A DIFFERENTIAL GEOMETRY-BASED ALGORITHM FOR SOLVING THE MINIMUM HELLINGER DISTANCE ESTIMATOR

Philip Anthony D'Ambrosio

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Electrical Engineering

Lamine Mili
A.A. (Louis) Beex
Y. (Joseph) Wang

April 21, 2008
Falls Church, VA

Keywords: Robust Estimation, Power Systems, Minimum Hellinger Distance, Information Geometry

A DIFFERENTIAL GEOMETRY-BASED ALGORITHM FOR SOLVING THE MINIMUM HELLINGER DISTANCE ESTIMATOR

Philip Anthony D'Ambrosio

ABSTRACT

Robust estimation of statistical parameters is traditionally believed to exist in a trade space between robustness and efficiency. This thesis examines the Minimum Hellinger Distance Estimator (MHDE), which is known to have desirable robustness properties as well as desirable efficiency properties. This thesis confirms that the MHDE is simultaneously robust against outliers and asymptotically efficient in the univariate location case. Robustness results are then extended to the case of simple linear regression, where the MHDE is shown empirically to have a breakdown point of 50%. A geometric algorithm for solution of the MHDE is developed and implemented. The algorithm utilizes the Riemannian manifold properties of the statistical model to achieve an algorithmic speedup. The MHDE is then applied to an illustrative problem in power system state estimation. The power system is modeled as a structured linear regression problem via a linearized direct current model; robustness results in this context have been investigated and future research areas have been identified from both a statistical perspective as well as an algorithm design standpoint.

Table of Contents

To Mama and Theo
Love, Papa

Acknowledgements

This thesis is the end result of a long process, undertaken while working at the Schafer Corporation in support of the Defense Advanced Research Projects Agency (DARPA). To that end, I extend my gratitude to personnel from both organizations for their unstinting support and supreme flexibility during my graduate studies.

I am indebted to Professor Lamine Mili, my advisor and committee chair, for introducing me to the fascinating research problems addressed in this thesis, for his support, and his seemingly endless energy, availability, and infectious enthusiasm. I extend my warm thanks to Professors Beex and Wang for making the time to serve on my committee.

My family and friends have buoyed me, encouraged me, and contributed immeasurably to this thesis; to every single one of them, I owe more of a debt than words could possibly express. In particular, my parents, Antonietta and Pasquale D'Ambrosio made all of this possible with their love and sacrifices throughout my life. One lifetime isn't enough to convey my thanks to them. Vincent and Chantel, Daniela and Todd, Cristina, and my in-laws, Thomas and Judith Kelly, have supported me, even during my absences, silences, and apparent disappearances from the face of the earth. I am looking forward to making up for lost time.

I would be remiss if I didn't offer acknowledgement to my cat, Jasmine, who stayed awake with me during my many thesis-related all-nighters.

Finally, my brilliant, amazing wife, Amy M. V. Bowers, has been an equal partner in this entire process, starting with the decision to write a thesis, and continuing throughout the research, writing, and defense of this work. I absolutely could not have continued without her enduring love, support, and deep understanding. This thesis is dedicated to her, and to the loving memory of our baby boy, Theodore Bowers D'Ambrosio.

Chapter 1

Introduction

1.0 Motivation and Previous Work

The statistical modeling of data and the estimation of parameters relevant to such models are nearly universal practices; one finds examples of such activities in engineering, natural science, social science, business and the liberal arts (Abur and Expósito, 2004; Fisher, 1925; Reif, 1965; Wilcox, 1987). Just as ubiquitous as the practice of statistical modeling is the phenomenon of outliers, which we may give a working definition of any experimental observations which cannot be predicted or accounted for by a given model. Outliers can occur spontaneously; they can also be introduced by the process of recording and processing data (*e.g.,* manual data entry errors, sensor failure).

A fundamental question in statistics concerns the handling of outliers. Often, such data is trying to tell the practitioner a story about limitations of an experiment, or novel behavior within the population being observed (Hoaglin, *et al.,* 1983). For this reason, it is desirable to identify outliers, so that we may either study them further or reject them completely (Rousseeuw and Leroy, 1987). However, the ability to do so often comes at the expense of efficiency at the probability distribution of the observed data (Hampel *et al.,* 1986). An estimator with a combination of robustness against outliers and efficiency at a particular distribution is desirable.

1.1 Contributions of the work

The main undertakings of this thesis are as follows:

- The MHDE is shown to be a robust regression estimator; specifically, the estimator is shown to be resistant to both vertical outliers and bad leverage points in 2 dimensions for the case of simple linear regression.
- A novel algorithm for solving the MHDE is introduced; this algorithm exploits the Riemannian geometry of the problem space to achieve a speedup over the steepest descent algorithm. For optimization problems such as finding the MHDE, it is common to employ an iterative approach such as the Newton method or gradient descent to solve the problem. Building upon foundational results by Amari (1998), we use the relationship between the Fisher information matrix and the Riemannian manifold structure of the space of probability densities to accelerate the convergence of the steepest descent algorithm.
- The regression MHDE is applied to a problem in power systems to illustrate how such an estimator might overcome current difficulties in state estimation. Following the work of (Mili et al., 1994; Mili and Coakley, 1996), we address state estimation in power systems as the solution of a structured linear regression model. This is the first application of the MHDE to power systems.
- As a contribution, it is shown that the algorithm implementing the MHDE cannot obtain the maximum exact fit point in structured linear regression problems. We have identified the difficulty as resulting from rounding errors when evaluating the MHDE cost function, which involves the product of many small exponentials; this is an area for future research.

1.2 Outline of the Thesis

1.2.1 Chapter 2

Chapter 2 will introduce the statistical figures of merit for the majority of this thesis. The relevant classical concepts – maximum likelihood estimation, Fisher consistency, unbiasedness, Fisher information, Cramer-Rao lower bound, and asymptotic efficiency, are reviewed here. Rudiments of the more modern robustness theory are discussed as well, including the concept of a contaminated model, bias, breakdown point, influence function, asymptotic bias and asymptotic variance, and gross error sensitivity. The chapter ends with a brief discussion of the difficulties encountered in obtaining an estimate that would be considered ideal by both classical and modern standards.

1.2.2 Chapter 3

Maximum likelihood estimation methods are widely used due to ease of computation (Rousseeuw and Leroy, 1987; Weisberg, 1985). However, such methods may not be robust to outliers, due to departures from the classical assumptions which they obey. On the other hand, robust estimation methods in general are predicated on the belief that resistance to gross errors can only be obtained at the expense of efficiency at the Gaussian distribution. Chapter 3 will review the minimum Hellinger distance estimator (MHDE), an estimator which has been a part of the statistical literature for at least three decades (Beran, 1977, Wolfowitz 1952, 1953, 1957). The estimator is shown to be both robust and efficient for univariate location; while these results are well-known to statisticians, they provide context for the novel results demonstrated in the latter part of the chapter. We extend the robustness result to show that the regression MHDE rejects a large number of bad leverage points.

1.2.3 Chapter 4

Chapter 4 focuses on the algorithmic aspect of our estimation procedure. A review of iterative methods is provided, focusing on the steepest descent algorithm. This algorithm has the desirable property of converging reliably to a local minimum (Kreyszig, 2006); however, the steepest descent algorithm can take many iterations to reach its stopping criterion, depending on the shape of the cost function (Forsythe, *et al.,* 1977; Press, *et al.,* 1992). Following the pioneering information-geometric works of Amari (1982, 1985, 1998, 2001) we introduce an algorithm for finding the regression MHDE in two or more dimensions; this algorithm provides a speedup over steepest descent algorithms by exploiting the Riemannian geometry of the problem space. The necessary geometric preliminaries are developed in this chapter as well, borrowing from the standard literature in information geometry (Amari, 1985; Amari and Nagaoka, 1993; Barndorf-Nielsen *et al.,* 1986; Murray and Rice, 1993) as well as references from mathematical physics (Arnold, 1989; Misner et al., 1973; Schutz, 1980).

1.2.4 Chapter 5

Chapter 5 applies the results of the previous chapters to the solution of a problem in power systems. In the context of the DC power flow estimation model, the MHDE is shown to provide robust state estimates for a system with sparse observed data. The secure, stable operation of power systems relies on estimation of the state variables of the power transmission system from a redundant collection of power and voltage measurements spread throughout the network (Abur and Exposito, 2004). However, due to calibration errors or missing meters, these measurements are often the source of biased or missing data; too many of these corrupt measurements may lead to unreliable state estimation results. The use of the MHDE provides the robustness needed in power system state estimation. In the following sections, we introduce the DC power flow model and apply the MHDE to this model on a 3-bus system. Based on results of (Mili *et al.,* 1994; Mili and Coakley, 1996) the model is a structured regression model, meaning that there are linear dependencies among the rows of the Jacobian matrix. The relevant concepts of surplus and exact fit point are introduced, and the relationship between maximum exact fit and breakdown point is discussed. The MHDE, as implemented in this simulation, has inherent weaknesses apart from the estimator itself; these suggest new avenues of study for the MHDE, not only from a robustness viewpoint, but algorithmically as well.

1.2.5 Chapter 6

Chapter 6 summarizes and discusses these results and offers several ways forward for a program of research. Future work relating to both the robustness theory of the MHDE and the algorithmic implementation of the estimator is proposed.

Chapter 2

Statistical Estimation of Location: Classical and Robustness Concepts

2.0 Introduction

The mathematical aspect of statistics is, at its heart, concerned with two main questions: what can we infer about a situation, or what credible information can we learn from a set of observations, and how certain can we be about this inference, given the data and the models that we apply. In statistics, the first question corresponds loosely to what are known as *point estimation*; a typical example is location estimation. The second question corresponds to *confidence interval estimation*, usually based on *dispersion estimation*. For the sake of space and intuitive development of later concepts, we will concentrate on the location problem, though much of the mathematics presented here can be applied to dispersion problems in a straightforward manner.

In estimating the location of a "true value" based on a set of observations, one might encounter any of the following situations: 1) we could apply an appropriate estimation model to data that is assumed to be exact, 2) we could apply an inappropriate model – either by accident or out of necessity – to a set of data that would fit another model quite well, 3) we could apply an appropriate model to data that contains imperfect or contaminated observations, or 4) we could apply an inappropriate model to imperfect data. The first situation is the chief domain of classical estimation, while other situations are addressed by robust statistics.

This chapter will introduce the commonly known and employed procedures for judging the viability of an estimator. There are two main sets of procedures. The first is "classical" in nature, and relies on strong assumptions about the conditions under which observations are made and estimates are obtained. Because it relies on such strong assumptions, the theory is very mature, but admits estimators that do not always work well in real-world situations.

The other theory, the "modern" or "robust" theory, is reliant on somewhat weaker assumptions. The theory is by no means a finished one; however, there is still a core set of tools for characterizing a robust estimation process that has been used for decades and is generally accepted, and the understanding of these tools is useful for gathering an intuition about the performance of various estimation procedures.

2.1 Classical Statistics Concepts

Classical treatments of statistical estimation require a set of strong assumptions; namely, that samples are drawn from an independent, identically distributed (iid) space. Furthermore, the tools employed in such treatments are frequently based on the supposition that samples are drawn from a given distribution, usually Gaussian. These strong assumptions often go unmet; to the extent that they hold, the following concepts can be used to describe a good statistical estimator. Our treatment is an introductory one, meant to provide context for the results obtained in later chapters.

2.1.1 Maximum Likelihood Estimation

The "maximum likelihood" method of parameter estimation devised by Fisher is the crowning achievement of classical statistics, notable for bringing a new level of rigor to the often-subjective task of parameter estimation (Fisher, 1925). In the most mature of his papers, Fisher quantifies the information available in a given sample, and shows that the maximum likelihood method asymptotically obtains a parameter estimate with no loss of information.

Estimation of a one-dimensional location parameter offers the most straightforward means of understanding the maximum likelihood method. Let Z be a random variable belonging to the probability density function $f(z)$, with location parameter θ. Take a sample of n observations $\{z_1,\ldots,z_m\}$, assumed to be independent and identically distributed. The *likelihood* is the joint probability that the observations belong to a distribution with probability density function $f(\mathbf{z};\theta)$ viewed as a function of the parameter θ given the sample $l(\theta;\mathbf{z}) = cf(\mathbf{z};\theta)$ where c is a constant typically put equal to one. For convenience, the observations are assumed to be i.i.d., yielding

$$l(\theta;\mathbf{z}) = \prod_{i=1}^{m} f(z_i;\theta). \tag{2.1}$$

The estimated parameter in the maximum likelihood sense, written $\hat{\theta}$, is one that maximizes the likelihood function:

$$\left.\frac{\partial l}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0. \tag{2.2}$$

For the sake of tractability, particularly when estimating the parameters of a distribution that has an exponential form, the likelihood l is frequently replaced by the quantity $\ln l$, which is also frequently called the likelihood. Therefore the problem is more commonly posed as:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{m} \rho(z_i;\theta) \tag{2.3}$$

where

$$\rho = -\ln f(z_i;\theta). \tag{2.4}$$

Our objective function is then

$$J(\theta) = \sum_{i=1}^{m} \rho(z_i;\theta). \tag{2.5}$$

To solve this optimization problem, we set the derivative of the objective function equal to zero and solve for *θ*:

$$\frac{\partial}{\partial\theta} \sum_{i=1}^{m} \rho(z_i;\theta) = 0. \tag{2.6}$$

A helpful example is that of a univariate Gaussian distribution

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\theta)^2}{2\sigma^2}}, \tag{2.7}$$

where *θ* is the central location of the distribution, and *σ* is the standard deviation. For the sake of simplicity we assume $\sigma = 1$. Using the steps just defined, we can find the maximum likelihood estimate of the location part of the Gaussian distribution. At this point, it is helpful to employ the substitution $r = z - \theta$. Rewriting equation (2.7), we have

$$f(r) = \frac{1}{\sqrt{2\pi}} e^{-\frac{r^2}{2}}. \tag{2.8}$$

To put this function in the appropriate form to solve for the maximum likelihood estimator, we take the logarithm and get

$$\begin{aligned} \rho(r) &= -\ln f(r) \\ &= \ln\left(\sqrt{2\pi}\right) + \frac{r^2}{2}. \end{aligned} \tag{2.9}$$

To find our maximum likelihood estimator $\hat{\theta}$, we must solve

$$\frac{\partial}{\partial\theta} \sum_{i=1}^{m} \rho(r_i) = 0. \tag{2.10}$$

Explicitly, we have

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta} \sum_{i=1}^{m} \rho(r_i) \\ &= \frac{\partial}{\partial\theta} \sum_{i=1}^{m} r_i^2 \\ &= 2\sum_{i=1}^{m} r_i \\ &= 2\sum_{i=1}^{m} \left(z_i - \hat{\theta}\right) \\ &= -m\hat{\theta} + \sum_{i=1}^{m} z_i; \\ m\hat{\theta} &= \sum_{i=1}^{m} z_i; \\ \hat{\theta} &= \frac{1}{m} \sum_{i=1}^{m} z_i. \end{aligned} \tag{2.11}$$

The upshot of this example is that for a Gaussian distribution, the maximum likelihood estimator of location is the sample mean.

2.1.2 Fisher Consistency

Perhaps the most intuitively desirable property for an estimator is the one commonly known as *Fisher consistency*, or *consistency*. In words, the more observations we have, the more our estimate will tend to the true value of any parameter. Formally, we can write

$$\lim_{m\to\infty} \hat{\theta}_m = \theta. \tag{2.12}$$

Figure 2.1 shows the consistency of the ML estimator of location at the normal distribution; the true value of the location is zero. Note that while the estimator becomes qualitatively more consistent for the highest values of *m*, it will provide an estimate of location that is, at least to the naked eye, centered around the true value with just a couple hundred observations of the distribution.

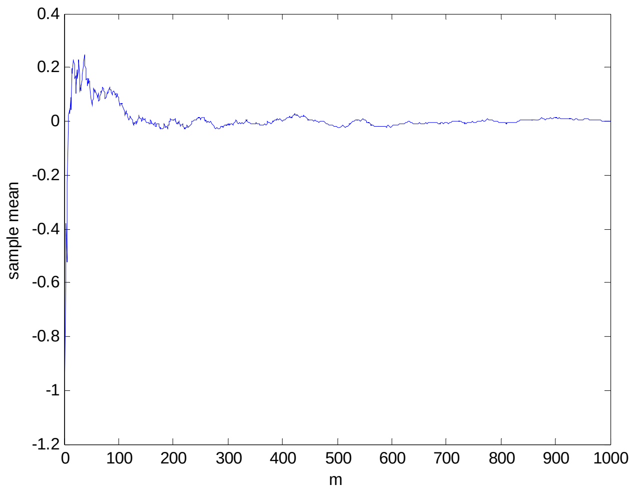

Figure 2.1: *Consistency of the sample mean at the standard Gaussian distribution.*

2.1.3 Unbiasedness

In the classical view of statistics, *unbiasedness* is a very important property for an estimator to have. The default definition of unbiasedness, commonly known as *mean unbiasedness*, calls for the expected value of an estimator to be equal to the true value of the parameter in the asymptotic limit. Define the bias as:

$$b_m = E[\hat{\theta}_m] - \theta. \tag{2.13}$$

Then an estimator is unbiased when $b_m = 0$. As shown in Section 2.1.1, the ML estimator of location for the normal Gaussian distribution is the sample mean; it can easily be shown that this estimator is mean unbiased.

2.1.4 Fisher Information

Let *Z* be a random variable with realized values $\{z_1,...,z_m\}$. *Z* contains an amount of information about θ. The information that *Z* contains about θ is known as the *Fisher Information* $I_f(\theta)$. Mathematically, it can be expressed as:

$$I_f(\theta) = \int_{-\infty}^{\infty} \left(\frac{f'(z;\theta)}{f(z;\theta)} \right)^2 f(z;\theta) dz = E\left\{ \left(\frac{f'(z;\theta)}{f(z;\theta)} \right)^2 \right\} = E\left\{ \left[\frac{\partial}{\partial \theta} (\ln f(z;\theta)) \right]^2 \right\}. \tag{2.14}$$

While all three of these expressions are equivalent, there are instances where any one will prove more computationally convenient than the others.

To illustrate the concept, let us calcluate the Fisher information contained within the Gaussian distribution with zero mean and unit variance. Recall that the Gaussian in this case has a probability density function of the form

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}. \tag{2.15}$$

From this, we calculate the derivative:

$$\frac{df_Z(z)}{dz} = -z \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}. \tag{2.16}$$

The Fisher information for the normal distribution is then

$$\begin{aligned} I_f(\theta) &= \int_{-\infty}^{\infty} \left(\frac{f'(z;\theta)}{f(z;\theta)} \right)^2 f(z;\theta) dz \\ &= \int_{-\infty}^{\infty} \frac{z^2 e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dx \\ &= 1. \end{aligned} \tag{2.17}$$

2.1.5 Cramèr-Rao Lower Bound

The inverse of the Fisher information serves an important statistical purpose; it provides the theoretical performance limit of any estimator in the sense that it is the lower bound of the variance that any estimator may have. This lower bound is called the *Cramèr-Rao Lower Bound*, and it is expressed via the *Cramèr-Rao Inequality* as:

$$Var\left(\sqrt{m}\, \hat{\theta}_m \right) = E\left\{ m\left(\hat{\theta}_m - \theta \right)^2 \right\} \geq \frac{1}{I_f(\theta)}. \tag{2.18}$$

2.1.6 Efficiency

The aforementioned bounds give a theoretical limit for the variance of an estimator. We can assess the performance of an estimator in terms of its *efficiency* e_m. We can define the quantity as follows:

$$e_m = \frac{\frac{1}{I_f(\theta)}}{Var\left(\sqrt{m}\hat{\theta}_m\right)}, \tag{2.19}$$

where $0 \le e_m \le 1$. An efficient estimator, then, obtains e_m=1.

2.1.7 Asymptotic Efficiency

A consistent estimator is *asymptotically efficient* if

$$\lim_{m\to\infty} Var\left(\sqrt{m}\hat{\theta}_m\right) = \frac{1}{I_f(\theta)}. \tag{2.20}$$

As an example, the sample mean is the maximum likelihood estimator of location at the Gaussian distribution. The asymptotic variance of this estimator is

$$\begin{aligned} \lim_{m\to\infty} Var\left(\sqrt{m}\hat{\theta}_m\right) &= \lim_{m\to\infty} E\left[m\left(\hat{\theta}_m - \theta\right)^2 \right] \\ &= \sigma^2. \end{aligned} \tag{2.21}$$

For the case of a Gaussian distribution with zero mean and unit variance, $\sigma^2 = 1$. The sample mean is therefore an asymptotically efficient estimator at the Gaussian distribution. Figure 2.2 below illustrates this fact.

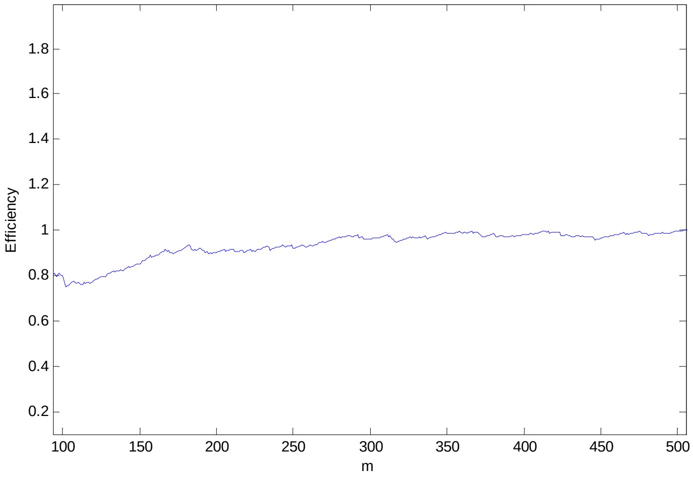

Figure 2.2: *Efficiency of the MLE vs. sample size m at the standard Gaussian distribution for m=100 to m=500 observations.*

2.2 Modern Robustness Concepts

The theory of robust estimation was first described in its modern form by Huber (Huber, 1964), and later extended by Hampel in his PhD thesis and related papers (Hampel, 1968, 1974; Hampel *et al*., 1986). Robustness in its contemporary statistical sense can be described as immunity to departures from any assumptions made about a set of observations; in particular, departures from the assumed probability distribution may be viewed as due to the presence of outliers among the observations. The greater the tolerable disparity between the assumed and the true probability distributions, the more robust the estimation method. This is in contrast to the Fisherian approach, which assumes that all observations belong to the same target distribution.

In this section we present an outline of concepts from robust statistics germane to our later discussions of minimum distance estimation. The notion of a *contaminated model* is considered, and various ways of quantifying robustness based on this model are introduced. Our motivation for choosing a robust statistical treatment of estimation is meant to address two major challenges in statistics: that Gaussianity is often assumed, whether it is appropriate or not, and that even Gaussian data can be contaminated in a number of ways.

2.2.1 Contaminated Models

The most commonly accepted mathematical method for treating contaminated data comes from (Huber, 1964, 1981); it is this model upon which much of the robustness approach is based. Broadly speaking, the *ε-contaminated model* addresses outliers by positing that the observed data belongs to a mixture of two distributions. The mixture is of the form:

$$G(r) = (1-\varepsilon)F(r) + \varepsilon H(r) \tag{2.22}$$

where $0 \le \varepsilon \le 1$ is the fraction of the observations regarded as outliers. By convention, *F* is generally the target distribution, which is assumed, and *H* is the distribution of contaminated data, assumed to be unknown.

2.2.2 Bias Revisited

In section 2.1, *bias* was defined as the difference between the mean of the estimator and the true value, yielding

$$b_m = E[\hat{\theta}_m] - \theta. \tag{2.23}$$

In our investigation of robustness, it will be helpful to define the bias in a different setting. Specifically, the notion of bias can be a useful tool for comparing the performance of an estimator under deviations from Fisherian assumptions. Define the bias as

$$b = \left| \hat{\theta}(F) - \hat{\theta}(G) \right|, \tag{2.24}$$

where $\overset{\wedge}{\theta}(F)$ is the functional form of the estimator at *F* in the von Mises sense (Fernholz, 1983). When the estimator is Fisher consistent at *F*, $\overset{\wedge}{\theta}(F)$ is equal to the true value, that is $\overset{\wedge}{\theta}(F) = \theta$. The *maximum bias* is obtained by taking the supremum over all *H*, that is,

$$b_{\max} = \sup_{H} \left| \overset{\wedge}{\theta}(F) - \overset{\wedge}{\theta}(G) \right|. \tag{2.25}$$

This quantity represents the worst-case effect of a contaminated population on the estimator.

2.2.3 Breakdown Point and Maximum Breakdown Point

For a given sample, the *breakdown point* of an estimator is the maximum fraction of contamination $\varepsilon *$ that the estimator can tolerate while still having a bounded bias. Formally, we write

$$\varepsilon^* = \left\{ \varepsilon \,; \sup \left| \hat{\theta}(F) - \hat{\theta}(G) \right| \text{ is finite} \right\} \tag{2.26}$$

In the case of the sample mean, we can show that the breakdown point is $\varepsilon *$=0; this means that a single arbitrarily placed outlier can lead to an arbitrarily biased estimate.

In contrast, let us consider the *sample median*, an order statistic defined for an ordered set $\{z_1, z_2, ..., z_m\}$ as

$$\hat{z}_{med} = \begin{cases} m \text{ odd}: & z_{[m/2]+1} \\ m \text{ even}: & \left(z_{[m/2]} + z_{[m/2]+1}\right)/2 \end{cases} \tag{2.27}$$

where [*m*/2] denotes the integer part of *m*/2. For the sample median, the overall effect of a fraction ε of outliers on a location estimate is bounded, and it can be shown that the sample median has a breakdown point of nearly 50%, making it a robust estimator of location.

The *maximum breakdown point* is the upper limit on the fraction of contaminated data that any estimator can tolerate. In the case of location estimation it is

$$\varepsilon^*_{\max} = \left[\frac{m-1}{2}\right] \Big/ m. \tag{2.28}$$

2.2.4 Influence Function

The idea for the influence function comes from Hampel (1968). As the name implies, the influence function of an estimator measures the impact that a single observation can have on an estimator. Assume that a sample has *m* observations (z_1, …, z_{m-1}, z_m), of which the first *m*-1 observations belong to a distribution *F*; z_m can then take on any value. The influence function IF$_m$(z) then measures the standardized difference between the two estimators:

$$IF_m(z) = \frac{\hat{\theta}_m(z_1, z_2, ..., z_{m-1}, z_m) - \hat{\theta}_{m-1}(z_1, z_2, ..., z_{m-1})}{\varepsilon}. \tag{2.29}$$

In this case, we are measuring the effect of a single contaminated observation in a sample of size *m*, so ε=1/*m*. Our influence function can then be written as:

$$\begin{aligned} IF_m(z) &= \frac{\hat{\theta}_m(z_1, z_2, ..., z_{m-1}, z_m) - \hat{\theta}_{m-1}(z_1, z_2, ..., z_{m-1})}{1/m} \\ &= m\left(\hat{\theta}_m - \hat{\theta}_{m-1}\right). \end{aligned} \tag{2.30}$$

We can relate the bias to the influence function as follows: $b_m \approx \varepsilon\left|IF_m(z)\right|$.

The above definition applies to finite samples, but can be generalized to the asymptotic case. Under some regularity conditions, the *asymptotic influence function* can be written as the limiting expression of the finite sample influence function:

$$
\begin{aligned}
IF(z) &= \lim_{m\to\infty} \frac{\overset{\wedge}{\theta}_m(z_1, z_2, \ldots, z_{m-1}, z) - \overset{\wedge}{\theta}_{m-1}(z_1, z_2, \ldots, z_{m-1})}{\varepsilon} \\
&= \left.\frac{\partial \hat{\theta}(F_m)}{\partial \varepsilon}\right|_{\varepsilon=0} .
\end{aligned}
\tag{2.31}
$$

2.2.5 Asymptotic Bias and Asymptotic Variance

Part of the reason for widespread acceptance of the influence function lies in its use as a Rosetta stone of sorts, effective to first order for approximating a number of useful statistical quantities. As noted above, the rate of change of bias and finite sample influence function have nearly identical functional forms. It then follows naturally that the *asymptotic bias* of an estimator can be approximated by the asymptotic influence function:

$$
b \approx \varepsilon |IF(z)|. \tag{2.32}
$$

On the other hand, the *asymptotic variance* of an estimator at a distribution *F* with no contamination can be written in terms of the IF as

$$
\begin{aligned}
\lim_{m\to\infty} Var\left(\sqrt{m}\hat{\theta}_m ; F\right) &= \int_{-\infty}^{\infty} IF^2(z) f(z) dz \\
&= Var\left(\hat{\theta}; F\right).
\end{aligned}
\tag{2.33}
$$

2.2.6 Gross Error Sensitivity

When the influence function exists, the asymptotic influence function is, in and of itself, a useful measure of the robustness of an estimator. Additionally, a number of other measures of robustness can be derived from the asymptotic IF. Of these, the most widely known and employed is the *gross error sensitivity* $\gamma *$, defined as follows:

$$\gamma^* = \sup\left|IF(z)\right|. \tag{2.34}$$

From this definition, the maximum asymptotic bias can be approximated by

$$b_{\max} \approx \varepsilon\gamma * . \tag{2.35}$$

In words, the gross error sensitivity measures the worst-case effect of an infinitesimal contamination on the estimator.

2.2.7 Maximum Bias Curve

Many of the concepts just discussed can be related to each other geometrically by a construction known as the *maximum bias curve*, which shows $b_{\max}$ as a function of ε. Through analysis of this curve, we can understand many properties of an estimator quickly. Note that because the maximum bias for the MLE at the normal distribution is unbounded for even a single outlier, no maximum bias curve can be drawn for this estimator.

Consider the case of the sample median, where F is the standard Gaussian distribution. Since $(1-\varepsilon)F\left(\hat{\theta}_{med}\right) = 1/2$, we have $b_{\max} = \left|\hat{\theta}_{med}\right| = F^{-1}(1/2(1-\varepsilon))$. Figure 2.3 shows the maximum bias curve for the sample median.

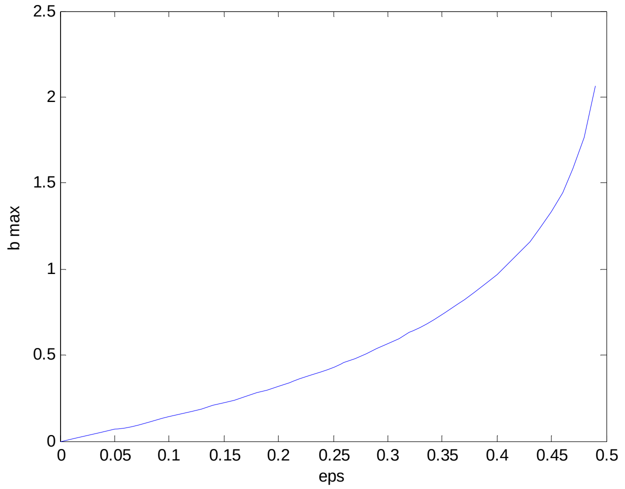

Figure 2.3: *Asymptotic maximum bias curve of the sample median at the standard Gaussian distribution.*

2.2.8 Efficiency Versus Robustness

Previously, it was shown that the sample median is the most robust estimator of location, capable of tolerating nearly 50% of the sample having unbounded values. However, by the Fisherian concept of efficiency at the Gaussian distribution, our estimator is less than optimal. Figure 2.4 shows the efficiency of the MLE, which is the sample mean, and the sample median for *m* samples drawn from the normal distribution: while the efficiency of the MLE tends toward unity, the efficiency of the median tends toward a lesser value. It is shown in (Fisher, 1925) that the asymptotic efficiency of the sample median is approximately 0.64.

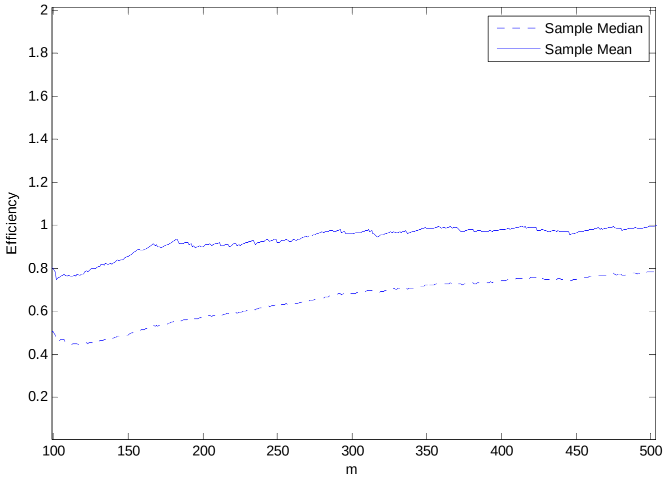

Figure 2.4: *Efficiency of sample median and sample mean at standard Gaussian distribution for m=100 to m=500 observations.*

This illustrates the point that there is no single estimator that works best for all situations; one of the many trade spaces in which practitioners must operate is that of efficiency vs. robustness. Put another way, it is generally believed that one must tolerate some loss of information (as measured by efficiency) in exchange for the ability to tolerate outliers – otherwise known as robustness. The common perception is that a balance needs to be struck between classical and modern notions of goodness. In the next chapter, we will discuss a notable exception: a class of estimators that exhibits asymptotic efficiency, as well as highly desirable robustness properties at a given distribution, including the normal distribution.

Chapter 3

Minimum Hellinger Distance Estimation

3.0 Introduction

Maximum likelihood methods, while widely used, may be non-robust due to disagreement between the assumptions upon which the models are based and the true probability distribution of observed data. In the years since robustness theory first came to prominence, several methods for mitigating the shortcomings of maximum likelihood estimation have been proposed, such as M-estimation (Huber, 1964). In general, robust estimation methods are predicated on the belief that resistance to gross errors can only be obtained at the expense of efficiency at the Gaussian distribution. In this chapter, we introduce an estimator based on *minimum distance* methods, and show that it has attractive robustness features, while maintaining asymptotic efficiency at any given distribution where the estimator is developed.

3.1 Minimum Hellinger Distance Estimator

The estimator we are about to examine requires a broadened consideration of how we estimate the parameters of a statistical model. In the majority of commonly used estimators, the practitioner relies on one of a small handful of techniques:

- Optimizing the model parameters to fit every observation to the same joint distribution. This is inherently non-robust because it supposes that all observations belong to the same joint distribution (Li, 1985).
- Estimation methods of probability density functions based on histograms or count data. There are inherent drawbacks to using this procedure for modeling probability density functions; the estimates are sensitive to the performance of the density estimator. Regarding the widely used non parametric methods of estimation, while literature exists on optimal bin width choices, subjectivity on the part of the statistical worker plays a larger part in selection of this parameter than in other techniques (Scott, 1979).
- Estimating parameters based on order statistics. This is robust in the one-dimensional case, but the generalization to higher dimensions is not straightforward (Small, 1990). It has been known that parameter estimation based on order statistics is not, generally speaking, efficient (Fisher, 1925; Mosteller, 1946).

3.1.1 Minimum Distance Estimators

We now discuss a class of estimators, called minimum distance estimators, which has been a part of the statistical literature for some time (Wolfowitz, 1953, 1954, 1957). A *minimum distance estimator* minimizes the distance between two density functions in an abstract functional space. Our discussion derives from (Beran, 1977), but treatments of this subject can be found in a number of publications (Basu, *et al.*, 1997; Hettmansperger, *et al.*, 1994).

3.1.2 Hellinger Distance, Hellinger Affinity, and MHDE

Minimum distance estimation has been a part of the statistical landscape for a long time. Early work by Kolmogorov (1933) played an important part in recognizing probability theory as an applied branch of measure theory and, in some sense, set the stage for the application of concepts such as distance to distributions of random variables. Kolmogorov proposed $\sup\left|G(r) - F(r)\right|$ as an early measure of distance between distributions. However, it was Wolfowitz who made a rigorous case for the use of estimating parameters by minimizing the distance between two curves. Though his original work initially focused on estimators based on the distribution function, our exposition will focus on density functions. Readers who are unfamiliar with the basic terminology of measure theory are advised to consult (Ash, 1972; Kolmogorov and Fomin, 1975; Pollard, 2002) for discussions of measure motivated by various points of view.

An estimation process based on the Hellinger distance was first put forth by Beran (1977); we explain the basic idea here. The motivating problem for this discussion is as follows: we observe *m* random variables $\{Z_1, Z_2, \ldots, Z_m\}$, with the intention of modeling their collective behavior. As discussed in Chapter 2, a common assumption is that the random variables are independent identically distributed (i.i.d.) with a density belonging to a parametric family $\{f_\theta : \theta \in \Theta\}$. At the same time, we realize that our assumption almost certainly will not fully coincide with the reality of our experiment, due to missing data, contaminated data, systematic errors in observation, short-memory and long-memory processes, and other real-world effects. The problem that we must solve, then, is how do we construct an estimation procedure that strikes a balance between the tractability afforded by our assumptions, and tolerating the challenging conditions just described?

Define $\mathfrak{F}$ as the set of all probability densities with respect to Lebesgue measure on the real line. For two probability distributions, *F* and *G*, with densities $f, g \in \mathfrak{F}$ respectively, the *squared Hellinger distance* between *f* and *g* is defined as

$$d_H(f, g) = \left\| f^{1/2} - g^{1/2} \right\| \tag{3.1}$$

where $\|\bullet\|$ is the L_2 norm. In the estimation procedure to be described, f is taken to be a parametric model; g is taken to be an "empirical" density based on the observations. One way of visualizing the Hellinger distance is shown in Figure 3.1; the empirical density g, as well as various parameterizations of f, written f_{θ_1} and f_{θ_2}, are points in a metric space .

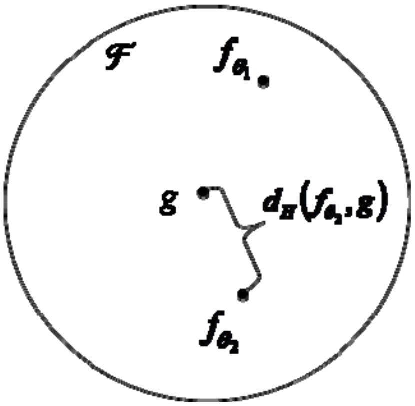

Figure 3.1: *The Hellinger distance between $f_{\theta 2}$ and g on the space of probability density functions.*

Let the parametric model f have a density $\{f_\theta : \theta \in \Theta\}$, where θ is the parameter to be estimated. The minimum Hellinger distance functional T is defined on $\mathcal{F}$ by the requirement that for every g in $\mathcal{F}$,

$$\left\| f_{T(g)}^{1/2} - g^{1/2} \right\| = \min_{\theta \in \Theta} \left\| f_\theta^{1/2} - g^{1/2} \right\|. \tag{3.2}$$

The value θ produced by this functional is known as the *minimum Hellinger distance estimator* (MHDE). Expanding the quantity d_H, we have

$$\begin{aligned} d_H(f_\theta, g) &= \int \left(f_\theta^{1/2} - g^{1/2} \right)^2 dz \\ &= \int f_\theta - 2\sqrt{f_\theta g} + g \, dz \\ &= 2 - 2\int \sqrt{f_\theta g} \, dz \end{aligned} \tag{3.3}$$

From equation (3.3) above, we can see that, in practice, minimizing the Hellinger distance over Θ is a matter of maximizing the quantity $\int \sqrt{f_\theta g}\ dz$, which is also known as the *Hellinger affinity* (van der Vaert, 1998; Basu *et al.*,1997).

3.2 Univariate Location Estimates with the MHDE

3.2.1 The Univarate Location Problem Defined

One of the most basic – and most universal– problems in statistics is that of the location of central tendency based on a set of observations. This problem is a special case of both multivariate location and multivariate regression (Anderson, 2003; Rousseeuw and Leroy, 1987) and is equally applicable to problems in a number of fields (Wilcox, 1987; Leon-Garcia, 1994; Reif, 1965).

To anchor our discussion, consider a set of *m* observations drawn from a Gaussian distribution with zero mean, unit variance. This distribution is described by the *probability density function*

$$f(z;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(z-\mu)^2}{2\sigma^2}}. \tag{3.4}$$

Substituting in explicit values $\mu=0,\sigma^2=1$, we write

$$f(z;\mu,\sigma)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}. \tag{3.5}$$

Given a set of *m* observations $\{z_1,z_2,...,z_m\}$, we ask the following question: what is the central value of the data set? The question of what constitutes a "central value" lies at the heart of our investigation, and each method of estimation described above represents a different answer to this question.

3.2.2 The Sample Mean in Univariate Location

Recall from Chapter 2 that for the univariate location problem given above, the sample mean is the maximum likelihood estimator of location at the normal distribution, and is given by

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} z_i . \tag{3.6}$$

This is one method of estimating central value, easily interpreted by anyone who has ever taken a class in Newtonian mechanics: the sample mean is computed via the exact same expression one would use to find a center-of-mass in one dimension. This analogy is not lost on statisticians, who talk about phenomena such as “leverage points” in regression.

The sample mean is, historically speaking, a popularly employed estimator because of its ease of computation (Weisberg, 1985). Additionally, as shown in Chapter 2, this estimator is asymptotically efficient. However, the very features that make this estimator both efficient and tractable make it highly susceptible to outliers. To see this, compare what happens to the sample mean for m=1000 observations of the standard Gaussian distribution, versus m=1001, where the last observation is an outlier z_{1001}=10.

For the m=1000 case, when all the observations come from the Gaussian distribution, $\hat{\theta} = \text{-}0.0431$. However, a single outlier at z_{1001}=10 changes the estimate to $\hat{\theta} = \text{-}0.0330$. When z_{1001}=1000, $\hat{\theta} = 0.9560$; in this way, it can be seen that the mean can be arbitrarily biased by moving an outlier arbitrarily far away from the bulk of the observations.

3.2.3 The Sample Median in Univariate Location

We now turn our attention to the sample median, an example of an estimator based on ordered data. This estimator, as well as the maximum likelihood estimator, was discussed in Chapter 2. In contrast to the sample mean, the sample median is robust against outliers. To return to the example given above, the sample median from 1000 observations of the Gaussian is $\hat{\theta} = \text{-}0.0131$. In this case, the sample median remains $\hat{\theta} = \text{-}0.0131$; this fact does not change whether the outlier is placed at z_{1001}=10, 100, or 1000. Figure 2.3 shows the maximum bias curve for the sample median. There are two very important pieces of information to be gleaned from this curve: the first is that the gross error sensitivity, given by the infinitesimal slope at ε =0, is $\sqrt{\pi/2}$. The second notable piece of information that this curve tells us is that, for the normal distribution *N*(0, 1), the asymptotic breakdown point of the median is 0.5.

There are, however, a couple of drawbacks associated with this approach to estimation. As previously discussed, this estimator does not have a high asymptotic efficiency at the Gaussian distribution. Furthermore, the sample median does not generalize to higher dimensions in any straightforward way.

3.2.4 The Univariate MHDE

In this thesis, we examine the properties of our estimator in the context of the same Gaussian distribution; this represents the ground truth of our data which we seek to recover via the MHDE. Recall from the definition of the MHDE in equation (3.3) that our estimation procedure requires a parametric model *f*, and an empirical density *g*, derived from the observations. Our parametric model is

$$f(z;\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-\theta)^2} \tag{3.7}$$

where θ is the parameter of location to be estimated.

Some care is required in our choice of the nonparametric density; *g* is constructed from individual observations of data points, but our estimator is defined on the space of probability distributions. To make our estimator work, we must somehow embed our data in this space. One way of achieving this end is by the use of a nonparametric *kernel density estimator*

$$g(y)=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{h}K\left(\frac{y-X_i}{h}\right), \tag{3.8}$$

where *K* is referred to as the *kernel* and *h* is a positive number known as the *bandwidth* (Scott, 1992). While the optimal choice of bandwidth value and kernel type are an important area of study in their own right, it does not hinder the exposition of our results to set *h*=1, and to define our kernel density estimator as

$$\begin{aligned} g(y)&=\frac{1}{mh}\sum_{j=1}^{m}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y-z_j}{h}\right)^2} \\ &=\frac{1}{m\sqrt{2\pi}}\sum_{j=1}^{m}e^{-\frac{1}{2}(y-z_j)^2}. \end{aligned} \tag{3.9}$$

This type of kernel is generally known as the *Gaussian kernel*.

Armed with these basic statements, we can define a cost function based on the Hellinger distance which must be optimized for choice of parameter *θ*:

$$\begin{aligned} J(\theta)&=d_H\left(f(y,\theta),g(y)\right) \\ &=2-2\left(\frac{1}{\sqrt{2\pi}}\right)\int_{-\infty}^{\infty}\left[e^{-\frac{1}{2}(y-\theta)^2}\left(\frac{1}{m}\sum_{j=1}^{m}e^{-\frac{1}{2}(y-z_j)^2}\right)\right]^{\frac{1}{2}}dy. \end{aligned} \tag{3.10}$$

Removing the constant additive terms and normalizing the multiplicative constants, we have

$$J(\theta)=-\frac{1}{\sqrt{m}}\int_{-\infty}^{\infty}\left[e^{-\frac{1}{2}(y-\theta)^2}\left(\sum_{j=1}^{m}e^{-\frac{1}{2}(y-z_j)^2}\right)\right]^{\frac{1}{2}}dy. \tag{3.11}$$

Because this integral has no closed form, we approximate it by the Riemann sum after an appropriate discretization; the cost function can then be rewritten as:

$$J(\theta) = -\frac{1}{\sqrt{m}} \sum_{i=1}^{N} e^{-\frac{1}{4}(y_i - \theta)^2} a_i \tag{3.12}$$

where

$$a_i = \left(\sum_{j=1}^{m} e^{-\frac{1}{2}(y_i - z_j)^2} \right)^{\frac{1}{2}}. \tag{3.13}$$

3.2.5 Empirical Investigation of MHDE Asymptotic Efficiency

Recall from Chapter 2 that the Fisher information of the normal distribution *N*(0, 1) is unity. In his original paper on the MHDE, Beran (1977) put forth an informal proof of the asymptotic efficiency of the MHDE. Rather than repeating the proof here, it will be more illuminating for our purposes to verify empirically that for our choice of Gaussian kernel, the univariate location MHDE is asymptotically efficient. We do this by calculating the variance of our estimator for increasing sample size *m*. Figure 3.2 shows that the efficiency of the MHDE tends asymptotically toward unity with increasing *m*.

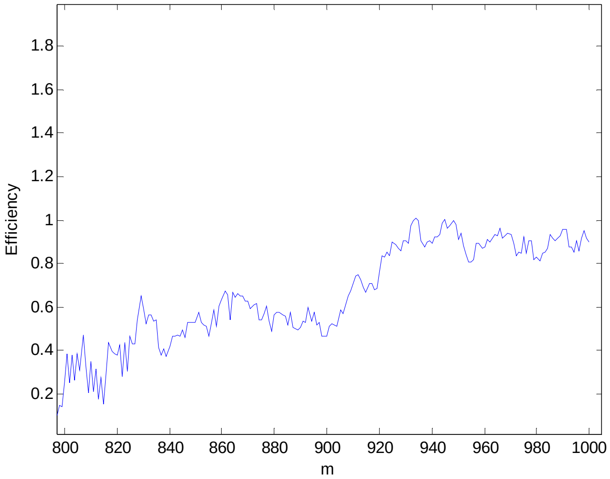

Figure 3.2: *Efficiency of the MHDE with increasing sample size under standard Gaussian distribution for m=800 to m=1000 observations.*

3.2.6 Robustness of the MHDE in Location

As was the case with discussions of efficiency, a graphical representation of the robustness of the univariate location MHDE will prove to be much more helpful than an analytical solution for the purpose of illustrating the properties of this estimator. The maximum bias curve for the univariate location MHDE at the standard normal distribution $N(0, 1)$, shown in Figure 3.3, is nearly identical to that of the sample median. The significance of this fact is this: whereas the sample median has no easy generalization to higher dimensions, the univariate MHDE can be extended for implementation in regression problems as well as multivariate location and covariance problems. A thorough analysis of the latter is given in (Tamura and Boos, 1986); for the purposes of this thesis, we dedicate the remainder of this chapter to an empirical investigation of extensions of the location MHDE for linear regression.

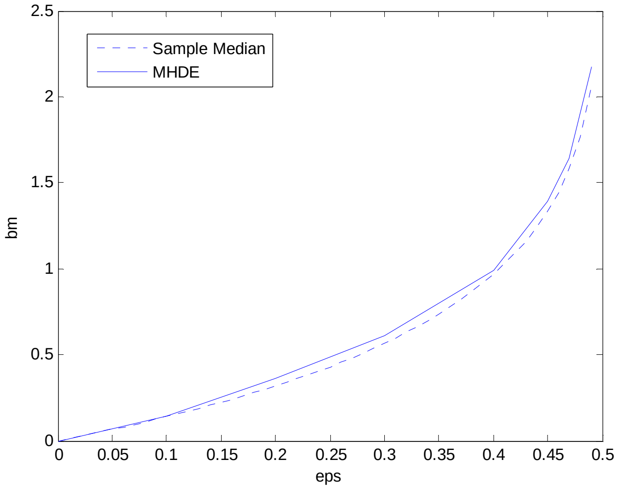

Figure 3.3: *Maximum bias curve of the univariate location MHDE compared to the sample median at the standard Gaussian distribution, m=1000*

3.3 Linear regression with the MHDE

Estimating the central location of a distribution is an important task, but generalizing the problem to multiple dimensions greatly extends the range of applicability of our estimator. A particular generalization of the univariate location problem is the *simple linear regression* problem, in which we investigate linear relationships in ordered pairs of observations. Good treatments of the subject can be found in (Weisberg, 1985; Montgomery and Peck, 1982; Seber and Lee, 2003; Cook and Weisberg, 1982).

3.3.1 The Simple Linear Regression Problem Defined

Consider a two-dimensional set of observations $\{(z_i, h_i)\}$, $i = 1,...,m.$ We hope to ascertain some kind of linear relationship between *z* and *h*. We assume that the relationship can be described by a *linear regression model* of the form

$$z_i = h_i x_1 + x_2 + e_i \text{ for } i = 1,...,m \tag{3.14}$$

where x_1 is our slope, x_2 is the intercept, and e_i is an additive error.

It is worth taking a moment to discuss the physical interpretation of this model. What do x_1 and x_2 really tell us? What does e_i really mean? The linear regression equation can be thought of as a method of modeling a physical process. The variables x_1 and x_2 parameterize the deterministic part of this process. We commonly refer to h as a *design variable* or *explanatory variable*. We call z our *dependent variable* or *response variable*. To the extent that the process is deterministic, and the relationship between h and z is linear, the first two terms of the equation will dominate. If the relationship between h and z is not linear, or if there is no deterministic relationship between the design and response variables, we expect e to be the dominant term.

As we explore methods of solving the linear regression problem, it will be helpful to employ more compact notation. We can rewrite equation (3.14) as $\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e}$, where

$$\mathbf{z} = \begin{bmatrix} z_1 \\ \vdots \\ z_m \end{bmatrix}, \; \mathbf{H} = \begin{bmatrix} \mathbf{h}_1^{\mathbf{T}} \\ \vdots \\ \mathbf{h}_m^{\mathbf{T}} \end{bmatrix} = \begin{bmatrix} h_1 & 1 \\ \vdots & \vdots \\ h_m & 1 \end{bmatrix}, \; \mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_m \end{bmatrix}, \; \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \qquad (3.15)$$

The elements of the column vector $\mathbf{x}$ are the *true regression parameters* which fit a line to our dataset. Our objective is to provide an estimate $\hat{\mathbf{x}} = \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix}$ of these regression parameters.

3.3.2 Ordinary Least Squares in 2-D Regression

Historically, the predominant method of handling regression problems has been the *method of least squares*, which we define here. For the regression model defined by equation (3.15), the least squares (LS) estimate is the one such that

$$\arg\min_{\mathbf{x}} \sum_{i=1}^{m} r_i^2 \qquad (3.16)$$

where the *residuals* r_i are given by $r_i = z_i - \mathbf{h}_i^T \mathbf{x}$.

Consider the regression estimates obtained by the LS method in three distinct cases. First we consider a sample of *m*=20 observations, with

$$\mathbf{z} = \begin{bmatrix} 1 \\ \vdots \\ 20 \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 20 & 1 \end{bmatrix}, \quad e_i \sim N(0, 0.01). \tag{3.17}$$

In this case, the true x_1 and x_2 are given by $\mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. We find that the LS estimator performs well in this case, as shown in Figure 3.4.

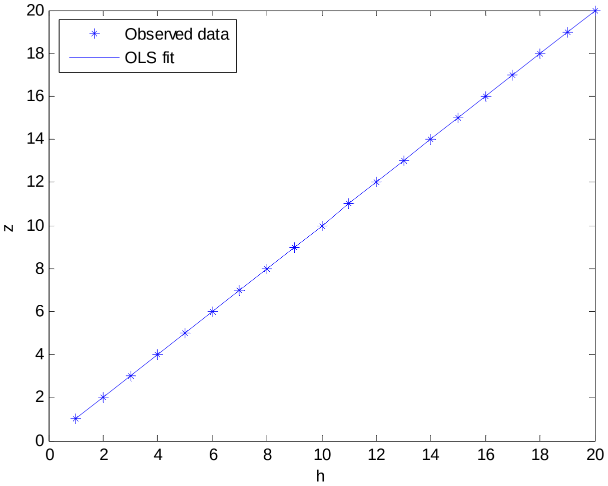

Figure 3.4: *Ordinary least squares linear regression fit to 20 data points.*

Now consider the case where a single observation has fallen off the line, either as the result of an interesting physical phenomenon, or as the result of poor data entry. Our data in such a case might look like the following:

$$\mathbf{z} = \begin{bmatrix} 1 \\ \vdots \\ 19 \\ 0 \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 19 & 1 \\ 100 & 1 \end{bmatrix}, \; e_i \sim N(0, 0.01). \tag{3.18}$$

An estimator that is robust against outliers would tend to the same line shown in Figure 3.4. But as observed in Figure 3.5, the LS fit breaks down. It deviates significantly from the true line. In fact, it can be arbitrarily pulled away by the action of a single outlier.

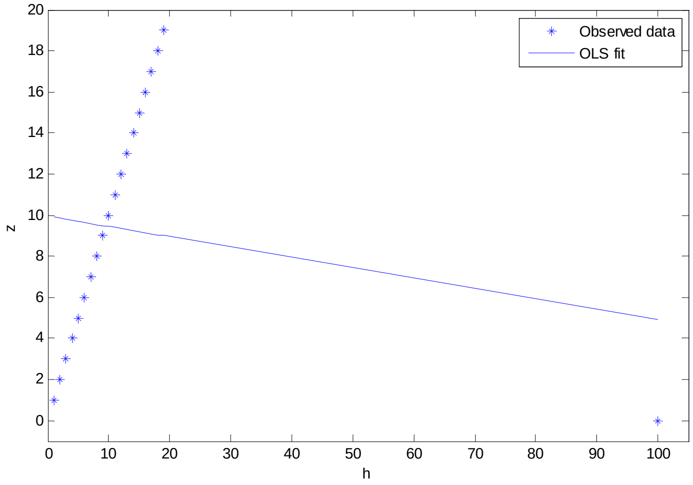

Figure 3.5: *Breakdown of the ordinary LS estimator in the presence of a single outlier in leverage position at (100, 0).*

To be more specific, the type of outlier shown above is commonly known as a *bad leverage point.* In the same way that we examine the robustness of an estimator in one dimension by placing a single observation arbitrarily far away from the bulk of the data on the number line, we can probe the robustness of a regression estimator by placing a single observation arbitrarily far away from the bulk of the data in 2-dimensional space; while placing the outlier so that its associated *h* value is an outlier along the *h* axis, resulting in a bad leverage point *in extremis* – the grossest of gross errors, as it were – the bad leverage point, which is shown in Figure 3.5, is qualitatively sufficient to show that the ordinary LS estimator cannot tolerate even a single outlier. As previously mentioned, the ordinary LS estimator is prized by practitioners in a number of fields for its computational ease; however, the associated cost is that of increased attention and care on the part of anyone who uses this estimator when performing aggregate analysis of real data.

3.3.3 Least Median Squares in 2-D Regression

It can be difficult to visualize how an estimator might *not* be affected by a gross error in the manner described above; however, such estimators do exist. As a group, they are known as *high breakdown estimators*. Some relevant examples are given in (Rousseeuw, 1984; Siegel, 1982). An exemplar of this class of estimators is the *least median of squares* (LMS), shown by Rousseeuw to be robust against a high fraction of outliers. In simple regression this estimate is given by

$$\arg\min_{\hat{\mathbf{x}}} \underset{i}{median}\left(r_i^{\,2}\right). \tag{3.19}$$

In higher dimensions, it minimizes the *v*-th ordered squared residual, where

$$v = \left[\frac{m}{2}\right] + \left[\frac{n+1}{2}\right], \tag{3.20}$$

where *m* is the total number of observations, and *n* is the number of parameters we are attempting to estimate.

To illustrate the ability of an estimator to reject bad leverage points, we revisit the example given above, placing a single bad leverage point at (100, 0). As we see in Figure 3.6, the LMS estimator rejects a single bad leverage point; the OLS fit is shown for comparison.

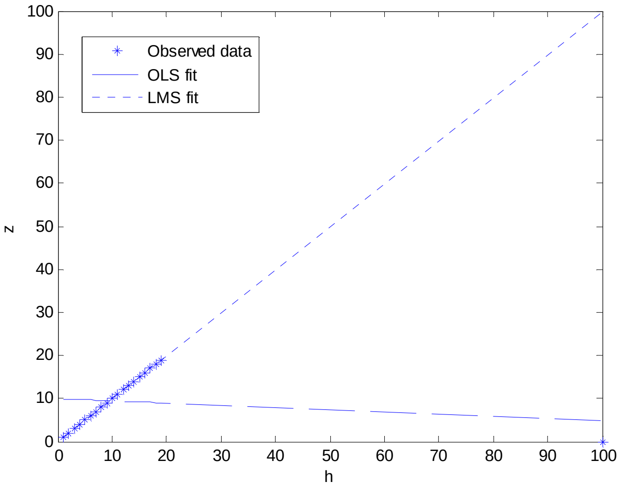

Figure 3.6: *LMS regression fit in the presence of a single outlier in leverage position at (100, 0).*

By definition, the regression parameters obtained by LMS estimates minimize the median residual among the observations in the data set. For the example given above, one would reasonably expect that the LMS estimator could tolerate more than one bad leverage point. Perhaps the most challenging problem that one might face in a linear regression setting is that of the *two-line problem*, which can be imagined in practical terms as the entry of data from two completely different populations into the same dataset. In such a situation, it is up to the practitioner to identify the origin and nature of the second line, be it an outlier or a legitimate part of the process to be modeled. Assuming there should only be one regression line, a two-line dataset leaves the practitioner with very few clues to distinguish between the data to be modeled and the outliers to be identified – or, in our case, rejected. The two-line problem, largely synthetic in nature, represents a worst-case scenario for any linear regression estimator. The breakdown point of the simple linear regression estimator can be regarded as the fraction of bad leverage points which can be placed along a line orthogonal to the line defined by our true regression parameters. As in the univariate case, a robust estimator is one which can tolerate a high fraction of bad leverage points, up to 50%. Figure 3.7 shows a data set of size m=20, where 9 of the observations in the data set have been deleted and replaced with bad leverage points. Even with nearly half the data corrupted, the LMS still manages to recover the true regression parameters.

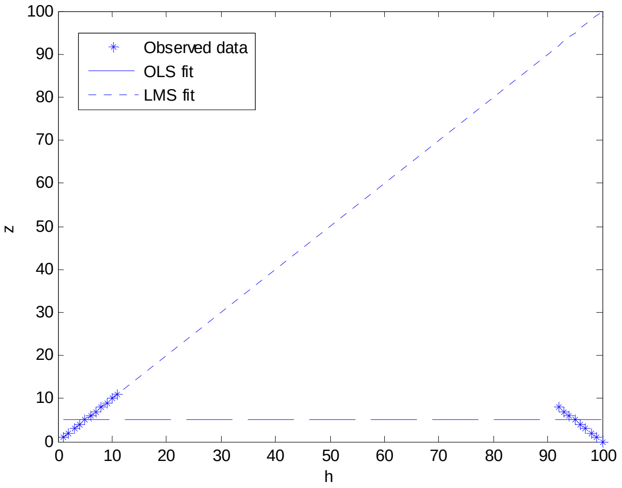

Figure 3.7: *LMS and OLS regression fits in the two-line problem*

At first blush, this would appear to be the end of the story; we have just described an estimator that is known to be robust, even when nearly 50% of the data is comprised of outliers. However, the LMS estimator exhibits two unattractive tendencies: the first is that the estimator is known to be inefficient. This makes intuitive sense, given that the LMS is the straight line lying at the middle of the narrowest strip covering half of the observations (Rousseeuw, 1984). This fact on its own is not enough to rule out the LMS estimator for many applications; there may be situations where the statistical worker is willing to sacrifice efficiency in exchange for a highly robust estimate. However, more damning in the case of the LMS estimator is its computational complexity, as briefly mentioned in (Mili *et al.*, 1991); the iterative search for an optimal order statistic becomes exponentially slower as the problem space increases in dimension. The implementation of the LMS estimator for large-scale problems would likely rely on some sort of approximate algorithm for tractability. The question still remains as to whether there exists a robust estimator, preferably an efficient one, which suffers from no obvious "curse of dimensionality" issues as the problem space scales up.

3.3.4 MHDE in 2-D regression

We can define an estimator based on minimizing the Hellinger distance between observation points and points on the line of best fit. This estimator is more computationally complex than ordinary least squares, but is far more tractable than the LMS estimator, has a remarkable resistance to outliers in any dimension. One major difference between our proposed estimator and the LS estimator is that the LS estimator is based on optimizing a function of the residuals, while the regression MHDE is based on optimizing a function of probability density functions (pdfs) centered at each observation point. This approach is developed further in the next few paragraphs.

Recall our definition of the Hellinger affinity from Section 3.1.2. In the regression case, we generalize the Hellinger affinity as

$$Aff(f,g)=\sum_{i=1}^{m}\int\sqrt{f_i g_i}\,dy_i. \tag{3.21}$$

This technique, while more computationally complex than least squares, provides a straightforward extension of our one-dimensional estimator, and exhibits a very high level of robustness against outliers, as we will see below.

We now define the minimum Hellinger distance estimator for the two-dimensional regression case. Following (Scott, 1992), our empirical pdf is defined as

$$g(\mathbf{y})=\frac{1}{m}\sum_{j=1}^{m}\frac{1}{(2\pi)^{m/2}}\exp\left[-\frac{1}{2}\sum_{i=1}^{m}\left(y_{ji}-z_j\right)^2\right] \tag{3.22}$$

and our parametric model is defined as

$$\begin{aligned}f(\mathbf{y};\mathbf{x})&=\frac{1}{(2\pi)^{m/2}}\exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{Hx})^T(\mathbf{y}-\mathbf{Hx})\right]\\&=\frac{1}{(2\pi)^{m/2}}\exp\left[-\frac{1}{2}\sum_{j=1}^{m}\left(y_{ji}-\mathbf{h}_\mathbf{j}^\mathbf{T}\mathbf{x}\right)^2\right]\\&=\prod_{j=1}^{m}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(y_{ji}-\mathbf{h}_j^T\mathbf{x}\right)^2\right].\end{aligned} \tag{3.23}$$

Returning to our definition of the regression MHDE given in equation 3.21, we can write our cost function as:

$$
\begin{aligned}
J(\mathbf{x}) &= -2\sum_{i=1}^{N}\left[\left(\prod_{j=1}^{m}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(y_{ji}-\mathbf{h}_j^T\mathbf{x}\right)^2\right]\right)\left(\frac{1}{m}\sum_{j=1}^{m}\frac{1}{(2\pi)^{m/2}}\exp\left[-\frac{1}{2}\sum_{i=1}^{m}\left(y_{ji}-z_j\right)^2\right]\right)\right]^{\frac{1}{2}} \\
&= -2\sum_{i=1}^{N}\left[\left(\frac{1}{(2\pi)^{m/4}}\exp\left[-\frac{1}{4}\sum_{j=1}^{m}\left(y_{ji}-\mathbf{h}_j^T\mathbf{x}\right)^2\right]\right)\left(\frac{1}{m(2\pi)^{m/2}}\sum_{j=1}^{m}\exp\left[-\frac{1}{2}\sum_{i=1}^{m}\left(y_{ji}-z_j\right)^2\right]\right)^{\frac{1}{2}}\right] \quad (3.24) \\
&= -2\left(\frac{1}{(2\pi)^{m/4}}\right)\sum_{i=1}^{N}a_i\exp\left[-\frac{1}{4}\sum_{j=1}^{m}\left(y_{ji}-\mathbf{h}_j^T\mathbf{x}\right)^2\right].
\end{aligned}
$$

The MHDE fits to the observed data in the presence of a single bad leverage point and two-line data are shown in Figure 3.8 and Figure 3.9 respectively. The lines of best fit obtained via the MHDE are close to those obtained via the LMS estimator; from this one can conclude that, for simple linear regression with normally distributed errors, the MHDE – an estimator previously shown to be asymptotically efficient – can have a breakdown point of up to 50%. Additionally, the regression MHDE is robust against vertical outliers, as shown in Figure 3.10.

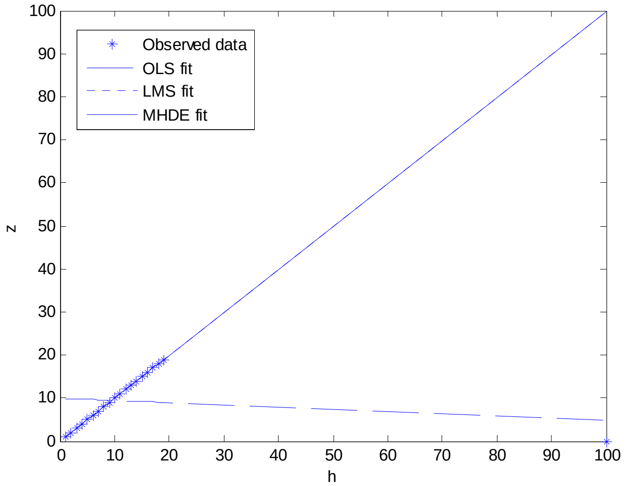

Figure 3.8: *MHDE, LMS and OLS regression fits in the presence of a single outlier in leverage position at (100, 0).*

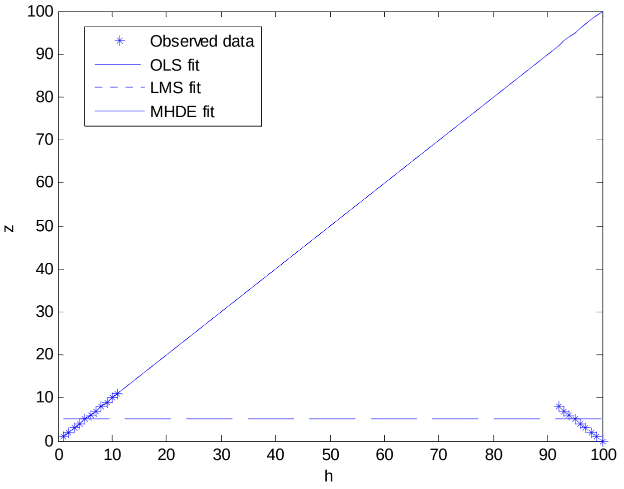

Figure 3.9: *MHDE, LMS and OLS regression fits in the two-line problem.*

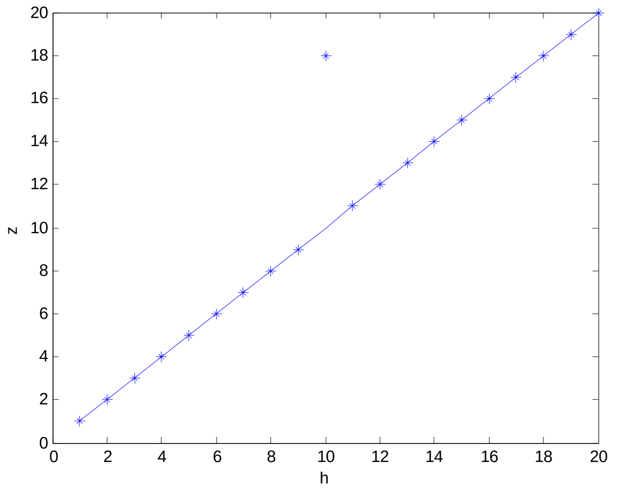

Figure 3.10: *MHDE fit of straight line in the presence of a single vertical outlier.*

3.4 Conclusion

In this chapter, we have confirmed previous robustness and efficiency results related to the minimum Hellinger distance estimator. We have extended these results to show that the MHDE is robust in the case of simple linear regression with normally distributed errors. In the next chapter, we will address the issue of tractability of the MHDE, and investigate a technique for algorithmic speedup in the regression case.

Chapter 4

The Natural Gradient Algorithm for MHDE

In the previous chapter we encountered three different estimators, namely the OLS, the LMS, and the MHDE. The OLS method involves the solution of a linear optimization problem, making it practical to implement, but undesirable due to its lack of robustness. The LMS method is able to reject a large fraction of outliers, but involves the repeated solution of a combinatorial optimization problem, making it difficult to implement for large-scale problems. Our estimator, the MHDE, exhibits the same level of robustness as the LMS for the problem described in Chapter 3, with a level of tractability that falls between the two other approaches mentioned: obtaining an estimate via the MHDE requires the solution of a nonlinear optimization problem. In this chapter, we provide a review of the underlying philosophy of iterative methods of equation solving. We then focus on a specific iterative method that will help us regard our problem space as having shape; the realization of this fact, along with the introduction of some abstract geometric concepts, will bring us full-circle to a surprising and elegant conclusion which can be applied to speed up the solution of the MHDE. Our review of iterative methods borrows from (Forsythe *et al*., 1977; Press *et al*., 1992; Kreyszig, 2006); our discussion of differential-geometric and information-geometric methods borrows from (Amari, 1982, 1985, 1998, 2001; Amari and Nagaoka, 1993; Arnold, 1989; Barndorf-Nielsen *et al*., 1986; Misner *et al*., 1973; Murray and Rice, 1993; Schutz, 1980).

4.1 Iterative Methods: A Review

It is often the case that we need to find the minimum of a function which has no analytical solution, or else finding the solution would require considerable time and resources. In such cases, we turn to one of any number of approximation methods. One particularly useful class of approximations is known as the class of *iterative methods*. While we will not attempt to catalog all such methods, we can say that they all adhere to the following general form:

```
Algorithm: Iterative solution
Inputs: (1) cost function $J(\mathbf{x})$, (2) initial estimate of $\mathbf{x}$
Outputs: (1) $\mathbf{x}_{k+1}$ such that $\mathbf{x}_{k+1} = \arg\min J(\mathbf{x})$
Procedure:
1. Hazard an initial guess at $\mathbf{x}$, called $\mathbf{x}_k$
2. Test if it minimizes the desired function
3. If so, halt. If not, put forth a modified guess $\mathbf{x}_{k+1}$
4. Repeat steps 2-3 until $\mathbf{x}_{k+1} \approx \arg\min J(\mathbf{x})$
```

While iterative algorithms follow this script, each varies in the prescribed manner that our guesses are to be modified, as well as additional means for accelerating our convergence to the final guess. In the next section, we shall describe one particular widely used method, the gradient descent.

4.2 Steepest Descent for MHDE

Recall that for a scalar function of a vector variable $J(\mathbf{x})$, the *gradient* $\nabla J(\mathbf{x})$ is a vector which points in the direction of most rapid ascent in the close vicinity of **x**. Alternately, $-\nabla J(\mathbf{x})$ gives the direction of maximal descent. In searching for the minimum of $J(\mathbf{x})$, it is the latter quantity that will prove most useful.

Iterative methods begin with a guess – more educated, or less – about the value of **x** that minimizes $J(\mathbf{x})$. If the initial guess $\mathbf{x}_0$ is away from the solution, the following algorithm can be implemented.

```
Algorithm: Steepest Descent
Inputs: (1) cost function $J(\mathbf{x})$, (2) initial estimate $\mathbf{x}_0$
Outputs: (1) $\mathbf{x}_{k+1}$ such that $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \le \varepsilon$
Procedure:
1. Initialize algorithm $\mathbf{x}_k = \mathbf{x}_\mathbf{0}$
2. Evaluate $\nabla J(\mathbf{x}_k)$
3. Calculate $\alpha$ such that
       $\frac{d}{d\alpha}[J(\mathbf{x}_k - \alpha \nabla J(\mathbf{x}_k))] = 0$
4. $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla J(\mathbf{x}_k)$
5. Repeat steps 2-4 until $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \le \varepsilon$
```

Each repetition of steps 2-4 constitutes one *iteration*. There are a few explanatory comments which need to be made about the steepest descent algorithm, starting with step 2 of the procedure. As previously stated, iterative algorithms require one to make a guess about the solution of an equation, and then modifying that guess repeatedly until guessing correctly. Since it is our goal to find the value of **x** that minimizes a given function $J(\mathbf{x})$, moving negatively along the gradient is a sound strategy for finding a minimum of $J(\mathbf{x})$.

One practical drawback to the gradient method is that for certain types of functions, the algorithm can take a large number of steps to find a solution. This occurs because the curves produced by such functions may not be very steep; think of a long, shallow, bowl-shaped curve such as the one shown in Figure 4.1. This is the purpose of step 3: if we can stretch or contract the gradient vector as needed, we can help our gradient descent algorithm converge much more rapidly. It is for this reason that the algorithm is referred to as steepest descent.

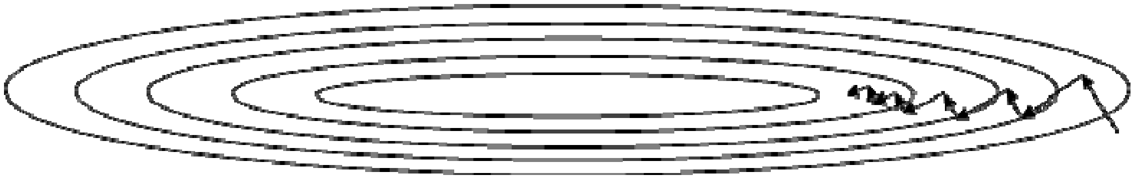

Figure 4.1: *An example of a curve for which gradient descent would converge slowly.*

Ideally, we would find an exact solution α for each iteration so that $J(\mathbf{x})$ is minimum in the direction of the gradient. There are two simpler strategies which can be employed at this step: the first is to pick a static value for α (setting $\alpha = 1$ is a safe strategy) and abide by the extended computing time as a result. Though this strategy, known as *ordinary gradient descent,* is suboptimal, in many instances it returns the correct output in an acceptable amount of time. The other strategy is to employ a fixed number of iterations of another optimization algorithm, such as Newton's method or a line search. This is a hedge between simply setting $\alpha = 1$ and finding an exact solution, which may provide a "good enough" speedup in the gradient descent algorithm. The chief danger is that a value of α that is too large will result in an "overshooting" phenomenon, where rather than gradually coalescing upon the correct answer, each iteration skips past the solution point. In our algorithm, we apply a small number of iterations of line search via the bisection method to find an approximation of α at each iteration.

In practice, both steepest descent and ordinary gradient descent have the drawback of a slow convergence rate, especially near the solution. To overcome this weakness, one would employ ordinary gradient descent or steepest descent to get close to a solution, and then switch to another algorithm such as Newton's method to rapidly gain several decimal places of accuracy. The purpose of this chapter is not to implement a full solution, but rather to show how geometric principles can be used to reach the neighborhood of the solution more rapidly.

Step 4 incorporates our halting criterion; ideally, we would like our algorithm to run until we get $\mathbf{x}_{k+1} = \mathbf{x}_k$, or equivalently, $\nabla J(\mathbf{x}_k) = \mathbf{0}$. In reality, each iteration of our algorithm might move only slightly closer to the solution point, which means that our algorithm could spend a long time generating values of $\mathbf{x}_{k+1}$ that are arbitrarily close to the solution without ever halting. It is much more practical to declare that we only need to know the solution to a few decimal places. The halting criterion then becomes $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \varepsilon$, where ε is quite small.

We now apply the gradient method to the solution of the regression MHDE. While the correctness of our solution was addressed in Chapter 3, here we will be strictly concerned with the speed at which that solution was obtained. Recall that the cost function in this case is

$$J(\mathbf{x}) = \kappa_1 \sum_{i=1}^{N} a_i \exp\left[-\frac{1}{4} \sum_{j=1}^{m} \left(y_{ji} - \mathbf{h}_j^T \mathbf{x} \right)^2 \right], \tag{4.1}$$

where

$$a_i = \left(\frac{1}{m(2\pi)^{m/2}} \sum_{j=1}^{m} \exp\left[-\frac{1}{2} \sum_{i=1}^{m} \left(y_{ji} - z_j \right)^2 \right] \right)^{\frac{1}{2}}, \text{ and } \kappa_1 = -2\left(\frac{1}{(2\pi)^{m/4}} \right). \tag{4.2}$$

The compact notation we applied early on allows us to compute the gradient of $J(\mathbf{x})$ with comparatively little difficulty. We write the gradient as

$$\nabla J(\mathbf{x}) = \kappa_1 \sum_{i=1}^{N} a_i \left[\frac{\partial}{\partial \mathbf{x}} \exp\left[-\frac{1}{4} \sum_{j=1}^{m} \left(y_{ji} - \mathbf{h}_j^T \mathbf{x} \right)^2 \right] \right]. \tag{4.3}$$

To compute this derivative, we must employ the chain rule; this requires computing the derivative of the argument of the exponential. Expanding terms, we have

$$-\frac{1}{4}\sum_{j=1}^{m}\left(y_{ji}-\mathbf{h}_j^T\mathbf{x}\right)^2=-\frac{1}{4}\sum_{j=1}^{m}y_{ji}^2-2y_{ji}\mathbf{h}_j^T\mathbf{x}+\left(\mathbf{h}_j^T\mathbf{x}\right)^2. \tag{4.4}$$

Taking the derivative of this expression gives us

$$\begin{aligned}\frac{\partial}{\partial\mathbf{x}}\left[-\frac{1}{4}\sum_{j=1}^{m}\left(y_{ji}-\mathbf{h}_j^T\mathbf{x}\right)^2\right]&=\frac{\partial}{\partial\mathbf{x}}\left[-\frac{1}{4}\sum_{j=1}^{m}y_{ji}^2-2y_{ji}\mathbf{h}_j^T\mathbf{x}+\left(\mathbf{h}_j^T\mathbf{x}\right)^2\right]\\&=-\frac{1}{4}\sum_{j=1}^{m}-2y_{ji}\mathbf{h}_j^T+2\left(\mathbf{h}_j^T\right)^2\mathbf{x}\\&=\frac{1}{2}\sum_{j=1}^{m}\mathbf{h}_j\left(y_{ji}-\mathbf{h}_j^T\mathbf{x}\right)\end{aligned} \tag{4.5}$$

Applying this to the gradient of the cost function, we have

$$\nabla J(\mathbf{x})=\kappa_1\sum_{i=1}^{N}a_i\left[\frac{1}{2}\sum_{j=1}^{m}\mathbf{h}_j\left(y_{ji}-\mathbf{h}_j^T\mathbf{x}\right)\right]\exp\left[-\frac{1}{4}\sum_{j=1}^{m}\left(y_{ji}-\mathbf{h}_j^T\mathbf{x}\right)^2\right]. \tag{4.6}$$

Thus far we have derived the objective function for the MHDE, as well as its gradient. In the previous chapter, we were concerned with the performance of our estimator in terms of statistical robustness, and found that for the cases we examined, the MHDE performs satisfactorily from this standpoint. In this chapter, we ask the question: for a given implementation of our estimator, how long does it take to find a solution? That is now our criterion for performance.

Without the benefit of a formal background in computational complexity, we are somewhat limited in what we can say about the performance of our algorithm; however, there are still a handful of constructive observations that we can make. Specifically, we can ask: how does the run time of the steepest descent algorithm for the MHDE vary as a function of the number of observations *m*? How does run time vary with the number of dimensions *n*?

To see how the execution time of the steepest descent algorithm varies with the number of observations *m*, we use a simple example: try to find the exact fit of a line with unit slope, zero intercept. In this case, the number of parameters is constant with *n*=2, while the number of observations vary over the range *m*=2 to *m*=50. The resulting execution time as a function of *m* is shown as a bilogarithmic plot in Figure 4.2. Taking the median slope between all points, we find that $t = O\left(m^{1.73}\right)$. The steepest descent algorithm for the MHDE has nearly quadratic computational complexity as a function of the number of observations.

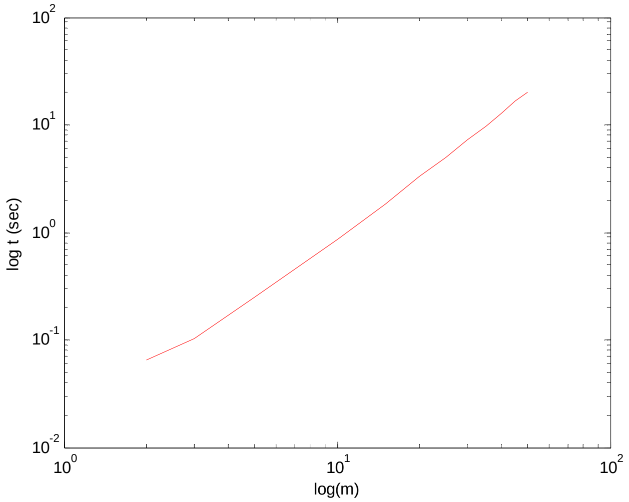

Figure 4.2: *Execution time of steepest descent as a function of the number of observations, m.*

To examine the run-time behavior of the steepest descent algorithm for the MHDE as a function of the number of parameters *n* we use a slightly different set of observations, for reasons that will become clear in Chapter 5. In this case, for *n*=2 we try to find an exact fit for

$$\mathbf{H}=\begin{bmatrix} 2 & -1 \\ 1 & -1 \\ -1 & 1 \\ -1 & 2 \\ 0 & 1 \\ 0 & -1 \\ -1 & -1 \\ -1 & 0 \\ 1 & 0 \end{bmatrix}; \mathbf{z}=\begin{bmatrix} 0 \\ -0.1 \\ 0.1 \\ 0.3 \\ 0.2 \\ -0.2 \\ -0.3 \\ -0.1 \\ 0.1 \end{bmatrix}; \mathbf{x}=\begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix}, \tag{4.7}$$

for *n*=3 we fit

$$\mathbf{H}=\begin{bmatrix} 3 & -1 & -1 \\ 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 3 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \\ -1 & 0 & -1 \\ -1 & 0 & 0 \end{bmatrix}; \mathbf{z}=\begin{bmatrix} -0.2 \\ -0.1 \\ 0 \\ -0.1 \\ 0.6 \\ 0.2 \\ 0.2 \\ -0.4 \\ -0.1 \end{bmatrix}; \mathbf{x}=\begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \end{bmatrix}, \tag{4.8}$$

and for *n* = 4 we fit

$$\mathbf{H}=\begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ 0 & 0 & -1 & 2 \\ 0 & -1 & 0 & -1 \\ 1 & 0 & -1 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}; \mathbf{z}=\begin{bmatrix} -0.3 \\ -0.2 \\ 0.2 \\ 0.5 \\ -0.6 \\ -0.2 \\ 0.1 \\ 0.1 \\ -0.1 \\ 0.4 \\ 0.2 \end{bmatrix}; \mathbf{x}=\begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \\ 0.4 \end{bmatrix}. \tag{4.9}$$

We note that these matrices describe measurements of 3, 4 and 5-bus power systems in specific configurations. This topic will be discussed in greater detail in Chapter 5.

The results, shown in Figure 4.3 below, indicate that the execution time of the steepest descent scales as approximately $t = O\left(n^{0.75}\right)$. We ask the following question: can we improve the run time of our algorithm?

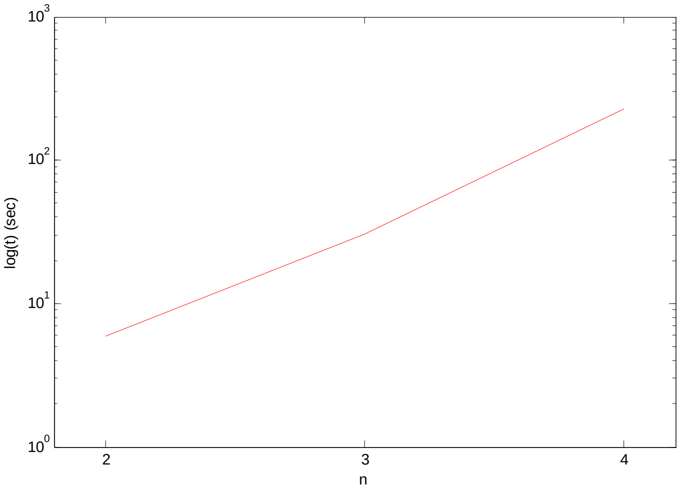

Figure 4.3: *Execution time of steepest descent as a function of the number of parameters, n.*

4.3 Differential Geometry: A review

Great innovations in the mathematical sciences come from the power of abstraction. The intuitive world around us has three spatial dimensions which we readily visualize and within which we comfortably move around. A more abstract concept of space allows us to find solutions to problems of increasing difficulty in an economical and elegant manner. The practice of imagining a parameter space as a physical space in aid of solving a complicated problem has numerous historical antecedents. In the physical sciences, the utility of Lagrange's equations – where combinations of Cartesian coordinates, angular coordinates, and other available, convenient degrees of freedom are combined to quickly and completely solve complicated equations of motion – are one such example of the revolutionary power of thinking beyond our ordinary three-dimensional world with great profit. More famously, Einstein forever altered our understanding of nature by advancing the notion that time, as a coordinate, is every bit as important as space, and that space-time in its most general sense must be considered not as some kind of four-dimensional Cartesian coordinate, but within the framework of differential geometry (Misner, *et. al.,* 1973).

It is the echoes of this latter work that we pick up when we consider problems in information geometry (Amari, 1985; Amari and Nagaoka, 1993; Murray and Rice, 1993). In the same way that many astonishing results fall out as a natural consequence of Einstein's view of nature, a geometric approach to information can yield interesting and helpful new ways of looking at problems. The field of information geometry is essentially an application of differential geometry to problems in statistics. While the contents of this field are rapidly expanding, only a few key concepts are needed to help us understand how we might better visualize our estimation problem, and how that might lead to an improved algorithm for solving the MHDE.

4.3.1 Manifolds and Probability Densities

For the purpose of our discussion, a few key concepts must be introduced. Absolute rigor is not necessary, but a general understanding of the underlying mathematical machinery combined with a few examples will give us what we need to move forward.

Central to our discussion, and to the use of natural gradient descent, is the concept of a *manifold* (Amari, 1985). While there are many definitions of a manifold which vary slightly from each other, one which is sufficient for our needs is the following: an n-dimensional manifold S is a set of points such that the open neighborhood around every point in S has a continuous one-to-one map to an open set of the n-dimensional real space $\mathbf{R}^n$. A manifold can be visualized as a space that locally resembles flat Euclidean space, but may look very different globally. As such, we can define a coordinate system for each part of S; a single coordinate system need not cover all of S.

Most smooth objects encountered in everyday life can be thought of as manifolds; to fix the concept in our minds, we introduce the canonical example: the surface of a 3-dimensional sphere. The surface of a sphere looks, locally, like flat, 2-dimensional space. Every point on the sphere can be mapped to the 2-dimensional plane $\mathbf{R}^2$ except for one of the poles, as shown in Figure 4.4. By introducing a second coordinate system, which maps all the points on the sphere to $\mathbf{R}^2$ except the opposite pole, we have a set of coordinate systems completely capable of mapping any point on the surface of a sphere to $\mathbf{R}^2$ (Arnold, 1989).

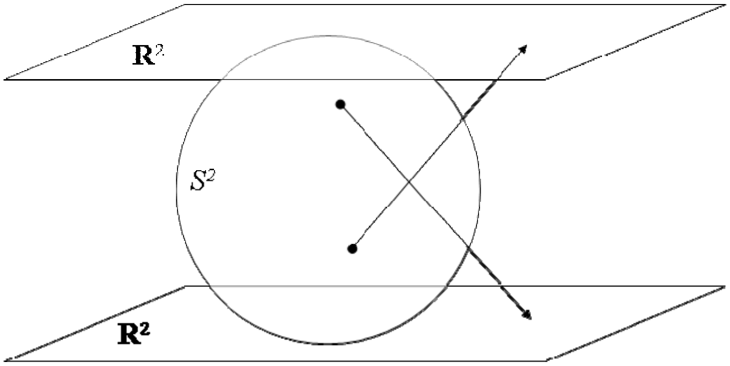

Figure 4.4: *The surface of a sphere can be mapped to the 2-dimensional real plane almost everywhere. A second set of coordinate systems allows us to cover the whole sphere. After (Arnold, 1989).*

In the context of this thesis, a more interesting and useful example of a manifold is the family of normal distributions whose probability density functions are expressed as

$$p(z;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-(z-\mu)^2}{2\sigma^2}\right] \tag{4.10}$$

For convenience of notation we set $\boldsymbol{\theta} = (\theta^1, \theta^2)$, where $\theta^1 = \mu$, $\theta^2 = \sigma$. The set $S = \{p(z;\boldsymbol{\theta})\}$ is then a manifold comprised of the normal probability density functions for all values of $\boldsymbol{\theta}$ for a given x. In this sense, we can say that $\boldsymbol{\theta}$ is the coordinate system for our manifold, and every point represents a density function.

4.3.2 Tangent Spaces and Natural Basis

In Euclidean space, the notion of a vector is intuitively understood as a straight line forming an arrow from one point to another. Having defined the concept of a manifold, we have no such intuition for the global behavior of a vector in such a space, since we do not necessarily have a single global system for addressing points on a manifold. Taking advantage of the “locally Euclidean” nature of a manifold, we choose to define vectors locally as well. This is achieved by constructing a linearized space tangential to a point on the manifold, and then defining a vector in that space. The following paragraphs outline how one does this formally.

At a point p on the manifold, we define a *tangent vector* as a vector tangential to a parameterized curve $\theta = \theta(t)$ passing through p. By convention, we say that the curve passes through p at t=0; a tangent vector A_p at point p is then expressed as

$$A_p = \left.\frac{d\theta}{dt}\right|_{t=0}. \qquad (4.11)$$

Once again using the surface of a sphere as an aid to visualization, one can imagine more than one curve passing through the same point from different directions, as shown in Figure 4.5. The vector space obtained by collecting the set of all tangent vectors passing through p is known as the *tangent space* T_p.

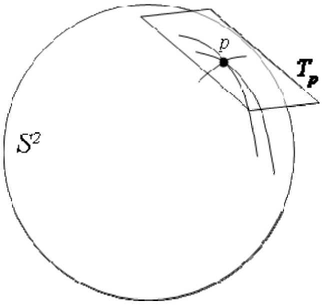

Figure 4.5: *The tangent space of a point on a manifold is defined by the set of all curves passing through that point.*

As with the vector spaces with which we are commonly familiar, the choice of basis vectors with which we represent the space may not be unique; however, as we will remind ourselves momentarily, some bases have more desirable characteristics than others. We choose the partial derivatives with respect to $\{\theta^1, \theta^2, ..., \theta^n\}$to form the basis of our tangent space, and abbreviate these partial derivatives $\partial_i \equiv \partial / \partial\theta^i$. The set $\{\partial_i\}$ is called the *natural basis* of the tangent space with respect to our chosen coordinate system. In this way we can represent any vector A_p in the tangent space T_p as a linear combination of the bases:

$$A_p = A_p{}^i \partial_i . \tag{4.12}$$

Equipped with the ability to express vectors in a tangent space, we now define a space that will become useful in explorations of statistics. Consider the manifold comprised of normal distributions $p(z;\mathbf{\theta})$. We can define a related manifold, built up from the logarithm of the normal distribution:

$$\ell(z;\mathbf{\theta}) = \log[p(z;\mathbf{\theta})]. \tag{4.13}$$

When we use the natural basis to describe points on our manifold, the tangent space of $\ell(z;\mathbf{\theta})$ is written $T_p^{(1)}$, and we call this the *1-representation* of the tangent space. We will use this representation extensively for the remainder of the chapter, because doing so allows us to highlight some interesting statistical properties. One such example involves the expectation of a statistical distribution. If

$$E[f(z)] = \int_R f(z)p(z;\mathbf{\theta})dz , \tag{4.14}$$

then

$$\begin{aligned}0 &= \partial_i \int_R p(z;\boldsymbol{\theta})dz \\ &= \int_R \partial_i p(z;\boldsymbol{\theta})dz \\ &= \int_R p(z;\boldsymbol{\theta})\partial_i \ell(z;\boldsymbol{\theta})dx \\ &= E[\partial_i \ell(z;\boldsymbol{\theta})].\end{aligned} \tag{4.15}$$

This tells us that, in this construction, $E[A_p(z)]=0$ for any vector $A_p(z) \in T_p^{(1)}$; for this reason it is an appropriate choice of basis.

We continue to consider the normal distribution defined in the previous section, with $\theta^1 = \mu$, $\theta^2 = \sigma$. What is the natural basis of this distribution? In our chosen coordinate system, the natural basis is written:

$$\begin{aligned}\partial_1 &= \frac{\partial}{\partial \mu}, \\ \partial_2 &= \frac{\partial}{\partial \sigma}.\end{aligned} \tag{4.16}$$

To get our 1-representation, we start by taking the log of the normal distribution:

$$\ell(z;\boldsymbol{\theta}) = -\frac{(z-\mu)^2}{2\sigma^2} - \log\left(\sqrt{2\pi}\sigma\right), \tag{4.17}$$

and the basis of the 1-representation is therefore

$$\begin{aligned}\partial_1 \ell &= \frac{z-\mu}{\sigma^2}, \\ \partial_2 \ell &= \frac{(z-\mu)^2}{\sigma^3} - \frac{1}{\sigma}.\end{aligned} \tag{4.18}$$

4.3.3 Metrics and Fisher Information

A tangent space gives us the ability to define vectors in a curved space. To work meaningfully with these vectors, it is helpful to have a way of defining length and angle on a manifold. This is accomplished by defining an inner product on the manifold.

When the inner product $\langle A,B \rangle$ of two tangent vectors *A* and *B* $\in T_p$ are defined, the manifold is called a *Riemannian manifold*. The choice of an inner product is not unique, but there is a natural way of defining the inner product that is helpful in studying statistics. Consider two random variables $A(z)$ and $B(z)$, which are the 1-representations of *A* and *B*. We define the inner product

$$\langle A,B \rangle = E[A(z)B(z)]. \tag{4.19}$$

Since, as shown in the previous section, *E*[*A*(*z*)]=*E*[*A*(*z*)]=0 for our choice of basis, the inner product is just the covariance Cov[*A*(*z*),*B*(*z*)]. At this point, it would be reasonable to ask what would happen if we were to take the inner product of the basis vectors ∂_i and ∂_j. When we do this, we get:

$$g_{ij}(\boldsymbol{\theta}) = \langle \partial_i, \partial_j \rangle = E\left[\partial_i \ell(z;\boldsymbol{\theta}) \partial_j \ell(z;\boldsymbol{\theta})\right]. \tag{4.20}$$

The result is a multilinear object known as a *tensor*. This particular tensor allows us to take the calculus-based algorithm for computing an inner product, and turn it into a simple algorithmic process. The quantity g_{ij} is called the *metric tensor*, and the inner product of two vectors can be expressed as a component sum in the following way:

$$\langle A,B \rangle = \langle A^i \partial_i, B^j \partial_j \rangle = A^i B^j g_{ij} \tag{4.21}$$

The metric tensor, as the name implies, directly gives us the means for defining distance between two vectors in a tangent space. But what is the statistical interpretation of such a construct? When applied to the 1-representation, the metric tensor is exactly the Fisher information matrix. This leads to a natural means of comparing the “distance” of two estimators or two probability distributions. In the metric framework, the distance between two estimators can be written as

$$ds^2 = g_{ij}(\boldsymbol{\theta})(\theta'^i - \theta^i)(\theta'^j - \theta^j) \tag{4.22}$$

If we let the difference between θ^i and θ'^i become very small, we are furnished with a natural calculus-based definition of distance: $ds^2 = g_{ij} d\theta^i d\theta^j$. If we go back to our parameterized view of $\theta=\theta(t)$, we can rewrite our definition of distance as $ds^2 = g_{ij}[\theta(t)]\dot{\theta}^i \dot{\theta}^j dt^2$. Therefore,

$$s = \int_{t_0}^{t_1} \sqrt{g_{ij}\dot{\theta}^i \dot{\theta}^j}\, dt \tag{4.23}$$

is the distance between two points in a tangent space, and the minimum-distance curve is referred to as the *Riemannian geodesic* connecting two points.

We can better understand some of these quantities with the concrete example of the normal distribution. For such a distribution, the metric tensor is

$$g_{ij} = \left\langle \partial_i, \partial_j \right\rangle = \frac{1}{\sigma^2}\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \tag{4.24}$$

and the Riemann distance between two points $\underset{1}{\boldsymbol{\theta}} = \left(\mu_1, \sigma_1\right)$ and $\underset{2}{\boldsymbol{\theta}} = \left(\mu_2, \sigma_2\right)$ is

$$d(\underset{1}{\boldsymbol{\theta}}, \underset{2}{\boldsymbol{\theta}}) = \sqrt{2}\cosh^{-1}\left[\frac{(\mu_1 - \mu_2)^2 + 2\left(\sigma_1^2 + \sigma_2^2\right)^2}{4\sigma_1\sigma_2}\right]. \tag{4.25}$$

4.4 Natural Gradient Descent

We return to the gradient descent algorithm. As previously mentioned, one drawback of the algorithm is the fact that for curves and surfaces that vary slowly in space, many iterations are required to optimize our cost function. From a geometric standpoint, it is important to note that we are searching for a minimum of our curve in a Euclidean space. The intuition to be gained from the previous section is that, by treating our problem space as a Riemannian manifold, and using information we may have about the structure of this manifold, we can achieve a speedup of the gradient descent algorithm.

The *natural gradient descent* (NGD) algorithm, first employed by Amari (1998), makes use of the manifold structure of a problem space by pre-multiplying the gradient of the cost function by the metric:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \mathbf{G}^{-1} \nabla J(\boldsymbol{\theta}_k), \tag{4.26}$$

where $\mathbf{G}^{-1} = \{g_{ij}\}^{-1}$. The net effect is that instead of taking steps along the surface defined by $J(\boldsymbol{\theta})$ in Euclidean space, each iteration of the natural gradient steps along the surface of the manifold of possible solutions.

In the case of finding the regression MHDE, we saw that steepest descent can take a linearly increasing amount of time to run to completion as the number of dimensions *N* increases, and a quadratically increasing amount of time as the number of observations *m* increases. Does the natural gradient approach offer any improvement in execution time?

To implement the natural gradient descent algorithm for the regression MHDE, we must compute the metric, also known as the Fisher information matrix:

$$\begin{aligned}\mathbf{G} &= I_f(z) \\ &= \int\left(\frac{f'(\mathbf{z};\boldsymbol{\theta})}{f(\mathbf{z};\boldsymbol{\theta})}\right)^2 f(\mathbf{z};\boldsymbol{\theta})d\mathbf{z} \\ &= E\left[\left(-\frac{\partial}{\partial\boldsymbol{\theta}}\ln(f(\mathbf{z};\boldsymbol{\theta}))\right)^2\right].\end{aligned} \tag{4.27}$$

In the previous chapter, we chose as our error model the standard Gaussian distribution. For a set of observations in vector form, we can write

$$f(\mathbf{z};\mathbf{x}) = \left(\frac{1}{(2\pi)^m|\mathbf{R}|}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{z}-\mathbf{Hx})^T\mathbf{R}^{-1}(\mathbf{z}-\mathbf{Hx})\right), \tag{4.28}$$

where $\mathbf{R}$ is the covariance. In the model chosen, $\mathbf{R}=\mathbf{I}$, the identity matrix. Taking the logarithm of both sides, we have

$$-\ln f(\mathbf{z};\boldsymbol{\theta}) = \frac{1}{2}(\mathbf{z}-\mathbf{H}\boldsymbol{\theta})^T\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\boldsymbol{\theta}) - \ln\left(\frac{1}{(2\pi)^m|\mathbf{R}|}\right)^{\frac{1}{2}}. \tag{4.29}$$

Differentiating both sides of this equation gives us

$$-\frac{\partial}{\partial\boldsymbol{\theta}}\ln f(\mathbf{z};\boldsymbol{\theta}) = \mathbf{H}^T\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\boldsymbol{\theta}). \tag{4.30}$$

The final step in computing the Fisher information is to calculate the expectation of the above quantity squared:

$$\begin{aligned}E\left[\left(-\frac{\partial}{\partial\boldsymbol{\theta}}\ln f(\mathbf{z};\boldsymbol{\theta})\right)^2\right] &= E\left[\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\boldsymbol{\theta})(\mathbf{z}-\mathbf{H}\boldsymbol{\theta})^T\mathbf{R}^{-1}\mathbf{H}\right] \\ &= \mathbf{H}^T\mathbf{R}^{-1}\mathbf{R}\mathbf{R}^{-1}\mathbf{H} \\ &= \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}.\end{aligned} \tag{4.31}$$

Since $\mathbf{R} = \mathbf{I}$ in our model, $\mathbf{G} = \mathbf{H}^T\mathbf{H}$. For linear regression when the error model is standard Gaussian, the natural gradient descent algorithm is then very easy to implement:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\nabla J\left(\boldsymbol{\theta}_k\right). \tag{4.32}$$

When we augment the gradient descent in this way, there is a noticeable improvement in the execution time of the algorithm. Recall that the run-time of the steepest descent algorithm scaled almost quadratically with the number of observations *m*; the NGD algorithm scales in a linear fashion, as shown in Figure 4.6. Taking the median of the slope between points on the curve in Figure 4.6, we find that the execution time of the NGD algorithm is approximately $t = O\left(m^{0.83}\right)$.

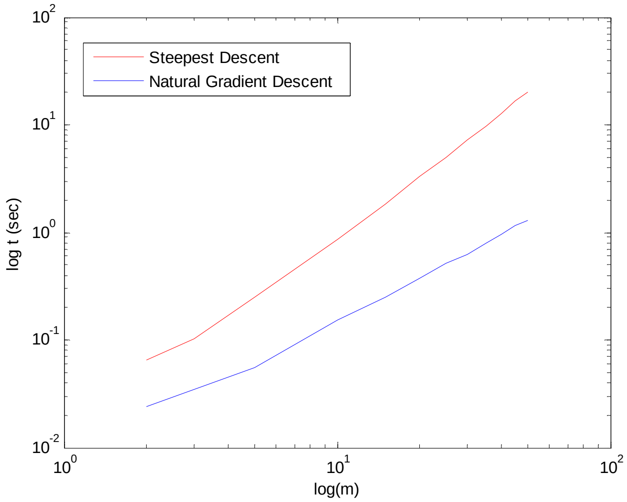

Figure 4.6: *Execution time of natural gradient descent and steepest descent as a function of the number of observations, m.*

As we vary the number of parameters *n*, we see that the NGD algorithm offers a more modest, but still notable improvement: our execution time scales as $t = O\left(n^{0.64}\right)$, compared with $t = O\left(n^{0.75}\right)$ for the steepest descent algorithm.

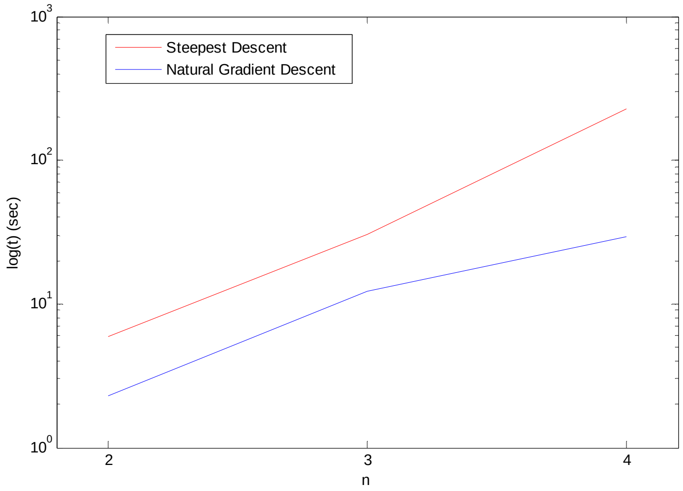

Figure 4.7: *Execution time of natural gradient descent and steepest descent as a function of the number of parameters, n.*

4.5 Conclusions

In this chapter, we discussed the basic theory behind iterative numerical methods. We covered a particular algorithm, the gradient descent algorithm, in detail. We introduced the basics of differential geometry, including manifolds, tangent spaces, and metrics, and showed how the fundamental components of probability and statistics relate to these geometric constructions. We then showed that by incorporating the manifold structure of the space of probability density functions into our algorithm, we can achieve a significant speedup in the solution of our estimator. In the next chapter, we apply our robust, efficient estimation scheme – now with improved execution time – to a rudimentary problem in power systems.

Chapter 5

Application of MHDE to Power Systems

5.0 Introduction

Having demonstrated the robustness of our estimator for problems of univariate location as well as simple linear regression, and having further implemented a novel algorithm for solution of these problems, we now examine one area of ongoing study where these results might have some practical use. The secure, stable operation of power systems relies on estimation of the state variables of the power transmission system from a redundant collection of power and voltage measurements spread throughout the network. However, these measurements are often the source of biased or missing data; too many of these corrupt measurements may lead to unreliable state estimation results. The use of the MHDE provides the robustness needed in power system state estimation. In the following sections, we introduce the DC power flow model and apply the MHDE to this model on a 3-bus system. Our discussion of relevant background is based on (Abur and Exposito, 2004; Mili *et al.*, 1994; Mili and Coakley, 1996; Schweppe and Rom, 1970).

5.1 DC State Estimation of a 3-Bus Power System

An electric power system exists as a complex network of subsystems, which provide electric energy over a wide area for a variety of uses. Such a system generally consists of transmission lines, shunt capacitors, transformers, and generators which are modeled via equivalent circuits. Modern electric power systems contain a collection of software programs whose purpose is monitoring and control of power transmission and generation. At their heart, there is the static state estimator, which processes a host of measurements collected every few minutes via a scanning procedure supposed to provide a true snapshot of the system state. The measurements include metered values on power flows and power injections together with some nodal voltage magnitudes. In theory, the power system is a very large circuit, subject to Kirchhoff's and Ohm's laws. However, many measurement devices which coexist on the grid contain components which vary in age, quality, and calibration; as a result some of them may provide strongly biased metered values. To make matters worse, subsystems in certain portions of the transmission grid operate without meters, which translate into unobservable subsystems. Bursts of large noise or failure of one or more measurements can translate into either missing or corrupted data, depending on the failure mode. Furthermore, topology errors may occur following unreported switching of circuit breakers. Given a descriptive model of the power system, our task is to estimate the overall state of the system in a reliable manner despite the presence of gross measurement errors.

5.2 Structured Regression in Power Systems

5.2.1 Regression Model of Networked Elements

The choice of design matrices in Chapter 4 was not accidental; these matrices correspond to 3, 4, and 5-bus power systems with specific measurement configurations. In this chapter, we will examine the behavior of the MHDE on a 3-bus power system with the topology shown in Figure 5.1; the small ovals on the diagram represent injection and flow measurements throughout the system. For the purposes of testing the robustness of our estimation scheme, a *DC state estimation model* is sufficient. In the DC model, we set all bus voltage magnitudes around 1.0 per unit, neglect all resistances and shunt capacitances, and assume small phase differences between nodal voltage phasors.

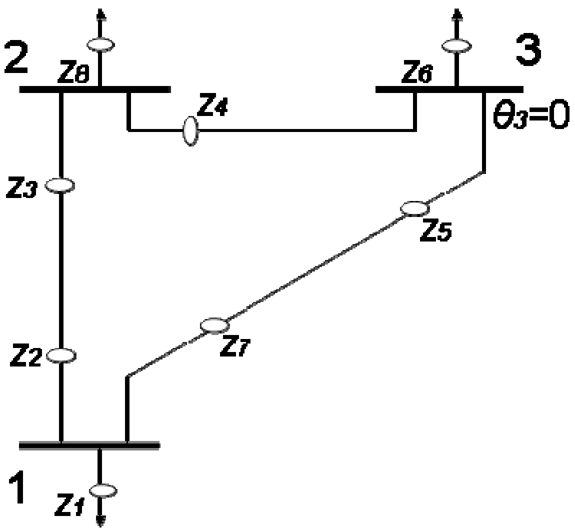

Figure 5.1: *One-line diagram of a 3-bus system with 7 measurements.*

The utility of the DC model is that it allows us to focus on the analysis of our state estimation scheme by linearizing the power system model. This approximation is only valid for small phase angles.

In the DC approximation, the power flow P_{ij} on a line from bus *i* to bus *j* is related to the voltage phase angles θ_i and θ_j by the equation

$$P_{ij} = \frac{\theta_i - \theta_j}{X_{ij}}, \tag{5.1}$$

and the injection at bus *i* is the sum of all flows incident to bus *i*:

$$P_i = \sum_j P_{ij}. \tag{5.2}$$

One nodal voltage phasor in the system is taken as the reference by setting its phase angle arbitrarily to zero; by convention, we take the highest-numbered bus as a reference and estimate voltage phasors at all other buses relative to the voltage phasor of that bus. The system of measurements can be expressed as a linear regression model of the form

$\mathbf{z} = \mathbf{Hx} + \mathbf{e}$, where **H** is the $m \times n$ dimensional Jacobian matrix $\mathbf{H} = \begin{bmatrix}\mathbf{h}_1^T & \cdots & \mathbf{h}_m^T\end{bmatrix}^T$, and $\mathbf{x} = \begin{bmatrix}\theta_1 & \cdots & \theta_N\end{bmatrix}^T$. Unlike the linear regression model presented in Chapter 3, the design space spanned by the row vectors of **H** no longer represents a slope-intercept equation, but the coordinates of a hyperplane passing through the origin.

5.2.2 Exact Fit Point and Breakdown Point

A set of *m* measurements lies in an (*n*+1) dimensional space $(z, \theta_1, ..., \theta_n)$. Each measurement is an ordered *n*+1-tuple consisting of an observation z_i and its corresponding row vector in the Jacobian matrix is $\mathbf{h}_i^T$. If an estimator of *n* unknown state variables fits an *n*-dimensional hyperplane when the majority of measurements lie on the hyperplane, the estimator is said to have the *exact fit property*. The *exact fit point* of an estimator is the maximum fraction δ^* of outliers that the estimator can tolerate before it no longer fit that hyperplane. Formally, this is expressed as

$$\delta^*(T, Z) = \max\{\delta\,; T(Z') = T(Z)\} \text{ for all } Z' \qquad (5.3)$$

where *Z* is a sample that lies on the hyperplane, *Z*' is a sample where some fraction *δ* of points lie outside the hyperplane.

5.2.3 General Position, Reduced Position, and Structured Regression

A set of *m* measurements is said to be in *general position* if any *n* row vectors of **H** are linearly independent. The practical implication of general position is that any *n*-row subset of **H** is sufficient to uniquely determine **x.** When a regression model contains measurements in general position, the maximum breakdown point of an estimator is

$$\varepsilon_{\max}^* = \left[\frac{m-n}{2}\right] \Big/ m, \qquad (5.4)$$

where $[\bullet]$ denotes the integer part of the argument (Rousseeuw and Leroy, 1987).

The *maximum exact fit point* $\delta^*_{\max}$ is defined as the largest exact fit point attainable by any estimator for a given design space. Rousseeuw and Leroy (1987) have conjectured that when the measurements are in general position, the maximum exact fit point is equal to the maximum breakdown point.

A set of observations are in *reduced position* if one or more rows of **H** are linearly dependent. If this is the case, there is no guarantee that an arbitrary subset of **H** can be used to uniquely determine **x**. When a design matrix is in reduced position, a linear regression model is called a *structured regression model* (Mili and Coakley, 1996).

5.2.4 Fundamental Sets, Critical and Surplus Measurements

A helpful concept for discussing breakdown and exact fit in structured regression models is that of the *fundamental set*. A fundamental set of a state variable θ_j is the set of all measurements whose row vectors $\mathbf{h}_i^T$ have non-zero elements associated with that state variable.

A *critical measurement* is one whose elimination drops the rank of the design matrix by at least one. The *surplus* s^* is the smallest number of measurements whose removal turns one or more measurements into a critical measurement. The concept of surplus is a global one; locally, each fundamental set has its own surplus s_i^*. The global surplus is then

$$s^* \le \min_i s_i^*. \tag{5.6}$$

For a structured regression model, an upper bound of the maximum exact fit point is given by (Mili, *et. al.*,1994):

$$\delta^*_{\max} \leq \left[\frac{s^*}{2}\right] \Big/ m. \tag{5.7}$$

The exact expression of the maximum exact fit point was derived in (Mili *et al.,* 1994; Mili and Coakley, 1996). Let *M* be the maximum number of measurements whose projections on the factor space lie on a (*n*-1)-dimensional vector space. *M* is given by

$$\delta^*_{\max} = \left[\frac{m - M - 1}{2}\right] \Big/ m. \tag{5.8}$$

5.2.5 Power System State Estimation as Structured Regression

Consider the 3-bus power system shown in Figure 5.1 above. For this system, the Jacobian matrix is given by

$$\mathbf{H} = \begin{bmatrix} 2 & -1 \\ 1 & -1 \\ -1 & 1 \\ 0 & 1 \\ -1 & 0 \\ 1 & 0 \\ -1 & -1 \\ -1 & 2 \end{bmatrix}. \tag{5.9}$$

In this example, the fit is a two-dimensional plane passing through the origin. The linear dependency between the second and third rows of the design matrix mean that this matrix is in reduced position; the model is a structured regression model. The elements of the Jacobian matrix are shown in a factor space in Figure 5.2. For the 3-bus system, *M* is the maximum number of row vectors of **H** that lie on a line passing through the origin. For our measurement configuration, *M*=2, yielding $\delta^*_{\max} = \left[(m - M - 1)/2\right]/m = 2/8$. This means that globally, only two outliers are required to cause any estimator to break down for a system-wide state estimate. In contrast, the maximum breakdown point, had the measurements been in general position, would be [(*m*-*n*)/2]/*m*= 3/8.

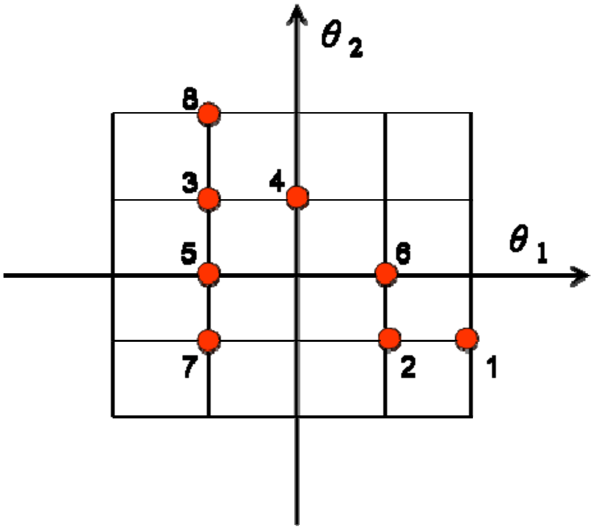

Figure 5.2: *Design space of 3-bus system with 8 measurements.*

5.3 MHDE in Power System State Estimation

5.3.1 Practical Considerations of Least Median of Squares Estimates in Power Systems

It was shown by Mili and Coakley (1996) that in structured regression models, the LMS must be tuned to minimize the square of a particular quantile based on the maximum exact fit point of the entire system. This adds more computational tasks to an estimation process that is already known to be computationally complex. One potential advantage that the MHDE might offer in the context of power system state estimation is a similar level of robustness against outliers without the need for tuning the estimator for a specific measurement configuration.

5.3.2 Exact Fit Point of MHDE

For this discussion, our Jacobian matrix **H** is the one given in Section 5.2.5, and our true parameter vector is $\mathbf{x} = [0.1 \quad 0.2]^T$. We examine the exact fit point of the MHDE by replacing some portion of our measurement vector **z** with outliers. Table 5.1 shows the effects of adding outliers on our estimator. The MHDE, in this case, does not obtain the maximum breakdown point in all cases. This cannot be attributed definitively to a limitation of the estimator, but rather it is due to numerical truncation errors inherent to the computer simulation which lead to a sensitivity of the algorithm to the position of outliers in our measurement space. An improved implementation of the MHDE which can circumvent this limitation of desktop computers is one area for future research.

z	ε	Outliers	$\hat{x}_1$	$\hat{x}_2$
$[0 \quad -0.1 \quad 0.1 \quad 0.2 \quad -0.1 \quad 0.1 \quad -0.3 \quad 0.3]^T$	0	uncontaminated	0.0998	0.1997
$[0 \quad -0.1 \quad 0.1 \quad 0.2 \quad -0.1 \quad 0.1 \quad -0.3 \quad 0]^T$	1/8	z_8	0.0998	0.1996
$[0 \quad 0 \quad 0 \quad 0.2 \quad -0.1 \quad 0.1 \quad -0.3 \quad 0.3]^T$	2/8	z_2, z_3	**0.0033**	**0.0061**
$[0 \quad -0.1 \quad 0 \quad 0 \quad -0.1 \quad 0.1 \quad -0.3 \quad 0.3]^T$	2/8	z_3, z_4	0.0998	0.1992
$[0 \quad -0.1 \quad 0 \quad 0.2 \quad 0 \quad 0.1 \quad -0.3 \quad 0.3]^T$	2/8	z_3, z_5	**0.0848**	**0.1695**
$[0 \quad -0.1 \quad 0.1 \quad 0 \quad -0.1 \quad 0.1 \quad -0.3 \quad 0]^T$	2/8	z_4, z_8	0.0998	0.1996
$[0 \quad -0.1 \quad 0.1 \quad 0 \quad 0 \quad 0 \quad -0.3 \quad 0.3]^T$	3/8	z_4, z_5, z_6	**0.0920**	**0.1850**

Table 5.1: *Regression estimates for 3-bus system with various configurations of outliers. Bold numbers represent estimates where breakdown is evident. Ground truth is [0.1 0.2]^T.*

5.3.3 Robustness of MHDE to Bad Leverage Points

A shortened line in a power system has the effect of adding a bad leverage point to our regression model. By shortening the line between Bus 2 and Bus 3 to one-tenth of its original length, the impedance becomes X_{23}=0.1 p.u. Our Jacobian matrix is given by:

$$\mathbf{H} = \begin{bmatrix} 2 & -1 \\ 1 & -1 \\ -1 & 1 \\ 0 & 10 \\ -1 & 0 \\ 1 & 0 \\ -1 & -1 \\ -1 & 11 \end{bmatrix}, \tag{5.10}$$

and our factor space is altered to look like Figure 5.3.

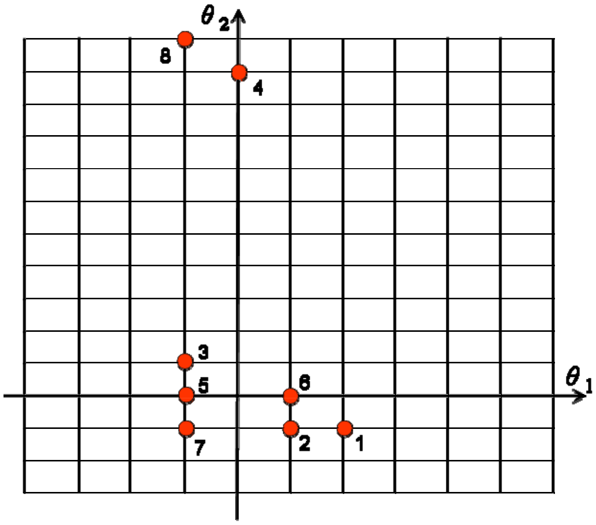

Figure 5.3: *Design space of 3-bus system with 8 measurements, altered to have 2 bad leverage points.*

In this case, the MHDE obtains an estimate of $\hat{\mathbf{x}} = [0.0080 \quad 0.0280]^T$; our estimator breaks down in the presence of two bad leverage points. If we remove the bad leverage point in the $\mathbf{h}_4^T$ position, the MHDE obtains $\hat{\mathbf{x}} = [0.0968 \quad 0.1941]^T$; it can withstand a single bad leverage point.

5.3.4 Discussion

The reason of the breakdown of the algorithm in the presence of two outliers stems from numerical approximations in the evaluation of the MHDE cost function, resulting in rounding errors. Specifically, the cost function involves the products of very small exponential numbers, resulting in even smaller numbers that are outside the range the computer double precision. One possible way to alleviate this problem is to perform an appropriate transformation of the cost function. This is an area that calls for further research and development.

5.4 Conclusions

In this chapter we examined the robustness properties of the MHDE in the context of power systems. Inherent difficulties with the cost function and its implementation prevent us from reaching any conclusions about the robustness of the MHDE in structured regression.

Chapter 6

Conclusions and Future Directions

6.1 Conclusions

In this thesis we have examined statistical procedures based on the notion of minimum distance functionals. We have introduced and analyzed the properties of an estimator, the minimum Hellinger distance estimator, and confirmed that it is both robust and efficient in the univariate location case. We have extended this result to show by example that the MHDE is highly robust against bad leverage points in linear regression.

The resulting estimator was implemented via two algorithms: the steepest descent algorithm and the natural gradient descent algorithm, which enhances steepest descent by taking into account the manifold structure of the space of parametric model probability density functions. The natural gradient descent decreased the execution time of the estimation procedure both as a function of the number of estimation parameters and the number of observations; for the latter, the optimization problem was reduced from nearly quadratic time to approximately linear time.

The concept of exact fit point and its relationship to breakdown point was discussed in the context of structured regression models. The MHDE was applied to a rudimentary problem in power systems, and the exact fit point of the estimator was observed. Due to limitations of the algorithm, it is not clear at this time whether the MHDE actually obtains the maximum exact fit point in structured regression. This result suggests one possible area of future research.

6.2 Future Directions

6.2.1 Robustness Theory of MHDE

Up until now, little has been done with minimum distance estimators for regression problems. There are interesting open questions on multiple fronts. The empirical results of this thesis suggest that the MHDE is highly robust against outliers which are far away from the bulk of the observed data.

The robustness and efficiency properties of the MHDE were considered in an empirical way; it is necessary to develop a formal understanding of the robustness of the MHDE from the perspective of influence functions. One topic which should be examined in the future is the robustness of the MHDE when the model density function is of a completely different family than the true probability density of the observed data. In a similar vein, the basic tools of differential geometry were applied to this estimation problem from an algorithmic standpoint. A more sophisticated treatment of the geometric structure of this estimator may give us a deeper insight into the general characteristics of this estimator, as well as its robustness and efficiency at different families of density functions.

6.2.2 Improved Algorithms for MHDE

The natural gradient descent algorithm offered a speedup over steepest descent, but still required an execution time on the order of tens of seconds to optimize the MHDE cost function for relatively small, low-dimensional data sets. Additionally, the current algorithm relies on repeated iterations of products of squared inverse exponentials –very small numbers multiplied together resulting in even smaller numbers; as a result the algorithm obtains a result quite slowly. A more pressing concern is that building up the repeated products of many observations quickly tests the limits of computer arithmetic; the truncation and rounding which takes place leads to the kind of algorithmic instability seen in the results of Chapter 5. To mitigate these characteristics, the problem needs to be recast. The investigation of transform methods to increase execution speed should be undertaken. Additionally, hybrid algorithms which combine natural gradient descent with Newton's method or other methods to more rapidly obtain a result should be investigated.

References

1. A. Abur and A. G. Expósito. *Power System State Estimation: Theory and Implementation*. New York, Marcel Dekker, 2004.

2. S. I. Amari. “Differential Geometry of Curved Exponential Families-Curvatures and Information Loss,” *The Annals of Statistics*, Vol. 10, No. 2 (June 1982), pp 357-385.

3. S. I. Amari. *Differential-Geometrical Methods in Statistics*. Berlin, Springer-Verlag, 1985.

4. S. I. Amari. ”Natural Gradient Works Efficiently in Learning,” *Neural Computation,* Vol. 10 (1998), pp 251-276.

5. S. I. Amari. “Information Geometry on Hierarchy of Probability Distributions,” *IEEE Transactions on Information Theory*, Vol. 47, No. 5 (July 2001), pp 1701-1711.

6. S. I. Amari and H. Nagaoka. *Methods of Information Geometry*. Providence, American Mathematical Society, 1993.

7. T. W. Anderson. *An Introduction toMultivariate Statistical Analysis/3e.* Hoboken, John Wiley and Sons, 2003.

8. V. I. Arnold. *Mathematical Methods of Classical Mechanics/2e*. New York, Springer-Verlag, 1989.

9. R. B. Ash. *Real Analysis and Probability.* New York, Academic Press, 1972.

10. O. E. Barndorff-Nielsen, D. R. Cox, and N. Reid. “The Role of Differential Geometry in Statistical Theory,” *International Statistical Review,* Vol. 54, No. 1 (1986), pp 83-96.

11. A. Basu, I. R. Harris and S. Basu. “Minimum Distance Estimatior: The Approach Using Density-based Distances,” in G. S. Maddala and C. R. Rao, *Robust Inference*. Amsterdam, Elsevier, 1997.

12. R. Beran. “Minimum Hellinger Distance Estimates for Parametric Models,” *The Annals of Statistics,* Vol. 5, No. 3 (May 1977), pp 445-463.

13. R. D. Cook and S. Weisberg. *Residuals and Influence in Regression*. New York, Chapman and Hall, 1982.

14. L. F. Fernholz. *von Mises Calculus for Statistical Functionals.* New York, Springer-Verlag, 1983.

15. R. A. Fisher. "Theory of Statistical Estimation," *Proceedings of the Cambridge Philosophical Society,* Vol. 22 (1925), pp. 700-725.

16. G.E. Forsythe, M. A. Malcolm, C. B. Moler. *Computer Methods for Mathematical Computations*. Englewood Cliffs, Prentice-Hall, 1977.

17. F. R. Hampel. *Contributions to the theory of robust estimation.* PhD. Thesis. University of California, Berkeley, 1968, as cited in (Hampel, *et al*., 1986).

18. F. R. Hampel. "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, Vol. 69 (1974), pp 383-393.

19. F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. New York, Wiley, 1986.

20. T. P. Hettmansperger, I. Hueter and J. Husler. "Minimum Distance Estimators," *Journal of Statistical Planning and Inference,* Vol. 41, No. 3 (1994), pp 291-302.

21. D. Hoaglin, F. Mosteller, and J. W. Tukey (eds.). *Understanding Robust and Exploratory Data Analysis*. New York, Wiley, 1983.

22. P. J. Huber. "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics,* Vol. 35, No. 1 (Mar. 1964), pp. 73-101.

23. P. J. Huber. *Robust Statistics.* New York, Wiley, 1981.

24. A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* 1933. Published in English as *Foundations of Probability*. New York, Chelsea, 1950. Available electronically at http://www.kolmogorov.com.

25. A. N. Kolmogorov and S. V. Fomin. *Introductory Real Analysis*. Translated to English by R. B. Silverman. New York, Dover, 1975.

26. E. Kreyszig. *Advanced Engineering Mathematics/9e.* New York, John Wiley and Sons, 2006.

27. A. Leon-Garcia. *Probability and Random Processes for Electrical Engineering/2e*. Reading, Addison Wesley Longman, 1994.

28. G. Li. "Robust Regression," in D. C. Hoaglin, F. Mosteller and J. W. Tukey (eds.), *Exploring Data Tables, Trends, and Shapes*. New York, John Wiley and Sons, 1985.

29. L. Mili and C. W. Coakley. “Robust Estimation in Structured Linear Regression,” *The Annals of Statistics,* Vol. 24, No. 6 (Dec. 1996), pp 2593-2607.

30. L. Mili, V. Phaniraj and P. J. Rousseeuw. “Least Median of Squares Estimation in Power Systems (with discussions),” *IEEE Transactions on Power Systems,* Vol. 6, No. 2 (May 1991), pp511-523.

31. C. W. Misner, K. S. Thorne, J. A. Wheeler. *Gravitation*. San Francisco, W. H. Freeman, 1973.

32. D. C. Montgomery and E. A. Peck. *Introduction to Linear Regression Analysis*. New York, John Wiley and Sons, 1982.

33. F. Mosteller. “On Some Useful “Inefficient” Statistics,” *Annals of Mathematical Statistics,* Vol. 17, No. 4 (Dec. 1946), pp. 377-408.

34. M. K. Murray and J. W. Rice. *Differential Geometry and Statistics*. Boca Raton, Chapman Hall, 1993.

35. D. Pollard. *A User’s Guide to Measure Theoretic Probability*. Cambridge, Cambridge University Press, 2002.

36. W.H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing/2e*. New York, Cambridge University Press, 1992.

37. F. Reif. *Fundamentals of Statistical and Thermal Physics*. New York, McGraw-Hill, 1965.

38. P. J. Rousseeuw. “Least Median of Squares Regression,” *Journal of the American Statistical Association*, Vol. 79, No. 388 (Dec. 1984), pp. 871-880.

39. P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. New York, John Wiley and Sons, 1987.

40. B. Schutz. *Geometric Methods of Mathematical Physics.* Cambridge, Cambridge University Press, 1980.

41. F. C. Schweppe and D. B. Rom. “Power System Static-State Estimation, Part II: Approximate Model,” *IEEE Transactions on Power Apparatus and Systems*, Vol. PAS-89, No. 1 (Jan. 1970), pp125-130.

42. D. W. Scott. “On Optimal and Data-Based Histograms,” *Biometrika,* Vol. 66, No. 3 (Dec. 1979), pp. 605-610.

43. D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* New York, Wiley, 1992.

44. G. A. F. Seber and A. J. Lee. *Linear Regression Analysis/2e*. New York, John Wiley and Sons, 2003.

45. A. F. Siegel. "Robust Regression Using Repeated Medians," *Biometrika,* Vol. 69, No. 1 (Apr. 1982), pp. 242-244.

46. C. G. Small. "A Survey of Multidimensional Medians," *International Statistical Review*, Vol. 58, No. 3 (Dec. 1990), pp. 263-277.

47. R. N. Tamura and D. D. Boos. "Minimum Hellinger Distance Estimation for Multivariate Location and Covariance," *Journal of the American Statistical Association,* Vol. 81, No. 393 (Mar. 1986), pp 223-229.

48. A. W. van der Vaert. *Asymptotic Statistics.* Cambridge, Cambridge University Press, 1998.

49. S. Weisberg. *Applied Linear Regression/2e*. New York, Wiley and Sons, 1985.

50. R. R. Wilcox. *New Statistical Procedures for the Social Sciences: Modern Solutions to Basic Problems*. Hillsdale, Lawrence Erlbaum Associates, 1987.

51. J. Wolfowitz. "Estimation by the minimum distance method," *Annals of the Institute of Statistical Mathematics* (Japan), Vol. 5 (1953), pp 9-23, as cited in (Wolfowitz, 1957).

52. J. Wolfowitz. "Estimation by the Minimum Distance Method in Non-Parametric Stochastic Difference Equations," *Annals of Mathematical Statistics,* Vol. 25 (1954), pp 203-217.

53. J. Wolfowitz. "The Minimum Distance Method," *Annals of Mathematical Statistics,* Vol. 28, No. 1 (Mar. 1957), pp 75-88.