Rethinking Biased Estimation: Improving Maximum Likelihood and the Cramér–Rao Bound
Rethinking Biased Estimation: Improving Maximum Likelihood and the Cramér–Rao Bound

Yonina C. Eldar

Department of Electrical Engineering
Technion — Israel Institute of Technology
Haifa 32000
Israel
yonina@ee.technion.ac.il

Full text available at: http://dx.doi.org/10.1561/2000000008
Foundations and Trends® in
Signal Processing
Volume 1 Issue 4, 2007
Editorial Board

Editor-in-Chief:
Robert M. Gray
Dept of Electrical Engineering
Stanford University
350 Serra Mall
Stanford, CA 94305
USA
rmgray@stanford.edu

Editors
Abeer Alwan (UCLA)  Jelena Kovacevic (CMU)
John Apostolopoulos (HP Labs)  B.S. Manjunath (UCSB)
Pamela Cosman (UCSD)  Urbashi Mitra (USC)
Michelle Effros (California Institute of Technology)  Thrasos Pappas (Northwestern University)
Yonina Eldar (Technion)  Mihaela van der Shaar (UCLA)
Yariv Ephraim (George Mason University)  Luis Torres (Technical University of Catalonia)
Sadaoki Furui (Tokyo Institute of Technology)  Michael Unser (EPFL)
Vivek Goyal (MIT)  P.P. Vaidyanathan (California Institute of Technology)
Sinan Gunturk (Courant Institute)  Rabab Ward (University of British Columbia)
Christine Guillemot (IRISA)  Susie Wee (HP Labs)
Shelia Hemami (Cornell)  Clifford J. Weinstein (MIT Lincoln Laboratories)
Lina Karam (Arizona State University)  Min Wu (University of Maryland)
Nick Kingsbury (Cambridge University)  Josiane Zerubia (INRIA)
Alex Kot (Nanyang Technical University)
Editorial Scope

Foundations and Trends® in Signal Processing will publish survey and tutorial articles on the foundations, algorithms, methods, and applications of signal processing including the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital and multirate signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multiresolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing
- Classification and detection
- Estimation and regression
- Tree-structured methods

Information for Librarians
Foundations and Trends® in Signal Processing, 2007, Volume 1, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.
Rethinking Biased Estimation: Improving Maximum Likelihood and the Cramér–Rao Bound

Yonina C. Eldar

Abstract

One of the prime goals of statistical estimation theory is the development of performance bounds when estimating parameters of interest in a given model, as well as constructing estimators that achieve these limits. When the parameters to be estimated are deterministic, a popular approach is to bound the mean-squared error (MSE) achievable within the class of unbiased estimators. Although it is well-known that lower MSE can be obtained by allowing for a bias, in applications it is typically unclear how to choose an appropriate bias.

In this survey we introduce MSE bounds that are lower than the unbiased Cramér–Rao bound (CRB) for all values of the unknowns. We then present a general framework for constructing biased estimators with smaller MSE than the standard maximum-likelihood (ML) approach, regardless of the true unknown values. Specializing the results to the linear Gaussian model, we derive a class of estimators that dominate least-squares in terms of MSE. We also introduce methods for choosing regularization parameters in penalized ML estimators that outperform standard techniques such as cross validation.
Contents

1 Introduction 1
   1.1 Estimation Model 5
   1.2 Minimum Variance Unbiased Estimation 7
   1.3 Maximum Likelihood Estimation 9
   1.4 Outline and Goals 12

2 The Cramér–Rao Bound and Extensions 15
   2.1 Cramér–Rao Bound (CRB) 15
   2.2 Bias Gradient Matrix 18
   2.3 Uniform Cramér–Rao Bound (UCRB) 21
   2.4 Achieving the UCRB 23

3 Mean-Squared Error Bounds 31
   3.1 MSE Bound 31
   3.2 Minimal MSE Bound with Linear Bias 35
   3.3 Dominating the CRB with Linear Bias 39
   3.4 Quadratic Inverse Fisher Information 44
   3.5 Affine Bias 56
   3.6 Application: SNR Estimation 60

4 Minimax and Blind Minimax Estimation 65
   4.1 Minimax MSE Estimation 66
4.2 Stein-Type Estimates
4.3 Blind Minimax Estimation
4.4 Relation to Stein-type Estimation
4.5 Numerical Results

5 The SURE Principle
5.1 MSE Estimation
5.2 Generalized SURE Principle
5.3 Application to Regularization Selection

6 Bounded Error Estimation
6.1 The Chebyshev Center
6.2 The Relaxed Chebyshev Center
6.3 Special Cases
6.4 Statistical Analysis

Acknowledgments

Notations and Acronyms

A Convex Optimization Methods
A.1 Convex Sets, Functions, and Problems
A.2 Duality Theory
A.3 Semidefinite Programs
A.4 Minimax Problems

References
The problem of estimating a set of unknown deterministic parameters is ubiquitous in a vast variety of areas in science and engineering including, for example, communication, economics, signal processing, seismology, and control. Many engineering systems rely on estimation theory to extract required information by estimating values of unknown parameters. Statisticians use parameter estimation techniques to extract and infer scientific, medical, and social conclusions from numerical data which are subject to random uncertainties.

Parameter estimation has a rich history dating back to Gauss and Legendre who used the least-squares (LS) method to predict movements of planets [62, 63, 97]. Mathematically, in an estimation problem, we are given a set of observations $x$ which we assume depend on an unknown parameter vector $\theta_0$. In this survey, we treat the setting in which $\theta_0$ is an unknown deterministic vector, i.e., the classical estimation setting as opposed to Bayesian inference. The problem then is to infer $\theta_0$ from the data using an estimate $\hat{\theta}$ which is a function of $x$, and to gain insight into the theoretical effects of the parameters on the system output.

One of the prime goals of statistical estimation theory is the development of bounds on the best achievable performance in inferring
parameters of interest in a given model, as well as determining estimators that achieve these limits. Such bounds provide benchmarks against which we can compare the performance of any proposed estimator, and insight into the fundamental limitations of the problem.

A classic performance benchmark is the Cramér–Rao bound (CRB) [27, 28, 30, 60, 119, 120], which characterizes the smallest achievable total variance of any unbiased estimator of $\theta_0$. Although other variance bounds exist in the literature, the CRB is relatively easy to determine, and can often be achieved by the maximum likelihood (ML) method [100, 120]. Despite its popularity, the CRB limits only the variance of unbiased estimators. However, in some problems, restricting attention to unbiased approaches leads to unreasonable solutions, that may, for example, be independent of the problem parameters [71, 98]. More importantly, in many cases the variance can be made smaller at the expense of increasing the bias, while ensuring that the overall estimation error is reduced. Therefore, even though unbiasedness may be appealing intuitively, it does not necessarily lead to a small estimation error $\hat{\theta} - \theta_0$ [34]. Consequently, the design of estimators is typically subject to a tradeoff between variance and bias [50, 58, 81, 107, 136].

In this survey, we discuss methods to improve the accuracy of unbiased estimators used in many signal processing problems. At the heart of the proposed methodology is the use of the mean-squared error (MSE) as the performance criteria. The MSE is the average of the squared-norm error $\|\hat{\theta} - \theta_0\|^2$, and is equal to the sum of the variance and the squared-norm of the bias. In an estimation context, where our prime concern is inferring $\theta_0$, the MSE (or weighted MSE) provides a direct measure of the relevant performance. Although herein we focus on the MSE, the essential ideas can be easily generalized to include weighted MSE criteria which measure the average weighted squared-norm error [48].

The approach we present is based on introducing a bias as a means of reducing the MSE. Biased estimation strategies are used extensively in a variety of different signal processing applications, such as image restoration [31, 108] where the bias corresponds to spatial resolution, smoothing techniques in time series analysis [115, 137],...
spectrum estimation \cite{131}, wavelet denoising \cite{33}, and diagonal loading in beamforming applications \cite{21, 26, 56}. Despite the fact that biasing as a method for improving performance is a mainstream approach, very often the choice of bias is rather \textit{ad-hoc}. In particular, although the biased algorithms mentioned above will improve the performance for certain choices of $\theta_0$, they can in fact deteriorate the MSE for other parameter values. Thus, in general, conventional biasing methods are not guaranteed to dominate ML, i.e., do not necessarily have lower MSE for all choices of $\theta_0$. Furthermore, many of these techniques include regularization parameters which are typically chosen by optimizing a data-error measure, i.e., an objective that depends on the estimated data $\hat{x}$ obtained by replacing $\theta_0$ by $\hat{\theta}$ in the model equations. Here, we focus on biasing in a way that is guaranteed to improve the MSE for all parameter values. This is achieved by using objectives that are directly related to the estimation error and are not data-error driven.

In their seminal work, Stein and James showed that for the independent, identically-distributed (iid) linear Gaussian model, it is possible to construct a nonlinear estimate of $\theta_0$ with lower MSE than that of ML for all values of the unknowns \cite{88, 128}. Such a strategy is said to dominate ML. In general an estimator $\hat{\theta}_1$ dominates a different estimator $\hat{\theta}_2$ if its MSE is no larger than that of $\hat{\theta}_2$ for all feasible $\theta_0$, and is strictly smaller for at least one choice of $\theta_0$; an estimator is admissible if it is not dominated by any other approach. Stein’s landmark idea has since been extended in many different directions and has inspired the work on ML-dominating methods which is the focus of this survey.

Here we go beyond the iid Gaussian model, and address a broad variety of estimation problems within an unified, systematic framework. To characterize the best possible bias-variance tradeoff in a general setting we would like to obtain a bound on the smallest achievable MSE in a given estimation problem. However, since $\theta_0$ is deterministic, the MSE will in general depend on $\theta_0$ itself. Therefore, the MSE cannot be used as a design criterion for choosing an optimal bias. Indeed, the point-wise minimum of the MSE is given by the trivial zero bound, which can be achieved with $\hat{\theta} = \theta_0$.

To overcome this obstacle, instead of attempting to minimize the MSE over all possible estimators, which includes the trivial solution.
\( \hat{\theta} = \theta_0 \), we restrict attention to methods that lie in a suitable class; the CRB is an example where we consider only methods with zero bias. Allowing for a broader set of bias vectors will result in MSE bounds that are lower than the CRB for all values of \( \theta_0 \). Furthermore, as part of the proposed framework we introduce explicit methods that achieve these lower bounds resulting in estimators with performance superior to unbiased approaches. In cases where the ML is efficient, namely it achieves the CRB, this methodology guarantees the existence of estimators that have lower MSE than ML for all values of \( \theta_0 \).

The strategy we outlined is based on first developing MSE performance bounds, and then designing estimators that achieve these limits, thus ensuring MSE improvement over existing unbiased solutions. An alternative technique to improve traditional estimates which is prevalent in the literature is the use of regularization, first systematically studied by Tikhonov \[135, 136\] and later extended to general estimation problems via the penalized ML (PML) approach \[65, 66\]. In general, regularization methods measure both the fit to the observed data and the physical plausibility of the estimate. Traditional applications of PML and regularization techniques have relied on data-error measures for selecting the regularization parameters \[17, 61, 64, 72, 73, 89, 110\].

As part of the proposed framework in this survey, we introduce methods for choosing the required regularization parameters based on measures of estimation error rather than data error. A popular design strategy in this spirit is to minimize Stein’s unbiased risk estimate (SURE) \[32, 122, 129, 130\], which is an unbiased estimate of the MSE. This method is appealing as it allows to directly approximate the MSE of an estimate from the data, without requiring knowledge of \( \theta_0 \). Besides leading to significant performance improvement over standard data-driven approaches in many practical problems, this technique can often be shown to dominate ML. In fact, the celebrated James–Stein estimate \[88, 128\], although originally derived based on different considerations, can be obtained from the SURE principle, as can many other ML-dominating approaches.

In most of the survey, we focus on problems in which the relationship between the data \( x \) and the unknown parameters \( \theta_0 \) is given by
1.1 Estimation Model

a statistical model. In the last section, we depart from this framework and discuss methods for bounded error estimation in which the statistical model is replaced by the assumption that $\theta_0$ is restricted to some deterministic set, defined by prior constraints. The link to the rest of the survey is that in this context as well, we can replace traditional data-error strategies by methods that are inherently based on the error between the estimate $\hat{\theta}$ and the true parameter $\theta_0$. Although this approach is deterministic in nature, it can also be used in a statistical setting where the constraints are dictated by the underlying statistical properties. For example, given measurements $x = \theta_0 + w$, where $w \in \mathbb{R}^n$ is a zero-mean random vector with covariance $\sigma^2 I$, we can assume that $\theta_0$ lies in the constraint set $\|x - \theta_0\|^2 \leq n\sigma^2$. Despite the fact that this restriction is not always satisfied, using it in conjunction with the proposed estimation strategy leads to an estimate that dominates the constrained ML solution. Therefore, this approach can also be used to develop MSE-dominating techniques when a statistical model exists.

Our focus here is on static models. In recent years, there has been increasing interest in inference techniques and performance bounds for dynamical systems [134]. We believe that the essential ideas introduced can be extended to the dynamical setting as well.

1.1 Estimation Model

Throughout the survey, our goal is to estimate a deterministic parameter vector $\theta_0$ from measurements $x$. For concreteness, we assume that $\theta_0$ is a real length-$m$ vector, and $x$ is a real length-$n$ vector. However, all the results are valid for the complex case as well with obvious modifications. The relationship between $x$ and $\theta_0$ is described by the probability density function (pdf) $p(x; \theta_0)$ of $x$ characterized by $\theta_0$. We emphasize that $\theta_0$ is a deterministic unknown vector, so that no Bayesian prior is assumed on $\theta_0$. Consequently, $p(x; \theta_0)$ is not a joint pdf, but rather a pdf of $x$ in which $\theta_0$ figures as an unknown parameter. As we will see throughout the survey, this renders the problem considerably more challenging, but at the same time more intriguing than its Bayesian counterpart.
As an example, suppose we have a Bernoulli random variable \( x_i \) which takes on the value 1 with probability (w.p.) \( \theta_0 \) and 0 w.p. \( 1 - \theta_0 \). Our goal is to estimate \( \theta_0 \) from \( n \) iid measurements. Denoting by \( \mathbf{x} = (x_1, \ldots, x_n)^T \) the vector whose components are the measurements \( x_i \), the pdf of \( \mathbf{x} \) can be written as

\[
p(\mathbf{x}; \theta_0) = \theta_0^{\sum_{i=1}^n x_i} (1 - \theta_0)^{n - \sum_{i=1}^n x_i}.
\] (1.1)

Another important class of examples, which we will study in detail in Section 4, is the linear Gaussian model. In this case the unknown vector \( \theta_0 \in \mathbb{R}^m \) is related to \( \mathbf{x} \in \mathbb{R}^n \) through the linear model:

\[
\mathbf{x} = \mathbf{H} \theta_0 + \mathbf{w}.
\] (1.2)

Here \( \mathbf{H} \) is a known \( n \times m \) model matrix with full column-rank, and \( \mathbf{w} \) is a zero-mean Gaussian random vector with covariance matrix \( \mathbf{C} \), which for simplicity is assumed to be positive definite. For the model (1.2), the pdf of \( \mathbf{x} \) is

\[
p(\mathbf{x}; \theta_0) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{C}|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{H} \theta_0)^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{H} \theta_0) \right\}.
\] (1.3)

Although we assume that \( \mathbf{H} \) is known in the model (1.2), similar ideas to those developed here can be used when \( \mathbf{H} \) is subject to deterministic or random uncertainty [8, 44, 51, 56, 144, 145].

A broader class of pdfs which includes (1.3) is the exponential family of distributions which can expressed in the form:

\[
f(\mathbf{x}; \theta_0) = r(\mathbf{x}) \exp \{ \theta_0^T \phi(\mathbf{x}) - g(\theta_0) \},
\] (1.4)

where \( r(\mathbf{x}) \) and \( \phi(\mathbf{x}) \) are functions of the data only, and \( g(\theta_0) \) depends on the unknown parameter \( \theta_0 \). Exponential pdfs play an important role in statistics due to the Pitman–Koopman–Darmois theorem [29, 94, 117], which states that among distributions whose domain does not vary with the parameter being estimated, a sufficient statistic with bounded dimension as the sample size increases can be found only in exponential families [100]. Furthermore, efficient estimators achieving the CRB exist only when the underlying model is exponential. Many known distributions are of the exponential form, such as Gaussian,
1.2 Minimum Variance Unbiased Estimation

Given data $x$ and a model $p(x; \theta_0)$ a pervasive inference strategy in signal processing applications is to seek a minimum variance unbiased (MVU) estimate of $\theta_0$. This is typically accomplished by using the theory of sufficient statistics or the attainment of the CRB [93]. Although an MVU solution is not guaranteed to exist, in many problems of interest such an estimate can be found, at least asymptotically. The constraint of unbiasedness is often a practical one, since in many cases the variance, or the MSE, can be minimized over this class using functions of the data that are truly estimators, i.e., the statistic does not depend on the unknown parameter. However, there are several severe limitations of unbiased methods.

First, unbiased estimators are not always guaranteed to exist. An example is when inferring the odds ratio $p = \theta_0 / (1 - \theta_0)$ from $n$ Bernoulli trials. It can be shown that there is no unbiased estimate for $p$ [124, Sec. 7.12]. On the other hand, there exist many reasonable approximations such as $p = \hat{\theta} / (1 - \hat{\theta})$, where $\hat{\theta} = (1/n) \sum_{i=1}^{n} x_i$.

Second, the unbiasedness requirement can sometimes produce nonsensical results. As an example, consider the problem of estimating the probability of success $\theta_0$ in a set of Bernoulli trials, from the number of experiments $x$ until success [25]. The pdf of $x$ is given by

$$p(x; \theta_0) = \theta_0 (1 - \theta_0)^{x-1}, \quad x = 1, 2, \ldots$$  \hspace{1cm} (1.5)

The only unbiased estimate for this problem, and hence the MVU solution, is

$$\hat{\theta}_0 = \begin{cases} 1, & x = 1; \\ 0, & \text{otherwise}. \end{cases}$$  \hspace{1cm} (1.6)

Clearly this is an unreasonable estimate of $\theta_0$. A more appealing choice is $\theta = 1/x$. 

gamma, chi-square, beta, Dirichlet, Bernoulli, binomial, multinomial, Poisson, and geometric distributions. Exponential families will play an important role in Section 5 in the context of estimation based on the SURE criterion.
As another example, suppose that \( x \) is a Poisson random variable with mean \( \theta_0 > 0 \), and we would like to estimate \( p = \exp\{-2\theta_0\} \), which is the probability that no events occur in two units of time. Clearly the true value of \( p \) satisfies \( p \in (0, 1) \). However, the only unbiased estimate is given by

\[
\hat{p} = \begin{cases} 
1, & x \text{ even;} \\
-1, & x \text{ odd},
\end{cases}
\]

(1.7)

which always falls outside the range \((0, 1)\), and is extremely unreasonable [99] [132, Exercise 17.26] [124, Sec. 7.16]. A somewhat more complex example, in which the only unbiased estimator always ends up considerably outside the problem bounds, can be found in [77].

Finally, the most important objection to the constraint of unbiasedness is that it produces estimators \( \hat{\theta} \) whose optimality is based on the error between \( \hat{\theta} \) and the average value, not \( \hat{\theta} \) and the true value as measured by the MSE. It is the latter that is actually of prime importance in an estimation context as it is a direct measure of estimation error. Specifically, the MSE is defined by

\[
E\{\|\hat{\theta} - \theta_0\|^2\} = \int \|\hat{\theta} - \theta_0\|^2 f(x; \theta_0) dx = \|b(\theta_0)\|^2 + v(\theta_0),
\]

(1.8)

where \( b(\theta_0) = E\{\hat{\theta}\} - \theta_0 \) is the bias of the estimate, and \( v(\theta_0) = E\{|\hat{\theta} - E\{\hat{\theta}\}|^2\} \) is its variance. Note that the MSE depends explicitly on \( \theta_0 \). An MVU method minimizes the MSE only over a constrained class for which \( b(\theta_0) = 0 \) for all \( \theta_0 \). Thus, even in problems in which the MVU approach leads to reasonable estimates, the MSE performance may still be improved using a biased technique.

The difficulty in using the MSE as a design objective is that in general it depends explicitly on \( \theta_0 \). This parameter dependency also renders comparison between different estimators a difficult (and often impossible) task. Indeed, one method may be better than another for some values of \( \theta_0 \), and worse for others. For instance, the trivial estimator \( \hat{\theta} = 0 \) achieves optimal MSE when \( \theta_0 = 0 \), but its performance is otherwise poor. Nonetheless, it is possible to impose a partial order among inference techniques [100] using the concepts of domination and admissibility. An estimator \( \hat{\theta}_1 \) dominates an estimator \( \hat{\theta}_2 \) on a given set...
The estimator $\hat{\theta}_1$ strictly dominates $\hat{\theta}_2$ on $\mathcal{U}$ if

$$E\{\|\hat{\theta}_1 - \theta\|^2\} < E\{\|\hat{\theta}_2 - \theta\|^2\}, \text{ for all } \theta \in \mathcal{U}. \quad (1.10)$$

If $\hat{\theta}_1$ dominates $\hat{\theta}_2$ then clearly it is better in terms of MSE. An estimator $\hat{\theta}$ is admissible if it is not dominated by any other method. If an estimator is inadmissible, then there exists another approach whose MSE is no larger than the given method for all $\theta$ in $\mathcal{U}$, and is strictly smaller for some $\theta$ in $\mathcal{U}$.

The study of admissibility is sometimes restricted to linear methods. A linear admissible estimator is one which is not dominated by any other linear strategy. The class of linear admissible techniques can be characterized by a simple rule [24, 43, 83, 121], and given any linear inadmissible estimator, it is possible to construct a linear admissible alternative which dominates it by using convex analysis tools [43]. However, the problem of admissibility is considerably more intricate when the linearity restriction is removed; generally, admissible estimators are either trivial (e.g., $\hat{\theta} = 0$) or exceedingly complex [105]. As a result, much research has focused on finding simple nonlinear techniques that dominate ML.

### 1.3 Maximum Likelihood Estimation

One of the most popular estimation strategies is the ML method in which the estimate $\hat{\theta}$ is chosen to maximize the likelihood of the observations:

$$\hat{\theta} = \arg \max_{\theta} p(x; \theta). \quad (1.11)$$

This approach was pioneered by Fisher between 1912 and 1922 [1, 59] and has widespread applications in various fields. The ML estimator enjoys several appealing properties, including asymptotic efficiency under suitable regularity conditions. Thus, asymptotically, and in many
non-asymptotic cases, the ML approach is MVU optimal. Nonetheless, its MSE can be improved upon in the non-asymptotic regime in many different settings.

As is evident from (1.11) the ML technique is data driven, meaning the quality of the estimator is determined by how well it describes the observations. However, the ML objective is not related to the MSE which is a direct measure of estimation error. This distinction is clearly seen when considering the linear Gaussian model (1.2). In this case the ML criterion coincides with the weighted LS objective:

$$\arg \max_{\theta} p(x; \theta) = \arg \min_{\theta} (x - H\theta)^T C^{-1} (x - H\theta).$$ (1.12)

Evidently, the ML solution is designed to minimize the error between the given data and the estimated data $\hat{x} = H\hat{\theta}$. Assuming $H$ has full column-rank, the resulting LS estimate is given by

$$\hat{\theta}_{LS} = (H^T C^{-1} H)^{-1} H^T C^{-1} x.$$ (1.13)

It is well known that $\hat{\theta}_{LS}$ is also MVU optimal for Gaussian noise [93].

To illustrate the fact that minimizing data error does not necessarily imply a small estimation error, in Figure 1.1 we consider an example of the model (1.2) in which $\theta_0$ represents the 2D signal in Figure 1.1(a). Our goal is to recover this image from the observation $x$ of Figure 1.1(b) which is obtained after shifting and blurring with a Gaussian kernel, and corruption by additive Gaussian noise. We assume that the distortion and noise variance are known. Using the LS estimate results in the image in Figure 1.1(c) in which the original signal

![Fig. 1.1 Image recovery using least-squares (LS) and a biased minimax estimate. (a) original 2D signal. (b) Corrupted image. (c) Recovery using LS. (d) Recovery based on a minimax strategy.](http://dx.doi.org/10.1561/2000000008)
is completely destroyed. On the other hand, using a minimax estimate, which we will discuss in Section 4, we obtain a pretty good recovery of the signal, as can be seen in Figure 1.1(d). Clearly the fact that the data error is smaller in Figure 1.1(c) is not sufficient to guarantee good signal recovery.

As another example, consider estimating a signal \( \theta_0(t) \) that is observed through the heat integral equation and corrupted by additive noise. The true and observed signals are shown in Figures 1.2(a) and 1.2(b), respectively. In Figure 1.2(c) we compare the estimated signal using LS and a bounded-error approach (RCC) based on controlling the minimax estimation error, which we present in Section 6. Evidently, the latter strategy, referred to as the Chebyshev center, leads to substantial performance improvement.

![Fig. 1.2 Signal recovery using least-squares (CLS) and the Chebyshev estimate (RCC).](http://dx.doi.org/10.1561/2000000008)
These examples illustrate that minimizing data error does not necessarily imply a small estimation error. From a statistical perspective, MVU methods do not guarantee satisfactory estimation performance, even when they exist and lead to reasonable strategies.

1.4 Outline and Goals

Stein’s discovery of ML-dominating techniques in the linear Gaussian model, half a century ago, shocked the statistics community. Since then many other examples of ML improvement have been discovered and analyzed. In this survey, we present a broad framework for constructing ML-dominating solutions in a broad variety of estimation problems. More specifically, we present general tools for reducing MSE by introducing a bias. An important aspect of the proposed approach is that the reduction in MSE is guaranteed for all choices of the unknown parameter vector. The methods we outline for constructing estimators are designed to explicitly optimize an objective based on estimation error rather than data error. The performance advantage of the algorithms we present is greatest in difficult problems, i.e., short data records or lower signal-to-noise ratios (SNRs). Applications include the design of estimation algorithms for sonar, radar, and communications, as well as a myriad of other disciplines that rely heavily on precise measurement of parameters.

It is our hope that this framework will provide additional support for ML dominating methods, both by supplying an intuitive understanding of this phenomenon, and by providing a wide class of powerful new estimators.

1.4.1 Outline

In Section 2 we begin by reviewing the standard unbiased CRB and then discuss extensions to biased estimation. In particular, we introduce the uniform CRB which provides a benchmark on the variance of any biased estimator with bias-gradient matrix whose norm is limited by a constant. This bound is asymptotically achieved by the PML method with a suitable regularization function. The uniform CRB is useful in problems in which the bias gradient norm has a physical interpretation;
this is the case in some imaging applications where the norm is related to image resolution [80, 108]. Furthermore, it requires the specification of only one parameter (the norm bound) rather than the entire bias gradient matrix, as in the standard biased CRB [130].

In Section 3, we study MSE bounds which directly limit the estimation error. These bounds depend on the unknown parameter vector \( \theta_0 \), as well as on the bias of the estimate \( \hat{\theta} \). In order to optimize the bound we first consider the class of estimates with linear bias vectors, and seek the member from this set that minimizes the bound. A nice aspect of this approach is that once an optimal bias of this form is found, it can be used to construct a linear modification of the ML estimate that dominates ML whenever the latter is efficient. We demonstrate this methodology through several examples which illustrate how scaling can be used to reduce the MSE. As we show, it is often possible to improve the MSE for all \( \theta_0 \) using a linear modification, without any prior knowledge on the true parameter values. This linear scaling is chosen as a solution to a minimax optimization problem.

Building on the linear results, in Section 4, we present the blind minimax technique which leads to nonlinear modifications of the ML solution. The approach is illustrated in the context of the linear Gaussian model and makes use of a two-stage process: first, a set is estimated from the measurements; next, a linear minimax method for this set is used to estimate the parameter itself. Surprisingly, the resulting estimate can be shown to dominate the ML solution even though no prior information is assumed. The blind minimax technique provides a framework whereby many different estimators can be generated, and provides insight into the mechanism by which these techniques out-perform ML. In particular, we show how the celebrated James–Stein estimate can be derived within this framework.

An alternative approach for deriving ML-dominating methods is to use the SURE principle. In Section 5, we introduce the SURE objective and illustrate how it can be applied to construct methods that have lower MSE than ML. The essential idea is to choose a class of estimates, and then select the member that minimizes the MSE estimate. We demonstrate, in particular, the use of the SURE design method for selecting regularization parameters in PML estimation.
Finally, in Section 6, we extend the estimation-error methodology to a deterministic setting. We treat estimation problems in which there are prior constraints on $\theta_0$, such as weighted norm restrictions or interval constraints on the individual components of $\theta_0$. The standard approach in such settings is constrained ML in which the likelihood is maximized subject to the given restrictions. Instead, we introduce the Chebyshev center estimator which is based on minimizing the worst-case estimation error $\|\hat{\theta} - \theta_0\|^2$ over all feasible solutions. As we show, this strategy can reduce the estimation error dramatically with respect to the constrained ML method. This design technique can also be used in a statistical setting by replacing the statistical model with an appropriate constraint on $\theta_0$. Even though this later restriction is not always satisfied in practice, the resulting estimate can be shown in some cases to dominate the constrained ML for the same problem setting.

The procedures we develop throughout the survey are based on convex optimization tools and minimax formulations. In the Appendix, we provide a brief overview of the basics of convex analysis, emphasizing the results needed in our presentation.
References

References


References

References


References

References


References


References


