Foundations and Trends® in
Signal Processing
Vol. 1, No. 4 (2007) 305–449

DOI: 10.1561/2000000008

now
the essence of knowledge

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

Yonina C. Eldar[1]

[1] *Department of Electrical Engineering, Technion — Israel Institute of Technology, Haifa 32000, Israel, yonina@ee.technion.ac.il*

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.

1

Introduction

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 $\mathbf{x}$ which we assume depend on an unknown parameter vector $\boldsymbol{\theta}_0$. In this survey, we treat the setting in which $\boldsymbol{\theta}_0$ is an unknown deterministic vector, i.e., the classical estimation setting as opposed to Bayesian inference. The problem then is to infer $\boldsymbol{\theta}_0$ from the data using an estimate $\hat{\boldsymbol{\theta}}$ which is a function of $\mathbf{x}$, 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 $\boldsymbol{\theta}_0$. 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 $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ [34]. Consequently, the design of estimators is typically subject to a tradeoff between variance and bias [50, 58, 81, 104, 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 $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2$, 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 $\boldsymbol{\theta}_0$, 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 [131], wavelet denoising [33], and diagonal loading in beamforming applications [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 *ad-hoc*. In particular, although the biased algorithms mentioned above will improve the performance for certain choices of $\boldsymbol{\theta}_0$, 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 $\boldsymbol{\theta}_0$. 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 $\hat{\mathbf{x}}$ obtained by replacing $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}}$ 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 $\boldsymbol{\theta}_0$ with lower MSE than that of ML for all values of the unknowns [88, 128]. Such a strategy is said to dominate ML. In general an estimator $\hat{\boldsymbol{\theta}}_1$ dominates a different estimator $\hat{\boldsymbol{\theta}}_2$ if its MSE is no larger than that of $\hat{\boldsymbol{\theta}}_2$ for all feasible $\boldsymbol{\theta}_0$, and is strictly smaller for at least one choice of $\boldsymbol{\theta}_0$; 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 $\boldsymbol{\theta}_0$ is deterministic, the MSE will in general depend on $\boldsymbol{\theta}_0$ 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 $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$.

To overcome this obstacle, instead of attempting to minimize the MSE over all possible estimators, which includes the trivial solution

$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$, 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 $\boldsymbol{\theta}_0$. 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 $\boldsymbol{\theta}_0$.

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 $\boldsymbol{\theta}_0$. 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 $\mathbf{x}$ and the unknown parameters $\boldsymbol{\theta}_0$ is given by

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 $\boldsymbol{\theta}_0$ 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 $\hat{\boldsymbol{\theta}}$ and the true parameter $\boldsymbol{\theta}_0$. 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 $\mathbf{x} = \boldsymbol{\theta}_0 + \mathbf{w}$, where $\mathbf{w} \in \mathbb{R}^n$ is a zero-mean random vector with covariance $\sigma^2\mathbf{I}$, we can assume that $\boldsymbol{\theta}_0$ lies in the constraint set $\|\mathbf{x} - \boldsymbol{\theta}_0\|^2 \leq n\sigma^2$. 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 $\boldsymbol{\theta}_0$ from measurements $\mathbf{x}$. For concreteness, we assume that $\boldsymbol{\theta}_0$ is a real length-m vector, and $\mathbf{x}$ is a real length-n vector. However, all the results are valid for the complex case as well with obvious modifications. The relationship between $\mathbf{x}$ and $\boldsymbol{\theta}_0$ is described by the probability density function (pdf) $p(\mathbf{x};\boldsymbol{\theta}_0)$ of $\mathbf{x}$ characterized by $\boldsymbol{\theta}_0$. We emphasize that $\boldsymbol{\theta}_0$ is a deterministic unknown vector, so that no Bayesian prior is assumed on $\boldsymbol{\theta}_0$. Consequently, $p(\mathbf{x};\boldsymbol{\theta}_0)$ is not a joint pdf, but rather a pdf of $\mathbf{x}$ in which $\boldsymbol{\theta}_0$ 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.) θ_0 and 0 w.p. $1-\theta_0$. Our goal is to estimate θ_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

$$p(\mathbf{x};\boldsymbol{\theta}_0) = \theta_0^{\sum_{i=1}^n x_i}(1-\theta_0)^{n-\sum_{i=1}^n x_i}. \tag{1.1}$$

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 $\boldsymbol{\theta}_0 \in \mathbb{R}^m$ is related to $\mathbf{x} \in \mathbb{R}^n$ through the linear model:

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta}_0 + \mathbf{w}. \tag{1.2}$$

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

$$p(\mathbf{x};\boldsymbol{\theta}_0) = \frac{1}{\sqrt{(2\pi)^n|\mathbf{C}|}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}_0)^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}_0)\right\}. \tag{1.3}$$

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:

$$f(\mathbf{x};\boldsymbol{\theta}_0) = r(\mathbf{x})\exp\{\boldsymbol{\theta}_0^T\phi(\mathbf{x}) - g(\boldsymbol{\theta}_0)\}, \tag{1.4}$$

where $r(\mathbf{x})$ and $\phi(\mathbf{x})$ are functions of the data only, and $g(\boldsymbol{\theta}_0)$ depends on the unknown parameter $\boldsymbol{\theta}_0$. 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,

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.

1.2 Minimum Variance Unbiased Estimation

Given data $\mathbf{x}$ and a model $p(\mathbf{x};\boldsymbol{\theta}_0)$ a pervasive inference strategy in signal processing applications is to seek a minimum variance unbiased (MVU) estimate of $\boldsymbol{\theta}_0$. 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 θ_0 in a set of Bernoulli trials, from the number of experiments x until success [25]. The pdf of x is given by

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

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

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

Clearly this is an unreasonable estimate of θ_0. A more appealing choice is $\hat{\theta} = 1/x$.

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

$$\hat{p} = \begin{cases} 1, & x \text{ even;} \\ -1, & x \text{ odd,} \end{cases} \tag{1.7}$$

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 $\hat{\boldsymbol{\theta}}$ whose optimality is based on the error between $\hat{\boldsymbol{\theta}}$ and the average value, not $\hat{\boldsymbol{\theta}}$ 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

$$E\left\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2\right\} = \int \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2 f(\mathbf{x};\boldsymbol{\theta}_0) d\mathbf{x} = \|\mathbf{b}(\boldsymbol{\theta}_0)\|^2 + v(\boldsymbol{\theta}_0), \tag{1.8}$$

where $\mathbf{b}(\boldsymbol{\theta}_0) = E\{\hat{\boldsymbol{\theta}}\} - \boldsymbol{\theta}_0$ is the bias of the estimate, and $v(\boldsymbol{\theta}_0) = E\{\|\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\}\|^2\}$ is its variance. Note that the MSE depends explicitly on $\boldsymbol{\theta}_0$. An MVU method minimizes the MSE only over a constrained class for which $\mathbf{b}(\boldsymbol{\theta}_0) = \mathbf{0}$ for all $\boldsymbol{\theta}_0$. 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 $\boldsymbol{\theta}_0$. 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 $\boldsymbol{\theta}_0$, and worse for others. For instance, the trivial estimator $\hat{\boldsymbol{\theta}} = \mathbf{0}$ achieves optimal MSE when $\boldsymbol{\theta}_0 = \mathbf{0}$, 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 $\hat{\boldsymbol{\theta}}_1$ dominates an estimator $\hat{\boldsymbol{\theta}}_2$ on a given set

$\mathcal{U}$ if

$$\begin{aligned} &E\{\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}\|^2\} \leq E\{\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}\|^2\}, \quad \text{for all } \boldsymbol{\theta} \in \mathcal{U}; \\ &E\{\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}\|^2\} < E\{\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}\|^2\}, \quad \text{for some } \boldsymbol{\theta} \in \mathcal{U}. \end{aligned} \tag{1.9}$$

The estimator $\hat{\boldsymbol{\theta}}_1$ strictly dominates $\hat{\boldsymbol{\theta}}_2$ on $\mathcal{U}$ if

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

If $\hat{\boldsymbol{\theta}}_1$ dominates $\hat{\boldsymbol{\theta}}_2$ then clearly it is better in terms of MSE. An estimator $\hat{\boldsymbol{\theta}}$ 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 $\boldsymbol{\theta}$ in $\mathcal{U}$, and is strictly smaller for some $\boldsymbol{\theta}$ 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., $\hat{\boldsymbol{\theta}} = \mathbf{0}$) 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 $\hat{\boldsymbol{\theta}}$ is chosen to maximize the likelihood of the observations:

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

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:

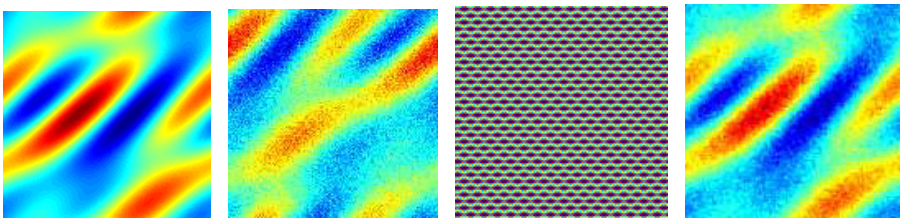
$$\arg\max_{\boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} (\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}). \tag{1.12}$$

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

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

It is well known that $\hat{\boldsymbol{\theta}}_{\rm LS}$ 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 $\boldsymbol{\theta}_0$ represents the 2D signal in Figure 1.1(a). Our goal is to recover this image from the observation $\mathbf{x}$ 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



(a) (b) (c) (d)

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.

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.

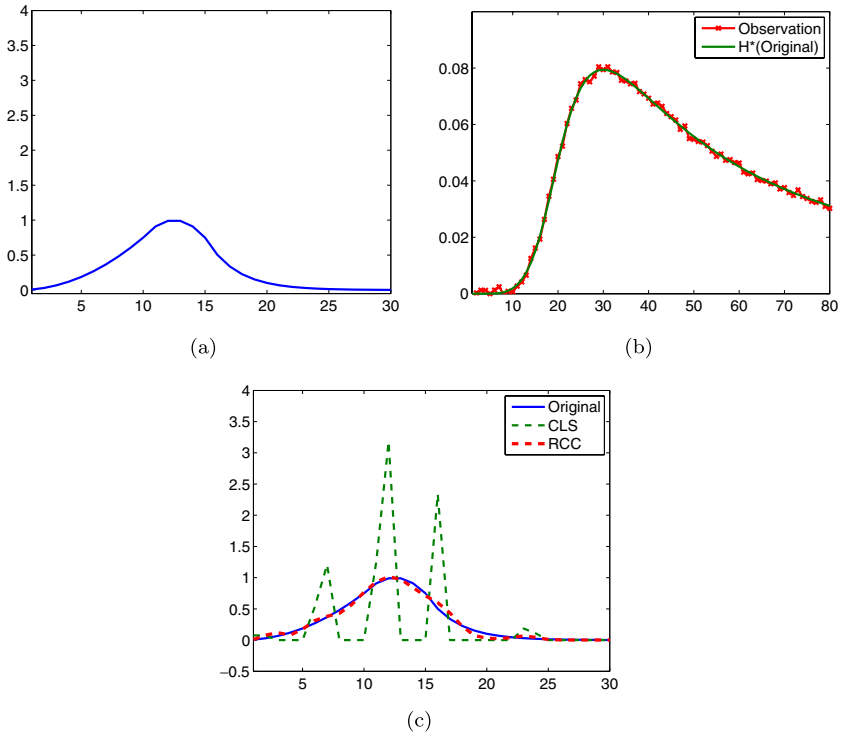


Fig. 1.2 Signal recovery using least-squares (CLS) and the Chebyshev estimate (RCC). (a) True signal. (b) Observed signal. (c) Recovery using CLS and RCC.

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 [140].

In Section 3, we study MSE bounds which directly limit the estimation error. These bounds depend on the unknown parameter vector $\boldsymbol{\theta}_0$, as well as on the bias of the estimate $\hat{\boldsymbol{\theta}}$. 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 $\boldsymbol{\theta}_0$ 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 outperform 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 $\boldsymbol{\theta}_0$, such as weighted norm restrictions or interval constraints on the individual components of $\boldsymbol{\theta}_0$. 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 $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2$ 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 $\boldsymbol{\theta}_0$. 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.

2

The Cramér–Rao Bound and Extensions

In this section we begin the search for good estimates of $\boldsymbol{\theta}_0$ by discussing bounds on the variance of different estimation strategies. Although several bounds appear in the literature, for concreteness we focus on the CRB since it is relatively easy to determine, and can often be achieved. In the next section we will see how these variance limits can be used to directly control the MSE rather than only the variability of the estimators. The ideas we develop can be applied to other performance benchmarks by following the same essential steps.

2.1 Cramér–Rao Bound (CRB)

2.1.1 Unbiased CRB

One of the primary approaches to recover $\boldsymbol{\theta}_0$ given the data $\mathbf{x}$ is to seek an MVU solution $\hat{\boldsymbol{\theta}}$. Any estimate $\hat{\boldsymbol{\theta}}$ (for which the MSE is defined) can be characterized by its bias

$$\mathbf{b}(\boldsymbol{\theta}_0) = E\{\hat{\boldsymbol{\theta}}\} - \boldsymbol{\theta}_0, \tag{2.1}$$

and covariance matrix

$$\mathbf{C}(\boldsymbol{\theta}_0) = E\{(\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\})(\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\})^T\}. \tag{2.2}$$

An unbiased estimate satisfies $\mathbf{b}(\boldsymbol{\theta}_0) = \mathbf{0}$ for all $\boldsymbol{\theta}_0$.

Under suitable regularity conditions on $p(\mathbf{x};\boldsymbol{\theta})$ (see e.g., [28, 119, 120]), the covariance of any *unbiased* estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ is bounded by the CRB which states that

$$\mathbf{C}(\boldsymbol{\theta}_0) \succeq \mathbf{J}^{-1}(\boldsymbol{\theta}_0). \tag{2.3}$$

Here $\mathbf{J}(\boldsymbol{\theta}_0)$ is the Fisher information matrix defined by

$$\mathbf{J}(\boldsymbol{\theta}_0) = E\left\{\left[\frac{d\log p(\mathbf{x};\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}}\right]^T\left[\frac{d\log p(\mathbf{x};\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}}\right]\right\}, \tag{2.4}$$

which depends in general on the true unknown parameter vector $\boldsymbol{\theta}_0$, and $\mathbf{A} \succeq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B} \succeq 0$, where $\mathbf{A} \succeq 0$ ($\mathbf{A} \succ 0$) denotes a symmetric and nonnegative (positive) definite matrix.

In order to simplify the derivations, we assume throughout the survey that $\mathbf{J}(\boldsymbol{\theta}_0)$ is invertible. The CRB was first published by Frechet [60] and later by Darmois [30], Cramér [27], and Rao [119]. Using (2.3) we can bound the total variance that is achievable using any unbiased technique, where the total variance is the sum of the variances in estimating the individual components of $\boldsymbol{\theta}_0$:

$$v(\boldsymbol{\theta}_0) = E\{\|\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\}\|^2\} = \sum_{i=1}^{m} E\{(\hat{\theta}_i - E\{\hat{\theta}_i\})^2\}. \tag{2.5}$$

From (2.2) and (2.3) we have immediately that

$$E\left\{\|\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\}\|^2\right\} = \mathrm{Tr}\left(\mathbf{C}(\boldsymbol{\theta}_0)\right) \geq \mathrm{Tr}\left(\mathbf{J}^{-1}(\boldsymbol{\theta}_0)\right). \tag{2.6}$$

An unbiased estimate achieving the CRB is called efficient. It can be shown that if an estimate is efficient, then it is necessarily ML optimal. Furthermore, under suitable regularity assumptions on $p(\mathbf{x};\boldsymbol{\theta}_0)$, the unique ML solution is asymptotically unbiased and achieves the CRB [100, 119, 120].

The importance of the CRB is that it allows the assessment of how close to optimality a given unbiased recovery method is. In particular, if the variance of an unbiased estimate $\hat{\boldsymbol{\theta}}$ is equal to the CRB, then it has minimal variance. Consequently, the ML approach is often MVU optimal.

As an example, consider the linear Gaussian model (1.2). The Fisher information for this problem can be readily computed and is equal to

$$\mathbf{J} = \mathbf{H}^T \mathbf{C}^{-1} \mathbf{H}. \tag{2.7}$$

Therefore the minimal attainable variance using any unbiased approach is $\mathrm{Tr}((\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1})$. To determine an MVU method, consider the ML estimate (1.13), which coincides with the LS solution. It is easy to see that $E\{\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\} = \boldsymbol{\theta}_0$ and

$$E\big\{\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}} - \boldsymbol{\theta}_0\|^2\big\} = \mathrm{Tr}\left(\mathbf{G}\mathbf{C}\mathbf{G}^T\right) = \mathrm{Tr}\left((\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1}\right), \tag{2.8}$$

where we defined $\mathbf{G} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}^{-1}$. Evidently, the ML estimate in this problem is MVU optimal.

2.1.2 Biased CRB

A simple modification of the CRB renders it applicable to biased estimates as well [140]. Specifically, let $\hat{\boldsymbol{\theta}}$ denote an arbitrary estimator of $\boldsymbol{\theta}_0$ with bias $\mathbf{b}(\boldsymbol{\theta}_0)$. Then its covariance must satisfy

$$\mathbf{C}(\boldsymbol{\theta}_0) \succeq (\mathbf{I} + \mathbf{D}(\boldsymbol{\theta}_0))\, \mathbf{J}^{-1}(\boldsymbol{\theta}_0)\, (\mathbf{I} + \mathbf{D}(\boldsymbol{\theta}_0))^T \stackrel{\triangle}{=} \mathbf{C}(\mathbf{D}), \tag{2.9}$$

where $\mathbf{D}(\boldsymbol{\theta}_0)$ is the bias gradient matrix defined by

$$\mathbf{D}(\boldsymbol{\theta}_0) = \frac{d\mathbf{b}(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}_0}. \tag{2.10}$$

Substituting $\mathbf{D}(\boldsymbol{\theta}_0) = \mathbf{0}$ in (2.9), the bound reduces to the unbiased CRB (2.3). Note that the biased CRB depends on the bias gradient and not the bias itself. This makes intuitive sense since any constant bias is removable, even if it is very large, and therefore should not affect the variance. From (2.9) it follows immediately that the total variance of any estimate with bias gradient matrix $\mathbf{D}(\boldsymbol{\theta}_0)$ is bounded below by

$$E\big\{\|\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\}\|^2\big\} \geq \mathrm{Tr}\left(\big(\mathbf{I} + \mathbf{D}(\boldsymbol{\theta}_0)\big)\mathbf{J}^{-1}(\boldsymbol{\theta}_0)\big(\mathbf{I} + \mathbf{D}(\boldsymbol{\theta}_0)\big)^T\right). \tag{2.11}$$

Continuing with the linear Gaussian model, suppose we now restrict attention to estimators with linear bias so that $\mathbf{b}(\boldsymbol{\theta}_0) = \mathbf{M}\boldsymbol{\theta}_0$ for some matrix $\mathbf{M}$. Then it is easy to see that the biased CRB is attained by

$$\hat{\boldsymbol{\theta}} = \mathbf{G}\mathbf{x} = (\mathbf{I} + \mathbf{M})\hat{\boldsymbol{\theta}}_{\mathrm{LS}}, \tag{2.12}$$

with $\mathbf{G} = (\mathbf{I} + \mathbf{M})(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}$. Indeed, the bias of $\hat{\boldsymbol{\theta}}$ is

$$\mathbf{b}(\boldsymbol{\theta}_0) = (\mathbf{GH} - \mathbf{I})\boldsymbol{\theta}_0 = \mathbf{M}\boldsymbol{\theta}_0 \tag{2.13}$$

so that the bias gradient matrix is $\mathbf{M}$. The total variance of $\hat{\boldsymbol{\theta}} = \mathbf{Gx}$ is

$$E\{\|\hat{\boldsymbol{\theta}} - \mathbf{GH}\boldsymbol{\theta}_0\|^2\} = E\{\|\mathbf{G}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}_0)\|^2\} = \mathrm{Tr}(\mathbf{GCG}^T), \tag{2.14}$$

which is equal to (2.9). When $\mathbf{M} = \mathbf{0}$ so that the estimator is unbiased, $\hat{\boldsymbol{\theta}}$ of (2.12) reduces to the LS solution, as we expect.

2.1.3 CRB with Constraints

In some applications, the possible values of $\boldsymbol{\theta}_0$ may be confined to a known subset of the parameter space through smooth functional constraints. If the restrictions are in the form of equality constraints, then the CRB under these limitations can be found in principle by reparameterizing the original problem. However, this approach can be difficult in general models, and obscures insight into the problem. Instead, a constrained CRB under general equality restrictions was derived in [67, 106]. When the constraint set is expressed as a smooth inequality restriction, the constrained CRB is identical to the unconstrained CRB at all regular points of the space, i.e., all points where the constraints are not active [67]. Therefore, in practice, in the presence of inequality restrictions on $\boldsymbol{\theta}_0$, the performance of an estimate with a particular bias structure is still limited by the CRB.

2.2 Bias Gradient Matrix

Typically in estimation problems there are two conflicting objectives: We would like to choose $\hat{\boldsymbol{\theta}}$ to achieve the smallest possible total variance *and* the smallest bias. However, generally, minimizing the bias results in an increase in variance and vice versa. The MSE allows an optimal tradeoff between the variance and bias in terms of estimation error by considering their sum rather than each separately; we discuss MSE optimization in the next section. In some signal processing applications the bias gradient is related to a physical parameter such as image resolution in the context of imaging [108]. In such settings the problem definition

may impose *a-priori* constraints on the bias gradient. It may then be useful to optimize the variance subject to the given bias constraints. Even when no prior restrictions are given, it can still be of interest to characterize the fundamental bias–variance tradeoff by analyzing the lowest possible variance achievable for different limitations on the bias. This idea was first introduced in [78, 80] for estimating a scalar parameter, and then extended in [41] to the vector case. A concrete application to pinhole SPECT system design is developed in [108].

To characterize the bias–variance tradeoff we need a measure of the bias. Choosing the bias itself is not useful since an estimate can be found that makes both the bias and the variance equal to zero at a given point. Indeed, for any fixed value of $\boldsymbol{\theta}_0$ we can choose $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$. Clearly, this estimate is not useful for values other than $\boldsymbol{\theta}_0$. However, when the unknown parameter $\boldsymbol{\theta}$ is equal to $\boldsymbol{\theta}_0$ this choice will lead to a zero bias and zero variance. An alternative suggestion which we adopt in this section, is to explore the smallest variance attainable when the bias gradient norm is limited by some constant [41, 78, 80, 108].

The uniform CRB (UCRB) is a bound on the variance achievable with any estimate whose bias gradient matrix $\mathbf{D}(\boldsymbol{\theta}_0)$ has bounded norm. Note that $\mathbf{D}(\boldsymbol{\theta}_0)$ is invariant to a constant bias term, so that in effect it characterizes the part of the bias that cannot be removed. To better understand the role of $\mathbf{D}(\boldsymbol{\theta}_0)$, and why we choose to bound its norm, recall that a fundamental difficulty in characterizing the bias–variance tradeoff is that we can reduce both measures to zero by using the estimate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$. Clearly, this choice is only meaningful for a particular value of $\boldsymbol{\theta}_0$. In order to focus attention on a class of reasonable estimates we constrain the bias function to be slowly varying so that it does not change too rapidly over a neighborhood of $\boldsymbol{\theta}_0$. More specifically, we restrict the squared-norm $\|\mathbf{b}(\boldsymbol{\theta}) - \mathbf{b}(\boldsymbol{\theta}_0)\|^2$ over the unit sphere

$$\mathcal{S} = \{\boldsymbol{\theta} | (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \leq 1\}. \tag{2.15}$$

The norm $\|\mathbf{b}(\boldsymbol{\theta}) - \mathbf{b}(\boldsymbol{\theta}_0)\|^2$ depends on the specific choice of $\boldsymbol{\theta} \in \mathcal{S}$. In our development, we consider two measures of the norm: A worst-case approach in which we treat the largest norm over $\mathcal{S}$, and an average strategy in which the norm is evaluated at an "average" point in $\mathcal{S}$.

To relate the resulting norm constraints to the bias gradient $\mathbf{D}(\boldsymbol{\theta}_0)$, we use a first-order Taylor approximation to write

$$\mathbf{b}(\boldsymbol{\theta}) - \mathbf{b}(\boldsymbol{\theta}_0) \approx \mathbf{D}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \stackrel{\triangle}{=} \mathbf{D}\mathbf{u}, \tag{2.16}$$

where $\mathbf{u} = \boldsymbol{\theta} - \boldsymbol{\theta}_0$, and for brevity we omitted the dependency of $\mathbf{D}$ on $\boldsymbol{\theta}_0$. As we now show, the largest and average norm variation over $\mathcal{S}$ are both related to matrix norms of $\mathbf{D}$.

Using (2.16), we can approximate the maximal variation of $\|\mathbf{b}(\boldsymbol{\theta}) - \mathbf{b}(\boldsymbol{\theta}_0)\|^2$ over $\mathcal{S}$ as

$$\max_{\boldsymbol{\theta} \in \mathcal{S}} \|\mathbf{b}(\boldsymbol{\theta}) - \mathbf{b}(\boldsymbol{\theta}_0)\|^2 \approx \max_{\|\mathbf{u}\|^2 \leq 1} \|\mathbf{D}\mathbf{u}\|^2 = \|\mathbf{D}\|^2, \tag{2.17}$$

where $\|\mathbf{D}\|$ denotes the spectral norm of $\mathbf{D}$ [84], i.e., the largest singular value. The worst case variation $\|\mathbf{D}\|^2$ occurs when $\mathbf{u}$ is chosen to be a unit-norm vector in the direction of the eigenvector corresponding to the largest eigenvalue of $\mathbf{D}^T\mathbf{D}$. To develop an average bias measure, instead of choosing $\mathbf{u}$ to be in the direction of the worst-case eigenvector, we select a weighted average of the eigenvectors $\mathbf{u} = \sum_{i=1}^m a_i \mathbf{v}_i$, where $\mathbf{v}_i, 1 \leq i \leq m$ are the eigenvectors of $\mathbf{D}^T\mathbf{D}$, and $a_i > 0$ are arbitrary coefficients satisfying $\sum_{i=1}^m a_i^2 = 1$, so that $\|\mathbf{u}\| = 1$. For this choice of $\mathbf{u}$,

$$\|\mathbf{D}\mathbf{u}\|^2 = \mathrm{Tr}\left(\mathbf{V}\mathbf{A}\mathbf{V}^T\mathbf{D}^T\mathbf{D}\right) = \mathrm{Tr}\left(\mathbf{D}^T\mathbf{D}\mathbf{Q}\right), \tag{2.18}$$

where $\mathbf{V}$ is the matrix of eigenvectors $\mathbf{v}_i$, $\mathbf{A} = \mathrm{diag}\,(a_1^2, \ldots, a_m^2)$ and $\mathbf{Q} = \mathbf{V}\mathbf{A}\mathbf{V}^T$.

Motivated by these observations, we consider the following two measures of bias gradient: an average measure corresponding to a weighted squared Frobenius norm,

$$D_{\mathrm{AVG}} = \mathrm{Tr}\left(\mathbf{D}^T\mathbf{D}\mathbf{W}\right), \tag{2.19}$$

where $\mathbf{W} \succ 0$, and a worst case bias gradient measure corresponding to a weighted squared spectral norm,

$$D_{\mathrm{WC}} = \max_{\|\mathbf{z}\|=1} \mathbf{z}^T\mathbf{S}\mathbf{D}^T\mathbf{D}\mathbf{S}\mathbf{z}, \tag{2.20}$$

for some $\mathbf{S} \succ 0$.

2.3 Uniform Cramér–Rao Bound (UCRB)

The UCRB limits the variance of any estimate with bias gradient norm bounded by a constant, where we treat both norms D_{AVG} and D_{WC}.

In the sequel, we omit the dependency of $\mathbf{b}, \mathbf{D}$, and $\mathbf{J}$ on $\boldsymbol{\theta}_0$, for simplicity of notation.

2.3.1 Average Bias Constraint

We first treat the problem of minimizing $\text{Tr}(\mathbf{C}(\mathbf{D}))$ with $\mathbf{C}(\mathbf{D})$ given by (2.9) subject to $D_{\text{AVG}} \leq \gamma$:

$$\begin{aligned} \min_{\mathbf{D}} \quad & (\mathbf{I}+\mathbf{D})\,\mathbf{J}^{-1}\,(\mathbf{I}+\mathbf{D})^T \\ \text{s.t.} \quad & \text{Tr}\left(\mathbf{D}^T\mathbf{D}\mathbf{W}\right) \leq \gamma. \end{aligned} \tag{2.21}$$

If $\gamma \geq \text{Tr}(\mathbf{W})$, then we can choose $\mathbf{D} = -\mathbf{I}$ which results in $\text{Tr}(\mathbf{C}(\mathbf{D})) = 0$. This corresponds to using the estimate $\hat{\boldsymbol{\theta}} = \mathbf{0}$. We next consider the case $0 < \gamma < \text{Tr}(\mathbf{W})$. Since $\text{Tr}(\mathbf{C}(\mathbf{D}))$ and D_{AVG} are both convex in $\mathbf{D}$, and (2.21) is strictly feasible, we can find the optimal $\mathbf{D}$ using the Karush–Kuhn–Tucker (KKT) conditions [16] (see Theorem A.2 in the Appendix). To this end, we form the Lagrangian

$$\mathcal{L} = \text{Tr}\left((\mathbf{I}+\mathbf{D})\,\mathbf{J}^{-1}\left(\mathbf{I}+\mathbf{D}\right)^T\right) + \alpha\left(\text{Tr}(\mathbf{D}^T\mathbf{D}\mathbf{W}) - \gamma\right), \tag{2.22}$$

where $\alpha \geq 0$. The optimal solution $\mathbf{D} = \widehat{\mathbf{D}}_{\text{AVG}}$ is determined by setting the derivative of $\mathcal{L}$ to $\mathbf{0}$ which results in

$$\widehat{\mathbf{D}}_{\text{AVG}} = -\mathbf{J}^{-1}\left(\mathbf{J}^{-1} + \alpha\mathbf{W}\right)^{-1} = -\mathbf{I} + \alpha\left(\mathbf{W}^{-1} + \alpha\mathbf{J}\right)^{-1}\mathbf{J}. \tag{2.23}$$

The last equality follows from the matrix inversion lemma. If $\alpha = 0$, then $\widehat{\mathbf{D}}_{\text{AVG}} = -\mathbf{I}$ which violates the constraint (2.21). Therefore, $\alpha > 0$ which from the KKT conditions implies that (2.21) is satisfied with equality so that

$$\text{Tr}(\widehat{\mathbf{D}}_{\text{AVG}}^T\widehat{\mathbf{D}}_{\text{AVG}}\mathbf{W}) = \text{Tr}\left((\mathbf{W}^{-1} + \alpha\mathbf{J})^{-2}\mathbf{W}^{-1}\right) = \gamma. \tag{2.24}$$

We conclude that the total variance of any estimator with bias gradient $\mathbf{D}$ satisfying (2.21) with $0 < \gamma < \text{Tr}(\mathbf{W})$ is bounded by

$$\begin{aligned} \text{Tr}(\mathbf{C}) &\geq \text{Tr}\big((\mathbf{I}+\widehat{\mathbf{D}}_{\text{AVG}})\mathbf{J}^{-1}(\mathbf{I}+\widehat{\mathbf{D}}_{\text{AVG}})^T\big) \\ &= \alpha^2\text{Tr}\left((\mathbf{W}^{-1}+\alpha\mathbf{J})^{-2}\mathbf{J}\right), \end{aligned} \tag{2.25}$$

where $\alpha > 0$ is the unique solution of (2.24).

2.3.2 Worst-Case Bias Constraint

We now treat the constraint

$$D_{\rm wc} = \max_{\|\mathbf{z}\|=1} \mathbf{z}^T \mathbf{S}\mathbf{D}^T\mathbf{D}\mathbf{S}\mathbf{z} \le \gamma, \tag{2.26}$$

for some positive definite matrix $\mathbf{S}$. Denoting by $\lambda_{\max}(\mathbf{S})$ the largest eigenvalue of $\mathbf{S}$, it is clear that for $\gamma \ge \lambda^2_{\max}(\mathbf{S})$ we can choose $\mathbf{D} = -\mathbf{I}$, which results in zero total variance. The derivation of the UCRB when $\gamma < \lambda^2_{\max}(\mathbf{S})$ is considerably more involved. Therefore, we omit the details here, and refer the interested reader to [41]. When $\mathbf{S}$ has the same eigenvectors as $\mathbf{J}$, the $\mathbf{D}$ minimizing the total variance subject to (2.26) is

$$\widehat{\mathbf{D}}_{\rm wc} = \big(\mathbf{I} - \sqrt{\gamma}\mathbf{S}^{-1}\big)\mathbf{P} - \mathbf{I}. \tag{2.27}$$

Here $\mathbf{P}$ is the orthogonal projection onto the space spanned by the eigenvectors of $\mathbf{S}$ corresponding to eigenvalues larger than $\sqrt{\gamma}$. The total variance is then bounded by

$$\mathrm{Tr}(\mathbf{C}) \ge \mathrm{Tr}\big(\big(\mathbf{I} - \sqrt{\gamma}\mathbf{S}^{-1}\big)^2\mathbf{P}\mathbf{J}^{-1}\big), \tag{2.28}$$

where we used the fact that $\mathbf{J}^{-1}, \mathbf{P}$, and $\mathbf{S}^{-1}$ all commute.

In the special case in which $\mathbf{S} = \mathbf{I}$, all the eigenvalues of $\mathbf{S}$, which are equal to 1, are larger than γ, which is constrained to be smaller than $\lambda^2_{\max}(\mathbf{S}) = 1$. Thus, $\mathbf{P} = \mathbf{I}$, and

$$\mathrm{Tr}(\mathbf{C}) \ge \mathrm{Tr}\big((1 - \sqrt{\gamma})^2\mathbf{J}^{-1}\big). \tag{2.29}$$

As we expect, when $\gamma \to 0$, the UCRB coincides with the CRB.

We summarize the UCRB in the following theorem.

Theorem 2.1. Let $\mathbf{x}$ denote measurements of a deterministic parameter vector $\boldsymbol{\theta}_0$ with pdf $p(\mathbf{x};\boldsymbol{\theta}_0)$ and Fisher information matrix $\mathbf{J} = \mathbf{J}(\boldsymbol{\theta}_0)$. Let $\hat{\boldsymbol{\theta}}$ be an arbitrary estimate of $\boldsymbol{\theta}_0$ with covariance matrix $\mathbf{C} = \mathbf{C}(\boldsymbol{\theta}_0)$ and bias gradient matrix $\mathbf{D} = \mathbf{D}(\boldsymbol{\theta}_0)$. Then we have the following:

(1) If $\mathrm{Tr}(\mathbf{D}^T\mathbf{D}\mathbf{W}) \le \gamma < \mathrm{Tr}(\mathbf{W})$ for some $\mathbf{W} \succ 0$ then

$$\mathrm{Tr}(\mathbf{C}) \ge \alpha^2\mathrm{Tr}\big((\mathbf{W}^{-1} + \alpha\mathbf{J})^{-2}\mathbf{J}\big),$$

where $\alpha > 0$ is chosen such that

$$\mathrm{Tr}\big((\mathbf{W}^{-1} + \alpha\mathbf{J})^{-2}\mathbf{W}^{-1}\big) = \gamma.$$

(2) If $\max_{\|\mathbf{z}\|=1} \mathbf{z}^T\mathbf{S}\mathbf{D}^T\mathbf{D}\mathbf{S}\mathbf{z} \leq \gamma < \lambda^2_{\max}(\mathbf{S})$ for some $\mathbf{S} \succ 0$ that has the same eigenvector matrix as $\mathbf{J}$ then

$$\mathrm{Tr}(\mathbf{C}) \geq \mathrm{Tr}\big((\mathbf{I} - \sqrt{\gamma}\mathbf{S}^{-1})^2\mathbf{P}\mathbf{J}^{-1}\big),$$

where $\mathbf{P}$ is the orthogonal projection onto the space spanned by the eigenvectors of $\mathbf{S}$ corresponding to eigenvalues larger than $\sqrt{\gamma}$.

2.3.3 Geometrical Interpretation

Geometrically, the UCRB divides the bias–variance plane into two regions, and limits the boundary of the achievable points, namely the points $(\gamma, \min_{\mathbf{D}\in\mathcal{V}} \mathrm{Tr}(\mathbf{C}(\mathbf{D})))$. Here $\mathcal{V}$ is the set of bias gradient matrices $\mathbf{D}$ for which the appropriate norm (average, or worst-case) is constrained by γ. All points on one part of the plane are unachievable, while the points on the other part can possibly be attained (depending on whether or not the UCRB is tight). As we will see in the next section, at least asymptotically, the boundary can typically be achieved. Since both the bias and the variance depend on $\boldsymbol{\theta}_0$, the curve will be different for varying choices of $\boldsymbol{\theta}_0$. Figure 2.1 illustrates an example for a particular $\boldsymbol{\theta}_0$. Two important points shown in Figure 2.1 are the circle on the x-axis, corresponding to $\mathbf{D} = -\mathbf{I}$ for which the total variance is 0, and the circle on the y-axis, which represents the unbiased CRB corresponding to $\gamma = 0$. The points on the boundary are the UCRB which follow from minimizing the total variance subject to the appropriate norm constraint.

2.4 Achieving the UCRB

An interesting aspect of the UCRB is that in many settings it can be achieved, at least asymptotically. We begin with the linear Gaussian model for which the bounds are achieved exactly by linear estimators. We then show that for general pdfs, the bounds can be approached asymptotically using a PML approach.

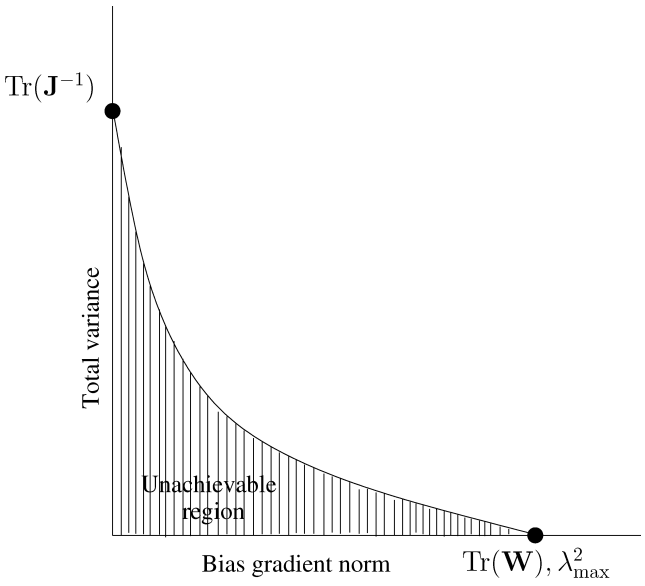


Fig. 2.1 A qualitative example of the optimal bias–variance tradeoff curve.

2.4.1 Linear Gaussian Model

We first treat the class of estimation problems represented by the linear Gaussian model (1.2).

As we have seen, in this setting $\mathbf{J} = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}$. Furthermore, any estimator of the form (2.12) achieves the biased CRB on estimators with bias gradient $\mathbf{M}$. Since in this case $\widehat{\mathbf{D}}_{\text{AVG}}$ is independent of $\boldsymbol{\theta}_0$, we can choose $\mathbf{M} = \widehat{\mathbf{D}}_{\text{AVG}}$, which leads to an estimate that achieves the average UCRB:

$$\hat{\boldsymbol{\theta}} = \begin{cases} (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H} + \delta\mathbf{W}^{-1})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}, & 0 \le \gamma < \mathrm{Tr}(\mathbf{W}); \\ 0, & \gamma \ge \mathrm{Tr}(\mathbf{W}), \end{cases} \tag{2.30}$$

where $\delta = 1/\alpha > 0$ satisfies $\mathrm{Tr}\left((\mathbf{W}^{-1} + (1/\delta)\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-2}\mathbf{W}^{-1}\right) = \gamma$. The estimator $\hat{\boldsymbol{\theta}}$ of (2.30) is equal to the ridge estimator proposed by Hoerl and Kennard [81] (also known as Tikhonov regularization [136]), and is widely used for solving inverse problems [70, 111].

Similarly, the worst-case UCRB is achieved with $\mathbf{M} = \widehat{\mathbf{D}}_{\mathrm{wc}}$, which leads to the estimate

$$\hat{\boldsymbol{\theta}} = \begin{cases} (\mathbf{I} - \sqrt{\gamma}\mathbf{S}^{-1})\mathbf{P}(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}, & 0 \le \gamma < \lambda^2_{\max}(\mathbf{S}); \\ 0, & \gamma \ge \lambda^2_{\max}(\mathbf{S}). \end{cases} \tag{2.31}$$

Here $\mathbf{P}$ is an orthogonal projection onto the space spanned by the eigenvectors of $\mathbf{S}$ corresponding to eigenvalues larger than $\sqrt{\gamma}$. When $\mathbf{S} = \mathbf{I}$, (2.31) is equal to the shrunken estimator proposed by Mayer and Willke [107], which is simply a scaled version of the LS solution.

We conclude that Tikhonov regularization and the shrunken estimator have a strong optimality property: among all linear and nonlinear estimators of $\boldsymbol{\theta}_0$ in the linear Gaussian model (1.2) with bounded bias gradient, they minimize the total variance. This provides further justification for these methods, which are used in many applications. It is also easy to see that when $\gamma \to 0$, both solutions converge to LS.

2.4.2 Asymptotic Optimality of the PML Estimator

As we have just seen, in the linear Gaussian model, the UCRB can be achieved with a linear estimator. When the average bias is considered, the estimator takes on the form of Tikhonov regularization. The Tikhonov solution also maximizes the penalized log-likelihood

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\max \left\{ \log p(\mathbf{x};\boldsymbol{\theta}) - \frac{\beta}{2}\boldsymbol{\theta}^T\mathbf{W}\boldsymbol{\theta} \right\} \\ &= \arg\min \left\{ (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T\mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) + \beta\boldsymbol{\theta}^T\mathbf{W}\boldsymbol{\theta} \right\}, \end{aligned} \tag{2.32}$$

where $p(\mathbf{x};\boldsymbol{\theta})$ is the Gaussian distribution with mean $\mathbf{H}\boldsymbol{\theta}$ and covariance $\mathbf{C}$, and β is a regularization parameter. When the worst-case bias is considered with weighting $\mathbf{S} = \mathbf{I}$, the shrunken estimator achieves the UCRB. We can immediately verify that this method also maximizes (2.32), with $\mathbf{W} = -\mathbf{H}^T\mathbf{H}$. A similar result holds when $\mathbf{S}$ has the same eigenvector matrix as $\mathbf{J}$. Evidently, the PML solution with an appropriate choice of penalizing function achieves the UCRB in the linear Gaussian model. In [41], it is shown that this optimality property of the PML approach holds more generally *asymptotically*.

The PML estimator of $\boldsymbol{\theta}_0$, denoted $\hat{\boldsymbol{\theta}}_{\text{PML}}$, is chosen to maximize the penalized log-likelihood function

$$\log p(\mathbf{x};\boldsymbol{\theta}) - \beta R(\boldsymbol{\theta}), \tag{2.33}$$

where $\beta > 0$ is a regularization parameter, and $R(\boldsymbol{\theta})$ is a penalizing function. The PML approach is equivalent to the maximum *a posteriori* method in Bayesian estimation if we interpret $e^{-\beta R(\boldsymbol{\theta})}$ as the prior pdf of $\boldsymbol{\theta}_0$. When N iid (vector) measurements $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are given, $\hat{\boldsymbol{\theta}}_{\text{PML}}$ is chosen to maximize

$$PL(\theta) = \sum_{i=1}^{N} \log p(\mathbf{x}_i;\boldsymbol{\theta}) - \beta_N R(\boldsymbol{\theta}), \tag{2.34}$$

where β_N is a regularization parameter that depends on N and satisfies $\beta_N/N \to \beta_0$ as $N \to \infty$.

The asymptotic distribution of any PML estimate with smooth penalizing function was derived in [41] based on which it was shown that in many cases the penalizing function $R(\boldsymbol{\theta})$ can be chosen such that the resulting PML solution asymptotically achieves the UCRB. This provides a method for selecting the penalizing function by achieving an optimal bias–variance tradeoff.

In the special case of estimating a scalar θ_0, we have the following proposition.

Proposition 2.2. Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ denote N iid measurements of a deterministic parameter θ_0 with pdf $p(\mathbf{x};\theta_0)$. Let $\hat{\theta}_{\text{PML}}$ denote the PML estimator of θ_0 that maximizes the penalized log-likelihood (2.34) with penalizing function $R(\theta)$. Suppose that $d\check{\theta}(\theta_0)/d\theta_0 \leq 1$, where

$$\check{\theta}(\theta) = \arg\max_{\theta} \left\{ E\left\{\log p(\mathbf{x};\theta)\right\} - \beta_0 R(\theta) \right\},$$

and $\beta_0 = \lim_{N\to\infty} \beta_N/N$. Then $\hat{\theta}_{\text{PML}}$ asymptotically achieves the UCRB if and only if $R(\theta)$ is chosen such that

$$\frac{d\log p(\mathbf{x};\check{\theta})}{d\theta} - E\left\{ \frac{d\log p(\mathbf{x};\check{\theta})}{d\theta} \right\} = c\frac{d\log p(\mathbf{x};\theta_0)}{d\theta}, \tag{2.35}$$

for some deterministic constant c.

The more general case can be found in [41].

2.4.3 Example of PML Optimality

We now consider an example, taken from [41], that illustrates the PML estimator and its asymptotic optimality.

Suppose we are given N scalar iid measurements $x_1, \ldots, x_N$ of an exponential random variable with unknown mean $1/\theta_0 > 0$. Thus,

$$p(x_i;\theta_0) = \theta_0 e^{-x_i\theta_0}, \quad x_i \geq 0, \quad 1 \leq i \leq N. \tag{2.36}$$

The PML estimate $\hat{\theta}_{\text{PML}}$ with penalizing function $R(\theta)$ is given by the value of θ that maximizes

$$PL(\theta) = N \log\theta - \theta \sum_{i=1}^{N} x_i - \beta_N R(\theta), \tag{2.37}$$

for some parameter $\beta_N > 0$ such that $\beta_N/N \to \beta_0$ as $N \to \infty$. We seek a penalizing function $R(\theta)$ that is optimal in the sense that the resulting estimator asymptotically achieves the UCRB.

From (2.36),

$$\frac{d \log p(x;\theta)}{d\theta} = \frac{1}{\theta} - x, \tag{2.38}$$

so that

$$E\left\{ \frac{d \log p(x;\check{\theta})}{d\theta} \right\} = \frac{1}{\check{\theta}} - \frac{1}{\theta_0}. \tag{2.39}$$

Therefore,

$$\frac{d \log p(x;\check{\theta})}{d\theta} - E\left\{ \frac{d \log p(x;\check{\theta})}{d\theta} \right\} = \frac{1}{\theta_0} - x, \tag{2.40}$$

and (2.35) is satisfied with $c = 1$. From Proposition 2.2 it follows that for any choice of $R(\theta)$ such that $d\check{\theta}/d\theta_0 \leq 1$, the resulting PML inference asymptotically achieves the UCRB. Note, however, that for finite values of N, the performance of the PML estimator will depend on the specific choice of $R(\theta)$.

As an example, suppose that $R(\theta) = \theta$. The resulting PML estimator is given by

$$\hat{\theta}_{\text{PML}} = \arg\max \left\{ N \log\theta - \theta \left(\sum_{i=1}^{N} x_i + \beta_N \right) \right\} = \frac{N}{\sum_{i=1}^{N} x_i + \beta_N}. \tag{2.41}$$

From the definition of $\check{\theta}$ we have that

$$\check{\theta} = \arg\max \left\{ \log\theta - \frac{\theta}{\theta_0} - \beta_0 R(\theta) \right\} = \frac{\theta_0}{1+\beta\theta_0}. \tag{2.42}$$

Therefore, $d\check{\theta}(\theta_0)/d\theta_0 \leq 1$, and from Proposition 2.2 the estimator of (2.41) asymptotically achieves the UCRB.

As another example, suppose that $R(\theta) = \log\theta$. In this case

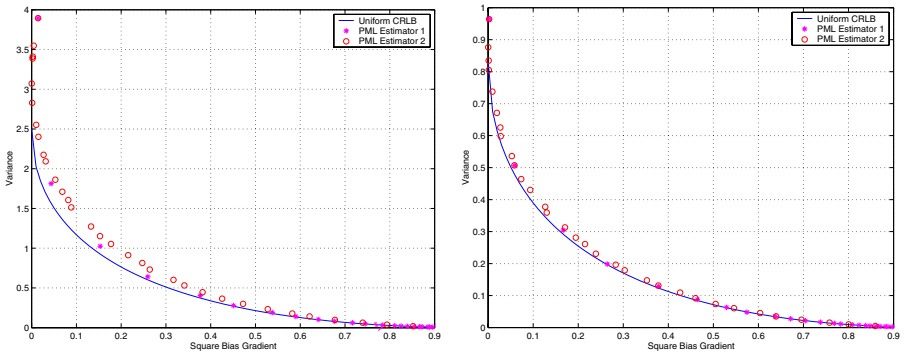
$$\check{\theta} = (1-\beta_0)\theta_0, \tag{2.43}$$

so that again $d\check{\theta}(\theta_0)/d\theta_0 \leq 1$. We therefore conclude that the resulting PML estimator, given by

$$\hat{\theta}_{\mathrm{PML}} = \arg\max \left\{ (N-\beta_N)\log\theta - \theta \sum_{i=1}^{N} x_i \right\} = \frac{N-\beta_N}{\sum_{i=1}^{N} x_i}, \tag{2.44}$$

asymptotically achieves the UCRB.

In Figure 2.2 we compare the performance of the PML methods of (2.41) and (2.44) with the UCRB, for different values of N. In the figures, the variance and the squared bias gradient of the estimators are approximated from the measurements. Specifically, for each γ we generate $L = 5{,}000$ PML estimators, where each one is based on N iid



(a) (b)

Fig. 2.2 Performance of the PML estimators (2.41) (denoted "1") and (2.44) (denoted "2") for different values of N in comparison with the UCRB. The line denotes the UCRB, the stars denote the performance of PML estimator 1, and the circles denote the performance of PML estimator 2. (a) $N = 10$. (b) $N = 30$.

measurements. The variance is approximated by the empirical variance; the squared bias gradient is estimated using the procedure detailed in [80] (see also an explanation in [41]). In Figure 2.2(a) we plot the variance of the PML estimators as a function of the squared bias gradient for $N = 10$, and in Figure 2.2(b) we choose $N = 30$.

From the figures it is apparent that even for small N the UCRB serves as a good approximation to the estimator's variance, particularly for large values of bias gradient norm. As we expect from our analysis, for increasing values of N the variance of both estimators approaches that of the UCRB for all values of squared bias gradient. Note, however, that for small values of N the performance of the two estimators is different. In particular, the estimator given by (2.41) results in smaller variance.

In this section we outlined tools to characterize the fundamental tradeoff between variance and bias, by deriving lower bounds on the minimal achievable total variance subject to constraints on the norm of the bias gradient matrix. We then showed that for an appropriate choice of penalizing function, the PML estimator asymptotically achieves the UCRB. In the next section we treat the more general scenario in which we do not know in advance properties of the desired bias gradient, and instead would like to directly optimize the MSE.

3

Mean-Squared Error Bounds

We now focus our attention on the MSE and develop bounds on the minimal achievable MSE of biased estimators. Our goal is to choose the bias in an optimal way such that the overall MSE is reduced with respect to unbiased estimation. In this section the biasing is accomplished by multiplying an unbiased estimator by a suitable matrix (or by scaling the unbiased method when the estimator is a scalar) which introduces some bias but lessens the variability in such a way that the overall MSE is reduced. The important aspect of the framework we introduce is that the reduction in MSE is guaranteed for all choices of the unknown parameter vector. In subsequent sections we will discuss nonlinear modifications of the MVU solution.

Although in our derivations we treat the (unweighted) MSE, the essential ideas introduced in this section can be generalized to include weighted MSE criteria which measure the average weighted squared-norm error [48].

3.1 MSE Bound

Recall that the MSE of a given estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ is given by

$$E\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2\} = \|\mathbf{b}(\boldsymbol{\theta}_0)\|^2 + \mathrm{Tr}(\mathbf{C}(\boldsymbol{\theta}_0)), \tag{3.1}$$

where $\mathbf{b}(\boldsymbol{\theta}_0)$ is the estimator bias and $\mathbf{C}(\boldsymbol{\theta}_0)$ is its covariance matrix. In the previous section we discussed bounds on $\mathrm{Tr}(\mathbf{C}(\boldsymbol{\theta}_0))$ for a given choice of $\mathbf{b}(\boldsymbol{\theta}_0)$, or for estimates with bias gradient in a suitable set. A method achieving the corresponding CRB has minimum variance among its class.

Instead of limiting only the variance of a biased estimator we can bound the MSE which is a direct measure of estimator performance. Substituting the biased CRB (2.11) into (3.1) it follows that the MSE of any estimator with bias $\mathbf{b}(\boldsymbol{\theta}_0)$ is bounded below by

$$\|\mathbf{b}(\boldsymbol{\theta}_0)\|^2 + \mathrm{Tr}\left((\mathbf{I} + \mathbf{D}(\boldsymbol{\theta}_0))\mathbf{J}^{-1}(\boldsymbol{\theta}_0)(\mathbf{I} + \mathbf{D}(\boldsymbol{\theta}_0))^T\right). \tag{3.2}$$

A similar expression can be obtained for the weighted MSE [48]. Ideally, to obtain the tightest possible MSE bound, we would like to minimize (3.2) over all bias vectors $\mathbf{b}(\boldsymbol{\theta}_0)$. For every fixed value of $\boldsymbol{\theta}_0$ the minimum can be obtained with $\mathbf{b}(\boldsymbol{\theta}) = \boldsymbol{\theta}_0 - \boldsymbol{\theta}$; for this choice $\mathbf{b}(\boldsymbol{\theta}_0) = 0$ and $\mathbf{D}(\boldsymbol{\theta}_0) = -\mathbf{I}$. The minimum of (3.2) is equal to 0 and is achieved at $\boldsymbol{\theta}_0$ by the estimate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$ which clearly cannot be implemented. Furthermore, $\hat{\boldsymbol{\theta}}$ optimizes the bound for any specific $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ but not for all values of $\boldsymbol{\theta}$. Thus, in general we cannot minimize (3.2) point-wise, for all $\boldsymbol{\theta}_0$. Nonetheless, in some cases, we may be able to minimize the bound over all bias vectors in a suitable class. In the more challenging setting, in which the bound cannot be minimized directly, it still may be possible to find a bias $\mathbf{b}(\boldsymbol{\theta}_0)$ such that the resulting MSE bound is smaller than the unbiased CRB for all possible values of $\boldsymbol{\theta}_0$. Our goal therefore is to minimize the MSE benchmark over all bias vectors in a suitable class (which includes the zero bias), when possible. Otherwise, we aim at finding a bias vector such that the resulting MSE limit is smaller than the unbiased CRB for all values of $\boldsymbol{\theta}_0$ in a predefined set [47].

The class of bias functions we consider in this section are linear bias vectors of the form:

$$\mathbf{b}(\boldsymbol{\theta}_0) = \mathbf{M}\boldsymbol{\theta}_0, \tag{3.3}$$

for some $m \times m$ matrix $\mathbf{M}$. With this choice, the MSE bound of (3.2) becomes:

$$\mathrm{MSEB}\,(\mathbf{M}, \boldsymbol{\theta}_0) = \boldsymbol{\theta}_0^T \mathbf{M}^T \mathbf{M} \boldsymbol{\theta}_0 + \mathrm{Tr}\left((\mathbf{I} + \mathbf{M})\mathbf{J}^{-1}(\boldsymbol{\theta}_0)(\mathbf{I} + \mathbf{M})^T\right). \tag{3.4}$$

If $\mathbf{M} = 0$, then as we expect the bound coincides with the CRB: $\mathrm{MSEB}\,(0, \boldsymbol{\theta}_0) = \mathrm{Tr}(\mathbf{J}^{-1}(\boldsymbol{\theta}_0))$. In the next section we show how these ideas can be extended to nonlinear bias functions via the blind minimax framework.

Throughout this section, we will assume that an estimate exists that achieves the unbiased CRB. The estimators we develop are obtained by linearly transforming the given efficient strategy. Nonetheless, there are a variety of problems in which the CRB cannot be achieved, but an MVU solution can be found. An example is when $p(\mathbf{x};\theta_0)$ is the uniform distribution on $[0,\theta_0]$ [93]. Our approach can be applied to this setting as well by replacing the CRB by the variance of an MVU solution. The proposed estimators are then linear transformations of the corresponding MVU method.

An advantage of restricting attention to linear bias vectors is that we can use results on unbiased estimation to find estimators that achieve the corresponding MSE bound. Specifically, if $\hat{\boldsymbol{\theta}}$ is an efficient unbiased method whose MSE is given by the CRB, then the MSE of

$$\hat{\boldsymbol{\theta}}_{\mathrm{b}} = (\mathbf{I} + \mathbf{M})\hat{\boldsymbol{\theta}} \tag{3.5}$$

is equal to $\mathrm{MSEB}\,(\mathbf{M}, \boldsymbol{\theta}_0)$. To see this, since $E\{\hat{\boldsymbol{\theta}}\} = \boldsymbol{\theta}_0$,

$$\mathbf{b}_{\hat{\boldsymbol{\theta}}_{\mathrm{b}}}(\boldsymbol{\theta}_0) = (\mathbf{I} + \mathbf{M})E\{\hat{\boldsymbol{\theta}}\} - \boldsymbol{\theta}_0 = \mathbf{M}\boldsymbol{\theta}_0. \tag{3.6}$$

Using the fact that $\hat{\boldsymbol{\theta}}_{\mathrm{b}} - E\{\hat{\boldsymbol{\theta}}_{\mathrm{b}}\} = (\mathbf{I} + \mathbf{M})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and $\mathbf{C}(\boldsymbol{\theta}_0) = \mathbf{J}^{-1}(\boldsymbol{\theta}_0)$,

$$\begin{aligned} \mathbf{C}_{\hat{\boldsymbol{\theta}}_{\mathrm{b}}}(\boldsymbol{\theta}_0) &= (\mathbf{I} + \mathbf{M})E\big\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\big\}(\mathbf{I} + \mathbf{M})^T \\ &= (\mathbf{I} + \mathbf{M})\mathbf{J}^{-1}(\boldsymbol{\theta}_0)(\mathbf{I} + \mathbf{M})^T, \end{aligned} \tag{3.7}$$

so that the MSE of $\hat{\boldsymbol{\theta}}_{\mathrm{b}}$ is given by $\mathrm{MSEB}\,(\mathbf{M}, \boldsymbol{\theta}_0)$. Therefore, if $\hat{\boldsymbol{\theta}}$ achieves the CRB and we find an $\mathbf{M}$ such that $\mathrm{MSEB}\,(\mathbf{M}, \boldsymbol{\theta}_0) < \mathrm{MSEB}\,(0, \boldsymbol{\theta}_0)$ for all feasible $\boldsymbol{\theta}_0$ in a given set, then the MSE of $\hat{\boldsymbol{\theta}}_{\mathrm{b}}$ will be smaller than that of $\hat{\boldsymbol{\theta}}$ for all $\boldsymbol{\theta}_0$ in the set. This allows us to reduce the MSE by a simple linear transformation. The important point is that this improvement is for *all choices of* $\boldsymbol{\theta}_0$ in the set (which can be the entire space $\mathbb{R}^m$). This essential concept is illustrated in Figure 3.1. The solid line represents the bound on unbiased estimates, and is assumed to be

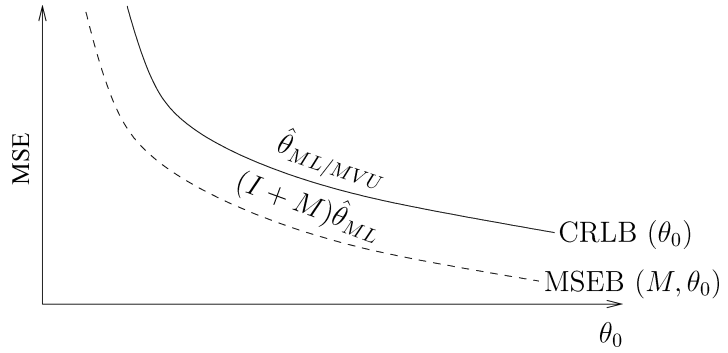


Fig. 3.1 Illustration of the linear bias bound.

achieved by an efficient ML estimate or an MVU solution. The dashed line is the bound corresponding to $\mathbf{M}$ which is lower for all choices of $\boldsymbol{\theta}_0$ and is achieved by a simple transformation of the MVU method.

In contrast, if we consider nonlinear bias vectors, then even if we find a bias that results in an MSE bound that is lower than the CRB, and an efficient estimator exists, it is still unclear in general how to construct a method achieving the resulting MSE bound.

3.1.1 Scaling to Reduce MSE

To illustrate the idea we just outlined, we now examine in more detail the use of scaling to reduce the MSE in a scalar setting [92].

Consider estimating a scalar θ_0 from the available data $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$. Suppose that $\hat{\theta}$ is an efficient (or MVU) estimator with variance $\mathrm{var}(\hat{\theta})$. To reduce the MSE, we bias $\hat{\theta}$ using a linear bias so that the biased estimate is

$$\hat{\theta}_{\mathrm{b}} = (1 + M)\hat{\theta}, \tag{3.8}$$

where M will be chosen to reduce the MSE $E\{(\hat{\theta}_{\mathrm{b}} - \theta_0)^2\}$. Using the fact that $E\{\hat{\theta}\} = \theta$, the MSE of $\hat{\theta}_{\mathrm{b}}$ becomes:

$$\mathrm{MSE}\,(\hat{\theta}_{\mathrm{b}}) = M^2\theta_0^2 + (1 + M)^2\mathrm{var}(\hat{\theta}). \tag{3.9}$$

Our goal is to choose M so that $\mathrm{MSEB}\,(\hat{\theta}_{\mathrm{b}})$ is less than the MSE of the original unbiased estimator $\hat{\theta}$, which is its variance $\mathrm{var}(\hat{\theta})$. Of course,

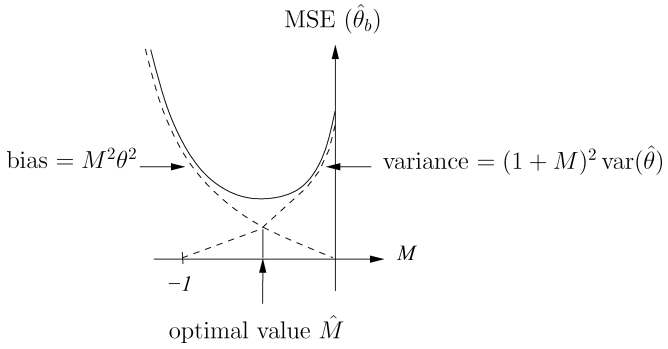


Fig. 3.2 Trading off bias for variance to reduce the MSE. The biased estimator is $\hat{\theta}_{\mathrm{b}} = (1+M)\hat{\theta}$, which is a scaled version of the unbiased MVU solution.

we would like the reduction in MSE to be as large as possible. From (3.9) we see immediately that to reduce the MSE we must have $-1 \leq M < 0$. In this interval, the bias $M^2\theta^2$ increases as M departs from zero, while the variance $(1+M)^2\mathrm{var}(\hat{\theta})$ decreases. This behavior is illustrated in Figure 3.2, where the MSE is plotted versus M over its allowable range $-1 \leq M \leq 0$ for a particular choice of θ_0. Evidently, there is a value of M that minimizes the overall MSE, trading off an increase in bias for a decrease in variability. Since with $M = 0$ we have $\hat{\theta}_{\mathrm{b}} = \hat{\theta}$, an optimal value of $M \neq 0$ will produce an estimator with a smaller MSE.

The key issue is whether the optimal M depends on the unknown value of θ_0. If it does not, then the biased estimator $\hat{\theta}_{\mathrm{b}}$ is realizable and the MSE can be minimized over linear-bias approaches. However, if M depends on θ_0, then the minimum MSE estimator cannot be implemented and it is not immediately obvious how to proceed. This latter case is the subject of Sections 3.3–3.6 in which we show that often it is still possible to choose a scaling such that the MSE of $\hat{\theta}_{\mathrm{b}}$ is reduced in comparison with $\mathrm{var}(\hat{\theta})$ for all θ_0.

3.2 Minimal MSE Bound with Linear Bias

Returning to the general vector problem, in this section we discuss cases in which the bound (3.4) can be minimized directly.

Since the objective in (3.4) is convex in $\mathbf{M}$, we can find the minimal value by setting the derivative to $\mathbf{0}$, which yields

$$\widehat{\mathbf{M}}(\mathbf{J}^{-1}(\boldsymbol{\theta}_0) + \boldsymbol{\theta}_0\boldsymbol{\theta}_0^T) = -\mathbf{J}^{-1}(\boldsymbol{\theta}_0). \tag{3.10}$$

Using the matrix inversion lemma the optimal $\mathbf{M}$ can be written as

$$\widehat{\mathbf{M}} = -\mathbf{I} + \frac{1}{1 + \boldsymbol{\theta}_0^T\mathbf{J}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^T\mathbf{J}(\boldsymbol{\theta}_0). \tag{3.11}$$

In general $\mathbf{M}$ will depend on $\boldsymbol{\theta}_0$ which is unknown, so that there is no constant value of $\mathbf{M}$ that optimizes the bound. However, if (3.11) is independent of $\boldsymbol{\theta}_0$, then this choice of $\mathbf{M}$ minimizes the bound for all possible values of $\boldsymbol{\theta}_0$. This occurs when $\boldsymbol{\theta}_0 = \theta_0$ is a scalar, and $J^{-1}(\theta_0) = \alpha\theta_0^2$ for some $\alpha > 0$. In this case the optimal choice of $\mathbf{M} = \widehat{M}$ follows from (3.11) as

$$\widehat{M} = -\frac{\alpha}{1+\alpha}, \tag{3.12}$$

and the corresponding bound is

$$\mathrm{MSEB}\,(\widehat{M}, \theta) = \frac{\alpha}{1+\alpha}\theta_0^2 = \frac{1}{1+\alpha}J^{-1}(\theta_0) < \mathrm{MSEB}\,(0, \theta_0) \tag{3.13}$$

for all θ_0 such that $J^{-1}(\theta_0) > 0$. If $\hat{\theta}$ achieves the CRB, then an estimator achieving $\mathrm{MSEB}\,(\widehat{M}, \theta_0)$ can be found using (3.5), which leads to the following theorem [47].

Theorem 3.1. Let $\mathbf{x}$ denote measurements of a deterministic parameter θ_0 with pdf $p(\mathbf{x};\theta_0)$. Assume that the Fisher information with respect to θ_0 has the form $J(\theta_0) = 1/(\alpha\theta_0^2)$ for some $\alpha > 0$. Then the MSE of any estimate $\hat{\theta}$ of θ_0 with linear bias satisfies

$$E\{|\hat{\theta} - \theta_0|^2\} \geq \frac{\alpha}{1+\alpha}\theta_0^2. \tag{3.14}$$

Furthermore, if there exists an efficient estimate $\hat{\theta}$, then

$$\hat{\theta}_\mathrm{b} = \frac{1}{1+\alpha}\hat{\theta}$$

achieves the bound (3.14), and has smaller MSE than $\hat{\theta}$ for all $\theta_0 \neq 0$.

We now consider several examples illustrating Theorem 3.1.

Example 3.1. I. Suppose we are given n iid measurements $x_i, 1 \leq i \leq n$ that are each distributed uniformly on $[0, \theta_0]$, and we wish to estimate θ_0. As mentioned in Section 3.1, in this case the CRB is not defined; however, an MVU estimator exists and is given by $\hat{\theta} = (1 + 1/n)x_{\max}$, where $x_{\max} = \max_i x_i$ [146, p. 108]. Its variance, which is the minimum variance achievable with an unbiased estimator, is equal to $[1/(n(n + 2))]\theta_0^2$. Using Theorem 3.1 with the minimum variance replacing the inverse Fisher information, we conclude that the estimator

$$\hat{\theta}_{\mathrm{b}} = \frac{n+2}{n+1} x_{\max} \tag{3.15}$$

has smaller MSE for all values of θ_0. The same estimator also minimizes the MSE among invariant estimates with the property that $\hat{\theta}(cx_1, \ldots, cx_n) = c\hat{\theta}(x_1, \ldots, x_n)$ for all $c > 0$ [98].

II. Consider the problem of estimating the variance σ^2 of a Gaussian random variable with known mean μ from n iid measurements $x_i, 1 \leq i \leq n$. An efficient estimate of σ^2 achieving the unbiased CRB $J^{-1}(\theta) = 2\sigma^4/n$ is

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2. \tag{3.16}$$

In this case $\alpha = 2/n$. Therefore, from Theorem 3.1 it follows that the estimator

$$\hat{\theta}_{\mathrm{b}} = \frac{1}{n+2} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{3.17}$$

has smaller MSE than $\hat{\theta}$ for all values of $\sigma^2 > 0$. This estimate is also minimum risk scale equivariant, i.e., it minimizes the MSE among all estimates satisfying $\hat{\theta}(b\mathbf{x}) = b^2\hat{\theta}(\mathbf{x})$ [100].

If μ is unknown, then the CRB cannot be achieved. However,

$$\hat{\theta} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.18}$$

with $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$ is MVU optimal with MSE $2\sigma^4/(n-1)$. Applying Theorem 3.1 with the minimum variance replacing the Fisher

information, we conclude that

$$\hat{\theta}_{\mathrm{b}} = \frac{1}{n+1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.19}$$

has smaller MSE than $\hat{\theta}$ of (3.18) for all values of μ and σ^2. This estimate is also minimum risk equivariant [100].

III. As a final example, suppose we wish to estimate the mean θ_0 of an exponential random variable from n iid measurements $x_i, 1 \leq i \leq n$ where

$$p(x;\theta_0) = \frac{1}{\theta_0}e^{-x/\theta_0}, \quad \theta_0 \geq 0. \tag{3.20}$$

An efficient estimator is the ensemble average $\hat{\theta} = (1/n)\sum_{i=1}^{n} x_i$, whose MSE is θ_0^2/n. From Theorem 3.1, the MSE of the estimator

$$\hat{\theta}_{\mathrm{b}} = \frac{1}{n+1}\sum_{i=1}^{n} x_i \tag{3.21}$$

is $\theta_0^2/(n+1)$, which is less than the CRB for all $\theta_0 > 0$.

The results of this example are easily extended to the scale parameter for any member of the exponential family for which a scale exists. As an example, the MVU estimator of the scale for a Gamma distribution with shape parameter $\beta > 0$ is $\hat{\theta} = 1/(\beta n)\sum_{i=1}^{n} x_i$. Our results imply that the MSE of $\hat{\theta}$ can be reduced for all θ by using the modified estimator $1/(\beta n + 1)\sum_{i=1}^{n} x_i$. The relative MSE improvement over the MVU method is $1 + 1/(n\beta)$, so that the smaller the values of n and β, the more substantial the gain.

In the more general case in which $\widehat{\mathbf{M}}$ of (3.11) depends on $\boldsymbol{\theta}_0$, we can approximate it by substituting the ML estimate $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_0$, resulting in

$$\hat{\boldsymbol{\theta}}_{\mathbf{b}} = (\mathbf{I} + \widehat{\mathbf{M}}(\hat{\boldsymbol{\theta}}))\hat{\boldsymbol{\theta}} = \frac{\hat{\boldsymbol{\theta}}^T\mathbf{J}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\theta}}}{1 + \hat{\boldsymbol{\theta}}^T\mathbf{J}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\theta}}}\hat{\boldsymbol{\theta}}. \tag{3.22}$$

We may next try and further improve $\hat{\boldsymbol{\theta}}_{\mathbf{b}}$ by plugging it into $\widehat{\mathbf{M}}$ and creating a new estimate $(\mathbf{I} + \widehat{\mathbf{M}}(\hat{\boldsymbol{\theta}}_{\mathbf{b}}))\hat{\boldsymbol{\theta}}_{\mathbf{b}}$. Continuing recursively we have that at the kth iteration,

$$\hat{\boldsymbol{\theta}}_k = (\mathbf{I} + \widehat{\mathbf{M}}(\hat{\boldsymbol{\theta}}_{k-1}))\hat{\boldsymbol{\theta}}_{k-1}, \tag{3.23}$$

where $\hat{\boldsymbol{\theta}}_0$ is the ML solution. The effect of each iteration is to multiply the previous estimate by a nonlinear shrinkage factor, resulting in a nonlinear modification of ML. This strategy is studied in detail in [52] for estimating $\boldsymbol{\theta}_0$ in the linear Gaussian model (1.2). In this setting it is shown that the above iterations converge to

$$\hat{\boldsymbol{\theta}}_{\mathbf{b}} = \begin{cases} \left(1 + \sqrt{1 - \frac{4}{a(\mathbf{x})}}\right)\hat{\boldsymbol{\theta}}, & a(\mathbf{x}) > 4; \\ \mathbf{0}, & a(\mathbf{x}) \leq 4, \end{cases} \tag{3.24}$$

where $a(\mathbf{x}) = \hat{\boldsymbol{\theta}}^T \mathbf{Q} \hat{\boldsymbol{\theta}}$ with $\mathbf{Q} = \mathbf{H}^T \mathbf{C}^{-1} \mathbf{H}$ and $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathrm{LS}}$ is the LS estimate given by (1.13). Furthermore, this estimate is shown to dominate the LS solution when the effective dimension $d_{\mathrm{eff}} = \mathrm{Tr}(\mathbf{Q}^{-1})/\lambda_{\max}(\mathbf{Q}^{-1})$ satisfies $d_{\mathrm{eff}} \geq 4$.

The iterations defined above lead to a nonlinear modification of the ML estimate. In the next section, we study linear corrections, that are guaranteed to dominate ML in a general (not necessarily Gaussian) setting.

3.3 Dominating the CRB with Linear Bias

We have seen in the previous section that in some special cases we can minimize the MSE over all linear bias vectors. Even when direct minimization is not possible, we may still be able to find a matrix $\mathbf{M}$ such that the resulting MSE bound is smaller than the unbiased CRB for all possible values of the true parameter $\boldsymbol{\theta}_0$.

Before developing a general theory to improve the CRB using a linear bias, we return to the scalar case and consider an example which illustrates the main ideas [92].

3.3.1 Constant Minimum Variance

When $\boldsymbol{\theta} = \theta_0$ is a scalar, the optimal scaling $\widehat{\mathbf{M}} = \widehat{M}$ of (3.10) becomes

$$\widehat{M} = -\frac{J^{-1}(\theta_0)}{J^{-1}(\theta_0) + \theta_0^2} = -\frac{1}{1 + \theta_0^2/J^{-1}(\theta_0)}. \tag{3.25}$$

If $\theta_0^2/J^{-1}(\theta_0)$ depends upon θ_0, then exact minimization is not possible. Nonetheless, we illustrate in the ensuing sections that the MSE can still be improved uniformly for all θ_0 in many cases.

A simple scenario in which $\widehat{M}$ depends on θ_0 is when $J^{-1}(\theta_0)$ is equal to a constant, denoted V. In this situation we can no longer improve the MSE over the entire range θ_0 in $\mathbb{R}$ using a linear bias. However, as we now show, it is still possible to derive a bias such that

$$\text{MSEB}\,(\hat{\theta}_\text{b}) = (1+M)^2 V + M^2\theta_0^2 < V, \tag{3.26}$$

over a restricted parameter range. (In fact, lower MSE can be attained for all θ_0 if we allow for a nonlinear bias as we show in the next section.)

Suppose it is known that $|\theta_0| \leq U$ for some $U > 0$. Since $\text{MSEB}\,(\hat{\theta}_\text{b})$ is monotonically increasing in $|\theta_0|$, it is enough to require that (3.26) holds for $|\theta_0| = U$. Thus, we would like to choose M such that

$$\max_{|\theta|\leq U} \text{MSEB}\,(\hat{\theta}_\text{b}) = (1+M)^2 V + M^2 U^2 < V \tag{3.27}$$

or equivalently

$$\max_{|\theta|\leq U} \{\text{MSEB}\,(\hat{\theta}_\text{b}) - \text{MSEB}\,(\hat{\theta})\} = (1+M)^2 V + M^2 U^2 - V < 0, \tag{3.28}$$

where we used the fact that $\text{MSEB}\,(\hat{\theta}) = \text{var}(\hat{\theta}) = V$ is independent of θ. After some simplification, (3.28) reduces to

$$1 + M > \frac{U^2 - V}{U^2 + V}. \tag{3.29}$$

Any estimator of the form (3.8) with M satisfying (3.29) will have lower MSE than $\hat{\theta}$.

As our goal is to reduce the MSE as much as possible, we can choose M that produces the most negative value of $\max_{|\theta|\leq U}\{\text{MSEB}\,(\hat{\theta}_\text{b}) - \text{MSEB}\,(\hat{\theta})\}$. Since $\text{MSEB}\,(\hat{\theta})$ is independent of θ, this approach is equivalent to selecting the M that minimizes (3.28), resulting in

$$1 + M = \frac{U^2}{U^2 + V}, \tag{3.30}$$

which satisfies (3.29). Thus, the estimator with linear bias that minimizes the maximum MSE over $|\theta_0| \leq U$ is

$$\hat{\theta}_\text{b} = (1+M)\hat{\theta} = \frac{U^2}{U^2 + V}\hat{\theta}. \tag{3.31}$$

Interestingly, if $U \to \infty$, then the biased estimator approaches the unbiased solution. Using (3.9) and (3.30) the resulting MSE can be shown to be

$$\mathrm{MSEB}\,(\hat{\theta}_{\mathrm{b}}) = V\left[\frac{U^4 + \theta_0^2 V}{(U^2+V)^2}\right]. \tag{3.32}$$

For $|\theta_0| \leq U$ the term in brackets is less than or equal to $U^2/(U^2+V)$. Thus, a sizable reduction in the MSE results if $U^2/V \ll 1$. We next consider a specific example in which the inverse Fisher is constant and θ_0 is constrained in range.

Example 3.2. Suppose we have n iid observations $x_i, 1 \leq i \leq n$, where each x_i is a Gaussian random variable with mean θ_0 and variance σ^2. Our goal is to estimate the location parameter θ_0.

The sample mean $\hat{\theta} = \bar{x} = (1/n)\sum_{i=1}^{n} x_i$ is an efficient estimate, whose variance is the constant $V = \sigma^2/n$. If θ_0 is restricted to $|\theta_0| \leq U$, then the MSE of $\hat{\theta}$ can be improved by using the biased estimator of (3.31)

$$\hat{\theta}_{\mathrm{b}} = (1+M)\hat{\theta} = \frac{U^2}{U^2 + \sigma^2/n}\bar{x}, \tag{3.33}$$

whose MSE is given by (3.32) with $V = \sigma^2/n$. The condition for a sizable reduction in MSE becomes $U^2/(\sigma^2/n) \ll 1$, namely a short data record or low SNR.

The last example treats a scalar linear Gaussian model. Extensions to the vector case will be discussed in detail in the next section.

3.3.2 The Minimax Approach

The preceding discussion illustrates that even when the optimal $\mathbf{M}$ of (3.11) depends on $\boldsymbol{\theta}_0$, we may still be able to reduce the MSE uniformly over all allowable $\boldsymbol{\theta}_0$ by employing a minimax strategy. To state these results more generally, suppose we have an efficient estimate $\hat{\boldsymbol{\theta}}$ with MSE given by $\mathbf{J}^{-1}(\boldsymbol{\theta}_0)$ which now may depend on $\boldsymbol{\theta}_0$. Our goal is to reduce the MSE of $\hat{\boldsymbol{\theta}}$ by considering biased estimators of the form (3.5),

where we choose $\mathbf{M}$ so that

$$\mathrm{MSEB}\left(\mathbf{M}, \boldsymbol{\theta}_0\right) < \mathrm{MSEB}\left(0, \boldsymbol{\theta}_0\right) \tag{3.34}$$

for all values of $\boldsymbol{\theta}_0$ in some set $\mathcal{U}$. If the matrix $\mathbf{M}$ satisfies (3.34), then we will say that $\mathbf{M}$ (strictly) dominates [100] the CRB on $\mathcal{U}$. This will ensure that if $\hat{\boldsymbol{\theta}}$ is an efficient estimator, then the estimator $\hat{\boldsymbol{\theta}}_{\mathrm{b}} = (\mathbf{I} + \mathbf{M})\hat{\boldsymbol{\theta}}$ will have smaller MSE than $\hat{\boldsymbol{\theta}}$ for all values of $\boldsymbol{\theta}_0 \in \mathcal{U}$. In addition to satisfying (3.34), we would like $\mathbf{M}$ to have the property that there is no other matrix $\mathbf{M}' \neq \mathbf{M}$ such that

$$\mathrm{MSEB}\left(\mathbf{M}', \boldsymbol{\theta}_0\right) \leq \mathrm{MSEB}\left(\mathbf{M}, \boldsymbol{\theta}_0\right) \tag{3.35}$$

for all $\boldsymbol{\theta}_0$ in $\mathcal{U}$. Such a matrix $\mathbf{M}$ will be called admissible [100]. Our problem therefore is to find an admissible $\mathbf{M}$ that dominates the CRB on $\mathcal{U}$.

The concepts of domination and admissability are intuitively desirable. However, it is not immediately obvious how to use them as a design criterion. Fortunately, it turns out that an admissible dominating matrix can be found as a solution to a convex optimization problem, as incorporated in the following theorem.

Theorem 3.2. Let $\mathbf{x}$ denote measurements of a deterministic parameter vector $\boldsymbol{\theta}_0$ with pdf $p(\mathbf{x}; \boldsymbol{\theta}_0)$. Let

$$\mathrm{MSEB}\left(\mathbf{M}, \boldsymbol{\theta}_0\right) = \boldsymbol{\theta}_0^T \mathbf{M}^T \mathbf{M} \boldsymbol{\theta}_0 + \mathrm{Tr}\left((\mathbf{I} + \mathbf{M})\mathbf{J}^{-1}(\boldsymbol{\theta}_0)(\mathbf{I} + \mathbf{M})^T\right),$$

be a bound on the MSE of any estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ with linear bias $\mathbf{b}(\boldsymbol{\theta}_0) = \mathbf{M}\boldsymbol{\theta}_0$, where $\mathbf{J}(\boldsymbol{\theta}_0)$ is the Fisher information matrix, and let $\mathcal{U} \subseteq \mathbb{R}^m$. Define

$$\widehat{\mathbf{M}} = \arg\min_{\mathbf{M}} \sup_{\boldsymbol{\theta} \in \mathcal{U}} \left\{\mathrm{MSEB}\left(\mathbf{M}, \boldsymbol{\theta}\right) - \mathrm{MSEB}\left(0, \boldsymbol{\theta}\right)\right\}. \tag{3.36}$$

Then

(1) $\widehat{\mathbf{M}}$ is unique;
(2) $\widehat{\mathbf{M}}$ is admissible on $\mathcal{U}$;
(3) If $\widehat{\mathbf{M}} \neq 0$, then $\mathrm{MSEB}\left(\mathbf{M}, \boldsymbol{\theta}_0\right) < \mathrm{MSEB}\left(0, \boldsymbol{\theta}_0\right)$ for all $\boldsymbol{\theta}_0 \in \mathcal{U}$.

Note that the minimum in (3.36) is well defined since the objective is continuous and coercive (a function $f(\mathbf{x})$ is coercive if $f(\mathbf{x}) \to \infty$ when $\|\mathbf{x}\| \to \infty$) [16].

Proof. The proof follows directly from the proof of [43, Theorem 1] by noting that $\mathrm{MSEB}(\mathbf{M}, \boldsymbol{\theta})$ is continuous, coercive, and strictly convex in $\mathbf{M}$. □

We point out that a dominating and admissible $\mathbf{M}$ can be obtained using other objectives besides the MSE difference. For example, we can consider the ratio between $\mathrm{MSEB}(\mathbf{M}, \boldsymbol{\theta})$ and $\mathrm{MSEB}(0, \boldsymbol{\theta})$. The only properties required are that the error is continuous, coercive, and strictly convex in $\mathbf{M}$.

From Theorem 3.2 we conclude that if the solution $\widehat{\mathbf{M}}$ of (3.36) is nonzero, and if $\hat{\boldsymbol{\theta}}$ achieves the CRB, then the MSE of $\hat{\boldsymbol{\theta}}_{\mathrm{b}} = (\mathbf{I} + \widehat{\mathbf{M}})\hat{\boldsymbol{\theta}}$ is smaller than that of $\hat{\boldsymbol{\theta}}$ for all $\boldsymbol{\theta}_0 \in \mathcal{U}$; furthermore, no other estimator with linear bias exists that has a smaller (or equal) MSE than $\hat{\boldsymbol{\theta}}_{\mathrm{b}}$ for all values of $\boldsymbol{\theta}_0 \in \mathcal{U}$. Our construction also ensures that the improvement in performance is the largest possible, for the worst-case choice of $\boldsymbol{\theta}_0$.

To interpret (3.36) note that it can be written equivalently in max–min form as $\max_{\mathbf{M}} g(\mathbf{M})$ where

$$g(\mathbf{M}) = \inf_{\boldsymbol{\theta} \in \mathcal{U}} \{\mathrm{MSEB}(0, \boldsymbol{\theta}) - \mathrm{MSEB}(\mathbf{M}, \boldsymbol{\theta})\}. \tag{3.37}$$

For every value of $\boldsymbol{\theta}$, there will be a certain difference between the MSEs of the unbiased and biased estimators. Our choice of $\mathbf{M}$ is guaranteed to make the smallest difference (with respect to $\boldsymbol{\theta}$) between the MSEs as large as possible (with respect to $\mathbf{M}$). Since with $\mathbf{M} = 0$, we have $g(\mathbf{M}) = 0$, the minimal value of $g(\mathbf{M})$ must be either positive, indicating a reduction in MSE using $\hat{\boldsymbol{\theta}}_{\mathrm{b}}$, or at worst zero, indicating no change in the worst-case MSE. Since $\widehat{\mathbf{M}}$ is unique due to the strict convexity of (3.36), it follows that if $\widehat{\mathbf{M}} \neq 0$, then we are guaranteed that $g(\widehat{\mathbf{M}}) > 0$ which means that $\mathrm{MSEB}(\mathbf{M}, \boldsymbol{\theta}) < \mathbf{J}^{-1}(\boldsymbol{\theta}) = \mathrm{MSEB}(0, \boldsymbol{\theta})$ for all $\boldsymbol{\theta}$.

An important observation is that even in the absence of constraints on $\boldsymbol{\theta}_0$, a biased estimator can yield reduced MSE over an unbiased method. Concrete examples will be given in the next section.

The problem (3.36) is convex in $\mathbf{M}$ for any constraint set $\mathcal{U}$ since the supremum of a convex function over any set $\mathcal{U}$ is convex. For arbitrary forms of $\mathbf{J}^{-1}(\boldsymbol{\theta}_0)$ we can solve (3.36) by using any one of the known iterative algorithms for solving minimax problems, such as subgradient algorithms [95] or the prox method [112]. To obtain more efficient solutions, in the following sections we restrict the structure of $\mathbf{J}^{-1}(\boldsymbol{\theta}_0)$ such that the resulting optimization problem can be converted into one of the standard convex forms for which very efficient software is available.

3.4 Quadratic Inverse Fisher Information

A broad class of convex problems for which polynomial-time algorithms exists are semidefinite programs (SDPs) [113, 141] (see also the Appendix). These are optimization problems that involve minimizing a linear function subject to linear matrix inequalities (LMIs), i.e., matrix inequalities of the form $\mathbf{G}(\mathbf{M}) \succeq 0$, where $\mathbf{G}(\mathbf{M})$ is linear in $\mathbf{M}$. Once a problem is formulated as an SDP, standard software packages, such as the Self-Dual-Minimization (SeDuMi) package [133] or CVX [68], can be used to solve the problem in polynomial time within any desired accuracy. Using principles of duality theory in vector space optimization, the SDP formulation can also be used to derive necessary and sufficient optimality conditions.

It turns out that for a large class of inverse Fisher matrices, the minimax problem (3.36) can be reduced to a single minimization that takes on the form of an SDP [47]. Specifically, we consider the class of problems represented by $\mathbf{J}^{-1}(\boldsymbol{\theta}_0)$ with quadratic form:

$$\mathbf{J}^{-1}(\boldsymbol{\theta}_0) = \sum_{i=1}^{\ell} \mathbf{B}_i \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^T \mathbf{B}_i^T + \sum_{i=1}^{k} (\mathbf{C}_i \boldsymbol{\theta}_0 \mathbf{z}_i^T + \mathbf{z}_i \boldsymbol{\theta}_0^T \mathbf{C}_i^T) + \mathbf{A}, \qquad (3.38)$$

for some matrices $\mathbf{A} \succeq 0$, $\mathbf{B}_i, \mathbf{C}_i$, and vectors $\mathbf{z}_i$. (Alternatively, when considering MVU estimators, we assume that the minimum variance has the form (3.38)). Besides leading to analytically tractable expressions, there are many cases in which the inverse Fisher information can be written as in (3.38). Several examples are presented below.

Example 3.3. I. A simple example is estimating the vector $\boldsymbol{\theta}_0$ in the linear Gaussian model (1.2). In this case the inverse Fisher information matrix is the constant $\mathbf{J}^{-1} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}$ which is clearly a special case of (3.38) with

$$\mathbf{A} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}, \quad \mathbf{C}_i = \mathbf{0}, \quad \mathbf{z}_i = \mathbf{0}, \quad \mathbf{B}_i = \mathbf{0}. \tag{3.39}$$

This setting will be discussed in Section 4.

II. Consider the problem of estimating the mean μ and variance σ^2 of a Gaussian random variable from n iid measurements. In this case $\boldsymbol{\theta}_0 = (\mu \quad \sigma^2)^T$, and

$$\mathbf{J}^{-1}(\boldsymbol{\theta}_0) = \frac{\sigma^2}{n}\begin{bmatrix} 1 & 0 \\ 0 & 2\sigma^2 \end{bmatrix}, \tag{3.40}$$

which has the form (3.38) with $\ell = 1$, $k = 1$,

$$\mathbf{A} = 0, \quad \mathbf{C}_1 = \frac{1}{n}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{z}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{B}_1 = \sqrt{\frac{2}{n}}\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3.41}$$

For known μ, the inverse Fisher information with respect to σ^2 is $J^{-1}(\sigma^2) = 2\sigma^4/n$, in which case $\ell = 1$, $B_1 = \sqrt{2/n}$ and all the remaining parameters are equal to 0.

If σ^2 is known and μ is Gaussian with zero mean and unknown variance σ_μ^2, then

$$J^{-1}(\sigma_\mu^2) = 2\left(\sigma_\mu^2 + \frac{\sigma^2}{n}\right)^2, \tag{3.42}$$

so that now $\ell = 1, k = 1$,

$$\mathbf{A} = 2\frac{\sigma^4}{N^2}, \quad \mathbf{C}_1 = 4\frac{\sigma^2}{n}, \quad \mathbf{z}_1 = 1, \quad \mathbf{B}_1 = \sqrt{2}. \tag{3.43}$$

III. As another example, suppose that μ is an unknown scalar in additive white Gaussian noise with unknown variance σ^2. The inverse Fisher information for estimating the SNR $\theta_0 = \mu^2/\sigma^2$ is

$$J^{-1}(\theta_0) = \frac{1}{n}(4\theta_0 + 2\theta_0^2), \tag{3.44}$$

which has the form (3.38) with $\ell = 1, k = 1$,

$$\mathbf{A} = 0, \quad \mathbf{C}_1 = \frac{4}{n}, \quad \mathbf{z}_1 = 1, \quad \mathbf{B}_1 = \sqrt{\frac{2}{n}}. \tag{3.45}$$

We will consider this example in detail in Section 3.6.

IV. As a final example, suppose that $\mathbf{x}$ is a vector of counts with mean $\mathbf{g}(\boldsymbol{\theta}_0)$ where

$$\mathbf{g}(\boldsymbol{\theta}_0) = \mathbf{H}\boldsymbol{\theta}_0 + \mathbf{c} \tag{3.46}$$

for some known invertible matrix $\mathbf{H}$ and known constant vector $\mathbf{c}$. The elements x_i of $\mathbf{x}$ are independent, with a Poisson distribution

$$\ln p(x_i; \boldsymbol{\theta}_0) = x_i \ln(g_i(\boldsymbol{\theta}_0)) - g_i(\boldsymbol{\theta}_0) + a,$$

where a is a known constant. This problem arises for example in emission-computed tomography [126]. The inverse Fisher information in this case is given by [79]

$$\mathbf{J}^{-1}(\boldsymbol{\theta}_0) = \mathbf{H}^{-1} \operatorname{diag}(g_1(\boldsymbol{\theta}_0), \ldots, g_m(\boldsymbol{\theta}_0)) \mathbf{H}^{-T}, \tag{3.47}$$

where $\mathbf{H}^{-T} = (\mathbf{H}^{-1})^T$. We can express $\mathbf{J}^{-1}(\boldsymbol{\theta}_0)$ of (3.47) in the form (3.38) with $\ell = 0, k = m$

$$\begin{aligned} \mathbf{A} &= \mathbf{H}^{-1} \operatorname{diag}(c_1, \ldots, c_m) \mathbf{H}^{-T}, \\ \mathbf{C}_i &= \mathbf{H}^{-1} \mathbf{E}_i([\mathbf{H}]_i), \quad \mathbf{z}_i = [\mathbf{H}^{-T}]_i, \quad 1 \le i \le m, \end{aligned} \tag{3.48}$$

where $[\mathbf{H}]_i$ denotes the ith row of $\mathbf{H}$ and $\mathbf{E}_i(\mathbf{d})$ is the matrix whose ith row is equal to $\mathbf{d}$, and whose remaining elements are equal to zero.

In Section 3.4.1 we treat the case in which $\mathcal{U} = \mathbb{R}^m$ so that $\boldsymbol{\theta}_0$ is not restricted, and show that with $\mathbf{J}^{-1}(\boldsymbol{\theta}_0)$ given by (3.38), the optimal $\mathbf{M}$ can be found as a solution to an SDP. We also develop necessary and sufficient optimality conditions on $\mathbf{M}$ that lead to further insights into the solution.

In some settings, we may have additional information on the parameter vector $\boldsymbol{\theta}_0$ which can result in a lower MSE bound. The set $\mathcal{U}$ is then chosen to capture these properties of $\boldsymbol{\theta}_0$. For example, we may know that the norm of $\boldsymbol{\theta}_0$ is bounded: $\boldsymbol{\theta}_0^T \boldsymbol{\theta}_0 \le U$ for some $U > 0$. There are

also examples where there are natural restrictions on the parameters, for example if θ_0 represents the variance or the SNR of a random variable, then $\theta_0 > 0$. More generally, θ_0 may lie in a specified interval $\alpha \leq \theta_0 \leq \beta$. These constraints can all be viewed as special cases of the quadratic constraint $\boldsymbol{\theta}_0 \in \mathcal{Q}$ where

$$\mathcal{Q} = \{\boldsymbol{\theta} | \boldsymbol{\theta}^T \mathbf{A}_1 \boldsymbol{\theta} + 2\mathbf{b}_1^T \boldsymbol{\theta} + c_1 \leq 0\}, \tag{3.49}$$

for some $\mathbf{A}_1, \mathbf{b}_1$, and c_1. Note that we do not require that $\mathbf{A}_1 \succeq 0$ so that the constraint set (3.49) is not necessarily convex. In Section 3.4.4, we discuss the scenario in which $\boldsymbol{\theta}_0 \in \mathcal{Q}$, and show that again an admissible dominating $\mathbf{M}$ can be found by solving an SDP. Using the results of [7], the ideas we develop can also be generalized to the case of two quadratic constraints of the form $\mathcal{Q}$.

Before proceeding to the detailed developments, it is important to note, that even in cases where $\mathbf{M}$ is computed numerically via an SDP, i.e., a closed form solution does not exist, the calculation of $\mathbf{M}$ does not depend on the data $\mathbf{x}$. Therefore, $\mathbf{M}$ can be computed off line. Once the data is received, to implement the proposed estimator all that is needed is to multiply the unbiased method by the matrix $\mathbf{I} + \mathbf{M}$ so that the additional cost incurred is negligible.

3.4.1 Dominating Bound on the Entire Space

We first treat the case in which $\mathcal{U} = \mathbb{R}^m$ so that $\boldsymbol{\theta}_0$ is not restricted and show how to solve (3.36) with $\mathbf{J}^{-1}(\boldsymbol{\theta}_0)$ given by (3.38).

Defining,

$$\begin{aligned}
\mathbf{A}_0(\mathbf{M}) &= \mathbf{M}^T\mathbf{M} + \sum_{i=1}^{\ell} \mathbf{B}_i^T \left((\mathbf{I}+\mathbf{M})^T(\mathbf{I}+\mathbf{M}) - \mathbf{I}\right)\mathbf{B}_i; \\
\mathbf{b}_0(\mathbf{M}) &= \sum_{i=1}^{k} \mathbf{C}_i^T \left((\mathbf{I}+\mathbf{M})^T(\mathbf{I}+\mathbf{M}) - \mathbf{I}\right)\mathbf{z}_i; \\
c_0(\mathbf{M}) &= \mathrm{Tr}\left(\left((\mathbf{I}+\mathbf{M})^T(\mathbf{I}+\mathbf{M}) - \mathbf{I}\right)\mathbf{A}\right),
\end{aligned} \tag{3.50}$$

we can write (3.36) as

$$\min_{\mathbf{M}} \max_{\boldsymbol{\theta}} \left\{\boldsymbol{\theta}^T \mathbf{A}_0(\mathbf{M})\boldsymbol{\theta} + 2\mathbf{b}_0^T(\mathbf{M})\boldsymbol{\theta} + c_0(\mathbf{M})\right\}. \tag{3.51}$$

The problem (3.51) is equivalent to

$$\begin{array}{ll} \min\limits_{t,\mathbf{M}} & t \\ \text{s.t.} & \boldsymbol{\theta}^T \mathbf{A}_0(\mathbf{M})\boldsymbol{\theta} + 2\mathbf{b}_0^T(\mathbf{M})\boldsymbol{\theta} + c_0(\mathbf{M}) \le t, \quad \text{for all } \boldsymbol{\theta}. \end{array} \tag{3.52}$$

Using [12, p. 163], we can rewrite (3.52) as

$$\begin{array}{ll} \min\limits_{t,\mathbf{M}} & t \\ \text{s.t.} & \mathbf{G}(\mathbf{M}) \triangleq \begin{bmatrix} \mathbf{A}_0(\mathbf{M}) & \mathbf{b}_0(\mathbf{M}) \\ \mathbf{b}_0^T(\mathbf{M}) & c_0(\mathbf{M}) - t \end{bmatrix} \preceq 0. \end{array} \tag{3.53}$$

Since the choice of parameters $\mathbf{M} = 0$, $t = 0$ satisfies the constraint (3.53), the problem is always feasible. In our development below, we consider the case in which (3.53) is strictly feasible, i.e., there exists a matrix $\mathbf{M}$ such that $\mathbf{G}(\mathbf{M}) \prec 0$. It can be shown that strict feasibility is equivalent to $\sum_{i=1}^{\ell} \mathbf{B}_i^T \mathbf{B}_i \succ 0$ [47, Lemma 1]. If (3.53) is not strictly feasible then as shown in [47, Appendix II], it can always be reduced to a strictly feasible problem with additional linear constraints on $\mathbf{M}$. A similar approach to that taken here can then be followed for the reduced formulation. Therefore, in the remainder of this section we assume that (3.53) is strictly feasible.

3.4.2 SDP Formulation

The constraint (3.53) is not written in convex form, so that we cannot directly apply standard convex algorithms or Lagrange duality theory to find the optimal $\mathbf{M}$. Fortunately, it can be converted into a convex constraint, leading to a convex formulation of (3.53). This result is incorporated in the following proposition [47, Lemma 2].

Proposition 3.3. Consider the setting of Theorem 3.2 with $\mathcal{U} = \mathbb{R}^m$ and $\mathbf{J}^{-1}(\boldsymbol{\theta}_0)$ given by (3.38). Then $\widehat{\mathbf{M}}$ is the solution to the SDP

$$\begin{array}{ll} \min\limits_{t,\mathbf{M},\mathbf{X}} & t \\ \text{s.t.} & \mathbf{Z}(\mathbf{M},\mathbf{X}) \preceq 0 \\ & \begin{bmatrix} \mathbf{X} & \mathbf{M}^T \\ \mathbf{M} & \mathbf{I} \end{bmatrix} \succeq 0, \end{array} \tag{3.54}$$

where

$$\mathbf{Z}(\mathbf{M},\mathbf{X}) = \begin{bmatrix} \mathbf{X} + \sum_{i=1}^{\ell} \mathbf{B}_i^T \Phi \mathbf{B}_i & \sum_{i=1}^{k} \mathbf{C}_i^T \Phi \mathbf{z}_i \\ \sum_{i=1}^{k} \mathbf{z}_i^T \Phi \mathbf{C}_i & \operatorname{Tr}(\mathbf{A}\Phi) - t \end{bmatrix}, \tag{3.55}$$

and for brevity we denoted $\Phi = \mathbf{X} + \mathbf{M} + \mathbf{M}^T$.

Note that the constraints in (3.54) are LMIs, since the unknown matrices appear linearly. Therefore, (3.54) is indeed an SDP and can be solved using standard optimization software.

3.4.3 Dual Problem

Once we have formulated our problem as an SDP we can use duality theory to gain more insight into the form of the optimal $\mathbf{M}$, and to provide an alternative method of solution.

Since (3.54) is convex and strictly feasible, the optimal value of t is equal to the optimal value of the dual problem. Direct calculation shows that the dual is

$$\begin{array}{ll} \min\limits_{\mathbf{w},\Pi} & \operatorname{Tr}\big(\mathbf{S}(\Pi,\mathbf{w})\left(\mathbf{S}(\Pi,\mathbf{w}) + \Pi\right)^{-1}\mathbf{S}(\Pi,\mathbf{w})\big) \\ \text{s.t.} & \begin{bmatrix} \Pi & \mathbf{w} \\ \mathbf{w}^T & 1 \end{bmatrix} \succeq 0, \end{array} \tag{3.56}$$

where

$$\mathbf{S}(\Pi,\mathbf{w}) = \sum_{i=1}^{\ell} \mathbf{B}_i \Pi \mathbf{B}_i^T + \sum_{i=1}^{k} (\mathbf{z}_i \mathbf{w}^T \mathbf{C}_i^T + \mathbf{C}_i \mathbf{w} \mathbf{z}_i^T) + \mathbf{A}. \tag{3.57}$$

The optimal matrix $\mathbf{M}$ is related to the optimal dual variables via

$$\mathbf{M} = -\mathbf{S}(\Pi,\mathbf{w})\left(\mathbf{S}(\Pi,\mathbf{w}) + \Pi\right)^{-1}. \tag{3.58}$$

Note that $\mathbf{S}(\Pi,\mathbf{w})$ is guaranteed to be invertible [47]. Using Schur's lemma (see Lemma A.3 in the Appendix), (3.56) can be written as

$$\begin{array}{ll} \min\limits_{\mathbf{Y},\mathbf{w},\Pi} & \operatorname{Tr}(\mathbf{Y}) \\ \text{s.t.} & \begin{bmatrix} \mathbf{Y} & \mathbf{S}(\Pi,\mathbf{w}) \\ \mathbf{S}(\Pi,\mathbf{w}) & \mathbf{S}(\Pi,\mathbf{w}) + \Pi \end{bmatrix} \succeq 0 \\ & \begin{bmatrix} \Pi & \mathbf{w} \\ \mathbf{w}^T & 1 \end{bmatrix} \succeq 0, \end{array}$$

which is again an SDP.

As illustrated in the following example, in some cases the dual problem may admit a closed form solution, leading to an explicit expression for $\mathbf{M}$ via (3.58).

Example 3.4. Suppose that $\boldsymbol{\theta}_0 = \theta_0$ is a scalar and $J^{-1}(\theta_0) = a + b^2\theta_0^2$ with $a > 0$. The dual problem (3.59) becomes

$$\min_{\pi \geq 0} \frac{(a + b^2\pi)^2}{a + (b^2 + 1)\pi}. \tag{3.59}$$

The optimal solution can be shown to be

$$\pi = \max\left(\frac{a(1 - b^2)}{b^2(b^2 + 1)}, 0\right), \tag{3.60}$$

leading to

$$\widehat{M} = \max\left(-\frac{2b^2}{b^2 + 1}, -1\right). \tag{3.61}$$

Therefore, if $\hat{\theta}_0$ achieves the CRB $J^{-1}(\theta_0) = a + b^2\theta_0^2$, then the estimator

$$\hat{\theta}_{\mathrm{b}} = \begin{cases} \dfrac{1 - b^2}{1 + b^2}\hat{\theta}_0, & |b| \leq 1; \\ 0, & |b| \geq 1 \end{cases} \tag{3.62}$$

achieves the MSE

$$\mathrm{MSEB}\left(\widehat{M}, \theta_0\right) = \begin{cases} a\dfrac{(1 - b^2)^2}{(1 + b^2)^2} + b^2\theta_0^2, & |b| \leq 1; \\ \theta_0^2, & |b| \geq 1, \end{cases} \tag{3.63}$$

which is smaller than $J^{-1}(\theta_0)$ for all θ_0.

The last example illustrates that the CRB can be improved uniformly even without any prior knowledge on $\boldsymbol{\theta}_0$.

Using the KKT conditions it can be shown that the matrix $\mathbf{M}$ is optimal if and only if there exists a matrix Π and a vector $\mathbf{w}$ such that $\Pi \succeq \mathbf{w}\mathbf{w}^T$ and the following conditions hold:

$$\begin{gathered} \mathbf{M} = -\mathbf{S}(\Pi, \mathbf{w})\left(\mathbf{S}(\Pi, \mathbf{w}) + \Pi\right)^{-1}; \\ \begin{bmatrix} \mathbf{A}_0(\mathbf{M}) & \mathbf{b}_0(\mathbf{M}) \\ \mathbf{b}_0^T(\mathbf{M}) & c_0(\mathbf{M}) - \mathrm{Tr}\left(\mathbf{M}\mathbf{S}(\Pi, \mathbf{w})\right) \end{bmatrix} \preceq 0, \end{gathered} \tag{3.64}$$

where $\mathbf{A}_0(\mathbf{M}), \mathbf{b}_0(\mathbf{M}), c_0(\mathbf{M})$ are defined by (3.50), and $\mathbf{S}(\Pi, \mathbf{w})$ is given by (3.57).

An important observation from (3.64) is that regardless of Π, the optimal $\mathbf{M}$ is not equal to 0. Therefore, from Theorem 3.2 it follows that as long as the problem is strictly feasible, we can improve the CRB for all values of $\boldsymbol{\theta}_0$ by a linear transformation. Moreover, it is proven in [47] that $\mathbf{M}$ is also nonzero when the problem is not strictly feasible, as long as $\mathbf{B}_i \neq 0$ for some i. We therefore have the following proposition.

Proposition 3.4. Consider the setting of Theorem 3.2 with $\mathcal{U} = \mathbb{R}^m$ and $\mathbf{J}^{-1}(\boldsymbol{\theta}_0)$ of (3.38). Then $\widehat{\mathbf{M}} = 0$ if and only if $\mathbf{B}_i = 0$ for all i.

An immediate consequence of Proposition 3.4 is that when $\mathbf{J}^{-1}(\boldsymbol{\theta}_0)$ is a constant, as in the linear Gaussian model, the CRB cannot be improved upon uniformly for all $\boldsymbol{\theta}_0$ using a linear bias. In the next section we will see that when the parameter values are restricted, uniform improvement is possible. On the other hand, if we allow for a nonlinear modification, then even in the linear Gaussian case we can improve the CRB and MVU estimation uniformly over the entire space (as long as the effective dimension is large enough). This will be proven in Sections 4 and 5.

3.4.4 Dominating Bound on a Quadratic Set

We now treat the scenario in which $\boldsymbol{\theta}_0$ is restricted to the quadratic set $\mathcal{Q}$ of (3.49). To find an admissible dominating matrix in this case we need to solve the minimax problem

$$\min_{\mathbf{M}} \max_{\boldsymbol{\theta} \in \mathcal{Q}} \{\mathrm{MSEB}\,(\mathbf{M}, \boldsymbol{\theta}) - \mathrm{MSEB}\,(0, \boldsymbol{\theta})\}. \tag{3.65}$$

We assume that there exists a $\boldsymbol{\theta}$ in the interior of $\mathcal{Q}$. However, we do not impose any further restrictions on the parameters $\mathbf{A}_1, \mathbf{b}_1$, and c_1.

The optimal $\mathbf{M}$ can be found following similar steps as those used in the previous section. The main difference is that now the inner maximization in (3.51) needs to be solved over a quadratic set. Omitting

the dependence on $\mathbf{M}$, the resulting problem is

$$\begin{array}{ll} \max\limits_{\boldsymbol{\theta}} & \boldsymbol{\theta}^T \mathbf{A}_0 \boldsymbol{\theta} + 2\mathbf{b}_0^T \boldsymbol{\theta} + c_0 \\ \text{s.t.} & \boldsymbol{\theta}^T \mathbf{A}_1 \boldsymbol{\theta} + 2\mathbf{b}_1^T \boldsymbol{\theta} + c_1 \leq 0. \end{array} \tag{3.66}$$

This is a special case of a trust region problem, for which strong duality holds (assuming that there is a strictly feasible point) even though the problem is not convex [20]. Using this strong duality result we can convert the problem into

$$\begin{array}{ll} \min\limits_{\lambda \geq 0, t, \mathbf{M}} & t \\ \text{s.t.} & \begin{bmatrix} \lambda \mathbf{A}_1 & \lambda \mathbf{b}_1 \\ \lambda \mathbf{b}_1^T & \lambda c_1 + t \end{bmatrix} \succeq \begin{bmatrix} \mathbf{A}_0(\mathbf{M}) & \mathbf{b}_0(\mathbf{M}) \\ \mathbf{b}_0^T(\mathbf{M}) & c_0(\mathbf{M}) \end{bmatrix}. \end{array} \tag{3.67}$$

The detailed derivation is given [47]. It is easy to see that (3.67) is always feasible, since both matrices in (3.67) can be made equal to 0 by choosing $\mathbf{M} = 0$, and $\lambda = t = 0$. A necessary and sufficient condition for strict feasibility is [40]

$$\sum_{i=1}^{\ell} \mathbf{B}_i^T \mathbf{B}_i \mathbf{v} = 0, \quad \mathbf{v} \neq 0 \Rightarrow \mathbf{v}^T \mathbf{A}_1 \mathbf{v} > 0. \tag{3.68}$$

In particular, (3.67) is strictly feasible if $\sum_{i=1}^{\ell} \mathbf{B}_i^T \mathbf{B}_i \succ 0$ or $\mathbf{A}_1 \succ 0$.

Assuming strict feasibility, (3.67) can be converted into an SDP, in a similar way to Proposition 3.3 [47, Lemma 3]:

Proposition 3.5. Consider the setting of Theorem 3.2 with $\mathcal{U} = \mathcal{Q}$ of (3.49) and $\mathbf{J}^{-1}(\boldsymbol{\theta}_0)$ given by (3.38). Then $\widehat{\mathbf{M}}$ is the solution to the SDP

$$\begin{array}{ll} \min\limits_{t, \lambda \geq 0, \mathbf{M}, \mathbf{X}} & t \\ \text{s.t.} & \mathbf{Z}(\mathbf{M}, \mathbf{X}) \preceq \lambda \mathbf{F} \\ & \begin{bmatrix} \mathbf{X} & \mathbf{M}^T \\ \mathbf{M} & \mathbf{I} \end{bmatrix} \succeq 0, \end{array} \tag{3.69}$$

where $\mathbf{Z}(\mathbf{M}, \mathbf{X})$ is defined in (3.55), and

$$\mathbf{F} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{b}_1 \\ \mathbf{b}_1^T & c_1 \end{bmatrix}. \tag{3.70}$$

Using Lagrange duality it can be shown that the optimal matrix $\mathbf{M}$ is given by (3.58), where Π and $\mathbf{w}$ are the solution to the dual problem

$$\begin{aligned} \min_{\mathbf{Y},\mathbf{w},\Pi} \quad & \operatorname{Tr}(\mathbf{Y}) \\ \text{s.t.} \quad & \begin{bmatrix} \mathbf{Y} & \mathbf{S}(\Pi,\mathbf{w}) \\ \mathbf{S}(\Pi,\mathbf{w}) & \mathbf{S}(\Pi,\mathbf{w})+\Pi \end{bmatrix} \succeq 0 \\ & \begin{bmatrix} \Pi & \mathbf{w} \\ \mathbf{w}^T & 1 \end{bmatrix} \succeq 0 \\ & \operatorname{Tr}(\Pi\mathbf{A}_1) + 2\mathbf{w}^T\mathbf{b}_1 + c_1 \leq 0, \end{aligned} \tag{3.71}$$

which is again an SDP.

Finally, from the KKT conditions we conclude that $\mathbf{M}$ is optimal if and only if there exists a matrix Π and a vector $\mathbf{w}$ such that $\Pi \succeq \mathbf{w}\mathbf{w}^T$ and the following conditions hold:

$$\begin{aligned} \mathbf{M} &= -\mathbf{S}(\Pi,\mathbf{w})\left(\mathbf{S}(\Pi,\mathbf{w})+\Pi\right)^{-1}; \\ \operatorname{Tr}(\Pi\mathbf{A}_1) + 2\mathbf{w}^T\mathbf{b}_1 + c_1 &\leq 0; \\ \lambda\left(\operatorname{Tr}(\Pi\mathbf{A}_1) + 2\mathbf{w}^T\mathbf{b}_1 + c_1\right) &= 0; \\ \begin{bmatrix} \mathbf{A}_0(\mathbf{M}) & \mathbf{b}_0(\mathbf{M}) \\ \mathbf{b}_0^T(\mathbf{M}) & c_0(\mathbf{M}) - \operatorname{Tr}\left(\mathbf{M}\mathbf{S}(\Pi,\mathbf{w})\right) \end{bmatrix} &\preceq \lambda \begin{bmatrix} \mathbf{A}_1 & \mathbf{b}_1 \\ \mathbf{b}_1^T & c_1 \end{bmatrix}, \end{aligned} \tag{3.72}$$

where $\mathbf{A}_0(\mathbf{M}), \mathbf{b}_0(\mathbf{M}), c_0(\mathbf{M})$ are defined by (3.50), and $\mathbf{S}(\Pi,\mathbf{w})$ is given by (3.57).

As in the unconstrained case, regardless of Π, $\mathbf{M}$ of (3.72) is not equal to 0. Therefore, from Theorem 3.2 it follows that as long as the problem is strictly feasible, we can improve the CRB for all values of $\boldsymbol{\theta} \in \mathcal{Q}$ by a linear transformation.

3.4.5 Constant Fisher Matrix

A special case of (3.38) is when $\mathbf{J}(\boldsymbol{\theta}_0) = \mathbf{A}^{-1}$ is a constant matrix. Examples include the linear Gaussian model, and location estimation.

In a location estimation problem we are given measurements

$$\mathbf{x} = \boldsymbol{\theta}_0 + \mathbf{w}, \tag{3.73}$$

where $\mathbf{w}$ is a random vector with pdf $p_w(\mathbf{w}) > 0$ that is supported on the entire space. The pdf of $\mathbf{x}$ is then $p(\mathbf{x};\boldsymbol{\theta}_0) = p_w(\mathbf{x} - \boldsymbol{\theta}_0)$. Now,

$$\frac{d \ln p(\mathbf{x};\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}_0} = -\frac{1}{p_w(\mathbf{x} - \boldsymbol{\theta}_0)} \frac{dp_w(\mathbf{x} - \boldsymbol{\theta}_0)}{d\mathbf{w}}, \tag{3.74}$$

where $dp_w(\mathbf{x} - \boldsymbol{\theta}_0)/d\mathbf{w}$ is the derivative of $p_w(\mathbf{w})$ with respect to $\mathbf{w}$ evaluated at $\mathbf{x} - \boldsymbol{\theta}_0$. From (3.74), the Fisher information is

$$\begin{aligned} \mathbf{J}(\boldsymbol{\theta}_0) &= \int \frac{1}{p_w(\mathbf{x} - \boldsymbol{\theta}_0)} \left[\frac{dp_w(\mathbf{x} - \boldsymbol{\theta}_0)}{d\mathbf{w}}\right]^T \left[\frac{dp_w(\mathbf{x} - \boldsymbol{\theta}_0)}{d\mathbf{w}}\right] d\mathbf{x} \\ &= \int \frac{1}{p_w(\mathbf{z})} \left[\frac{dp_w(\mathbf{z})}{d\mathbf{w}}\right]^T \left[\frac{dp_w(\mathbf{z})}{d\mathbf{w}}\right] d\mathbf{z}, \end{aligned} \tag{3.75}$$

where we used the change of variables $\mathbf{z} = \mathbf{x} - \boldsymbol{\theta}_0$ and the last equality follows from the fact that the integral boundaries do not depend on $\boldsymbol{\theta}_0$. It is evident from (3.75) that $\mathbf{J}(\boldsymbol{\theta}_0)$ is independent of $\boldsymbol{\theta}_0$.

From Proposition 3.4 it follows that the CRB cannot be improved upon in this case over the entire space $\mathbb{R}^m$ using a *linear* modification. However, if $\boldsymbol{\theta}_0$ is restricted to a quadratic set, then the CRB can be reduced, as incorporated in the following proposition.

Proposition 3.6. Let $\mathbf{x}$ denote measurements of a deterministic parameter vector $\boldsymbol{\theta}_0$ with pdf $p(\mathbf{x};\boldsymbol{\theta}_0)$. Assume that the Fisher information with respect to $\boldsymbol{\theta}_0$ has the form $\mathbf{J}(\boldsymbol{\theta}_0) = \mathbf{A}^{-1}$, and that $\|\boldsymbol{\theta}_0\|^2 \leq c$. If there exists an efficient estimator $\hat{\boldsymbol{\theta}}$, then

$$\hat{\boldsymbol{\theta}}_{\mathrm{b}} = \frac{c}{\mathrm{Tr}(\mathbf{A}) + c} \hat{\boldsymbol{\theta}}$$

has smaller MSE than $\hat{\boldsymbol{\theta}}$ for all $\|\boldsymbol{\theta}_0\|^2 \leq c$.

Proof. The proof follows from showing that $\mathbf{M} = -\mathrm{Tr}(\mathbf{A})/(\mathrm{Tr}(\mathbf{A}) + c)$ satisfies the optimality conditions (3.72) with

$$\Pi = \frac{c}{\mathrm{Tr}(\mathbf{A})}\mathbf{A}, \quad t = -\frac{\mathrm{Tr}^2(\mathbf{A})}{\mathrm{Tr}(\mathbf{A}) + c}, \quad \lambda = \frac{\mathrm{Tr}^2(\mathbf{A})}{(\mathrm{Tr}(\mathbf{A}) + c)^2}. \tag{3.76}$$

This can be established by direct substitution. □

The estimator $\hat{\boldsymbol{\theta}}_{\rm b}$ of Proposition 3.6 is a shrinkage method, i.e., a constant multiple of the unbiased solution $\hat{\boldsymbol{\theta}}$. Estimators of this type have been used extensively in the literature [11, 18, 51, 103, 107] following the seminal work of James and Stein [88].

Closed form expressions for $\hat{\boldsymbol{\theta}}_{\rm b}$ can also be obtained in the case of a weighted norm constraint of the form $\boldsymbol{\theta}_0^T \mathbf{T} \boldsymbol{\theta}_0 \leq c$ for certain choices of $\mathbf{T} \succ 0$ using similar techniques as those used in [43]. These results will be illustrated in the context of the linear Gaussian model (1.2) in the next section.

3.4.6 Feasibility

In general, the minimax estimate $\hat{\boldsymbol{\theta}}_{\rm b} = (\mathbf{I} + \widehat{\mathbf{M}})\hat{\boldsymbol{\theta}}$ obtained over the set $\mathcal{Q}$, does not always lie in $\mathcal{Q}$. This is because we are constrained to a linear modification. A simple way to render the resulting estimate feasible so that it is in $\mathcal{Q}$, when $\mathcal{Q}$ is a convex set, is to project it onto the set. Thus, instead of $\hat{\boldsymbol{\theta}}_{\rm b}$ we may use $\hat{\boldsymbol{\theta}}_{\rm p}$ which is the solution to

$$\min_{\hat{\boldsymbol{\theta}}_{\rm p} \in \mathcal{Q}} \|\hat{\boldsymbol{\theta}}_{\rm b} - \hat{\boldsymbol{\theta}}_{\rm p}\|^2. \tag{3.77}$$

An important observation is that if $\hat{\boldsymbol{\theta}}_{\rm b}$ dominates $\hat{\boldsymbol{\theta}}$, then we are guaranteed that $\hat{\boldsymbol{\theta}}_{\rm p}$ will as well. To see this, note that for any $\boldsymbol{\theta}$,

$$\|\hat{\boldsymbol{\theta}}_{\rm b} - \boldsymbol{\theta}\|^2 = \|\hat{\boldsymbol{\theta}}_{\rm b} - \hat{\boldsymbol{\theta}}_{\rm p}\|^2 + \|\hat{\boldsymbol{\theta}}_{\rm p} - \boldsymbol{\theta}\|^2 + 2(\hat{\boldsymbol{\theta}}_{\rm b} - \hat{\boldsymbol{\theta}}_{\rm p})^T(\hat{\boldsymbol{\theta}}_{\rm p} - \boldsymbol{\theta}). \tag{3.78}$$

Now, by the projection theorem onto convex sets [16] we have that for any $\boldsymbol{\theta} \in \mathcal{Q}$,

$$(\hat{\boldsymbol{\theta}}_{\rm b} - \hat{\boldsymbol{\theta}}_{\rm p})^T(\hat{\boldsymbol{\theta}}_{\rm p} - \boldsymbol{\theta}) \geq 0. \tag{3.79}$$

Combining (3.78) and (3.79),

$$\|\hat{\boldsymbol{\theta}}_{\rm p} - \boldsymbol{\theta}\|^2 \leq \|\hat{\boldsymbol{\theta}}_{\rm b} - \boldsymbol{\theta}\|^2, \quad \text{for all } \boldsymbol{\theta} \in \mathcal{Q}, \tag{3.80}$$

which immediately implies that

$$E\{\|\hat{\boldsymbol{\theta}}_{\rm p} - \boldsymbol{\theta}\|^2\} \leq E\{\|\hat{\boldsymbol{\theta}}_{\rm b} - \boldsymbol{\theta}\|^2\}, \quad \text{for all } \boldsymbol{\theta} \in \mathcal{Q}. \tag{3.81}$$

Therefore, the projection can only decrease the MSE, with respect to $\hat{\boldsymbol{\theta}}_{\rm b}$.

3.5 Affine Bias

Until now we considered linear bias vectors. An extension that can further reduce the MSE is to allow for an *affine* bias of the form $\mathbf{M}\boldsymbol{\theta}_0 + \mathbf{u}$ rather than just a linear term [40]. In this case the biased estimator will take on the form:

$$\hat{\boldsymbol{\theta}}_{\mathrm{b}} = (\mathbf{I} + \mathbf{M})\hat{\boldsymbol{\theta}} + \mathbf{u}, \tag{3.82}$$

and the MSE bound becomes

$$\begin{aligned}\mathrm{MSEB}\,(\mathbf{M}, \mathbf{u}, \boldsymbol{\theta}_0) \;=\; & (\mathbf{M}\boldsymbol{\theta}_0 + \mathbf{u})^T(\mathbf{M}\boldsymbol{\theta}_0 + \mathbf{u}) \\ & + \mathrm{Tr}((\mathbf{I} + \mathbf{M})\mathbf{J}^{-1}(\boldsymbol{\theta}_0)(\mathbf{I} + \mathbf{M})^T).\end{aligned} \tag{3.83}$$

Note that although the constant part of the bias $\mathbf{u}$ will not affect the variance, it does influence the MSE through the bias.

To find an admissible and dominating pair $(\mathbf{M}, \mathbf{u})$ we can optimize the MSE difference (3.36) over both $\mathbf{M}$ and $\mathbf{u}$:

$$(\widehat{\mathbf{M}}, \hat{\mathbf{u}}) = \arg\min_{\mathbf{M},\mathbf{u}} \sup_{\boldsymbol{\theta}\in\mathcal{U}} \left\{\mathrm{MSEB}\,(\mathbf{M}, \mathbf{u}, \boldsymbol{\theta}) - \mathrm{MSEB}\,(0, 0, \boldsymbol{\theta})\right\}. \tag{3.84}$$

In [40] it is shown that often allowing for an affine bias can substantially improve the performance. Furthermore, when the inverse Fisher contains a linear term, or the constraint set is not symmetric around 0, including an affine bias leads to bounds that are intuitively more appealing.

The development of the optimal choice of $\mathbf{M}$ and $\mathbf{u}$ is similar to that of the optimal linear bias. The details can be found in [40]. Some examples of the use of an affine biased estimator are given next.

Example 3.5. Let $\mathbf{x}$ be a random vector with pdf $p(\mathbf{x}; \theta_0)$ such that the Fisher information with respect to θ_0 has the form $J^{-1}(\theta_0) = b^2\theta_0^2 + 2c\theta_0 + a$, where b, c are real constants and $a > 0$. Then the minimax M and u that are the solution to (3.84) with $\mathcal{U} = \mathbb{R}$ are given by

$$M = \begin{cases} -\dfrac{2b^2}{1+b^2}, & |b| < 1; \\ -1, & |b| \geq 1, \end{cases} \qquad u = \begin{cases} -\dfrac{2c}{1+b^2}, & |b| < 1; \\ -\dfrac{c}{b^2}, & |b| \geq 1. \end{cases} \tag{3.85}$$

Furthermore, if there exists an efficient estimator $\hat{\theta}$, then

$$\hat{\theta}_{\mathrm{b}} = \begin{cases} \dfrac{1-b^2}{1+b^2}\hat{\theta} - \dfrac{2c}{1+b^2}, & |b| < 1; \\ -\dfrac{c}{b^2}, & |b| \geq 1 \end{cases}$$

has smaller MSE than $\hat{\theta}$ for all θ_0.

The minimax linear modification for the Fisher information of Example 3.5 with $c = 0$ was treated in Example 3.4. Substituting $c = 0$ in (3.85) we see that the values of M coincide in both examples.

3.5.1 Constant Fisher Matrix

We next revisit the case of a constant Fisher matrix discussed in Section 3.4.5, with $\boldsymbol{\theta}_0$ restricted to a quadratic set. Specifically, suppose that $\mathbf{J}(\boldsymbol{\theta}_0) = \mathbf{A}^{-1}$, and $\boldsymbol{\theta}_0 \in \mathcal{Q}$ with

$$\mathcal{Q} = \left\{ \boldsymbol{\theta}_0 : \boldsymbol{\theta}_0^T \boldsymbol{\theta}_0 + 2\mathbf{b}_1^T \boldsymbol{\theta}_0 + c_1 \leq 0 \right\}. \tag{3.86}$$

Then it can be shown that the optimal $\mathbf{M}$ and $\mathbf{u}$ that are the solution to (3.84) are given by

$$\mathbf{M} = -\frac{\mathrm{Tr}(\mathbf{A})}{\mathrm{Tr}(\mathbf{A}) + \mathbf{b}_1^T \mathbf{b}_1 - c_1}\hat{\boldsymbol{\theta}}, \quad \mathbf{u} = -\frac{\mathrm{Tr}(\mathbf{A})}{\mathrm{Tr}(\mathbf{A}) + \mathbf{b}_1^T \mathbf{b}_1 - c_1}\mathbf{b}_1, \tag{3.87}$$

which leads to the following proposition [40].

Proposition 3.7. Let $\mathbf{x}$ denote measurements of a deterministic parameter vector $\boldsymbol{\theta}_0$ with pdf $p(\mathbf{x}; \boldsymbol{\theta}_0)$. Assume that the Fisher information with respect to $\boldsymbol{\theta}_0$ has the form $\mathbf{J}(\boldsymbol{\theta}_0) = \mathbf{A}^{-1}$, and that $\boldsymbol{\theta}_0 \in \mathcal{Q}$ of (3.86). If there exists an efficient estimator $\hat{\boldsymbol{\theta}}$, then the estimator

$$\hat{\boldsymbol{\theta}}_{\mathrm{b}} = \frac{\mathbf{b}_1^T \mathbf{b}_1 - c_1}{\mathrm{Tr}(\mathbf{A}) + \mathbf{b}_1^T \mathbf{b}_1 - c_1}\hat{\boldsymbol{\theta}} - \frac{\mathrm{Tr}(\mathbf{A})}{\mathrm{Tr}(\mathbf{A}) + \mathbf{b}_1^T \mathbf{b}_1 - c_1}\mathbf{b}_1 \tag{3.88}$$

has smaller MSE than $\hat{\boldsymbol{\theta}}$ for all $\boldsymbol{\theta}_0 \in \mathcal{Q}$. The corresponding affine MSE bound is

$$\frac{\mathrm{Tr}(\mathbf{A})}{(\mathrm{Tr}(\mathbf{A}) + \mathbf{b}_1^T \mathbf{b}_1 - c_1)^2} \left(\mathrm{Tr}(\mathbf{A}) \|\boldsymbol{\theta}_0 + \mathbf{b}_1\|^2 + (\mathbf{b}_1^T \mathbf{b}_1 - c)^2 \right). \tag{3.89}$$

As we expect intuitively, when $\mathbf{b}_1 = \mathbf{0}$ so that the constraint set is symmetric around 0, the optimal choice of $\mathbf{u}$ is $\mathbf{u} = \mathbf{0}$. In this case the minimax affine bias coincides with the minimax linear solution given by Proposition 3.6. Another interesting case is when the set $\mathcal{Q}$ is given by $\|\boldsymbol{\theta} - \mathbf{v}\|^2 \leq c$. For this choice,

$$\mathbf{M} = -\frac{\mathrm{Tr}(\mathbf{A})}{\mathrm{Tr}(\mathbf{A}) + c}\mathbf{I}, \quad \mathbf{u} = \frac{\mathrm{Tr}(\mathbf{A})}{\mathrm{Tr}(\mathbf{A}) + c}\mathbf{v}. \tag{3.90}$$

The value of $\mathbf{M}$ is the same as that obtained with the linear minimax correction when $\mathbf{v} = \mathbf{0}$ (see Proposition 3.6). Therefore, the effect of shifting the center of the set is to shift the estimator in the direction of the center, with magnitude that takes into account both the set (via c) and the Fisher information (via $\mathbf{A}$). Note, however, that the linear minimax $\mathbf{M}$ with respect to the shifted set $\|\boldsymbol{\theta} - \mathbf{v}\|^2 \leq c$ will be different than that given by (3.90).

It is interesting to consider the relative improvement over the CRB afforded by using the optimal affine bias. Denoting by $r(\mathbf{A}, \boldsymbol{\theta}_0)$ the ratio between the affine bound (3.89) and the CRB (which is equal to $\mathrm{Tr}(\mathbf{A})$) we have

$$r(\mathbf{A}, \boldsymbol{\theta}_0) = \frac{\mathrm{Tr}(\mathbf{A})\|\boldsymbol{\theta}_0 + \mathbf{b}_1\|^2 + (\mathbf{b}_1^T\mathbf{b}_1 - c)^2}{(\mathrm{Tr}(\mathbf{A}) + \mathbf{b}_1^T\mathbf{b}_1 - c_1)^2}. \tag{3.91}$$

It is easy to see that the derivative of $r(\mathbf{A}, \boldsymbol{\theta}_0)$ with respect to $\mathrm{Tr}(\mathbf{A})$ is negative, as long as

$$\|\boldsymbol{\theta}_0 + \mathbf{b}_1\|^2 < 2(\mathbf{b}_1^T\mathbf{b}_1 - c_1). \tag{3.92}$$

Since $\boldsymbol{\theta}_0 \in \mathcal{Q}$ with $\mathcal{Q}$ defined by (3.86) we have that $\|\boldsymbol{\theta}_0 + \mathbf{b}_1\|^2 \leq \mathbf{b}_1^T\mathbf{b}_1 - c_1$ and therefore (3.92) is satisfied as long as $\mathbf{b}_1^T\mathbf{b}_1 - c_1 > 0$, or equivalently, as long as there is more than one possible value of $\boldsymbol{\theta}_0$, which is our standing assumption. Thus, $r(\mathbf{A}, \boldsymbol{\theta}_0)$ is monotonically decreasing in $\mathrm{Tr}(\mathbf{A})$ and consequently the relative improvement is more pronounced when the CRB is large. This makes intuitive sense: When the estimation problem is difficult (such as small sample size, low SNR), we can benefit from biased methods.

Example 3.6. Consider the linear Gaussian model (1.2) for which $\mathbf{J}^{-1}(\boldsymbol{\theta}_0) = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}$. Suppose we know that $\|\boldsymbol{\theta}_0 - \mathbf{v}\|^2 \leq c$ for some $\mathbf{v}$ and $c > 0$. From Proposition 3.7 the minimax MSE estimator under this constraint is

$$\hat{\boldsymbol{\theta}} = \frac{c}{c + \mathrm{Tr}((\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1})}\hat{\boldsymbol{\theta}}_{\mathrm{LS}} + \frac{\mathrm{Tr}((\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1})}{c + \mathrm{Tr}((\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1})}\mathbf{v}. \tag{3.93}$$

As an example of the improvement afforded by the affine modification (3.93), in Figure 3.3 we compare its MSE with the MSE of the minimax linear estimator which is a solution to (3.36), and the MVU solution which coincides with the LS estimate (1.13). Note that the resulting transformations $\mathbf{M}$ are different for the linear and affine modifications. We assume that $\mathbf{C} = \sigma^2\mathbf{I}$, where σ^2 is varied to achieve the desired SNR, defined by $\mathrm{SNR}\ [\mathrm{dB}] = 10\log\|\boldsymbol{\theta}_0\|^2/\sigma^2$. Here $\mathbf{v} = (1,1,1,1)^T, c = 4, \boldsymbol{\theta}_0 = 2\mathbf{v}$, and $\mathbf{H}^T\mathbf{H}$ was generated as a realization of a random matrix. As can be seen from the figure, allowing for an affine transformation improves the performance significantly. It is also apparent that as σ^2 increases, the relative improvement is more

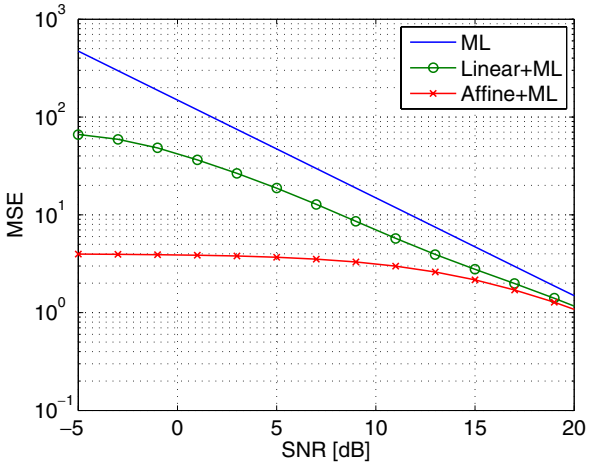


Fig. 3.3 MSE in estimating $\boldsymbol{\theta}_0$ in a linear Gaussian model as a function of the SNR using the least-squares, linear modification and affine modifications of the least-squares estimator.

pronounced. This follows from our general analysis in which we have shown that the relative advantage increases when the CRB is large.

3.6 Application: SNR Estimation

Up until this point we have shown *analytically* that the CRB can be uniformly improved upon using an affine bias. We also discussed how to construct an estimator whose MSE is uniformly lower than a given efficient method. Here we demonstrate that these results can be used in practical settings even when an efficient approach is unknown. Specifically, we propose an affine modification of the ML estimator regardless of whether the ML strategy is efficient. We illustrate this basic idea in the context of SNR estimation.

Suppose we wish to estimate the SNR of a constant signal in Gaussian noise, from n iid measurements

$$x_i = \mu + w_i, \quad 1 \le i \le n \tag{3.94}$$

where w_i is a zero-mean Gaussian random variable with variance σ^2, and the SNR is defined by $\theta_0 = \mu^2/\sigma^2$. The ML estimate of θ_0 is

$$\hat{\theta} = \frac{\hat{\mu}^2}{\hat{\sigma}^2}, \tag{3.95}$$

where

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2. \tag{3.96}$$

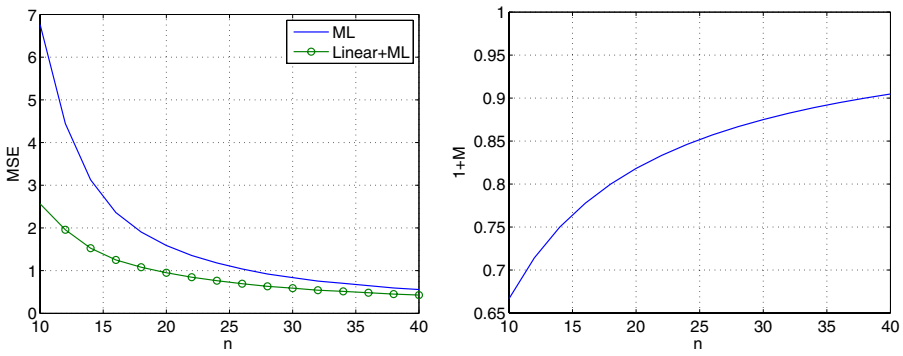
In general $\hat{\theta}$ is biased and does not achieve the CRB.

As we have seen in Example 3.3. III in Section 3.4, the inverse Fisher information in this case is

$$J^{-1}(\theta_0) = \frac{1}{n}(4\theta_0 + 2\theta_0^2). \tag{3.97}$$

In addition, we know that $\theta_0 \ge 0$ for all choices of μ and σ^2. Thus, to obtain a lower bound than the CRB we may seek the scalar $\hat{M}$ that is the solution to

$$\min_{M} \max_{\theta \ge 0} \left\{ \theta^2 M^2 + ((1+M)^2 - 1) J^{-1}(\theta) \right\}. \tag{3.98}$$



(a) (b)

Fig. 3.4 Estimating SNR using the ML and linearly transformed ML estimators. (a) MSE as a function of the number of observations n for an SNR of 2. (b) $1 + \widehat{M}$ as a function n.

The optimal value of M can be found using the SDP formulation of Section 3.4.4. For our estimator, we then use $(1 + \hat{M})\hat{\theta}$. (It can be shown that in this example the optimal affine choice is $\hat{u} = 0$.)

In Figure 3.4(a) we compare the MSE of the ML and linearly modified ML estimators as a function of the number of observations n for an SNR of $\theta_0 = 2$. For each value of n, the MSE is averaged over $10,000$ noise realizations. As can be seen from the figure, the MSE of the linearly modified ML approach is smaller than that of the ML estimator for all values of n. In Figure 3.4(b) we plot $1 + \widehat{M}$ as a function of n.

In some cases we may have prior information on the range of SNR values possible, which can be exploited to further improve the performance. Suppose that the SNR satisfies $\alpha \leq \theta_0 \leq \beta$ for some values of α and β. The ML solution is then

$$\hat{\theta}_c = \begin{cases} \hat{\theta}, & \alpha \leq \hat{\theta} \leq \beta; \\ \alpha, & \hat{\theta} \leq \alpha; \\ \beta, & \hat{\theta} \geq \beta, \end{cases} \tag{3.99}$$

where $\hat{\theta} = \hat{\mu}^2/\hat{\sigma}^2$. To develop an affine modification of ML we note that the constraint $\alpha \leq \theta_0 \leq \beta$ can be written as

$$(\theta_0 - \alpha)(\theta_0 - \beta) = \theta_0^2 - (\alpha + \beta)\theta_0 + \alpha\beta \leq 0. \tag{3.100}$$

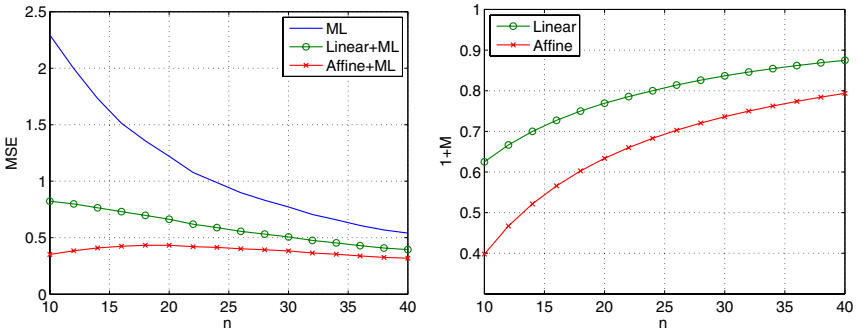


Fig. 3.5 Estimating SNR using the ML, linearly transformed ML and affine transformed ML estimators subject to the constraint (3.100). (a) MSE as a function of the number of observations n for an SNR of 2. (b) $1 + \widehat{M}$ as a function n.

Since the constraint is quadratic, the optimal M and u can be found using an SDP formulation. They are then applied to the constrained ML solution (3.99) to yield the estimate $\hat{\theta} = (1 + \widehat{M})\hat{\theta}_c + \hat{u}$.

In Figure 3.5 we compare the MSE of the ML, the linearly modified ML and affine modified ML estimators subject to (3.100), for an SNR of $\theta_0 = 2$ and SNR bounds $\alpha = 1, \beta = 5$. For each value of n, the MSE is averaged over $10,000$ noise realizations. As can be seen from the figure, the affine modification of the ML estimator performs significantly better than the ML approach and also better than the linearly modified ML method.

This example illustrates the fact that even when no efficient estimator exists, our general ideas can still be used in practice to reduce the MSE. Furthermore, we can improve the performance of constrained estimates as well. In particular, here the linear and affine modification were applied to the constrained ML solution. This is despite the fact that the CRB may no longer be a bound on the constrained ML performance, since the latter is heavily biased due to the parameter restrictions. Nonetheless, we were able to improve the performance for all values of n.

Another interesting point is that in this particular example, when using the affine modification the resulting estimate resided in the interval $[\alpha, \beta]$. This is not true for the linear modification whose performance

can be slightly improved by projection onto the constraint set. However, the impact on performance is small, and is therefore not plotted.

Before concluding this section, we emphasize the key idea presented herein: the CRB can be dominated for all parameter values by solving a certain minimax optimization problem. Although we focused here on the CRB, the basic concepts and tools we proposed are relevant in a more general context and can be used for other variance measures, as well as other classes of bias vectors.

In this section, we primarily discussed linear improvements of the ML solution. In the next sections, we will consider nonlinear corrections that may further improve the MSE. There are several strategies to obtain nonlinear modifications that dominate the ML solution:

(1) Use the optimal linear correction (3.11) to generate a sequence of iterations as in (3.23). This method is studied for the linear Gaussian model in [52].
(2) Combine the linear minimax framework presented in this section with a constraint set on $\boldsymbol{\theta}_0$ that is estimated from the data. This blind minimax technique is discussed in detail in the next section.
(3) Instead of computing the exact MSE, use an estimate of the MSE in order to design nonlinear estimates. This is the topic of Section 5.
(4) Replace the MSE, which is hard to compute for nonlinear estimates, by a deterministic error measure in conjunction with a minimax approach for constrained estimation problems. This strategy will be discussed in Section 6.

4

Minimax and Blind Minimax Estimation

In this section we depart from the linearity assumption that was prevalent in the last section, and explore the use of nonlinear improvements of the ML estimate. This will be the theme in the remainder of the survey. The general strategy proposed here is the blind minimax framework in which we first estimate a constraint set from the data, and then design a linear minimax solution matched to the estimated set. Although this essential idea is applicable to a broad class of problems, for concreteness, we demonstrate the details of this approach on the linear Gaussian model. Our starting point is the linear minimax approach presented in the previous section to improve the CRB. As we show, when the CRB is constant, this strategy reduces to finding the minimax MSE estimate. We first study this problem in detail, and derive closed form solutions. The heart of the section is in showing how the linear minimax strategy can be used as a basis for the development of nonlinear improvements of the ML design method. This will lead to ML domination *with no prior information on* $\boldsymbol{\theta}_0$ in the Gaussian setting when the dimension is large enough, even though the Fisher information is constant.

4.1 Minimax MSE Estimation

Theorem 3.2 provides a general recipe for improving the CRB using a linear modification. In this section we are concerned with the case in which $\mathbf{J}(\boldsymbol{\theta}_0) = \mathbf{Q}$ is constant, independent of $\boldsymbol{\theta}_0$. In this setting (3.36) reduces to

$$\begin{aligned}
&\arg\min_{\mathbf{M}} \sup_{\boldsymbol{\theta}\in\mathcal{U}} \left\{\mathrm{MSEB}\left(\mathbf{M},\boldsymbol{\theta}\right) - \mathbf{Q}^{-1}\right\} \\
&\quad = \arg\min_{\mathbf{M}} \sup_{\boldsymbol{\theta}\in\mathcal{U}} \mathrm{MSEB}\left(\mathbf{M},\boldsymbol{\theta}\right) \\
&\quad = \arg\min_{\mathbf{M}} \sup_{\boldsymbol{\theta}\in\mathcal{U}} \left\{\boldsymbol{\theta}^T\mathbf{M}^T\mathbf{M}\boldsymbol{\theta} + \mathrm{Tr}\left((\mathbf{I}+\mathbf{M})\mathbf{Q}^{-1}(\mathbf{I}+\mathbf{M})^T\right)\right\}. \quad (4.1)
\end{aligned}$$

If $\hat{\boldsymbol{\theta}}$ is an efficient estimate with variance $\mathbf{Q}^{-1}$, then the term in the brackets in (4.1) is the MSE of $\hat{\boldsymbol{\theta}}_{\mathrm{b}} = (\mathbf{I}+\mathbf{M})\hat{\boldsymbol{\theta}}$. Thus, our approach is equivalent to finding the linear modification $(\mathbf{I}+\mathbf{M})\hat{\boldsymbol{\theta}}$ that minimizes the worst-case MSE over the given parameter set $\mathcal{U}$. If $\mathcal{U}$ is quadratic, then the solution can be obtained using the results of Section 3.4.4.

4.1.1 Linear Gaussian Model

An interesting special case is the linear Gaussian model (1.2) for which

$$\mathbf{Q} = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}. \quad (4.2)$$

Our primary focus in this section is on this setting although the ideas easily extend to the more general constant-Fisher case.

The MVU estimate for the linear Gaussian model is the LS solution (1.13). An important property of the LS estimate is that it is linear in the data $\mathbf{x}$. Therefore, we can write

$$\hat{\boldsymbol{\theta}}_{\mathrm{b}} = (\mathbf{I}+\mathbf{M})\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = \mathbf{G}\mathbf{x}, \quad (4.3)$$

where $\mathbf{G} = (\mathbf{I}+\mathbf{M})(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}$ is an $n \times m$ matrix. Since $\mathbf{G}\mathbf{H} = \mathbf{I}+\mathbf{M}$, the problem (4.1) can be written in terms of $\mathbf{G}$ as

$$\min_{\mathbf{G}} \sup_{\boldsymbol{\theta}\in\mathcal{U}} \left\{\boldsymbol{\theta}^T(\mathbf{I}-\mathbf{G}\mathbf{H})^T(\mathbf{I}-\mathbf{G}\mathbf{H})\boldsymbol{\theta} + \mathrm{Tr}\left(\mathbf{G}\mathbf{H}(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{G}^T\right)\right\}. \quad (4.4)$$

It can be shown that (4.4) is equivalent to

$$\min_{\mathbf{G}} \sup_{\boldsymbol{\theta}\in\mathcal{U}} \left\{ \boldsymbol{\theta}^T (\mathbf{I}-\mathbf{GH})^T (\mathbf{I}-\mathbf{GH})\boldsymbol{\theta} + \mathrm{Tr}\left(\mathbf{GCG}^T\right) \right\}. \tag{4.5}$$

This follows from the fact that the optimal solution to (4.5) satisfies $\mathbf{G} = \mathbf{GH}(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}$ (see the proof of Proposition 1 in [51]).

To interpret (4.5), note that the MSE of any linear estimate $\hat{\boldsymbol{\theta}} = \mathbf{Gx}$ of $\boldsymbol{\theta}_0$ in the model (1.2) is given by the objective in (4.5). Therefore, this problem can be interpreted as finding the linear estimate that minimizes the worst-case MSE over $\mathcal{U}$. This strategy is referred to in the statistical literature as linear minimax MSE estimation [3, 43, 51, 82, 96, 116, 127].

The most common type of restriction treated in this context is an ellipsoidal set of the form $\|\boldsymbol{\theta}_0\|_{\mathbf{T}}^2 = \boldsymbol{\theta}_0^T\mathbf{T}\boldsymbol{\theta}_0 \leq U^2$ for some matrix $\mathbf{T} \succ 0$ and constant $U > 0$. Earlier references derived the minimax MSE solution for the iid case in which $\mathbf{H} = \mathbf{C} = \mathbf{I}$, and $\mathbf{T}$ is a diagonal matrix. In [3] an approximate solution is developed for colored noise with $\mathbf{H} = \mathbf{I}$. Several iterative algorithms were proposed in [82, 96]. However, these methods are computationally demanding and have no convergence proof. Furthermore, even upon convergence, they are not guaranteed to yield the minimax solution.

To obtain an efficient numerical algorithm to find the minimax MSE estimator for general choices of $\mathbf{H}$ and $\mathbf{C}$, note that (4.4) is a special case of (3.36). Therefore, if $\boldsymbol{\theta}_0 \in \mathcal{Q}$, where $\mathcal{Q}$ is a quadratic set as in (3.49), then an efficient SDP formulation of the minimax linear estimate can be obtained by using Proposition 3.5 with values $\mathbf{B}_i = 0, \mathbf{C}_i = 0$, and $\mathbf{A} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}$. The advantage of this representation is that the solution of the SDP is guaranteed to be the minimax estimate, and can be found in polynomial time within any desired accuracy using off-the-shelf numerical packages.

4.1.2 Closed-Form Solutions

Once the minimax MSE estimate is formulated as a solution to an SDP, the KKT conditions can be used to derive closed-form solutions. This approach was used in [51] and [43] to derive explicit expressions for the

minimax MSE estimate in the Gaussian model with a constraint set of the form $\boldsymbol{\theta}_0^T \mathbf{T} \boldsymbol{\theta}_0 \leq U^2$.

One class of examples in which a closed form solution exists is when $\mathbf{T}$ and $\mathbf{Q}$ of (4.2) have the same eigenvector matrix:

Proposition 4.1. Let $\boldsymbol{\theta}_0$ denote the deterministic unknown parameters in the model $\mathbf{x} = \mathbf{H}\boldsymbol{\theta}_0 + \mathbf{w}$, where $\mathbf{H}$ is a known $n \times m$ matrix with rank m, and $\mathbf{w}$ is a zero-mean random vector with covariance $\mathbf{C} \succ 0$. Let $\mathbf{Q} = \mathbf{H}^T \mathbf{C}^{-1} \mathbf{H} = \mathbf{V} \Sigma \mathbf{V}^T$, where $\mathbf{V}$ is unitary and $\Sigma = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_m)$ and let $\mathbf{T} = \mathbf{V} \Lambda \mathbf{V}^T$, where $\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_m)$ with $\lambda_1 \geq \cdots \geq \lambda_m > 0$. Then the solution to (4.5) with $\mathcal{U} = \{\boldsymbol{\theta}_0 : \boldsymbol{\theta}_0^T \mathbf{T} \boldsymbol{\theta}_0 \leq U^2\}$ is given by

$$\hat{\boldsymbol{\theta}} = \mathbf{P}(\mathbf{I} - \alpha \mathbf{T}^{1/2})(\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x} = \mathbf{P}(\mathbf{I} - \alpha \mathbf{T}^{1/2}) \hat{\boldsymbol{\theta}}_{\mathrm{LS}}\,,$$

where

$$\mathbf{P} = \mathbf{V} \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I}_{m-k} \end{bmatrix} \mathbf{V}^T$$

is an orthogonal projection onto the space spanned by the last $m - k$ columns of $\mathbf{V}$,

$$\alpha = \frac{\sum_{i=k+1}^{m} (\lambda_i^{1/2} / \sigma_i)}{U^2 + \sum_{i=k+1}^{m} (\lambda_i / \sigma_i)},$$

and k is the smallest index such that $0 \leq k \leq m - 1$ and

$$\alpha \lambda_{k+1}^{1/2} < 1. \tag{4.6}$$

For $\mathbf{T} = \mathbf{I}$,

$$\hat{\boldsymbol{\theta}} = \frac{U^2}{U^2 + \mathrm{Tr}((\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1})} \hat{\boldsymbol{\theta}}_{\mathrm{LS}}\,. \tag{4.7}$$

Note that there always exists a $0 \leq k \leq m - 1$ satisfying the condition (4.6). Indeed, for $k = m - 1$ we have that

$$\alpha = \frac{\sum_{i=k+1}^{m} (\lambda_i^{1/2} / \sigma_i)}{U^2 + \sum_{i=k+1}^{m} (\lambda_i / \sigma_i)} = \frac{\lambda_m^{1/2} / \sigma_m}{U^2 + \lambda_m / \sigma_m} < \lambda_m^{-1/2} \tag{4.8}$$

so that $\alpha \lambda_m^{1/2} < 1$. For particular values of λ_i and σ_i, there may be smaller values of k for which (4.6) holds.

Another case in which a closed-form solution can be determined was studied in [43], and is given in the following proposition.

Proposition 4.2. Consider the setting of Proposition 4.1 with arbitrary $\mathbf{T}$ satisfying

$$\lambda_{\min}\big(\mathbf{Q}^{-1}(\mathbf{Q}\mathbf{T}^{-1}\mathbf{Q})^{1/2}\big) \geq \alpha, \tag{4.9}$$

where

$$\alpha = \frac{\mathrm{Tr}\left((\mathbf{Q}^{-1}\mathbf{T}\mathbf{Q}^{-1})^{1/2}\right)}{U^2 + \mathrm{Tr}(\mathbf{Q}^{-1}\mathbf{T})}. \tag{4.10}$$

Then the solution to (4.5) is

$$\hat{\boldsymbol{\theta}} = (\mathbf{I} - \alpha\left(\mathbf{Q}^{-1}\mathbf{T}\mathbf{Q}^{-1}\right)^{1/2}\mathbf{Q})\hat{\boldsymbol{\theta}}_{\mathrm{LS}}. \tag{4.11}$$

The minimax estimator over any compact constraint set $\mathcal{U}$ was shown in [10] to dominate the LS solution for all $\boldsymbol{\theta}_0 \in \mathcal{U}$. This also follows from our general analysis in the previous section, since this estimate is a special case of (3.36). To illustrate this result, consider the case in which $\mathbf{T} = \mathbf{I}$. The minimax solution $\hat{\boldsymbol{\theta}}$ follows from Proposition 4.1, and has MSE given by

$$E\big\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2\big\} = \frac{\mathrm{Tr}^2(\mathbf{Q}^{-1})\|\boldsymbol{\theta}_0\|^2 + U^4\mathrm{Tr}(\mathbf{Q}^{-1})}{(U^2 + \mathrm{Tr}(\mathbf{Q}^{-1}))^2}. \tag{4.12}$$

Since $\|\boldsymbol{\theta}_0\|^2 \leq U^2$, the MSE is upper bounded by

$$E\big\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2\big\} \leq \frac{U^2\mathrm{Tr}(\mathbf{Q}^{-1})}{U^2 + \mathrm{Tr}(\mathbf{Q}^{-1})} \leq \mathrm{Tr}(\mathbf{Q}^{-1}), \tag{4.13}$$

which is the MSE of $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$.

In this section we focus on the case in which $\mathbf{H}$ and $\mathbf{C}$ are known. However, similar minimax ideas can be used to estimate $\mathbf{x}$ when both these variables are subject to uncertainty [6, 9, 44, 45, 51]. Minimax approaches have also been thoroughly investigated in the Bayesian setting in which $\boldsymbol{\theta}_0$ is random but the statistics are not known exactly [42, 46, 53, 54, 85, 91, 142].

4.1.3 Minimax Regret Estimation

Although the minimax approach has enjoyed widespread use in the design of robust methods for signal processing and communication [90, 91], it may be overly conservative since it optimizes the performance for the worst possible choice of unknowns. In [50] a new competitive approach to linear estimation was proposed, based on the concept of minimax regret. The idea is to seek a linear estimator whose performance is as close as possible to that of the optimal linear approach, i.e., the one minimizing the MSE when $\boldsymbol{\theta}_0$ is known. More specifically, the minimax regret estimator minimizes the worst-case difference between the MSE of a linear method $\hat{\boldsymbol{\theta}} = \mathbf{G}\mathbf{x}$, which does not know $\boldsymbol{\theta}_0$, and the smallest attainable MSE with a linear estimator $\hat{\boldsymbol{\theta}} = \mathbf{G}(\boldsymbol{\theta}_0)\mathbf{x}$ that knows $\boldsymbol{\theta}_0$, so that $\mathbf{G}$ can depend explicitly on $\boldsymbol{\theta}_0$. Since we are restricting ourselves to linear estimators of the form $\hat{\boldsymbol{\theta}} = \mathbf{G}\mathbf{x}$, we cannot achieve zero MSE even when $\boldsymbol{\theta}_0$ is known. The best possible MSE, which we denote by MSE_0, is illustrated schematically in Figure 4.1. Instead of choosing an estimate to minimize the worst-case MSE, we propose designing $\hat{\boldsymbol{\theta}}$ to minimize the worst-case difference between its MSE and the best

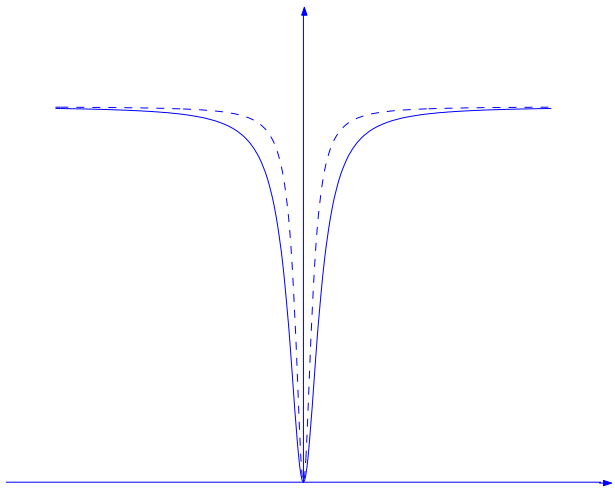


Fig. 4.1 The line represents the best attainable MSE as a function of $\boldsymbol{\theta}_0$ when $\boldsymbol{\theta}_0$ is known, and the dashed line represents a desirable graph of MSE with small regret as a function of $\boldsymbol{\theta}_0$ using some linear estimator that does not depend on $\boldsymbol{\theta}_0$.

possible MSE, as illustrated in Figure 4.1. By considering the *difference* between the MSE and the optimal MSE rather than the MSE directly, we can, in some cases, counterbalance the conservative character of the minimax strategy.

To develop an explicit expression for MSE_0 we first determine the estimator $\hat{\boldsymbol{\theta}} = \mathbf{G}(\boldsymbol{\theta}_0)\mathbf{x}$ that minimizes the MSE when $\boldsymbol{\theta}_0$ is known. Differentiating the MSE (given by the objective in (4.5)) with respect to $\mathbf{G}$ and equating to $\mathbf{0}$, results in

$$\mathbf{G}(\boldsymbol{\theta}_0)\mathbf{C} + (\mathbf{G}(\boldsymbol{\theta}_0)\mathbf{H} - \mathbf{I})\boldsymbol{\theta}_0\boldsymbol{\theta}_0^T\mathbf{H}^T = \mathbf{0}, \tag{4.14}$$

so that, after applying the matrix inversion lemma,

$$\mathbf{G}(\boldsymbol{\theta}_0) = \frac{1}{1 + \boldsymbol{\theta}_0^T\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\boldsymbol{\theta}_0}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^T\mathbf{H}^T\mathbf{C}^{-1}. \tag{4.15}$$

Substituting $\mathbf{G}(\boldsymbol{\theta}_0)$ back into the expression for the MSE

$$\text{MSE}_0 = \frac{\boldsymbol{\theta}_0^T\boldsymbol{\theta}_0}{1 + \boldsymbol{\theta}_0^T\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\boldsymbol{\theta}_0}. \tag{4.16}$$

Since $\boldsymbol{\theta}_0$ is unknown, we cannot implement the optimal estimator (4.15). Instead we design $\hat{\boldsymbol{\theta}} = \mathbf{G}\mathbf{x}$ to minimize the worst-case regret $\mathcal{R}(\boldsymbol{\theta}_0, \mathbf{G})$, where

$$\begin{aligned}\mathcal{R}(\boldsymbol{\theta}_0, \mathbf{G}) &= E\left\{\|\mathbf{G}\mathbf{x} - \boldsymbol{\theta}_0\|^2\right\} - \text{MSE}_0 \\ &= \text{Tr}(\mathbf{G}\mathbf{C}\mathbf{G}^T) + \boldsymbol{\theta}_0^T(\mathbf{I} - \mathbf{G}\mathbf{H})^T(\mathbf{I} - \mathbf{G}\mathbf{H})\boldsymbol{\theta}_0 \\ &\quad - \frac{\boldsymbol{\theta}_0^T\boldsymbol{\theta}_0}{1 + \boldsymbol{\theta}_0^T\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\boldsymbol{\theta}_0},\end{aligned}$$

subject to the constraint $\|\boldsymbol{\theta}_0\|_{\mathbf{T}} \leq U$. Thus we seek the matrix $\mathbf{G}$ that is the solution to

$$\min_{\mathbf{G}} \max_{\boldsymbol{\theta}^T\mathbf{T}\boldsymbol{\theta} \leq U^2} \mathcal{R}(\boldsymbol{\theta}, \mathbf{G}). \tag{4.17}$$

Problem (4.17) is a nonconvex, difficult optimization problem. Nonetheless, in many cases it can be transformed into convex form. One class of examples is when $\mathbf{T}$ and $\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}$ have the same eigenvector matrix, as incorporated in the following proposition.

Proposition 4.3. Consider the setting of Proposition 4.1. Then the minimax regret estimator has the form

$$\hat{\boldsymbol{\theta}} = \mathbf{V}\mathbf{D}\mathbf{V}^T(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x},$$

with $\mathbf{D} = \mathrm{diag}\,(d_1, \ldots, d_m)$ where d_i are the solution to the convex optimization problem

$$\begin{aligned} \min_{\tau, d_i} \quad & \tau \\ \text{s.t.} \quad & \sum_{i=1}^{m} \frac{d_i^2}{\sigma_i} \le \tau \\ & \max_{s_i \in \mathcal{S}} \left\{ \sum_{i=1}^{m} (1-d_i)^2 s_i - \frac{\sum_{i=1}^{m} s_i}{1 + \sum_{i=1}^{m} \sigma_i s_i} \right\} + \sum_{i=1}^{m} \frac{d_i^2}{\sigma_i} \le \tau, \end{aligned}$$

with

$$\mathcal{S} = \left\{ s_i : s_i \ge 0, \sum_{i=1}^{m} \lambda_i s_i = U^2 \right\}.$$

In the special case in which $\boldsymbol{\theta}_0 = \theta_0$ is a scalar so that $\mathbf{x} = \mathbf{h}\theta_0 + \mathbf{w}$ for some known vector $\mathbf{h}$, the minimax regret estimate over the interval $L \le |\theta_0| \le U$ is given by [55, 57]

$$\hat{\boldsymbol{\theta}} = \left(1 - \frac{1}{\sqrt{(1 + L^2\mathbf{h}^T\mathbf{C}^{-1}\mathbf{h})(1 + U^2\mathbf{h}^T\mathbf{C}^{-1}\mathbf{h})}} \right) \hat{\boldsymbol{\theta}}_{\mathrm{LS}}. \tag{4.18}$$

Further special cases are discussed in [50].

The minimax regret concept has recently been used to develop competitive beamforming approaches [55, 57]. It has also been applied to linear models in which $\boldsymbol{\theta}_0$ is random with unknown covariance [42, 46, 53, 54].

In the sequel, we use the blind minimax framework to develop nonlinear estimation strategies based on the linear minimax solution. The linear minimax regret estimate can also be used in a similar way as a basis for blind regret techniques; this is an interesting direction for further study.

4.2 Stein-Type Estimates

Until now we have limited our discussion to *linear* modifications of the ML or MVU method. In the remainder of this section, and in the succeeding next sections, we consider nonlinear extensions. In many cases this allows for a more pronounced performance advantage. Furthermore, as we will show, using a nonlinear modification of LS, we can dominate it in the linear Gaussian setting over all choices of $\boldsymbol{\theta}_0$, not only norm bounded values.

In his seminal paper, Stein showed that for $\mathbf{H} = \mathbf{I}$ and white noise, the LS strategy is inadmissible when the parameter dimension is larger than 2 [128] meaning there exist estimates that dominate it for all $\boldsymbol{\theta}_0$. Several years later, James and Stein developed a nonlinear shrinkage of the conventional LS and proved that it dominates the LS solution [88]. The James–Stein class of estimators is given by

$$\hat{\boldsymbol{\theta}}_{\mathrm{JS}} = \left(1 - \frac{r\sigma^2}{\|\mathbf{x}\|^2}\right)\mathbf{x}, \tag{4.19}$$

where σ^2 is the noise variance, and $0 \leq r \leq 2(m-2)$ (Stein chose $r = m$ and James and Stein in [88] used $r = (m-2)$ which minimizes the MSE among this class). A drawback of the James–Stein choice is that the shrinkage factor can be negative. To remedy this deficiency, the positive-part James–Stein estimate was suggested by Baranchik [4] and is given by

$$\hat{\boldsymbol{\theta}} = \left[1 - \frac{r\sigma^2}{\|\mathbf{x}\|^2}\right]_+ \mathbf{x}, \tag{4.20}$$

where we used the notation

$$[x]_+ = \begin{cases} x, & x \geq 0; \\ 0, & x \leq 0. \end{cases} \tag{4.21}$$

This estimate yields lower MSE than the conventional James–Stein method (4.19).

Various "extended" James–Stein methods were later constructed for the general (non-iid) case [11, 15, 18, 36, 103]. One of the common

strategies is Bock's estimator [18] which is given by

$$\hat{\boldsymbol{\theta}} = \left(1 - \frac{d_{\text{eff}} - 2}{\hat{\boldsymbol{\theta}}_{\text{LS}}^T \mathbf{Q} \hat{\boldsymbol{\theta}}_{\text{LS}}}\right) \hat{\boldsymbol{\theta}}_{\text{LS}}, \tag{4.22}$$

where $\mathbf{Q} = \mathbf{H}^T \mathbf{C}^{-1} \mathbf{H}$ and

$$d_{\text{eff}} = \frac{\text{Tr}(\mathbf{Q}^{-1})}{\lambda_{\max}(\mathbf{Q}^{-1})} \tag{4.23}$$

is the effective dimension [103], and may be roughly described as the number of independently measured parameters in the system. Indeed, when $\mathbf{Q} = \mathbf{I}$, we have $d_{\text{eff}} = m$. However, none of these approaches has become a standard alternative to the LS estimator, and they are rarely used in practice in engineering applications [37, 103]. Perhaps one reason for this is that some of the estimators are poorly justified and seem counterintuitive, and as such they are sometimes regarded with skepticism (see discussion following [35]). Another reason is that many of these approaches (including Bock's method) result in shrinkage estimators, consisting of a gain factor multiplying LS. Shrinkage techniques can certainly be used to reduce MSE; however, in the non-iid case, some measurements are noisier than others, and thus a single shrinkage factor for all measurements can be considered suboptimal. Furthermore, in some applications, a gain factor has no effect on final system performance: for example, in image reconstruction, multiplying the entire image by a constant does not improve quality.

In the next section, we provide a framework for generating a wide class of low-complexity, LS-dominating estimators, which are constructed from a simple, intuitive principle, called the blind minimax approach [11]. This method is used as a basis for selecting and generating techniques tailored for given problems. Many blind minimax estimators (BMEs) reduce to Stein-type methods in the iid case, and they continue to dominate the LS solution for all $\boldsymbol{\theta}_0$ in the general, non-iid setting as well. Thus, we show analytically that the proposed technique achieves lower MSE than LS, when an appropriate condition on the problem setting is satisfied. Unlike Bock's approach, BMEs may be constructed so that they are non-shrinkage, which improves their

performance. Furthermore, extensive simulations show that BMEs considerably outperform Bock's method.

4.3 Blind Minimax Estimation

BMEs are based on linear minimax estimators over a bounded parameter set, introduced in Section 4.1. As we have seen, as long as some bounded set is known to contain $\boldsymbol{\theta}_0$, minimax techniques outperform LS. However, in our setting now, no prior information about the parameter set is assumed. Instead, the blind minimax approach makes use of a two-stage process:

(1) A parameter set $\mathcal{U}$ is estimated from the measurements;
(2) A minimax estimator designed for $\mathcal{U}$ is used to infer $\boldsymbol{\theta}_0$.

The result may be viewed as a simple decision rule, independent of this two-stage construction. In particular, the dominance results do not rely on the parameter actually lying within the estimated set. Thus, the blind minimax technique provides a framework whereby many different estimators can be generated, and provides insight into the mechanism by which these techniques outperform the LS approach.

BMEs differ in the way the parameter set $\mathcal{U}$ is estimated. Here, we consider sets of the form $\{\boldsymbol{\theta}_0 : \boldsymbol{\theta}_0^T \mathbf{T} \boldsymbol{\theta}_0 \leq U^2\}$ for some $\mathbf{T} \succ 0$. In Section 4.3.1, we study the case in which $\mathbf{T} = \mathbf{I}$ so that the estimated set is a sphere; Section 4.3.2 derives estimators based on an ellipsoidal parameter set corresponding to $\mathbf{T} = \mathbf{Q}^b$ for some real number b. In Section 4.4, we demonstrate that several existing Stein-type methods can also be derived within the blind minimax framework.

4.3.1 The Spherical Blind Minimax Estimator

We begin by applying the blind minimax technique using a spherical parameter set $\mathcal{U} = \{\boldsymbol{\theta}_0 : \|\boldsymbol{\theta}_0\|^2 \leq U^2\}$. For given U, the linear minimax estimator is given by (4.7):

$$\hat{\boldsymbol{\theta}}_{\mathrm{M}} = \frac{U^2}{U^2 + \epsilon_0} \hat{\boldsymbol{\theta}}_{\mathrm{LS}}, \tag{4.24}$$

where

$$\epsilon_0 = \mathrm{Tr}((\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}) = \mathrm{Tr}(\mathbf{Q}^{-1}) \tag{4.25}$$

is the MSE of $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$. The resulting spherical BME (SBME) will have the form (4.24), where U^2 is estimated from the measurements. A natural estimate of U^2 is obtained by using the LS solution as $\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2$. Substituting into (4.24), the SBME is

$$\hat{\boldsymbol{\theta}}_{\mathrm{SBM}} = \frac{\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2}{\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2 + \epsilon_0}\hat{\boldsymbol{\theta}}_{\mathrm{LS}}. \tag{4.26}$$

Up to this point, we have arbitrarily chosen the parameter set to be centered on the origin. As we shall see, the proposed BMEs outperform the LS estimator. This demonstrates the fact that the LS approach results in an overestimate: reducing the norm of $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$ improves its performance. However, the choice of a parameter set centered on the origin is completely arbitrary; shrinkage estimates may be constructed around any constant center point $\boldsymbol{\theta}_p$ [69]. This will result in a weighted average between $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$ and $\boldsymbol{\theta}_p$, which may be useful if the parameter vector is expected to lie near a particular point. Thus, the off-center SBME is given by

$$\hat{\boldsymbol{\theta}} = \left(\frac{\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2}{\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2 + \epsilon_0}\right)\hat{\boldsymbol{\theta}}_{\mathrm{LS}} + \left(\frac{\epsilon_0}{\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2 + \epsilon_0}\right)\boldsymbol{\theta}_p. \tag{4.27}$$

All dominance results continue to hold for the off-center techniques as well. For notational simplicity, in the sequel we assume $\boldsymbol{\theta}_p = \mathbf{0}$.

The following theorem demonstrates that the SBME is guaranteed to outperform LS in terms of MSE.

Theorem 4.4. Suppose that $d_{\mathrm{eff}} > 4$, where d_{eff} is defined by (4.23). Then, the SBME (4.26) strictly dominates the LS estimator.

The condition of Theorem 4.4 can be roughly stated as a requirement for a sufficient number of independent parameters.

Note that the SBME is a special case of the estimator

$$\hat{\boldsymbol{\theta}}_c = \left(1 - \frac{\epsilon_0}{c + \|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2}\right)\hat{\boldsymbol{\theta}}_{\mathrm{LS}}, \tag{4.28}$$

in which $c = \epsilon_0$. The proof of Theorem 4.4 follows from the more general proposition below.

Proposition 4.5. Under the conditions of Theorem 4.4, the estimator $\hat{\boldsymbol{\theta}}_c$ given by (4.28) strictly dominates the LS estimator, for any $c \geq 0$.

To prove the proposition we rely on the following lemma [39, Proposition 1], which we will discuss in more detail in the next section in the context of SURE estimation.

Lemma 4.6. Let $\mathbf{x} = \mathbf{H}\boldsymbol{\theta}_0 + \mathbf{w}$ denote measurements of an unknown parameter vector $\boldsymbol{\theta}_0$ where $\mathbf{H}$ is an $n \times m$ matrix of full column rank, and $\mathbf{w}$ is a zero-mean Gaussian random vector with covariance $\mathbf{C} \succ 0$. Let $\mathbf{h}(\mathbf{u})$ with $\mathbf{u} = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$ be an arbitrary function of $\mathbf{u}$ that is weakly differentiable[1] in $\mathbf{u}$ and such that $E\{|h_i(\mathbf{u})|\}$ is bounded. Then

$$E\big\{\mathbf{h}^T(\mathbf{u})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\mathrm{LS}})\big\} = -E\left\{\mathrm{Tr}\left(\frac{d\mathbf{h}(\mathbf{u})}{d\mathbf{u}}\right)\right\}, \tag{4.29}$$

where $\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$ is the LS estimate.

Proof. To prove Proposition 4.5, first note that the MSE $R(\hat{\boldsymbol{\theta}}_c) = E\big\{\|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_c\|^2\big\}$ of $\hat{\boldsymbol{\theta}}_c$ is given by

$$R(\hat{\boldsymbol{\theta}}_c) = \epsilon_0 + E\left\{\frac{\epsilon_0^2\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2}{(c + \|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2)^2}\right\} + 2E\big\{\mathbf{h}^T(\mathbf{u})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\mathrm{LS}})\big\}, \tag{4.30}$$

where

$$\mathbf{h}(\mathbf{u}) = \frac{\epsilon_0}{c + \|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2}\hat{\boldsymbol{\theta}}_{\mathrm{LS}}. \tag{4.31}$$

Applying Lemma 4.6 we obtain

$$\begin{aligned} &E\left\{\frac{\epsilon_0}{c + \|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2}\hat{\boldsymbol{\theta}}_{\mathrm{LS}}^T(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\mathrm{LS}})\right\} \\ &\quad = -\epsilon_0 E\left\{\frac{\mathrm{Tr}(\mathbf{Q}^{-1})}{c + \|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2}\right\} + 2\epsilon_0 E\left\{\frac{\hat{\boldsymbol{\theta}}_{\mathrm{LS}}^T\mathbf{Q}^{-1}\hat{\boldsymbol{\theta}}_{\mathrm{LS}}}{(c + \|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2)^2}\right\}. \end{aligned} \tag{4.32}$$

[1] Roughly speaking, a function is weakly differentiable if it has a derivative almost everywhere, as long as the points that are not differentiable are not delta functions; see [101] for a more formal definition.

Substituting this result back into (4.30), we have

$$R(\hat{\boldsymbol{\theta}}_c) = \epsilon_0 + E\left\{ \frac{\epsilon_0}{c + \|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2} \cdot \left(\epsilon_0 \frac{\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2}{c + \|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2} - 2\epsilon_0 + 4\frac{\hat{\boldsymbol{\theta}}_{\mathrm{LS}}^T \mathbf{Q}^{-1} \hat{\boldsymbol{\theta}}_{\mathrm{LS}}}{c + \|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2} \right) \right\}. \quad (4.33)$$

Since $c \geq 0$,

$$R(\hat{\boldsymbol{\theta}}_c) \leq \epsilon_0 + E\left\{ \frac{\epsilon_0}{c + \|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2} \left(-\epsilon_0 + 4\lambda_{\max}(\mathbf{Q}^{-1}) \right) \right\}. \quad (4.34)$$

If $\epsilon_0 > 4\lambda_{\max}$, then the expectation is taken over a strictly negative range. Therefore $R(\hat{\boldsymbol{\theta}}_c) < \epsilon_0$, and $\hat{\boldsymbol{\theta}}_c$ strictly dominates $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$. □

As we have shown, in terms of MSE, the SBME outperforms LS over the entire space, providing us with a first example of the power of blind minimax estimation. The SBME is a shrinkage estimator, i.e., it consists of the LS multiplied by a gain factor smaller than one. This illustrates the fact that the LS technique tends to be an overestimate, and shrinkage can improve its performance.

4.3.2 The Ellipsoidal Blind Minimax Estimator

Since the covariance of $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$ is $\mathbf{Q}^{-1} = \left(\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H}\right)^{-1}$, not all elements of $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$ are equally trustworthy. In this sense, the scalar shrinkage of the SBME and other extended Stein estimators seems inadequate.

Indeed, several researchers have proposed shrinking each measurement according to its variance. Efron and Morris [36] suggest an empirical Bayes technique, in which high-variance components are shrunk more than low-variance ones. However, obtaining an estimate requires iteratively solving a set of nonlinear equations. Furthermore, it is not known whether this method dominates LS. By contrast, Berger [15] provides an estimator in which more shrinkage is applied to low-variance measurements, despite the fact that low-noise components are those for which the LS is most accurate. Berger's technique is constructed such that the shrinkage of all components is negligible whenever there is a substantial difference between the variances of different components.

As a result, LS dominance is guaranteed, but the MSE gain is insubstantial unless all noise components have similar variances.

Minimax estimators can easily be adapted for non-scalar shrinkage. Specifically, consider an ellipsoidal parameter set of the form $\mathcal{U} = \{\boldsymbol{\theta}_0 : \|\boldsymbol{\theta}_0\|_{\mathbf{T}}^2 = \boldsymbol{\theta}_0^T \mathbf{T} \boldsymbol{\theta}_0 \leq U^2\}$ (see Figure 4.2). Let $\hat{\boldsymbol{\theta}}_{\mathrm{M}}$ represent the linear minimax estimator for this set, which is a linear function of $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$. We can examine its effect on each component of $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$. Consider first components of $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$ in the direction of narrow axes of the ellipsoid $\mathcal{U}$. These components correspond to large eigenvalues of $\mathbf{T}$, and are denoted $\lambda_{\max}(\mathbf{T})$ in Figure 4.2. The parameter set imposes a tight constraint in these directions, and there will thus be considerable shrinkage of these elements. By contrast, components in the direction of wide axes of $\mathcal{U}$ (small eigenvalues of $\mathbf{T}$) are not constrained as tightly. Less shrinkage will be applied in this case, since the LS method is the linear minimax estimator for an unbounded set. In Figure 4.2, the shrinkage of wide-axis and narrow-axis components is illustrated schematically for a particular value of $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$.

Typically, one would want to obtain higher shrinkage for high-variance components. Since the covariance of $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$ is $\mathbf{Q}^{-1}$, we propose a BME based on a parameter set of the form:

$$\mathcal{U} = \{\boldsymbol{\theta}_0 : \boldsymbol{\theta}_0^T \mathbf{Q}^b \boldsymbol{\theta}_0 \leq U^2\} \tag{4.35}$$

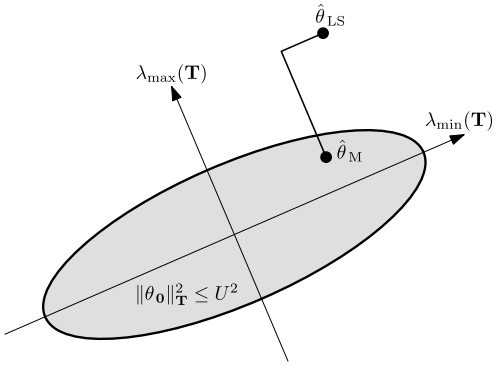


Fig. 4.2 Illustration of the adaptive shrinkage of the minimax estimator $\hat{\boldsymbol{\theta}}_{\mathrm{M}}$ for the set $\boldsymbol{\theta}_0^T \mathbf{T} \boldsymbol{\theta}_0 \leq U^2$. Low shrinkage is applied to elements of $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$ corresponding to small eigenvalues of $\mathbf{T}$, while components in directions of large eigenvalues obtain higher shrinkage.

for some constant $b < 0$. The bound U^2 is estimated as $U^2 = \hat{\boldsymbol{\theta}}_{\mathrm{LS}}^T \mathbf{Q}^b \hat{\boldsymbol{\theta}}_{\mathrm{LS}}$. We refer to the resulting technique as the ellipsoidal BME (EBME). Note that highly negative values of b yield an eccentric ellipsoid, and hence result in a larger disparity between the shrinkage of different measurements. Contrariwise, a choice of $b = 0$ yields scalar shrinkage, and the resulting estimator is identical to the SBME. As we will demonstrate, the EBME dominates the LS method under a condition similar to that of the SBME. However, the dominance condition of the EBME becomes stricter as b becomes more negative. Thus, there exists a trade-off between selective shrinkage and a broad dominance condition. Following [11], in the numerical examples below we will choose a value of $b = -1$ as a compromise.

As an additional motivation for the use of the EBME, consider the following example illustrated in Figure 4.3, which is taken from [11]. Here, a 100-sample signal is to be estimated from measurements of its discrete cosine transform (DCT). Each component of the DCT is corrupted by Gaussian noise: high-variance noise is added to the 10 highest-frequency components, while the remaining components contain much lower noise levels. In this example, $\mathbf{H}$ is the DCT matrix and $\mathbf{C}$ is diagonal in the DCT domain. Consequently, the LS estimator is equivalent to an inverse DCT transform, and thus ignores the differences in noise level between measurements. This causes substantial estimation error, as observed in Figure 4.3(a). The error is reduced by the SBME (Figure 4.3(b)), which multiplies the LS estimate by an appropriately chosen scalar; in this example, the squared error was reduced by 20%. Evidently, merely multiplying the result of the LS technique by an appropriately chosen scalar can significantly reduce estimation error. However, the most significant advantage is obtained by the EBME (Figure 4.3(c)), which shrinks the high-noise coefficients. The choice $b = -1$ resulted in shrinkage of 0.44 for the high-noise coefficients, and shrinkage of only 0.98 for low-noise coefficients. The resulting squared error was 83% lower than that of LS.

4.3.3 Dominance

We begin our analysis by obtaining an expression for the EBMEs. Since for $\mathbf{T} = \mathbf{Q}^b$, $\mathbf{T}$ and $\mathbf{Q}$ are jointly diagonalizable, we can use the results of

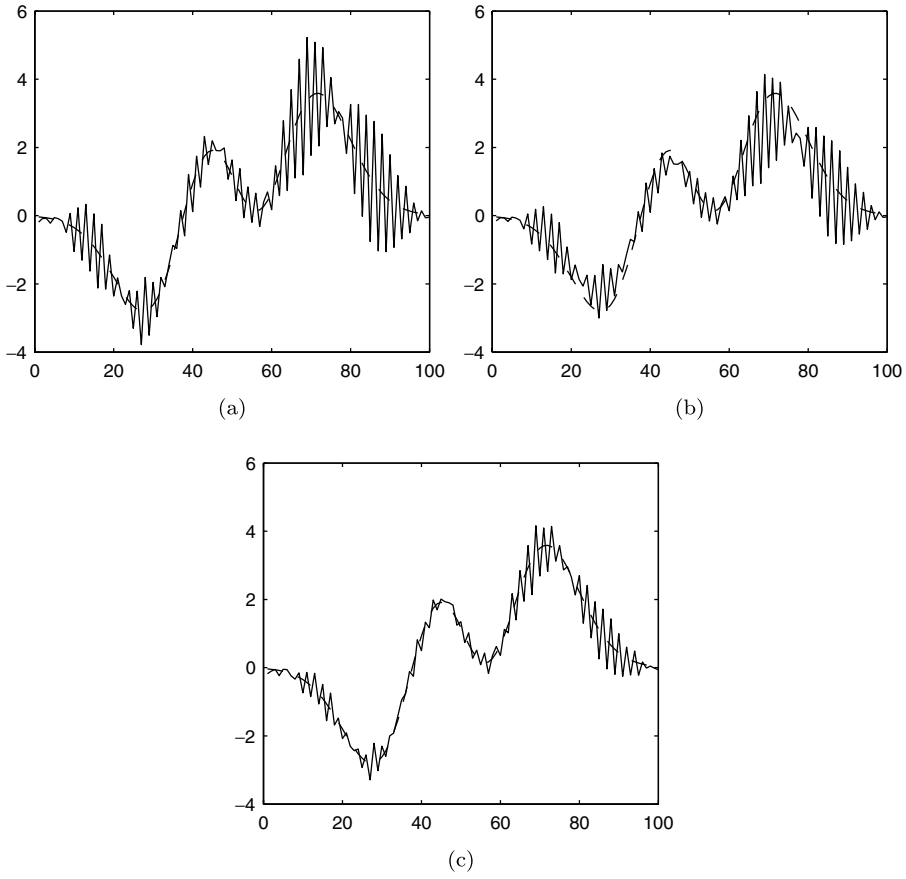


Fig. 4.3 Estimation of a signal from measurements of its DCT. Dashed line indicates original signal; solid line indicates estimate. (a) LS estimate. (b) Spherical BME. (c) Ellipsoidal BME.

Proposition 4.1 to obtain a closed-form solution for the corresponding minimax estimate. By substituting the value of U^2 into this closed form, we obtain the following result:

Proposition 4.7. Let $\mathbf{V}\Sigma\mathbf{V}^T$ be the eigenvalue decomposition of $\mathbf{Q} = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}$, where $\mathbf{V}$ is unitary and $\Sigma = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_m)$. Let $b \in \mathbb{R}$ be any constant, and suppose the eigenvalues Σ are ordered such that $\sigma_1^b \geq \sigma_2^b \geq \cdots \geq \sigma_m^b > 0$. Then, the EBME for the parameter set $\mathcal{U} = \{\boldsymbol{\theta}_0 : \|\boldsymbol{\theta}_0\|^2_{\mathbf{Q}^b} \leq U^2\}$ with $U^2 = \|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2_{\mathbf{Q}^b}$ is

given by

$$\hat{\boldsymbol{\theta}}_{\mathrm{EBM}} = \mathbf{V}\,\mathrm{diag}\,([1-\alpha\sigma_1^{b/2}]_+,\ldots,[1-\alpha\sigma_m^{b/2}]_+)\mathbf{V}^T\hat{\boldsymbol{\theta}}_{\mathrm{LS}} \tag{4.36}$$

when $\hat{\boldsymbol{\theta}}_{\mathrm{LS}} \neq \mathbf{0}$, and by $\hat{\boldsymbol{\theta}}_{\mathrm{EBM}} = \mathbf{0}$ when $\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = \mathbf{0}$. Here

$$\alpha = \frac{r_1}{\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2_{\mathbf{Q}^b} + r_2}, \quad r_1 = \sum_{i=k+1}^{m} \sigma_i^{b/2-1}, \quad r_2 = \sum_{i=k+1}^{m} \sigma_i^{b-1} \tag{4.37}$$

and k is chosen as the smallest index $0 \leq k \leq m-1$ such that

$$\alpha\sigma_{k+1}^{b/2} < 1. \tag{4.38}$$

Proof. In the case $\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = \mathbf{0}$, we need to find the linear minimax estimator for the set $\mathcal{U} = \{\mathbf{0}\}$, which is clearly given by $\hat{\boldsymbol{\theta}} = \mathbf{0}$. For all other values of $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$, we seek the linear minimax estimator for the set $\mathcal{U} = \{\boldsymbol{\theta}_0 : \boldsymbol{\theta}_0^T\mathbf{Q}^b\boldsymbol{\theta}_0 \leq U^2\}$, where $U^2 = \hat{\boldsymbol{\theta}}_{\mathrm{LS}}^T\mathbf{Q}^b\hat{\boldsymbol{\theta}}_{\mathrm{LS}} > 0$. Substituting this value of U^2 into Proposition 4.1 yields

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{EBM}} &= \mathbf{V}\,\mathrm{diag}\,(0,\ldots,0,1,\ldots,1)\mathbf{V}^T(\mathbf{I}-\alpha\mathbf{Q}^{b/2})\hat{\boldsymbol{\theta}}_{\mathrm{LS}} \\ &= \mathbf{V}\,\mathrm{diag}\,(0,\ldots,0,1-\alpha\sigma_{k+1}^{b/2},\ldots,1-\alpha\sigma_m^{b/2})\mathbf{V}^T\hat{\boldsymbol{\theta}}_{\mathrm{LS}}, \end{aligned} \tag{4.39}$$

where there are k zeros in the diagonal matrix. From (4.38), $1-\alpha\sigma_i^{b/2} < 0$ for all $i \leq k$, and therefore (4.39) can be written as (4.36). □

We note that, as long as $\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2_{\mathbf{Q}^b} > 0$, it is always possible to find a value k which satisfies (4.38).

Like the SBME, the EBME also dominates the LS estimator under suitable conditions, as shown in the following theorem. The proof is quite involved and can be found in [11].

Theorem 4.8. Let $\hat{\boldsymbol{\theta}}_{\mathrm{EBM}}$ be the EBME (4.36) and suppose that

$$\mathrm{Tr}(\mathbf{Q}^{b/2-1}) > 4\lambda_{\max}(\mathbf{Q}^{b/2-1}), \tag{4.40}$$

where $\mathbf{Q} = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}$. Then, $\hat{\boldsymbol{\theta}}_{\mathrm{EBM}}$ strictly dominates the LS estimator.

The dominance condition (4.40) is satisfied in many reasonable problems. Assuming a sufficient number of parameters, the only case in which this condition does *not* hold is the situation in which a small number of parameters (less than four) have much higher variance than all others; in this case, the LS method is admissible or nearly so.

4.4 Relation to Stein-type Estimation

Thus far, we have presented two examples of BMEs which dominate the LS method under suitable conditions. We now demonstrate that other BMEs extend Stein's estimator (4.19) and Baranchik's positive-part improvement (4.20).

In Section 4.3.1, the SBME (4.26) was constructed by using $U^2 = \|\hat{\boldsymbol{\theta}}_{\rm LS}\|^2$ as an estimate of $\|\boldsymbol{\theta}_0\|^2$. However, the fact that shrinkage techniques such as the SBME dominate LS indicates that $\hat{\boldsymbol{\theta}}_{\rm LS}$ is in fact an overestimate of $\boldsymbol{\theta}_0$. It is arguably more accurate to use a smaller value than $\|\hat{\boldsymbol{\theta}}_{\rm LS}\|^2$ to estimate $\|\boldsymbol{\theta}_0\|^2$. In particular, it is readily shown that

$$E\big\{\|\hat{\boldsymbol{\theta}}_{\rm LS}\|^2\big\} = \|\boldsymbol{\theta}_0\|^2 + \epsilon_0. \tag{4.41}$$

Hence, one may opt to use

$$U^2 = \|\hat{\boldsymbol{\theta}}_{\rm LS}\|^2 - \epsilon_0 \tag{4.42}$$

as an estimate of $\|\boldsymbol{\theta}_0\|^2$. Substituting (4.42) into the spherical minimax method (4.24) yields the balanced BME

$$\hat{\boldsymbol{\theta}}_{\rm BBM} = \left(1 - \frac{\epsilon_0}{\|\hat{\boldsymbol{\theta}}_{\rm LS}\|^2}\right)\hat{\boldsymbol{\theta}}_{\rm LS}. \tag{4.43}$$

The balanced BME reduces to Stein's estimator [128] in the iid case. Both techniques are well-defined unless $\hat{\boldsymbol{\theta}}_{\rm LS} = \mathbf{0}$, an event which has zero probability. Furthermore, the balanced BME extends Stein's method, in that it continues to dominate LS for the non-iid case, under suitable conditions. This is shown by the following theorem.

Theorem 4.9. Suppose $d_{\rm eff} > 4$, where $d_{\rm eff}$ is given by (4.23). Then, the balanced BME (4.43) strictly dominates the LS estimator.

Proof. The theorem follows by substituting $c = 0$ in Proposition 4.5. □

A well-known drawback of Stein's approach is that it sometimes causes negative shrinkage, i.e., the shrinkage factor in (4.43) is negative with nonzero probability. This is known to increase the MSE [4]. From

the blind minimax perspective, this negative shrinkage is a result of the fact that U^2 can become negative. Thus, it is natural to replace (4.42) with

$$U^2 = \left[\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2 - \epsilon_0\right]_+. \tag{4.44}$$

Substituting this value of U^2 into the spherical minimax estimator yields the positive-part BME, given by

$$\hat{\boldsymbol{\theta}}_{\mathrm{PBM}} = \left(\frac{[\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2 - \epsilon_0]_+}{[\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2 - \epsilon_0]_+ + \epsilon_0}\right)\hat{\boldsymbol{\theta}}_{\mathrm{LS}}. \tag{4.45}$$

Note that when $\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2 - \epsilon_0 < 0$, the estimator $\hat{\boldsymbol{\theta}}_{\mathrm{PBM}}$ equals $\mathbf{0}$; in all other cases, $\hat{\boldsymbol{\theta}}_{\mathrm{PBM}} = \hat{\boldsymbol{\theta}}_{\mathrm{BBM}}$. Thus, (4.45) may be written as

$$\hat{\boldsymbol{\theta}}_{\mathrm{PBM}} = \left[1 - \frac{\epsilon_0}{\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|^2}\right]_+ \hat{\boldsymbol{\theta}}_{\mathrm{LS}}. \tag{4.46}$$

In other words, $\hat{\boldsymbol{\theta}}_{\mathrm{PBM}}$ is the positive part of the balanced BME, and coincides with the Baranchik estimate in the iid case.

The balanced method presented in this section for estimating the parameter set radius results in a value (4.42) of U^2 which is smaller than that of the SBME. As a result, the balanced approach causes more shrinkage toward the origin. This tends to improve performance for low SNR at the expense of performance degradation at high SNR. In particular, $\hat{\boldsymbol{\theta}}_{\mathrm{PBM}}$ has a positive probability of yielding an estimate of $\mathbf{0}$. This may indeed reduce the MSE when the parameter is exceedingly small with respect to the noise variance, but will sacrifice high-SNR performance. In general, the positive-part BME tends to perform as well or worse than the SBME at SNR values above 0 dB, and better for lower SNR values. Thus, in most applications, use of the SBME is probably preferable. However, the fact that Stein's estimator can be derived and extended using blind minimax considerations illustrates the versatility of this approach.

4.5 Numerical Results

We now present several computer simulations, taken from [11], that illustrate the performance of the SBME and EBME. In these tests,

a value of $b = -1$ was used for the parameter set (4.35) of the EBME. Application of the minimax ideas presented here to beamforming in the context of array processing can be found in [55, 57, 125].

In the first example, we show a typical scenario, in which the number of parameters m and the number of measurements n are both 15. In addition $\mathbf{H} = \mathbf{I}$ and the noise covariance is given by

$$\mathbf{C} = \sigma^2 \operatorname{diag}(1,1,1,1,0.5,0.2,0.2,0.2,0.2,0.1,0.1,0.1,0.1,0.05,0.05) \tag{4.47}$$

resulting in an effective dimension of (5.8). Here σ^2 was selected to achieve the desired SNR. To illustrate the dependence on the value of the parameter vector $\boldsymbol{\theta}_0$, in Figure 4.4(a), $\boldsymbol{\theta}_0$ is in the direction of the maximum eigenvector of $\mathbf{Q}^{-1}$, while in Figure 4.4(b), $\boldsymbol{\theta}_0$ is chosen in the direction of the minimum eigenvector. This corresponds to parameters in the direction of maximal and minimal noise, respectively. Estimates of the MSE were calculated for a range of SNR values by generating 10,000 random realizations of noise per SNR value.

It is evident from Figure 4.4 that substantial improvement in MSE can be achieved by using BMEs in place of the LS approach: in some cases the MSE of the LS estimator is nearly three times larger than that of the BMEs. The performance gain is particularly noticeable at low and moderate SNR. At infinite SNR, the LS technique is known to

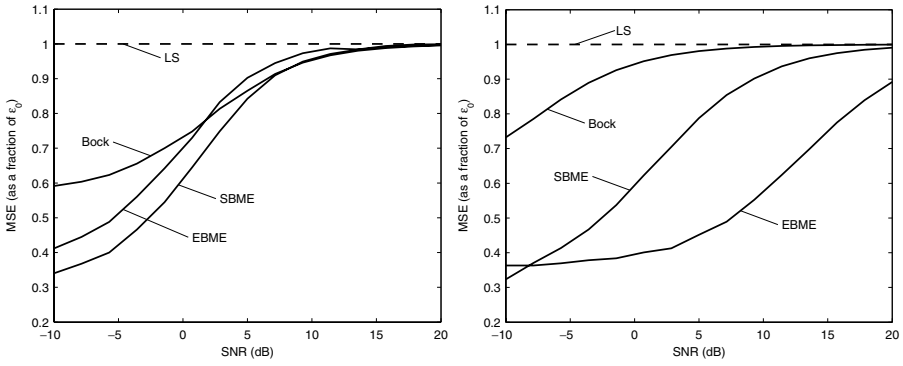


Fig. 4.4 MSE vs. SNR for a typical operating condition: effective dimension 5.8, $m = n = 15$. (a) $\boldsymbol{\theta}_0$ in direction of maximum noise. (b) $\boldsymbol{\theta}_0$ in direction of minimum noise.

be optimal [93], and all other methods converge to the value of the LS estimate; as a result, performance gain is smaller at high SNR, although substantial improvement can be obtained even at 10–15 dB.

To further compare the BMEs with Bock's estimator (4.22), a large set of parameter values $\boldsymbol{\theta}_0$ were generated for different SNRs. For each approach, and for each SNR, the lowest and highest MSE were determined, resulting in a measure of the performance range. This range is displayed in Figure 4.5 for two different choices of $\mathbf{C}$, which are indicated in the figure caption. Evidently, both BMEs outperform Bock's estimator under nearly all circumstances. It is also interesting to note that while the MSE of the EBME is highly dependent on the value of $\boldsymbol{\theta}_0$, the performance of the SBME is fairly constant. This is a result of the symmetric form of the SBME. On the other hand, the EBME achieves considerably lower MSE for most values of $\boldsymbol{\theta}_0$.

It is insightful to compare the performance of the SBME and EBME in Figures 4.4 and 4.5. While the worst-case performance of the two blind minimax techniques is similar, the EBME performs considerably better for some values of $\boldsymbol{\theta}_0$. This is a result of the fact that the EBME selectively shrinks the noisy measurements, whereas the SBME uses an identical shrinkage factor for all elements. If one measurement contains very little noise, the SBME is forced to reduce the shrinkage of all

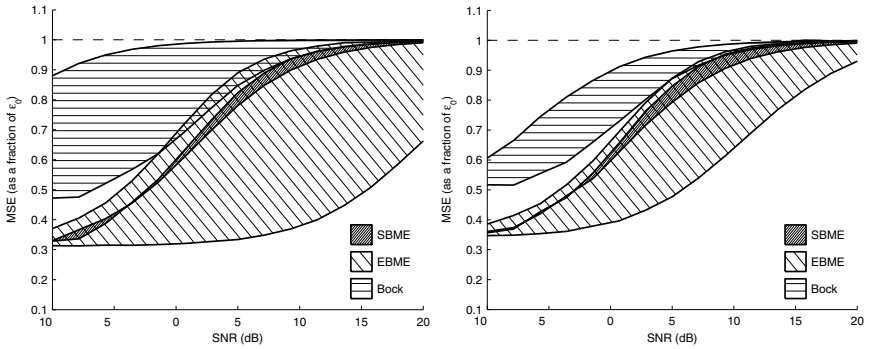


Fig. 4.5 Range of possible MSE values for different values of $\boldsymbol{\theta}_0$, as a function of SNR. $\mathbf{H} = \mathbf{I}$ for both figures. (a) $m = n = 15$, with eigenvalues of $\mathbf{C}$ distributed uniformly between 1 and 0.01, resulting in an effective dimension of 7.6. (b) $m = n = 10$, with $\mathbf{C}$ containing five eigenvalues of 1 and five eigenvalues of 0.1, resulting in an effective dimension of 5.5.

other measurements. The EBME, by contrast, can effectively reduce the effect of noisy measurements without shrinking the clean elements. As a result, the EBME is superior by far if $\boldsymbol{\theta}_0$ is orthogonal to the noisiest measurements; its performance gain is less substantial when $\boldsymbol{\theta}_0$ is in the direction of high shrinkage, since in these cases, shrinkage is applied to the parameter as well as the noise.

Another important advantage of the blind minimax approach over Bock's estimator is that the latter converges to the LS technique when the matrix $\mathbf{Q}$ is ill-conditioned, i.e., when some eigenvalues are much larger than others. This is because the shrinkage in Bock's method (4.22) is a function of $1/\|\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\|_{\mathbf{Q}}^2$. As a result, when $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$ contains a significant component in the direction of a large eigenvalue of $\mathbf{Q}$, shrinkage becomes negligible. Yet, in this case, shrinkage is still desirable for the remaining eigenvalues. This effect is demonstrated in Figure 4.6, which plots the performance of the various approaches for matrices $\mathbf{Q}$ having condition numbers between 1 and 1000. Here, 10 parameters and 10 measurements are used, $\mathbf{H} = \mathbf{I}$, and $\mathbf{C}$ is chosen such that the first five eigenvalues equal 1 and the remaining five eigenvalues equal a value v, selected to obtain the desired condition number. For each condition

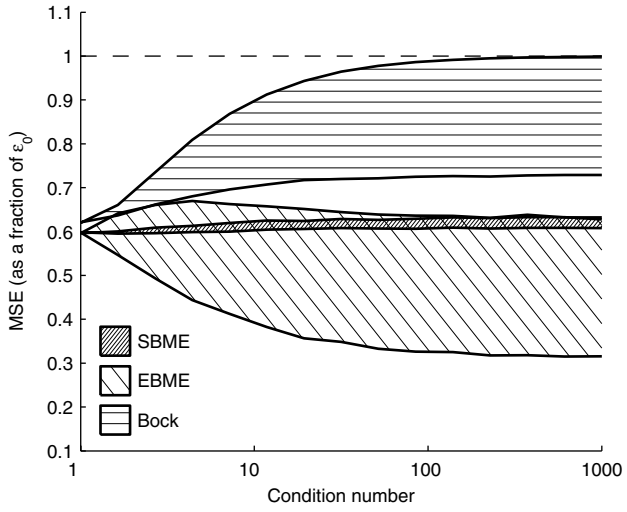


Fig. 4.6 Range of possible MSE values obtained for different values of $\boldsymbol{\theta}_0$, as a function of the condition number of $\mathbf{Q}$, with SNR = 0 dB and $m = n = 10$.

number, a large set of values $\boldsymbol{\theta}_0$ are chosen such that the SNR is 0 dB; as in Figure 4.5, the range of MSE values obtained for each estimate is plotted. It is evident that Bock's estimator approaches the LS method for ill-conditioned matrices. The performance of the EBME improves relative to the LS estimator for ill-conditioned matrices, since the high-noise components are further reduced in this case.

In this section, we explored the idea of blind minimax estimation, whereby one uses linear minimax estimators whose parameter set is itself inferred from measurements. This simple concept was examined in the setting of a linear system of measurements with colored Gaussian noise, where we have shown that the BMEs dominate the LS solution. Consequently, in any such problem, the proposed estimators can be used in place of LS, with a guaranteed performance gain. Apart from being useful in and of themselves, the proposed estimators support the underlying concept of blind minimax estimation. This principle can be applied to many other problems, such as estimation with uncertain system matrices, estimation with non-Gaussian noise, and sequential estimation.

5

The SURE Principle

In this section we continue our exploration of nonlinear ML modifications. The strategy we discuss here is based on the Stein unbiased risk estimate (SURE), which is an unbiased assessment of the MSE. Since the MSE in general depends on the true unknown parameter values it cannot serve as a design objective. Instead, the SURE principle provides a method to directly approximate the MSE of an estimate from the data, without requiring knowledge of the true parameter values. To use this approach as a design method, we choose a class of parameterized estimates, and then seek the values that minimize the SURE objective. We illustrate that this approach can be very effective in choosing regularization parameters in different signal recovery problems, and can outperform standard selection techniques such as cross-validation and the discrepancy method.

5.1 MSE Estimation

A common theme throughout this survey has been the desire to control the MSE of an estimate $\hat{\boldsymbol{\theta}}$. Unfortunately, since the MSE depends explicitly on the unknown parameter vector $\boldsymbol{\theta}_0$, it cannot be used as a

design objective. The idea behind the SURE method, which we outline in this section, is to approximate the MSE of $\hat{\boldsymbol{\theta}}$ from the given data $\mathbf{x}$. This MSE assessment can then be used to select between different estimation strategies.

Suppose we are given some estimate $\hat{\boldsymbol{\theta}}'$ whose performance we would like to improve. For concreteness, and in the spirit of the previous sections, we assume that $\hat{\boldsymbol{\theta}}' = \hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is an ML solution, however the ideas we outline hold more generally. To refine $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$, we consider estimators of the form

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathrm{ML}} + \mathbf{h}(\mathbf{x}), \tag{5.1}$$

for some vector function $\mathbf{h}(\mathbf{x})$, which we would like to choose such that the MSE is minimized. In practice, $\mathbf{h}(\mathbf{x})$ is typically chosen to have a particular structure, parameterized by some vector $\boldsymbol{\alpha}$. For example, $\mathbf{h}(\mathbf{x}) = \alpha\mathbf{x} - \hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ where α is a scalar, or $h_i(\mathbf{x}) = \psi_\alpha(x_i) - \hat{\boldsymbol{\theta}}_{\mathrm{ML}}$, where

$$\psi_\alpha(x) = \mathrm{sign}(x)[|x| - \alpha]_+ \tag{5.2}$$

is a soft-threshold with cut-off α. Ideally, we would like to select $\boldsymbol{\alpha}$ to minimize the MSE. Since this is impossible, in the SURE approach $\boldsymbol{\alpha}$ is designed to minimize an unbiased estimate (referred to as the SURE estimate) of the MSE.

To develop the SURE principle, we first compute the MSE of $\hat{\boldsymbol{\theta}}$ given by (5.1). Denoting by $\epsilon(\boldsymbol{\theta}_0)$ the MSE of $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$,

$$\begin{aligned} E\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2\} &= E\{\|\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}_0 + \mathbf{h}(\mathbf{x})\|^2\} \\ &= \epsilon(\boldsymbol{\theta}_0) + E\{\|\mathbf{h}(\mathbf{x})\|^2\} - 2E\{\mathbf{h}^T(\mathbf{x})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\mathrm{ML}})\}. \end{aligned} \tag{5.3}$$

In order to minimize the MSE over $\mathbf{h}(\mathbf{x})$ we need to evaluate explicitly

$$f(\mathbf{h}, \boldsymbol{\theta}_0) = E\{\|\mathbf{h}(\mathbf{x})\|^2\} - 2E\{\mathbf{h}^T(\mathbf{x})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\mathrm{ML}})\}, \tag{5.4}$$

which unfortunately depends in general on $\boldsymbol{\theta}_0$. Instead, we may seek an unbiased estimate of $f(\mathbf{h}, \boldsymbol{\theta}_0)$ and then choose $\mathbf{h}$ to minimize this estimate. The difficult expression to approximate is $E\{\mathbf{h}^T(\mathbf{x})\boldsymbol{\theta}_0\}$ as the dependency on $\boldsymbol{\theta}_0$ is explicit. Therefore, we concentrate on estimating this term. If this can be done, then we can easily obtain an unbiased

MSE estimate. Indeed, suppose we construct a function $g(\mathbf{h}(\mathbf{x}))$ that depends only on $\mathbf{x}$ (and not on $\boldsymbol{\theta}_0$), such that

$$E\{g(\mathbf{h}(\mathbf{x}))\} = E\{\mathbf{h}^T(\mathbf{x})\boldsymbol{\theta}_0\} \stackrel{\triangle}{=} E_{\mathbf{h},\boldsymbol{\theta}_0}. \tag{5.5}$$

Then

$$\hat{f}(\mathbf{h}) = \|\mathbf{h}(\mathbf{x})\|^2 - 2g(\mathbf{h}(\mathbf{x})) + 2\mathbf{h}^T(\mathbf{x})\hat{\boldsymbol{\theta}}_{\mathrm{ML}}, \tag{5.6}$$

is an unbiased estimate of $f(\mathbf{h},\boldsymbol{\theta}_0)$, since clearly $E\{\hat{f}(\mathbf{h})\} = f(\mathbf{h},\boldsymbol{\theta}_0)$. A reasonable strategy therefore is to select $\mathbf{h}(\mathbf{x})$ to minimize our assessment $\hat{f}(\mathbf{h})$ of the MSE.

The design framework proposed above reduces to finding an unbiased estimate of $E_{\mathbf{h},\boldsymbol{\theta}_0}$. Clearly, any such estimate will depend on the pdf $p(\mathbf{x};\boldsymbol{\theta}_0)$. In this section we broaden our scope with respect to the previous section, and assume that $p(\mathbf{x};\boldsymbol{\theta}_0)$ belongs to the exponential family of distributions (1.4) discussed in Section 1.1. For this class of pdfs we develop an unbiased approximation of $E_{\mathbf{h},\boldsymbol{\theta}_0}$. Before addressing the general setting, we illustrate the main idea by first considering the simpler iid Gaussian case in which we seek to estimate a vector $\boldsymbol{\theta}_0 \in \mathbb{R}^m$ from measurements $\mathbf{x} = \boldsymbol{\theta}_0 + \mathbf{w}$, where $\mathbf{w}$ is a zero-mean Gaussian random vector with iid components of variance σ^2. In Section 5.2.1 we treat the more difficult scenario in which $\phi(\mathbf{x})$ lies in a subspace $\mathcal{A}$ of $\mathbb{R}^m$, and the pdf (1.4) depends on $\boldsymbol{\theta}_0$ only through its orthogonal projection onto $\mathcal{A}$. This situation arises, for example, in the linear Gaussian model (1.2) when $\mathbf{H}$ is rank deficient. For this setup, we develop a SURE estimate of the MSE in estimating the projected parameter.

5.1.1 IID Gaussian Model

In order to develop an unbiased estimate of $E_{\mathbf{h},\boldsymbol{\theta}_0}$ in the iid Gaussian setting, we exploit the fact that for the Gaussian pdf $p(\mathbf{x};\boldsymbol{\theta}_0)$

$$(x_i - \theta_i)p(\mathbf{x};\boldsymbol{\theta}_0) = -\sigma^2 \frac{dp(\mathbf{x};\boldsymbol{\theta}_0)}{dx_i}, \tag{5.7}$$

where θ_i is the ith component of $\boldsymbol{\theta}_0$. Assuming that $E\{|h_i(\mathbf{x})|\}$ is bounded and $h_i(\mathbf{x})$ is weakly differentiable in $\mathbf{x}$, we have

that

$$\begin{aligned} E_{\mathbf{h},\boldsymbol{\theta}_0} &= \sum_{i=1}^{m} \int_{-\infty}^{\infty} h_i(\mathbf{x}) \theta_i p(\mathbf{x};\boldsymbol{\theta}_0) d\mathbf{x} \\ &= \sum_{i=1}^{m} \int_{-\infty}^{\infty} h_i(\mathbf{x}) \left(x_i p(\mathbf{x};\boldsymbol{\theta}_0) + \sigma^2 \frac{dp(\mathbf{x};\boldsymbol{\theta}_0)}{dx_i} \right) d\mathbf{x} \\ &= E\{\mathbf{h}^T(\mathbf{x})\mathbf{x}\} + \sigma^2 \sum_{i=1}^{m} \int_{-\infty}^{\infty} h_i(\mathbf{x}) \frac{dp(\mathbf{x};\boldsymbol{\theta}_0)}{dx_i} d\mathbf{x}, \end{aligned} \tag{5.8}$$

where the second equality is a result of (5.7). To evaluate the second term in (5.8), we use integration by parts:

$$\int_{-\infty}^{\infty} h_i(\mathbf{x}) \frac{dp(\mathbf{x};\boldsymbol{\theta}_0)}{dx_i} d\mathbf{x} = -\int_{-\infty}^{\infty} h_i'(\mathbf{x}) p_i(\mathbf{x};\boldsymbol{\theta}_0) d\mathbf{x} = -E\{h_i'(\mathbf{x})\}, \tag{5.9}$$

where we denoted $h_i'(\mathbf{x}) = dh_i(\mathbf{x})/dx_i$, and used the fact that $|h_i(\mathbf{x})p(\mathbf{x};\boldsymbol{\theta}_0)| \to 0$ for $|x_i| \to \infty$ since $E\{|h_i(\mathbf{x})|\}$ is bounded. We conclude from (5.8) and (5.9) that

$$E_{\mathbf{h},\boldsymbol{\theta}_0} = E\{\mathbf{h}^T(\mathbf{x})\mathbf{x}\} - \sigma^2 \sum_{i=1}^{m} E\{h_i'(\mathbf{x})\} \tag{5.10}$$

and therefore, $\mathbf{h}^T(\mathbf{x})\mathbf{x} - \sigma^2 \sum_{i=1}^{m} h_i'(\mathbf{x})$ is an unbiased estimate of $E_{\mathbf{h},\boldsymbol{\theta}_0}$. Plugging this expression into (5.3), and using the fact that in our setting $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \mathbf{x}$, we arrive at the following SURE assessment of the MSE:

$$\epsilon(\boldsymbol{\theta}_0) + \|\mathbf{h}(\mathbf{x})\|^2 + 2\sigma^2 \sum_{i=1}^{m} \frac{dh_i(\mathbf{x})}{dx_i}. \tag{5.11}$$

The MSE estimate (5.11) was first proposed by Stein in [129, 130].

We next illustrate how (5.11) can be used to design estimates.

Example 5.1. Suppose we are given Gaussian measurements $\mathbf{x} = \boldsymbol{\theta}_0 + \mathbf{w}$ from which we want to recover $\boldsymbol{\theta}_0$. We consider shrinkage estimates of the form $\hat{\boldsymbol{\theta}} = \alpha\mathbf{x}$ and want to select a good choice of α. To this end, we propose minimizing the SURE estimate of the MSE (5.11). Since $\hat{\boldsymbol{\theta}}$ corresponds to $\mathbf{h}(\mathbf{x}) = (\alpha - 1)\mathbf{x}$, the optimal value $\widehat{\alpha}$ minimizes

$$S(\alpha) = (1-\alpha)^2 \|\mathbf{x}\|^2 + 2m\sigma^2(\alpha - 1), \tag{5.12}$$

and is given by

$$\widehat{\alpha} = 1 - \frac{m\sigma^2}{\|\mathbf{x}\|^2}. \tag{5.13}$$

The resulting estimate is

$$\hat{\boldsymbol{\theta}} = \left(1 - \frac{m\sigma^2}{\|\mathbf{x}\|^2}\right)\mathbf{x}, \tag{5.14}$$

which coincides with the Stein method (4.19) [128], discussed in Section 4.2. This provides further justification for this approach. If in addition we require that $\alpha \geq 0$, then

$$\hat{\boldsymbol{\theta}} = \left[1 - \frac{m\sigma^2}{\|\mathbf{x}\|^2}\right]_+ \mathbf{x}, \tag{5.15}$$

which is equal to the positive-part Stein estimate (4.20).

The SURE strategy for the iid-Gaussian case has been applied to a variety of different denoising techniques [13, 33, 39, 102, 147]. The difference between the proposed recovery strategies is in the parametrization of $\mathbf{h}(\mathbf{x})$. For example, in the context of wavelet denoising [33], $\boldsymbol{\theta}_0$ represents the wavelet coefficients of some underlying signal. Motivated by the observation that these coefficients are often sparse, it was suggested in [33] to recover $\boldsymbol{\theta}_0$ using a component-wise soft-threshold corresponding to the choice $\hat{\theta}_i = \psi_\alpha(x_i)$, where $\psi_\alpha(x)$ is given by (5.2). The popular SUREShrink wavelet denoising strategy results when α is selected to minimize[1] (5.11). In [102], the parametrization in the wavelet domain was chosen as $\hat{\theta}_i = \sum_{j=1}^{k} a_j \phi_j(x_i)$, where $\phi_j(x)$ are given nonlinear functions of x, k is the number of parameters, and a_j are the coefficients that are optimized by SURE.

5.2 Generalized SURE Principle

In the previous section we illustrated the use of the SURE denoising strategy in the context of an iid Gaussian problem. Extensions to independent variables from an exponential family are treated in [14, 86, 87].

[1] More precisely, in SUREShrink α is determined by SURE only if it is lower than some upper limit.

All of these generalizations are confined to the independent case which precludes a variety of important applications such as image deblurring.

We now extend the approach, following the results of [39], to the general class of exponential pdfs

$$f(\mathbf{x};\boldsymbol{\theta}_0) = r(\mathbf{x})\exp\{\boldsymbol{\theta}_0^T\phi(\mathbf{x}) - g(\boldsymbol{\theta}_0)\}. \tag{5.16}$$

In order to address this model, some modifications to the basic technique outlined in the previous section are necessary. First, we note that a sufficient statistic for estimating $\boldsymbol{\theta}_0$ in the model (5.16) is [100]

$$\mathbf{u} = \phi(\mathbf{x}). \tag{5.17}$$

Therefore, any reasonable estimate of $\boldsymbol{\theta}_0$ will be a function of $\mathbf{u}$. More specifically, from the Rao–Blackwell theorem [93] it follows that if $\hat{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}_0$ which is not only a function of $\mathbf{u}$, then the estimate $E\{\hat{\boldsymbol{\theta}}|\mathbf{u}\}$ has smaller or equal MSE than $\hat{\boldsymbol{\theta}}$, for all $\boldsymbol{\theta}_0$. Therefore, in the sequel, we only consider methods that depend on the data via $\mathbf{u}$. This enables the use of integration by parts, similar to the iid Gaussian setting for which $\mathbf{u} = \phi(\mathbf{x}) = (1/\sigma^2)\mathbf{x}$. Note that $\mathbf{u}$ and $\boldsymbol{\theta}_0$ have the same length.

For our class of estimates, we choose

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathrm{ML}} + \mathbf{h}(\mathbf{u}), \tag{5.18}$$

for some function $\mathbf{h}(\mathbf{u})$. Note that from (5.16), $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is a solution of

$$\frac{dg(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \phi(\mathbf{x}) = \mathbf{u}, \tag{5.19}$$

and therefore depends only on $\mathbf{u}$. The MSE of $\hat{\boldsymbol{\theta}}$ is computed as in (5.3):

$$E\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2\} = \epsilon(\boldsymbol{\theta}_0) + E\{\|\mathbf{h}(\mathbf{u})\|^2\} - 2E\{\mathbf{h}^T(\mathbf{u})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\mathrm{ML}})\}. \tag{5.20}$$

To assess the MSE, we seek an unbiased estimate of $E\left\{\mathbf{h}^T(\mathbf{u})\boldsymbol{\theta}_0\right\}$. The following theorem provides such an estimate, using similar ideas to those used in the previous section [39].

Theorem 5.1. Let $\mathbf{x}$ denote a random vector with exponential pdf given by (5.16), and let $\mathbf{u} = \phi(\mathbf{x}) \in \mathbb{R}^m$ be a sufficient statistic for estimating $\boldsymbol{\theta}_0 \in \mathbb{R}^m$ from $\mathbf{x}$. Let $\mathbf{h}(\mathbf{u})$ be an arbitrary function of $\mathbf{u}$ that is weakly differentiable in $\mathbf{u}$ and such that $E\{|h_i(\mathbf{u})|\}$ is bounded. Then

$$E\left\{\mathbf{h}^T(\mathbf{u})\boldsymbol{\theta}_0\right\} = -\sum_{i=1}^{m} E\left\{\frac{dh_i(\mathbf{u})}{du_i}\right\} - E\left\{\mathbf{h}^T(\mathbf{u})\frac{d\ln q(\mathbf{u})}{d\mathbf{u}}\right\}, \quad (5.21)$$

where

$$q(\mathbf{u}) = \int r(\mathbf{x})\delta(\mathbf{u} - \phi(\mathbf{x}))d\mathbf{x}, \quad (5.22)$$

and $\delta(\mathbf{x})$ is the Dirac delta function. Therefore,

$$-\sum_{i=1}^{m} \frac{dh_i(\mathbf{u})}{du_i} - \mathbf{h}^T(\mathbf{u})\frac{d\ln q(\mathbf{u})}{d\mathbf{u}} \quad (5.23)$$

is an unbiased estimate of $E\{\mathbf{h}^T(\mathbf{u})\boldsymbol{\theta}_0\}$.

Note that the pdf $f(\mathbf{u};\boldsymbol{\theta}_0)$ of $\mathbf{u}$ is given by

$$f(\mathbf{u};\boldsymbol{\theta}_0) = \exp\{\boldsymbol{\theta}_0^T\mathbf{u} - g(\boldsymbol{\theta}_0)\}q(\mathbf{u}). \quad (5.24)$$

Therefore, an alternative to computing $q(\mathbf{u})$ using (5.22) is to evaluate the pdf of $\mathbf{u}$ and then use (5.24).

Based on Theorem 5.1 we can develop a generalized SURE principle for estimating an unknown parameter vector $\boldsymbol{\theta}_0$ in an exponential model. Specifically, let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathrm{ML}} + \mathbf{h}(\mathbf{u})$ be an arbitrary estimate of $\boldsymbol{\theta}_0$ where $\mathbf{h}(\mathbf{u})$ satisfies the regularity conditions of Theorem 5.1. Then, combining (5.20) and Theorem 5.1, an unbiased estimate of the MSE of $\hat{\boldsymbol{\theta}}$ is given by

$$\epsilon(\boldsymbol{\theta}_0) + \|\mathbf{h}(\mathbf{u})\|^2 + 2\sum_{i=1}^{m} \frac{dh_i(\mathbf{u})}{du_i} + 2\mathbf{h}^T(\mathbf{u})\left(\frac{d\ln q(\mathbf{u})}{d\mathbf{u}} + \hat{\boldsymbol{\theta}}_{\mathrm{ML}}\right). \quad (5.25)$$

We may then design $\hat{\boldsymbol{\theta}}$ by choosing $\mathbf{h}(\mathbf{u})$ to minimize (5.25). We refer to this technique as the generalized SURE principle.

To recover the iid Gaussian case, note that in this setting $\mathbf{u} = (1/\sigma^2)\mathbf{x}$. Therefore, $d\ln q(\mathbf{u})/d\mathbf{u} = -\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = -\mathbf{x}$, and $dh_i(\mathbf{u})/du_i = \sigma^2 h_i'(\mathbf{x})$. With these relations, (5.25) reduces to (5.11).

5.2.1 Rank-Deficient Models

In some settings, the sufficient statistic $\mathbf{u}$ lies in a subspace $\mathcal{A}$ of $\mathbb{R}^m$. As an example, suppose that in the Gaussian model (1.2) $\mathbf{H}$ is rank-deficient. In this case $\mathbf{u} = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$ lies in the range space $\mathcal{R}(\mathbf{H}^T)$ of $\mathbf{H}^T$, which is a subspace of $\mathbb{R}^m$. If $\boldsymbol{\theta}_0$ is not restricted to a subspace, then we do not expect to be able to reliably estimate $\boldsymbol{\theta}_0$ from $\mathbf{u}$, unless some additional information on $\boldsymbol{\theta}_0$ is known. Nonetheless, we may still obtain a reliable assessment of the part of $\boldsymbol{\theta}_0$ that lies in $\mathcal{A}$.

Denote by $\mathbf{P}$ the orthogonal projection onto $\mathcal{A}$. The MSE in estimating $\boldsymbol{\theta}_0$ can be written as

$$E\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2\} = E\{\|\mathbf{P}\hat{\boldsymbol{\theta}} - \mathbf{P}\boldsymbol{\theta}_0\|^2\} + E\{\|(\mathbf{I}-\mathbf{P})\hat{\boldsymbol{\theta}} - (\mathbf{I}-\mathbf{P})\boldsymbol{\theta}_0\|^2\}. \tag{5.26}$$

As we show below, if $\mathbf{u}$ depends on $\boldsymbol{\theta}_0$ only through $\mathbf{P}\boldsymbol{\theta}_0$, and in addition $\mathbf{u}$ has an exponential pdf, then we can obtain a SURE estimate of the error in $\mathcal{A}$, defined by

$$\mathrm{MSE}_{\mathcal{A}} = E\{\|\mathbf{P}\hat{\boldsymbol{\theta}} - \mathbf{P}\boldsymbol{\theta}_0\|^2\}. \tag{5.27}$$

If $\hat{\boldsymbol{\theta}}$ lies in $\mathcal{A}$, then $(\mathbf{I}-\mathbf{P})\hat{\boldsymbol{\theta}} = \mathbf{0}$ and the second term in (5.26) is constant, independent of $\hat{\boldsymbol{\theta}}$. Therefore, up to a constant, an approximation of $\mathrm{MSE}_{\mathcal{A}}$ is also an unbiased estimate of the true MSE $E\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2\}$ so that to optimize $\hat{\boldsymbol{\theta}}$ it is sufficient to estimate $\mathrm{MSE}_{\mathcal{A}}$. Even if $\hat{\boldsymbol{\theta}}$ does not lie in $\mathcal{A}$, the SURE estimate we develop may be used to approximate $\mathrm{MSE}_{\mathcal{A}}$. Since $\mathbf{u}$ depends only on $\mathbf{P}\boldsymbol{\theta}_0$, it is reasonable to restrict attention to estimates $\hat{\boldsymbol{\theta}} = \mathbf{h}_{\boldsymbol{\alpha}}(\mathbf{u})$, where the parameters $\boldsymbol{\alpha}$ are tuned to minimize $\mathrm{MSE}_{\mathcal{A}}$, subject to any other prior information we may have, such as norm constraints on $\boldsymbol{\theta}_0$. In such cases we can use a regularized SURE criterion with the SURE objective being an unbiased estimate of $\mathrm{MSE}_{\mathcal{A}}$ and the regularization term provided by the prior information, as discussed further in [39].

To derive a SURE estimate of $\mathrm{MSE}_{\mathcal{A}}$ we first note that if $\mathbf{u}$ lies in $\mathcal{A}$, then

$$\boldsymbol{\theta}_0^T\mathbf{u} = (\mathbf{P}\boldsymbol{\theta}_0)^T(\mathbf{P}\mathbf{u}). \tag{5.28}$$

Suppose that $\mathcal{A}$ has dimension $r < m$. Since $\mathbf{P}\boldsymbol{\theta}_0$ lies in an r-dimensional space, it can be expressed in terms of r components in

an appropriate basis. Denoting by $\mathbf{V}$ an $m \times r$ matrix with orthonormal vectors that span $\mathcal{A} = \mathcal{R}(\mathbf{P})$, the vector $\mathbf{P}\boldsymbol{\theta}_0$ can be expressed as $\mathbf{P}\boldsymbol{\theta}_0 = \mathbf{V}\boldsymbol{\theta}_0'$ for an appropriate length-r vector $\boldsymbol{\theta}_0'$. Similarly, $\mathbf{P}\mathbf{u} = \mathbf{V}\mathbf{u}'$. Therefore, we can write $\boldsymbol{\theta}_0^T\mathbf{u} = \boldsymbol{\theta}_0'^T\mathbf{u}'$, where we used the fact that $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. We assume that $\mathbf{u}'$ is a sufficient statistic for $\boldsymbol{\theta}_0'$ and that $f(\mathbf{u}';\boldsymbol{\theta}_0')$ has an exponential pdf:

$$f(\mathbf{u}';\boldsymbol{\theta}_0') = q(\mathbf{u}')\exp\{\boldsymbol{\theta}_0'^T\mathbf{u}' - g(\boldsymbol{\theta}_0')\}. \tag{5.29}$$

Let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathrm{ML}} + \mathbf{h}(\mathbf{u})$, where $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is some ML estimate (since the model is rank deficient the ML solution may not be unique). Then under the model (5.29) it can be shown [39] that an unbiased estimate of the MSE $E\{\|\mathbf{P}\hat{\boldsymbol{\theta}} - \mathbf{P}\boldsymbol{\theta}_0\|^2\}$ is given by

$$\epsilon_{\mathcal{A}}(\boldsymbol{\theta}_0) + \|\mathbf{P}\mathbf{h}(\mathbf{u})\|^2 + 2\mathrm{Tr}\left(\mathbf{P}\frac{d\mathbf{h}(\mathbf{u})}{d\mathbf{u}}\right) + 2\mathbf{h}^T(\mathbf{u})\left(\mathbf{V}\frac{d\ln q(\mathbf{u}')}{d\mathbf{u}'} + \hat{\boldsymbol{\theta}}_{\mathrm{ML}}\right), \tag{5.30}$$

with $\mathbf{u} = \mathbf{V}\mathbf{u}'$ and $\mathbf{V}$ denoting an orthonormal basis for $\mathcal{A}$, $\mathbf{P} = \mathbf{V}\mathbf{V}^T$ and $\epsilon_A(\boldsymbol{\theta}_0) = E\{\|\mathbf{P}\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \mathbf{P}\boldsymbol{\theta}_0\|^2\}$. When $\mathcal{A} = \mathbb{R}^m$, $\mathbf{P} = \mathbf{I}, \mathbf{V} = \mathbf{I}$ and (5.30) reduces to (5.25). The proof of (5.30) follows from noting that

$$E\{\mathbf{h}^T(\mathbf{u})\mathbf{P}\boldsymbol{\theta}_0\} = E\{\mathbf{h}^T(\mathbf{u})\mathbf{V}\boldsymbol{\theta}_0'\} = E\{(\mathbf{V}^T\mathbf{h}(\mathbf{u}))^T\boldsymbol{\theta}_0'\}, \tag{5.31}$$

and applying Theorem 5.1 to $\mathbf{V}^T\mathbf{h}(\mathbf{u})$.

5.2.2 Linear Gaussian Model

We now specialize the SURE principle to the linear Gaussian model (1.2). We begin by treating the case in which $\mathbf{H}$ is an $n \times m$ matrix with $n \geq m$ and full column rank. We then discuss the rank-deficient scenario.

To use Theorem 5.1 we need to compute the pdf $q(\mathbf{u})$ of $\mathbf{u}$. Since $\mathbf{u} = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$, it is a Gaussian random vector with mean $\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\boldsymbol{\theta}_0$ and covariance $\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}$. As $q(\mathbf{u})$ is the function multiplying the exponential in the pdf of $\mathbf{u}$, it follows that

$$q(\mathbf{u}) = K\exp\{-(1/2)\mathbf{u}^T(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{u}\}, \tag{5.32}$$

where K is a constant, independent of $\mathbf{u}$. Therefore,

$$\frac{d\ln q(\mathbf{u})}{d\mathbf{u}} = -(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{u} = -\hat{\boldsymbol{\theta}}_{\mathrm{ML}}, \tag{5.33}$$

where $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is the ML estimate of $\boldsymbol{\theta}_0$ given by

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}. \tag{5.34}$$

It then follows from Theorem 5.1 that

$$E\{\mathbf{h}^T(\mathbf{u})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{\mathrm{ML}})\} = -\sum_{i=1}^{m} E\left\{\frac{dh_i(\mathbf{u})}{du_i}\right\}. \tag{5.35}$$

An unbiased estimate of the MSE of $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathrm{ML}} + \mathbf{h}(\mathbf{u})$ is

$$\epsilon(\boldsymbol{\theta}_0) + \|\mathbf{h}(\mathbf{u})\|^2 + 2\sum_{i=1}^{m} \frac{dh_i(\mathbf{u})}{du_i}. \tag{5.36}$$

Next, suppose that $\mathbf{H}$ is rank deficient. In this case, the covariance of $\mathbf{u}$ is not invertible and therefore $q(\mathbf{u})$ can no longer be written as in (5.32). Instead, we use the results of Section 5.2.1 and consider a sufficient statistic for estimating the projection of $\boldsymbol{\theta}_0$ onto $\mathcal{R}(\mathbf{H}^T)$.

Let $\mathbf{H}$ have a singular value decomposition $\mathbf{H} = \mathbf{U}\Sigma\mathbf{Q}^T$ for some unitary matrices $\mathbf{U}$ and $\mathbf{Q}$ and Σ a diagonal $n \times m$ matrix with the first r diagonal elements equal to $\sigma_i > 0$ and the remaining elements equal 0 so that $\mathbf{H}$ has rank r. In this case, $\mathbf{V}$ is equal to the first r columns of $\mathbf{Q}$ and $\boldsymbol{\theta}'_0 = \mathbf{V}^T\boldsymbol{\theta}_0$. A sufficient statistic for estimating $\boldsymbol{\theta}'_0$ is $\mathbf{u}' = \mathbf{V}^T\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$. Indeed, $\mathbf{u}'$ is a Gaussian random vector with mean $\mu' = \mathbf{V}^T\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\boldsymbol{\theta}_0$ and covariance $\mathbf{C}' = \mathbf{V}^T\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\mathbf{V}$. Using the SVD of $\mathbf{H}$ we have that

$$\begin{aligned} \mu' &= \Lambda[\mathbf{U}^T\mathbf{C}^{-1}\mathbf{U}]_r\boldsymbol{\theta}'_0, \\ \mathbf{C}' &= \Lambda[\mathbf{U}^T\mathbf{C}^{-1}\mathbf{U}]_r, \end{aligned} \tag{5.37}$$

where Λ is an $r \times r$ diagonal matrix with diagonal elements $\sigma_i^2 > 0$ and $[\mathbf{A}]_r$ is the $r \times r$ top-left principle block of size r of the matrix $\mathbf{A}$. Since $\mathbf{C} \succ 0$, $\mathbf{C}'$ is invertible. Therefore, $f(\mathbf{u}';\boldsymbol{\theta}'_0)$ has the form (5.29) with

$$q(\mathbf{u}') = K\exp\{-(1/2)\mathbf{u}'^T\mathbf{C}'^{-1}\mathbf{u}'\}, \tag{5.38}$$

where K is a constant. Using (5.30) together with the fact that

$$\mathbf{V}\mathbf{C}'^{-1}\mathbf{u}' = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{\dagger}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x} = \hat{\boldsymbol{\theta}}_{\mathrm{ML}}, \tag{5.39}$$

leads to the following proposition.

Proposition 5.2. Let $\mathbf{x}$ denote measurements of an unknown parameter vector $\boldsymbol{\theta}_0$ in the linear Gaussian model (1.2), where $\mathbf{w}$ is a zero-mean Gaussian random vector with covariance $\mathbf{C} \succ 0$. Let $\mathbf{h}(\mathbf{u})$ with $\mathbf{u} = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$ be an arbitrary function of $\boldsymbol{\theta}_0$ that is weakly differentiable in $\mathbf{u}$ and such that $E\{|h_i(\mathbf{u})|\}$ is bounded, and let $\mathbf{P}$ be an orthogonal projection onto $\mathcal{R}(\mathbf{H}^T)$. Then

$$E\{\mathbf{h}^T(\mathbf{u})\mathbf{P}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\mathrm{ML}})\} = -E\left\{\mathrm{Tr}\left(\mathbf{P}\frac{d\mathbf{h}(\mathbf{u})}{d\mathbf{u}}\right)\right\},$$

where

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{\dagger}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$$

is an ML estimate of $\boldsymbol{\theta}_0$. An unbiased estimate of the MSE $E\{\|\mathbf{P}\hat{\boldsymbol{\theta}} - \mathbf{P}\boldsymbol{\theta}_0\|^2\}$ is

$$S(\mathbf{h}) = \epsilon_{\mathcal{A}}(\boldsymbol{\theta}_0) + \|\mathbf{P}\mathbf{h}(\mathbf{u})\|^2 + 2\mathrm{Tr}\left(\mathbf{P}\frac{d\mathbf{h}(\mathbf{u})}{d\mathbf{u}}\right), \tag{5.40}$$

where $\epsilon_A(\boldsymbol{\theta}_0) = E\{\|\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \mathbf{P}\boldsymbol{\theta}_0\|^2\}$.

The next example extends Example 5.1 to the non-iid setting.

Example 5.2. Suppose we are given measurements $\mathbf{x}$ that obey the linear Gaussian model (1.2). We seek an estimate of $\boldsymbol{\theta}_0$ of the form $\hat{\boldsymbol{\theta}} = \alpha\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$, where $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is given by (5.34). To choose α, we minimize the SURE unbiased MSE estimate (5.40). Note that in this case $\mathbf{h}(\mathbf{u}) = (\alpha - 1)\hat{\boldsymbol{\theta}}_{\mathrm{ML}} \in \mathcal{R}(\mathbf{H}^T)$ so that $S(\mathbf{h}) + \|(\mathbf{I} - \mathbf{P})\boldsymbol{\theta}_0\|^2$ is an unbiased estimate of the total MSE $E\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2\}$ and therefore it suffices to minimize $S(\mathbf{h})$, which is equivalent to minimizing

$$(1 - \alpha)^2\|\hat{\boldsymbol{\theta}}_{\mathrm{ML}}\|^2 + 2(\alpha - 1)\mathrm{Tr}((\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{\dagger}). \tag{5.41}$$

The optimal α is given by

$$\widehat{\alpha} = 1 - \frac{\mathrm{Tr}\left((\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{\dagger}\right)}{\|\hat{\boldsymbol{\theta}}_{\mathrm{ML}}\|^2}. \tag{5.42}$$

The resulting estimate is

$$\hat{\boldsymbol{\theta}} = \left(1 - \frac{\mathrm{Tr}\left((\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^\dagger\right)}{\|\hat{\boldsymbol{\theta}}_{\mathrm{ML}}\|^2}\right)\hat{\boldsymbol{\theta}}_{\mathrm{ML}}, \tag{5.43}$$

which coincides with the balanced blind minimax method proposed in (4.43), based on a minimax framework. Here we see that the same technique results from applying the generalized SURE criterion. If in addition we require that $\alpha \geq 0$, then (5.43) becomes

$$\hat{\boldsymbol{\theta}} = \left[1 - \frac{\mathrm{Tr}\left((\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^\dagger\right)}{\|\hat{\boldsymbol{\theta}}_{\mathrm{ML}}\|^2}\right]_+ \hat{\boldsymbol{\theta}}_{\mathrm{ML}}, \tag{5.44}$$

which is the positive-part balanced method of (4.46).

5.3 Application to Regularization Selection

A popular strategy for solving inverse problems of the form (1.2) is to use regularization techniques in conjunction with a LS objective. Specifically, the estimate $\hat{\boldsymbol{\theta}}$ is chosen to minimize a regularized LS criterion:

$$(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}})\mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}) + \lambda\|\mathbf{L}\hat{\boldsymbol{\theta}}\|, \tag{5.45}$$

where the norm is arbitrary. Here $\mathbf{L}$ is some regularization operator such as the discretization of a first- or second-order differential operator that accounts for smoothness properties of $\boldsymbol{\theta}_0$, and λ is the regularization parameter [72, 73]. An important problem in practice is the selection of λ, which strongly effects the recovery performance. One of the most popular approaches to choosing λ when the estimate is linear (as is the case when a squared-ℓ_2 norm is used in (5.45)) is the generalized cross-validation (GCV) method [64]. When the estimate takes on a more complicated nonlinear form, a popular selection method is the discrepancy principle [61].

Based on the generalized SURE criterion, we choose λ to minimize the SURE objective (5.40). This allows SURE-based optimization of a broad class of deblurring and deconvolution methods including both linear and nonlinear techniques. As we demonstrate for the cases in which the norm in (5.45) is the squared-ℓ_2 or ℓ_1 norms, this method

can dramatically outperform GCV and the discrepancy technique in practical applications. When the estimate is not given explicitly but rather as a solution of an optimization problem we can still employ the SURE strategy by using a Monte-Carlo approach to approximate the derivative of the estimate, which figures in the SURE expression [118]. Specifically,

$$\sum_{i=1}^{m} \frac{dh_i(\mathbf{u})}{du_i} \approx \frac{1}{\epsilon^2}\mathbf{z}^T(\mathbf{h}(\mathbf{u}+\mathbf{z}) - \mathbf{h}(\mathbf{u})), \tag{5.46}$$

where ϵ is a small constant and $\mathbf{z}$ is a zero-mean iid random vector, independent of $\mathbf{u}$, with covariance $\epsilon^2\mathbf{I}$.

Using several test images and a deconvolution problem, taken from [39], we demonstrate below that this strategy often leads to significant performance improvement over the standard GCV and discrepancy selection criteria in the context of image deblurring and deconvolution.

5.3.1 Image Deblurring

We first consider the case in which the squared-ℓ_2 norm is used in (5.45). The solution then has the form:

$$\hat{\boldsymbol{\theta}} = (\mathbf{Q} + \lambda\mathbf{L}^T\mathbf{L})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}, \tag{5.47}$$

where for brevity we denoted

$$\mathbf{Q} = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}. \tag{5.48}$$

The estimate (5.47) is commonly referred to as Tikhonov regularization [136].

In the GCV method, λ is chosen to minimize

$$G(\lambda) = \frac{1}{\mathrm{Tr}^2(\mathbf{I} - (\mathbf{Q} + \lambda\mathbf{L}^T\mathbf{L})^{-1}\mathbf{Q})} \sum_{i=1}^{n} (\mathbf{x}_i - [\mathbf{H}\hat{\boldsymbol{\theta}}]_i)^2. \tag{5.49}$$

To apply the SURE criterion, we rewrite the estimate (5.47) as $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathrm{ML}} + \mathbf{h}(\mathbf{u})$, where $\mathbf{u} = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$ and

$$\mathbf{h}(\mathbf{u}) = -\lambda(\mathbf{Q} + \lambda\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T\mathbf{L}\mathbf{Q}^{-1}\mathbf{u}. \tag{5.50}$$

We then suggest choosing the value of λ that minimizes the SURE objective (5.40), which is equivalent to minimizing

$$S(\lambda) = \|\mathbf{h}(\mathbf{u})\|^2 - 2\lambda \mathrm{Tr}\left((\mathbf{Q} + \lambda \mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T\mathbf{L}\mathbf{Q}^{-1}\right). \tag{5.51}$$

The optimal value can be determined numerically. This approach was first studied in the special case of Tikhonov regularization with white noise in [32, 61, 122]. In our simulations below, minimization of the GCV and SURE objectives were performed by using the `fmincon` function on Matlab.

We now demonstrate the performance of the SURE-based regularization method with examples taken from [39]. Specifically, we consider an image deblurring problem using the HNO deblurring package for Matlab[2] based on [76]. We chose several test images, and blurred them using a Gaussian point-spread function of dimension 9 with standard deviation 6. We then added zero-mean, Gaussian white noise with variance σ^2. In Figures 5.1 and 5.2 we compare the deblurred images resulting from using the Tikhonov estimate (5.47) with $\mathbf{L} = \mathbf{I}$ where the regularization parameter is chosen according to our new SURE criterion (left) and the GCV method (right), for different noise levels.

As can be seen from the figures, the SURE based approach leads to a substantial performance improvement over the standard GCV criterion. This can also be seen in Tables 5.1 and 5.2 in which we report the resulting MSE values.

5.3.2 Deconvolution Example

As another application of the SURE, consider the standard deconvolution problem in which a signal $\theta[\ell]$ is convolved by an impulse response $h[\ell]$ and contaminated by additive white Gaussian noise with variance σ^2. The observations $x[\ell]$ can be written in the form of the linear model (1.2) where $\mathbf{x}$ is the vector containing the observations $x[\ell]$, $\boldsymbol{\theta}_0$ consists of the input signal $\theta[\ell]$, and $\mathbf{H}$ is a Toeplitz matrix, representing convolution with the impulse response $h[\ell]$.

To recover $\theta[\ell]$ we may use a penalized LS approach (5.45) where we assume that the original signal $\theta[\ell]$ is smooth. This can be accounted

[2] The package is available at http://www2.imm.dtu.dk/~pch/HNO/.

Fig. 5.1 Deblurring of Lena using Tikhonov regularization with SURE (left) and GCV (right) choices of regularization and different noise levels: (a), (b) $\sigma = 0.01$; (c), (d) $\sigma = 0.05$; (e), (f) $\sigma = 0.1$.

for by choosing a penalization of the form $\|\mathbf{L}\boldsymbol{\theta}\|_1$, where $\mathbf{L}$ represents a second order derivative operator. The resulting penalized LS estimate can be determined by solving a quadratic optimization problem. In our simulations, we used `CVX`, a package for specifying and solving convex programs in Matlab [68].

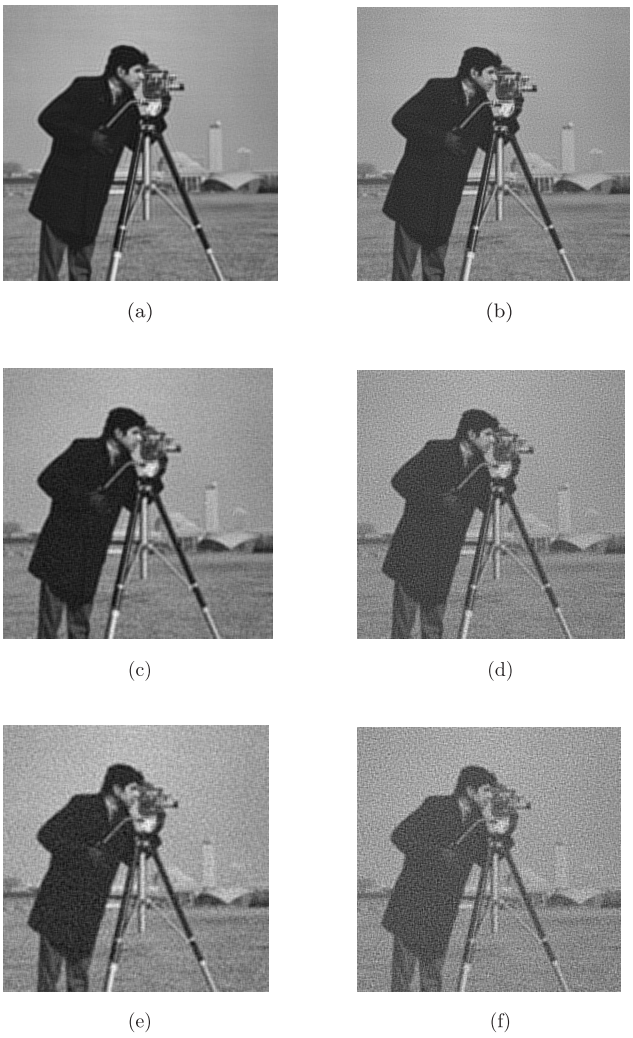


Fig. 5.2 Deblurring of Cameraman using Tikhonov regularization with SURE (left) and GCV (right) choices of regularization and different noise levels: (a), (b) $\sigma = 0.01$; (c), (d) $\sigma = 0.05$; (e), (f) $\sigma = 0.1$.

Since the resulting estimate is non-linear, due to the ℓ_1 penalization, we cannot apply the GCV equation (5.49). Instead, a popular approach to tune the parameter λ is to use the discrepancy principle in which λ is chosen such that the residual $\|\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}\|^2$ is equal to the noise level $n\sigma^2$ [32, 61].

Table 5.1 MSE for Tikhonov Deblurring of Lena.

	$\sigma = 0.01$	$\sigma = 0.05$	$\sigma = 0.1$
GCV	0.0022	0.0077	0.0133
SURE	0.0011	0.0025	0.0042

Table 5.2 MSE for Tikhonov Deblurring of Cameraman.

	$\sigma = 0.01$	$\sigma = 0.05$	$\sigma = 0.1$
GCV	0.0033	0.0121	0.0221
SURE	0.0016	0.0039	0.0064

To evaluate the performance of the SURE principle in this context, we consider an example from the Regularization Tools [74] for Matlab, also taken from [39]. All the problems in this toolbox are discretized versions of the Fredholm integral equation of the first kind:

$$g(s) = \int_a^b K(s,t)\theta(t)dt, \tag{5.52}$$

where $K(s,t)$ is the kernel and $\theta(t)$ is the solution for a given $g(s)$. The problem is to estimate $\theta(t)$ from noisy samples of $g(s)$. Using a midpoint rule with n points, (5.52) reduces to an $n \times n$ linear system

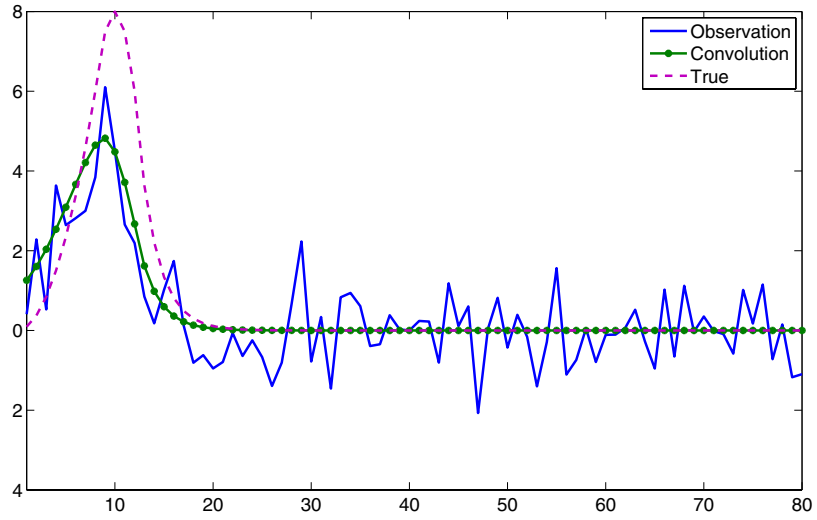


Fig. 5.3 The original signal $\boldsymbol{\theta}_0$ (dashed), the clean convolved signal (star) and the observations $\mathbf{x}$ with $\sigma = 1$.

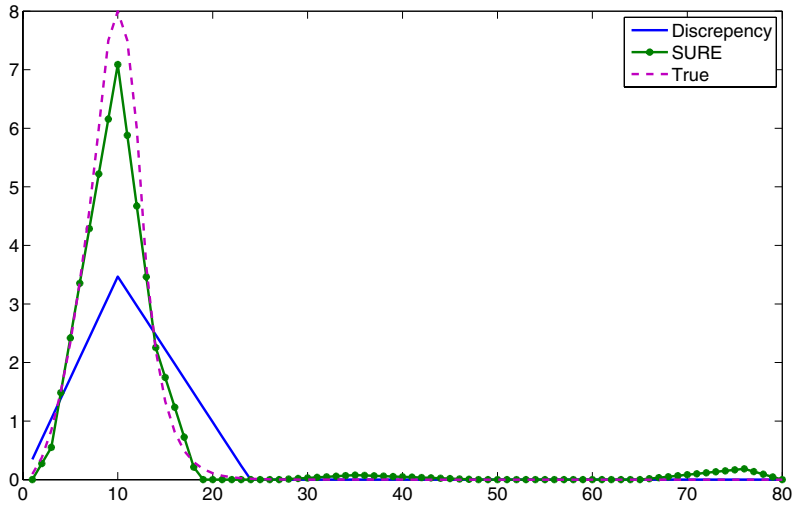


Fig. 5.4 Deconvolution using weighted ℓ_1 regularization with the discrepancy principle, SURE (star) and the original signal $\boldsymbol{\theta}_0$ (dashed) with $\sigma = 1$.

$\mathbf{x}_T = \mathbf{H}\boldsymbol{\theta}_0$. The functions in this toolbox differ in $K(t,s)$ and $\theta(s)$. Below we consider the function heat(n) with $n = 80$. The output of the function is the matrix $\mathbf{H}$ and the true vector $\boldsymbol{\theta}_0$ (which represents $\theta(t)$). The observations are $\mathbf{x} = \mathbf{x}_T + \mathbf{w}$, where $\mathbf{w}$ is a white Gaussian noise vector with variance $\sigma^2 = 1$.

In Figure 5.3 we plot the original signal along with the observations $\mathbf{x}$, and the clean convolved signal $\mathbf{x}_T = \mathbf{H}\boldsymbol{\theta}_0$. The original signal along with the estimates using the SURE principle and the discrepancy method are plotted in Figure 5.4. To evaluate the gradient of the estimate we used (5.46). Evidently, the SURE method leads to superior performance. The MSE using the SURE approach in this example is 0.10 while the discrepancy strategy leads to an MSE of 1.16.

To summarize, we developed an unbiased estimate of the MSE in multivariate exponential families by extending the SURE method. This generalized principle can now be used in exponential multivariate estimation problems to develop estimators with improved performance over existing approaches. As an application, we demonstrated the use of this technique in choosing regularization parameters in penalized inverse problems.

6

Bounded Error Estimation

In this last section, we depart from the statistical model that was an integral part of our discussion so far. Instead, we consider a bounded-error setting in which the measurement error is bounded. Our goal is to illustrate that estimation error approaches, as well as minimax methods, are relevant in this context as well and can be used to improve the performance substantially over standard LS-based objectives. Furthermore, this approach can be applied even when a statistical model exists by replacing the statistical relationship between the unknown parameter and the observed data by an appropriate bounded-error constraint. As we show, even though the constraint will not necessarily be satisfied for all realizations, the bounded error methodology can still result in estimates that dominate constrained ML in an MSE sense.

6.1 The Chebyshev Center

Our focus here is on the linear regression model $\mathbf{x} = \mathbf{H}\boldsymbol{\theta}_0 + \mathbf{w}$, where $\mathbf{w}$ is a noise vector which is not assumed to have any particular pdf. Instead, we adopt the bounded error approach, also referred to as set-membership estimation [109, 114], in which it is assumed that the noise

is norm bounded. We further suppose that there is prior deterministic information on $\boldsymbol{\theta}_0$, in the form of constraints. For example, a popular assumption is that $\|\mathbf{L}\boldsymbol{\theta}_0\| \leq \eta$ for some $\eta > 0$, where $\mathbf{L}$ is the discretization of a first or second-order differential operator that accounts for smoothness properties of $\boldsymbol{\theta}_0$ [73, 75]. Another example are interval restrictions on the components of $\boldsymbol{\theta}_0$. More generally, we assume that $\boldsymbol{\theta}_0$ lies in the set $\mathcal{C}$ defined by the intersection of k ellipsoids:

$$\mathcal{C} = \{\boldsymbol{\theta}_0 : f_i(\boldsymbol{\theta}_0) \stackrel{\triangle}{=} \boldsymbol{\theta}_0^T \mathbf{Q}_i \boldsymbol{\theta}_0 + 2\mathbf{g}_i^T \boldsymbol{\theta}_0 + d_i \leq 0, 1 \leq i \leq k\}, \tag{6.1}$$

where $\mathbf{Q}_i \succeq \mathbf{0}, \mathbf{g}_i \in \mathbb{R}^m$, and $d_i \in \mathbb{R}$.

In these settings, the most popular estimation strategy is the constrained LS (CLS) approach, which minimizes the data error $\|\hat{\mathbf{x}} - \mathbf{x}\|^2$ between the estimated data $\hat{\mathbf{x}} = \mathbf{H}\hat{\boldsymbol{\theta}}$ and the measurement vector $\mathbf{x}$, subject to $\hat{\boldsymbol{\theta}} \in \mathcal{C}$ [17]. Thus $\hat{\boldsymbol{\theta}}_{\text{CLS}}$ is the solution to

$$\min_{\boldsymbol{\theta} \in \mathcal{C}} \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2. \tag{6.2}$$

Clearly $\hat{\boldsymbol{\theta}}_{\text{CLS}}$ is feasible, namely it satisfies the constraints defining our prior knowledge. However, the fact that it minimizes the data error over $\mathcal{C}$ does not mean that it leads to a small estimation error $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|$. In fact, the simulations in Section 6.2.3 demonstrate that the resulting error can be quite large. Furthermore, the CLS solution often lies on the boundary of the set $\mathcal{C}$. It is well known that if the noise is random, then the MSE of such an estimate can be improved by moving away from the boundary toward the center of $\mathcal{C}$ [22].

To design an estimator with improved performance, we would like to directly control the estimation error $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$. Since no statistical assumptions are made at this stage, to achieve this goal, we assume that the noise is norm-bounded $\|\mathbf{w}\|^2 \leq \rho$; note however that the estimator we develop can also be used when $\mathbf{w}$ is random by choosing ρ proportional to its variance. In fact, in Section 6.4 we show that in the iid Gaussian setting such a choice results in an estimate that dominates the corresponding CLS method. Our framework can also easily incorporate other constraints on $\mathbf{w}$ such as bounds on the magnitudes of the individual components. The key to improving the performance, is that instead of minimizing the data error, we suggest minimizing the worst-case estimation error $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2$ over all feasible solutions.

Combining the restrictions on $\boldsymbol{\theta}_0$ and $\mathbf{w}$, the feasible parameter set, which is the set of all possible values of $\boldsymbol{\theta}_0$, is given by

$$\mathcal{Q} = \{\boldsymbol{\theta}_0 : \boldsymbol{\theta}_0 \in \mathcal{C}, \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}_0\|^2 \leq \rho\}. \tag{6.3}$$

We assume throughout for simplicity that there is at least one point in the interior of $\mathcal{Q}$, and that $\mathbf{H}^T\mathbf{H}$ is invertible. Our criterion then becomes

$$\min_{\hat{\boldsymbol{\theta}}} \max_{\boldsymbol{\theta}\in\mathcal{Q}} \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2, \tag{6.4}$$

which is equivalent to finding the Chebyshev center (CC) [2, 139] of $\mathcal{Q}$ defined as the center of the minimal radius ball enclosing the set. To see this, note that (6.4) can be written equivalently as

$$\min_{\hat{\boldsymbol{\theta}},r} \{r : \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \leq r \text{ for all } \boldsymbol{\theta} \in \mathcal{Q}\}. \tag{6.5}$$

For a given r, the set of all values of $\boldsymbol{\theta}$ satisfying $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \leq r$ defines a ball with radius $\sqrt{r}$ and center $\hat{\boldsymbol{\theta}}$. Thus, the constraint in (6.5) is equivalent to the requirement that the ball defined by r and $\hat{\boldsymbol{\theta}}$ encloses the set $\mathcal{Q}$. Since the minimization is over the squared-radius r, it follows that the solution is the center of the minimum radius ball enclosing $\mathcal{Q}$. The squared radius of the ball is the optimal minimax value of (6.4). This CC is illustrated in Figure 6.1, taken from [49], as the dotted center. The filled area is the intersection of three ellipsoids and the dotted circle is the minimum inscribing circle of the intersection of the ellipsoids. Evidently, in contrast with the CLS method, the CC will lie in the center of the set.

The discussion above motivates the use of the CC as a viable alternative to CLS. Unfortunately, however, computing the center (6.4) is a hard optimization problem. To better understand the intrinsic difficulty, note that the inner maximization is a non-convex quadratic optimization problem since we need to *maximize* a convex function. Nonetheless, as we show in the ensuing sections, using semidefinite relaxation ideas we can develop a pretty good approximation of the Chebyshev solution, referred to as the relaxed CC (RCC). The RCC can often be computed efficiently and leads to good squared-error performance. Furthermore, in many cases it actually coincides with the true CC [7, 8, 38, 49].

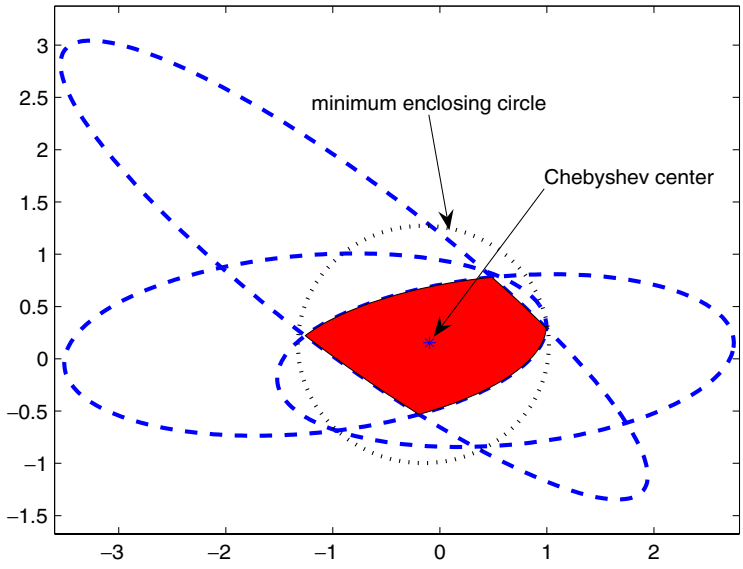


Fig. 6.1 The Chebyshev center of the intersection of three ellipsoids.

6.2 The Relaxed Chebyshev Center

To develop the RCC estimator, denoted $\hat{\boldsymbol{\theta}}_{\text{RCC}}$, consider the inner maximization in (6.4):

$$\max_{\boldsymbol{\theta}}\{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|^2 : f_i(\boldsymbol{\theta}) \leq 0, 0 \leq i \leq k\}, \tag{6.6}$$

where $f_i(\boldsymbol{\theta})$, $1 \leq i \leq k$ are defined by (6.1), and $f_0(\boldsymbol{\theta}) = \|\mathbf{x}-\mathbf{H}\boldsymbol{\theta}\|^2 - \rho$ is defined similarly with $\mathbf{Q}_0 = \mathbf{H}^T\mathbf{H}$, $\mathbf{g}_0 = -\mathbf{H}^T\mathbf{x}, d_0 = \|\mathbf{x}\|^2 - \rho$. Thus, the set $\mathcal{Q}$ can be written as

$$\mathcal{Q} = \{\boldsymbol{\theta} : f_i(\boldsymbol{\theta}) \leq 0, 0 \leq i \leq k\}. \tag{6.7}$$

Denoting $\Delta = \boldsymbol{\theta}\boldsymbol{\theta}^T$, and using the fact that $\boldsymbol{\theta}^T\mathbf{Q}\boldsymbol{\theta} = \text{Tr}(\Delta\mathbf{Q})$ for any $\mathbf{Q}$, (6.6) can be written equivalently as

$$\max_{(\Delta,\boldsymbol{\theta})\in\mathcal{G}}\{\|\hat{\boldsymbol{\theta}}\|^2 - 2\hat{\boldsymbol{\theta}}^T\boldsymbol{\theta} + \text{Tr}(\Delta)\}, \tag{6.8}$$

where

$$\mathcal{G} = \{(\Delta,\boldsymbol{\theta}) : f_i(\Delta,\boldsymbol{\theta}) \leq 0, 0 \leq i \leq k, \Delta = \boldsymbol{\theta}\boldsymbol{\theta}^T\}, \tag{6.9}$$

and we defined

$$f_i(\Delta,\boldsymbol{\theta}) = \mathrm{Tr}(\mathbf{Q}_i\Delta) + 2\mathbf{g}_i^T\boldsymbol{\theta} + d_i, \quad 0 \le i \le k. \tag{6.10}$$

The objective in (6.8) is concave (linear) in $(\Delta,\boldsymbol{\theta})$, but the set $\mathcal{G}$ is not convex, so that the problem (6.8) is not convex. To obtain a convex relaxation of (6.8) we may replace $\mathcal{G}$ by the convex set

$$\mathcal{T} = \{(\Delta,\boldsymbol{\theta}) : f_i(\Delta,\boldsymbol{\theta}) \le 0, 0 \le i \le k, \Delta \succeq \boldsymbol{\theta}\boldsymbol{\theta}^T\}. \tag{6.11}$$

Here we have changed the nonconvex constraint $\Delta = \boldsymbol{\theta}\boldsymbol{\theta}^T$ to a convex restriction $\Delta \succeq \boldsymbol{\theta}\boldsymbol{\theta}^T$. Indeed, using Schur's lemma (see Lemma A.3 in the Appendix) the latter constraint can be written as an LMI. The RCC is the solution to the resulting minimax problem:

$$\min_{\hat{\boldsymbol{\theta}}} \max_{(\Delta,\boldsymbol{\theta})\in\mathcal{T}} \{\|\hat{\boldsymbol{\theta}}\|^2 - 2\hat{\boldsymbol{\theta}}^T\boldsymbol{\theta} + \mathrm{Tr}(\Delta)\}. \tag{6.12}$$

The objective in (6.12) is concave (linear) in Δ and $\boldsymbol{\theta}$ and convex in $\hat{\boldsymbol{\theta}}$. Furthermore, the set $\mathcal{T}$ is convex and bounded. Therefore, we can replace the order of the minimization and maximization (see Appendix), resulting in the equivalent problem

$$\max_{(\Delta,\boldsymbol{\theta})\in\mathcal{T}} \min_{\hat{\boldsymbol{\theta}}} \{\|\hat{\boldsymbol{\theta}}\|^2 - 2\hat{\boldsymbol{\theta}}^T\boldsymbol{\theta} + \mathrm{Tr}(\Delta)\}. \tag{6.13}$$

The inner minimization is a simple quadratic problem, whose optimal value is $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$. Thus, (6.13) reduces to

$$\max_{(\Delta,\boldsymbol{\theta})\in\mathcal{T}} \{-\|\boldsymbol{\theta}\|^2 + \mathrm{Tr}(\Delta)\}, \tag{6.14}$$

which is a convex optimization problem with a concave objective and LMI constraints. The RCC estimate is the $\boldsymbol{\theta}$-part of the solution to (6.14).

The RCC is not generally equal to the CC of $\mathcal{Q}$. An exception is when $k = 1$ with the problem defined over the complex domain [8]. Since clearly $\mathcal{G} \subseteq \mathcal{T}$, we have that

$$\begin{aligned}\min_{\hat{\boldsymbol{\theta}}} \max_{\boldsymbol{\theta}\in\mathcal{Q}} \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 &= \min_{\hat{\boldsymbol{\theta}}} \max_{(\Delta,\boldsymbol{\theta})\in\mathcal{G}\{\|\hat{\boldsymbol{\theta}}\|^2 - 2\hat{\boldsymbol{\theta}}^T\boldsymbol{\theta} + \mathrm{Tr}(\Delta)\}} \\ &\le \min_{\hat{\boldsymbol{\theta}}} \max_{(\Delta,\boldsymbol{\theta})\in\mathcal{T}} \{\|\hat{\boldsymbol{\theta}}\|^2 - 2\hat{\boldsymbol{\theta}}^T\boldsymbol{\theta} + \mathrm{Tr}(\Delta)\}. \end{aligned} \tag{6.15}$$

Therefore the RCC provides an upper bound on the minimax value.

In Theorem 6.1 below we present an explicit representation of the RCC, which is obtained by computing the dual of (6.14) [49]. Before we present this result, we note that if we denote $\mathbf{Z} = \Delta - \boldsymbol{\theta}\boldsymbol{\theta}^T$ in (6.14), then the RCC can be written more explicitly as the solution to

$$\begin{array}{ll} \max\limits_{\mathbf{Z},\boldsymbol{\theta}} & \mathrm{Tr}(\mathbf{Z}) \\ \text{s.t.} & \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2 - \rho + \mathrm{Tr}(\mathbf{H}^T\mathbf{H}\mathbf{Z}) \leq 0; \\ & f_i(\boldsymbol{\theta}) + \mathrm{Tr}(\mathbf{Q}_i\mathbf{Z}) \leq 0, \quad 1 \leq i \leq k; \\ & \mathbf{Z} \succeq 0. \end{array} \tag{6.16}$$

Thus, the RCC satisfies the constraints $f_i(\boldsymbol{\theta}) \leq 0$ with margins given by $\mathrm{Tr}(\mathbf{H}^T\mathbf{H})$ and $\mathrm{Tr}(\mathbf{Q}_i\mathbf{Z})$, which are maximized in some sense (these constraints include both the prior restrictions and the measurement-error bound). This ensures that the RCC approaches the center of the constraint set.

Theorem 6.1. The RCC which is the solution to (6.14), is given by

$$\hat{\boldsymbol{\theta}}_{\mathrm{RCC}} = -\left(\sum_{i=0}^{k} \alpha_i \mathbf{Q}_i\right)^{-1} \left(\sum_{i=0}^{k} \alpha_i \mathbf{g}_i\right), \tag{6.17}$$

where $(\alpha_0, \ldots, \alpha_k)$ is an optimal solution of the following convex optimization problem in $k+1$ variables:

$$\begin{array}{ll} \min\limits_{\alpha_i} & \left\{ \left(\sum\limits_{i=0}^{k} \alpha_i \mathbf{g}_i\right)^T \left(\sum\limits_{i=0}^{k} \alpha_i \mathbf{Q}_i\right)^{-1} \left(\sum\limits_{i=0}^{k} \alpha_i \mathbf{g}_i\right) - \sum\limits_{i=0}^{k} d_i \alpha_i \right\} \\ \text{s.t.} & \sum\limits_{i=0}^{k} \alpha_i \mathbf{Q}_i \succeq \mathbf{I}, \\ & \alpha_i \geq 0, \quad 0 \leq i \leq k. \end{array} \tag{6.18}$$

Note that (6.18) can be cast as an SDP

$$\begin{aligned} \min_{\alpha_i} \quad & \left\{ t - \sum_{i=0}^{k} d_i \alpha_i \right\} \\ \text{s.t.} \quad & \begin{pmatrix} \sum_{i=0}^{k} \alpha_i \mathbf{Q}_i & \sum_{i=0}^{k} \alpha_i \mathbf{g}_i \\ \sum_{i=0}^{k} \alpha_i \mathbf{g}_i^T & t \end{pmatrix} \succeq \mathbf{0}, \\ & \sum_{i=0}^{k} \alpha_i \mathbf{Q}_i \succeq \mathbf{I}, \\ & \alpha_i \geq 0, \quad 0 \leq i \leq k. \end{aligned} \tag{6.19}$$

Therefore, the RCC can be determined efficiently using standard SDP solvers such as SeDuMi [133], SDPT3 [138] or CVX [68].

An important feature of the RCC estimate is that it is unique, and feasible, meaning it resides in the set $\mathcal{Q}$ [49].

6.2.1 Modeling of Linear Constraints

There are many signal processing examples in which there are interval constraints on the elements of $\boldsymbol{\theta}_0$. In this section we address the question of how to best represent such restrictions.

Specifically, suppose that one of the constraints defining the set $\mathcal{Q}$ is a double-sided linear inequality of the form:

$$\ell \leq \mathbf{a}^T \boldsymbol{\theta}_0 \leq u, \tag{6.20}$$

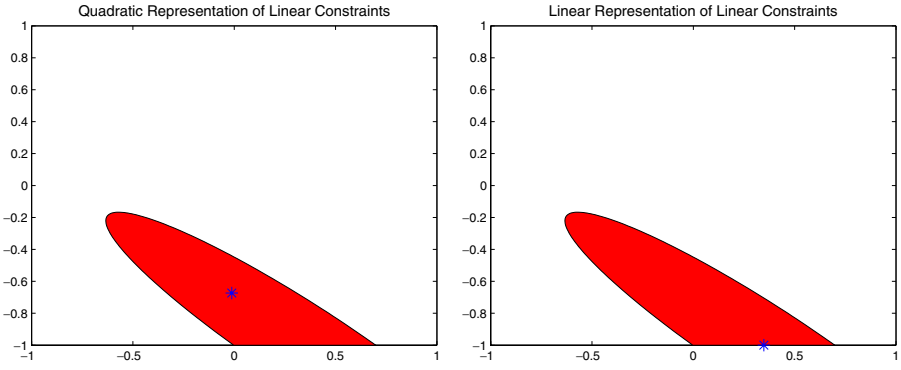
where $\ell < u$ and $\mathbf{a} \in \mathbb{R}^m$ is a nonzero vector. The constraint (6.20) can also be written in quadratic form as

$$(\mathbf{a}^T \boldsymbol{\theta}_0 - \ell)(\mathbf{a}^T \boldsymbol{\theta}_0 - u) \leq 0. \tag{6.21}$$

An important question that arises is whether or not the RCC depends on the specific representation of the set $\mathcal{Q}$. Clearly the CLS and CC estimates are independent of the representation of $\mathcal{Q}$, as they depend only the set $\mathcal{Q}$ itself. However, the RCC estimate is more involved

as it is a result of a relaxation of $\mathcal{Q}$, so that different characterizations may lead to different relaxed sets. In [49] it is shown that indeed the RCC depends on the specific form of $\mathcal{Q}$ chosen. Furthermore, the quadratic representation (6.21) is better than the linear characterization (6.20) in the sense that the resulting minimax value is smaller. Consequently, in the presence of several double sided linear constraints, it is best to represent all of them as quadratic restrictions.

In practice, the linear representation often provides a much looser bound on the squared radius of the minimum enclosing circle. A typical example, taken from [49], can be seen in Figure 6.2. The filled region describes the intersection of a randomly generated ellipsoid $\mathcal{E}$ with the box $[-1,1] \times [-1,1]$. The asterisk in Figure 6.2(a) is the RCC when the box constraints are modeled as $x_i^2 \leq 1, i = 1,2$, while the asterisk in Figure 6.2(b) is the RCC using the representation $-1 \leq x_i \leq 1, i = 1,2$. Clearly, the RCC using the linear representation is far from the center of the filled region (actually, it is on the boundary of the area!). In contrast, the RCC corresponding to the quadratic representation seems like a good measure of the center of the set. The minimax value in the linear choice was approximately 37% higher than that resulting from the quadratic representation.



(a) (b)

Fig. 6.2 The RCC of the intersection of an ellipsoid with the box $[-1,1] \times [-1,1]$ using a quadratic representation (a) and a linear representation (b).

6.2.2 Relation To The CLS

The RCC estimate was based on the prior information $\boldsymbol{\theta}_0 \in \mathcal{C}$ with $\mathcal{C}$ given by (6.1), and the bounded error constraint $\|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}_0\|^2 \leq \rho$. The CLS estimate, which is the solution to (6.2), tries to minimize the noise (or the data error) over $\boldsymbol{\theta}_0 \in \mathcal{C}$. As we show in this section, there are some interesting relationships between the two approaches.

To obtain a more explicit expression for the CLS estimate we can compute the dual of (6.2), from which we conclude that

$$\hat{\boldsymbol{\theta}}_{\text{CLS}} = -\left(\mathbf{Q}_0 + \sum_{i=1}^{k} \alpha_i \mathbf{Q}_i\right)^{-1} \left(\mathbf{g}_0 + \sum_{i=1}^{k} \alpha_i \mathbf{g}_i\right), \tag{6.22}$$

where $(\alpha_1, \ldots, \alpha_k)$ is an optimal solution of the following convex optimization problem in k variables:

$$\begin{aligned} \min_{\alpha_i} \quad & \left\{ \left(\mathbf{g}_0 + \sum_{i=1}^{k} \alpha_i \mathbf{g}_i\right)^T \left(\mathbf{Q}_0 + \sum_{i=1}^{k} \alpha_i \mathbf{Q}_i\right)^{-1} \left(\mathbf{g}_0 + \sum_{i=1}^{k} \alpha_i \mathbf{g}_i\right) \right. \\ & \left. -d_0 - \sum_{i=1}^{k} d_i \alpha_i \right\} \\ \text{s.t.} \quad & \alpha_i \geq 0, \quad 1 \leq i \leq k. \end{aligned}$$

Comparing with Theorem 6.1 we see that the RCC and CLS estimators have very similar structures. However, in the CLS, $\alpha_0 = 1$ whereas in the RCC it is a solution to an optimization problem. Furthermore, the RCC has an additional LMI constraint.

Another interesting observation is that the CLS can also be obtained as a relaxation of the CC [49]. However, this relaxation is looser than that of the RCC meaning that the resulting bound on the minimax value is larger. To see this, note that the RCC was obtained by replacing $\boldsymbol{\theta} \in \mathcal{Q}$ (corresponding to $\boldsymbol{\theta} \in \mathcal{C}$ and $\|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2 \leq \rho$) by the equivalent set $(\boldsymbol{\theta}, \Delta) \in \mathcal{G}$ with $\mathcal{G}$ given by (6.9), and then relaxing the non-convex constraint in $\mathcal{G}$ to obtain the convex set $\mathcal{T}$ of (6.11). The CLS can be viewed as a relaxed CC where $\mathcal{G}$ is replaced by a different convex set $\mathcal{V}$, that is larger than $\mathcal{T}$. To obtain $\mathcal{V}$, note that $\mathcal{G}$ can be written as

$$\begin{aligned} \mathcal{G} = \{(\Delta, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{C}, \\ \text{Tr}(\mathbf{H}^T\mathbf{H}\Delta) - 2\mathbf{x}^T\mathbf{H}^T\boldsymbol{\theta} + \|\mathbf{x}\|^2 - \rho \leq 0, \Delta = \boldsymbol{\theta}\boldsymbol{\theta}^T\}. \end{aligned} \tag{6.23}$$

In this representation we substituted $\Delta = \boldsymbol{\theta}\boldsymbol{\theta}^T$ only in the noise constraint, but not in the restrictions $\boldsymbol{\theta} \in \mathcal{C}$. Relaxing the non-convex equality leads to the relaxed convex set

$$\begin{aligned}\mathcal{V} = \{(\Delta, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{C}, \\ \mathrm{Tr}(\mathbf{H}^T\mathbf{H}\Delta) - 2\mathbf{x}^T\mathbf{H}^T\boldsymbol{\theta} + \|\mathbf{x}\|^2 - \rho \leq 0, \Delta \succeq \boldsymbol{\theta}\boldsymbol{\theta}^T\}.\end{aligned} \quad (6.24)$$

Evidently, the relaxation affected only the noise constraint and not those in $\mathcal{C}$. This results in a set $\mathcal{V}$ that includes the set $\mathcal{T}$ of (6.11). The following theorem establishes that $\hat{\boldsymbol{\theta}}_{\text{CLS}}$ is the solution to the resulting minimax problem.

Theorem 6.2. The CLS estimate of (6.2) is the same as the relaxed Chebyshev center

$$\min_{\hat{\boldsymbol{\theta}}} \max_{(\Delta,\boldsymbol{\theta})\in\mathcal{V}} \{\|\hat{\boldsymbol{\theta}}\|^2 - 2\hat{\boldsymbol{\theta}}^T\boldsymbol{\theta} + \mathrm{Tr}(\Delta)\}. \quad (6.25)$$

As in (6.16), we can substitute $\mathbf{Z} = \Delta - \boldsymbol{\theta}\boldsymbol{\theta}^T$ into (6.25) and obtain a more explicit representation:

$$\begin{array}{ll} \max\limits_{\mathbf{Z},\boldsymbol{\theta}} & \mathrm{Tr}(\mathbf{Z}) \\ \text{s.t.} & \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2 - \rho + \mathrm{Tr}(\mathbf{H}^T\mathbf{H}\mathbf{Z}) \leq 0; \\ & f_i(\boldsymbol{\theta}) \leq 0, \quad 1 \leq i \leq k; \\ & \mathbf{Z} \succeq 0. \end{array} \quad (6.26)$$

Here we see that the CLS solution satisfies the measurement error constraint with a margin dictated by $\mathrm{Tr}(\mathbf{H}^T\mathbf{H}\mathbf{Z})$. However, no margin is enforced on the other prior requirements. This leads to the fact that the CLS often lies on the boundary of the prior constraint set, in contrast to the RCC solution (6.16) in which a margin is enforced on the prior constraints as well.

6.2.3 Example

We now demonstrate the performance of the RCC approach via an example.

We consider a discretization of the heat integral equation implemented in the function `heat(90,1)` from the "Regularization Tools"

Matlab package [74]. In this case, $\mathbf{H}\boldsymbol{\theta}_0 = \mathbf{g}$, where $\mathbf{H} \in \mathbb{R}^{90\times 90}$ and $\boldsymbol{\theta}_0, \mathbf{g} \in \mathbb{R}^{90}$. The true vector $\boldsymbol{\theta}_0$ is shown in Figure 6.3 (True Signal) and resides in the set

$$\mathcal{C} = \{\boldsymbol{\theta} \in \mathbb{R}^{90} : \boldsymbol{\theta} \geq \mathbf{0}, \boldsymbol{\theta}^T\mathbf{1} \leq \eta\}, \tag{6.27}$$

where $\mathbf{1}$ is the vector of all ones. The observed vector is given by $\mathbf{x} = \mathbf{g} + \mathbf{w}$, where the elements of $\mathbf{w}$ are zero-mean, independent Gaussian random variables with standard deviation 0.001. Both $\mathbf{g}$ and $\mathbf{x}$ are shown in Figure 6.3 (Observation).

To compute the RCC, we chose the set $\mathcal{Q}$ as

$$\mathcal{Q} = \{\|\mathbf{H}\boldsymbol{\theta} - \mathbf{x}\|^2 \leq \rho, \boldsymbol{\theta} \in \mathcal{C}\}$$

with $\rho = \alpha\|\mathbf{w}\|^2$ for some constant α, and $\eta = \alpha(\sum_{i=1}^{90} \theta_i)$. We then used the following quadratic representation of $\mathcal{C}$:

$$\{\boldsymbol{\theta} \in \mathbb{R}^{90} : \theta_i(\theta_i - \eta) \leq 0, (\boldsymbol{\theta}^T\mathbf{1})^2 \leq \eta^2, i = 1, \ldots, 90\}.$$

For comparison, we computed the CLS estimate which is the solution to $\min\{\|\mathbf{H}\boldsymbol{\theta} - \mathbf{x}\|^2 : \boldsymbol{\theta} \in \mathcal{C}\}$.

The results of the RCC and CLS estimates for $\alpha = 2$ and $\alpha = 10$ are shown at the bottom of Figure 6.3. Evidently, the RCC approach leads to the best performance. The squared error of the CLS estimate was 196 times larger than that of the RCC solution for $\alpha = 2$, and 55 times larger when $\alpha = 10$. As expected, the performance of both methods is better when $\alpha = 2$. However, it is interesting to note that even when $\alpha = 10$, so that extremely loose prior information is used, the RCC results in very good behavior.

6.3 Special Cases

In this section we treat some special cases in which a more explicit expression for the RCC can be obtained.

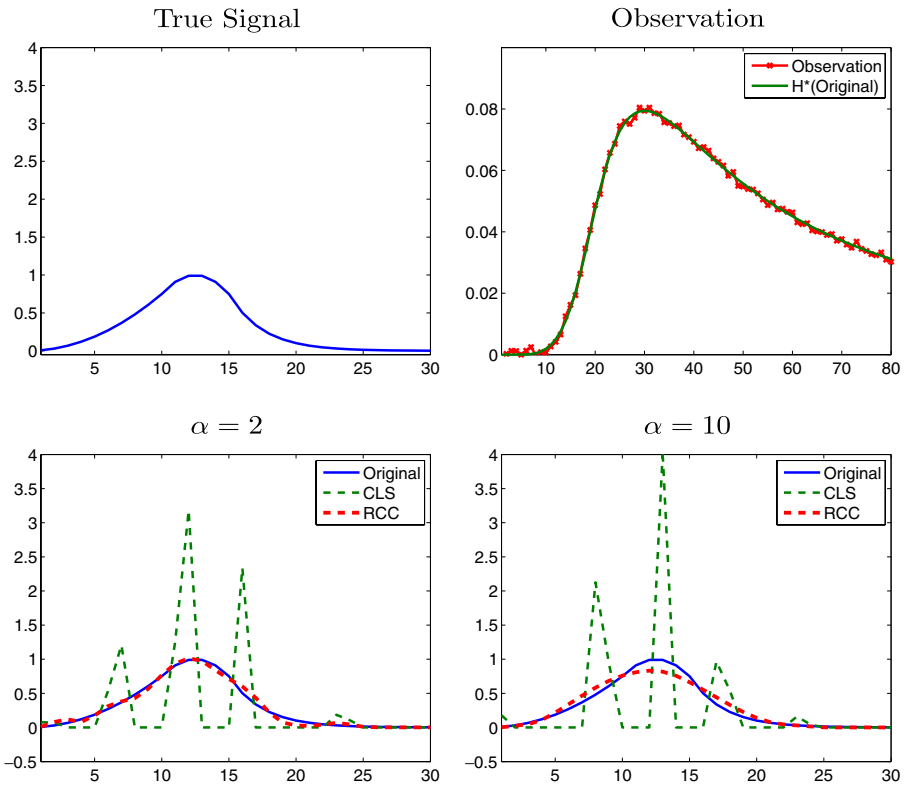


Fig. 6.3 Comparison between the RCC and CLS estimates.

6.3.1 IID Setting

Consider estimating $\boldsymbol{\theta}_0$ from $\mathbf{x} = \boldsymbol{\theta}_0 + \mathbf{w}$, where $\|\mathbf{w}\|^2 \leq \rho$, and $\|\boldsymbol{\theta}_0\|^2 \leq \eta$. A simple calculation shows that the CLS estimate is given by

$$\hat{\boldsymbol{\theta}}_{\text{CLS}} = \begin{cases} \mathbf{x}, & \|\mathbf{x}\|^2 \leq \eta; \\ \sqrt{\dfrac{\eta}{\mathbf{x}^T\mathbf{x}}}\mathbf{x}, & \|\mathbf{x}\|^2 \geq \eta. \end{cases} \tag{6.28}$$

In this special case the CC can be computed exactly. When $m = 1$ the set $\mathcal{Q}$ is just an interval in $\mathbb{R}$, and therefore the CC is its mid-point. When $m \geq 2$, it can be shown that the RCC is the exact CC [5, 38].

From (6.16), the Chebyshev estimate is therefore the solution to

$$\begin{array}{ll} \max\limits_{\boldsymbol{\theta},t\geq 0} & t \\ \text{s.t.} & \|\mathbf{x}-\boldsymbol{\theta}\|^2+t\leq\rho \\ & \|\boldsymbol{\theta}\|^2+t\leq\eta, \end{array} \tag{6.29}$$

where we denoted $t=\mathrm{Tr}(\mathbf{Z})$. Therefore, the CC tries to move the solution toward the center of the constraints: each constraint is satisfied with a margin of t and the aim is to maximize this gap. Clearly, this will move the solution away from the boundary of the set (unless the optimal t is $t=0$).

Interestingly, from (6.26) the CLS solution can be written in a similar form:

$$\begin{array}{ll} \max\limits_{\boldsymbol{\theta},t\geq 0} & t \\ \text{s.t.} & \|\mathbf{x}-\boldsymbol{\theta}\|^2+t\leq\rho \\ & \|\boldsymbol{\theta}\|^2\leq\eta. \end{array} \tag{6.30}$$

Comparing (6.29) and (6.30) highlights the fundamental difference between the two strategies: In the CLS the margin is only on the data error constraint while in the CC formulation both restrictions are satisfied with a gap.

Problem (6.29) is a simple convex optimization problem and therefore can be solved using duality theory which leads to the solution

$$\hat{\boldsymbol{\theta}}_{\mathrm{CC}}=\frac{1}{2}\left[1-\frac{\gamma}{\|\mathbf{x}\|^2}\right]_{[0,2]}\mathbf{x}. \tag{6.31}$$

Here $\gamma=\rho-\eta$, and we used the notation

$$x_{[a,b]}=\begin{cases} x, & a\leq x\leq b; \\ a, & x\leq a; \\ b, & x\geq b. \end{cases} \tag{6.32}$$

It is interesting to note that $\hat{\boldsymbol{\theta}}_{\mathrm{CC}}$ has a similar form to the James–Stein estimate (4.19). An important difference is the factor of $1/2$ that appears in (6.31). This factor can be explained in an empirical Bayes setting [36, 38]. Specifically, suppose that $\boldsymbol{\theta}_0$ is a Gaussian vector consisting of iid elements with variance τ, and that $\mathbf{w}=\mathbf{x}-\boldsymbol{\theta}_0$ is comprised of iid Gaussian variables with variance σ. If τ and σ are known,

then the minimum MSE estimate of $\boldsymbol{\theta}_0$ from $\mathbf{x}$ is

$$\hat{\boldsymbol{\theta}} = \frac{\tau}{\sigma + \tau}\mathbf{x}. \tag{6.33}$$

Empirical Bayes methods are based on using (6.33) in conjunction with estimates for τ and σ.

Since from (6.31) the Chebyshev estimate depends only on the difference $\rho - \eta$, we assume that $\sigma - \tau$ is given. Using the fact that for large m, $\|\boldsymbol{\theta}_0\|^2 \to m\sigma$ and $\|\mathbf{w}\|^2 \to m\tau$, we choose

$$\sigma - \tau = \frac{\rho - \eta}{m}. \tag{6.34}$$

Expressing the minimum MSE estimate of (6.33) as:

$$\hat{\boldsymbol{\theta}} = \frac{\tau}{\sigma + \tau}\mathbf{x} = \frac{1}{2}\left(1 - \frac{\sigma - \tau}{\sigma + \tau}\right)\mathbf{x}, \tag{6.35}$$

and using (6.34) together with the fact that in the limit of large m, $\mathbf{x}^T\mathbf{x} \to m(\sigma + \tau)$, results in the unrestricted Chebyshev estimate (i.e., without limiting the shrinkage to the interval $[0, 2]$).

The iid setting considered here reveals some of the essential properties of the Chebyshev approach: the estimate is in the center of the set rather than the boundary, the measurement constraints and prior constraints are treated equally, namely both are required to be satisfied with a gap and not only the measurement error restriction, and finally the CC can be interpreted in a Bayesian setting where both the error and the unknown vector are random, with unknown variances. We only assume that the variance difference is given. In Section 6.4 we show that when $\mathbf{w}$ is an iid Gaussian vector and ρ is chosen appropriately, the CC dominates the CLS solution.

6.3.2 Bounded-Norm Prior

The second special case we focus on is when the prior knowledge on $\boldsymbol{\theta}_0$ is a single bounded-norm constraint of the form $\|\mathbf{L}\boldsymbol{\theta}_0\|^2 \leq \eta$. In this case it is shown in [8] that the RCC is the exact CC when $\boldsymbol{\theta}_0$ is defined over the complex domain. Over the reals, a sufficient condition is developed to guarantee that the RCC is the exact CC. Furthermore, an efficient

algorithm based on the ellipsoid method [12] is provided to compute the RCC.

When $\mathbf{L} = \mathbf{I}$, the task of calculating the RCC estimator reduces to a single-variable convex minimization problem [8].

Proposition 6.3. Let $\mathbf{L} = \mathbf{I}$ and denote $\delta = \lambda_{\min}(\mathbf{H}^T\mathbf{H})$. Then the RCC estimator is given by

$$\hat{\boldsymbol{\theta}}_{\text{RCC}} = \begin{cases} (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})^{-1}\mathbf{H}^T\mathbf{x}, & 0 \le \lambda < \infty \\ \mathbf{0}, & \lambda = \infty, \end{cases}$$

where λ is determined as follows[1]:

(i) if $\delta > 0$, then $\lambda = 1/\mu - \delta$, where μ is the solution of the convex minimization problem

$$\min_{0 \le \mu \le 1/\delta} \big\{ (1 - \delta\mu)\eta + \mu(\rho - \|\mathbf{x}\|^2) + \mu^2 \mathbf{x}^T\mathbf{H}(\mu(\mathbf{H}^T\mathbf{H} - \delta\mathbf{I}) + \mathbf{I})^{-1}\mathbf{H}^T\mathbf{x} \big\}; \tag{6.36}$$

(ii) if $\delta = 0$, then $\lambda = 1/\xi$, where ξ is the solution of the convex minimization problem

$$\min_{\xi \ge 0} \big\{ \xi(\rho - \|\mathbf{x}\|^2) + \xi^2 \mathbf{x}^T\mathbf{H}(\xi\mathbf{H}^T\mathbf{H} + \mathbf{I})^{-1}\mathbf{H}^T\mathbf{x} \big\}. \tag{6.37}$$

6.4 Statistical Analysis

6.4.1 Domination in the IID Gaussian Setting

Until now we did not assume a specific statistical model on the noise $\mathbf{w}$. Interestingly, it can be shown that for the iid Gaussian model $\mathbf{x} = \boldsymbol{\theta}_0 + \mathbf{w}$, where $\mathbf{w}$ is a vector consisting of zero-mean iid Gaussian random variables, the value of ρ can be chosen such that the resulting CC dominates the CLS solution in terms of MSE over all $\|\boldsymbol{\theta}_0\|^2 \le \eta$. Thus, despite the fact that the CC is derived based on deterministic

[1] We use the standard terminology $a/0 = \infty$ whenever $a > 0$.

considerations, it can be used as a viable alternative to constrained ML techniques when a statistical model is given.

Domination of the CC estimate (6.31) over the CLS solution (6.28) is discussed in detail in [23]. Clearly the performance of the CC will depend on the choice of ρ. When the noise is Gaussian, $\mathbf{w}$ is not norm bounded, and therefore in order to satisfy the constraint $\|\mathbf{x} - \boldsymbol{\theta}_0\|^2 \leq \rho$ for all possible $\mathbf{x}$ we would need to choose $\rho = \infty$ which yields the trivial estimate $\hat{\boldsymbol{\theta}} = \mathbf{0}$. In practice, we do not enforce the constraint for all $\mathbf{x}$, but rather choose ρ large enough so that for "reasonable" noise vectors it is satisfied. An intuitive choice is to select ρ equal to the expected value of the constraint:

$$\rho = E\left\{\|\mathbf{x} - \boldsymbol{\theta}_0\|^2\right\} = E\left\{\|\mathbf{w}\|^2\right\} = n\sigma^2, \tag{6.38}$$

where n is the length of $\boldsymbol{\theta}_0$. Interestingly, this choice is sufficient to guarantee domination of the resulting CC over the CLS solution.

The proof of domination is based on the fact that under certain technical conditions, for general shrinkage estimates, i.e., estimates of the form $\hat{\boldsymbol{\theta}} = \mu(\|\mathbf{x}\|^2)\mathbf{x}$ for some function μ, domination over the sphere $\|\boldsymbol{\theta}_0\|^2 \leq \eta$ follows from domination over the boundary $\|\boldsymbol{\theta}_0\|^2 = \eta$. Since the MSE of $\hat{\boldsymbol{\theta}}$ in our case depends on $\boldsymbol{\theta}_0$ only through its norm, this implies that to guarantee domination over the CLS solution it is sufficient to select ρ such that domination holds for some value $\boldsymbol{\theta}_0$ with $\|\boldsymbol{\theta}_0\|^2 = \eta$. As shown in [23], there are many possible choices of ρ that will ensure domination. One such possibility is the value given by (6.38).

It would be interesting to extend the statistical analysis beyond the iid setting, to the general linear model $\mathbf{x} = \mathbf{H}\boldsymbol{\theta}_0 + \mathbf{w}$, and to include more general restrictions on the parameter vector $\boldsymbol{\theta}_0$. Whether or not domination over CLS continues to hold in this more general setting has not yet been established. However, the example in Section 6.2.3 demonstrates that the RCC can substantially improve the MSE performance in Gaussian noise even in more general linear models.

6.4.2 Extension to General Statistical Models

In the previous section we showed that even though the CC is based on deterministic considerations, in the Gaussian setting it can be used

to derive dominating methods in terms of MSE. We now suggest some ideas on how to extend the CC approach to more general statistical models.

Suppose we are given data $\mathbf{x}$ that is related to the unknown parameter $\boldsymbol{\theta}_0$ through a pdf $p(\mathbf{x};\boldsymbol{\theta}_0)$. This is the model we considered throughout the survey. In addition, $\boldsymbol{\theta}_0$ is known to lie in the set $\mathcal{C}$ defined by (6.1). The constrained ML approach in this setting is to seek the estimate $\hat{\boldsymbol{\theta}}$ that maximizes the likelihood over $\mathcal{C}$:

$$\max_{\boldsymbol{\theta}\in\mathcal{C}} p(\mathbf{x};\boldsymbol{\theta}). \tag{6.39}$$

In order to improve its MSE performance, we would like to use the CC methodology. The question is how to translate the relationship between $\mathbf{x}$ and $\boldsymbol{\theta}$ into a constraint that can be included in the set $\mathcal{C}$. One possibility is to bound the likelihood so as to ensure that $\boldsymbol{\theta}$ explains the data to some extent. Thus we may construct the set $\mathcal{Q}$ of (6.7) by adding the restriction $p(\mathbf{x};\boldsymbol{\theta}) \geq \alpha$ for some suitable choice of α. Alternatively, we can rely on the fact that under suitable regularity conditions the ML solution to the unconstrained problem $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \max p(\mathbf{x};\boldsymbol{\theta})$ is asymptotically unbiased and distributed as a Gaussian random vector with covariance given by the inverse Fisher information matrix. Therefore, the ellipsoid

$$(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta})^T \mathbf{J}^{-1}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}) \leq \rho, \tag{6.40}$$

serves as a confidence interval for the true unknown value $\boldsymbol{\theta}$. Thus, we can use this constraint together with the set $\mathcal{C}$ in order to determine the CC estimate. Note, that in general the Fisher information matrix $\mathbf{J}$ depends on $\boldsymbol{\theta}$. If the dependency results in a non-convex constraint, we can replace $\boldsymbol{\theta}$ by the ML estimate and define the quadratic restriction

$$(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta})^T \mathbf{J}^{-1}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}) \leq \rho. \tag{6.41}$$

To conclude this section, we focused here on a new estimation strategy that attempts to minimize the estimation error when there are constraints on the true parameter value, in contrast to previous regularization strategies which invoke a data-error based criterion. We showed how to convert the resulting minimax problem into a convex formulation, which is a pretty good approximation of the original problem.

Several further extensions of these ideas as well as numerical issues are treated in [8]. We also discussed how this methodology can be extended to a statistical setting, and showed that in the iid Gaussian case this method dominates the standard CLS solution. Thus, despite the fact that the CC approach is deterministic in nature, it can yield improved MSE performance over constrained ML techniques when a statistical model exists.

Acknowledgments

The author is indebted to Dr. Amir Beck, Zvika Ben-Haim, Prof. Aharon Ben-Tal, Slava Chernoi, Prof. Steven Kay, Prof. Arkadi Nemirovski, and Prof. Marc Teboulle for collaborating on parts of this survey. She also gratefully acknowledges her graduate students Zvika Ben-Haim, Raja Giryes, Tomer Michaeli, and Moshe Mishali, for going over a draft of this survey and providing helpful comments.

Notations and Acronyms

We summarize here the notation and acronyms used throughout the survey.

We denote vectors in $\mathbb{R}^m$ by boldface lowercase letters, e.g., $\mathbf{x}$, and matrices in $\mathbb{R}^{n\times m}$ by boldface uppercase letters, e.g., $\mathbf{A}$. The identity matrix of appropriate dimension is written as $\mathbf{I}$, $\mathrm{diag}\,(\delta_1,\ldots,\delta_m)$ is an $m\times m$ diagonal matrix with diagonal elements δ_i, $(\cdot)^T$ is the transpose of the corresponding matrix, and $\hat{(\cdot)}$ is an estimated vector or matrix. The ith component of a vector $\boldsymbol{\theta}$ is denoted by θ_i. The true value of an unknown vector parameter $\boldsymbol{\theta}$ is written as $\boldsymbol{\theta}_0$, and the true value of an unknown scalar parameter θ is denoted by θ_0. The gradient of the function $f(\boldsymbol{\theta})$ evaluated at the point $\boldsymbol{\theta}$ is written as $df(\boldsymbol{\theta})/d\boldsymbol{\theta}$, and is a row vector with jth element equal to $df(\boldsymbol{\theta})/d\theta_j$. The gradient of a vector $d\mathbf{b}(\boldsymbol{\theta})/d\boldsymbol{\theta}$ is a matrix, with ijth element equal to $db_i(\boldsymbol{\theta})/d\theta_j$, i.e., the derivative of the ith component of the vector $\mathbf{b}(\boldsymbol{\theta})$ with respect to θ_j. For a square matrix $\mathbf{A}$, $\mathrm{Tr}(\mathbf{A})$ is the trace of $\mathbf{A}$, $\mathbf{A}\succ 0$ ($\mathbf{A}\succeq 0$) means that $\mathbf{A}$ is symmetric and positive (nonnegative) definite, and $\mathbf{A}\succeq\mathbf{B}$ means that $\mathbf{A}-\mathbf{B}\succeq 0$. The largest and smallest eigenvalues of $\mathbf{A}$ are denoted by $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$. The standard Euclidean norm is denoted $\|\mathbf{x}\|^2=\mathbf{x}^T\mathbf{x}$ and $\|\mathbf{x}\|^2_{\mathbf{Q}}=\mathbf{x}^T\mathbf{Q}\mathbf{x}$ is the weighted

norm with weighting $\mathbf{Q}$. The range space of a matrix $\mathbf{A}$ is written as $\mathcal{R}(\mathbf{A})$.

Following is a list of the most frequently used acronyms:

- BME — blind minimax estimator
- CC — Chebyshev center
- CRB — Cramér–Rao bound
- EBME — ellipsoidal BME
- iid — independent, identically-distributed
- LMI — linear matrix inequality
- LS — least squares
- ML — maximum likelihood
- MSE — mean-squared error
- MVU — minimum variance unbiased
- RCC — relaxed Chebyshev center
- SBME — spherical BME
- SURE — Stein's unbiased risk estimate
- UCRB — uniform CRB.

A

Convex Optimization Methods

The mathematical machinery behind the estimation ideas presented in this survey is that of convex optimization. In this appendix, we briefly review the basics of this theory with an emphasize on the tools needed in our presentation. Most of the material is taken from [12, 16, 20]. The presentation largely follows [143].

A.1 Convex Sets, Functions, and Problems

We begin with the formal definitions of convex sets and convex functions (see Figure A.1):

Definition A.1. A set C is convex if for any $\mathbf{x}_1 \in C$, $\mathbf{x}_2 \in C$ and $0 < \lambda < 1$, we have $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in C$.

A function $f(\mathbf{x})$ is (strictly) convex in $\mathbf{x}$ if for every $\mathbf{x}_1$ and $\mathbf{x}_2$ in its domain and $0 < \lambda < 1$, we have $f(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2)(<) \leq \lambda f(\mathbf{x}_1) + (1-\lambda)f(\mathbf{x}_2)$.

These two definitions are the building blocks for convex optimization methods which are aimed at minimizing convex functions over convex sets.

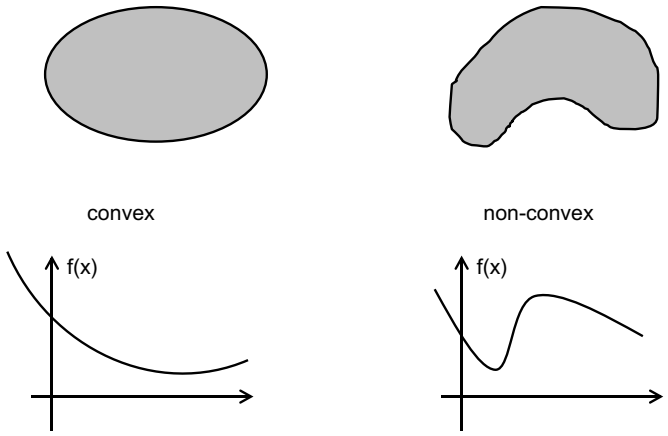


Fig. A.1 Convex sets and functions.

An important property of a convex function is that any local minimum is also a global minimum. Therefore, if $f(\mathbf{x})$ is a differentiable convex function of $\mathbf{x}$, and $df(\mathbf{x})/d\mathbf{x} = \mathbf{0}$ at a point $\mathbf{x}^*$, then $\mathbf{x}^*$ is a global minimum of $f(\mathbf{x})$. Furthermore, if $f(\mathbf{x})$ is a strictly convex function, then the minimum is unique. In several occasions throughout this survey, we determine an optimal solution by explicitly setting the derivative of a convex objective to $\mathbf{0}$. In this context, we make use of the following rule: For any symmetric matrix $\mathbf{A}$,

$$\frac{d\mathrm{Tr}(\mathbf{B}\mathbf{A}\mathbf{B}^T)}{d\mathbf{B}} = 2\mathbf{B}\mathbf{A}. \tag{A.1}$$

Often, in optimization problems there are constraints on the possible values of the input $\mathbf{x}$, resulting in constrained optimization. Consider the following general optimization problem

$$P : \begin{cases} \min\limits_{\mathbf{x}} & f(\mathbf{x}) \\ \text{s.t.} & f_i(\mathbf{x}) \leq 0, \quad 1 \leq i \leq N, \end{cases} \tag{A.2}$$

and denote its optimal solution by $\mathbf{x}^{\mathrm{opt}}$, assuming it is attained. If $f(\mathbf{x})$ and $f_i(\mathbf{x})$ for $1 \leq i \leq N$ are convex functions then we call P a standard convex optimization problem. In general there may be equality constraints which must be linear in order to ensure convexity; however, here we focus on inequalities. The problem is said to be feasible if

there exists an $\mathbf{x}$ satisfying $f_i(\mathbf{x}) \leq 0, 1 \leq i \leq N$. It is strictly feasible if $f_i(\mathbf{x}) < 0, 1 \leq i \leq N$ for some $\mathbf{x}$. It is well known that in such programs a local minimum is also a global minimum. Therefore, we can resort to local optimization methods and under relatively mild conditions solve the problem efficiently.

A.2 Duality Theory

Convex optimization methods provide efficient numerical algorithms but also analytical insight through the use of Lagrange duality theory. The first step in deriving this duality is associating a Lagrangian with problem (A.2):

$$\mathcal{L}(\mathbf{x}; \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{i=1}^{N} \lambda_i f_i(\mathbf{x}), \tag{A.3}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)^T$ is a vector with the Lagrange dual variables. The dual function is defined as

$$g(\boldsymbol{\lambda}) = \min_{\mathbf{x}} L(\mathbf{x}; \boldsymbol{\lambda}), \tag{A.4}$$

and the dual program by

$$D : \begin{cases} \max_{\boldsymbol{\lambda}} & g(\boldsymbol{\lambda}) \\ \text{s.t.} & \lambda_i \geq 0, \quad 1 \leq i \leq N. \end{cases} \tag{A.5}$$

The importance of duality theory stems from the following weak inequality:

$$g(\boldsymbol{\lambda}^{\text{opt}}) \leq f(\mathbf{x}^{\text{opt}}), \tag{A.6}$$

where $\boldsymbol{\lambda}^{\text{opt}}$ is the optimal solution to D (assuming it is attained). Thus, the dual program provides a lower bound on the optimal value of the primal program. This is useful as the dual is always a convex optimization problem which may be easier to solve than the original primal problem. In fact, the inequality holds for any feasible dual variable $\boldsymbol{\lambda}$ and not only for $\boldsymbol{\lambda}^{\text{opt}}$. This is an important property that allows to bound the gap between the optimal solution and any suboptimal approach.

Moreover, under simple technical conditions strong duality holds, namely there exists equality in (A.6). To state the results more formally we need the following definition:

Definition A.2 (Constraint qualification). A convex problem of the form (A.2) satisfies the constraint qualification if it is strictly feasible, and bounded from below.

Using this definition, we have the following important result:

Theorem A.1 (Strong duality in convex programming). Let (A.2) be a convex program which satisfies the constraint qualification in Definition A.2. Then the optimal value of (A.5) is attained and strong duality holds, i.e., $g(\boldsymbol{\lambda}^{\text{opt}}) = f(\mathbf{x}^{\text{opt}})$.

Strong duality allows us to find the optimal value of P by solving D. This may be advantageous when P is more complicated, or when we prefer to solve a maximization over a minimization. For example, duality can transform a minimax problem into a double minimization by replacing the inner maximization with its dual. One of the main disadvantages of solving P via D is that although their optimal values are the same, it is not always trivial to find the optimal solution $\mathbf{x}^{\text{opt}}$ as a function of $\boldsymbol{\lambda}^{\text{opt}}$.

Duality theory also leads to necessary and sufficient optimality conditions that often lead to further insight into the problem, or to closed form solutions. The most common are the Karush–Kuhn–Tucker (KKT) conditions[1]:

Definition A.3 (KKT conditions). The KKT conditions associated with (A.2) are that there exist a dual vector $\boldsymbol{\lambda} \succeq \mathbf{0}$ such that

(1) Complementary slackness: $\lambda_i f_i(\mathbf{x}) = 0$ for $1 \leq i \leq N$.
(2) Zero derivative of the Lagrangian: $\frac{dL(\mathbf{x};\boldsymbol{\lambda})}{d\mathbf{x}} = \mathbf{0}$, where $L(\mathbf{x};\boldsymbol{\lambda})$ is defined by (A.3).

---

[1] We assume that all the functions are differentiable.

The importance of this condition in convex programming lies in the following result:

Theorem A.2 (KKT conditions in convex programming). Let (A.2) be a convex program, and let $\mathbf{x}^*$ be a feasible solution. Then the KKT conditions are sufficient for $\mathbf{x}^*$ to be optimal. Moreover, if the constraint qualification in Definition A.2 holds, then the KKT conditions are also necessary.

A.3 Semidefinite Programs

The most common convex program is probably the linear program (LP), i.e., an optimization with a linear objective function and linear (affine) constraints:

$$\text{LP}: \begin{cases} \min\limits_{\mathbf{x}} & \mathbf{f}^T\mathbf{x} \\ \text{s.t.} & \mathbf{A}_i\mathbf{x} + \mathbf{b}_i \succeq \mathbf{0}, \quad 1 \leq i \leq N. \end{cases} \tag{A.7}$$

Here the inequality is an element-wise inequality. One of the main advances in modern convex optimization methods is the generalization of the results and algorithms for LP to conic programming, where the scalar inequalities are replaced by generalized conic inequalities.

In this survey, the main conic program we discuss is the semidefinite program (SDP), which is based on the notion of positive semi-definite matrices [141]. SDPs rely on the fact that the set $\mathbf{X} \succeq \mathbf{0}$ is convex in $\mathbf{X}$. The standard form of an SDP is

$$\text{SDP}: \begin{cases} \min\limits_{\mathbf{x}} & \mathbf{f}^T\mathbf{x} \\ \text{s.t.} & \mathbf{A}_0 + \sum_{i=1}^{N} x_i\mathbf{A}_i \succeq 0, \end{cases} \tag{A.8}$$

where $\mathbf{A}_i$ for $0 \leq i \leq N$ are symmetric matrices. The constraint in (A.8) is called a linear matrix inequality (LMI). LMIs are inequalities of the form $\mathbf{G}(\mathbf{x}) \succeq 0$, where $\mathbf{G}(\mathbf{x})$ is linear in $\mathbf{x}$. Once a problem is formulated as an SDP, standard software packages, such as the Self–Dual–Minimization (SeDuMi) [133], SDPT3 [138] or CVX [68], can be used to solve the problem in polynomial time within any desired accuracy. In practice though, SDPs with large matrices are difficult to solve.

Except for some slight differences, duality theory can be generalized to deal with conic programs. The main difference is the definition of the dual variables and the Lagrangian. With each conic inequality we associate a dual variable of the dual cone (in LP and SDP the dual cone is the cone itself). The Lagrangian is obtained by using the inner product with respect to that cone. Duality theory for SDP is used frequently and there are many good references on this topic [141]. The main idea is to associate a dual matrix variable $\Pi \succeq 0$ to each LMI constraint, and define the Lagrangian as

$$L(\mathbf{x};\Pi) = \mathbf{f}^T\mathbf{x} - \mathrm{Tr}\left(\Pi\left(\mathbf{A}_0 + \sum_{i=1}^{N} x_i\mathbf{A}_i\right)\right). \tag{A.9}$$

Note that the Lagrangian is formulated by subtracting the term resulting from the constraint. This element is subtracted instead of added (as in regular convex programming) because the cone is defined as a "*greater than or equal*" generalized inequality and not as a "*less than or equal*" inequality. The dual function is defined as before $g(\Pi) = \min_{\mathbf{x}} L(\mathbf{x};\Pi)$, and the dual program is obtained by maximizing $g(\Pi)$ over $\Pi \succeq \mathbf{0}$. The resulting dual turns out to also be an SDP. Conic duality theory states that under the constraint qualification condition in Definition A.2, the optimal values of the dual and primal programs are equal and attained; furthermore, the KKT conditions are necessary and sufficient for optimality of a feasible $\mathbf{x}$.

Although not always immediate or trivial, many practical optimization problems can be transformed into a standard conic program, and in particular to an SDP. An important lemma which often allows such a transformation is Schur's lemma [19, p. 28].

Lemma A.3 (Schur's Lemma). Let

$$\mathbf{M} = \begin{pmatrix} \mathbf{X} & \mathbf{Y}^T \\ \mathbf{Y} & \mathbf{Z} \end{pmatrix}$$

be a Hermitian matrix. Then $\mathbf{M} \succeq (\succ)\, 0$ if and only if $\mathbf{Z} \succeq (\succ)\, 0$, $\mathbf{X} - \mathbf{Y}^*\mathbf{Z}^\dagger\mathbf{Y} \succeq (\succ)\, 0$ and $\mathbf{Y}^T(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\dagger) = 0$. Equivalently, $\mathbf{M} \succeq (\succ)\, 0$ if and only if $\mathbf{X} \succeq (\succ)\, 0$, $\mathbf{Z} - \mathbf{Y}\mathbf{X}^\dagger\mathbf{Y}^T \succeq (\succ)\, 0$, and $\mathbf{Y}(\mathbf{I} - \mathbf{X}\mathbf{X}^\dagger) = 0$.

One of the uses of this lemma is in transforming quadratic constraints into standard LMIs. As a simple example, Schur's lemma may be used to transform the quadratic constraint $\Pi \succeq \mathbf{w}\mathbf{w}^T$ into the LMI

$$\begin{pmatrix} \Pi & \mathbf{w} \\ \mathbf{w}^T & \mathbf{I} \end{pmatrix} \succeq 0. \tag{A.10}$$

Another important lemma is the following [12, p. 163]:

Lemma A.4. Let $\mathbf{A}$ be a symmetric matrix. The condition $\mathbf{x}^T\mathbf{A}\mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c \geq 0$ holds for all $\mathbf{x}$ if and only if

$$\begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^T & c \end{pmatrix} \succeq 0. \tag{A.11}$$

A.3.1 Semidefinite Relaxation

In many practical problems it is difficult (or even provably impossible) to transform the problem into a convex form. In this case, a promising alternative is convex relaxation which approximates the problem by omitting the non-convex constraints. This approach is rather straightforward but will not necessarily lead to acceptable behavior. One of the tricks to improve the performance is to first lift the problem into a higher dimension and only then apply the relaxation. In many cases, this considerably improves the result.

A particular method is semidefinite relaxation (SDR) [12, 141] in which non-convex quadratic constraints are linearized in the space of semidefinite matrices. For example, consider the following non-convex problem:

$$P : \begin{cases} \max\limits_{\mathbf{x}} & \mathbf{x}^T\mathbf{A}_0\mathbf{x} \\ \text{s.t.} & \mathbf{x}^T\mathbf{A}_i\mathbf{x} + 2\mathbf{b}_i^T\mathbf{x} + c_i \leq 0, \quad 1 \leq i \leq N. \end{cases} \tag{A.12}$$

where $\mathbf{A}_i \succeq \mathbf{0},\ 0 \leq i \leq N$. The problem is non-convex due to the maximization of a convex function. To try and solve this problem, we can

reformulate it by defining $\mathbf{X} = \mathbf{x}\mathbf{x}^T$:

$$\begin{cases} \max\limits_{\mathbf{x},\mathbf{X}} & \mathrm{Tr}\mathbf{A}_0\mathbf{X} \\ \text{s.t.} & \mathrm{Tr}(\mathbf{X}\mathbf{A}_i) + 2\mathbf{b}_i^T\mathbf{x} + c_i \leq 0, \quad 1 \leq i \leq N \\ & \mathbf{X} \succeq \mathbf{0} \\ & \mathrm{rank}(\mathbf{X}) = 1. \end{cases} \tag{A.13}$$

So far, we just complicated the problem using additional matrix variables. However, now the objective is linear in $\mathbf{X}$, and the constraints are linear in $\mathbf{x}$ and $\mathbf{X}$. The only non-convex constraint is $\mathrm{rank}(\mathbf{X}) = 1$. Omitting it yields the SDR:

$$\mathrm{SDR}: \begin{cases} \max\limits_{\mathbf{x},\mathbf{X}} & \mathrm{Tr}\mathbf{A}_0\mathbf{X} \\ \text{s.t.} & \mathrm{Tr}(\mathbf{X}\mathbf{A}_i) + 2\mathbf{b}_i^T\mathbf{x} + c_i \leq 0, \quad 1 \leq i \leq N \\ & \mathbf{X} \succeq \mathbf{0}, \end{cases} \tag{A.14}$$

which is a convex relaxation of (A.12). If the optimal solution of the SDR is of rank-one, then we say that the relaxation is tight and we obtain an optimal solution for the original problem. Otherwise, we get a bound on the optimal value of the original problem.

One of the interesting properties of the SDR is that, in some cases, the same relaxation may be obtained using Lagrange duality theory. In particular, the SDR is often the convex bidual (dual of the dual) of the original problem. Indeed, for any optimization problem (not necessarily convex) there is a convex Lagrange dual program. The optimal value of the dual program is a bound on the optimal value of the original program. If the original problem was convex, then the bidual is usually the original problem itself (or a very similar problem with some change of variables). In non-convex programs, the bidual cannot be exactly the original problem, since it is always a convex program. Therefore, the bidual is considered as a standard technique to *convexify* non-convex problems. The interesting result is that in the example above, the bidual is exactly the SDR. There are a few examples of non-convex problems with a tight SDR. If it is also the Lagrange bidual, then this means that strong duality holds even though the problem is non-convex [7].

A.4 Minimax Problems

One of the important tools in convex optimization is minimax theory. It is a main ingredient in the derivation of Lagrange duality (where we minimize with respect to the primal variables, and maximize with respect to the dual variables). Furthermore, it is applicable in robust optimization problems which aim in optimizing worst case performance. The main result in this theory is given in the following proposition [123].

Proposition A.5. Let $\mathcal{X}$ and $\mathcal{Y}$ be convex compact sets, and let $f(\mathbf{x},\mathbf{y})$ be a continuous function which is convex in $\mathbf{x} \in \mathcal{X}$ for every fixed $\mathbf{y} \in \mathcal{Y}$ and concave in $\mathbf{y} \in \mathcal{Y}$ for every fixed $\mathbf{x} \in \mathcal{X}$. Then,

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}) = \max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\mathbf{y}),$$

and we can replace the order of the minimization and the maximization.

There are many variants of Proposition A.5 under weaker conditions. In particular, it is sufficient that only one of the sets will be compact, and convexity may be replaced by quasi-convexity.

References

[1] J. Aldrich, "R. A. Fisher and the making of maximum likelihood 1912–1922," *Statistical Science*, vol. 12, no. 3, pp. 162–176, 1997.

[2] D. Amir, "Chebychev centers and uniform convexity," *Pacific Journal of Mathematics*, vol. 77, no. 1, pp. 1–6, 1978.

[3] R. K. Bahr and J. A. Bucklew, "Minimax estimation of unknown deterministic signals in colored noise," *IEEE Transactions on Information Theory*, vol. 34, pp. 632–641, 1988.

[4] A. J. Baranchik, "Multiple regression and estimation of the mean of a multivariate normal distribution," Technical Report, 51, Stanford University, 1964.

[5] A. Beck, *Convexity Properties Associated with Nonconvex Quadratic Matrix Functions and Applications to Quadratic Programming,* preprint.

[6] A. Beck, A. Ben-Tal, and Y. C. Eldar, "Robust mean-squared error estimation of multiple signals in linear systems affected by model and noise uncertainties," *Mathematical Programming*, vol. 107, no. 1, pp. 155–187, 2006.

[7] A. Beck and Y. C. Eldar, "Strong duality in nonconvex quadratic optimization with two quadratic constraints," *SIAM Journal of Optimization*, vol. 17, no. 3, pp. 844–860, 2006.

[8] A. Beck and Y. C. Eldar, "Regularization in regression with bounded noise: A Chebyshev center approach," *SIAM Journal of Matrix Analysis Applications*, vol. 29, no. 2, pp. 606–625, 2007.

[9] A. Beck, Y. C. Eldar, and A. Ben-Tal, "Mean-squared error estimation for linear systems with block circulant uncertainty," *SIAM Journal of Matrix Analysis Applications*, vol. 29, pp. 712–730, 2007.

[10] Z. Ben-Haim and Y. C. Eldar, "Maximum set estimators with bounded estimation error," *IEEE Transactions on Signal Processing*, vol. 53, pp. 3172–3182, August 2005.

[11] Z. Ben-Haim and Y. C. Eldar, "Blind minimax estimation," *IEEE Transactions on Information Theory*, vol. 53, pp. 3145–3157, September 2007.

[12] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization*. MPS-SIAM Series on Optimization, 2001.

[13] A. Benazza-Benyahia and J.-C. Pesquet, "Building robust wavelet estimators for multicomponent images using Stein's principle," *IEEE Transactions on Image Processing*, vol. 14, no. 11, pp. 1814–1830, November 2005.

[14] J. Berger, "Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters," *Annals of Statistics*, vol. 8, no. 3, pp. 545–571, 1980.

[15] J. O. Berger, "Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss," *Annals of Statistics*, vol. 4, no. 1, pp. 223–226, January 1976.

[16] D. P. Bertsekas, *Nonlinear Programming*. Belmont MA: Athena Scientific, Second Edition, 1999.

[17] A. Björck, *Numerical Methods for Least-Squares Problems*. Philadelphia, PA: SIAM, 1996.

[18] M. E. Bock, "Minimax estimators of the mean of a multivariate normal distribution," *Annals of Statistics*, vol. 3, no. 1, pp. 209–218, January 1975.

[19] S. Boyd, L. E. Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*. Philadelphia, PA: SIAM, 1994.

[20] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[21] B. D. Carlson, "Covariance matrix estimation errors and diagonal loading in adaptive arrays," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 24, pp. 397–401, July 1988.

[22] A. Charras and C. van Eeden, "Bayes and admissibility properties of estimators in truncated parameter spaces," *The Canadian Journal of Statistics*, vol. 19, no. 2, pp. 121–114, June 1991.

[23] J. Chernoi and Y. C. Eldar, "Improving maximum likelihood for constrained denoising: An extended Chebyshev center approach," submitted to *IEEE Transactions on Signal Processing*.

[24] A. Cohen, "All admissible linear estimates of the normal mean," *Annals of Mathematical Statistics*, vol. 37, pp. 458–463, 1966.

[25] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*. Chapman and Hall, 1974.

[26] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. ASSP-35, pp. 1365–1376, October 1987.

[27] H. Cramér, "A contribution to the theory of statistical estimation," *Skandinavisk Aktuaries Tidskrif*, vol. 29, pp. 458–463, 1946.

[28] H. Cramér, *Mathematical Methods of Statistics*. Princeton, NJ: Princeton Univ. Press, 1946.

[29] G. Darmois, "Sur les lois de probabilites a estimation exhaustive," *C.R. Academy of Science Paris*, vol. 200, pp. 1265–1266, 1935.

[30] G. Darmois, "Sur les lois limites de la dispersion de certaines estimations," *Rev. Inst. Znt. Stat.*, vol. 13, pp. 9–15, 1945.

[31] G. Demoment, "Image reconstruction and restoration: Overview of common estimation structures and problems," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 37, pp. 2024–2036, 1989.

[32] L. Desbat and D. Girard, "The "minimum reconstruction error" choice of regularization parameters: Some effective methods and their application to deconvolution problems," *SIAM Journal of Science Computation*, vol. 16, no. 6, pp. 1387–1403, November 1995.

[33] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, December 1995.

[34] B. Efron, "Biased versus unbiased estimation," *Advances Mathematics*, vol. 16, pp. 259–277, 1975.

[35] B. Efron and C. Morris, "Combining possibly related estimation problems," *Journal of Royal Statistical Society B*, vol. 35, no. 3, pp. 379–421, 1973.

[36] B. Efron and C. Morris, "Steins estimation rule and its competitors: An empirical Bayes approach," *Journal of American Statistical Association*, vol. 68, pp. 117–130, 1973.

[37] B. Efron and C. Morris, "Stein's paradox in statistics," *Scientific American*, vol. 236, pp. 119–127, 1977.

[38] Y. C. Eldar, "The Chebyshev estimate: Statistical aspects," to appear in *Journal of Statistical Planning and Inference*.

[39] Y. C. Eldar, "Generalized SURE for exponential families: Applications to regularization," to appear in *IEEE Transactions on Signal Processing*; available at http://arxiv.org/abs/0804.3010.

[40] Y. C. Eldar, "MSE bounds with affine bias dominating the Cramér-Rao bound," to appear in *IEEE Transactions on Signal Processing*.

[41] Y. C. Eldar, "Minimum variance in biased estimation: Bounds and asymptotically optimal estimators," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1915–1930, July 2004.

[42] Y. C. Eldar, "Robust deconvolution of deterministic and random signals," *IEEE Transactions on Information Theory*, vol. 51, pp. 2921–2929, August 2005.

[43] Y. C. Eldar, "Comparing between estimation approaches: Admissible and dominating linear estimators," *IEEE Transactions on Signal Processing*, vol. 54, pp. 1689–1702, May 2006.

[44] Y. C. Eldar, "Minimax estimation of deterministic parameters in linear models with a random model matrix," *IEEE Transactions on Signal Processing*, vol. 54, no. 2, pp. 601–612, February 2006.

[45] Y. C. Eldar, "Minimax MSE estimation with noise covariance uncertainties," *IEEE Transactions on Signal Processing*, vol. 54, no. 1, pp. 138–145, January 2006.

[46] Y. C. Eldar, "Robust competitive estimation with signal and noise covariance uncertainties," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4532–4547, October 2006.

[47] Y. C. Eldar, "Uniformly improving the Cramér-Rao bound and maximum-likelihood estimation," *IEEE Transactions on Signal Processing*, vol. 54, pp. 2943–2956, August 2006.

[48] Y. C. Eldar, "Universal weighted MSE improvement of the least-squares estimator," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1788–1800, 2008.

[49] Y. C. Eldar, A. Beck, and M. Teboulle, "A minimax Chebyshev estimator for bounded error estimation," *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1388–1397, April 2008.

[50] Y. C. Eldar, A. Ben-Tal, and A. Nemirovski, "Linear minimax regret estimation of deterministic parameters with bounded data uncertainties," *IEEE Transactions on Signal Processing*, vol. 52, pp. 2177–2188, August 2004.

[51] Y. C. Eldar, A. Ben-Tal, and A. Nemirovski, "Robust mean-squared error estimation in the presence of model uncertainties," *IEEE Transactions on Signal Processing*, vol. 53, pp. 168–181, January 2005.

[52] Y. C. Eldar and J. Chernoi, "A pre-test like estimator dominating the least-squares method," *Journal of Statistical Planning and Inference*, vol. 138, no. 10, pp. 3069–3085, 1 October 2008.

[53] Y. C. Eldar and N. Merhav, "A competitive minimax approach to robust estimation of random parameters," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1931–1946, July 2004.

[54] Y. C. Eldar and N. Merhav, "Minimax MSE–ratio estimation with signal covariance uncertainties," *IEEE Transactions on Signal Processing*, vol. 53, pp. 1335–1347, April 2005.

[55] Y. C. Eldar and A. Nehorai, "Mean-squared error beamforming for signal estimation: A competitive approach," in *Robust Adaptive Beamforming*, (J. Li and P. Stoica, eds.), pp. 259–298, John Wiley & Sons, Inc., 2006.

[56] Y. C. Eldar, A. Nehorai, and P. S. La Rosa, "An expected least-squares beamforming approach to signal estimation with steering vector uncertainties," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 288–291, May 2006.

[57] Y. C. Eldar, A. Nehorai, and P. S. La Rosa, "A competitive mean-squared error approach to beamforming," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5143–5154, November 2007.

[58] Y. C. Eldar and A. V. Oppenheim, "Covariance shaping least-squares estimation," *IEEE Transactions on Signal Processing*, vol. 51, pp. 686–697, March 2003.

[59] R. A. Fisher, "On an absolute criterion for fitting frequency curves," *Messenger of Mathematics*, vol. 41, pp. 155–160, 1912.

[60] M. Fréchet, "Sur l'extension de certaines evaluations statistiques de petits echantillons," *Rev. Inst. Znt. Stat.*, vol. 11, pp. 128–205, 1943.

[61] N. P. Galatsanos and A. K. Katsaggelos, "Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation," *IEEE Transactions on Image Processing*, vol. 1, no. 3, pp. 322–336, 1992.

[62] K. F. Gauss, *Theoria Combination is Obsercationunt Erronbus Minimis Obnoxiae*. 1821.
[63] K. F. Gauss, *Theory of Motion of Heavenly Bodies Moving about the Sun in Conic Sections*. New York, NY: Dover, 1963.
[64] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, May 1979.
[65] I. J. Good and R. A. Gaskins, "Nonparametric roughness penalties for probability densities," *Biometrika*, vol. 58, pp. 255–277, 1971.
[66] I. J. Good and R. A. Gaskins, "Density estimation and bump hunting by the penalized likelihood method exemplified by scattering and meteorite data (with discussion and rejoinder)," *Journal of American Statistics*, vol. 10, pp. 811–824, 1980.
[67] J. D. Gorman and A. O. Hero, "Lower bounds for parametric estimation with constraints," *IEEE Transactions on Information Theory*, vol. 26, pp. 1285–1301, November 1990.
[68] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming (web page and software)," http://stanford.edu/˜boyd/cvx, March 2008.
[69] E. Greenberg and C. E. Webster Jr., *Advanced Econometrics*. New York: Wiley, Second Edition, 1983.
[70] M. H. J. Gruber, *Regression Estimators: A Comparative Study*. San Diego, CA: Academic Press Inc., 1990.
[71] P. R. Halmos, "The theory of unbiased estimation," *Annals of Mathematical Statistics*, vol. 17, pp. 34–43, 1946.
[72] M. Hanke and P. C. Hansen, "Regularization methods for large-scale problems," *Surveys Mathematical Industry*, vol. 3, no. 4, pp. 253–315, 1993.
[73] P. C. Hansen, "The use of the L-curve in the regularization of discrete ill-posed problems," *SIAM Journal of Science Statistical Computation*, vol. 14, pp. 1487–1503, 1993.
[74] P. C. Hansen, "Regularization tools, a matlab package for analysis of discrete regularization problems," *Numerical Algorithms*, vol. 6, pp. 1–35, 1994.
[75] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. Philadelphia, PA: SIAM, 1998.
[76] P. C. Hansen, J. G. Nagy, and D. P. O'Leary, *Deblurring Images: Matrices, Spectra, and Filtering*. Philadelphia, PA: SIAM, 2006.
[77] M. Hardy, "An illuminating counterexample," vol. 110, no. 3, pp. 234–238, 2003.
[78] A. O. Hero, "A Cramer-Rao type lower bound for essentially unbiased parameter estimation," Technical Report, 890, DTIC AD-A246666, MIT Lincoln Lab, Lexington, MA, January 1992.
[79] A. O. Hero and J. A. Fessler, "A recursive algorithm for computing CR-type bounds on estimator covariance," *IEEE Transactions on Information Theory*, vol. 40, pp. 1205–1210, July 1994.
[80] A. O. Hero, J. A. Fessler, and M. Usman, "Exploring estimator bias-variance tradeoffs using the uniform CR bound," *IEEE Transactions on Signal Processing*, vol. 44, no. 8, pp. 2026–2041, August 1996.

[81] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, February 1970.
[82] K. Hoffmann, "Characterization of minimax linear estimators in linear regression," *Statistics*, vol. 10, no. 1, pp. 19–26, 1979.
[83] K. Hoffmann, "Admissible improvements of the least squares estimator," *Statistics*, vol. 11, pp. 373–388, 1980.
[84] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, UK: Cambridge Univ. Press, 1985.
[85] P. J. Huber, *Robust Statistics*. New York: NY, John Wiley & Sons, Inc., 1981.
[86] H. M. Hudson, "A natural identity for exponential families with applications in multiparameter estimation," *Annals of Statistics*, vol. 6, no. 3, pp. 473–484, 1978.
[87] J. T. Hwang, "Improving upon standard estimators in discrete exponential families with applications to poisson and negative binomial cases," *Annals of Statistics*, vol. 10, no. 3, pp. 857–867, 1982.
[88] W. James and C. Stein, "Estimation of quadratic loss," in *Proceedings of Fourth Berkeley Symposium on Mathematical Statistics Probability*, vol. 1, pp. 361–379, Berkeley: University of California Press, 1961.
[89] W. C. Karl, "Regularization in image restoration and reconstruction," in *Handbook of Image and Video Processing*, (A. Bovik, ed.), pp. 183–202, ELSEVIER, Second Edition, 2005.
[90] S. A. Kassam and H. V. Poor, "Robust signal processing for communication systems," *IEEE Communication Magazine*, vol. 21, pp. 20–28, 1983.
[91] S. A. Kassam and H. V. Poor, "Robust techniques for signal processing: A survey," *IEEE Proceedings*, vol. 73, pp. 433–481, March 1985.
[92] S. Kay and Y. C. Eldar, "Rethinking biased estimation," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 133–136, May 2008.
[93] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ: Prentice Hall Inc., 1993.
[94] B. Koopman, "On distribution admitting a sufficient statistic," *Transactions on American Mathematical Society*, vol. 39, pp. 399–409, 1936.
[95] T. Larsson, M. Patriksson, and A.-B. Strömberg, "On the convergence of conditional ϵ-subgradient methods for convex programs and convex-concave saddle-point problems," *European Journal of Operations Research*, vol. 151, no. 3, pp. 461–473, 2003.
[96] H. Later, "A minimax linear estimator for linear parameters under restrictions in form of inequalities," *Statistics*, vol. 6, pp. 689–695, 1975.
[97] A.-M. Legendre, *Nouvelles méthodes pour la détermination des orbites des cometes*. 1806.
[98] E. L. Lehmann, "A general concept of unbiasedness," *Annals of Mathematical Statistics*, vol. 22, pp. 587–592, December 1951.
[99] E. L. Lehmann, "Estimation with inadequate information," vol. 78, no. 383, pp. 624–627, September 1983.
[100] E. L. Lehmann and G. Casella, *Theory of Point Estimation*. New York, NY: Springer-Verlag, Inc., Second Edition, 1998.
[101] E. H. Lieb and M. Loss, *Analysis*. American Mathematical Society, Second Edition, 2001.

[102] F. Luisier, T. Blu, and M. Unser, "A new SURE approach to image denoising: Interscale orthonormal wavelet thresholding," *IEEE Transactions on Image Processing*, vol. 16, no. 3, pp. 593–606, 2007.

[103] J. H. Manton, V. Krishnamurthy, and H. V. Poor, "James-Stein state filtering algorithms," *IEEE Transactions on Signal Processing*, vol. 46, pp. 2431–2447, September 1998.

[104] D. W. Marquardt, "Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation," *Technometrics*, vol. 12, no. 3, pp. 592–612, August 1970.

[105] Y. Maruyama, "A unified and broadened class of admissible minimax estimators of a multivariate normal mean," *Journal of Multivariate Analysis*, vol. 64, pp. 196–205, 1998.

[106] T. L. Marzetta, "A simple derivation of the constrained multiple parameter Cramér-Rao bound," *IEEE Transactions on Signal Processing*, vol. 41, pp. 2247–2249, June 1993.

[107] L. S. Mayer and T. A. Willke, "On biased estimation in linear models," *Technometrics*, vol. 15, pp. 497–508, August 1973.

[108] L. J. Meng and N. H. Clinthorne, "A modified uniform Cramer-Rao bound for multiple pinhole aperture design," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 896–902, July 2004.

[109] M. Milanese and G. Belforte, "Estimation theory and uncertainty intervals evaluation in presence of unknown but bounded errors: Linear families of models and estimators," *IEEE Transactions on Automatic Control*, vol. 27, no. 2, pp. 408–414, 1982.

[110] R. Molina, A. K. Katsaggelos, and J. Mateos, "Bayesian and regularization methods for hyperparameter estimation in image restoration," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 231–246, 1999.

[111] V. A. Morozov, *Methods for Solving Incorrectly Posed Problems*. New York, NY: Springer-Verlag, 1984.

[112] A. Nemirovski, "Prox-method with rate of convergence $o(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems," *SIAM Journal of Optimization*, vol. 15, pp. 229–251, 2004.

[113] Y. Nesterov and A. Nemirovski, *Interior-Point Polynomial Algorithms in Convex Programming*. Philadelphia, PE: SIAM, 1994.

[114] J. P. Norton, "Identification and application of bounded parameter models," *Automatica*, vol. 23, pp. 497–507, 1987.

[115] F. O'Sullivan, "A statistical perspective on ill-posed inverse problems," *Statistical Science*, vol. 1, no. 4, pp. 502–527, 1986.

[116] M. S. Pinsker, "Optimal filtering of square-integrable signals in Gaussian noise," *Problems Information on Transactions*, vol. 16, pp. 120–133, 1980.

[117] E. Pitman, "Sufficient statistics and intrinsic accuracy," *Proceedings of Cambridge Philosophical Society*, vol. 32, pp. 567–579, 1936.

[118] S. Ramani, T. Blu, and M. Unser, "Blind optimization of algorithm parameters for signal denoising by Monte-Carlo SURE," in *Proceedings of International Conference on Acoustics, Speech, Signal Processing (ICASSP-2008), (Las-Vegas, NV)*, April 2008.

[119] C. R. Rao, "Minimum variance and the estimation of several parameters," in *Proceedings of Cambridge Philosophical Society*, pp. 280–283, 1946.

[120] C. R. Rao, *Linear Statistical Inference and Its Applications*. New York, NY: John Wiley & Sons, Inc., Second Edition, 1973.

[121] C. R. Rao, "Estimation of a parameter in linear models," *Annals of Statistics*, vol. 4, pp. 1023–1037, 1976.

[122] J. Rice, "Choice of smoothing parameter in deconvolution problems," *Contemporary Mathematics*, vol. 59, pp. 137–151, 1986.

[123] R. T. Rockafellar, *Convex Analysis*. Princeton NJ: Princeton Univ. Press, 1970.

[124] J. P. Romano and A. F. Siegel, *Counterexamples in Probability and Statistics*. Monterey, CA: Wadsworth & Brooks, 1985.

[125] Y. Rong, Y. C. Eldar, and A. B. Gershman, "Performance tradeoffs among adaptive beamforming criteria," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 651–659, December 2007.

[126] D. L. Snyder and M. I. Miller, *Random Point Processes in Time and Space*. New York: Springer-Verlag, 1991.

[127] P. Speckman, "Spline smoothing and optimal rates of convergence in nonparametric regression models," *Annals of Statistics*, vol. 13, no. 3, pp. 970–983, 1985.

[128] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," in *Proceedings on Third Berkeley Symposium on Mathematical Statistical Probability*, vol. 1, pp. 197–206, Berkeley: University of California Press, 1956.

[129] C. M. Stein, "Estimation of the mean of a multivariate distribution," *Proceedings of Prague Symposium on Asymptotic Statistics*, pp. 345–381, 1973.

[130] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *Annals of Statistics*, vol. 9, no. 6, pp. 1135–1151, November 1981.

[131] P. Stoica and R. L. Moses, *Introduction to Spectral Analysis*. Upper Saddle River, NJ: Prentice Hall Inc., 1997.

[132] A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics*. Vol. 2, London: Edward Arnold, Fifth Edition, 1991.

[133] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11–12, pp. 625–653, 1999.

[134] P. Tichavsky, C. H. Muravchik, and A. Nehorai, "Posterior Cramer-Rao bounds for discrete-time nonlinear filtering," *IEEE Transactions on Signal Processing*, vol. 46, no. 5, pp. 1386–1396, May 1998.

[135] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Soviet Mathematique Dokldy*, vol. 5, pp. 1035–1038, 1963.

[136] A. N. Tikhonov and V. Y. Arsenin, *Solution of Ill-Posed Problems*. Washington, DC: V.H. Winston, 1977.

[137] D. M. Titterington, "Common structure of smoothing techniques in statistics," *International Statistics Review*, vol. 53, pp. 141–170, 1985.

[138] K. C. Toh, M. J. Todd, and R. H. Tütüncü, "SDPT3 — a MATLAB software package for semidefinite programming, version 1.3," *Optimization Methods Software*, vol. 11/12, no. 1–4, pp. 545–581, 1999 (Interior point methods).

[139] J. F. Traub, G. Wasikowski, and H. Wozinakowski, *Information-Based Complexity*. New York: Academic, 1988.

[140] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*. John Willey and Sons Inc., 1968.

[141] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Review*, vol. 38, no. 1, pp. 40–95, March 1996.

[142] K. S. Vastola and H. V. Poor, "Robust Wiener-Kolmogorov theory," *IEEE Transactions on Information Theory*, vol. IT-30, pp. 316–327, March 1984.

[143] A. Wiesel, "Convex optimization methods in MIMO communication systems," Technical Report, Ph.D. thesis, Technion — Israel Institute of Technology, 2007.

[144] A. Wiesel, Y. C. Eldar, and A. Beck, "Maximum likelihood estimation in linear models with a Gaussian model matrix," *IEEE Signal Processing Letter*, vol. 13, no. 5, pp. 292–295, May 2006.

[145] A. Wiesel, Y. C. Eldar, and Y. Yeredor, "Linear regression with Gaussian model uncertainty: Algorithms and bounds," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2194–2205, June 2008.

[146] S. Zacks, *The Theory of Statistical Inference*. John Willey and Sons, Inc., 1971.

[147] X. P. Zhang and M. D. Desai, "Adapting denoising based on SURE risk," *IEEE Signal Processing Letters*, vol. 5, no. 10, pp. 265–267, 1998.