 & \varepsilon^{-\frac{1}{2}} & \cdots & 0 \\ \vdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & \varepsilon^{-\frac{1}{2}} \end{pmatrix} = \varepsilon^{-\frac{1}{2}}\mathbf{I}_p.$$

Therefore, the quadratic approximation is given as

$$k(\mathbf{x}) = p\varepsilon^{\frac{1}{2}} + \frac{1}{2}\mathbf{x}^T\varepsilon^{-\frac{1}{2}}\mathbf{x}\mathbf{I}_p.$$

Thus, Equation (4.3) expressed in the form of Equation (4.1) gives $g(\alpha)$ as

$$g(\alpha) = \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda\left(p\varepsilon^{\frac{1}{2}} + \frac{1}{2}\alpha^T\varepsilon^{-\frac{1}{2}}\alpha\right). \tag{4.5}$$

Finding the gradient of $g(\alpha)$ and equating to zero gives

$$\begin{aligned} 2\mathbf{X}^T\mathbf{X}\alpha - 2\mathbf{X}^T\mathbf{y} + \lambda\alpha\varepsilon^{-\frac{1}{2}} &= 0 \\ \mathbf{X}^T\mathbf{X}\alpha + \frac{1}{2}\lambda\alpha\varepsilon^{-\frac{1}{2}} &= \mathbf{X}^T\mathbf{y} \\ \left(\mathbf{X}^T\mathbf{X} + \frac{1}{2}\lambda\varepsilon^{-\frac{1}{2}}\mathbf{I}\right)\alpha &= \mathbf{X}^T\mathbf{y} \\ \alpha_\mu = \left(\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}, & \end{aligned} \tag{4.6}$$

where

$$\mu = \frac{1}{2}\lambda\varepsilon^{-\frac{1}{2}},$$

and λ is the regularization parameter in the L_1-norm regularized least squares in Equation (4.1). Notice that the solution in Equation (4.6) is of the form

$$\alpha = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

which is the solution of the L_2-norm regularization.

Numerical Experiment

To illustrate our results, we make use of the 12×7 Hilbert sub-matrix of the 12×12 Hilbert matrix, which constitutes an overdetermined system. As introduced in Chapter One, Hilbert matrices are known to be very ill-conditioned because the coefficient matrix $\mathbf{X}^T\mathbf{X}$ is almost near zero. The problem is to find $\alpha \in \Re^p$ such that $\mathbf{X}\alpha = \mathbf{y}$. The vector $\mathbf{y}$ is chosen such that the true solution is $\alpha = [1,\ 1,\ 1,\ 1,\ 1, 1,\ 1]^T$.

Several iterations are performed to obtain an optimal regularization parameter μ which will hopefully give a solution close to the true solution. Table 7 shows the solutions corresponding to $\mu = 10^{-35},\ 10^{-30}, 10^{-25}, \ldots, 10^{0}$ computed for regularization of order zero using SVD. See Appendix A for the implementation of the algorithm for computing the regularized solution.

Table 7: Quadratic-SVD Regularized Solution Using 12×7 Hilbert Matrix

Parameter μ	Approximate Solution $\hat{\alpha}$	Error $\|\alpha_{exact} - \hat{\alpha}\|$
10^{-35}	0.999999999998873 1.000000000038512 0.999999999666076 1.000000001202600 0.999999997920514 1.000000001715320 0.999999999457774	$2.07948647190648e-009$
10^{-30}	0.999999999998873 1.000000000038512 0.999999999666076 1.000000001202600 0.999999997920514 1.000000001715320 0.999999999457774	$2.07948647190648e-009$
10^{-25}	0.999999999998873 1.000000000038517 0.999999999666034 1.000000001202753 0.999999997920250 1.000000001715537 0.999999999457705	$2.07975037191943e-009$
10^{-20}	0.999999999985505 1.000000000513768 0.999999995471529 1.000000016419557 0.999999971548982 1.000000023462818 0.999999992593616	$2.84510176529196e-008$

Table 7 Continued

Parameter μ	Approximate Solution $\hat{\alpha}$	Error $\|\|\alpha_{exact} - \hat{\alpha}\|\|$
10^{-15}	1.000000005382488 1.000000311138334 0.999994094825253 1.000030102812234 0.999935968486692 1.000060745682307 0.999978746314639	$6.40315133079161e-005$
10^{-10}	0.999921828945325 1.001005379765547 0.997630528523892 0.999928642059073 1.002328582218346 1.001747414368654 0.997383337346468	0.00261666265353178
10^{-5}	0.977300836168243 1.039833840286250 1.052974774228701 1.030493186016140 0.993147747772983 0.950701679037129 0.907598316425139	0.0924016835748606
10^{-1}	0.608292793467497 0.383729725831142 0.291266270999275 0.237993058004172 0.202532009830954 0.176916847526424 0.157405776526399	0.842594223473601
10^{0}	0.0921918039495352 0.0575537783146415 0.0434500839595747 0.0353811004241751 0.0300364447633618 0.0261899735487342 0.0232685527604642	0.976731447239536

From Table 7, as the values of μ increase, the regularized solution deteriorates. Therefore, smaller values of μ yields a good solution. The regularized solution converges at $\mu = 10^{-30}$ with the smallest error of $2.07948647190648e-009$.

A method usually considered in the literature after obtaining a smoothing approximation to replace the L_1-norm functional is an unconstrained optimization method known as the Modified Newton's method. To compare our results using regularization with the Modified Newton's method, we now implement this method.

The algorithm based on the implementation of Modified Newton's Method is formulated as

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k \mathbf{H}(\mathbf{x}_k)^{-1} \nabla g(\mathbf{x}_k),$$

where $\mathbf{x}_k$ is the current iterate, $\mathbf{x}_{k+1}$ is the next iterate, $\nabla g(\mathbf{x}_k)$ is the gradient at the current iterate $\mathbf{x}_k$, $\beta_k > 0$ is the step size and $\mathbf{H}(\mathbf{x}_k)$ is the Hessian at the current iterate.

From Equation (4.3), the gradient of $g(\alpha)$ is given as

$$\nabla g(\alpha) = 2\mathbf{X}^T(\mathbf{X}\alpha - \mathbf{y}) + \lambda G(\alpha),$$

where

$$G(\alpha) = \left[\alpha_1(\alpha_1^2+\varepsilon)^{-\frac{1}{2}},\ \alpha_2(\alpha_2^2+\varepsilon)^{-\frac{1}{2}},\ \cdots,\ \alpha_p(\alpha_p^2+\varepsilon)^{-\frac{1}{2}}\right]^T$$

and the Hessian is also given as

$$\mathbf{H}(\alpha) = 2\mathbf{X}^T\mathbf{X} + \varepsilon\lambda\mathbf{h}(\alpha),$$

where

$$\mathbf{h}(\alpha) = \text{diag}\left[(\alpha_1^2+\varepsilon)^{-\frac{3}{2}},\ (\alpha_2^2+\varepsilon)^{-\frac{3}{2}},\ \cdots,(\alpha_p^2+\varepsilon)^{-\frac{3}{2}}\right].$$

Script-files are created in OCTAVE 3.8.2 to compute the solutions at various values of the regularization parameter. The following algorithm shows the general structure of the implementation of the Modified Newton's method.

Algorithm for Modified Newton's Method
Set initial guess
Set the function parameter
Initialize $\alpha := \text{ones}(7,1)$
Define the approximating function: $g = f(\alpha) + \mu * j(\alpha)$
Define the number of iterations, $i = 1:100$
Define Modified Newtons method, $\alpha(:,\ i+1) = \alpha(:,\ i) - \beta * H/\text{Grd}\ (:,\ i)$
end

See Appendix B for the implementation of the Modified Newton's method based on Lee et al. approximation. In the algorithm based on the Lee et al. approximation, the initial guess is $0.25 * \text{ones}(7,\ 1)$, the function parameter $\varepsilon = 0.0001$, and the stepsize $\beta = 2$.

A number of iterations are performed and the best approximate solution is obtained at the 81st iterate. Table 8 shows various approximate solutions $(\hat{\alpha})$ using various values of the parameter μ.

Table 8: Solution of Modified Newton's Method based on Lee et al. Approximation at 81st Iterate

μ	$\hat{\alpha}$	$\|\|\alpha_{\text{exact}} - \hat{\alpha}\|\|$
10^{-16}	1.000000680934881 0.999975996291894 1.000210550708668 0.999239773222684 1.001312540621827 0.998920937855809 1.000339707272387	0.00131254062182729
10^{-15}	0.999978098438806 1.000772246382488 0.993224729563827 1.024467575479284 0.957750022379654 1.034738943150938 0.989062357494188	0.0422499776203459
10^{-14}	0.999737295807271 1.009255257320510 0.918846675479713 1.292941242914096 0.494330233039847 1.415658427743051 0.869158951998543	0.505669766960153

From Table 8, by increasing the value of μ from 10^{-16}, the solution seems to be deteriorating. The iterations show that as we move away from the 81st iterate, there is not much difference between the solutions from the 82nd to the 100th iteration. The error in the computed solution corresponding to $\mu = 10^{-16}$ is

$$\|\alpha_{\text{exact}} - \hat{\alpha}\|_2 = 0.00131254062182729,$$

which is about 2 digits accurate. The loss in the accuracy of the solution is due to the fact that the coefficient matrix $\mathbf{X}^T\mathbf{X}$ in the regularized solution (Equation (4.6)), is ill-conditioned, with a condition number $\kappa \approx 2.31648078701200e+$

015. Thus, the accuracy of the computed solution is reduced by about 14 digits, which is $d-(k-1)$, where $d=16$ and $k=15$.

Table 9 shows the solutions corresponding to the optimal regularization parameter of Modified Newton's method (MNM) and regularization method (RM).

Table 9: Modified Newton's Method versus Regularization Method with Quadratic Approximation

Method	μ	$\hat{\alpha}$	$\|\alpha_{\text{exact}}-\hat{\alpha}\|$
MNM	10^{-16}	1.000000680934881 0.999975996291894 1.000210550708668 0.999239773222684 1.001312540621827 0.998920937855809 1.000339707272387	$1.31254062182729e-003$
RM	10^{-30}	0.999999999998873 1.000000000038512 0.999999999666076 1.000000001202600 0.999999997920514 1.000000001715320 0.999999999457774	$2.07948647190648e-009$

From Table 9, it is clear that the best approximate solution for the Modified Newton's method occurred at $\mu=10^{-16}$, with the step size $\beta=2$, and at the 81st iteration. For the Regularization method, the best approximate solution occurred at $\mu=10^{-30}$ with 9 digit accuracy.

Sigmoid Function Approximation

In this section, we consider the Sigmoid Function approximation to the L_1-norm functional. The approximation takes advantage of the non-negative projection operators

$$(x)_+ = \max(x,0) \quad \text{and} \quad (-x)_+ = \max(-x,0).$$

This projection function can be smoothly approximated by the integral of a sigmoid function (Chen & Mangasarian, 1996) given as

$$(x)_+ \approx p(x,\kappa) = x + \frac{1}{\kappa}\log(1+e^{-\kappa x}) \quad \text{and}$$

$$(-x)_+ \approx p(-x,\kappa) = -x + \frac{1}{\kappa}\log(1+e^{\kappa x}).$$

The functions $p(x,\kappa)$ and $p(-x,\kappa)$ are members of a class of smoothing functions presented by Chen and Mangasarian (1996). These smoothing approximations of the projections have been used to transform the standard L_1-norm formulation into an efficiently-solved unconstrained problem.

By combining $p(x,\kappa)$ and $p(-x,\kappa)$, we obtain the identity

$$|x| = (x)_+ + (-x)_+.$$

We arrive at a smoothing approximation for the absolute value function that consists of the sum of the integral of two sigmoid functions given by

$$\begin{aligned} |x| &\approx (x)_+ + (-x)_+ = p(x,\kappa) + p(-x,\kappa) \\ &= \frac{1}{\kappa}[\log(1+e^{-\kappa x}) + \log(1+e^{\kappa x})] \\ &\stackrel{def}{=} |x|_\kappa . \end{aligned}$$

The graphs of the projection operators are given in Figure 12.

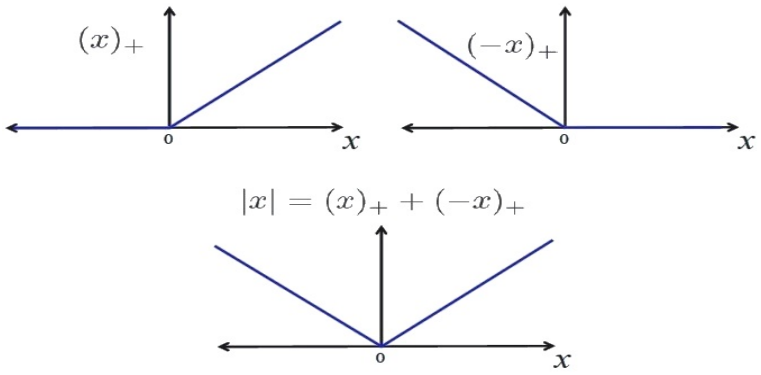

Figure 12: Projection Operators of the Absolute Value Function.

A graph of various values of the parameter κ in approximating the absolute value function is given in the Figure 13. Approximations are given for different values $10,\ 100,\ 1000$ and 10000, of κ.

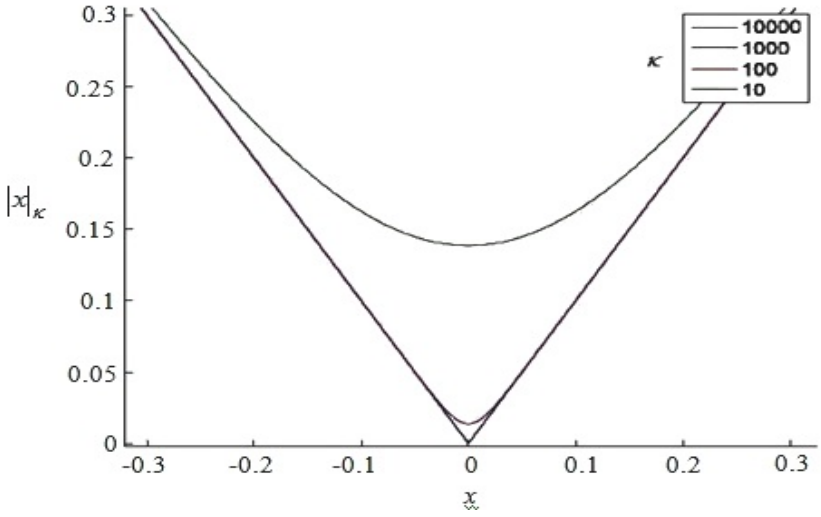

Figure 13: Sigmoid Approximation of $|x|_\kappa$ for Various Values of Approximating Parameter, κ.

From Figure 13, $|x|_\kappa \to |x|$ as $\kappa \to \infty$. That is,

$$\lim_{\kappa\to\infty} |x|_\kappa = |x|\,.$$

Thus, it would be suitable to choose $\kappa = 1000$ for the subsequent implementation.

Given the smoothing approximation, the gradient $\nabla(|x|_\kappa)$ and the Hessian $\nabla^2(|x|_\kappa)$

in single variable form is derived as follows:

$$\begin{aligned}
|x|_\kappa &= \frac{1}{\kappa}\Big[\log(1+e^{-\kappa x}) + \log(1+e^{\kappa x})\Big] \\
\nabla(|x|_\kappa) &= \frac{1}{\kappa}\Big[\frac{-\kappa e^{-\kappa x}}{1+e^{-\kappa x}} + \frac{\kappa e^{\kappa x}}{1+e^{\kappa x}}\Big] \\
&= \frac{-e^{-\kappa x}}{1+e^{-\kappa x}} + \frac{e^{\kappa x}}{1+e^{\kappa x}} \\
&= \frac{1+e^{\kappa x}-1-e^{-\kappa x}}{(1+e^{-\kappa x})(1+e^{\kappa x})} \\
&= \frac{(1+e^{\kappa x})-(1+e^{-\kappa x})}{(1+e^{-\kappa x})(1+e^{\kappa x})} \\
&= \frac{(1+e^{\kappa x})}{(1+e^{-\kappa x})(1+e^{\kappa x})} - \frac{(1+e^{-\kappa x})}{(1+e^{-\kappa x})(1+e^{\kappa x})} \\
&= \frac{1}{1+e^{-\kappa x}} - \frac{1}{1+e^{\kappa x}}.
\end{aligned}$$

Therefore,

$$\nabla(|x|_\kappa) = (1+e^{-\kappa x})^{-1} - (1+e^{\kappa x})^{-1},$$

and

$$\begin{aligned}
\nabla^2(|x|_\kappa) &= (1+e^{-\kappa x})^{-2}\kappa e^{-\kappa x} + (1+e^{\kappa x})^{-2}\kappa e^{\kappa x} \\
&= \kappa\Big[(1+e^{-\kappa x})^{-2}e^{-\kappa x} + (1+e^{\kappa x})^{-2}e^{\kappa x}\Big] \\
&= \frac{\kappa e^{\kappa x}}{(1+e^{\kappa x})^2}\Big[\frac{(1+e^{\kappa})^2}{(1+e^{-\kappa x})^2}(e^{-\kappa x})^2 + 1\Big] \\
&= \frac{\kappa e^{\kappa x}}{(1+e^{\kappa x})^2}\Big[\frac{(e^{-\kappa x}(1+e^{\kappa x})^2}{(1+e^{-\kappa x})^2} + 1\Big] \\
&= \frac{\kappa e^{\kappa x}}{(1+e^{\kappa x})^2}\Big[\frac{(e^{-\kappa x}+1)^2}{(1+e^{-\kappa x})^2} + 1\Big] \\
&= \frac{2\kappa e^{\kappa x}}{(1+e^{\kappa x})^2}.
\end{aligned}$$

Therefore,

$$\nabla^2(|x|_\kappa) = \frac{2\kappa e^{\kappa x}}{(1+e^{\kappa x})^2}.$$

For $\mathbf{x} \in \Re^p$,

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{p} |x_i| \approx \sum_{i=1}^{p} |x_i|_\kappa.$$

The loss function in Equation (4.3) therefore becomes

$$g(\alpha) = \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda \sum_{i}^{p} \frac{1}{\kappa}\Big[\log(1+e^{-\kappa\alpha_i}) + \log(1+e^{\kappa\alpha_i})\Big]. \tag{4.7}$$

The regularized solution is given as

$$\alpha_\mu = (\mathbf{X}^T\mathbf{X} + \mu\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}, \tag{4.8}$$

where $\mu = \frac{1}{4}\lambda\kappa$.

The regularized solution α_μ written in terms of singular value decomposition (SVD) is given in component form as

$$\alpha_\mu = \sum_{i=1}^{p} \frac{\sigma_i}{\sigma_i^2 + \mu}(\mathbf{U}_i^T\mathbf{y})\mathbf{V}_i.$$

Proof

Let

$$\mathbf{X} = \mathbf{USV}^T.$$

Now, from Equation (4.8),

$$\begin{aligned}
\alpha_\mu &= \Big[(\mathbf{USV}^T)^T\mathbf{USV}^T + \mu\mathbf{I}\Big]^{-1}(\mathbf{USV}^T)^T\mathbf{y} \\
&= \Big[(\mathbf{VSU}^T)(\mathbf{USV}^T) + \mu\mathbf{I}\Big]^{-1}(\mathbf{VSU}^T)\mathbf{y} \\
&= \Big[\mathbf{VS}^2\mathbf{V}^T + \mu\mathbf{I}\Big]^{-1}(\mathbf{VSU}^T)\mathbf{y} \\
&= \Big[\mathbf{VS}^2\mathbf{V}^T + \mu\mathbf{VIV}^T\Big]^{-1}(\mathbf{VSU}^T)\mathbf{y} \\
&= \Big[\mathbf{V}(\mathbf{S}^2 + \mu\mathbf{I})\mathbf{V}^T\Big]^{-1}(\mathbf{VSU}^T)\mathbf{y} \\
&= \mathbf{V}\Big(\mathbf{S}^2 + \mu\mathbf{I}\Big)^{-1}\mathbf{V}^{-1}(\mathbf{VSU}^T)\mathbf{y} \\
&= \Big(\mathbf{S}^2 + \mu\mathbf{I}\Big)^{-1}(\mathbf{VSU}^T)\mathbf{y}
\end{aligned}$$

Therefore, the singular value decomposition solution in component form is given as

$$\alpha_\mu = \sum_{i=1}^{p} \frac{\sigma_i}{\sigma_i^2 + \mu}(\mathbf{U}_i^T \mathbf{y})\mathbf{V}_i,$$

which ends the proof.

Derivation of Analytic Solution using Sigmoid Approximation

Minimizing Equation (4.7) gives,

$$\nabla g(\alpha, \kappa) = 2\mathbf{X}^T(\mathbf{X}\alpha - \mathbf{y}) + \lambda \sum_{i=1}^{p} (1 + e^{-\kappa\alpha_i})^{-1} - (1 + e^{\kappa\alpha_i})^{-1} = 0.$$

A linear approximation to

$$(1 + e^{-\kappa\alpha})^{-1} - (1 + e^{\kappa\alpha})^{-1}$$

is obtained as

$$\begin{aligned}
k(\alpha) &= \left[1 + 1 + (-\kappa\alpha) + \frac{(-\kappa\alpha)^2}{2} + \frac{(-\kappa\alpha)^3}{3!} + \cdots\right]^{-1} \\
&\quad - \left[1 + 1 + (\kappa\alpha) + \frac{(\kappa\alpha)^2}{2} + \frac{(\kappa\alpha)^3}{3!} + \cdots\right]^{-1} \\
&= \frac{1}{2}\left[1 + \frac{(-\kappa\alpha)}{2} + \frac{(-\kappa\alpha)^2}{2\times 2} + \frac{(-\kappa\alpha)^3}{2\times 3!} + \cdots\right]^{-1} \\
&\quad - \frac{1}{2}\left[1 + \frac{(\kappa\alpha)}{2} + \frac{(\kappa\alpha)^2}{2\times 2!} + \frac{(\kappa\alpha)^3}{2\times 3!} + \cdots\right]^{-1} \\
&= \frac{1}{2}\left[\left(1 - \frac{(-\kappa\alpha)}{2} - \frac{(-\kappa\alpha)^2}{2\times 2} - \frac{(-\kappa\alpha)^3}{2\times 3!} + \cdots\right)\right. \\
&\quad \left. - \left(1 - \frac{(\kappa\alpha)}{2} - \frac{(\kappa\alpha)^2}{2\times 2!} - \frac{(\kappa\alpha)^3}{2\times 3!} + \cdots\right)\right]
\end{aligned}$$

after some expansion and simplification. By ignoring terms of higher order, we obtain the linear approximation

$$k(\alpha) = \frac{1}{2}\kappa\alpha.$$

Now,

$$g(\alpha) \Rightarrow \mathbf{X}^T\mathbf{X}\alpha + \frac{1}{2}\lambda k(\alpha) = \mathbf{X}^T\mathbf{y}.$$

Using the linear approximation for $k(\alpha)$, we obtain the minimization of $g(\alpha)$ as

$$\mathbf{X}^T\mathbf{X}\alpha + \frac{1}{4}\lambda\kappa\alpha = \mathbf{X}^T\mathbf{y}.$$

Thus,

$$\alpha_\mu = \left(\mathbf{X}^T\mathbf{X} + \mu\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}, \tag{4.9}$$

where

$$\mu = \frac{1}{4}\lambda\kappa.$$

Table 10 shows the solutions corresponding to $\mu = 10^{-35}, 10^{-30}, 10^{-25}, \ldots, 10^{0}$ computed for regularization of order zero using SVD. See Appendix C for the implementation of the regularized solution based on the Sigmoid function approximation.

From Table 10, as the value of μ increases, the regularized solution deteriorates. Therefore, smaller μ values yields a good solution. The regularized solution converges at $\mu = 10^{-30}$ with the smallest error of $2.07948647190648e-009$.

Table 10: Sigmoid-SVD Regularized Solution Using 12×7 Hilbert Matrix

Parameter μ	Approximate Solution $\hat{\alpha}$	Error $\|\|\alpha_{exact} - \hat{\alpha}\|\|$
10^{-35}	0.999999999998873 1.000000000038512 0.999999999666076 1.000000001202600 0.999999997920514 1.000000001715320 0.999999999457774	$2.07948647190648e-009$
10^{-30}	0.999999999998873 1.000000000038512 0.999999999666076 1.000000001202600 0.999999997920514 1.000000001715320 0.999999999457774	$2.07948647190648e-009$
10^{-25}	0.999999999998872 1.000000000038536 0.999999999665866 1.000000001203362 0.999999997919194 1.000000001716408 0.999999999457431	$2.08080563890434e-009$
10^{-20}	0.999999999932141 1.000000002410994 0.999999978726610 1.000000077167407 0.999999866269820 1.000000110282785 0.999999965190480	$1.33730180484903e-007$
10^{-15}	1.000000154999835 0.999996947058063 1.000011476931069 1.000000897189852 0.999940285071780 1.000088199159491 0.999961952592952	$8.81991594909870e-005$

Table 10 Continued

Parameter μ	Approximate Solution $\hat{\alpha}$	Error $\|\|\alpha_{exact} - \hat{\alpha}\|\|$
10^{-10}	0.999878803391175 1.001735967604173 0.995460247265296 1.000607052908816 1.004208253857990 1.002515755068508 0.995476486511775	0.00453975273470353
10^{-5}	0.957069895104291 1.114406029985795 1.085210819207213 1.024193719073540 0.959629838376702 0.898713366587744 0.843197096416448	0.156802903583552
10^{-1}	0.1744572114330376 0.1090387435111180 0.0823690250242499 0.0670987879006640 0.0569786259823465 0.0496922582789419 0.0441564066393135	0.955843593360686
10^{0}	0.01931625248486419 0.01204744148330808 0.00909070489003618 0.00740016644625126 0.00628090172088117 0.00547565239439158 0.00486422042842840	0.995135779571572

To implement the Modified Newton's method for sigmoid approximation. The problem is to find $\alpha \in \Re^p$ such that $\mathbf{X}\alpha = \mathbf{y}$. From Equation (4.3), the gradient of $g(\alpha)$ is given by

$$\nabla g(\alpha) = 2\mathbf{X}^T(\mathbf{X}\alpha - \mathbf{y}) + \lambda G(\alpha)$$

where $G(\alpha) = \left[(1+e^{-\kappa\alpha_1})^{-1} - (1+e^{\kappa\alpha_1})^{-1}, \cdots, (1+e^{-\kappa\alpha_p})^{-1} - (1+e^{\kappa\alpha_p})^{-1}\right]^T$

and the Hessian is given by

$$\nabla^2(g(\alpha)) = 2\mathbf{X}^T\mathbf{X} + 2\kappa\lambda\, \mathbf{h}(\alpha),$$

where $\mathbf{h}(\alpha) = \text{diag}\left[\frac{e^{\kappa\alpha_1}}{(1+e^{\kappa\alpha_1})^2}, \frac{e^{\kappa\alpha_2}}{(1+e^{\kappa\alpha_2})^2}, \cdots, \frac{e^{\kappa\alpha_p}}{(1+e^{\kappa\alpha_p})^2}\right]$.

A script-file is created in OCTAVE 3.8.2 to compute the solutions at various iterations given the regularization parameter $\mu = 10^{-16}$, with stepsize $\beta = 3$, and the parameter $\kappa = 300$ in the approximation of the sigmoid function. The result of the implementation of the algorithm for the Modified Newton's method based on sigmoid approximation is given in Table 11. See Appendix D for the implementation of the Modified Newton's method based on the Sigmoid function approximation.

Table 11: Solution of Modified Newton's Method of the Sigmoid Function Approximation at 84th Iterate

μ	$\hat{\alpha}$	$\|\alpha_{exact} - \hat{\alpha}\|$
	1.000005092022318	
	0.999820688361910	
	1.001571753611182	
10^{-16}	0.994327784550849	0.00978940876437595
	1.009789408764376	
	0.991954381867874	
	1.002532280742229	
	1.000152474616397	
	0.994647378535933	
	1.046814834231628	
10^{-15}	0.831334575553153	0.825415447248076
	1.290708563243729	
	0.761330714507619	
	1.075052243753027	
	0.999841879133815	
	1.005575703218988	
	0.951079243534985	
10^{-14}	1.176675179695010	0.866261600864291
	0.694911508033166	
	1.250858148937579	
	0.921014800285320	

From Table 11, a number of iterations are performed and the best approximate

solution is obtained at the 84th iterate. Increasing the value of the parameter from $\mu = 10^{-16}$, the solution seems to be deteriorating. The error in the computed solution corresponding to $\mu = 10^{-16}$ is

$$\|\alpha_{\text{exact}} - \hat{\alpha}\|_2 = 0.00978940876437595,$$

which is about two digits accurate. As we move away from that iterate, there is not much difference between the solutions from the 85th to the 100th iterate. Table 12 gives the best approximate solution for the Modified Newton's Method and the Regularization Method.

Table 12: Modified Newton's Method versus Regularization Method with Sigmoid Approximation

Method	μ	$\hat{\alpha}$	$\|\alpha_{\text{exact}} - \hat{\alpha}\|$
MNM	10^{-16}	1.000005092022318 0.999820688361910 1.001571753611182 0.994327784550849 1.009789408764376 0.991954381867874 1.002532280742229	$9.78940876437595e-003$
RM	10^{-30}	0.999999999998873 1.000000000038512 0.999999999666076 1.000000001202600 0.999999997920514 1.000000001715320 0.999999999457774	$2.07948647190648e-009$

From Table 12, the accuracy in the computed solution of MNM corresponding to $\mu = 10^{-16}$ is just about 3 digits. The accuracy in that of the RM is up to about 9 digits.

Cubic Hermite Approximation

The Cubic Hermite approximation is a spline where each piece is a third-degree polynomial specified in Hermite form: that is, by its values and first derivatives at the end points of the corresponding domain interval. The Hermite form of a cubic polynomial defines the polynomial $p(x)$ by specifying two distinct points $[-\gamma,\ \gamma]$, and providing values for the following four equations gives

$$\begin{bmatrix} 0 & 1 & 2\gamma & 3\gamma^2 \\ 0 & 1 & -2\gamma & 3\gamma^2 \\ 1 & \gamma & \gamma^2 & \gamma^3 \\ 1 & -\gamma & \gamma^2 & -\gamma^3 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ \gamma \\ \gamma \end{bmatrix} \tag{4.10}$$

Solving for the unknown parameters in Equation (4.10), gives

$$a_0 = \frac{\gamma}{2}, \quad a_1 = 0, \quad a_2 = \frac{1}{2\gamma}, \quad a_3 = 0.$$

Therefore,

$$P(\mathbf{x}) = \frac{\gamma}{2} + \frac{1}{2\gamma}\mathbf{x}^2.$$

To determine the best approximate solution, we first examine the nature of the plot of the absolute value function for various values of γ. The graph of the abs(x) is given in Figure 14 for various values of the parameter γ.

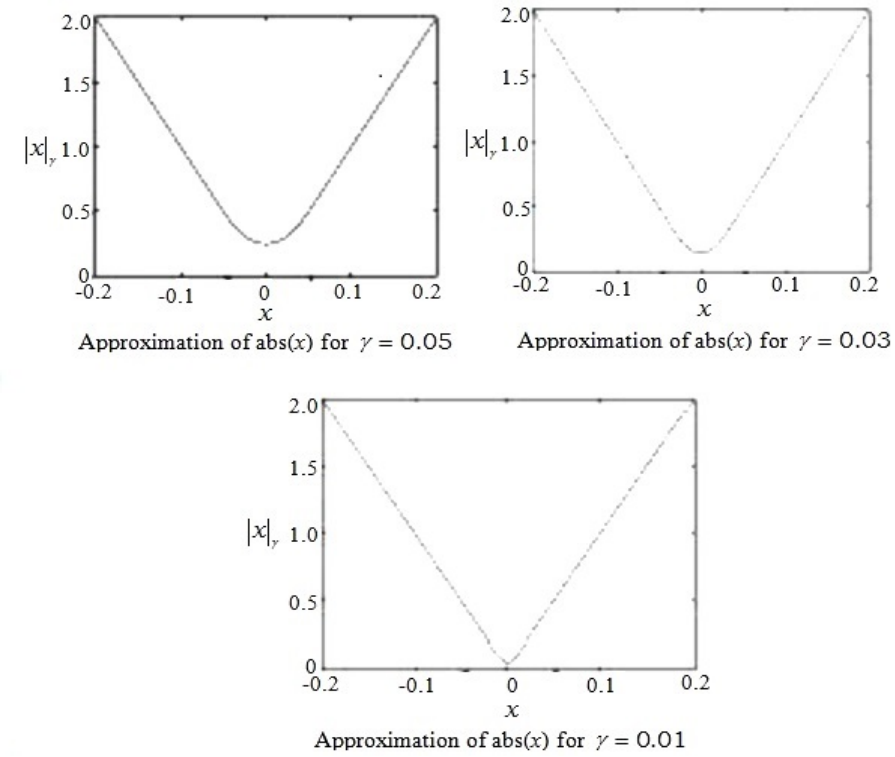

Figure 14: Cubic Hermite Approximation of $|x|_\gamma$ for Various Values of Approximating Parameter, γ.

From Figure 14, $|x|_\gamma \to |x|$ as $\gamma \to 0$. That is,

$$\lim_{\gamma \to 0} |x|_\gamma = |x|.$$

Thus, it would be suitable to choose $\gamma = 0.05$.

It will be shown that a scalar Cubic Hermite approximation to the absolute value function is given as

$$|x|_\gamma \approx \frac{\gamma}{2} + \frac{1}{2\gamma}x^2.$$

The gradient $\nabla(|x|_\gamma)$ and the Hessian $\nabla^2(|x|_\gamma)$ are derived as follows:

$$\nabla(|x|_\gamma) = \frac{x}{\gamma} \quad \text{and} \quad \nabla^2(|x|_\gamma) = \frac{1}{\gamma}.$$

For $\mathbf{x} \in \Re^p$,

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{p} |x_i| \approx \sum_{i=1}^{p} |x_i|_\gamma.$$

The loss function in Equation (4.3) therefore becomes,

$$g(\alpha) = \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda \sum_i^p \left(\frac{\gamma}{2} + \frac{1}{2\gamma}\alpha_i^2\right). \tag{4.11}$$

The regularized solution is given as

$$\alpha_\mu = (\mathbf{X}^T\mathbf{X} + \mu\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}, \tag{4.12}$$

where $\mu = \frac{1}{2\gamma}\lambda$.

The regularized solution α_μ written in terms of singular value decomposition (SVD) is given in component form as

$$\alpha_\mu = \sum_{i=1}^p \frac{\sigma_i}{\sigma_i^2 + \mu}(\mathbf{U}_i^T\mathbf{y})\mathbf{V}_i.$$

Proof

Let

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T.$$

Now, from Equation (4.1), we have

$$\begin{aligned}
\alpha_\mu &= \left[(\mathbf{U}\mathbf{S}\mathbf{V}^T)^T\mathbf{U}\mathbf{S}\mathbf{V}^T + \mu\mathbf{I}\right]^{-1}(\mathbf{U}\mathbf{S}\mathbf{V}^T)^T\mathbf{y} \\
&= \left[(\mathbf{V}\mathbf{S}\mathbf{U}^T)(\mathbf{U}\mathbf{S}\mathbf{V}^T) + \mu\mathbf{I}\right]^{-1}(\mathbf{V}\mathbf{S}\mathbf{U}^T)\mathbf{y} \\
&= \left[\mathbf{V}\mathbf{S}^2\mathbf{V}^T + \mu\mathbf{I}\right]^{-1}(\mathbf{V}\mathbf{S}\mathbf{U}^T)\mathbf{y} \\
&= \left[\mathbf{V}\mathbf{S}^2\mathbf{V}^T + \mu\mathbf{V}\mathbf{I}\mathbf{V}^T\right]^{-1}(\mathbf{V}\mathbf{S}\mathbf{U}^T)\mathbf{y} \\
&= \left[\mathbf{V}\left(\mathbf{S}^2 + \mu\mathbf{I}\right)\mathbf{V}^T\right]^{-1}(\mathbf{V}\mathbf{S}\mathbf{U}^T)\mathbf{y} \\
&= \mathbf{V}\left(\mathbf{S}^2 + \mu\mathbf{I}\right)^{-1}\mathbf{V}^{-1}(\mathbf{V}\mathbf{S}\mathbf{U}^T)\mathbf{y} \\
&= \left(\mathbf{S}^2 + \mu\mathbf{I}\right)^{-1}(\mathbf{V}\mathbf{S}\mathbf{U}^T)\mathbf{y}.
\end{aligned}$$

Therefore, the singular value decomposition solution in component form is given as

$$\alpha_\mu = \sum_{i=1}^p \frac{\sigma_i}{\sigma_i^2 + \mu}(\mathbf{U}_i^T\mathbf{y})\mathbf{V}_i$$

which ends the proof.

Table 13 shows the solutions corresponding to each μ_i for $\mu = 10^{-35}, 10^{-30}, 10^{-25}, \ldots, 10^{0}$ for regularization of order zero using SVD.

Table 13: Cubic Hermite-SVD Regularized Solution Using 12×7 Hilbert Matrix

Parameter μ	Approximate Solution $\hat{\alpha}$	Error $\|\|\alpha_{exact} - \hat{\alpha}\|\|$
10^{-35}	0.999999999998873 1.000000000038512 0.999999999666076 1.000000001202600 0.999999997920514 1.000000001715320 0.999999999457774	$2.07948647190648e-009$
10^{-30}	0.999999999998873 1.000000000038512 0.999999999666076 1.000000001202600 0.999999997920514 1.000000001715320 0.999999999457774	$2.07948647190648e-009$
10^{-25}	0.999999999998873 1.000000000038513 0.999999999666067 1.000000001202631 0.999999997920461 1.000000001715363 0.999999999457760	$2.07953931852245e-009$
10^{-20}	0.999999999996198 1.000000000133594 0.999999998826900 1.000000004246953 0.999999992644548 1.000000006066182 0.999999998084514	$7.35545158114803e-009$

Table 13 Continued

Parameter μ	Approximate Solution $\hat{\alpha}$	Error $\|\|\alpha_{exact} - \hat{\alpha}\|\|$
10^{-15}	0.999999976883887 1.000000914030655 0.999991353305120 1.000032960674909 0.999940702249965 1.000050356850425 0.999983724337111	$5.92977500348812e-005$
10^{-10}	0.999977802872302 1.000254859290361 0.999575004547348 0.999462342623030 1.000769728470951 1.001050955585264 0.998892184490866	0.00110781550913430
10^{-5}	0.997083859443683 0.988757350223393 1.027233238770432 1.030059635253386 1.011348445727649 0.981955740065443 0.947888075611324	0.0521119243886761
10^{-1}	1.182764491144766 0.778733472069915 0.603822066884575 0.499971946128715 0.429416918328446 0.377683032353304 0.337821193887241	0.662178806112759
10^{0}	0.3753186032580785 0.2354033817425414 0.1781502630022617 0.1452917881795200 0.1234792009450131 0.1077549779624429 0.0957968325361443	0.904203167463856

From Table 13, as the values of μ increase, the regularized solution deteriorates. Therefore, smaller μ values yields a good solution. The regularized solution converges at $\mu = 10^{-30}$ with the smallest error of $2.07948647190648e-009$. It can be observed that as we reduce the value of μ further, the error remains the same. Thus, the solution converges at $\mu = 10^{-30}$. See Appendix E for the implementation of the regularized solution of the Cubic Hermite approximation.

To implement the Modified Newton's method for Cubic Hermite approximation, we want to find $\alpha \in \Re^p$ such that $\mathbf{X}\alpha = \mathbf{y}$. From Equation (4.11), the gradient of $g(\alpha)$ is given by

$$\nabla g(\alpha) = 2\mathbf{X}^T(\mathbf{X}\alpha - \mathbf{y}) + \lambda G(\alpha),$$

where

$G(\alpha) = (\frac{\alpha_1}{\gamma}, \frac{\alpha_2}{\gamma}, \ldots, \frac{\alpha_p}{\gamma})^T$

and the Hessian is also given as

$$\nabla^2(g(\alpha)) = 2\mathbf{X}^T\mathbf{X} + \frac{1}{\gamma}\lambda\mathbf{I}_p.$$

The result of the implementation of the algorithm of the Modified Newton's method based on Cubic Hermite approximation is given in Table 14.

A script-file is created in OCTAVE 3.8.2 to compute the solutions at various iterations given the regularization parameter $\mu = 10^{-16}$, with the step size $\beta = 3$, and the parameter $\gamma = 0.05$ in the cubic Hermite approximation. A number of iterations are performed and the best approximate solution is obtained at the 86th iterate. Table 14 shows the various solutions using various values of the parameter μ. See Appendix F for the implementation of the Modified Newton's method based on the Cubic Hermite approximation.

Table 14: Solution of Modified Newton's Method of the Cubic Hermite Approximation at 86th Iterate

μ	$\hat{\alpha}$
	0.999983772069521
	1.000576215743305
	0.994919717190969
10^{-16}	1.018414062557087
	0.968111442843396
	1.026280481466755
	0.991709642462265
	0.999655665671350
	1.012136463880960
	0.893550223924022
10^{-15}	1.384344473224146
	0.336429480638737
	1.545533573625836
	0.828255590349513
	0.998566256577911
	1.050515505579027
	0.557039423069666
10^{-14}	2.599028960108513
	-1.760299621017860
	3.269011614438952
	0.285745185125661

From Table 14, there is not much difference between the solutions from

the 87th to the 100th iterate. Increasing the value of the parameter from $\mu = 10^{-16}$, the solution deteriorate from the true solution. The error in the computed solution corresponding to $\mu = 10^{-16}$ is

$$\|\alpha_{\text{exact}} - \hat{\alpha}\|_2 = 0.0318885571566037,$$

which is about one digit accurate.

Table 15 gives the best approximate solution for the Modified Newton's method and the Regularization Method.

Table 15: Modified Newton's Method versus Regularization Method with Cubic Hermite Approximation

Method	μ	$\hat{\alpha}$	$\|\alpha_{\text{exact}} - \hat{\alpha}\|$
MNM	10^{-16}	0.999983772069521 1.000576215743305 0.994919717190969 1.018414062557087 0.968111442843396 1.026280481466755 0.991709642462265	$3.18885571566037e-002$
RM	10^{-30}	0.999999999998873 1.000000000038512 0.999999999666076 1.000000001202600 0.999999997920514 1.000000001715320 0.999999999457774	$2.07948647190648e-009$

From the numerical simulations, the Modified Newton's method solutions vary for the three smoothing approximations considered. However, the regularized solutions are the same at $\mu = 10^{-30}$.

The relation among the parameters of the Tikhonov regularization and the smoothing approximations considered are summarized in Table 16.

Table 16 : Relation among Parameters of Tikhonov Regularization and Smoothing Approximations

Method	Parameter of Smoothing Method	Regularization Parameter	Parameter in terms of λ
Tikhonov	-	λ	λ
Quadratic	ε	μ	$\mu = \frac{1}{2}\lambda\varepsilon^{-\frac{1}{2}}$
Sigmoid	κ	μ	$\mu = \frac{1}{4}\lambda\kappa$
Cubic	γ	μ	$\mu = \frac{1}{2\gamma}\lambda$

In Table 16, λ is the regularization parameter in Equation (4.1). Column 4 of the table gives the relation for the regularization parameter μ in the minimization problem in terms of the smoothing approximation of the L_1-norm penalty. We now compare the three smoothing approximations using regularization with a non-smooth method which makes use of Truncated Newton Interior-Point method described by Kim et al., (2007). In that paper, they developed a MATLAB Solver for large-scale L_1-regularized least squares problems called $l1\_ls$.

Using our value of the parameter $\mu = 10^{-30}$ in the $l1\_ls$, we display in Table 17 the result of all three regularization methods (RM) and that of the $l1\_ls$.

Table 17: Summary of Methods and their Solutions at $\mu = 10^{-30}$

Method	Solution corresponding to $\mu = 10^{-30}$	$\|\alpha_{exact} - \hat{\alpha}\|$
RM	0.999999999998873 1.000000000038512 0.999999999666076 1.000000001202600 0.999999997920514 1.000000001715320 0.999999999457774	$2.07948647190648e-009$
l1_ls	1.000000000000292 0.999999999990242 1.000000000081881 0.999999999715937 1.000000000472646 0.999999999624568 1.000000000114463	$4.72645700355656e-010$

From Table 17, it is seen that the solutions from the three smoothing approximations by regularization method is as good as the non-smooth method, which is accurate to about ten digits.

Chapter Summary

In this chapter, we have considered three smoothing methods for approximating the L_1-norm penalty in the L_1-norm regularized least squares problem. It is an attempt to obtain an approximate differentiable function for the non-differentiable L_1-norm penalty term. The three methods considered are Quadratic approximation of the Lee et al. approximation, the Sigmoid function approximation, and the Cubic Hermite. For each approximation, we have obtained the regularized solution and the corresponding solution in terms of the singular value decomposition. In each case, the result of the approximation has be assessed using the Hilbert sub-matrix of dimension 12×7. The regularized smoothing

approximation solution produced almost the same results that are accurate to nine digits at the same parameter value of $\mu = 10^{-30}$. Subsequently, for each of the three methods, we have obtained optimal approximate solutions by means of the Newton's method. It is observed that the results of the Newton's method under all three methods show visible differences and produced solutions that are accurate only to at most three digits.

Therefore, our results from the smoothing approximation to the L_1-norm regularization functional are as good as the non-smooth methods used in developed solvers.

CHAPTER FIVE

NON-SMOOTHING OPTIMIZATION WITH SPARSITY INDUCED SYSTEMS

Introduction

Least squares optimization with L_1-norm regularization can be cast as an unconstrained problem by finding a good smoothing approximation to the L_1-norm which have been discussed in Chapter Four.

In this chapter, we consider another way of addressing the same problem, this time by considering a non-smoothing approximation of the non-differentiable L_1-norm penalty term. Least squares minimization subject to an L_1-norm was presented and popularised independently under the names "Least Absolute Shrinkage and Selection Operator" (LASSO) by Tibshirani, (1994). This acronym has become a dominant expression describing the L_1-norm function. Tibshirani, (1994) presented several different methods for optimizing the LASSO. The 'LASSO' minimizes the residual sum of squares (RSS) subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly zero and hence gives interpretable models.

While L_2-norm regularization is an effective means of achieving numerical stability and increasing predictive performance, it does not address a problem with least squares estimates, which does not ensure parsimony of the model and interpretability of the coefficients. While the size of the coefficient values is bounded, minimizing the RSS with a penalty on the L_2-norm does not encourage sparsity, and the resulting models typically have non-zero values associated with all coefficients. It has been proposed that, rather than simply achieving the goal of 'shrinking' the coefficients, higher values for the L_2-norm penalty force the coefficients to be more similar to each other in order to minimize their joint

L_2-norm. An alternative approach has been to replace the L_2-norm penalty with an L_1-norm. This L_1-norm regularization has many of the beneficial properties of L_2-norm regularization, but yields sparse models that are more easily interpreted. An additional advantage of L_1-norm penalties is that the models often outperform those produced with an L_2-norm penalty, when irrelevant features are present in the solution α. This property provides an alternate motivation for the use of an L_1-norm penalty. It provides a regularized feature selection method, and thus can give low variance feature selection, compared to the high variance performance of typical subset selection techniques.

The LASSO for Linear Models

In this section, we introduce the LASSO estimator for linear regression. We describe the basic LASSO method, and outline a simple approach for its implementation. We relate the LASSO to ridge regression, and also view it as a Bayesian estimator.

In the linear regression setting, we are given m samples $(x_i, y_i)_{i=1}^m$, where each $x_i = (x_{i1}, \ldots, x_{ip})$ is a p-dimensional vector of features or predictors, and each $y_i \in \Re$ is the associated response variable. Our goal is to approximate the response variable y_i using a linear combination of the predictors

$$\eta(x_i) = \alpha_0 + \sum_{j=1}^{p} x_{ij}\alpha_j. \tag{5.1}$$

The model is parametrised by the vector of regression weights $\alpha = (\alpha_1, \ldots, \alpha_p) \in \Re^p$ and an intercept (or "bias") term $\alpha_0 \in \Re$. The usual "least squares" estimator for the pair (α_0, α) is based on minimizing squared-error loss given by

$$\underset{\alpha_0,\alpha}{\text{minimize}} \left\{ \sum_{i=1}^{m} (y_i - \alpha_0 - \sum_{j=1}^{p} x_{ij}\alpha_j)^2 \right\}. \tag{5.2}$$

There are two reasons why we might consider an alternative to the least-squares

estimate. The first reason is prediction accuracy: the least squares estimate often has low bias but large variance, and prediction accuracy can sometimes be improved by shrinking the values of the regression coefficients, or setting some coefficients to zero. By doing so, we introduce some bias but reduce the variance of the predicted values, and hence may improve the overall prediction accuracy (as measured in terms of the mean-squared error). The second reason is for the purposes of interpretation. With a large number of predictors, we often would like to identify a smaller subset of these predictors that exhibit the strongest effects.

This chapter is devoted to discussion of the LASSO, a method that combines the least squares loss in Equation (5.2) with an L_1-constraint, or bound on the sum of the absolute values of the coefficients. Relative to the least squares solution, this constraint has the effect of shrinking the coefficients, and even setting some to zero. In this way, it provides an automatic way for doing feature selection in linear regression. Moreover, unlike some other criteria for feature selection, the resulting optimization problem is convex, and can be solved efficiently for large problems.

The LASSO Estimator

Given a collection of m predictor-response pairs $(x_i, y_i)_{i=1}^m$, the LASSO finds the solution $(\hat{\alpha_0}, \hat{\alpha})$ to the optimization problem

$$\min_{\alpha_0,\alpha} \left\{ \sum_{i=1}^{m} (y_i - \alpha_0 - \sum_{j=1}^{p} x_{ij}\alpha_j)^2 \right\} \tag{5.3}$$

$$\text{subject to} \quad \sum_{j=1}^{p} |\alpha_j| \leq t.$$

The constraint $\sum_{j=1}^{p} |\alpha_j| \leq t$ can be written more compactly as the L_1-norm constraint $\|\alpha_j\| \leq t$. Furthermore, Equation (5.3) is often represented using matrix-vector notation. Let $\mathbf{y} = (y_1, \ldots, y_m)$ denote the m-vector of responses, and $\mathbf{X}$

be an $m \times p$ matrix with $x_i \in \Re^p$ in its ith row, then the optimization problem in Equation (5.3) can be rewritten as a minimization problem of the form

$$\min_{\alpha_0,\alpha} \left\{ \|\mathbf{y} - \alpha_0\mathbf{1} - \mathbf{X}\alpha\|_2^2 \right\} \tag{5.4}$$

$$\text{subject to} \quad \left\|\alpha_j\right\| \leq t,$$

where $\mathbf{1} = \{\mathbf{1}, \mathbf{1}, \ldots, \mathbf{1}\} \in \Re^{\mathbf{m}}$ is the vector of m ones, and $\|.\|_2$ denotes the usual Euclidean norm on vectors. The bound t is a kind of "budget": it limits the sum of the absolute values of the parameter estimates. Since a shrunken parameter estimate corresponds to a more heavily-constrained model, this budget limits how well we can fit the data. It must be specified by an external procedure such as cross-validation, which we discuss later in the chapter. Typically, we first standardize the data matrix $\mathbf{X}$ so that each column is centered with mean zero and variance one. These centering conditions are convenient, since they mean that we can omit the intercept term α_0 in the LASSO optimization. Given an optimal LASSO solution $\hat{\alpha}$ on the centered data, we can recover the optimal solutions for the un-centered data: $\hat{\alpha}$ is the same, and the intercept $\hat{\alpha}_0$ is given by

$$\hat{\alpha_0} = \bar{y} - \sum_{j=1}^{p} \bar{x_j}, \hat{\alpha_j},$$

where $\bar{y}$ and $\bar{x_j}_1^p$ are the original means. For this reason, we omit the intercept α_0 from the LASSO for the rest of this chapter. It is often convenient to rewrite the LASSO problem in the Lagrangian form as

$$\underset{\alpha \in \Re^p}{\text{minimize}} \left\{ \|\mathbf{y} - \mathbf{X}\alpha\|_2^2 + \lambda \|\alpha\|_1 \right\}, \tag{5.5}$$

for some $\lambda \geq 0$. By Lagrangian duality, there is a one-to-one correspondence between the constrained problem in Equation (5.3) and the Lagrangian form in Equation (5.5): for each value of t in the range where the constraint $\|\alpha\|_1 \leq t$ is active, there is a corresponding value of λ that yields the same solution from the Lagrangian form in Equation (5.5). Conversely, the solution $\hat{\alpha_\lambda}$ to the problem

in Equation (5.5) solves the bound problem with $t = \|\hat{\alpha_\lambda}\|_1$. There are many descriptions for the LASSO, the factor 1 appearing in Equation (5.3) and Equation (5.5) can be replaced by $\frac{1}{2m}$ or $\frac{1}{2}$. Although this makes no difference in Equation (5.3), it corresponds to a simple re-parametrisation of λ in Equation (5.5), this kind of standardisation makes λ values comparable for different sample sizes (useful for cross-validation). The theory of convex analysis tells us that necessary and sufficient conditions for a solution to problem in Equation (5.5) take the form

$$-\langle x_j,\ \mathbf{y} - \mathbf{X}\alpha\rangle + \lambda s_j = 0, \quad j = 1,\ \ldots,\ n. \tag{5.6}$$

Here, each s_j is an unknown quantity equal to $\text{sign}(\alpha_j)$, if $\alpha_j \neq 0$ and some value lying in $[-1,\ 1]$ otherwise, it is a sub-gradient for the absolute value function (as described in Chapter Three). In other words, the solutions $\hat{\alpha}$ to problem in Equation (5.5) are the same as solutions $(\hat{\alpha}, \hat{s})$ to Equation (5.6). This system is a form of the Karush Kuhn Tucker (KKT) conditions for problem in Equation (5.5). Expressing a problem in sub-gradient form can be useful for designing algorithms for finding its solutions.

As an example of the LASSO, let us consider the data given in Table 18, taken from Thomas (1990) on reported violent crime rate data.

Table 18: Violent Crime Rate and Predictors for 50 U.S.A Cities

city	funding	hs	not-hs	college	college4	violent crime rate
1	40	74	11	31	20	184
2	32	72	18	43	18	213
3	57	70	11	16	16	347
4	31	71	11	25	19	565
5	67	72	9	29	24	327
⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	66	67	26	18	16	1244

Source: Thomas (1990)

The outcome is the reported violent crime rate per $100,000$ residents in 50 U.S.A cities. There are five predictors: annual police funding in dollars per resident, percent of people 25 years and older with four years of high school, percent of 16- to 19-year olds not in high school and not high school graduates, percent of 18- to 24-year olds in college, and percent of people 25 years and older with at least four years of college. This small example is for illustration only, but helps to demonstrate the nature of the LASSO solutions. Typically, the LASSO is most useful for much larger problems, including "wide" data for which $p \gg m$.

The subplot (a) of Figure 15 shows the result of applying the LASSO with the bound t varying from zero on the left, all the way to a large value on the right, where it has no effect. The horizontal axis has been scaled so that the maximal bound, corresponding to the least squares estimates $\tilde{\alpha}$, is one. We see that for much of the range of the bound, many of the estimates are exactly zero, and hence, the corresponding predictor(s) would be excluded from the model. The LASSO has this feature selection property due to the geometry that underlies

the L_1 constraint $\|\alpha\|_1 \leq t$. To understand this better, the subplot (b) shows the estimates from ridge regression also known as L_2-norm regularization discussed in Chapter Three. This technique predates the LASSO. It solves a criterion very similar to Equation (5.3). This is given as

$$\min_{\alpha_0,\,\alpha} \left\{ \sum_{i=1}^{m} \left(y_i - \alpha_0 - \sum_{j=1}^{p} x_{ij}\alpha_j \right)^2 \right\} \tag{5.7}$$

$$\text{subject to } \sum_{j=1}^{p} \alpha_j^2 \leq t^2$$

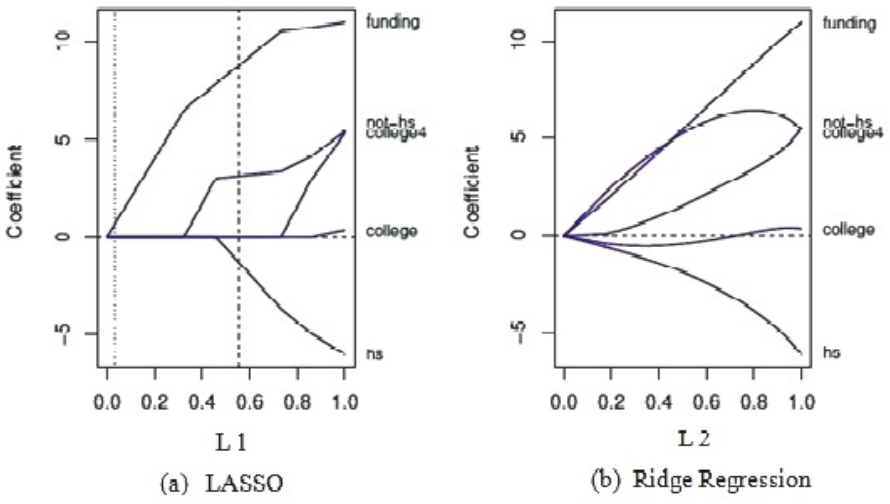

Figure 15: Coefficient Path for LASSO and Ridge Regression.

The ridge profiles in subplot (b) have roughly the same shape as the LASSO profiles in (a), but are not equal to zero except at the left end. Figure 16 contrasts the two constraints used in the LASSO and ridge regression by Hastie, Tibshirani and Friedman (2009).

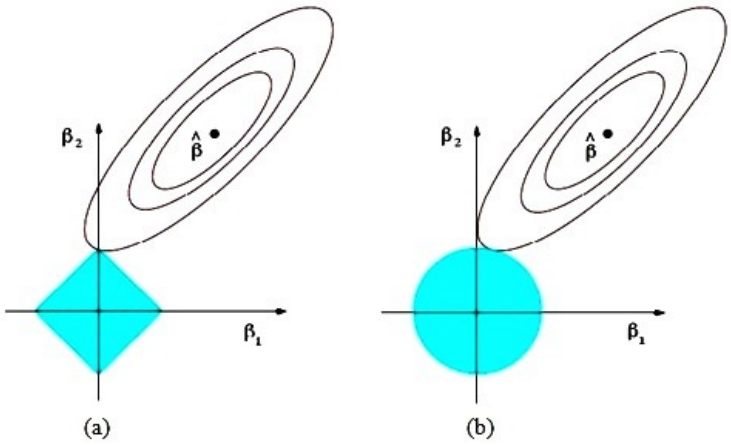

Figure 16: Graphs showing (a) L_1 Constraint and (b) L_2 Constraint.

The residual sum of squares has elliptical contours, centered at the full least squares estimates. The constraint region for ridge regression is the disk $\beta_1^2 + \beta_2^2 \leq t^2$, in (b) while that for LASSO is the diamond $|\beta|_1 + |\beta|_2 \leq t$, in (a). Both methods find the first point where the elliptical contours hit the constraint region. Unlike the disk, the diamond has corners; if the solution occurs at a corner, then it has one parameter j equal to zero. When $p > 2$, the diamond becomes a rhomboid, and has many corners, flat edges, and faces; there are many more opportunities for the estimated parameters to be zero. We use the term sparse for a model with few non-zero coefficients. Hence, a key property of the L_1 constraint is its ability to yield sparse solutions. This idea can be applied in many different statistical models, and is the central part of this thesis.

LASSO Solution

Obtaining a regression model for a dataset $\mathbf{X}$ $(m \times p)$ usually encounters challenges when $p > m$. In this case, rank$(\mathbf{X}) < p$. This condition produces a non-unique solution such that there is a variable $x_i, \in \{1, 2, \ldots, p\}$ whose coef-

ficient is positive on one solution and negative on another. Adding an element of null space of $\mathbf{X}$ to a least squares solution produces another solution. This phenomena makes the interpretation of the regression parameters generally problematic. The main problem here is that there are inconsistencies in signs of the variables on different solutions.

Unconstrained Formulation of LASSO

Formulation of the LASSO can be cast as either constrained or unconstrained. The two formulations are equivalent. The proof is given as follows.

Proof

The unconstrained formulation of the LASSO is given as

$$\min_{\alpha} \mathbf{g}\,(\alpha,\, \lambda_1) = \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda_1 \|\alpha\|_1 \tag{5.8}$$

whiles the constrained formulation is given as

$$\min_{\alpha} \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 \tag{5.9}$$

$$\text{subject to} \|\alpha\|_1 \leq t.$$

The Lagrangian for Equation (5.9) is given by

$$\mathbf{g}(\alpha,\, \lambda_2) = \|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda_2(\|\alpha\|_1 - t). \tag{5.10}$$

Now, Equation (5.10) is equivalent to Equation (5.8) except for the constant term $-\lambda_2 t$. The necessary conditions for optimality for the problem in Equation (5.8) is given by

$$\nabla_\alpha \mathbf{g}_{\lambda_1}(\alpha^*) = 0 = \nabla_\alpha \mathbf{g}_{\lambda_2}(\alpha^*),$$

where $\alpha^*(\lambda_1)$ is the optimal solution for a given λ_1, (see Equation (5.11)). Therefore, the gradient of Equation (5.10) is given as

$$\nabla_{\alpha}\mathbf{g}(\alpha^*,\lambda_2^*) \Rightarrow 2\mathbf{X}^T(\mathbf{X}\alpha-\mathbf{y})+\lambda_2 \text{sign}(\alpha)=0$$

whereas that for Equation (5.8) is given as

$$\nabla_{\alpha}\mathbf{g}(\alpha^*,\lambda_1^*) \Rightarrow 2\mathbf{X}^T(\mathbf{X}\alpha-\mathbf{y})+\lambda_1 \text{sign}(\alpha)=0.$$

The sign(α) is the sub-gradient of the L_1-norm penalty. The KKT conditions imply that we have:

$$\begin{aligned} \nabla_{\alpha}\mathbf{g}_{\lambda_2}(\alpha^*) &= \nabla_{\alpha}\mathbf{g}_{\lambda_1}(\alpha^*)=0 \\ \lambda_2^*(\|\alpha\|_1-t) &= 0. \end{aligned}$$

What this means is that the gradient of the Lagrangian with respect to α should be null and the other is the complementary condition. We observe that the gradient of the Lagrangian is equal to the gradient of $\mathbf{g}_{\lambda_2}$, that is, the objective function in Equation (5.8) but with λ_2 instead of λ_1. Now let suppose we solve the problem in Equation (5.8) for a given λ_1 and obtain its solution as $\alpha^*(\lambda_1)$ and also let $t=\|\alpha^*(\lambda_1)\|_1$, the L_1-norm of the solution to the problem in Equation (5.8), then $\lambda_2^*=\lambda_1$ and $\alpha^*=\alpha^*(\lambda_1)$ satisfy the KKT conditions for the problem in Equation (5.9), showing that both the problems have the same solution.

Conversely, by setting $\lambda_1=\lambda_2^*$ in Equation (5.10), we can retrieve the same solution by solving the problem in Equation (5.8). Therefore, both problems are equivalent when $t=\|\alpha^*(\lambda_1)\|_1$.

Linear Regression Using LASSO

We consider an L_1-norm regularization functional in Equation (5.8) which we refer to as Method 1.

The gradient of $g(\alpha)$ is given by

$$\begin{aligned}\nabla g_i(\alpha) &= 2\mathbf{X}^T(\mathbf{X}\alpha-\mathbf{y})+\lambda\sum_{i=1}^{p}|\alpha_i| \\ &= (2\mathbf{X}^T\mathbf{X}\alpha)_i-(2\mathbf{X}^T\mathbf{y})_i+\lambda_i\text{sign}(\alpha_i),\end{aligned}$$

where

$$\text{sign}(\alpha_i)=\begin{cases}+1 & \text{if } \alpha_i>0\\ -1 & \text{if } \alpha_i<0\\ [-1,\ +1] & \text{if } \alpha_i=0.\end{cases}$$

Then,

$$\nabla g_i(\alpha)=\begin{cases}(2\mathbf{X}^T\mathbf{X}\alpha)_i-(2\mathbf{X}^T\mathbf{y})_i+\lambda_i & \text{if} \quad \alpha_i>0\\ (2\mathbf{X}^T\mathbf{X}\alpha)_i-(2\mathbf{X}^T\mathbf{y})_i-\lambda_i & \text{if} \quad \alpha_i<0\\ \Big[-(2\mathbf{X}^T\mathbf{y})_i-\lambda_i,\ -(2\mathbf{X}^T\mathbf{y})_i+\lambda_i\Big] & \text{if} \quad \alpha_i=0.\end{cases}$$

We consider each of the three cases of $\nabla g_i(\alpha)$. In the following, $\mathbf{I}_i$ is the ith column of the $p\times p$ identity matrix.

Case 1: When $\alpha_i>0$.

By first equating the gradient to zero,

$$(2\mathbf{X}^T\mathbf{X}\alpha)_i-(2\mathbf{X}^T\mathbf{y})_i+\lambda_i=0.$$

This implies that

$$\alpha_i=\frac{1}{2}(\mathbf{X}^T\mathbf{X})^{-1}(2\mathbf{X}^T\mathbf{y}-\lambda\mathbf{I}_i).$$

Then applying the condition $\alpha_i>0$, we obtain

$$\begin{aligned}&(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y})-\frac{1}{2}(\mathbf{X}^T\mathbf{X})^{-1}\lambda\mathbf{I}_i>0\\ &(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y})>\frac{1}{2}(\mathbf{X}^T\mathbf{X})^{-1}\lambda\mathbf{I}_i\\ &(\mathbf{X}^T\mathbf{y})>\frac{1}{2}\lambda\mathbf{I}_i\\ &2(\mathbf{X}^T\mathbf{y})_i>\lambda_i.\end{aligned}$$

Case 2: When $\alpha_i < 0$.

By equating the corresponding gradient to zero,

$$(2\mathbf{X}^T\mathbf{X}\alpha)_i - (2\mathbf{X}^T\mathbf{y})_i - \lambda_i = 0,$$

which implies that

$$\alpha_i = \frac{1}{2}(\mathbf{X}^T\mathbf{X})^{-1}(2\mathbf{X}^T\mathbf{y} + \lambda\mathbf{I}_i).$$

Then applying the condition $\alpha_i < 0$, we obtain

$$\begin{aligned}
&(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y}) + \frac{1}{2}(\mathbf{X}^T\mathbf{X})^{-1}\lambda\mathbf{I}_i < 0 \\
&(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y}) < -\frac{1}{2}(\mathbf{X}^T\mathbf{X})^{-1}\lambda\mathbf{I}_i \\
&(\mathbf{X}^T\mathbf{y}) < -\frac{1}{2}\lambda\mathbf{I}_i \\
&2(\mathbf{X}^T\mathbf{y})_i < -\lambda_i.
\end{aligned}$$

Case 3: When $\alpha_i = 0$.

By equating the corresponding gradient to zero,

$$\left[-(2\mathbf{X}^T\mathbf{y})_i - \lambda_i,\ -(2\mathbf{X}^T\mathbf{y})_i + \lambda_i\right] = 0$$

This implies that

$$-\lambda_i < 2(\mathbf{X}^T\mathbf{y})_i < \lambda_i.$$

Combining the solutions in the three cases, we obtain the composite solution as follows

$$\alpha_i = \begin{cases} \frac{1}{2}(\mathbf{X}^T\mathbf{X})^{-1}(2\mathbf{X}^T\mathbf{y} - \lambda\mathbf{I}_i), & \lambda_i < 2(\mathbf{X}^T\mathbf{y})_i \\ \frac{1}{2}(\mathbf{X}^T\mathbf{X})^{-1}(2\mathbf{X}^T\mathbf{y} + \lambda\mathbf{I}_i), & -\lambda_i > 2(\mathbf{X}^T\mathbf{y})_i \\ 0, & -\lambda_i < 2(\mathbf{X}^T\mathbf{y})_i < \lambda_i. \end{cases} \tag{5.11}$$

We will subsequently refer to Equation (5.11) as Method 1.

In Equation (5.11), a condition for the value of α_i depends on the value $2(\mathbf{X}^T\mathbf{y})_i$. We relate this value to the KKT condition which is given by

$$\mathbf{X}_j^T(\mathbf{y} - \mathbf{X}\hat{\alpha}) = \lambda_j\gamma_j,$$

Notice that the inner product of $\mathbf{X}_j^T$ (which is the jth column of the data matrix $\mathbf{X}$) and the residual $\mathbf{r}_j = (\mathbf{y} - \mathbf{X}\hat{\alpha})_j$ is the left hand side of the KKT condition. Thus, we can express

$$2(\mathbf{X}^T\mathbf{y})_j = 2\mathbf{X}_j^T(\mathbf{y} - \mathbf{X}\hat{\alpha}|_{\alpha_j=0})_j,$$

which may be simplified as

$$2(\mathbf{X}^T\mathbf{y})_j = 2(\mathbf{X}^T\mathbf{y})_j - 2(\mathbf{X}^T\mathbf{X})_j\alpha + 2(\mathbf{X}^T\mathbf{X})_{(j,\,j)}\alpha_j. \tag{5.12}$$

In Equation (5.12), $(\mathbf{X}^T\mathbf{X})_{(j,\,j)}$ is the jth diagonal element of $\mathbf{X}^T\mathbf{X}$ and $(\mathbf{X}^T\mathbf{X})_j$ is the jth row. Using this relation, we develop an algorithm for finding the solution to Equation (5.11).

Algorithm for Method 1

```
Initialize regularization parameter lambda;
alpha=zeros(n,1);
xy = x′*y*2;
xx = x′*x*2;
alphaold=alpha;
for j=1:n
cj=xy(j)-sum(xx(j,  :)* alpha)+xx(j, j)*alpha(j);
aj=xx(j, j);
if cj >lambda
alpha(j,1)=(cj-lambda)/aj;
elseif cj <-lambda
alpha  (j, 1)= (cj + lambda)/aj;
else
alpha  (j, 1)= 0;
end
end
```

In the algorithm, to ensure that $\alpha_j = 0$, we set an initial solution to be $\alpha = 0$.

Applications

Using an L_1-norm regularization has been found to be particularly effective in many other applications like feature selection by Tibshirani (1996), compressed sensing by Chen et al. (1998), sparse coding by Gregor and LeCun (2010), and discovery of graph connectivity by Hsieh, Sustik, Dhillon and Ravikumar (2011). In this section, we will focus on demonstrating the application of least squares with regularization to data fitting. Many problems in machine learning and data fitting can be cast as a least squares problem with a regularization term in order to limit over-fitting. The model is stated as the LASSO problem in Equation (1.13). Apart from limiting over-fitting, the L_1 regularizer tends to produce a sparse solution while reducing the computational cost.

Data Fitting

In Chapter One, we illustrated the need for regularization in order to obtain a good fit to a given datasets. The illustrations made use of data involving few dimensions. In this section, we use datasets of high dimensions. These datasets are those that have been introduced in Chapter One. As described in that chapter, the datasets involve both real and hypothetical ones. We will make use of a hypothetical data given by the Hilbert matrix of dimension 20×12. The real datasets are the level of atmospheric ozone concentration with dimensions 330×10 and the Housing datasets with dimensions 506×14. For the purpose of these illustrations, the selected datasets involve only continuous variables.

In each of the illustrations, we provide the solution of the least squares method, the L_2-norm regularization, the solution which is based on our derivation in Chapter Five, the *l1_ls* solver and finally the LASSO solution from an

inbuilt MATLAB command. In order to arrive at these solutions, we first obtain solutions based on Method 1 for various values of the regularization parameter. We provide a description of each of the datasets used in the thesis at the point of application.

Application to Ozone Concentration Data

The data covers the level of atmospheric ozone concentration from eight daily meteorological measurements made in the Los Angeles basin in 1976. The response, referred to as ozone, is actually the logarithm of the daily maximum of the hourly-average ozone concentrations in Upland, California. It involves 330 complete cases of measurements that were made every day that year. Thus, the data covers 330 observations on a total of ten variables which are described as follows:

Ozone : Upland Maximum Ozone

VH : Vandenberg 500 mb Height

Wind : Wind Speed (mph)

Humidity : Humidity

Temp : Sandburg AFB Temperature

IBH : Inversion Base Height

DPG : Daggot Pressure Gradient

IBT : Inversion Base Temperature

Vis : Visibility (miles)

DOY : Day of the Year

For this data, we intend to determine how well the method studied provides a good fit for the Ozone concentration in terms of the nine variables.

Table 19 gives the fit to the Ozone data for various regularization parameters and corresponding norm of the solution. We notice in Table 19 that $\|\alpha\|$ is much lower for $\lambda = 10^{-0.5}$ or higher values of λ. However, for these values the

standard method, *l*1_*ls* gives extremely low $\|\alpha\|$. The value of $\lambda = 10^{-1}$ is thus chosen so that our result and that of the standard method are quite close. By this selection, we are sure to avoid being prone to over-penalising the variables and hence obtain a more optimal solution. In addition, it can be observed that the solution begins to converge for values of λ lower than $\lambda = 10^{-1}$ with slightly higher solution norm.

Table 19: Method 1 Solution for various λ Values with Corresponding Norms for Ozone Data

Parameter λ	Method 1 α	Solution Norm $\|\|\alpha\|\|_2$
10^{1}	$1.17606060606061e+001$ $1.81349156058298e-005$ $-2.28159860683744e-002$ $1.92919896734494e-002$ $5.05882313003997e-003$ $-1.20306826495284e-003$ $5.36771906998305e-002$ $1.41708185337842e-002$ $-1.53044942009277e-002$ $2.42354956735694e-003$	11.7607864046703
10^{0}	$1.17742424242424e+001$ $1.57647849094480e-005$ $-2.32895205570284e-002$ $1.93324806257359e-002$ $5.06114647018805e-003$ $-1.20306208889537e-003$ $5.36790705621900e-002$ $1.41696660017421e-002$ $-1.53040603588725e-002$ $2.42297925201176e-003$	11.7744235604703
$10^{-0.5}$	$1.17752784427788e+001$ $1.55847149582150e-005$ $-2.33254971942704e-002$ $1.93355569133917e-002$ $5.06132298524176e-003$ $-1.20306161967131e-003$ $5.36792133841550e-002$ $1.41695784384744e-002$ $-1.53040273978557e-002$ $2.42293592247868e-003$	11.7754596399001
10^{-1}	$1.17756060606061e+001$ $1.55277718398098e-005$ $-2.33368740058939e-002$ $1.93365297209646e-002$ $5.06137880420284e-003$ $-1.20306147128963e-003$ $5.36792585484259e-002$ $1.41695507485378e-002$ $-1.53040169746670e-002$ $2.42292222047723e-003$	11.7757872770045

Table 19 Continued

Parameter λ	Method 1 α	Norm $\|\|\alpha\|\|_2$
10^{-5}	$1.17757575606061e+001$ $1.55014396877731e-005$ $-2.33421349740630e-002$ $1.93369795754445e-002$ $5.06140461651187e-003$ $-1.20306140267363e-003$ $5.36792794336968e-002$ $1.41695379439069e-002$ $-1.53040121546818e-002$ $2.42291588427374e-003$	11.7759387859221
10^{-10}	$1.17757575757574e+001$ $1.55014370543208e-005$ $-2.33421355002071e-002$ $1.93369796204340e-002$ $5.06140461909337e-003$ $-1.20306140266677e-003$ $5.36792794357855e-002$ $1.41695379426263e-002$ $-1.53040121541997e-002$ $2.42291588364006e-003$	11.7759388010743
10^{-15}	$1.17757575757576e+001$ $1.55014370542946e-005$ $-2.33421355002127e-002$ $1.93369796204345e-002$ $5.06140461909341e-003$ $-1.20306140266677e-003$ $5.36792794357855e-002$ $1.41695379426263e-002$ $-1.53040121541997e-002$ $2.42291588364006e-003$	11.7759388010745
10^{-20}	$1.17757575757576e+001$ $1.55014370542946e-005$ $-2.33421355002127e-002$ $1.93369796204345e-002$ $5.06140461909341e-003$ $-1.20306140266677e-003$ $5.36792794357855e-002$ $1.41695379426263e-002$ $-1.53040121541997e-002$ $2.42291588364006e-003$	11.7759388010745

Figure 17 displays the sequence of cross-validated mean square errors (MSE) of LASSO fit associated with each of 100 λ values. It also shows the line segments for each point that represent intervals of estimate for each MSE value. The right vertical line identifies the value of λ of about 10^{-1} as the value that minimizes the MSE. However, the left vertical line identifies the value of λ of about 10^{0} as the highest value that gives an MSE which is within one standard error of the minimium MSE. These results buttress our choice of $\lambda = 10^{-1}$ in Table 19.

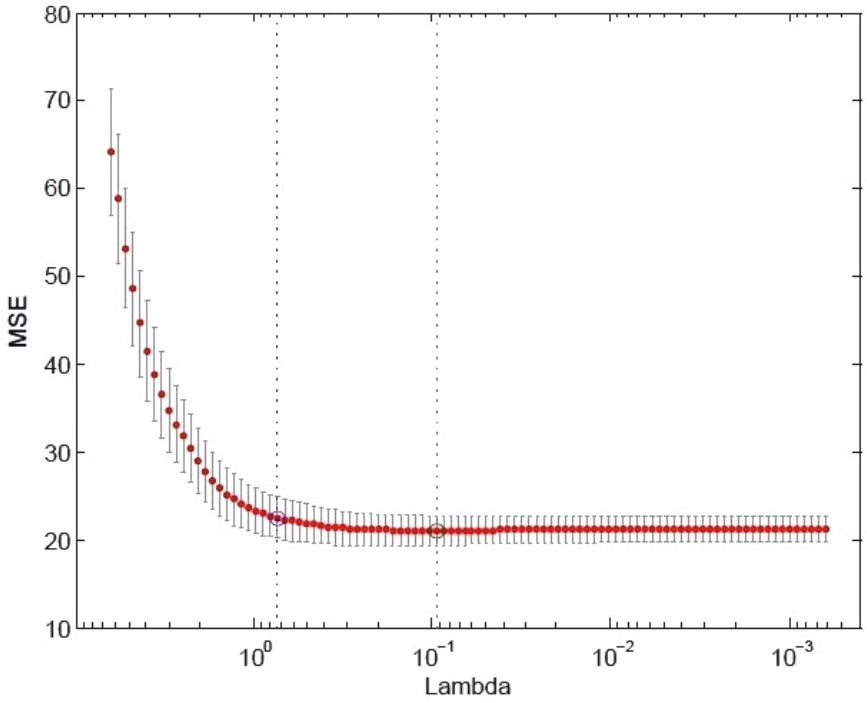

Figure 17: Cross-Validation of LASSO Fit of Ozone Data.

Table 20 gives the solution for the various methods described in the introductory part of this section for our optimal value of $\lambda = 10^{-1}$ selected in Table 19. In Table 20, it can be observed that for the selected λ, our method provides an improved solution over the least squares and also compares favourably with the standard method. The solution with the L_2-norm regularization is also given by the Ridge method. It is noted that the Ridge solution gives a smaller $\|\alpha\|$.

Under this circumstance, we will proceed to obtain a LASSO solution. It is observed that the LASSO solution sets three variables to zero and gives a smaller norm than the Ridge solution. Thus, for this data, the VH (Vandenberg 500 mb Height), Wind (Wind Speed (mph)) and DPG (Daggot Pressure Gradient) are found to be insignificant for determining level of ozone concentration.

Table 20: Solutions for Various Methods at Optimal Value of $\lambda = 10^{-1}$ of Ozone Data

	Least Squares	*l1_ls*	Method 1
	$1.83792938185221e+001$	$1.62141319837849e+001$	$1.17756060606061e+001$
	$-5.13398580625873e-003$	$-4.73897994067565e-003$	$1.55277718398098e-005$
	$-1.98303689571137e-002$	$-1.79736059630038e-002$	$-2.33368740058939e-002$
	$8.04923402085268e-002$	$8.07102530637668e-002$	$1.93365297209646e-002$
	$2.74334915209124e-001$	$2.73662965466202e-001$	$5.06137880420284e-003$
	$-2.49718504760582e-004$	$-2.55747006463849e-004$	$-1.20306147128963e-003$
	$-3.69681669971093e-003$	$-3.66948744553552e-003$	$5.36792585484259e-002$
	$2.92640252680749e-002$	$2.88210078126767e-002$	$1.41695507485378e-002$
	$-8.07416281628950e-003$	$-8.04730021982255e-003$	$-1.53040169746670e-002$
	$-8.84903284042084e-003$	$-8.86899200795415e-003$	$2.42292222047723e-003$
Norm	18.3815563307604	16.2166832200284	11.7757872770045

Table 20 Continued

Term	Least Squares	Ridge	LASSO
Intercept	$1.83792938185221e+001$	$3.39481358442390e+000$	-9.252495130026199
VH	$-5.13398580625873e-003$	$-2.40030901411901e-003$	0
Wind	$-1.98303689571137e-002$	$-6.99091193173656e-003$	0
Humidity	$8.04923402085268e-002$	$8.20014487872422e-002$	0.076869400178832
Temperature	$2.74334915209124e-001$	$2.69699018781978e-001$	0.257003761800373
IBH	$-2.49718504760582e-004$	$-2.91474679042056e-004$	-0.000341674416691
DPG	$-3.69681669971093e-003$	$-3.50952133997424e-003$	0
IBT	$2.92640252680749e-002$	$2.61949691581035e-002$	0.023655735407189
Viscosity	$-8.07416281628950e-003$	$-7.88818285680392e-003$	-0.006931522401284
DOY	$-8.84903284042084e-003$	$-8.98725605196803e-003$	-0.007601613843453
Norm	18.3815563307604	3.40662843262426	0.269490963407463

It is important to determine how the size of the solution norm reflect in the error in prediction, that is, the residual norm. Thus, in Table 21, we present the solution norm with corresponding residual norm for each of the five methods presented in Table 20.

Table 21: Solution and Residual Norms of Ozone Data

Method	Solution Norm $\|\|\alpha\|\|_2$	Residual Norm $\|\|\mathbf{y}-\mathbf{X}\alpha\|\|_2$
Least Squares	18.3815563307604	79.4431199160718
l1_ls	16.2166832200284	79.4437883803807
Method 1	11.7757872770045	102.574788852876
Ridge	3.40662843262426	79.4751306014532
LASSO	0.269490963407463	185.9927445319219

It can be seen from Table 21 that a small solution norm does not necessarily translate into a small residual norm. For example, the LASSO solution which

has the smallest solution norm rather has the largest residual norm. This means that even though the method eliminates variables that may be considered as closely related to other variables, it rather yields a very large error in prediction compared to even the un-regularized least squares solution. Incidentally, Method 1 solution also yields a high residual norm even though its solution norm is much better than the standard method $l1\_ls$.

Figure 18 also shows a plot of all coefficients values in the LASSO solution against the L_1-norm of the solution.

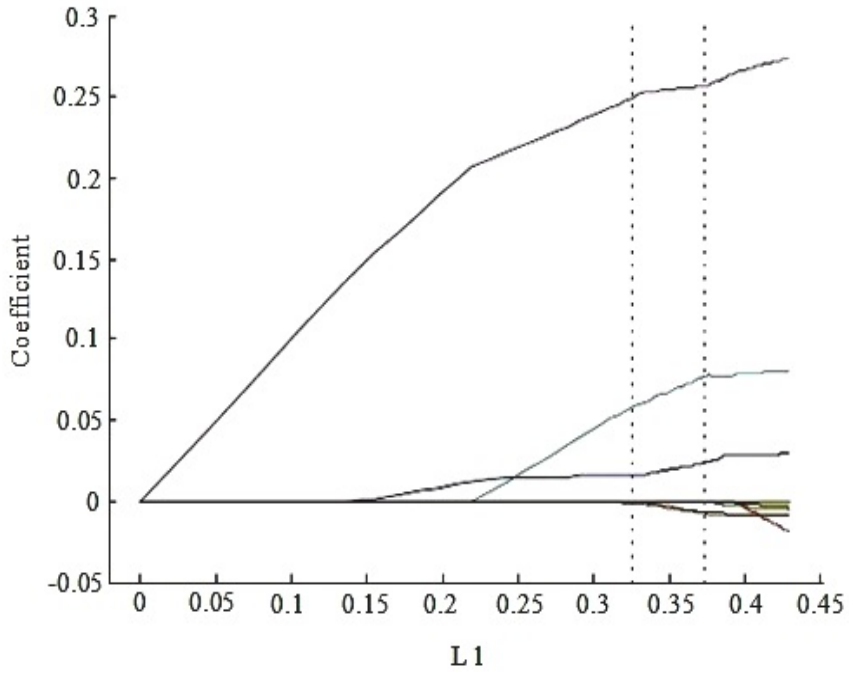

Figure 18: Trace Plot of Coefficients Fit by LASSO of Ozone Data.

Application to Boston Housing Data

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. It is created by Harrison and Rubinfeld (1978) on 'Hedonic prices and the demand for clean air'. The data consists of 506 observations on 14 variables. It seeks to explain the crime rate in Boston using

13 explanatory housing variables. The description of the variables are given as follows:

CRIM: per capita crime rate by town

ZN: proportion of residential land zoned for lots over 25,000 sq.ft.

INDUS: proportion of non-retail business acres per town

CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

NOX: nitric oxides concentration (parts per 10 million)

RM: average number of rooms per dwelling

AGE: proportion of owner-occupied units built prior to 1940

DIS: weighted distances to five Boston employment centres

RAD: index of accessibility to radial highways

TAX : full-value property-tax rate per $10,000$

PTRATIO: pupil-teacher ratio by town

B: $1000(Bk-0.63)^2$, where Bk is the proportion of blacks by town

LSTAT: lower status of the population

MEDV: Median value of owner-occupied homes in $1000's$

The data has been studied in Belsley, Kuh and Welsch (1980) and in Quinlan (1993).

Table 22 gives the fit for the Housing data for various regularization parameters and corresponding norm of the LASSO solution. We notice in Table 22 that $\|\alpha\|$ keeps increasing for lower values of λ. The value of $\lambda = 10^0$ is chosen so that our result and that of the standard method are quite close. By this selection, we are sure to avoid being prone to over-penalising the variables and hence obtain a more optimal solution. In addition, it can be observed that the solution will begin to converge only for values of λ which are much smaller and with higher solution norm.

Table 22: Method 1 Solution for various λ Values with Corresponding Norms of Housing Data

Parameter λ	Method 1 α	Norm $\|\alpha\|_2$
10^0	3.61253541501976327 −0.05970924825930624 0.15451268362766490 −3.25123411046304645 −0.62660281742045709 −0.09088620673210188 0.00980137478059614 −0.28314944728974384 0.23300269371763446 −0.00230218557618490 −0.03884210801860746 −0.00131264428488578 0.07181356966404394 −0.03384433504859750	4.91850022007651
$10^{-0.5}$	3.61321107928259266 −0.05972167942143729 0.15447563363223096 −3.26111080777584306 −0.62763891957459339 −0.09072133627756522 0.00980444068624631 −0.28323967955077950 0.23302037974320400 −0.00230220209480044 −0.03884278967034339 −0.00131231899481522 0.07180908376725045 −0.03383850365218625	4.92566392404247
10^{-1}	3.61342474308300421 −0.05972561050006699 0.15446391739494397 −3.26423409370269368 −0.62796656384415839 −0.09066919966204523 0.00980541021074090 −0.28326821349710946 0.23302597255555960 −0.00230220731844534 −0.03884300522754888 −0.00131221612906290 0.07180766520212875 −0.03383665960272630	4.92793130471480

Table 22 Continued

Parameter λ	Method 1 α	Norm $\|\alpha\|_2$
$10^{-1.5}$	3.61349230950928746 −0.05972685361628009 0.15446021239540053 −3.26522176343397241 −0.62807017405957422 −0.09065271261659140 0.00980571680130591 −0.28327723672321298 0.23302774115811656 −0.00230220897030689 −0.03884307339272246 −0.00131218360005584 0.07180721661244946 −0.03383607646308518	4.92864851488084
10^{-2}	3.61351367588932826 −0.05972724672414305 0.15445904077167188 −3.26553409202665845 −0.62810293848652898 −0.09064749895503950 0.00980581375375535 −0.28328009011784577 0.23302830043935210 −0.00230220949267139 −0.03884309494844294 −0.00131217331348061 0.07180707475593723 −0.03383589205813920	4.92887533678041
$10^{-2.5}$	3.61352043253195676 −0.05972737103576437 0.15445867027171753 −3.26563285899978606 −0.62811329950807115 −0.09064585025049395 0.00980584441281185 −0.28328099244045624 0.23302847729960782 −0.00230220965785754 −0.03884310176496043 −0.00131217006057991 0.07180702989696935 −0.03383583374417502	4.92894706617569

Figure 19 displays the sequence of cross-validated mean square errors (MSE) of LASSO fit associated with each of 100 λ values. It also shows line segments for each point that represent intervals of estimate for each MSE value. We notice that the MSE values are associated with large standard errors since the vertical segments of each MSE point is quite long. This indicates wide variation in measurement errors associated with the various models.

The right vertical line identifies the value of λ of about 10^{-1} as the value that minimizes the MSE. However, it can be noticed that this value does not appear to be any different from the MSE produced by the model with regularization parameter of about 10^{0}. The left vertical line identifies the value of λ which is more than 10^{0} as the highest value that gives an MSE which is within one standard error of the minimium MSE. Since a λ value greater than 10^{0} is prone to set several important variables to zero (that is, over-penalise), it is expedient to choose $\lambda = 10^{0}$. These results buttress our choice of $\lambda = 10^{0}$ in Table 22.

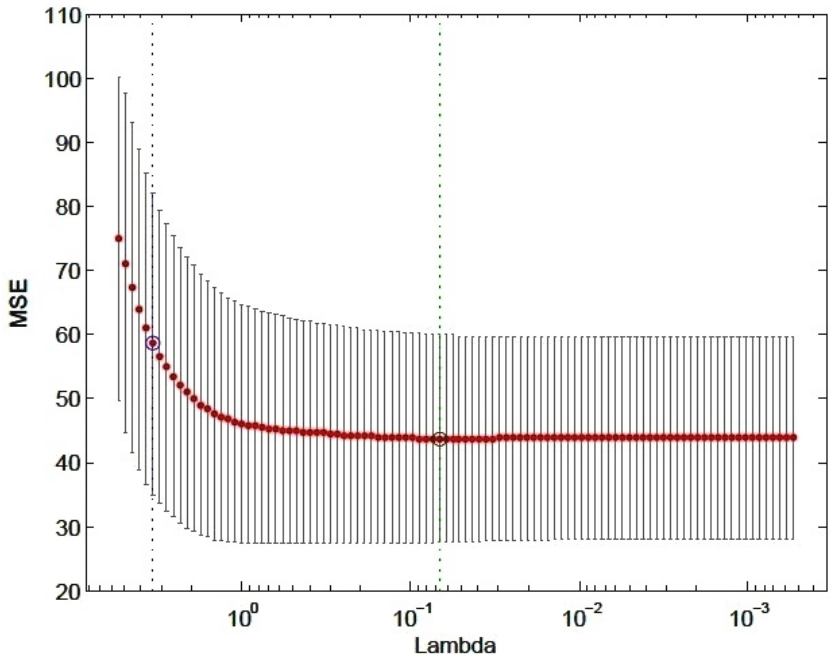

Figure 19: Cross-Validation of LASSO Fit of Housing Data.

Table 23 gives the solution for the five methods under consideration for our optimal value of $\lambda = 10^0$ selected in Table 22. In Table 23, it can be observed that for the selected λ, our method provides the lowest solution norm, much less than those of the least squares and the standard method. The solution norm for Method 1 is only smaller than that of the LASSO. It is observed that the LASSO solution sets as many as nine variables to zero leading to a very small solution norm. Thus, for this data RAD (index of accessibility to radial highways), B (proportion of blacks by town), LSTAT (lower status of the population), MEDV (Median value of owner-occupied homes) are found to be significant for determining level of crime.

Table 23: Solutions for Various Methods at Optimal Value of $\lambda = 10^0$ of Housing Data

	Least Squares	*l1_ls*	Method 1
	$1.70332275226348e+001$	$1.61496476456627e+001$	3.61253541501976327
	$4.48552146700335e-002$	$4.48446259457245e-002$	−0.05970924825930624
	$-6.38548235876919e-002$	$-6.58075922285515e-002$	0.15451268362766490
	$-7.49133610510609e-001$	$-7.39975772889635e-001$	−3.25123411046304645
	$-1.03135349120668e+001$	$-9.71263589007439e+000$	−0.62660281742045709
	$4.30130505864050e-001$	$4.53691081700494e-001$	−0.09088620673210188
	$1.45164343617989e-003$	$1.05641781079593e-003$	0.00980137478059614
	$-9.87175725502892e-001$	$-9.70533397198260e-001$	−0.28314944728974384
	$5.88208591473500e-001$	$5.84765416530147e-001$	0.23300269371763446
	$-3.78001638485947e-003$	$-3.72008670386168e-003$	−0.00230218557618490
	$-2.71080558472259e-001$	$-2.55884447661753e-001$	−0.03884210801860746
	$-7.53750488849195e-003$	$-7.46760238287949e-003$	−0.00131264428488578
	$1.26211376459500e-001$	$1.28858240764139e-001$	0.07181356966404394
	$-1.98886821265622e-001$	$-1.96404876788857e-001$	−0.03384433504859750
Norm	19.9675159568184	18.9026656584187	4.91850022007651

Table 23 Continued

Term	Least Squares	Ridge	LASSO
Intercept	$1.70332275226348e+001$	6.59146430570290587	−0.501184631019582
ZN	$4.48552146700335e-002$	0.04468522799826077	0
INDUS	$-6.38548235876919e-002$	−0.08159214601203457	0
CHAS	$-7.49133610510609e-001$	−0.78994249530158500	0
NOX	$-1.03135349120668e+001$	−4.10877314705836483	0
RM	$4.30130505864050e-001$	0.78899261230167572	0
AGE	$1.45164343617989e-003$	−0.00261627248458661	0
DIS	$-9.87175725502892e-001$	−0.80635873813479919	0
RAD	$5.88208591473500e-001$	0.54906051680831913	0.434511936263102
TAX	$-3.78001638485947e-003$	−0.00302518086347479	0
PTRATIO	$-2.71080558472259e-001$	−0.09864873954302035	0
B	$-7.53750488849195e-003$	−0.00655927780131929	−0.002856346847977
LSTAT	$1.26211376459500e-001$	0.16204706777770939	0.115979583629562
MEDV	$-1.98886821265622e-001$	−0.17191561900917043	−0.021450299648581
Norm	19.9675159568184	7.91213483807271	0.450244556487168

In order to determine how the size of the solution norm reflect in the residual norm, we present the solution norm with corresponding residual norm for each of the five methods presented in Table 24.

Table 24: Solution and Residual Norms of Housing Data

Method	Solution Norm $\|\alpha\|_2$	Residual Norm $\|\mathbf{y}-\mathbf{X}\alpha\|_2$
Least Squares	19.9675159568184	142.828327058284
l1_ls	18.9026656584187	142.830954528280
Method 1	4.91850022007651	161.450864746004
Ridge	7.91213483807271	143.157667620119
LASSO	0.450244556487168	149.1301149250945

It can be seen from Table 24 that a small solution norm does not necessarily

translate into a small residual norm. For example, even though the two LASSO solutions have the smallest solution norms, they are rather associated with the largest residual norms. For the LASSO solution in particular, even though the method produces a very sparse solution, it rather yields a very large error in prediction compared to even the un-regularized least squares solution. It is interesting to note that generally, the residual norms are high for all five methods.

This is an indication that the predictor variables considered for determining the crime data possibly do not include several other suitable predictors. In this case, therefore, the problem is due to the suitability of the explanatory variables rather than the method considered.

Application to the Hilbert Matrix

Table 25 gives the fit for the Hilbert matrix of dimension 20×12 for various regularization parameters and corresponding norm of Method 1 solution. We notice in Table 25 that $\|\alpha\|$ keeps increasing for lower values of λ. The value of $\lambda = 10^{-1}$ is chosen so that our result and that of the standard method are quite close. By this selection, we are sure to avoid being prone to over-penalising the variables and hence obtain a more optimal solution. In addition, it can be observed that the solution will begin to converge only for values of λ which are much smaller and with higher solution norm.

Table 25: Method 1 Solution for various λ Values with Corresponding Norms of Hilbert Matrix

Parameter λ	Method 1 α	Error $\|\|\alpha - \hat{\alpha}\|\|_2$
10^0	2.008926248628344 1.461103287721909 0.000000000000000 0.000000000000000 0.000000000000000 0.000000000000000 0.000000000000000 0.000000000000000 0.000000000000000 0.000000000000000 0.000000000000000 0.000000000000000	3.497220098466669
$10^{-0.5}$	2.026020554477923 1.636765304745830 0.000000000000000 −0.035321385610693 −0.136510010874235 −0.102419294346652 −0.073300303702161 −0.047874166304634 −0.025255105360424 −0.004817647252793 0.000000000000000 0.000000000000000 0.000000000000000	3.653475366557333
10^{-1}	2.031426248628344 1.692314511963317 −0.185049404419455 −0.249075629580111 −0.215154247980896 −0.187310074750930 −0.164081830404797 −0.144347422898135 −0.127299781105246 −0.112356985666977 −0.099092364112063 −0.087186978823423 −0.076397947988117	4.014392738082734

Table 25 Continued

Parameter λ	Method 1 α	Error $\|\|\alpha - \hat{\alpha}\|\|_2$
	2.033676248628344	
	1.715435634387458	
	−0.306988601955555	
	−0.279795672087906	
	−0.248577184122090	
	−0.223207946408644	
10^{-2}	−0.202344411211156	4.173646508171998
	−0.184914754848291	
	−0.170136707941386	
	−0.157441842387843	
	−0.146411358032089	
	−0.136731158716903	
	−0.128161462827724	
	2.033901248628344	
	1.717747746629872	
	−0.319182521709167	
	−0.282867676338682	
	−0.251919477736208	
	−0.226797733574420	
10^{-3}	−0.206170669291795	4.189673586303237
	−0.188971488043309	
	−0.174420400624996	
	−0.161950328059938	
	−0.151143257424078	
	−0.141685576706246	
	−0.133337814311699	
	2.033923748628344	
	1.717978957854113	
	−0.320401913684528	
	−0.283174876763763	
	−0.252253707097618	
	−0.227156712290996	
10^{-4}	−0.206553295099855	4.191277268083369
	−0.189377161362807	
	−0.174848769893365	
	−0.162401176627145	
	−0.151616447363273	
	−0.142181018505201	
	−0.133855449460077	

Figure 20 displays the sequence of cross-validated mean square errors (MSE) of LASSO fit associated with each of 100 values of λ. It also shows line segments for each point that represent intervals of estimate for each MSE value. The right vertical line identifies the value of λ of about 10^{-4} as the value that minimizes the MSE. However, the left vertical line identifies the value of λ of about 10^{-3} as the highest value that gives an MSE which is within one standard error of the minimium MSE.

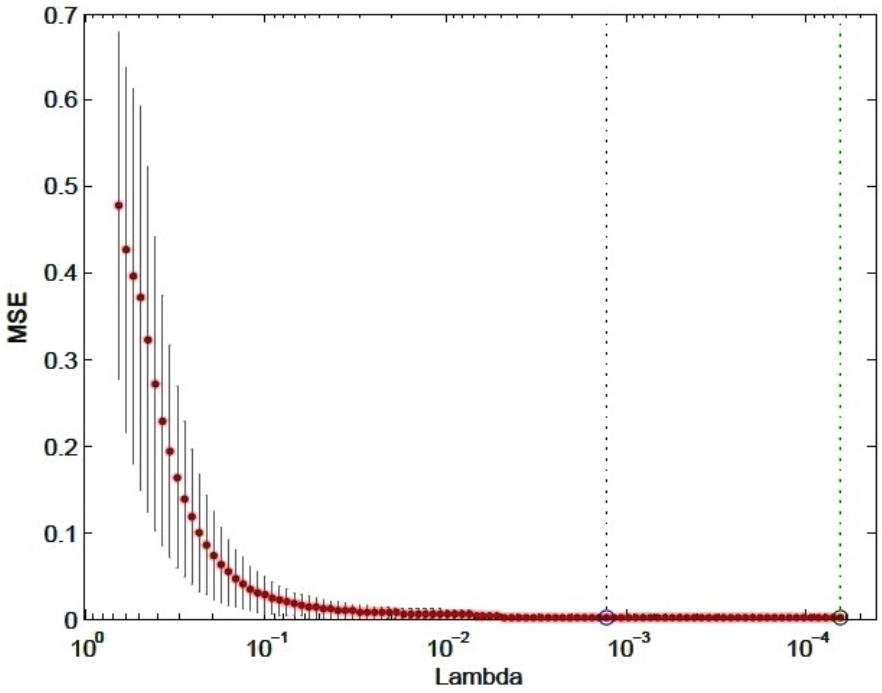

Figure 20: Cross-Validation of LASSO Fit of Hilbert Matrix.

Generally, a value of λ is chosen to lie between the two vertical lines. In this case, the MSE corresponding to $\lambda = 10^{-3}$ and $\lambda = 10^{-4}$ appear the same. We see however that the MSE corresponding to the value of $\lambda = 10^{-1}$ is also quite close to those of the recommended interval. We will therefore examine the models for both $\lambda = 10^{-1}$ and $\lambda = 10^{-3}$.

Table 26 gives the solution for the five methods for value of $\lambda = 10^{-1}$ selected in Table 25. It should be noted that for this data which is hypothetical, the

solution is already known as a vector of ones, that is, $\alpha = \{1,\ 1, \ldots, 1\} \in \Re^{12}$. Thus, a good solution is one that is closest to this vector. In Table 26, it can be observed that Method 1 solution has the second smallest solution norm after the Ridge, and the actual components of the solutions appear closest to the known solution only for the Ridge. Thus, we expect the Ridge solution to produce the smallest error (see Table 27). It should be noted however that Method 1 solution is a much improvement over the least squares. In this hypothetical case, a spares solution is not desired. The solution for the LASSO, which is highly sparse, therefore shows that the value of $\lambda = 10^{-1}$ cannot yield the desired result. We will therefore proceed to examine the solutions for $\lambda = 10^{-3}$.

For the selected λ, Method 1 provides an improved solution over the least squares and the standard *l*1_*ls* method. It is noted that the Ridge solution gives a smaller $\|\alpha\|$ and a good approximation. The best approximation however is given by the LASSO with the smallest error (see Table 26).

Table 26: Solutions for Various Methods at Optimal Value of $\lambda = 10^{-1}$ of Hilbert Matrix

	Least Squares	*l*1_*ls*	Method 1
	0.999909011935704	$1.31730637952852e+000$	2.031426248628344
	0.998621060241791	$4.27007234176251e-001$	1.692314511963317
	1.047481743425080	$4.81828163201003e+000$	−0.185049404419455
	0.636891017236956	$1.06991021373304e-005$	−0.249075629580111
	1.989495715932235	$5.13053329528883e-006$	−0.215154247980896
	−0.450015116481764	$3.43808758534863e-006$	−0.187310074750930
	4.342018810935787	$2.58952559879464e-006$	−0.164081830404797
	−5.146877164376640	$2.07614816571882e-006$	−0.144347422898135
	4.391665284020061	$1.73498557236404e-006$	−0.127299781105246
	−7.008171290483862	$1.48693578693960e-006$	−0.112356985666977
	33.220261306821953	$1.29866328441194e-006$	−0.099092364112063
	−37.634478931267573	$1.15012954781024e-006$	−0.087186978823423
	15.622654339213533	$1.03135259747115e-006$	−0.076397947988117
Norm	53.716033659057175	5.01332914948974	2.695946075368425

Table 26 Continued

Variable	Least Squares	Ridge	LASSO
Intercept	0.999909011935704	1.268744397624831	0.297982191449106
1	0.998621060241791	1.456703500480356	0
2	1.047481743425080	1.064366903647787	2.312245859146897
3	0.636891017236956	0.829879679532324	3.926469511577865
4	1.989495715932235	0.674638640083075	0
5	−0.450015116481764	0.564429366515625	0
6	4.342018810935787	0.482319066126151	0
7	−5.146877164376640	0.418941424464201	0
8	4.391665284020061	0.368677563330069	0
9	−7.008171290483862	0.327945521126813	0
10	33.220261306821953	0.294351731848317	0
11	−37.634478931267573	0.266235478533724	0
12	15.622654339213533	0.242408544447667	0
Norm	53.716033659057175	2.682329884835270	4.556714138333947

Table 27: Solution Norm and Error of Hilbert Matrix

Method	Solution Norm $\|\alpha\|_2$	Error $\|\alpha - \hat{\alpha}\|_2$
Least Squares	53.716033659057175	53.594713922985505
l1_ls	5.01332914948974	5.00082167236213
Method 1	2.695946075368425	4.014392738082734
Ridge	2.682329884835270	1.917188038654140
LASSO	4.556714138333947	4.504021869068018

Table 28: Solutions for Various Methods at Optimal Value of $\lambda = 10^{-3}$ of Hilbert Matrix

	Least Squares	*l*1_*ls*	Method 1
	0.999909011935704	$1.07522669623127e+000$	2.033901248628344
	0.998621060241791	$1.10791834114244e+000$	1.717747746629872
	1.047481743425080	$1.14981040716622e-005$	−0.319182521709167
	0.636891017236956	$4.18743902276477e-005$	−0.282867676338682
	1.989495715932235	$4.85188824026312e+000$	−0.251919477736208
	−0.450015116481764	$3.52346217355443e+000$	−0.226797733574420
	4.342018810935787	$8.00358211768244e-005$	−0.206170669291795
	−5.146877164376640	$3.15501280427493e-005$	−0.188971488043309
	4.391665284020061	$1.83103378824663e-005$	−0.174420400624996
	−7.008171290483862	$1.25533459352247e-005$	−0.161950328059938
	33.220261306821953	$9.43696318303170e-006$	−0.151143257424078
	−37.634478931267573	$7.51549960806700e-006$	−0.141685576706246
	15.622654339213533	$6.22370760421098e-006$	−0.133337814311699
Norm	53.716033659057175	6.19186568760697	2.753138002832837

Table 28 Continued

Variable	Least Squares	Ridge	LASSO
Intercept	0.999909011935704	1.041071426118285	0.015931619370373
1	0.998621060241791	0.881097886290349	0.911046406883395
2	1.047481743425080	1.241546431683054	1.184740432775627
3	0.636891017236956	1.245271429936149	1.107473379077761
4	1.989495715932235	1.164997821283271	1.051374210706618
5	−0.450015116481764	1.066849765400640	1.007005953033415
6	4.342018810935787	0.970688701052709	0.970650539353211
7	−5.146877164376640	0.882551469931652	0.940184082059951
8	4.391665284020061	0.803773704658071	0.914219461317036
9	−7.008171290483862	0.734029985583656	0.891790971964557
10	33.220261306821953	0.672449334002610	0.872198972746363
11	−37.634478931267573	0.618043206610607	0.854921588264491
12	15.622654339213533	0.569867956366867	0.839560410125281
Norm	53.716033659057175	3.393020904809317	3.351602906897268

Table 29: Solution Norm and Error of Hilbert Matrix

Method	Solution Norm $\|\alpha\|_2$	Error $\|\alpha - \hat{\alpha}\|_2$
Least Squares	53.716033659057175	53.594713922985505
*l*1_*ls*	6.19186568760697	5.49743320053883
Method 1	2.753138002832837	4.189673586303237
Ridge	3.393020904809317	0.853295155639124
LASSO	3.351602906897268	0.378033370096928

It should be noted in the case of the Hilbert matrix that a small solution norm necessarily translate into a small error and vice versa. A slight exception is in the case of Method 1 which has the smallest solution norm but does not translate into the smallest error. This is because the components of the solution

are mostly negative.

Further Assessment of Properties of Method 1

By its very construction, Method 1 just like any other LASSO method must have a feature selection property. In order to investigate this property, we will augment the 20×12 Hilbert matrix to include two additional columns that are linearly dependent on others. A 13th column is obtained as the average of the first three columns. The 14th column is created so that it is orthogonal to the first two columns using Gram-Schmidt Orthogonalisation process.

Using the cross-validated MSE of LASSO fit for this augmented matrix, it can be seen in Figure 21 that $\lambda = 10^{-3}$ provides a suitable regularization parameter. As can be seen in Table 30, Method 1 sets the two created columns to zero as expected. However, the LASSO method sets only one of the two new columns to zero. Thus, Method 1 identifies all linearly dependent variables as dispensable, which is a desirable statistical property. The LASSO however may consider a linearly dependent variable as indispensable. It is further observed that in the presence of linearly dependent variables, the least squares solution produces a very poor fit.

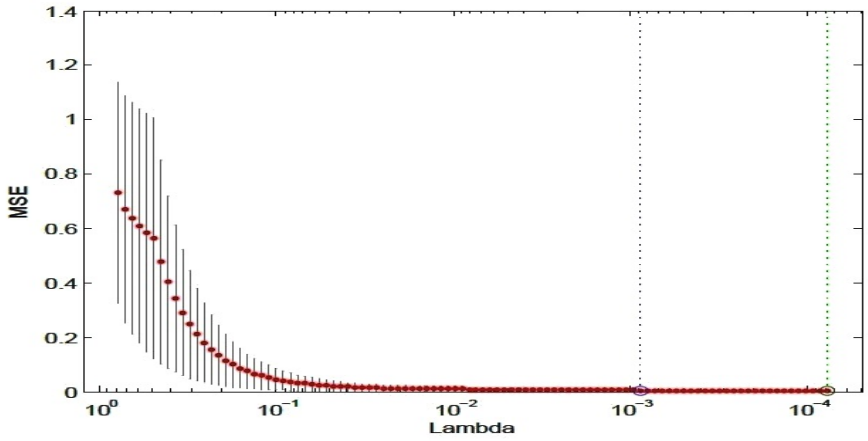

Figure 21: Cross-Validation of LASSO Fit of Augmented Hilbert Matrix.

Table 30: Solutions for Various Methods at Optimal Value of $\lambda = 10^{-3}$ of Augmented Matrix

	Least Squares($\times 1.0e+002$)	*l1_ls*	Method 1
	0.009987377897659	$1.07224342897854e+000$	2.175961038746634
	0.209549801354391	$1.59730101721228e+000$	2.076676150836919
	0.503466147654840	$1.55714147799468e-005$	−0.401369581926816
	−0.064500499229864	$5.70215532530348e-005$	−0.349116525315617
	0.527347821565438	$4.61591586991824e+000$	−0.307981051675371
	−1.139481533340739	$4.46195982301911e+000$	−0.275612062889607
	1.514573671821963	$1.09787121527510e-004$	−0.249502059377293
	−0.767372082362858	$4.33749093106751e-005$	−0.227982164013520
	−1.081843421521241	$2.52071302584233e-005$	−0.209924848124022
	1.078851144441624	$1.72841156333174e-005$	−0.194545298836001
	2.562158557008309	$1.29969413355205e-005$	−0.181280973264948
	−4.227301633868075	$1.03532954156533e-005$	−0.169717964211714
	1.735646062752233	$8.57819811984210e-006$	−0.159544718012593
	−0.738648259281753	$3.86049067478017e-005$	0
	0.210681219592631	$1.91176000057233e-006$	0
Norm	592.7263465629269	6.70200278358500	3.128010463559020

Table 30 Continued

Variable	Least Squares($\times 1.0e+002$)	Ridge	LASSO
Intercept	0.009987377897659	1.045506279278075	0.009073476731169
1	0.209549801354391	0.909709978504240	1.381920892531583
2	0.503466147654840	1.129897396747912	1.178218364190754
3	−0.064500499229864	1.174811169708568	1.166793543524774
4	0.527347821565438	1.131026545245665	1.133606784522456
5	−1.139481533340739	1.056855517151317	1.097139230075671
6	1.514573671821963	0.975753376593894	1.062560215380405
7	−0.767372082362858	0.896949556778248	1.031171116041779
8	−1.081843421521241	0.823867721746281	1.003073001472601
9	1.078851144441624	0.757489743392347	0.977998657299729
10	2.562158557008309	0.697765688571583	0.955592119908517
11	−4.227301633868075	0.644230252329005	0.935504241697110
12	1.735646062752233	0.596278820157640	0.917422736294870
13	−0.738648259281753	1.071472848317259	0.041121466380678
14	0.210681219592631	0.096577171767843	0
Norm	592.7263465629269	3.521032629893173	3.732928436495016

Table 31: Solution Norm and Error of Augmented Hilbert Matrix

Method	Solution Norm $\|\|\alpha\|\|_2$	Error $\|\|\alpha - \hat{\alpha}\|\|_2$
Least Squares	592.7263465629269	592.6827979914198
*l*1_*ls*	6.70200278358500	6.03500784177055
Method 1	3.128010463559020	4.661794673327281
Ridge	3.521032629893173	1.175281518697833
LASSO	3.732928436495016	1.473265072331274

Population Data

Finally in this section, we examine the features of Method 1 using data generated for polynomial fit of the population data presented in Example 1 in

Chapter One. The full polynomial fit (of order $n-1$) is first considered for Method 1 in relation to the other four methods. Figure 22 shows the least squares polynomial fit of the highest order of eighteen. The curve is seen to pass through all the points. The first column of Table 32 which gives the solution for the least squares method shows that this curve over-fits the data as coefficients of terms in powers higher than 9 are all almost zero. This means that we do not need to fit a polynomial with several higher order terms.

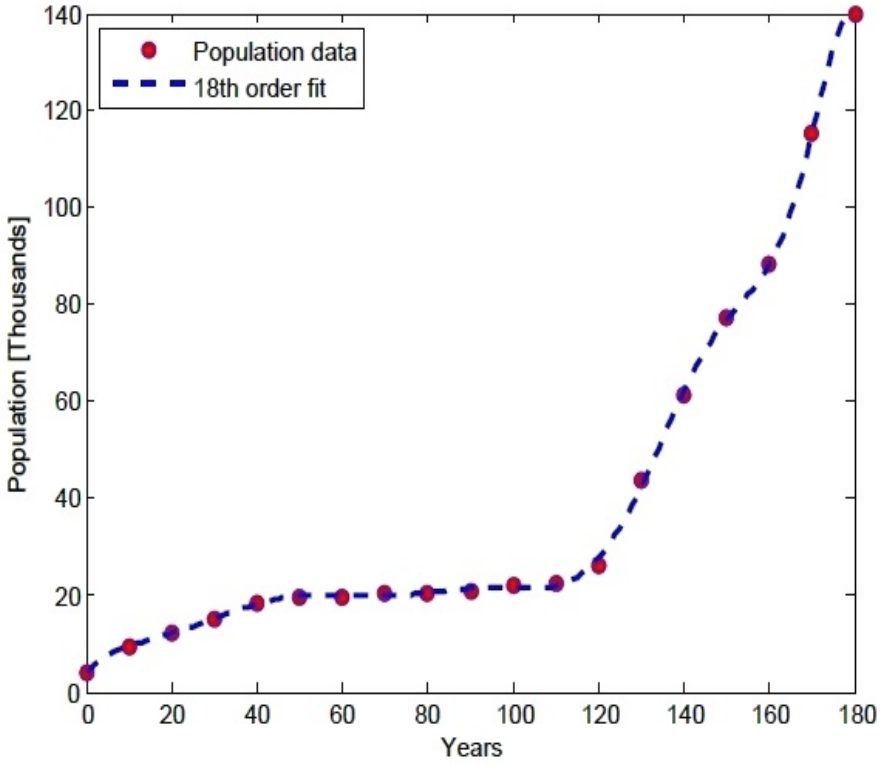

Figure 22: Least Squares Polynomial Fit of Order Eighteen.

We further examine a regularization of the least squares solution. Figure 23 is the Cross-Validated MSE for the highest degree polynomial for the data. The value of $\lambda = 10^{-1.5}$ is seen to be the optimal regularization parameter. It is interesting to note that for this data, a very low value of λ rather distort the fit.

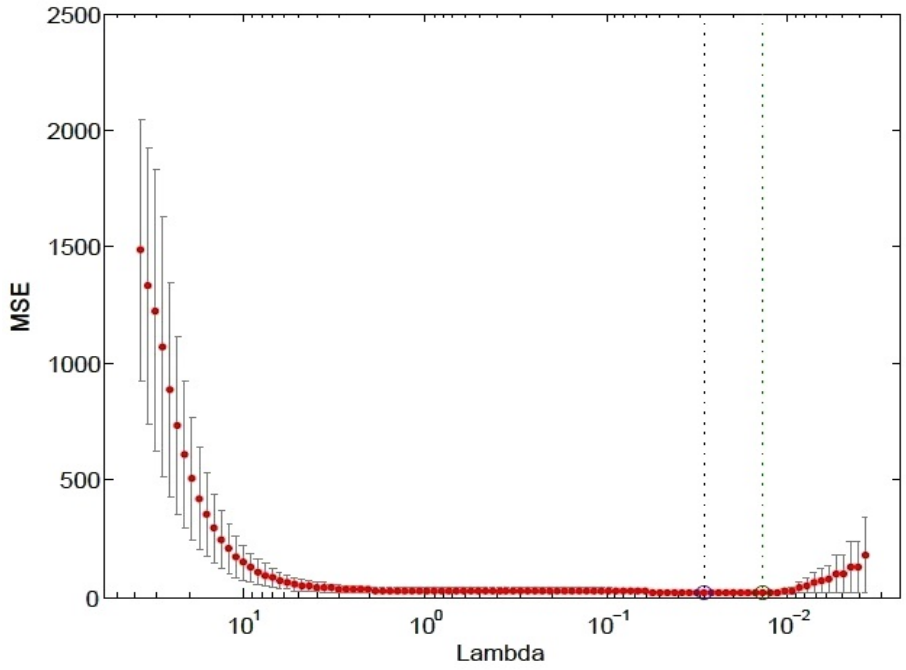

Figure 23: Cross-Validation of LASSO Fit of Population Data.

Using the chosen value of $\lambda = 10^{-1.5}$, we obtain the L_2 regularized least squares fit in Figure 24. The curve looks just like the least squares fit.

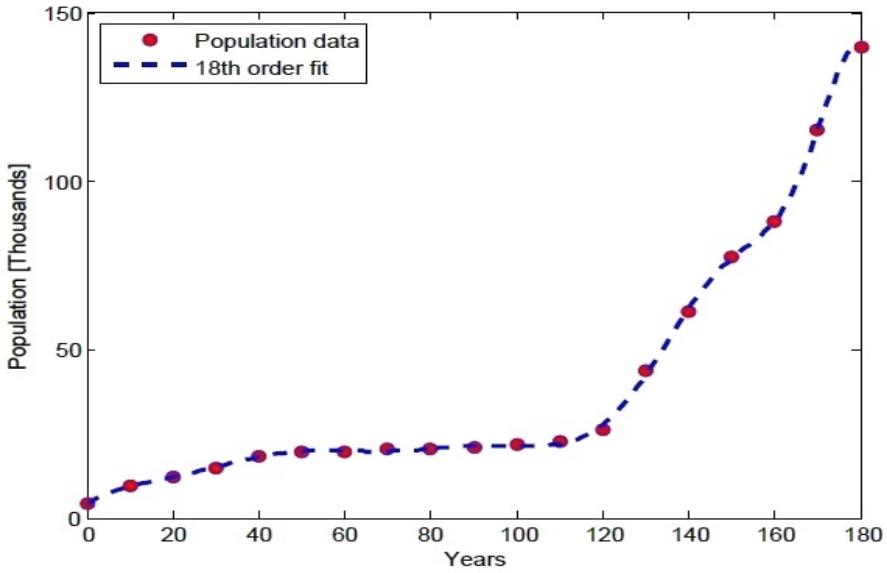

Figure 24: Polynomial Fit of Order Eighteen to the Population Data for the Ridge Method.

The closeness of the two solutions is seen in Table 32 and Table 33 which give almost the same value of the solution norm and residual norm for the two solutions. The ridge solution also buttresses a case of over-fitting. Method 1 (Table 32) also sets to almost zero coefficients of higher powers of the model. However, the corresponding fit (in Figure 25) does not fit the data at all. The figure thus shows clearly that an over-fitted model does not actually fit the model.

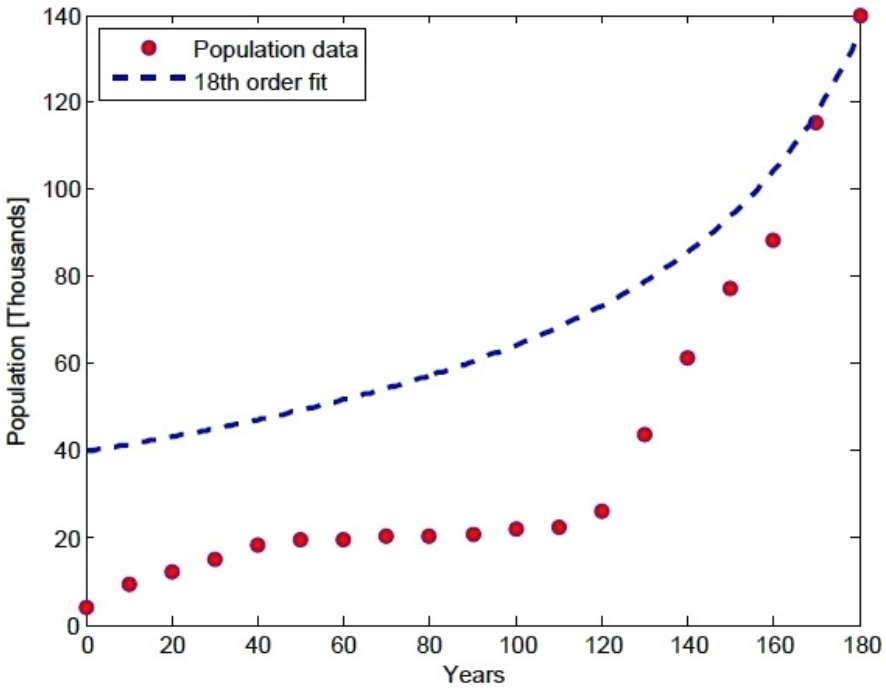

Figure 25: Method 1 Polynomial Fit of Order Eighteen.

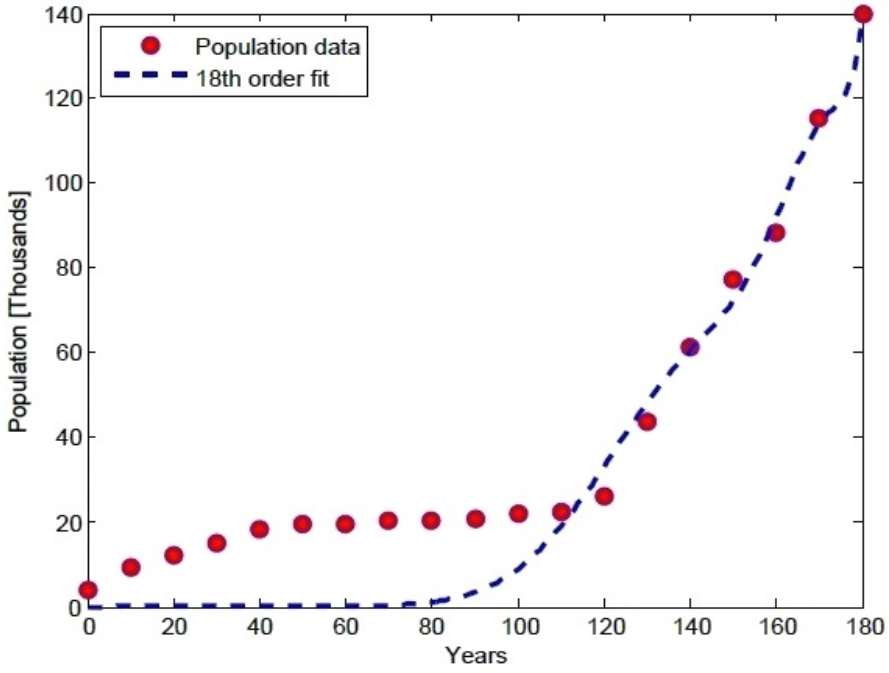

Figure 26: Polynomial Fit of Order Eighteen for *l*1_*ls* Method.

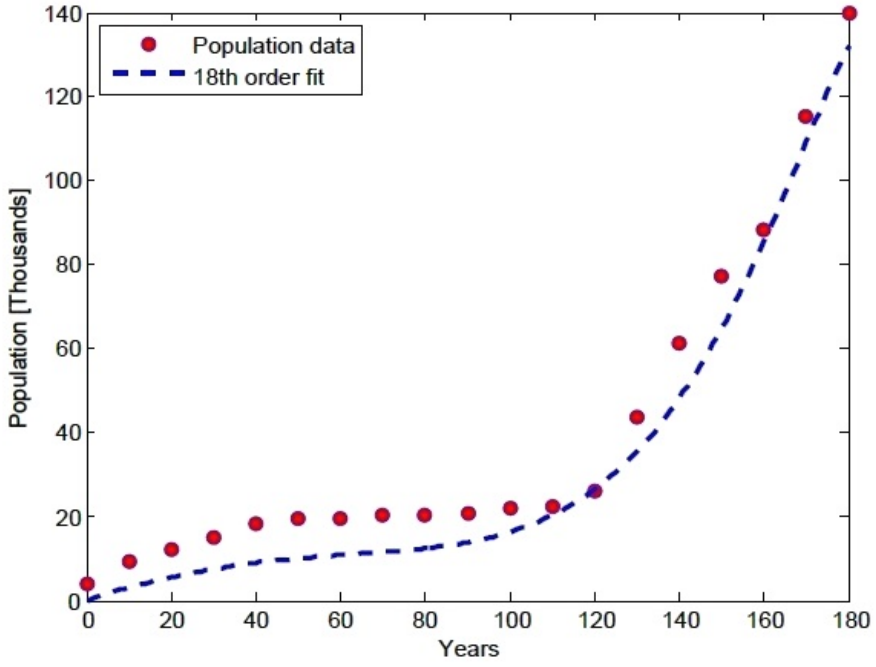

Figure 27: LASSO Polynomial Fit of Order Eighteen of Population Data.

The fit provided by the *l*1_*ls* method is presented in Figure 26. The figure shows a bad fit. The corresponding model, in Table 32, sets all terms of the model to almost zero as though a polynomial model is inappropriate.

The LASSO identifies only the first, second and the fifth terms of the model as influential and sets all others to zero. The graph of the LASSO in Figure 27 also shows a poor fit even though it generally shows the pattern of growth of the population. Thus, the three LASSO methods show that an over-fitted model does not actually fit the data. Method 1 emphasises this point. Based on the results and the results in Table 4 (Chapter One), we present polynomial fit of order three for all five methods. See Appendix G for polynomial fit of order five for all the five methods.

Table 32: Solutions for Various Methods at Optimal Value of $\lambda = 10^{-1.5}$ of Population Data

	Least Squares	*l1_ls*	Method 1
	4.017426257483916	$4.97154835416185e-052$	39.714009926931539
	0.956581154954685	$3.00915141828701e-050$	0.156962054118344
	−0.068506221310917	$2.22540503085932e-048$	0.000520640452908
	0.003629427345186	$1.79757359349947e-046$	0.000002059419805
	−0.000130394119539	$1.52055406746665e-044$	0.000000008404908
	0.000003870279427	$1.31494479239696e-042$	0.000000000034978
	−0.000000090411075	$1.14313165287354e-040$	0.000000000000148
	0.000000001423191	$9.84995593220107e-039$	0.000000000000001
	−0.000000000013663	$8.29537776919442e-037$	0.000000000000000
	0.000000000000073	$6.71910674236393e-035$	0.000000000000000
	−0.000000000000000	$5.12574670185295e-033$	0.000000000000000
	0.000000000000000	$3.57172899059658e-031$	0.000000000000000
	−0.000000000000000	$2.16130129595661e-029$	0.000000000000000
	0.000000000000000	$1.02903393224228e-027$	0.000000000000000
	−0.000000000000000	$2.96979351169587e-026$	0.000000000000000
	0.000000000000000	$-7.00395213146629e-028$	0.000000000000000
	0.000000000000000	$6.21301256760081e-030$	0.000000000000000
	−0.000000000000000	$-2.45521643213018e-032$	0.000000000000000
	0.000000000000000	$3.64443587767096e-035$	0.000000000000000
Norm	4.130310827583612	$2.97240192496144e-026$	39.714320110432723

Table 32 Continued

Variable	Least Squares	Ridge	LASSO
Intercept	4.017426257483916	3.901703281297118	7.034107646863767
1	0.956581154954685	0.920694620960775	0.328973580483269
2	−0.068506221310917	−0.057341535511464	−0.002723215512980
3	0.003629427345186	0.002602814351453	0
4	−0.000130394119539	−0.000083082858203	0
5	0.000003870279427	0.000002596082674	0.000000001057759
6	−0.000000090411075	−0.000000069012874	0
7	0.000000001423191	0.000000001193944	0
8	−0.000000000013663	−0.000000000012103	0
9	0.000000000000073	0.000000000000066	0
10	−0.000000000000000	−0.000000000000000	0
11	0.000000000000000	0.000000000000000	0
12	−0.000000000000000	−0.000000000000000	0
13	0.000000000000000	0.000000000000000	0
14	−0.000000000000000	−0.000000000000000	0
15	0.000000000000000	0.000000000000000	0
16	0.000000000000000	0.000000000000000	0
17	−0.000000000000000	−0.000000000000000	0
18	0.000000000000000	0.000000000000000	0
Norm	4.130310827583612	4.009271992968009	0.328984851564189

Table 33: Solution and Residual Norms of Polynomial Fit of Population Data

Method	Solution Norm $\|\alpha\|_2$	Residual Norm $\|\mathbf{y} - \mathbf{X}\alpha\|_2$
Least Squares	4.130310827583612	2.486823367148147
l1_ls	$2.97240192496144e-026$	54.1362406879371
Method 1	39.714320110432723	139.4715629026298
Ridge	4.009271992968009	2.489847268745522
LASSO	0.328984851564189	33.673526126802990

Table 34 shows the least squares solution for the degree three with the corresponding fit given in Figure 28. To determine the right regularization para-

meter, we obtain Figure 28. The figure shows that a value of λ equal to 10^{-2} is appropriate. Using this value, solutions for the four regularization methods are given in Tables 34 and 35. The graphs for the respective fits are given in Figures 29 to 33. The results show that the fits provided by the *l*1_*ls*, Ridge have almost the same estimates with the least squares. However, the LASSO and Method 1 (see Figures 31 and 33) continue to show that they are unsuitable for the kind of data.

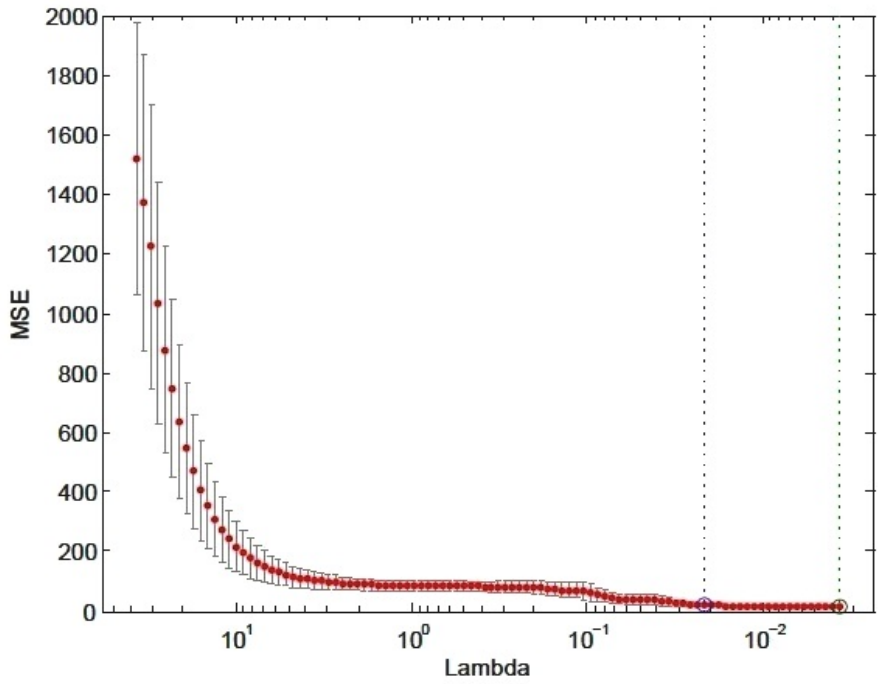

Figure 28: Cross-Validation LASSO Fit of Polynomial of Degree Three.

Table 34: Solutions for Various Methods at Optimal Value of $\lambda = 10^{-2}$ of Population Data

	Least Squares	*l*1_*ls*	Method 1
	3.367464798361461	3.364718229881230	39.714578947368423
	0.743102737235111	0.743211174169917	0.156957491702229
	−0.012518154242556	−0.012519312441434	0.000520643555525
	0.000070448111142	0.000070451740089	0.000002059429493
Norm	3.448503669048140	3.445845085614494	39.714889108394026

Table 34 Continued

Variable	Least Squares	Ridge	LASSO
Intercept	3.367464798361461	3.348163500142492	4.724037897351302
1	0.743102737235111	0.743876748268971	0.648449597305884
2	−0.012518154242556	−0.012526488226801	−0.011234809077931
3	0.000070448111142	0.000070474352428	0.000065845822752
Norm	3.448503669048140	3.429826286880369	0.648546918516196

Table 35: Solution and Residual Norms of Polynomial Fit of Population Data

Method	Solution Norm $\|\alpha\|_2$	Residual Norm $\|\mathbf{y} - \mathbf{X}\alpha\|_2$
Least Squares	3.448503669048140	13.846617984349315
*l*1_*ls*	3.445845085614494	13.846618453015891
Method 1	39.714889108394026	140.9277126962934
Ridge	3.429826286880369	13.846641112009655
LASSO	0.648546918516196	24.949421186858480

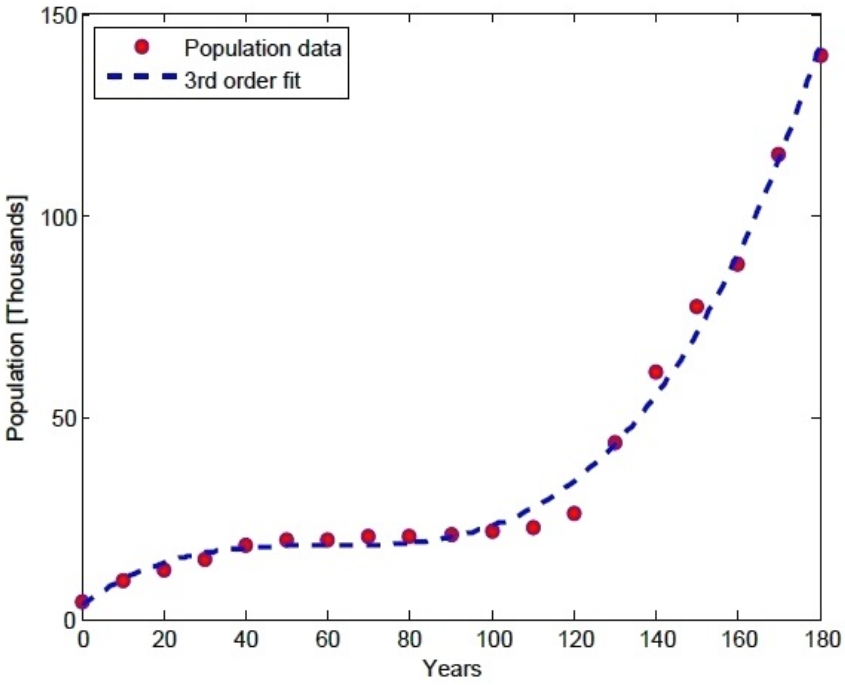

Figure 29: Least Squares Polynomial Fit of Order Three of Population Data.

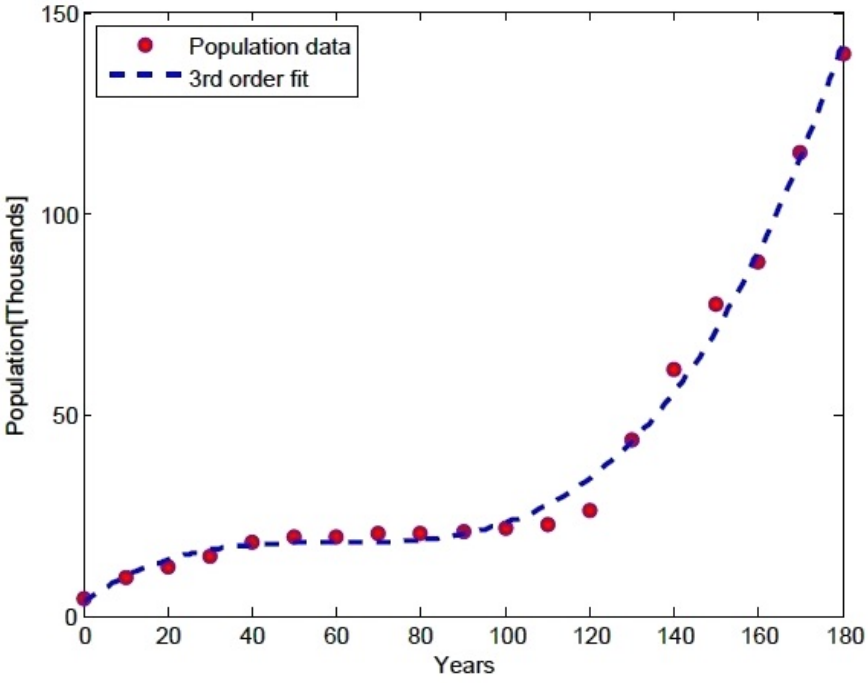

Figure 30: Ridge Polynomial Fit of Order Three of Population Data.

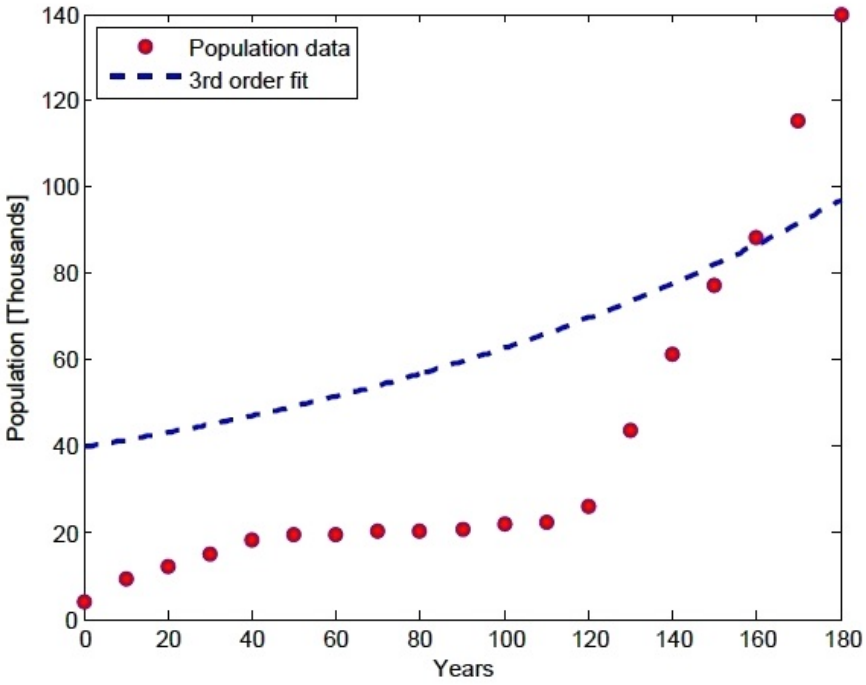

Figure 31: Method 1 Polynomial Fit of Order Three of Population Data.

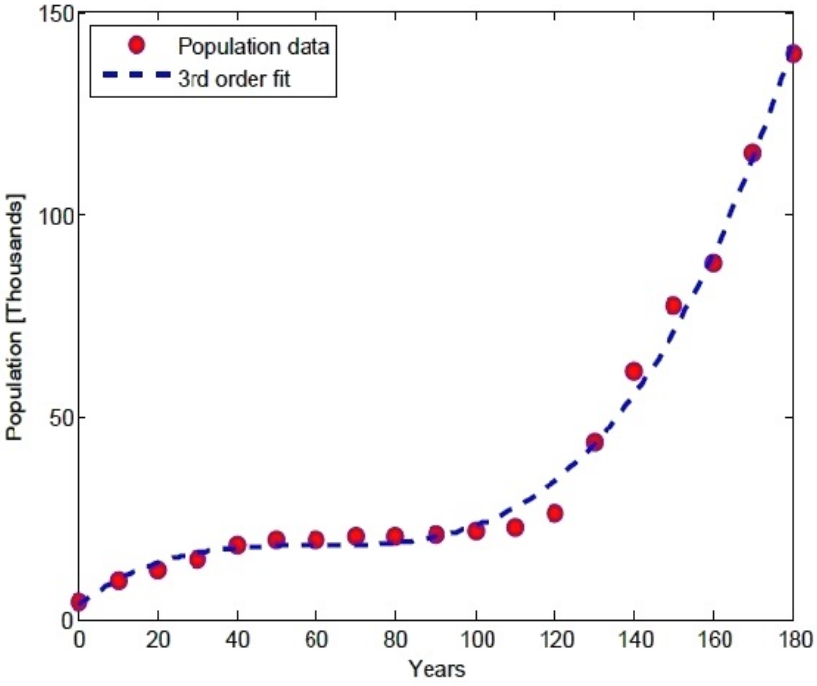

Figure 32: The *l*1_*ls* Polynomial Fit of Order Three of Population Data.

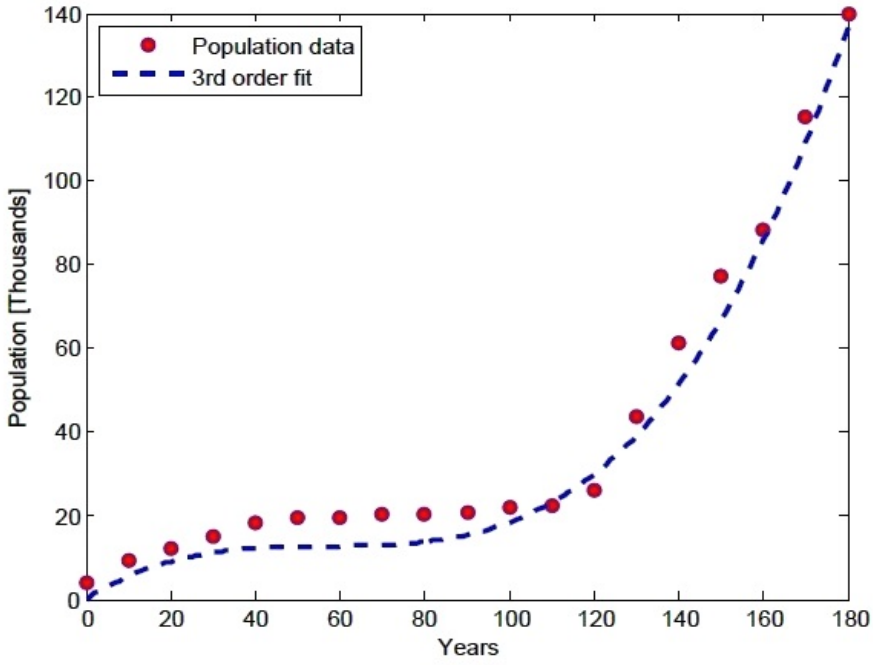

Figure 33: LASSO Polynomial Fit of Order Three of Population Data.

Appendix G shows the results for polynomial fit of order five for all five methods. The results do not show much difference from that of order three. The polynomial regression fitting has shown that L_1-regularization is not appropriate in this case. The lack of suitability of the methods is highlighted by Method 1. This buttresses why L_2-regularization is mostly used in polynomial data fitting.

We further examine the effect of trend component in the data on the performance of LASSO solutions by using data with more data points. We make use of the Global Temperature Anomaly introduced in Chapter One and plotted in Figure 4. In Table 36, we have the parameter estimates for order 20 polynomial model fit of the data for all five methods for optimal parameter value of $\lambda = 10^{-3}$. The least squares fit, which does not depend on regularization parameter, sets higher terms beyond the fifth power to almost zero. Thus, the least squares finds the first five terms of the polynomial as significant. The corresponding graph (in Figure 34) shows that this fit reflect rather a reverse of the actual

trend. The solution provided by the Ridge regularization (see Figure 34) looks similar to that of the least squares.

Table 36: Solutions for Various Methods at Optimal Value of $\lambda = 10^{-3}$ of Global Temperature Anomaly Data

	Least Squares $(1.0e+005*)$	$l1\_ls(1.0e-053*)$	Method 1
	2.200703621889671	−0.000000000000000	−0.105081325301205
	0.014519230762429	−0.000000000000000	0.000003001369790
	−0.000010777971979	−0.000000000000000	0.000000001562665
	−0.000000006932975	−0.000000000000000	0.000000000000811
	0.000000000002283	−0.000000000000000	0.000000000000000
	0.000000000000002	−0.000000000000000	0.000000000000000
	0.000000000000000	−0.000000000000000	0.000000000000000
	−0.000000000000000	−0.000000000000000	0.000000000000000
	−0.000000000000000	−0.000000000000000	0.000000000000000
	−0.000000000000000	−0.000000000000000	0.000000000000000
	0.000000000000000	−0.000000000000000	0.000000000000000
	−0.000000000000000	−0.000000000000000	0.000000000000000
	−0.000000000000000	−0.000000000000001	0.000000000000000
	0.000000000000000	−0.000000000001344	0.000000000000000
	0.000000000000000	−0.000000001536672	0.000000000000000
	0.000000000000000	−0.000001536542160	0.000000000000000
	−0.000000000000000	−0.001228204200790	0.000000000000000
	−0.000000000000000	−0.613154917337675	0.000000000000000
	−0.000000000000000	0.000940681916668	0.000000000000000
	−0.000000000000000	−0.000000485213500	0.000000000000000
	0.000000000000000	0.000000000084114	0.000000000000000
Norm	$2.200751516999684e+005$	$6.131568690196614e-054$	0.105081325344068

Table 36 Continued

Variable	Least Squares $(1.0e+005*)$	Ridge	LASSO
Intercept	2.200703621889671	−0.000339238607326	8.224183331673803
1	0.014519230762429	−0.661589116627188	−0.004822159593487
2	−0.000010777971979	−0.088146940252592	0
3	−0.000000006932975	0.000070793740067	0
4	0.000000000002283	0.000000012666109	0
5	0.000000000000002	0.000000000007774	0
6	0.000000000000000	−0.000000000000018	0
7	−0.000000000000000	0	0
8	−0.000000000000000	0	0
9	−0.000000000000000	0	0
10	0.000000000000000	0	0
11	−0.000000000000000	0	0
12	−0.000000000000000	0	0
13	0.000000000000000	0	0
14	0.000000000000000	0	0
15	0.000000000000000	0	0
16	0.000000000000000	0	0
17	−0.000000000000000	0	0
18	−0.000000000000000	0	0
19	−0.000000000000000	0	0
20	−0.000000000000000	0	0
21	0.000000000000000	0	0
Norm	2.200751516999684$e+005$	0.667435511798723	0.004822159593487

Table 37: Solution and Residual Norms of Polynomial Fit of Global Temperature Anomaly Data

Method	Solution Norm $\|\alpha\|_2$	Residual Norm $\|\mathbf{y}-\mathbf{X}\alpha\|_2$
Least Squares	$2.200751516999684e+005$	2.222518451821552
l1_ls	$6.131568690196614e-054$	1.547064734768672
Method 1	0.105081325344068	3.645445031865414
Ridge	0.667435511798723	1.463419999310810
LASSO	0.004822159593487	105.9726143089173

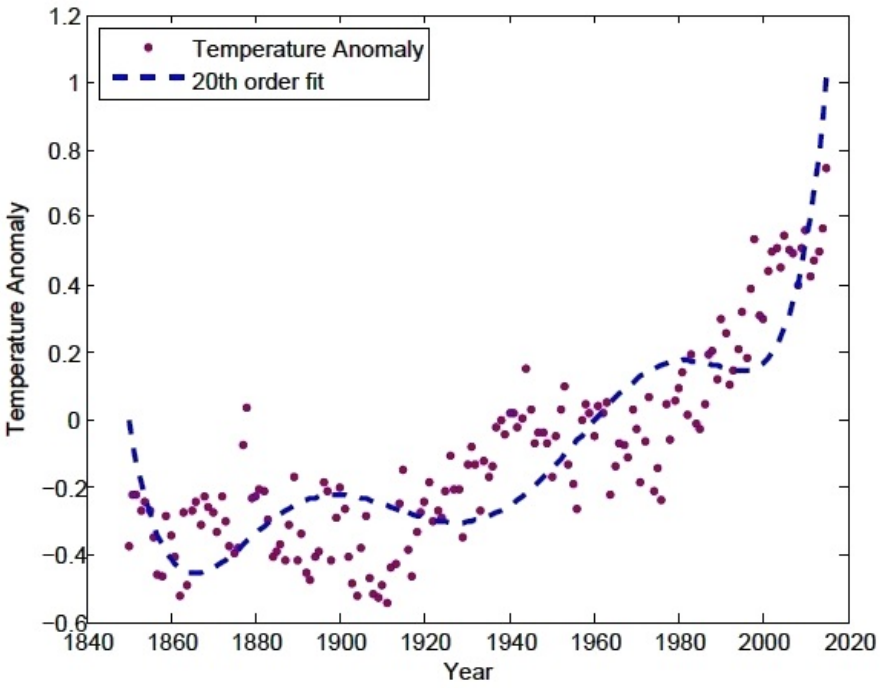

Figure 34: Least Squares Polynomial Fit of Order 20 of Global Temperature Anomaly Data.

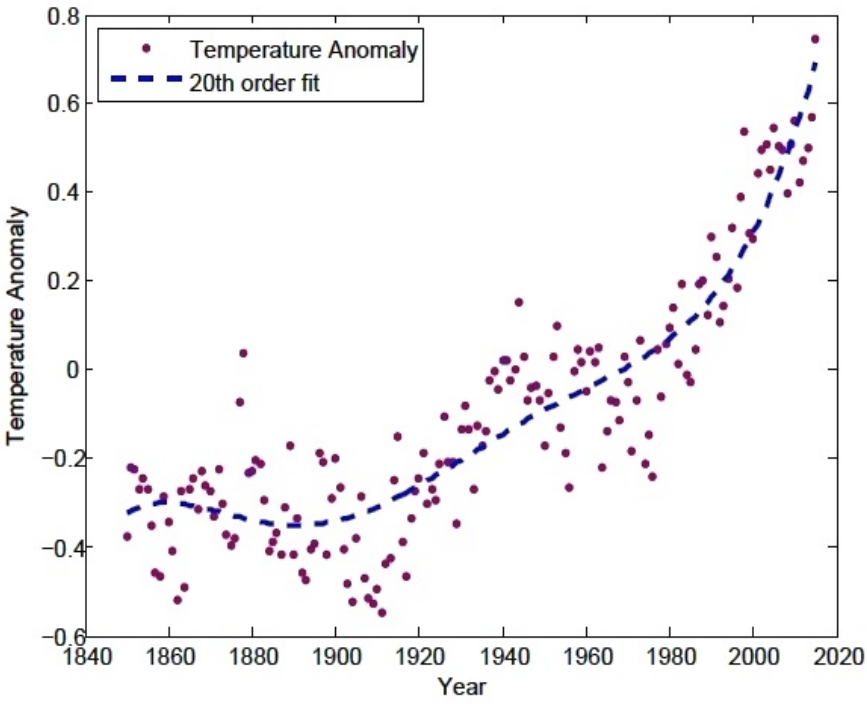

Figure 35: Ridge Polynomial Fit of Order 20 of Global Temperature Anomaly Data.

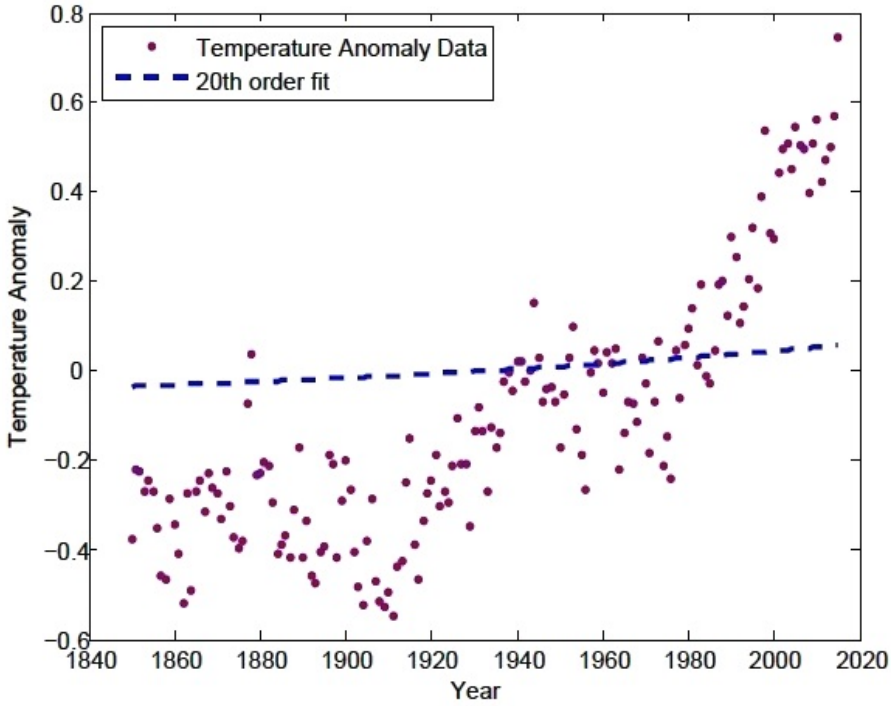

Figure 36: Method 1 Polynomial Fit of Order 20 of Global Temperature Anomaly Data.

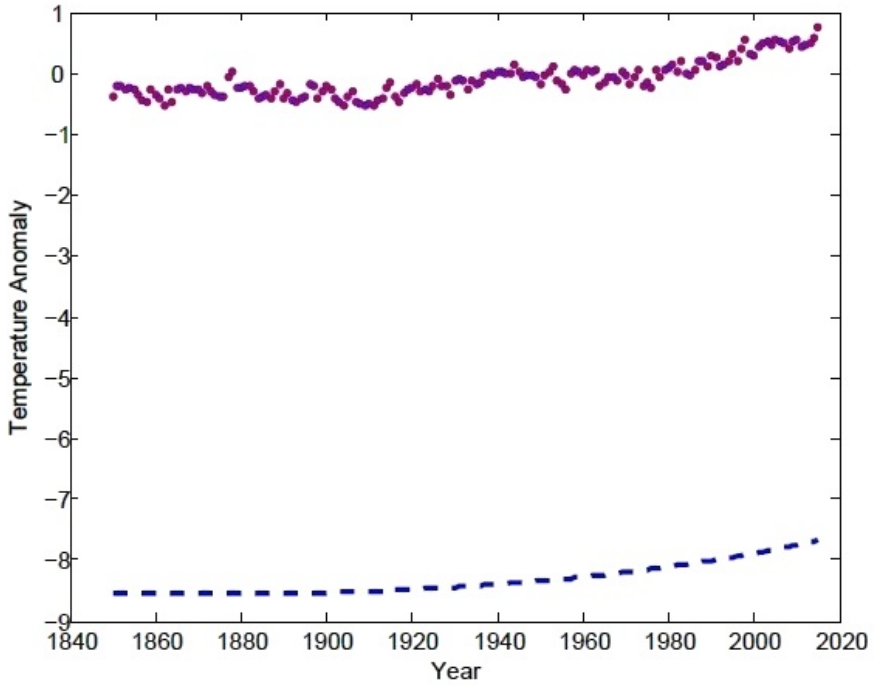

Figure 37: LASSO Polynomial Fit of Order 20 of Global Temperature Anomaly Data.

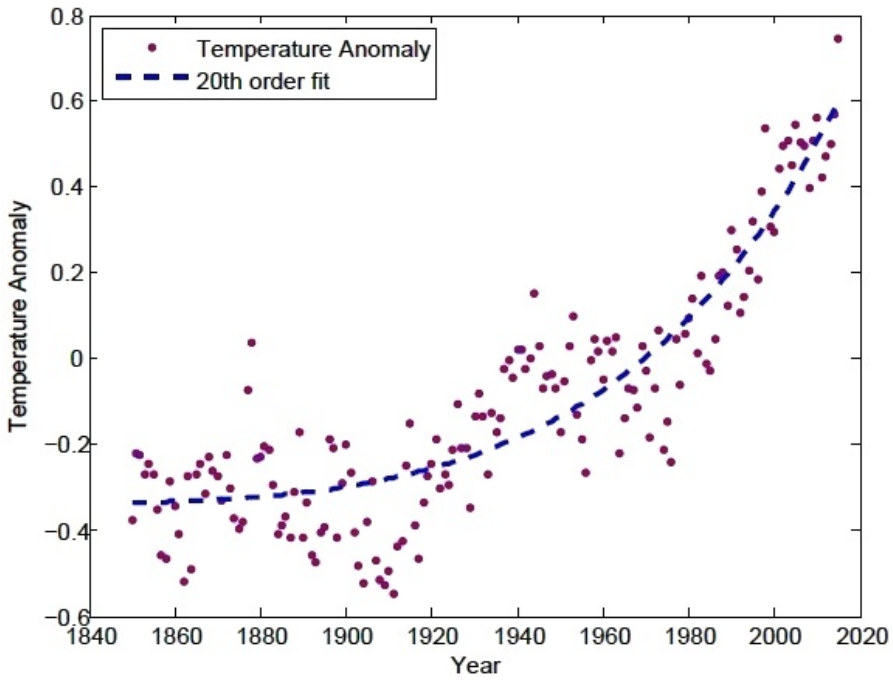

Figure 38: The *l*1_*ls* Polynomial Fit of Order 20 of Global Temperature Anomaly Data.

The fit given by the *l*1_*ls* sets almost all terms to zero. However, the graph in Figure 38 shows that this solution provides the best parameter estimates for the data among all methods considered. The residual norms in Table 37 buttresses this point. It however does not reflect the true trend of the data. This table and the graphs in Figures 36 and 37 for Method 1 and the LASSO show that these methods are simply unsuitable for the data. The results for the Global Temperature Anomaly shows that in general, L_1-norm regularized least squares methods are unsuitable for fitting data with monotone trends.

Singular Value Decomposition Form of Method 1

The solution for Method 1 has been obtained in Equation (5.11) as

$$\alpha_i = \begin{cases} \frac{1}{2}(\mathbf{X}^T\mathbf{X})^{-1}(2\mathbf{X}^T\mathbf{y} - \lambda\mathbf{I}_i), & \lambda_i < 2(\mathbf{X}^T\mathbf{y})_i \\ \frac{1}{2}(\mathbf{X}^T\mathbf{X})^{-1}(2\mathbf{X}^T\mathbf{y} + \lambda\mathbf{I}_i), & -\lambda_i > 2(\mathbf{X}^T\mathbf{y})_i \\ 0, & -\lambda_i < 2(\mathbf{X}^T\mathbf{y})_i < \lambda_i. \end{cases} \tag{5.13}$$

In this section, we derive the SVD representation of this solution. The SVD of $\mathbf{X}$ is given as

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T.$$

Making substitution for $\mathbf{X}$ in Equation (5.13), we derive the SVD version of Method 1.

For Case 1,

$$\begin{aligned} \alpha_i &= \frac{1}{2}\left[(\mathbf{U}\mathbf{S}\mathbf{V}^T)^T\mathbf{U}\mathbf{S}\mathbf{V}^T\right]^{-1}\left[2(\mathbf{U}\mathbf{S}\mathbf{V}^T)^T\mathbf{y} - \lambda\mathbf{I}_i\right] \\ &= \frac{1}{2}\left[(\mathbf{V}\mathbf{S}\mathbf{U}^T)\mathbf{U}\mathbf{S}\mathbf{V}^T\right]^{-1}\left[2(\mathbf{V}\mathbf{S}^T\mathbf{U}^T)\mathbf{y} - \lambda\mathbf{I}_i\right] \\ &= \frac{1}{2}\left[\mathbf{V}\mathbf{S}^2\mathbf{V}^T\right]^{-1}\left[2(\mathbf{V}\mathbf{S}^T\mathbf{U}^T)\mathbf{y} - \lambda(\mathbf{V}\mathbf{I}\mathbf{V}^T)_i\right] \\ &= \frac{1}{2}\mathbf{V}\left(\mathbf{S}^2\right)^{-1}\mathbf{V}^{-1}\left[\mathbf{V}(2\mathbf{S}^T\mathbf{U}^T\mathbf{y} - \lambda(\mathbf{V}^T)_i)\right] \\ &= \frac{1}{2}\mathbf{V}\left(\mathbf{S}^2\right)^{-1}\left(2\mathbf{S}^T\mathbf{U}^T\mathbf{y} - \lambda(\mathbf{V}^T)_i\right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\left(\mathbf{S}^2\right)^{-1}\left(2\mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{y} - \lambda\mathbf{I}_i\right) \\
\alpha_i &= \frac{1}{2\sigma_i^2}\left[2\sigma_i(\mathbf{U}_i^T\mathbf{y})\mathbf{V}_i - \lambda_i\right]
\end{aligned}
$$

Similarly, for Cases 2 and 3, the SVD in component forms are given as

$$
\begin{aligned}
\alpha_i &= \frac{1}{2}\left(\mathbf{S}^2\right)^{-1}\left(2\mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{y} + \lambda\mathbf{I}_i\right) \\
&= \frac{1}{2\sigma_i^2}\left[2\sigma_i(\mathbf{U}_i^T\mathbf{y})\mathbf{V}_i + \lambda_i\right]
\end{aligned}
$$

and 0, respectively. Combining the solutions in the three cases, the composite solution for Method 1 in terms of singular value decomposition is obtained as

$$
\alpha_i = \begin{cases}
\frac{1}{2\sigma_i^2}\left[2\sigma_i(\mathbf{U}_i^T\mathbf{y})\mathbf{V}_i - \lambda_i\right], & \lambda_i < 2(\mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{y})_i \\
\frac{1}{2\sigma_i^2}\left[2\sigma_i(\mathbf{U}_i^T\mathbf{y})\mathbf{V}_i + \lambda_i\right], & -\lambda_i > 2(\mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{y})_i \\
0, & -\lambda_i < 2(\mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{y})_i < \lambda_i.
\end{cases} \tag{5.14}
$$

Algorithm for Method 1 Using SVD

```
Initialize regularization parameter lambda;
alpha=zeros(n,1);
[USV] = svd(x);
sigma=diag(S);
SS=S′*S;
Uty=U′*y;
xy=2*(V*S′*Uty);
xx=2*(V*SS*V′);
alphaold=alpha;
for j=1:n
cj=xy(j)-sum(xx(j,  :)* alpha)+xx(j, j)*alpha(j);
aj=xx(j, j);
if cj >lambda
alpha(j,1)=(cj-lambda)/aj;
elseif cj <-lambda
alpha  (j, 1)= (cj + lambda)/aj;
else
alpha  (j, 1)= 0;
end
end
```

We will refer to Method 1 using SVD as Truncated Singular Value Decomposition (TSVD) or Method 2.

Application to Crime Data

We refer to Table 18 on crime data by Thomas (1990) and provide an TSVD solution to that data. The complete data (see Appendix H) has five predictor

variables are two response variables. Descriptions of the variables are given as $(Y_1, Y_2, X_1, X_2, X_3, X_4, X_5)$ for each city.

Y_1 is the total overall reported crime rate per 1 million residents

Y_2 is the reported violent crime rate per 100,000 residents

X_1 is annual police funding in dollars per resident

X_2 is the percentage of people 25 years and older with 4 years of high school

X_3 is the percentage of 16 to 19 year-olds not in high school and not high school graduates

X_4 is the percentage of 18 to 24 year-olds in college

X_5 is the percentage of people 25 years and older with at least 4 years of college.

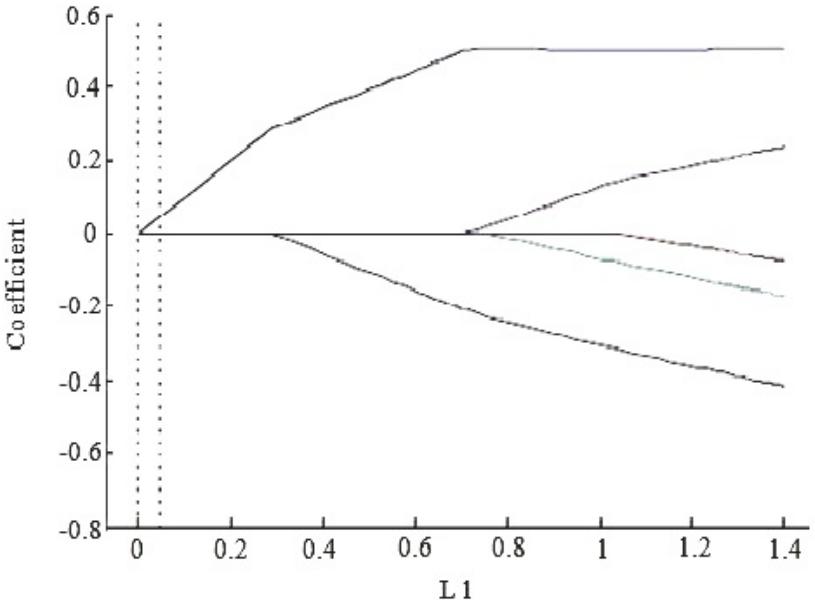

Figure 39: LASSO Shrinkage of Coefficients of Crime Data.

In Figure 39, the horizontal axis represents the absolute value of each coefficient which tends to 0 as the bound on the constraint increases. The vertical axis represents the coefficients of the variables. It can be seen that for much of the range of the bound most of the estimates tend to zero and hence corresponding predictor variables would be excluded from the model.

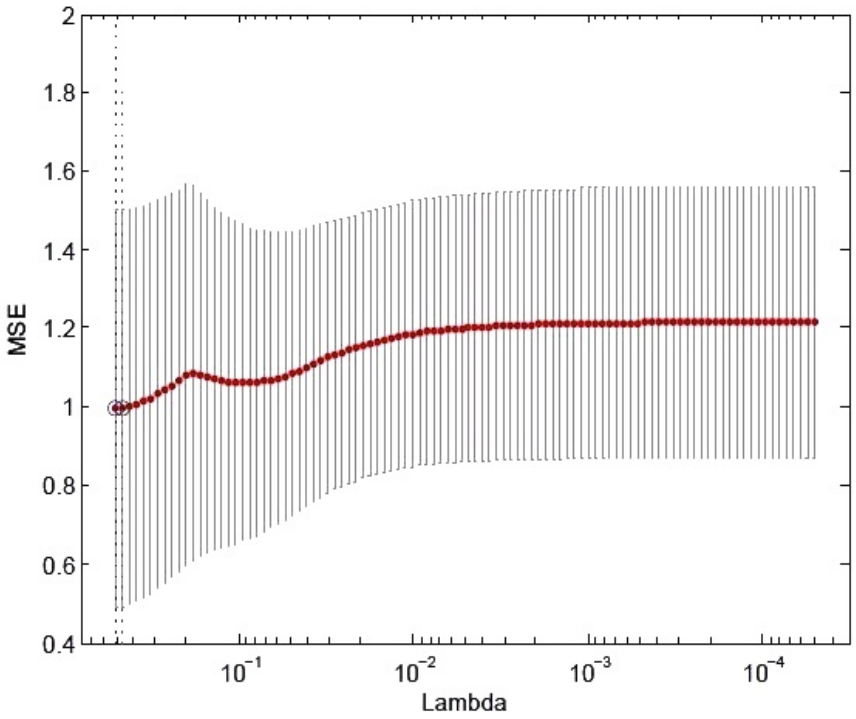

Figure 40: Cross-Validation of Crime Data.

Figure 40 displays the sequence of cross-validated mean square errors (MSE) of LASSO fit. It can be seen that both dotted vertical line appear to coincide on the same λ value of about $10^{-0.3}$. Therefore we choose $\lambda = 10^{-0.3}$ in determining the exact models in Table 39.

The principal components (PC) and the corresponding singular values obtained from a decomposition of the crime data is given in Table 38.

Table 38: Principal Component of Crime Data

Original Variables	1	2	3	4	5
x_1	−0.1106	0.7763	−0.4105	0.4627	0.0498
x_2	0.4496	0.4148	0.5730	−0.1369	0.5279
x_3	−0.5294	0.2056	−0.2259	−0.7096	0.3505
x_4	0.4745	−0.3107	−0.6307	0.0182	0.5293
x_5	0.5293	0.2940	−0.2331	−0.5131	−0.5620
Singular Value	11.2807	7.9964	5.6342	4.0677	2.3476

In this data, the first two PCs explains (78%) a little more than $\frac{3}{4}$ of the total variation $(11.2807^2 + \cdots + 2.3476^2 = 244.9982)$, which is quite high. The first PC alone accounts for about 52 % $(11.2807^2/244.9982)$ of variation. The results show that a solution in terms of the PCs would adequately represent information in the data. We therefore proceed to obtain a solution using TSVD. The solution that gives the smallest residual norm (5.9174) is given in Table 39. Thus, beyond the second PC, the third to the last singular values are found to contaminate this solution. The last three singular values are therefore truncated.

Table 39: TSVD Results for Crime Data

PC	TSVD Solution
1	0.388308154010534
2	−0.346418051073539
3	0.000000000000000
4	0.000000000000000
5	0.000000000000000
Residual Norm	5.91937624964826

Application to Prostate Cancer Data

The prostate cancer data comes from a study by Stamey et al. (1989) that examines the correlation between the level of prostate specific antigen and a number of clinical measures, in men who were about to receive a radical prostatectomy. The variables are described as follows:

log (cancer volume) (lcavol)

log (prostate weight) (lweight)

age

log (benign prostatic hyperplasia amount) (lbph)

seminal vesicle invasion (svi)

log (capsular penetration) (lcp)

Gleason score (gleason)

percentage Gleason scores 4 or 5 (pgg45)

We fit a linear model to log (prostate specific antigen) (lpsa) after first standardising the predictors. (See Appendix I for the dataset).

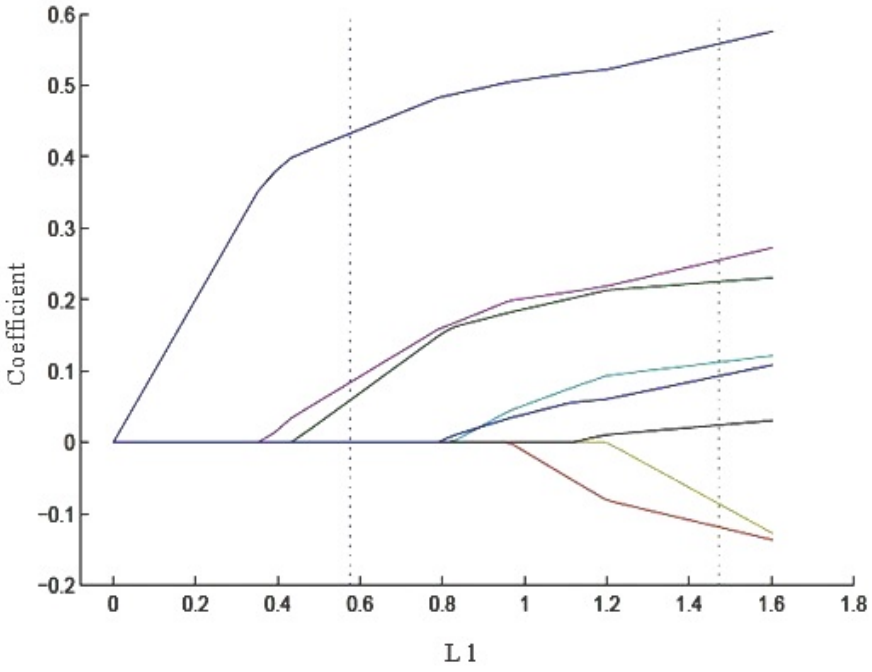

Figure 41: LASSO Shrinkage of Coefficients of Prostate Cancer Data.

Each curve represents a coefficient as a function of the (scaled) LASSO parameter; the broken line represents the model selected by generalised cross-validation. We notice that the absolute value of each coefficient tends to 0 as the bound on the constraint increases. The graph shows that we expect most of the predictors to be excluded from the model as for much of the range of the bound, many of the coefficients approach zero.

Figure 42 displays the sequence of cross-validated mean square errors (MSE) of LASSO fit associated with each of 100 λ values. It also shows the line segments for each point that represent intervals of estimate for each MSE value. The right vertical line and the left vertical line identifies the value of λ of about 10^{-1} which is the optimal value for all the λ values, therefore we choose $\lambda = 10^{-1}$ in determining the exact models in Tables 41 and 42.

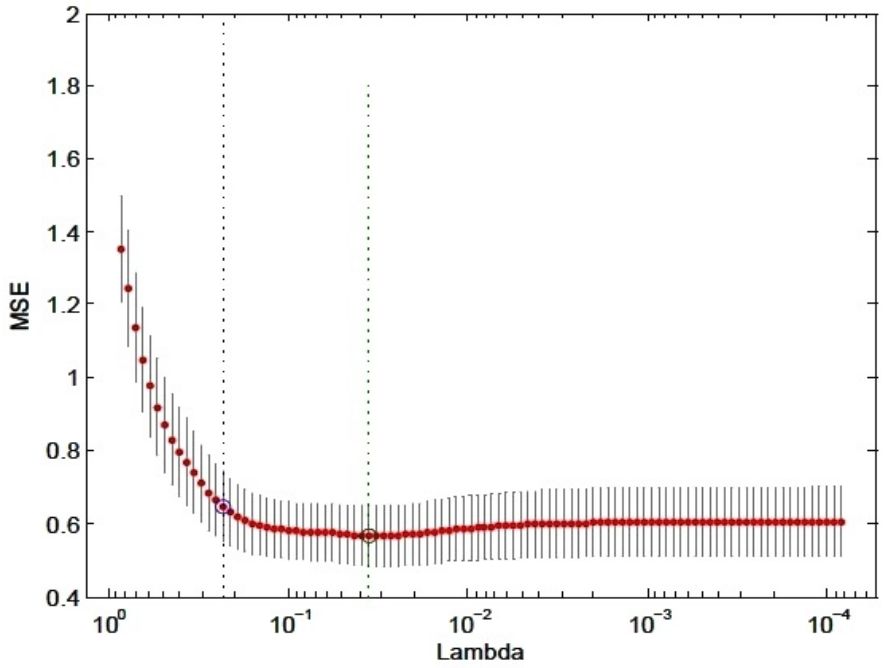

Figure 42: Cross-Validation of Prostate Cancer Data.

The principal components (PC) and the corresponding singular values obtained from a decomposition of the prostate cancer data are given in Table 40.

Table 40: Principal Component of Prostate Cancer Data

Original Variables	1	2	3	4	5	6	7	8
x_1	0.4222	−0.0537	0.3316	−0.1006	0.4059	−0.6250	−0.1735	0.3365
x_2	0.1871	0.5388	0.4225	0.1318	0.4379	0.5324	0.0105	−0.0598
x_3	0.2232	0.4686	−0.2424	−0.7928	−0.1421	−0.0586	0.1136	−0.0823
x_4	0.0856	0.6289	−0.0834	0.5105	−0.4358	−0.3626	−0.0828	0.0373
x_5	0.3902	−0.2074	0.3952	−0.1204	−0.5834	0.2827	−0.4604	0.0431
x_6	0.4642	−0.1901	0.1869	0.1360	−0.1205	−0.1240	0.6321	−0.5153
x_7	0.4057	0.0720	−0.5382	0.1581	0.2786	0.0492	−0.4911	−0.4409
x_8	0.4441	−0.0860	−0.4063	0.1593	−0.0310	0.3056	0.3119	0.6429
Singular Value	17.9619	12.5792	9.6780	7.7893	6.8115	6.5156	5.0211	4.3283

In this data, the first two PCs explains (0.63) close to $\frac{2}{3}$ of the total variation $(17.9619^2 + \cdots + 4.3283^2 = 768)$, which is quite high. The first PC alone accounts for about 42 % $(17.9619^2/768)$ of variation. The results show that a solution in terms of the PCs would adequately represent information in the data. We therefore proceed to obtain a solution using TSVD. The solution that gives the smallest residual norm (6.6233) is given in Table 41. Thus, beyond the second PC, the third to the last singular values are found to contaminate this solution. The last six singular values are therefore truncated.

Table 41: TSVD Results for Prostate Cancer Data

PC	TSVD Solution
1	0.436468124579854
2	0.375961055886647
3	0.000000000000000
4	0.000000000000000
5	0.000000000000000
6	0.000000000000000
7	0.000000000000000
8	0.000000000000000
Residual Norm	6.62327527246044

The result in Table 41 is compared with the results of the least squares method and the *l*1_*ls* solver for solving the non-smooth optimization problem given in Table 42. The table also shows a standard LASSO solution in MATLAB. The results show that a solution achieved under the TSVD could be preferred to the other methods as it achieves greater sparsity with equally small residual norm.

Table 42: Results for Prostate Cancer Data

Predictor	Least Squares	*l1_ls*	LASSO
lcavol	0.5762192831970794	0.5733143281345199	0.450526333979467
lweight	0.2308529420732906	0.2299117748855551	0.090767080155974
age	−0.1370451705556009	−0.1340392070333956	0
lbph	0.1215521381794642	0.1200507318437699	0
svi	0.2731706998275261	0.2702493221294212	0.110321946710557
lcp	−0.1284604953474022	−0.1215433677232128	0
gleason	0.0307963915113301	0.0297207123337756	0
pgg45	0.1089115924370662	0.1063076354285202	0
Residual Norm	5.68459336383028	5.68478361439537	6.481631353699084

In the next section, we consider another method for deriving a solution to the optimization problem under two specified conditions on the covariates.

Method 3

Orthonormal Covariates

We consider some basic properties of the LASSO estimator. Assuming first that the covariates are orthonormal so that $(x_i|x_j) = \delta_{ij}$, where $(.|.)$ is the inner product and δ_{ij} is the Kronecker delta, or, equivalently, $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, then using sub-gradient methods it can be shown that

$$\hat{\alpha_j} = \mathbf{S}m\lambda(\hat{\alpha_j}^{OLS}) = \hat{\alpha_j}^{OLS}\max\left(0, \quad 1 - \frac{m\lambda}{\left|\hat{\alpha_j}^{OLS}\right|}\right),$$

where $\hat{\alpha_j}^{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. The expression

$$\mathbf{S}_\alpha = \max\left(0, \quad 1 - \frac{m\lambda}{\left|\hat{\alpha_j}^{OLS}\right|}\right)$$

is referred to as the soft thresholding operator since it translates values towards

zero (making them exactly zero if they are small enough) instead of setting smaller values to zero and leaving larger ones untouched, which is the hard thresholding operator.

Proof

Given

$$g(\alpha)_{\alpha_0,\alpha} = \frac{1}{2m}\|\mathbf{X}\alpha - \mathbf{y}\|_2^2 + \lambda\|\alpha\|_1$$

and minimizing with respect to α, we obtain

$$\nabla g(\alpha) = \frac{1}{m}(\mathbf{X}^T\mathbf{X}\alpha - \mathbf{X}^T\mathbf{y}) = -\lambda S_j,$$

where

$$S_j = \begin{cases} 1, & \alpha_j^0 > 0 \\ -1, & \alpha_j^0 < 0 \\ 0, & \alpha_j^0 = 0 \end{cases}$$

is the sub-gradient of the L_1-norm regularization functional. Solving for α gives

$$\begin{aligned} \mathbf{X}^T\mathbf{X}\alpha &= \mathbf{X}^T\mathbf{y} - m\lambda S_j \\ \hat{\alpha}_j &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} - (\mathbf{X}^T\mathbf{X})^{-1}m\lambda S_j. \end{aligned}$$

Now,

$$\hat{\alpha}_j = \hat{\alpha}_j{}^{OLS} - (\mathbf{X}^T\mathbf{X})^{-1}m\lambda S_j. \tag{5.15}$$

In a special case where $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$, we consider three possibilities for $\hat{\alpha}_j$. These are:

1. If $\hat{\alpha}_j = \left|\hat{\alpha}_j{}^{OLS}\right| > 0, \quad$ set $S_i = 1$

2. If $\hat{\alpha}_j = -\left|\hat{\alpha}_j{}^{OLS}\right| < 0, \quad$ set $S_i = -1$

3. If $\hat{\alpha}_j = 0, \quad$ set $S_i = 0,$

If a coefficient from the least squares solution is already zero, that is, $\hat{\alpha}_j{}^{OLS} = 0,$ then we set $\hat{\alpha}_j = 0$ as well.

Then, Equation (5.15) now becomes

$$\begin{aligned} \hat{\alpha}_j &= \hat{\alpha}_j{}^{OLS} - m\lambda S_j \\ &= \hat{\alpha}_j{}^{OLS}\left(1 - \frac{m\lambda S_j}{\left|\hat{\alpha}_j{}^{OLS}\right|}\right). \end{aligned}$$

For $S_i = 1,$

$$\hat{\alpha}_j = \hat{\alpha}_j{}^{OLS}\left(1 - \frac{m\lambda}{\left|\hat{\alpha}_j{}^{OLS}\right|}\right). \tag{5.16}$$

For $S_i = -1,$

$$\begin{aligned} \hat{\alpha}_j &= -\left|\hat{\alpha}_j{}^{OLS}\right| - \lambda m(-1) \\ &= -\left|\hat{\alpha}_j{}^{OLS}\right| + \lambda m \end{aligned}$$

Therefore,

$$\hat{\alpha}_j = -\left|\hat{\alpha}_j{}^{OLS}\right|\left(1 - \frac{m\lambda}{\left|\hat{\alpha}_j{}^{OLS}\right|}\right). \tag{5.17}$$

Now, Equations (5.16) and (5.17) are the same, depending on the sign of S_j. Therefore, combining Equation (5.16) and $\hat{\alpha}_j = \hat{\alpha}_j{}^{OLS}$ gives

$$\hat{\alpha}_j = \hat{\alpha}_j{}^{OLS}\max\left(0, \quad 1 - \frac{m\lambda}{\left|\hat{\alpha}_j{}^{OLS}\right|}\right).$$

Thus, for the coefficients of the LASSO solution to go to zero, the expression of the threshold

$$1 - \frac{m\lambda}{\left|\hat{\alpha}_j{}^{OLS}\right|} < 0, \quad \text{or} \quad \lambda > \frac{\left|\hat{\alpha}_j{}^{OLS}\right|}{m},$$

and we achieve sparsity.

Similarly, if the threshold,

$$1 - \frac{m\lambda}{\left|\hat{\alpha}_j{}^{OLS}\right|} \geq 0, \quad \text{or} \quad \lambda \leq \frac{\left|\hat{\alpha}_j{}^{OLS}\right|}{m},$$

we achieve shrinkage by that factor.

Correlated Covariates

We consider the case where $\mathbf{X}^T\mathbf{X} \neq \mathbf{I}_p$. From Equation (5.15),

$$\hat{\alpha}_j = \hat{\alpha}_j{}^{OLS} - (\mathbf{X}^T\mathbf{X})^{-1} m\lambda S_j.$$

Now, let $\mathbf{X} = \mathbf{USV}^T$ be the SVD of $\mathbf{X}$.

$$\begin{aligned}
\mathbf{X}^T\mathbf{X} &= (\mathbf{USV}^T)^T(\mathbf{USV}^T) \\
&= \mathbf{V}(\mathbf{S}^T\mathbf{S})\mathbf{V}^T \\
(\mathbf{X}^T\mathbf{X})^{-1} &= (\mathbf{V}(\mathbf{S}^T\mathbf{S})\mathbf{V}^T)^{-1} \\
&= \mathbf{V}(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{V}^T \\
&= \mathbf{V}\begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & 0 & \cdots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \cdots & \frac{1}{\sigma_p^2} \end{pmatrix}\mathbf{V}^T
\end{aligned}$$

Therefore, considering the jth component of the coefficient, we have

$$\hat{\alpha}_j = \hat{\alpha}_j{}^{OLS} - (\mathbf{V}_j \frac{1}{\sigma_j^2} \mathbf{V}_j^T) m\lambda S_j.$$

Application of Method 3

We will consider both the crime and prostate cancer datasets in our illustrations. Table 43 gives a sparse solution of the crime data using the method. It identifies the first, second and the last variables as significant in the model. In addition, its associated residual norm is as small as the full model provided by

the least squares.

Table 43 : Results of Method 3 of Crime Data

	Least Squares	Method 3
	0.5061015532377721	0.306101553237772
	−0.4153060012346106	−0.215306001234611
	−0.0745112062624355	0.000000000000000
	−0.1693559131562814	0.000000000000000
	0.2361785503313412	0.036178550331341
Norm	5.70556065636911	5.97962522410037

Table 44 gives a sparse solution of the prostate data using Method 3. It identifies only three out of the eight variables as significant in the model. In addition, its associated norm is also as small as the full model provided by the least squares.

Table 44 : Results of Method 3 of Prostate Cancer Data

	Least Squares	Method 3
	0.5762192831970794	0.382219283197079
	0.2308529420732906	0.036852942073291
	−0.1370451705556009	0.000000000000000
	0.1215521381794642	0.000000000000000
	0.2731706998275261	0.079170699827526
	−0.1284604953474022	0.000000000000000
	0.0307963915113301	0.000000000000000
	0.1089115924370662	0.000000000000000
Norm	5.68459336383028	7.01149897737364

Chapter Summary

In this chapter, the focus has been on non-smoothing approximation of the L_1-norm problem. The first attempt is to cast the constrained formulation of the L_1-regularized least squares problem as an unconstrained formulation. This approach yields a clear L_1-norm regularization functional. By applying the sub-gradient method, an analytic optimal solution is obtained using two approaches. The derived methods specifies solutions for various scenarios of the component α_i of the solution vector. The solution has equivalently been presented under cases that depend on the optimal regularization parameter λ.

To assess the performance of the methods in determining the parameters for minimizing L_1-regularized least squares, a number of datasets have been used. These datasets are among those that are mostly known in the literature to have important inherent problems.

The problem of these datasets mainly regards the ill-conditioning which makes it difficult to obtain desirable solution for any method, and are therefore used as test data. A typical ill-conditioned data is the Hilbert matrix. Various dimensions of the Hilbert matrix, which is hypothetical in nature have been used with pre-assigned solution to the regularization problem, to determine how close the methods studied come close to retrieving the exact solution. It is observed that a regularization parameter of 10^{-3} yields the best result for Method 1. For this value of the parameter, Method 1 particularly has the smallest solution norm. However, the solution is not necessarily the best approximation but provides the best minimization of the residual norm. Under a non-optimal regularization parameter value of 10^{-1}, the L_2 regularization produces the best approximation. For this value, the LASSO that is based on the Elastic Net produces a feature selection solution which is not desired. Other datasets which are real are also selected to have similar properties as the Hilbert matrix. These

are datasets that appear to have monotonic trends with time series components. The datasets used for such study involve population growth for 18 years and global temperature anomalies for 166 years. They reveal that Method 1 is not suitable for determining trends in datasets with monotone trends. This property is found to be consistent with other LASSO methods, particularly the Elastic-Net-based LASSO. This result is not unexpected because there are no embedded techniques in the construction of those LASSO methods to detect monotonicity in datasets. It is however noticed that the L_2-norm regularization performs quite well in minimizing errors in such datasets.

The techniques are applied to another set of real datasets that involve crime data, prostate cancer data, level of ozone concentration and Boston housing crime dataset. The ozone concentration data yields different solutions with different techniques and produces the highest residual norm for the Elastic Net LASSO and Method 1. The two however are associated with small solutions norms. The Housing Crime data is observed to have a high residual even for the full model, and irrespective of the minimization method. It thus requires a every high regularization parameter ($\lambda = 10^0$) to achieve a desirable minimization of the least squares. Even under this condition, Method 1 produces a very small solution norm. The Elastic Net LASSO produces a highly sparse solution for this dataset.

An important observation made under the applications is that Method 1 does not appear to exhibit clear feature selection property for all the datasets used. A special data is therefore designed to determine the suitability of Method 1 for feature selection. For such dataset, the Hilbert matrix of order 20×12 is augmented by including two additional columns that are linear combinations of other columns. The 13th column is created to be the average of the first three columns, and the 14th is designed by the Gram Schmidt Orthogonalisation process such that it is orthogonal to the first two columns and are also linear

combinations of the first two. It is found that only Method 1 is able to set to zero the two additional columns. The Elastic Net based LASSO sets only one of the two columns to zero. This result indicates that Method 1 is very sensitive to variable dependence and treats them as dispensable, which is an important statistical property. However, the other LASSO methods may consider a dependent variable as indispensable. The result further indicates that the derived LASSO Method 1 is actually robust to over-regularization that has the tendency to set even relevant predictors to zero.

On the other hand, Method 3 is found to exhibit clear feature selection properties just like the standard LASSO method in MATLAB. Unlike the Method 1, Method 3 however is not robust to over-regularization. As a result, the choice of regularization parameter for this method does not follow from results of cross-validation used in Method 1.

CHAPTER SIX

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

In this chapter, we consider a summary of the entire thesis and then provide the conclusions and recommendations based on the findings.

Summary

The thesis has been motivated by the non-differentiability of the L_1-norm penalty in the L_1-regularized least squares problem. It focuses on optimization of the least squares function with the L_1-norm regularization of the parameters. The study has relied on a wide range of preliminary concepts and techniques. The concept involve orthogonalisation process, singular value decomposition, sub-gradient of a function and application to non-differentiable functions, polynomial fitting, optimality conditions and duality. The techniques include basically the L_1-norm regularization (LASSO) and Tikhonov regularization (Ridge). These preliminary tools have been reviewed within the context of over-determined systems.

The study has examined various smoothing approximations of the L_1-norm regularization functional, which include the Quadratic, Sigmoid and Cubic Hermite functional. Tikhonov regularization is applied to each of the resulting smooth least squares minimization problem using the Hilbert 12×12 matrix. The solution is compared with that of the Modified Newton's method which is mostly used in the literature. The approximations basically is a modification of the Lee's approximation to the L_1-norm term $\|\alpha\|_1$ given by $\sqrt{x^2+\varepsilon}$ for a small ε. The regularized solution using this approximation has been presented and also specified and proved in terms of the singular value decomposition. A quadratic approximation to the Lee's (2006) approximation is derived. The L_1-norm regularized least squares problem is then formulated in terms of this approximation

and the corresponding solution is derived. The solution so obtained appears similar to the L_2-norm regularized solution with a slight difference which is the constant multiple $\mu\mathbf{I}$ in the order zero Tikhonov regularization. The performance of the approximation has been assessed by using the Hilbert sub-matrix of order 12×7 and a chosen solution of $\alpha = \{1,\ 1, \ldots, 1\} \in \Re^7$. The assessment provides a quadratic singular value decomposition regularized solution of order zero for various values of the parameter μ in the set $10^{-35}, 10^{-30}, \ldots, 10^{0}$. It is found that smaller values of μ yields a good approximation to the exact solution and converges at $\mu = 10^{-30}$. The solution has a very small error $\|\alpha_{\text{exact}} - \hat{\alpha}\|$ which shows an accuracy to about nine digits. The results of the smoothing approximation are compared with the Modified Newton's method based on the Lee's approximation. A suitable initial guess of $\alpha_0 = 0.25 \times \text{ones}(7,1)$ is selected. It is also demonstrated that for $\varepsilon = 0.0001$,

$$\lim_{\varepsilon \to 0} |\alpha|_\varepsilon = \|\alpha\|_1 .$$

These values are used to perform a number of iterations which yields the best approximate solution at the 81st iterate at a parameter value of $\mu = 10^{-16}$ and a step size $\beta = 2$. The solution produces a small error which shows an accuracy of about 2 digits. Another smoothing approximation considered is the Sigmoid function approximation to the L_1-norm functional. The absolute value function is first expressed as the sum of the integral of two sigmoid functions. The gradient and the Hessian for the approximation are derived. It is demonstrated that for the parameter value of $\kappa = 1000$,

$$\lim_{\kappa \to \infty} |\alpha|_\kappa = \|\alpha\|_1 .$$

A formulation is given for the L_1-norm regularization in terms of the sigmoid approximation and corresponding regularized solution is derived and written in terms of the singular value decomposition. Analysis of the results gives the

sigmoid-SVD regularized solution for various values of the parameter μ. The solution converges at $\mu = 10^{-30}$ with error that shows an accuracy to nine digits. The Modified Newton's method based on the sigmoid approximation formulation yields the best approximate result at the 84th iterate with $\mu = 10^{-16}$ and step size $\beta = 3$ and function parameter value $\kappa = 300$. The error in the Modified Newton's method shows an accuracy to two digits.

The Cubic Hermite is the last smoothing approximation studied. We have derived an expression for Hermite form of the Cubic polynomial in terms of the parameter γ. It is demonstrated that for $\gamma = 0.05$,

$$\lim_{\gamma \to 0} |\alpha|_\gamma = \|\alpha\|_1 .$$

A formulation of the L_1-norm problem in terms of the Cubic Hermite is derived with corresponding regularized solution and its SVD version. An approximation of the Cubic Hermite-SVD regularized solution converges for parameter value $\mu = 10^{-30}$ with a small error that shows an accuracy to nine digits. The Modified Newton's method based on the Cubic Hermite approximation formulation yields the best solution at the 86th iterate with $\mu = 10^{-16}$, a step size $\beta = 3$ and function parameter $\gamma = 0.05$. The Modified Newton's method shows an accuracy to just one digit.

It is observed that the loss in accuracy in the results of the Modified Newton's method is as a result of the ill-conditioning of the matrix $\mathbf{X}^T\mathbf{X}$ in the solution. All three smoothing methods give results with the same level of accuracy (to nine digits) at the same parameter value of $\mu = 10^{-30}$. The solutions of the three smoothing methods are compared with the *l1_ls* method, a non-smooth method which is based on the Truncated Newton Interior Point method. For the same value of $\mu = 10^{-30}$, the accuracy in the solution of the *l1_ls* method is up to ten digits, almost the same as the smoothing approximations.

In order to specify a specific direction for our study on non-smoothing

approximation of the L_1-norm problem, we first attempt to cast the constrained formulation of the L_1-regularized least squares problem as an unconstrained formulation. Thus, unconstrained formulation gives a clear L_1-norm regularization functional. By applying the sub-gradient method, an analytic optimal solution is obtained. Method 1 specifies solution for three scenarios of the component α_i of the solution vector. The three cases are solution for $\alpha_i > 0$, $\alpha_i < 0$ and $\alpha_i = 0$. Equivalently, the solution has been presented under cases that depend on the optimal regularization parameter λ. In this case, the solution is provided so that it specifies the condition for which components of the coefficient parameter are not affected by large values of the regularization parameter leading to over-regularization. The derivation set a component (that is, a variable) to zero if the regularization parameter exceeds $\left\|2(\mathbf{X}^T\mathbf{y})_i\right\|_\infty$.

To assess the performance of Method 1 in determining the parameters for minimizing L_1-regularized least squares, a number of datasets have been used. These datasets are among those that are mostly known in the literature to have important inherent problems.

The problem of these datasets mainly regards the ill-conditioning which makes it difficult to obtain desirable solution for any method, and are therefore used as test data. A typical ill-conditioned data is the Hilbert matrix. Various dimensions of the Hilbert matrix, which is hypothetical in nature have been used with pre-assigned solution to the regularization problem, to determine how close the methods studied come close to retrieving the exact solution. It is observed that a regularization parameter of 10^{-3} yields the best result. For this value of the parameter, Method 1 particularly has the smallest solution norm but not necessarily the best approximation. The LASSO however, provides the best minimization of the residual norm. Under a non-optimal regularization parameter value of 10^{-1}, the L_2-norm regularization produces the best approximation. For this value, the Elastic Net based LASSO produces a feature selection solution

which is not desired in the case which involves the retrieval of an exact solution. Other datasets which are real are also selected to have similar properties as the Hilbert matrix. These are datasets that appear to have monotonic trends. They are mainly those that have time series components and appears suitable for polynomial fitting. The datasets used for this study, which involve population growth for 18 years and global temperature anomalies for 166 years, reveal that Method 1 is just not suitable for determining trends in datasets with such components. This property is found to be consistent with other LASSO methods, particularly the Elastic Net based LASSO. This result is not unexpected as by the construction of the LASSO methods, there are no embedded techniques to detect monotonicity in datasets. This result shows that for predicting trends, L_1-regularized least squares methods fused with additional constraints for trend detection and prediction should be used. It is however noticed that the L_2-norm regularization performs quite well in minimizing errors in such datasets.

The techniques are applied to another set of real datasets that involve level of Ozone Concentration and Boston Housing Crime dataset. The ozone concentration data which appears to yield different solutions with different techniques produces the highest residual norm for the Elastic Net LASSO and Method 1. The two however are associated with small solutions norms. The Housing crime data is noted to have a high residual even for the full model, and irrespective of the minimization method. It requires a every high regularization parameter $(\lambda = 10^0)$ for a desirable minimization of the least squares. Even under this condition, Method 1 produces a very small solution norm, only higher than the Elastic Net LASSO which even retains only 4 out of 13 predictor variables.

Throughout the applications to various datasets, it is observed that Method 1 does not appear to exhibit clear feature selection property. In most of the illustrations, using polynomial fit data, the Hilbert matrix with high (non-optimal) regularization parameter of 10^{-1}, the Ozone Concentration data, and

particularly the Housing Crime data, the Elastic Net LASSO sets some variables to zero. However, Method 1 does not set any variable completely to zero. A special data is therefore designed to determine the suitability of Method 1 for feature selection. To design such a data, the Hilbert matrix of order 20×12 is augmented by including two additional columns that are linear combinations of other columns. The 13th column is created to be the average of the first three columns, and the 14th is designed by the Gram-Schmidt Orthogonalisation process such that it is orthogonal to the first two columns and are also linear combinations of the first two. The two additional columns are thus linearly dependent on the other columns of the matrix. It is found that only Method 1 is able to set to zero the two additional columns. The Elastic Net based LASSO sets only one of the two columns to zero. This result indicates that Method 1 is very sensitive to variable dependence and treats them as dispensable, which is an important statistical property. However, the Elastic Net based LASSO may consider a dependent variable as indispensable. The result further indicates that Method 1 is actually robust to over-regularization that has the tendency to set even relevant predictors to zero. Method 1 sets to zero only when the variable is obviously redundant.

Conclusions

The thesis focused on optimization of the least squares function with the L_1-norm regularization of the parameters. Studies on the subject has gained overwhelming interest as a result of the need to improve upon the accuracy of models for various types of datasets. However, the non-differentiability of the L_1-norm penalty poses a major challenge to obtaining an analytic solution to the L_1-norm regularized least squares problem. Particular attention of the study is therefore directed at exploring smoothing and non-smoothing approximations that yields differentiable loss functional and ensures a close-form solution.

The study has relied on a wide range of preliminary concepts and techniques which involve orthogonalisation process, singular value decomposition, sub-gradient of a function, data fitting, optimality conditions, the L_1-norm regularization (LASSO) and Tikhonov regularization (Ridge).

Three main smoothing approximations of the L_1-norm regularization functional have been examined which include the Quadratic, Sigmoid and Cubic Hermite functional. Tikhonov regularization is then applied to each of the resulting smooth least squares minimization problem. Using the Hilbert 12×12 matrix, the solution of each approach is compared with that of the Modified Newton's method. The approximations basically are a modification of the Lee's approximation to the L_1-norm term $\|\alpha\|_1$ given by $\sqrt{x^2+\varepsilon}$ for a small ε. The regularized solution using this approximation has been presented and also specified and proved in terms of the singular value decomposition. Each smoothing approximation to the L_1-norm regularization functional is derived. The L_1-norm regularized least squares problem is then formulated in terms of the approximation and the corresponding regularized solution is derived and written in terms of the singular value decomposition. The performance of the approximation has been assessed by using the Hilbert sub-matrix of order 12×7 and a chosen solution of $\alpha = \{1,\ 1, \ldots, 1\} \in \Re^7$. It is found that for all three methods, smaller values of the regularization parameter μ yields a good approximation to the exact solution and converges at $\mu = 10^{-30}$. The solutions have a very small error $\|\alpha_{\text{exact}} - \hat{\alpha}\|$ which shows an accuracy to about nine digits. In each approximation, a suitable value of the approximation parameter is obtained as $\varepsilon = 0.0001$, $\kappa = 1000$, and $\gamma = 0.05$, respectively, for the Quadratic, Sigmoid and Cubic Hermite. For each of these values, it demonstrated that

$$\lim_{\tau \to \phi} |\alpha|_\tau = \|\alpha\|_1 ,$$

where τ represents the smoothing approximation parameter and ϕ is the corre-

sponding limiting value. The results of the smoothing approximations are compared with the Modified Newton's method based on the L_1-norm regularization functional. With a suitable initial guess, the Modified Newton's method yields approximate solution at various iterates for each smoothing method at the same parameter value of $\mu = 10^{-16}$ and various step sizes. The solution produces an error which shows an accuracy of not more than two digits. It is observed that the loss in accuracy in the results of the Modified Newton's method is as a result of the ill-conditioning of the matrix $\mathbf{X}^T\mathbf{X}$ in the solution.

The solutions of the three smoothing methods are compared with the *l*1_*ls* method, a non-smooth method which is based on the Truncated Newton Interior Point method. For the same value of $\mu = 10^{-30}$, the accuracy in the solution of the *l*1_*ls* method is almost the same as the smoothing approximations.

The study on non-smoothing approximation of the L_1-norm problem, focuses on casting the constrained formulation of the L_1-regularized least squares problem as an unconstrained formulation. By applying the sub-gradient method, an analytic optimal solution is obtained.

The performance of Method 1, in determining the parameters for minimizing L_1-regularized least squares, is assessed using a number of datasets. These datasets, which includes various dimensions of the Hilbert matrix, are among those that are mostly known in the literature to have important inherent problems and can therefore be used as test datasets for derived methods. In order to expose the inherent problems of the datasets used, which number six in all, four other known methods have been assessed in addition to Method 1. The use of these other methods also provides a basis for assessing the performance of Method 1.

In almost all the datasets, Method 1 particularly is associated with the smallest solution norm but not necessarily the best approximation. Another important feature of Method 1 is that it is not suitable for determining trends in data-

sets with monotonic trends. This property is found to be consistent with other LASSO methods, particularly the Elastic Net based LASSO. This result is not unexpected as by the construction of the LASSO methods, there are no embedded techniques to detect monotonicity in datasets. This result shows that for predicting trends, L_1-regularized least squares methods fused with additional constraints such as isotonic regression for trend detection and prediction should be used. It is however noticed that the L_2-norm regularization performs quite well in minimizing errors in such datasets.

Attempts at achieving sparsity of the analytic solution has been made in two ways. The initial solution is expressed in terms of the singular value decomposition so that by truncating smaller singular values, the desired sparsity is achieved using suitable regularization parameter obtained by the K-fold cross-validation of the fit. In another way, the solution itself has been induced to ensure sparsity by designing the algorithm to enforce sparsity with a suitable choice of the regularization parameter.

The results show that the LASSO formulation and solution must be appropriately designed for certain type of datasets, particularly those that are severely ill-conditioned and those with monotone trends.

Recommendations

The results of the study show that the LASSO method may not be appropriate for all datasets. For example, for datasets that are severely ill-conditioned and those with monotone trends, the LASSO formulation and its solution should be designed appropriately. In this case, the choice of the regularization parameter must be carefully selected.

Using a specially designed data matrix with dependent columns, the applications show that Method 1 exhibits clear feature selection property only when predictor variables are linearly dependent. It is found that under this condition,

Method 1 is able to set to zero all linearly dependent columns of the data matrix. The Elastic Net-based LASSO, which is observed to produce sparse solution in most cases, surprisingly does not set all dependent columns to zero. This result indicates that Method 1 is very sensitive to variable dependence and treats them as dispensable, which is an important statistical property. It is therefore recommended in datasets that are prone to have dependent predictors. However, the Elastic Net based LASSO may consider a dependent variable as indispensable. The result further indicates that a method for solving the LASSO problem could be robust to over-regularization which may not easily set relevant predictors to zero. However, other methods may not be robust to over-regularization. Such methods may require a very careful choice of the regularization parameter.

This study is carried out in the context of over-determined systems. Further studies on the subject could focus on approximation methods in under-determined systems.

REFERENCES

Annergren, M. (2012). ADMM for L_1-regularized optimization problems and applications oriented input design for MPC (Unpublished doctoral dissertation). KTH School of Electrical Engineering, Sweden.

Argyriou, A., Evgeniou, T., & Pontil, M. (2007). Multi-task feature learning. *Advances in Neural Information Processing Systems*, 19, (pp. 41-48).

Banerjee, O., El Ghaoui, L., & Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9, 485-516.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2(1):183-202.

Belsley, P., Kuh, M., & Welsch, A. (1980). 'Regression diagnostics', Wiley, 244-261.

Bemporad, A., Morari, M., Dua, V., & Pistikopoulos, E. (2002). The explicit linear quadratic regulator for constrained systems. *Automatica*, 38(1), 320.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer, New York.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press, New York, USA.

Candes, E., & Tao, T. (2007). The dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6), 2313-2351.

Carroll, R., Ruppert, D., & Stefanski, L. (1995). *Measurement error in nonlinear models*. Chapman & Hall/CRC, London.

Chen, C., & Mangasarian, O. L. (1996). A class of smoothing functions for nonlinear and mixed complementarity problems. *Computational Optimization and Applications*, 5(2), 97-138.

Chen, S. S., Donoho, D. L., & Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. on Scientific Computing*, 20(1), 33-61.

Chen, S. S., Donoho, D. L., & Saunders, M. A. (1999). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1), 33-61.

Degroat, R., & Dowling, E. (1991). The data least squares problem and channel equalization. *IEEE Trans. Signal Process*, 41, 407-411.

Donoho, D. (2004). For most large under-determined systems of linear equations the minimal L_1-norm solution is also the sparsest solution. (Technical Report). Statistics Dept., Stanford University.

Duchi, J., Shalev-Shwartz, S., Singer, Y., & Chandra, T. (2008). Efficient projections onto the L_1-ball for learning in high dimensions. In Proc. Int. Conf. Mach. Learn. (ICML), Helsinki, Finland.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2002). Least angle regression (Tech. Rep.). Technical report, Stanford University.

Efron, B., Johnstone, T. H., & Tibshirani, R. (2004). Least angle regression. Analysis of Statistics, 32(2), 407-499.

Figueiredo, M. A. T., Nowak, R. D., & Wright, S. J. (2007) "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing*, 1(4), pp. 586-597.

Farahmand, M., Ghavamzadeh, A. M., Szepesvari, C., & Mannor, S. (2009). Regularized policy iteration. Neural information processing systems.

Ferreau, H. J., Bock, H. G., & Diehl, M. (2008). An online active set strategy to overcome the limitations of explicit MPC. *International Journal of Robust and Nonlinear Control*, 18(8), 816-830.

Friedman, J., Hastie, T., Hoefling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*; 2(1):302-332.

Friedman, J., Hastie, T., Hoefling, H., & Tibshirani, R. (2008). Pathwise coor-

dinate optimization. *Annals of Applied Statistics*, 2(1), 302-332.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*., 33(1), 1-22.

Fu, W. (1998). Penalized rregression: The bridge verses the Lasso. *Journal of Computational and Graphical Statistics*, 7(3), 397-416.

Fuller, W. (1987). *Measurement Error Models*. Wiley, New York.

Gander, W., Gander, M. J., & Kwok, F. (2014). *Scientific Computing-An Introduction using MAPLE and MATLAB*. XVIII 905, Springer.

Genkin, A., Lewis, D., & Madigan, D. (2007). Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3), 291-304.

Gleser, L. (1981). Estimation in a multivariate errors in variables regression model: large sample results. *The Annals of Statistics*, 9, 24-44.

Golub, G. H. & Reinsch, C. (1970). "Singular value decomposition and least squares solutions," *Numer. Math*., 14, 403-420.

Golub, G. H. & Van Loan, C. F. (1980). An analysis of the total least squares problem. *SIAM J. Numer. Anal*., 17(6): 883-893.

Goodman, J. (2004). Exponential priors for maximum entropy models. In HLT-NAACL, 305-312.

Grandvalet, Y., & Canu, S. (1998). Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. In NIPS, 445-451.

Gregor, K., & LeCun, Y. (2010). Learning fast approximations of sparse coding. ICML, 399-406.

Harrison, D., & Rubinfeld, D. L. (1978). 'Hedonic prices and the demand for clean air', *J. Environ. Economics & Management*, 5, 81-102.

Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5, 1391-1415.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The Elements of Statistical*

learning, Springer.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, (2nd Ed.), Springer Verlag, New York.

Hsieh, M. A., Sustik, C. J., Dhillon, I. S., & Ravikumar, P. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. NIPS, 2330-2338.

James, G. M., Paulson, C., & Rusmevichientong, P. (2013). Penalized and Constrained Regression. Unpublished Manuscript, University of Southern California.

Jones, P. D., Osborn, T. J., & Briffa, K. R. (2016). Monthly and Annual Temperature Anomalies (degrees C), 1850-2015. Climate Research Unit, School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, United Kingdom.

Kim, Y., & Kim, J. (2004). Gradient Lasso for feature selection. In Proceedings of the twenty-first international conference on machine learning, page 60, New York, USA Press.

Kim, S. J., Koh, K., Lustig, M., Boyd, S., & Gorinevsky, D. (2007). "An Interior-Point method for large-scale L_1-regularized least squares," *IEEE Journal of Selected Topics in Signal Processing*, 4(1), 606-617.

Krishnapuram, B., & Hartemink, A. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 957-968.

Lawson, C. L. & Hanson, R. J. (1974). *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, N. J.

Lee, S. I., Lee, H., Abbeel, P., & Ng, A. Y. (2006). "Efficient L_1-regularized logistic regression," In Proceedings of the Twenty-first National Conference on Artificial Intelligence (AAAI), 1-9.

Lee, H., Battle, A., Raina, R., & Ng, A. Y. (2007). "Efficient sparse coding algorithms." In Advances in Neural Information Processing Systems 19, MIT Press.

Low, S., & Lapsley, D. E. (1999). Optimization flow control: Basic algorithm and convergence. *IEEE/ACM Transactions on Networking*, 7, 861 -874.

Meier, L., van de Geer, S., & Buhlmann, P. (2006). The group Lasso for logistic regression (Technical Report). ETH Seminar fur Statistik.

Meier, L., Van De Geer, S., & Buhlmann, P. (2008). The group Lasso for logistic regression, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1), 53-71.

Ng, A. Y. (2004). Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In Proceedings of the twenty-first international conference on machine learning, pp. 78, New York, USA.

Obozinski, G., Taskar, B., & Jordan, M. (2006). Multi- task feature selection (Technical Report). Statistics Dept., University of California, Berkeley.

Osborne, M., Presnell, B., & Turlach, B. (2000). On the Lasso and its dual. *Journal of Computational and Graphical Statistics*, 9, 319-337.

Park, M. Y., & Hastie, T. (2006). Regularization path algorithms for detecting gene interactions (Technical Report). Stanford University.

Park, M. Y., & Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society* Series B; 69: 659-677.

Parker, D. E. (2016). Monthly and Annual Temperature Anomalies (degrees C), 1850-2015. Hadley Center for Climate Prediction and Research. Meteorological Office. Bracknell, Berkshire, United Kingdom.

Perkins, S., Lacker, K., & Theiler, J. (2003). Grafting: Fast, incremental feature selection by gradient descent in function space. *Journal of Machine Learn-*

ing Research, 3, 1333-1356.

Quattoni, A., Carreras, X., Collins, M., & Darrell, T. (2009). An efficient projection for L_1 regularization. In Proceedings of the 26th International Conference on Machine Learning, 857-864.

Quinlan, R. (1993). Combining Instance-Based and Model-Based Learning. In Proceedings on the Tenth International Conference of Machine Learning, 236-243, University of Massachusetts, Amherst. Morgan Kaufmann.

Rao, C. V., Wright, S. J., & Rawlings, J. B. (1998). Application of interior-point methods to model predictive control. *Journal of Optimization Theory and Applications*, 99, 723-757.

Richter, S., Jones, C. N., & Morari, M. (2009). Real-time input constrained MPC using fast gradient methods. In Proceedings of the 48th IEEE Conference on Decision and Control, 7387-7393.

Rosset, S., Zhu, J., & Hastie, T. (2004). Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5, 941-973.

Rosset, S., & Zhu, J. (2007). Adaptable, efficient and robust methods for regression and classification via piecewise linear regularized coefficient paths.

Roth, V. (2004). The Generalized Lasso. In IEEE transactions on neural networks, 15(1).

Rudin, L. I., Osher, S., & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms.*Physica D: Nonlinear Phenomena*, 60(1-4), 259 -268.

Sardy, S., Bruce, A., & Tseng, P. (1998). Block coordinate relaxation methods for nonparametric signal denoising with wavelet dictionaries (Tech. Rep.). Technical report, Seattle, WA.

Sardy, S., Tseng, P. & Bruce, A.G. (2001). Robust wavelet denoising. *IEEE Transactions on Signal Processing*, 49(6).

Schmidt, M. (2005). Least Squares Optimization with L_1-Norm Regularization. CS542B Project Report.

Schmidt, M., Murphy, K., Fung, G., & Rosale, R. (2008). Structure learning in random fields for heart motion abnormality detection. Proc. of Conf. on Computer Vision and Pattern Recognition.

Schmidt, M., Van Den Berg, E., Friedlander, M., & Murphy, K. (2009). Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. Proc. of Conf. on Artificial Intelligence and Statistics (pp. 456463).

Shevade, K., & Keerthi, S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19, 2246-2253.

Simila, T., & Tikka, J. (2007). Input selection and shrinkage in multi-response linear regression. *Computational Statistics and Data Analysis*, 52, 406 422.

Srikant, R. (2004). Mathematics of Internet Congestion Control. Birkhauser.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. & Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II radical prostatectomy treated patients, *Journal of Urology* , 16, 1076-1083.

Thomas, G. S. (1990). *The Rating Guide to Life in Americas Small Cities*, Prometheus books.

Tibshirani, R. (1994). Regression shrinkage and selection via the Lasso. In Technical report, University of Toronto.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*. Series B (Methodological), 58(1), 267-288.

Tibshirani, R. J. (2013). The Lasso problem and uniqueness. *Electronic Journal of Statistics*, 7(0):1456- 1490.

Tibshirani, R. J. & Taylor, J. (2011). The Solution Path of the Generalized

Lasso, *The Annals of Statistics*, 39, 1335-1371.

Tibshirani R. (1997). The Lasso method for variable selection in the cox model. Statistics in Medicine; 16:385-395. [PubMed: 9044528].

Turlach, B., Venables, W., & Wright, S. (2005). Simultaneous variable selection. *Technometrics*, 47, 349-363.

Van der Kooij, A. (2007). Prediction accuracy and stability of regression with optimal scaling transformations. A Technical report. Leiden University: Dept. of Data Theory.

Van Huffel, S., & Vandewalle, J. (1989). Analysis and properties of the generalized total least squares problem when some or all columns are subject to error. *SIAM J. Matrix Anal.*, 10, 294-315.

Wang, Y., & Boyd, S. (2010). Fast model predictive control using online optimization. *IEEE Transactions on Control Systems Technology*, 18(2), 267-278.

Wu, T., Chen, Y., Hastie, T., Sobel, E., & Lange, K. (2009). Genome-wide association analysis by penalized logistic regression. *Bioinformatics*, 25(6), 714-721.

Wu, T., & Lange, K. (2008a). Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics*, 2(1), 224-244.

Wu, W. B., Woodroofe, M., & Mentz, G. (2001). Isotonic Regression: Another Look at the Changepoint Problem, *Biometrica*, 88, 793 - 804.

Yongdai, K. & Jinseog, K. (2004). Gradient Lasso for Feature Selection. In ICML: Proceedings of the twenty-first international conference on Machine learning, 60, New York, NY, USA. ACM Press.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, Series B, 68(1), 49-67.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with

grouped variables, *Journal of the Royal Statistical Society*: Series B (Statistical Methodology), 68(1), 49-67.

Zhang, H., Wahba, Y., G. Lin, Voelker, M., Ferris, M., Klein, R., & Klein, B. (2002). Variable selection and model building via likelihood basis pursuit (Tech. Rep.). Technical Report 1059, University of Wisconsin, Department of Statistics.

Zheng, A. Z., Jordan, M. I., Liblit, B., & Aiken, A. (2004). Statistical debugging of sampled programs. Advances in neural information processing systems 16. mit press, Cambridge, MA.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Royal. Stat. Soc*. B, 67(2), 301-320.

APPENDICES

APPENDIX A

REGULARIZED SOLUTION OF HILBERT MATRIX USING QUADRATIC APPROXIMATION

Function File

$\mathbf{H} = \text{hilb}\ (12)$

$\mathbf{X} = \mathbf{H}(:, 1:7)$

$\alpha = \text{ones}\ (7,1)$

$\mathbf{y} = \mathbf{X} * \alpha$

$[\mathbf{USV}] = \text{svd}\ (\mathbf{X})$

$\sigma = \text{diag}\ (S)$

$\text{Utb} = \mathbf{U}' * \mathbf{y}$

$k = -30$

$\lambda = 10^k$

$\varepsilon = 0.0001$

for j=1:7

$q(:, j) = (\sigma(j)/((\sigma(j))^2 + \lambda * 0.5 * \varepsilon^{-0.5})) * (Utb(j) * \mathbf{V}(:, j))$

end

$\text{qt} = \text{sum}\ (q')'$

$\text{error}_{\text{qt}} = \text{norm}\ ((x - \text{qt}), \text{inf})$

APPENDIX B

CODE FOR IMPLEMENTING MODIFIED NEWTON'S METHOD FOR QUADRATIC APPROXIMATION

Function File

$\alpha = [\alpha(1)\ \alpha(2)\ \alpha(3)\ \alpha(4)\ \alpha(5)\ \alpha(6)\ \alpha(7)]'$

$\mathbf{H} = \text{hilb}\ (12)$

$\mathbf{X} = \mathbf{H}(:, 1:7)$

$\alpha 0 = \text{ones}\ (7,1)$

$\mathbf{y} = \mathbf{X} * \alpha 0$

$\mathbf{b} = \mathbf{X} * \alpha - \mathbf{y}$

$L = \mathbf{b}' * \mathbf{b}$

Define Lee's Approximation function

$\varepsilon = 0.0001$

for $i = 1:7$

$q = \sqrt{\alpha(i)^2 + \varepsilon}$

end

$Q = \text{sum}\ (q')$

$\lambda = 10^{-16}$

$g = L + \lambda * Q$

Modified Newton's Method File

Initial guess, $\alpha 0 = 0.25 * \text{ones}\ (7,1)$

Pre-allocation for the iterations

$\alpha = \text{zeros}\ \ (\text{length}\ (\alpha 0), 101)$

$\alpha(:,1) = \alpha 0$

Pre-allocation for computing the gradients

$\text{Grd} = \text{zeros}\ \ (\text{length}(\alpha 0), 100)$

100 Iterations

for $i = 1 : 100$

Gradient of the Objective function

$\text{Grd}\ (:,i) = \text{fdjac}\ (\text{@Function File})$

Hessian of the Objective function

$\mathbf{H} = \text{fdhess}\ (\text{@Function File})$

Newton's method

$\alpha(:,i+1) = \alpha(:,i) - 2 * \mathbf{H}\ Grd(:,i)$

end

Solution vector X

$\alpha 100 = \alpha(:,81 : 100)$ Gradient at the end of the 100th iteration

$\text{Grad}\ \ = \text{Grd}\ (:,81 : 100)$

Hessian at the end of the 100th iteration

$\mathbf{H}$

Test for Positive Definiteness

$\text{Lt} = \text{chol}\ (\mathbf{H})$

APPENDIX C

REGULARIZED SOLUTION OF HILBERT MATRIX USING SIGMOID FUNCTION APPROXIMATION

Function File

$\mathbf{H} = \text{hilb}\ (12)$

$\mathbf{X} = \mathbf{H}(:, 1:7)$

$\alpha = \text{ones}\ (7,1)$

$\mathbf{y} = \mathbf{X} * \alpha$

$[\mathbf{USV}] = \text{svd}\ (\mathbf{X})$

$\sigma = \text{diag}\ (\mathbf{S})$

$\text{Utb} = \mathbf{U}' * \mathbf{y}$

$\kappa = 1000$

$k = -30$

$\lambda = 10.^{k}$

for $j = 1:7$

$\alpha sol7(:, j) = (\sigma(j)/((\sigma(j))^2 + 1/4 * \lambda * \alpha)) * (Utb(j) * \mathbf{V}(:, j))$

end

$\alpha tSVD7 = \text{sum}\ (\alpha sol7')'$

$\text{error}_{\alpha} tSVD7 = \text{norm}\ ((\alpha - \alpha tSVD7), \text{inf})$

APPENDIX D

CODE FOR IMPLEMENTING MODIFIED NEWTON'S METHOD FOR SIGMOID FUNCTION APPROXIMATION

Function File

$\alpha = [\alpha(1)\ \alpha(2)\ \alpha(3)\ \alpha(4)\ \alpha(5)\ \alpha(6)\ \alpha(7)]'$

$\mathbf{H} = \text{hilb}\ (12)$

$\mathbf{X} = \mathbf{H}(:, 1:7)$

$\alpha 0 = \text{ones}\ (7,1)$

$\mathbf{y} = \mathbf{X} * \alpha 0$

$\mathbf{b} = \mathbf{X} * \alpha - \mathbf{y}$

$L = \mathbf{b}' * \mathbf{b}$

Define the Sigmoid function

$\kappa = 300$

for $i = 1:7$

$s = 1/\kappa * (\log\ (1 + \exp\ (-\kappa * x(i))) + \log\ (1 + \exp\ (\kappa * x(i))))$

end

$S = \text{sum}\ (s')$

$\lambda = 10^{-16}$

$g = L + \lambda * S$

Modified Newton's Method File
Initial guess $\alpha 0 = 0.25 * \text{ones}\ (7,1)$
Pre-allocation for the iterations
$\alpha = \text{zeros}\ (\text{length}\ (\alpha 0), 101)$
$\alpha(:,1) = \alpha 0$
Pre-allocation for computing the gradients
$\text{Grd} = \text{zeros}\ (\text{length}\ (\alpha 0), 100)$
100 Iterations
for $i = 1 : 100$
Gradient of the Objective function
$\text{Grd}\ (:,i) = \text{fdjac}\ (\text{@Function File})$
Hessian of the Objective function
$\mathbf{H} = \text{fdhess}\ (\text{@Function File})$
Newton's method
$\alpha(:,i+1) = \alpha(:,i) - 3 * \mathbf{H}\ \text{Grd}\ (:,i)$
end
Solution vector X
$\alpha 100 = \alpha(:, 81 : 100)$
Gradient at the end of the 100th iteration
$\text{Grad} = \text{Grd}\ (:, 81 : 100)$
Hessian at the end of the 100th iteration
$\mathbf{H}$
Test for Positive Definiteness
$\text{Lt} = \text{chol}\ (\mathbf{H})$

APPENDIX E

REGULARIZED SOLUTION OF HILBERT MATRIX USING CUBIC HERMITE APPROXIMATION

Function File

$\mathbf{H} = \text{hilb}\ (12)$

$A = H(:, 1:7)$

$\alpha = \text{ones}\ (7, 1)$

$\mathbf{y} = \mathbf{X} * \alpha$

$[\mathbf{USV}] = \text{svd}\ (\mathbf{X})$

$\sigma = \text{diag}(\mathbf{S})$

$Utb = \mathbf{U}' * \mathbf{y}$

$\gamma = 0.05$

$n = 1/(2 * \gamma)$

$k = -30$

$\lambda = 10.^{k}$

for $j = 1:7$

$c(:, j) = (\sigma(j)/(((\sigma(j))^2 + \lambda * n))) * (Utb(j) * \mathbf{V}(:, j))$

end

$ct = \text{sum}(c')'$

$\text{error}_{\text{ct}} = \text{norm}\ ((\alpha - \text{ct}), \text{inf})$

APPENDIX F

CODE FOR IMPLEMENTING MODIFIED NEWTON'S METHOD FOR CUBIC HERMITE APPROXIMATION

Function File

$\alpha = [\alpha(1)\ \alpha(2)\ \alpha(3)\ \alpha(4)\ \alpha(5)\ \alpha(6)\ \alpha(7)]'$

$\mathbf{H} = \text{hilb}\ (12)$

$\mathbf{X} = \mathbf{H}(:, 1:7)$

$\alpha 0 = \text{ones}\ (7,1)$

$\mathbf{y} = \mathbf{X} * \alpha 0$

$\mathbf{b} = \mathbf{X} * \alpha - \mathbf{y}$

$L = \mathbf{b}' * \mathbf{b}$

Define the Cubic Hermite Approximation

$\gamma = 0.05$

for $i = 1:7$

$c = \gamma/2 + (1/(2 * \gamma)) * \alpha(i)^2$

end

$C = \text{sum}\ (c')$

$\lambda = 10^{-16}$

$g = L + \lambda * C$

Modified Newton's Method

Initial guess $\alpha 0 = 0.25 * \text{ones}\ (7,1)$

Pre-allocation for the iterations

$\alpha = \text{zeros}\ (\text{length}\ (\alpha 0), 101)$

$\alpha(:,1) = \alpha 0$

Pre-allocation for computing the gradients

$\text{Grd} = \text{zeros}\ (\text{length}\ (\alpha 0), 100)$

100 Iterations

for$i = 1 : 100$

Gradient of the Objective function

$\text{Grd}(:,i) = \text{fdjac}\ (\text{@Function File})$

Hessian of the Objective function

$\mathbf{H} = \text{fdhess}\ (\text{@Function File})$

Newton's method

$\alpha(:,i+1) = \alpha(:,i) - 3 * H\ Grd(:,i)$

end

Solution vector X

$\alpha 100 = \alpha(:,81:100)$ Gradient at the end of the 100th iteration

$\text{Grad} = \text{Grd}\ (:,81:100)$

Hessian at the end of the 100th iteration

H

Test for Positive Definiteness

$\text{Lt} = \text{chol}(H)$

APPENDIX G

POLYNOMIAL FIT OF DEGREE 5 OF POPULATION DATA

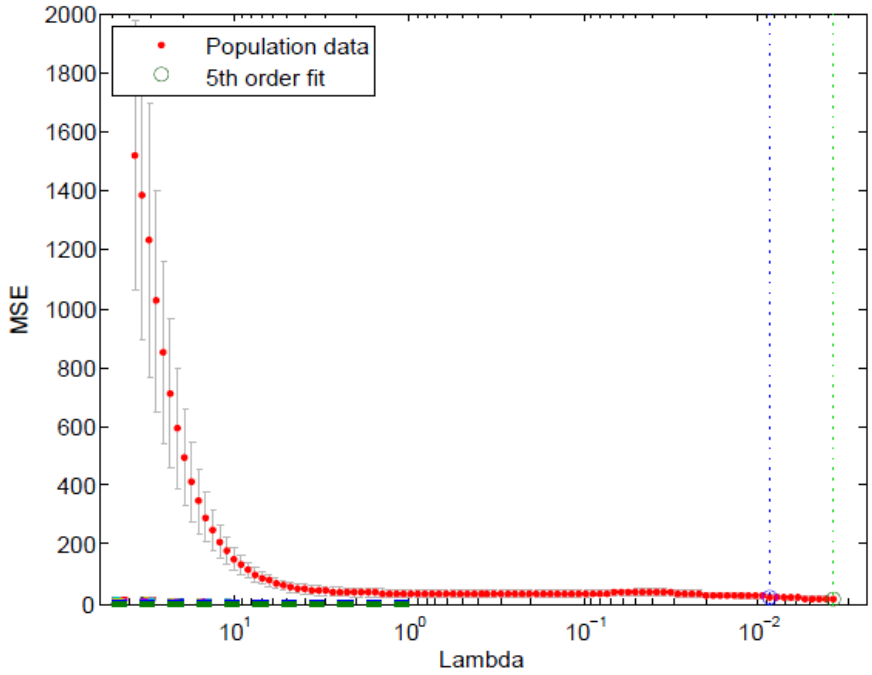

Figure G1: Cross-Validation LASSO Fit of Polynomial of Degree Five.

Table G1: Solutions for Various Methods at Optimal Value of $\lambda = 10^{-2}$ of Population Data

	Least Squares	*l1_ls*	Method 1
	4.532808404662905	4.529397922858414	39.714578947368423
	0.270778590462724	0.271084245838057	0.156957491702229
	0.011213985497136	0.011205384831343	0.000520643555525
	−0.000330004731447	−0.000329901365519	0.000002059429493
	0.000002703345127	0.000002702789559	0.000000008404938
	−0.000000006305520	−0.000000006304421	0.000000000034978
Norm	4.540902877271965	4.537516708666442	39.714889108394026

Table G1 Continued

Variable	Least Squares	Ridge	LASSO
Intercept	4.532808404662905	4.494278702234064	5.216745350108717
1	0.270778590462724	0.274260197310556	0.535161413592931
2	0.011213985497136	0.011115578759935	−0.007426407969439
3	−0.000330004731447	−0.000328818710862	0.000022636191828
4	0.000002703345127	0.000002696958227	0.000000189301638
5	−0.000000006305520	−0.000000006292871	−0.000000000261031
Norm	4.540902877271965	4.502652926154060	0.535212939535794

Table G2: Solution and Residual Norms of Polynomial Fit of Population Data

Method	Solution Norm $\|\|\alpha\|\|_2$	Residual Norm $\|\|\mathbf{y} - \mathbf{X}\alpha\|\|_2$
Least Squares	4.540902877271965	11.086831207477713
l1_ls	4.537516708666442	11.086831816079259
Method 1	39.714889108394026	139.8848942555254
Ridge	4.502652926154060	11.086908871167171
LASSO	0.535212939535794	26.911746501461966

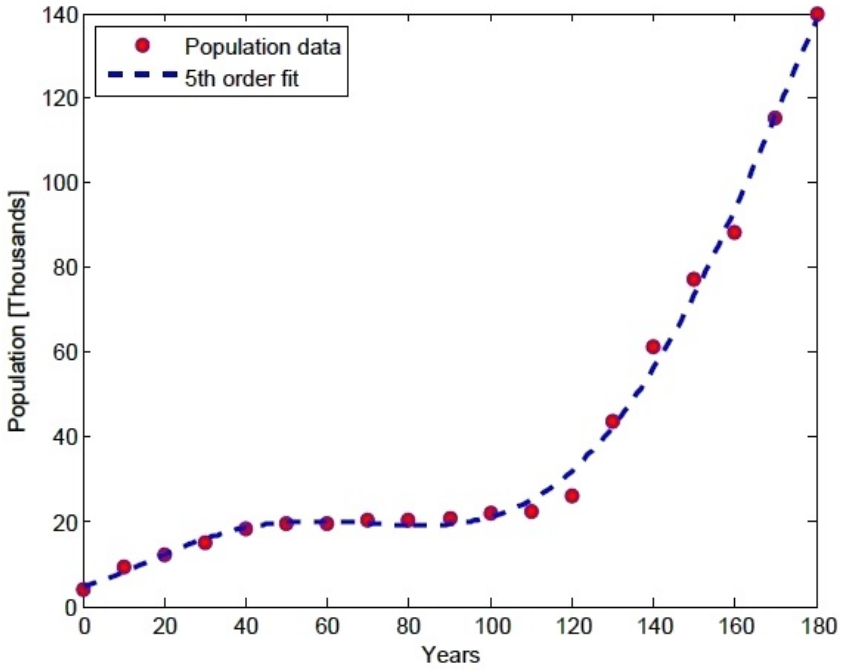

Figure G2 : Least Squares Polynomial Fit of Order Five.

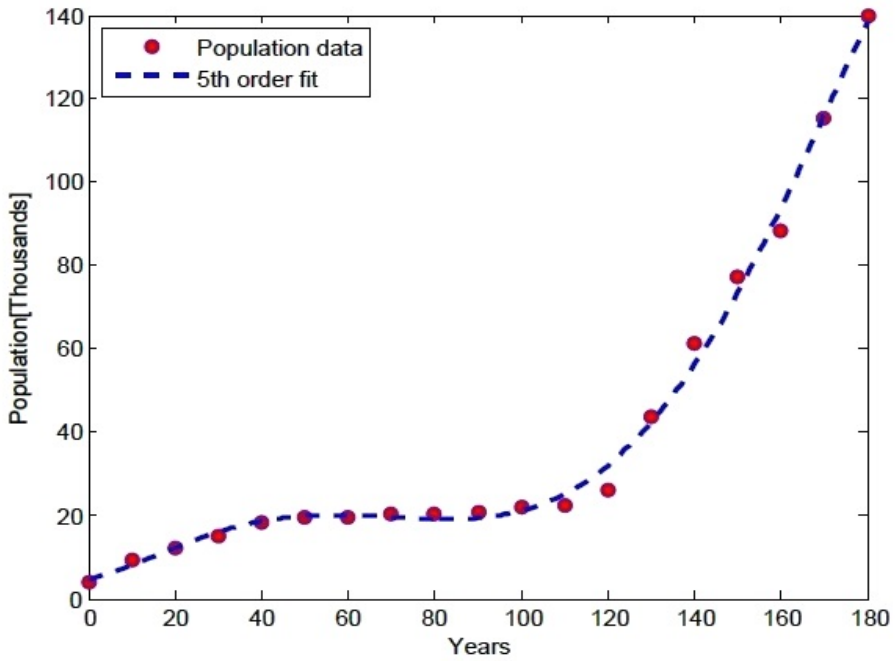

Figure G3: Ridge Polynomial Fit of Order Five.

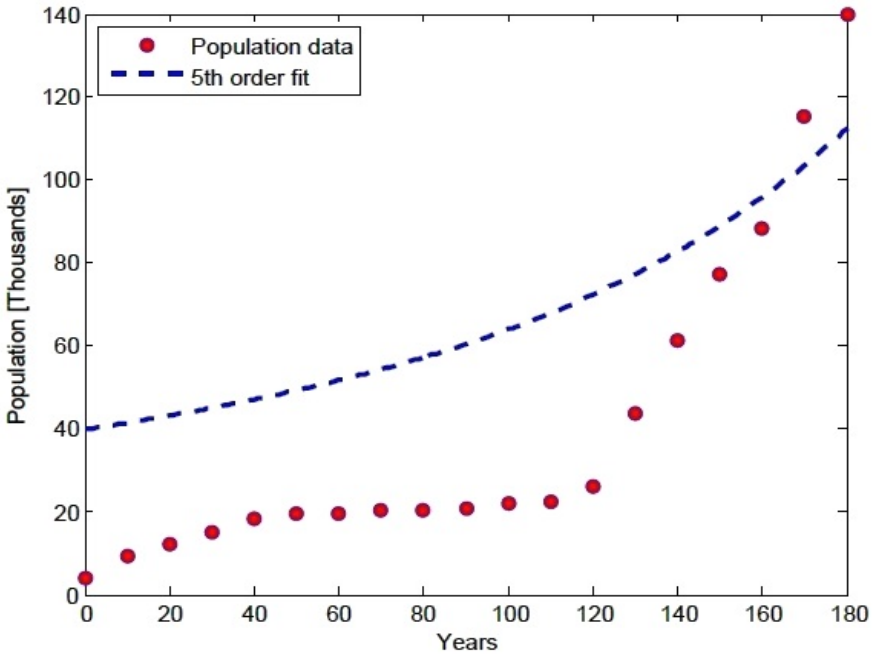

Figure G4 :Method 1 Polynomial Fit of Order Five.

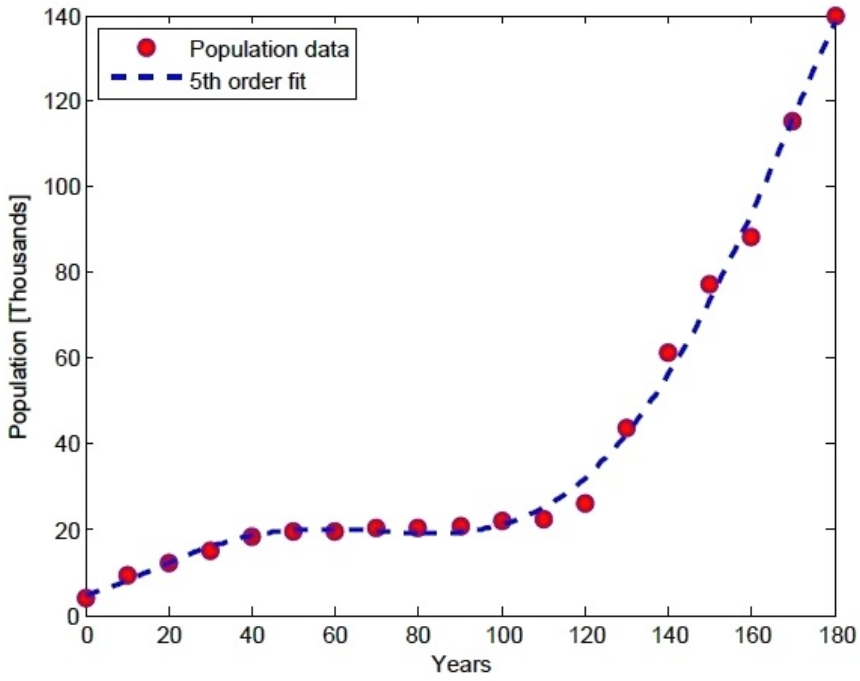

Figure G5: The *l*1_*ls* Polynomial Fit of Order Five.

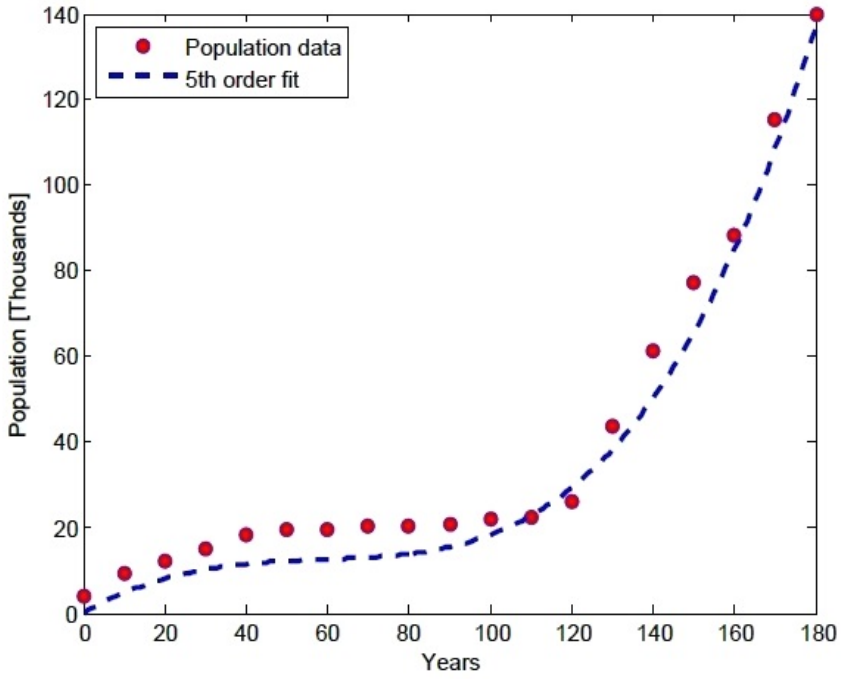

Figure G6 : LASSO Polynomial Fit of Order Five.

APPENDIX H

CRIME DATASET

Table H1: Crime Dataset

Y_1	Y_2	X_1	X_2	X_3	X_4	X_5
478	184	40	74	11	31	20
494	213	32	72	11	43	18
643	347	57	70	18	16	16
341	565	31	71	11	25	19
773	327	67	72	9	29	24
603	260	25	68	8	32	15
484	325	34	68	12	24	14
546	102	33	62	13	28	11
424	38	36	69	7	25	12
548	226	31	66	9	58	15
506	137	35	60	13	21	9
819	369	30	81	4	77	36
541	109	44	66	9	37	12
491	809	32	67	11	37	16
514	29	30	65	12	35	11
371	245	16	64	10	42	14
457	118	29	64	12	21	10
437	148	36	62	7	81	27
570	387	30	59	15	31	16
432	98	23	56	15	50	15
619	608	33	46	22	24	8
357	218	35	54	14	27	13

Table H1: Crime Dataset Continued

Y_1	Y_2	X_1	X_2	X_3	X_4	X_5
623	254	38	54	20	22	11
547	697	44	45	26	18	8
792	827	28	57	12	23	11
799	693	35	57	9	60	18
439	448	31	61	19	14	12
867	942	39	52	17	31	10
912	1017	27	44	21	24	9
462	216	36	43	18	23	8
859	673	38	48	19	22	10
805	989	46	57	14	25	12
652	630	29	47	19	25	9
776	404	32	50	19	21	9
919	692	39	48	16	32	11
732	1517	44	49	13	31	14
657	879	33	72	13	13	22
1419	631	43	59	14	21	13
989	1375	22	49	9	46	13
821	1139	30	54	13	27	12
1740	3545	86	62	22	18	15
815	706	30	47	17	39	11
760	451	32	45	34	15	10
936	433	43	48	26	23	12

Table H1: Crime Dataset Continued

Y_1	Y_2	X_1	X_2	X_3	X_4	X_5
863	601	20	69	23	7	12
783	1024	55	42	23	23	11
715	457	44	49	18	30	12
1504	1441	37	57	15	35	13
1324	1022	82	72	22	15	16
940	1244	66	67	26	18	16

APPENDIX I

PROSTATE CANCER DATASET

Table I1: Prostate Cancer Dataset

lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
-0.579818495	2.769459	50	-1.38629436	0	-1.38629436	6	0	-0.4307829
-0.994252273	3.319626	58	-1.38629436	0	-1.38629436	6	0	-0.1625189
-0.510825624	2.691243	74	-1.38629436	0	-1.38629436	7	20	-0.1625189
-1.203972804	3.282789	58	-1.38629436	0	-1.38629436	6	0	-0.1625189
0.751416089	3.432373	62	-1.38629436	0	-1.38629436	6	0	0.3715636
-1.049822124	3.228826	50	-1.38629436	0	-1.38629436	6	0	0.7654678
0.737164066	3.473518	64	0.61518564	0	-1.38629436	6	0	0.7654678
0.693147181	3.539509	58	1.53686722	0	-1.38629436	6	0	0.8544153
-0.776528789	3.539509	47	-1.38629436	0	-1.38629436	6	0	1.0473190
0.223143551	3.244544	63	-1.38629436	0	-1.38629436	6	0	1.0473190
0.254642218	3.604138	65	-1.38629436	0	-1.38629436	6	0	1.2669476
-1.347073648	3.598681	63	1.26694760	0	-1.38629436	6	0	1.2669476
1.613429934	3.022861	63	-1.38629436	0	-0.59783700	7	30	1.2669476
1.477048724	2.998229	67	-1.38629436	0	-1.38629436	7	5	1.3480731
1.205970807	3.442019	57	-1.38629436	0	-0.43078292	7	5	1.3987169
1.541159072	3.061052	66	-1.38629436	0	-1.38629436	6	0	1.4469190
-0.415515444	3.516013	70	1.24415459	0	-0.59783700	7	30	1.4701758
2.288486169	3.649359	66	-1.38629436	0	0.37156356	6	0	1.4929041
-0.562118918	3.267666	41	-1.38629436	0	-1.38629436	6	0	1.5581446

APPENDIX I CONTINUED

Table I1: Prostate Cancer Dataset Continued

lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
0.182321557	3.825375	70	1.65822808	0	-1.38629436	6	0	1.5993876
1.147402453	3.419365	59	-1.38629436	0	-1.38629436	6	0	1.6389967
2.059238834	3.501043	60	1.47476301	0	1.34807315	7	20	1.6582281
-0.544727175	3.375880	59	-0.79850770	0	-1.38629436	6	0	1.6956156
1.781709133	3.451574	63	0.43825493	0	1.17865500	7	60	1.7137979
0.385262401	3.667400	69	1.59938758	0	-1.38629436	6	0	1.7316555
1.446918983	3.124565	68	0.30010459	0	-1.38629436	6	0	1.7664417
0.512823626	3.719651	65	-1.38629436	0	-0.79850770	7	70	1.8000583
-0.400477567	3.865979	67	1.81645208	0	-1.38629436	7	20	1.8164521
1.040276712	3.128951	67	0.22314355	0	0.04879016	7	80	1.8484548
2.409644165	3.375880	65	-1.38629436	0	1.61938824	6	0	1.8946169
0.285178942	4.090169	65	1.96290773	0	-0.79850770	6	0	1.9242487
0.182321557	3.804438	65	1.70474809	0	-1.38629436	6	0	2.0082140
1.275362800	3.037354	71	1.26694760	0	-1.38629436	6	0	2.0082140
0.009950331	3.267666	54	-1.38629436	0	-1.38629436	6	0	2.0215476
-0.010050336	3.216874	63	-1.38629436	0	-0.79850770	6	0	2.0476928
1.308332820	4.119850	64	2.17133681	0	-1.38629436	7	5	2.0856721
1.423108334	3.657131	73	-0.57981850	0	1.65822808	8	15	2.1575593
0.457424847	2.374906	64	-1.38629436	0	-1.38629436	7	15	2.1916535

Table I1: Prostate Cancer Dataset Continued

lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
2.660958594	4.085136	68	1.37371558	1	1.83258146	7	35	2.2137539
0.797507196	3.013081	56	0.93609336	0	-0.16251893	7	5	2.2772673
0.620576488	3.141995	60	-1.38629436	0	-1.38629436	9	80	2.2975726
1.442201993	3.682610	68	-1.38629436	0	-1.38629436	7	10	2.3075726
0.582215620	3.865979	62	1.71379793	0	-0.43078292	6	0	2.3272777
1.771556762	3.896909	61	-1.38629436	0	0.81093022	7	6	2.3749058
1.486139696	3.409496	66	1.74919985	0	-0.43078292	7	20	2.5217206
1.663926098	3.392829	61	0.61518564	0	-1.38629436	7	15	2.5533438
2.727852828	3.995445	79	1.87946505	1	2.65675691	9	100	2.5687881
1.163150810	4.035125	68	1.71379793	0	-0.43078292	7	40	2.5687881
1.745715531	3.498022	43	-1.38629436	0	-1.38629436	6	0	2.5915164
1.220829921	3.568123	70	1.37371558	0	-0.79850770	6	0	2.5915164
1.091923301	3.993603	68	-1.38629436	0	-1.38629436	7	50	2.6567569
1.660131027	4.234831	64	2.07317193	0	-1.38629436	6	0	2.6775910
0.512823626	3.633631	64	1.49290410	0	0.04879016	7	70	2.6844403
2.127040520	4.121473	68	1.76644166	0	1.44691898	7	40	2.6912431
3.153590358	3.516013	59	-1.38629436	0	-1.38629436	7	5	2.7047113

Table I1: Prostate Cancer Dataset Continued

lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
1.266947603	4.280132	66	2.12226154	0	-1.38629436	7	15	2.7180005
0.974559640	2.865054	47	-1.38629436	0	0.50077529	7	4	2.7880929
0.463734016	3.764682	49	1.42310833	0	-1.38629436	6	0	2.7942279
0.542324291	4.178226	70	0.43825493	0	-1.38629436	7	20	2.8063861
1.061256502	3.851211	61	1.29472717	0	-1.38629436	7	40	2.8124102
0.457424847	4.524502	73	2.32630162	0	-1.38629436	6	0	2.8419982
1.997417706	3.719651	63	1.61938824	1	1.90954250	7	40	2.8535925
2.775708850	3.524889	72	-1.38629436	0	1.55814462	9	95	2.8535925
2.034705648	3.917011	66	2.00821403	1	2.11021320	7	60	2.8820035
2.073171929	3.623007	64	-1.38629436	0	-1.38629436	6	0	2.8820035
1.458615023	3.836221	61	1.32175584	0	-0.43078292	7	20	2.8875901
2.022871190	3.878466	68	1.78339122	0	1.32175584	7	70	2.9204698
2.198335072	4.050915	72	2.30757263	0	-0.43078292	7	10	2.9626924
-0.446287103	4.408547	69	-1.38629436	0	-1.38629436	6	0	2.9626924
1.193922468	4.780383	72	2.32630162	0	-0.79850770	7	5	2.9729753
1.864080131	3.593194	60	-1.38629436	1	1.32175584	7	60	3.0130809
1.160020917	3.341093	77	1.74919985	0	-1.38629436	7	25	3.0373539
1.214912744	3.825375	69	-1.38629436	1	0.22314355	7	20	3.0563569
1.838961071	3.236716	60	0.43825493	1	1.17865500	9	90	3.0750055
2.999226163	3.849083	69	-1.38629436	1	1.90954250	7	20	3.2752562
3.141130476	3.263849	68	-0.05129329	1	2.42036813	7	50	3.3375474
2.010894999	4.433789	72	2.12226154	0	0.50077529	7	60	3.3928291
2.537657215	4.354784	78	2.32630162	0	-1.38629436	7	10	3.4355988
2.648300197	3.582129	69	-1.38629436	1	2.58399755	7	70	3.4578927
2.779440197	3.823192	63	-1.38629436	0	0.37156356	7	50	3.5130369
1.467874348	3.070376	66	0.55961579	0	0.22314355	7	40	3.5160131
2.513656063	3.473518	57	0.43825493	0	2.32727771	7	60	3.5307626
2.613006652	3.888754	77	-0.52763274	1	0.55961579	7	30	3.5652984
2.677590994	3.838376	65	1.11514159	0	1.74919985	9	70	3.5709402
1.562346305	3.709907	60	1.69561561	0	0.81093022	7	30	3.5876769
3.302849259	3.518980	64	-1.38629436	1	2.32727771	7	60	3.6309855

Table I1: Prostate Cancer Dataset Continued

lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
2.024193067	3.731699	58	1.63899671	0	-1.38629436	6	0	3.6800909
1.731655545	3.369018	62	-1.38629436	1	0.30010459	7	30	3.7123518
2.807593831	4.718052	65	-1.38629436	1	2.46385324	7	60	3.9843437
1.562346305	3.695110	76	0.93609336	1	0.81093022	7	75	3.9936030
3.246490992	4.101817	68	-1.38629436	0	-1.38629436	6	0	4.0298060
2.532902848	3.677566	61	1.34807315	1	-1.38629436	7	15	4.1295508
2.830267834	3.876396	68	-1.38629436	1	1.32175584	7	60	4.3851468
3.821003607	3.896909	44	-1.38629436	1	2.16905370	7	40	4.6844434
2.907447359	3.396185	52	-1.38629436	1	2.46385324	7	10	5.1431245
2.882563575	3.773910	68	1.55814462	1	1.55814462	7	80	5.4775090
3.471966453	3.974998	68	0.43825493	1	2.90416508	7	20	5.5829322