Blind Minimax Estimation

Zvika Ben-Haim and Yonina C. Eldar, *Member, IEEE*

***Abstract*—We consider the linear regression problem of estimating an unknown, deterministic parameter vector based on measurements corrupted by colored Gaussian noise. We present and analyze blind minimax estimators (BMEs), which consist of a bounded parameter set minimax estimator, whose parameter set is itself estimated from measurements. Thus, our approach does not require any prior assumption or knowledge, and the proposed estimator can be applied to any linear regression problem. We demonstrate analytically that the BMEs strictly dominate the least-squares (LS) estimator, i.e., they achieve lower mean-squared error (MSE) for any value of the parameter vector. Both Stein's estimator and its positive-part correction can be derived within the blind minimax framework. Furthermore, our approach can be readily extended to a wider class of estimation problems than Stein's estimator, which is defined only for white noise and nontransformed measurements. We show through simulations that the BMEs generally outperform previous extensions of Stein's technique.**

***Index Terms*—Biased estimation, James–Stein estimation, minimax estimation, linear regression model.**

I. Introduction

THE problem of estimating a parameter vector from noisy measurements has countless applications in science and engineering. Such estimation problems are typically modeled either in a Bayesian setting, in which a prior distribution on the parameter is assumed, or in a deterministic setting, in which such a prior is not assumed [1]. This paper examines the deterministic estimation problem. We further assume that the measurements $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}+\boldsymbol{w}$ are linear combinations of the parameter vector $\boldsymbol{x}$, to which Gaussian noise $\boldsymbol{w}$ is added. Here the transformation matrix $\boldsymbol{H}$ and the noise covariance are assumed to be known. We seek an estimate $\hat{\boldsymbol{x}}$ which approximates $\boldsymbol{x}$ in the sense of minimal mean-squared error (MSE).

This ubiquitous problem was first addressed by Gauss [2] and Legendre [3], who proposed the classical least-squares (LS) estimator. Several lines of reasoning can be used to support the LS approach. One argument is that the LS estimator minimizes the squared error between the measurements $\boldsymbol{y}$ and the transformed estimate $\hat{\boldsymbol{y}} = \boldsymbol{H}\hat{\boldsymbol{x}}$. The LS estimator is also the maximum-likelihood solution for Gaussian noise. However, neither of these criteria are directly related to the MSE, or to any other measure of the distance between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$. Another property of the LS solution is that it is the unbiased estimator achieving minimal MSE. Yet by removing the requirement of unbiasedness, estimators yielding lower MSE can be constructed. While linearity and unbiasedness may be intuitively appealing properties, they are not directly related to the primary goal at hand, namely, achieving low estimation error. Indeed, there are many examples in which the requirement of unbiasedness results in absurd estimators [4].

Because the parameter vector $\boldsymbol{x}$ is deterministic, the MSE $E\{\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|^2\}$ is generally a function of $\boldsymbol{x}$. In other words, one method may be better than another for some values of $\boldsymbol{x}$, and worse for other values. For instance, the trivial estimator $\hat{\boldsymbol{x}} = \boldsymbol{0}$ achieves optimal MSE when $\boldsymbol{x} = \boldsymbol{0}$, but its performance is otherwise poor. Nonetheless, it is possible to impose a partial order among estimation techniques [5], as follows. An estimator $\hat{\boldsymbol{x}}_1$ is said to *strictly dominate* a different estimator $\hat{\boldsymbol{x}}_2$ if the MSE of $\hat{\boldsymbol{x}}_1$ is lower than that of $\hat{\boldsymbol{x}}_2$, for all values of $\boldsymbol{x}$. If the MSE of $\hat{\boldsymbol{x}}_1$ is never higher than that of $\hat{\boldsymbol{x}}_2$, and is strictly lower for at least one parameter value, then $\hat{\boldsymbol{x}}_1$ is said to *dominate* $\hat{\boldsymbol{x}}_2$. An estimator is said to be *admissible* if it is not dominated by any other estimator. Surprisingly, when the parameter vector contains three or more elements, the LS method turns out to be inadmissible, i.e., some techniques *always* achieve lower MSE [6]. Thus, it is of interest to characterize the class of admissible estimators, and to find techniques which dominate LS.

The study of admissibility is sometimes restricted to linear methods $\hat{\boldsymbol{x}} = \boldsymbol{G}\boldsymbol{y}$. A linear admissible estimator is one which is not dominated by any other linear strategy. A simple rule characterizes the class of linear admissible techniques [7], and, given any linear inadmissible estimator, it is possible to construct a linear admissible alternative which dominates it [8]. However, the problem of admissibility is considerably more intricate when the linearity restriction is removed; generally, admissible estimators are either trivial (e.g., $\hat{\boldsymbol{x}} = \boldsymbol{0}$) or exceedingly complex [9], [10]. As a result, much research has focused on finding simple nonlinear techniques which dominate LS.

Early work on LS-dominating strategies considered the independent and identically distribution (i.i.d.) case, for which $\boldsymbol{H} = \boldsymbol{I}$ and the noise is white. Among these, the James–Stein estimator [5], [11] is the best known example; other approaches include the works of Stein [6] and Thompson [12]. Various "extended" James–Stein methods were later constructed for the general (non-i.i.d.) case [13]–[16]. Of these, Bock's technique [13] is quoted most often [16], [17]. However, none of these approaches has become a standard alternative to the LS estimator, and they are rarely used in practice in engineering applications [16]. 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 [18]). Another reason is that many of these approaches (including Bock's method) result in shrinkage estimators, consisting of a gain factor multiplying the LS estimate. Shrinkage techniques can certainly be used to reduce MSE; however, in the

Manuscript received April 23, 2006; revised May 22, 2007. This work was supported by the Israel Science Foundation under Grant 536/04.

The authors are with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Technion City, Haifa 32000, Israel (e-mail: zvikabh@technion.ac.il; yonina@ee.technion.ac.il).

Communicated by X. Wang, Associate Editor for Detection and Estimation.
Digital Object Identifier 10.1109/TIT.2007.903118

non-i.i.d. 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 an image reconstruction problem, multiplying the entire image by a constant does not improve quality.

In this paper, 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 [19], [20]. 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 i.i.d. case, and they continue to dominate the LS solution in the general, non-i.i.d. case 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 nonshrinkage, which improves their performance. Furthermore, extensive simulations show that BMEs considerably outperform Bock's method.

BMEs are based on linear minimax estimators over a bounded parameter set [21], [22]. These are linear methods designed for a slightly different problem, in which the parameter is known to belong to a given set. The minimax approach has been thoroughly studied in this setting, and closed-form solutions are known for many types of sets [8], [22]. In our case, however, no prior information about the parameter set is assumed. Instead, the blind minimax approach makes use of a two-stage process (Section II): first, a set is estimated from the measurements; next, a minimax method for this set is used to estimate the parameter itself. The result may be viewed as a simple decision rule, independent of this two-stage construction process. Indeed, our LS-dominance proofs do not rely on the method by which the techniques are generated. In particular, the dominance results do not depend 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 method by which the parameter set is estimated. In Section III, we study the case in which the estimated set is a sphere; Section IV derives estimators based on an ellipsoidal parameter set. Section V demonstrates that several existing Stein-type methods can also be derived in the blind minimax framework. Section VI compares the blind minimax approach with LS regularization techniques, while in Section VII, the BMEs are compared with other Stein-type decision rules. The paper concludes with a discussion in Section VIII.

Throughout this paper, vectors are denoted by lowercase boldface letters, and matrices by uppercase boldface letters. The ith component of a vector $\boldsymbol{v}$ is written as v_i. $\boldsymbol{T}^{1/2}$ indicates the (unique) positive semidefinite square root of a positive semidefinite matrix $\boldsymbol{T}$. The notation $\tilde{\boldsymbol{u}} \sim \mathcal{N}_p(\boldsymbol{u}, \boldsymbol{Q})$ signifies that $\tilde{\boldsymbol{u}}$ is a random vector of length p, distributed normally with mean $\boldsymbol{u}$ and covariance $\boldsymbol{Q}$. $\|\boldsymbol{x}\|^2$ is the Euclidean norm $\boldsymbol{x}^*\boldsymbol{x}$, and $\|\boldsymbol{x}\|_{\boldsymbol{T}}^2$ is the $\boldsymbol{T}$-norm $\boldsymbol{x}^*\boldsymbol{T}\boldsymbol{x}$, where $\boldsymbol{T}$ is a positive definite matrix. Finally, $\mathrm{diag}(a_1, \ldots, a_n)$ refers to the $n \times n$ diagonal matrix whose diagonal elements are $a_1, \ldots, a_n$.

II. Blind Minimax Estimation

Consider the problem of estimating an unknown deterministic parameter vector $\boldsymbol{x} \in \mathbb{C}^m$ from measurements $\boldsymbol{y} \in \mathbb{C}^n$ given by

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{w} \tag{1}$$

where $\boldsymbol{H} \in \mathbb{C}^{n \times m}$ is a known matrix and $\boldsymbol{w}$ is a Gaussian random vector with zero mean and covariance $\boldsymbol{C}_{\boldsymbol{w}}$. For simplicity, we assume that $\boldsymbol{H}$ is full-rank and that $\boldsymbol{C}_{\boldsymbol{w}}$ is positive definite.

The standard solution to this regression problem is the LS approach

$$\hat{\boldsymbol{x}}_{\mathrm{LS}} = \left(\boldsymbol{H}^* \boldsymbol{C}_{\boldsymbol{w}}^{-1} \boldsymbol{H}\right)^{-1} \boldsymbol{H}^* \boldsymbol{C}_{\boldsymbol{w}}^{-1} \boldsymbol{y}. \tag{2}$$

The MSE of $\hat{\boldsymbol{x}}_{\mathrm{LS}}$ does not depend on the value of $\boldsymbol{x}$, and is given by

$$\epsilon_0 = E\{\|\hat{\boldsymbol{x}}_{\mathrm{LS}} - \boldsymbol{x}\|^2\} = \mathrm{Tr}(\boldsymbol{Q}^{-1}) \tag{3}$$

where

$$\boldsymbol{Q} = \boldsymbol{H}^* \boldsymbol{C}_{\boldsymbol{w}}^{-1} \boldsymbol{H}. \tag{4}$$

Despite the popularity of the LS method, other estimators are known to achieve lower MSE. We propose a novel strategy leading to such LS-dominating techniques, namely, the blind minimax approach. To illustrate this concept, suppose for a moment that $\boldsymbol{x}$ is known to lie within a compact parameter set $\mathcal{S}$. In this case, a linear minimax estimator over the set $\mathcal{S}$ may be constructed [8], [21], [22]. This is the linear estimator $\hat{\boldsymbol{x}}_{\mathrm{M}} = \boldsymbol{G}\boldsymbol{y}$ minimizing the worst case MSE among all possible values of $\boldsymbol{x}$ in $\mathcal{S}$

$$\hat{\boldsymbol{x}}_{\mathrm{M}} = \arg \min_{\hat{\boldsymbol{x}} = \boldsymbol{G}\boldsymbol{y}} \max_{\boldsymbol{x} \in \mathcal{S}} E\{\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|^2\}. \tag{5}$$

A closed-form solution of (5) has been previously derived for many cases of interest. Furthermore, it has been shown that any linear minimax estimator achieves lower MSE than that of the LS method, for all values of $\boldsymbol{x}$ in $\mathcal{S}$ [8], [19]. Thus, as long as *some* bounded set is known to contain $\boldsymbol{x}$, minimax techniques outperform the LS estimator.

BMEs utilize minimax estimators when no parameter set is known. This is done in a two-stage process:

1) A parameter set $\mathcal{S}$ is estimated from the measurements.
2) A minimax estimator designed for $\mathcal{S}$ is used to estimate the parameter vector $\boldsymbol{x}$.

Various methods for estimating the parameter set $\mathcal{S}$ can be used, resulting in a variety of BMEs. In this paper, we consider sets of the form $\{\boldsymbol{x} : \boldsymbol{x}^*\boldsymbol{T}\boldsymbol{x} \leq L^2\}$. In the next section, we examine the case $\boldsymbol{T} = \boldsymbol{I}$, in which the parameter set is spherical, resulting in a shrinkage estimator. Subsequently, in Section IV, we discuss the more general case in which $\boldsymbol{T} = (\boldsymbol{H}^* \boldsymbol{C}_{\boldsymbol{w}}^{-1} \boldsymbol{H})^b$ for some real number b. In both cases, closed forms are provided, and dominance over the LS method is demonstrated.

III. The Spherical Blind Minimax Estimator

In this section, we apply the blind minimax technique using a spherical parameter set $\mathcal{S}$ whose radius L will be estimated from measurements. We assume for now that the sphere is centered

on the origin, $\mathcal{S} = \{\boldsymbol{x} : ||\boldsymbol{x}||^2 \leq L^2\}$. For a given value of L, the linear minimax estimator is [22]

$$\hat{\boldsymbol{x}}_{\rm M} = \frac{L^2}{L^2 + \epsilon_0}\hat{\boldsymbol{x}}_{\rm LS} \tag{6}$$

where $\hat{\boldsymbol{x}}_{\rm LS}$ is the LS estimator (2) and ϵ_0 is the MSE (3) of $\hat{\boldsymbol{x}}_{\rm LS}$. The resulting spherical BME (SBME) will have the form (6), where L^2 is estimated from the measurements.

As an estimate of L^2, we seek a value as close as possible to $||\boldsymbol{x}||^2$: a smaller value would exclude the true vector $\boldsymbol{x}$ from the parameter set, while a larger value would yield an overly conservative estimator. Since $\boldsymbol{x}$ is unknown, a natural alternative is to use $\hat{\boldsymbol{x}}_{\rm LS}$ instead. Thus, we propose to estimate L^2 as $||\hat{\boldsymbol{x}}_{\rm LS}||^2$. Substituting into (6), the SBME is then given by

$$\hat{\boldsymbol{x}}_{\rm SBM} = \frac{||\hat{\boldsymbol{x}}_{\rm LS}||^2}{||\hat{\boldsymbol{x}}_{\rm LS}||^2 + \epsilon_0}\hat{\boldsymbol{x}}_{\rm LS}. \tag{7}$$

In the i.i.d. case, the SBME reduces to the well-known Thompson estimator [12]. Under suitable conditions, Thompson's technique is known to strictly dominate the LS estimator, meaning that it achieves lower MSE for all values of $\boldsymbol{x}$ [23]. However, the SBME is equally well-defined for the non-i.i.d. case. As we shall see, the SBME strictly dominates LS in the non-i.i.d. case, and can thus be viewed as a generalization of Thompson's results. In Section V, we will demonstrate that the blind minimax approach can be used to derive generalizations of additional well-known methods, including Stein's estimator.

Up to this point, we have arbitrarily chosen the parameter set to be centered on the origin. The result was a weighted average between the LS estimate and $\mathbf{0}$. Averaging with a constant value $\mathbf{0}$ may be viewed as a restraint, which lessens the effect of measurement noise. As we shall see, the proposed BMEs outperform the LS estimator. This result demonstrates the fact that the LS approach results in an overestimate: reducing the norm of $\hat{\boldsymbol{x}}_{\rm LS}$ improves its performance. However, the choice of a parameter set centered on the origin is completely arbitrary; BMEs may be constructed around any constant center point $\boldsymbol{x}_0$ [17]. This will result in a weighted average between $\hat{\boldsymbol{x}}_{\rm LS}$ and $\boldsymbol{x}_0$, which may be useful if the parameter vector is expected to lie near a particular point. Thus, the "off-center" SBME is given by

$$\hat{\boldsymbol{x}} = \left(\frac{||\hat{\boldsymbol{x}}_{\rm LS}||^2}{||\hat{\boldsymbol{x}}_{\rm LS}||^2 + \epsilon_0}\right)\hat{\boldsymbol{x}}_{\rm LS} + \left(\frac{\epsilon_0}{||\hat{\boldsymbol{x}}_{\rm LS}||^2 + \epsilon_0}\right)\boldsymbol{x}_0. \tag{8}$$

All dominance results continue to hold for the off-center techniques as well. In the sequel, we assume $\boldsymbol{x}_0 = \mathbf{0}$ merely for the sake of notational simplicity.

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

Theorem 1: Suppose $\epsilon_0/\epsilon_{\max} > 4$, where ϵ_0 is given by (3), $\epsilon_{\max}$ is the largest eigenvalue of $\boldsymbol{Q}^{-1}$, and $\boldsymbol{Q} = \boldsymbol{H}^*\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H}$. Then, the SBME (7) strictly dominates the LS estimator.

The value $\epsilon_0/\epsilon_{\max}$ is known as the effective dimension [16], and may be roughly described as the number of independently measured parameters in the system. In the i.i.d. case, for example, the effective dimension simply equals the length of the vector $\boldsymbol{x}$. Thus, the condition of Theorem 1 can be roughly stated as a requirement for a sufficient number of independent parameters. This requirement is a result of the fact that the LS estimator is admissible when up to two parameters are estimated [6]. However, since many estimation problems contain dozens or hundreds of parameters and measurements, the requirement on the effective dimension holds for a variety of applications.

Note that the SBME is a special case of the estimator

$$\hat{\boldsymbol{x}}_c = \left(1 - \frac{\epsilon_0}{c + ||\hat{\boldsymbol{x}}_{\rm LS}||^2}\right)\hat{\boldsymbol{x}}_{\rm LS} \tag{9}$$

in which $c = \epsilon_0$. Thus, rather than proving Theorem 1, we prove the following, more general proposition, which will also be used in Section V.

Proposition 1: Under the conditions of Theorem 1, the estimator $\hat{\boldsymbol{x}}_c$ given by (9) strictly dominates the LS estimator, for any $c \geq 0$.

The proof of Proposition 1 makes use of the following lemma, which is due to Stein [5, Theorem 1.5.15].

Lemma 1 (Stein): Let $\hat{v} \sim \mathcal{N}_p(\boldsymbol{v}, \boldsymbol{I})$, and let $g(\hat{\boldsymbol{v}})$ be a differentiable function such that $E\{|\frac{\partial g(\hat{\boldsymbol{v}})}{\partial \hat{v}_i}|\} < \infty$ for all i. Then

$$E\left\{\frac{\partial g(\hat{\boldsymbol{v}})}{\partial \hat{v}_i}\right\} = -E\{g(\hat{\boldsymbol{v}})(v_i - \hat{v}_i)\}. \tag{10}$$

Proof of Proposition 1: To prove the proposition, first note that the MSE $R(\hat{\boldsymbol{x}}_c) = E\{||\boldsymbol{x} - \hat{\boldsymbol{x}}_c||^2\}$ of $\hat{\boldsymbol{x}}_c$ is given by

$$\begin{aligned} R(\hat{\boldsymbol{x}}_c) = \epsilon_0 &+ E\left\{\frac{\epsilon_0^2||\hat{\boldsymbol{x}}_{\rm LS}||^2}{(c + ||\hat{\boldsymbol{x}}_{\rm LS}||^2)^2}\right\} \\ &+ 2E\left\{\frac{\epsilon_0}{c + ||\hat{\boldsymbol{x}}_{\rm LS}||^2}\hat{\boldsymbol{x}}_{\rm LS}^*(\boldsymbol{x} - \hat{\boldsymbol{x}}_{\rm LS})\right\}. \end{aligned} \tag{11}$$

Let $\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^*$ be the eigenvalue decomposition of $\boldsymbol{Q}$, such that $\boldsymbol{V}$ is unitary and $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_m)$. Define $\hat{\boldsymbol{v}} = \boldsymbol{V}^*\boldsymbol{Q}^{1/2}\hat{\boldsymbol{x}}_{\rm LS}$ and $\boldsymbol{v} = \boldsymbol{V}^*\boldsymbol{Q}^{1/2}\boldsymbol{x}$. With these definitions, we have

$$\begin{aligned} \hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{v} &= \hat{\boldsymbol{x}}_{\rm LS}^*\boldsymbol{x} \\ \hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{v}} &= ||\hat{\boldsymbol{x}}_{\rm LS}||^2 \\ \hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{-2}\hat{\boldsymbol{v}} &= ||\hat{\boldsymbol{x}}_{\rm LS}||^2_{\boldsymbol{Q}^{-1}}. \end{aligned} \tag{12}$$

Using these properties, the third term in (11) becomes

$$\begin{aligned} &E\left\{\frac{\epsilon_0}{c + ||\hat{\boldsymbol{x}}_{\rm LS}||^2}\hat{\boldsymbol{x}}_{\rm LS}^*(\boldsymbol{x} - \hat{\boldsymbol{x}}_{\rm LS})\right\} \\ &\quad = E\left\{\frac{\epsilon_0}{c + \hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{v}}}\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{-1}(\boldsymbol{v} - \hat{\boldsymbol{v}})\right\} \\ &\quad = \epsilon_0\sum_{i=1}^{p}\sigma_i^{-1}E\left\{\frac{\hat{v}_i(v_i - \hat{v}_i)}{c + \hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{v}}}\right\}. \end{aligned} \tag{13}$$

To evaluate (13), let

$$g_i(\hat{\boldsymbol{v}}) \triangleq \frac{\hat{v}_i}{c + \hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{v}}} \tag{14}$$

and note that $\hat{\boldsymbol{v}}$ is distributed normally with mean $\boldsymbol{v}$ and covariance $\boldsymbol{I}$. We can thus apply Lemma 1 to obtain

$$\begin{aligned} &E\left\{\frac{\epsilon_0}{c+\|\hat{\boldsymbol{x}}_{\mathrm{LS}}\|^2}\hat{\boldsymbol{x}}_{\mathrm{LS}}^*(\boldsymbol{x}-\hat{\boldsymbol{x}}_{\mathrm{LS}})\right\} \\ &\quad=-\epsilon_0\sum_i \sigma_i^{-1}E\left\{\frac{1}{c+\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{v}}}-2\frac{\sigma_i^{-1}\hat{v}_i^2}{(c+\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{v}})^2}\right\} \\ &\quad=-\epsilon_0 E\left\{\frac{\mathrm{Tr}(\boldsymbol{\Sigma}^{-1})}{c+\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{v}}}\right\}+2\epsilon_0 E\left\{\frac{\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{-2}\hat{\boldsymbol{v}}}{(c+\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{v}})^2}\right\} \\ &\quad=-\epsilon_0 E\left\{\frac{\mathrm{Tr}(\boldsymbol{Q}^{-1})}{c+\|\hat{\boldsymbol{x}}_{\mathrm{LS}}\|^2}\right\}+2\epsilon_0 E\left\{\frac{\|\hat{\boldsymbol{x}}_{\mathrm{LS}}\|^2_{\boldsymbol{Q}^{-1}}}{(c+\|\hat{\boldsymbol{x}}_{\mathrm{LS}}\|^2)^2}\right\}. \end{aligned} \tag{15}$$

Substituting this result back into (11), we have

$$\begin{aligned} R(\hat{\boldsymbol{x}}_c)=\epsilon_0+E\Bigg\{&\frac{\epsilon_0}{c+\|\hat{\boldsymbol{x}}_{\mathrm{LS}}\|^2} \\ &\cdot\left(\epsilon_0\frac{\|\hat{\boldsymbol{x}}_{\mathrm{LS}}\|^2}{c+\|\hat{\boldsymbol{x}}_{\mathrm{LS}}\|^2}-2\epsilon_0+4\frac{\|\hat{\boldsymbol{x}}_{\mathrm{LS}}\|^2_{\boldsymbol{Q}^{-1}}}{c+\|\hat{\boldsymbol{x}}_{\mathrm{LS}}\|^2}\right)\Bigg\}. \end{aligned} \tag{16}$$

Since $c \geq 0$

$$R(\hat{\boldsymbol{x}}_c)\leq \epsilon_0+E\left\{\frac{\epsilon_0}{c+\|\hat{\boldsymbol{x}}_{\mathrm{LS}}\|^2}(-\epsilon_0+4\epsilon_{\max})\right\}. \tag{17}$$

If $\epsilon_0 > 4\epsilon_{\max}$, then the expectation is taken over a strictly negative range, and hence, $R(\hat{\boldsymbol{x}}_c)$ is always lower than ϵ_0, so that $\hat{\boldsymbol{x}}_c$ strictly dominates $\hat{\boldsymbol{x}}_{\mathrm{LS}}$. □

As we have shown, in terms of MSE, the SBME outperforms LS, 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 estimator multiplied by a gain factor smaller than one. The SBME thus illustrates the fact that the LS technique tends to be an overestimate, and shrinkage can improve its performance.

IV. The Ellipsoidal Blind Minimax Estimator

A. Motivation

Not all elements of the LS estimate are equally trustworthy. Rather, $\hat{\boldsymbol{x}}_{\mathrm{LS}}$ is a Gaussian random vector with mean $\boldsymbol{x}$ and covariance $\boldsymbol{Q}^{-1}=(\boldsymbol{H}^*\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H})^{-1}$. Thus, some components of $\hat{\boldsymbol{x}}_{\mathrm{LS}}$ have lower variance than others. In this sense, the scalar shrinkage factor of the SBME (7) and other extended Stein estimators [13] seems inadequate.

Indeed, several researchers have proposed shrinking each measurement according to its variance. Efron and Morris [14] propose an empirical Bayes technique, in which high-variance components are shrunk more than low-variance ones. However, no closed form is available for this estimator, and 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 approach 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, dominance over the LS method is guaranteed, but the MSE gain is insubstantial unless all noise components have similar variances.

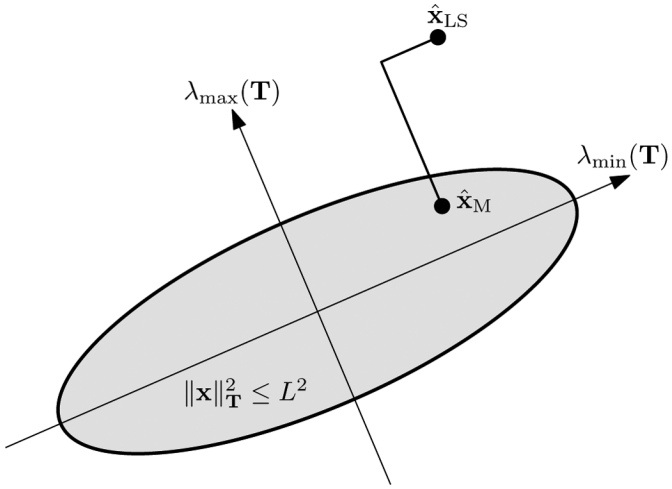

Fig. 1. Illustration of the adaptive shrinkage of the minimax estimator $\hat{\boldsymbol{x}}_{\mathrm{M}}$ for the parameter set $\boldsymbol{x}^*\boldsymbol{T}\boldsymbol{x} \leq L^2$. Low shrinkage is applied to components of $\hat{\boldsymbol{x}}_{\mathrm{LS}}$ corresponding to small eigenvalues of $\boldsymbol{T}$, while components in directions of large eigenvalues obtain higher shrinkage.

Minimax estimators can easily be adapted for nonscalar shrinkage. Specifically, consider an ellipsoidal parameter set of the form $\mathcal{S}=\{\boldsymbol{x}:\|\boldsymbol{x}\|^2_{\boldsymbol{T}}\leq L^2\}$, for some positive definite matrix $\boldsymbol{T}$ (see Fig. 1). Let $\hat{\boldsymbol{x}}_{\mathrm{M}}$ represent the linear minimax estimator for this set. It can be shown that $\hat{\boldsymbol{x}}_{\mathrm{M}}$ is a linear function of $\hat{\boldsymbol{x}}_{\mathrm{LS}}$, and one can therefore examine its effect on each component of $\hat{\boldsymbol{x}}_{\mathrm{LS}}$. Consider first components of $\hat{\boldsymbol{x}}_{\mathrm{LS}}$ in the direction of narrow axes of the ellipsoid $\mathcal{S}$. These components correspond to large eigenvalues of $\boldsymbol{T}$, and are denoted $\lambda_{\max}(\boldsymbol{T})$ in Fig. 1. 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 $\mathcal{S}$ (small eigenvalues of $\boldsymbol{T}$) 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 Fig. 1, the shrinkage of wide-axis and narrow-axis components is illustrated schematically for a particular value of $\hat{\boldsymbol{x}}_{\mathrm{LS}}$.

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

$$\mathcal{S}=\left\{\boldsymbol{x}:\|\boldsymbol{x}\|^2_{\boldsymbol{Q}^b}\leq L^2\right\} \tag{18}$$

for some constant $b<0$. The bound L^2 is estimated as $L^2=\|\hat{\boldsymbol{x}}_{\mathrm{LS}}\|^2_{\boldsymbol{Q}^b}$. 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

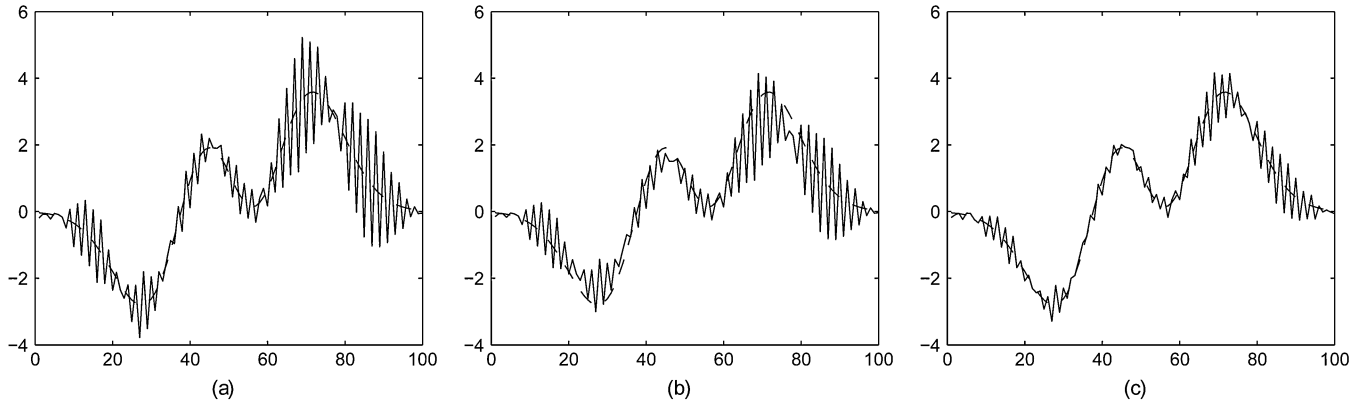

Fig. 2. Estimation of a signal from measurements of its DCT. In this example, high-frequency components have a much higher noise variance than low-frequency components. Dashed line indicates original signal; solid line indicates estimate. (a) LS estimate; (b) spherical BME, resulting in a shrinkage factor of 0.79; (c) ellipsoidal BME, with shrinkage in the range 0.44–0.98.

a tradeoff between selective shrinkage and a broad dominance condition. 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 application example (Fig. 2). 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 ten highest frequency components, while the remaining components contain much lower noise levels. Thus, $\boldsymbol{C_w}$ is diagonal, and $\boldsymbol{H}$ is the DCT matrix. The condition number of $\boldsymbol{H}^*\boldsymbol{C_w}^{-1}\boldsymbol{H}$ is 1000.

Since $\boldsymbol{C_w}$ is diagonal, 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 Fig. 2(a). The error is reduced by the SBME (Fig. 2(b)), which multiplies the LS estimate by an appropriately chosen scalar; in the example above, the squared error was reduced by 20% compared with that of the LS estimate. Hence, 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 (Fig. 2(c)), which shrinks the high-noise coefficients. Specifically, in this example, 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 the LS estimate.

Thus, our preliminary example demonstrates that it is possible to achieve substantial improvements over the LS technique by using nonscalar shrinkage. As we will demonstrate presently, this empirical finding is only an example of the wide range of cases in which the EBME is guaranteed to improve on the LS approach.

B. Dominance

We begin our analysis by obtaining an expression for the EBMEs. A closed-form solution for minimax estimators of an ellipsoidal parameter set was developed in [22]. By substituting the value of L^2 into this closed form, we obtain the following result.

Proposition 2 (Closed-Form EBME): Let $\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^*$ be the eigenvalue decomposition of $\boldsymbol{Q} = \boldsymbol{H}^*\boldsymbol{C_w}^{-1}\boldsymbol{H}$, where $\boldsymbol{V}$ is orthonormal and $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_m)$. Let $b \in \mathbb{R}$ be any constant, and suppose the eigenvalues $\boldsymbol{\Sigma}$ 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 $\mathcal{S} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_{\boldsymbol{Q}^b}^2 \leq L^2\}$ with $L^2 = \|\hat{\boldsymbol{x}}_{\mathrm{LS}}\|_{\boldsymbol{Q}^b}^2$ is given by

$$\hat{\boldsymbol{x}}_{\mathrm{EBM}} = \boldsymbol{V}\operatorname{diag}\left(\left(1-\alpha\sigma_1^{b/2}\right)_+, \ldots, \left(1-\alpha\sigma_m^{b/2}\right)_+\right)\boldsymbol{V}^*\hat{\boldsymbol{x}}_{\mathrm{LS}} \quad (19)$$

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

$$\begin{aligned}(\cdot)_+ &= \max(\cdot, 0)\\ \alpha &= \frac{r_1}{\|\hat{\boldsymbol{x}}_{\mathrm{LS}}\|_{\boldsymbol{Q}^b}^2 + r_2}\\ r_1 &= \sum_{i=k+1}^{m}\sigma_i^{b/2-1}\\ r_2 &= \sum_{i=k+1}^{m}\sigma_i^{b-1}\end{aligned} \quad (20)$$

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

$$\alpha\sigma_{k+1}^{b/2} < 1. \quad (21)$$

Proof: In the case $\hat{\boldsymbol{x}}_{\mathrm{LS}} = \boldsymbol{0}$, we need to find the linear minimax estimator for the set $\mathcal{S} = \{\boldsymbol{0}\}$. Clearly, the solution in this case is $\hat{\boldsymbol{x}} = \boldsymbol{0}$. For all other values of $\hat{\boldsymbol{x}}_{\mathrm{LS}}$, we seek the linear minimax estimator for the set $\mathcal{S} = \{\boldsymbol{x} : \boldsymbol{x}^*\boldsymbol{Q}^b\boldsymbol{x} \leq L^2\}$, where $L^2 = \hat{\boldsymbol{x}}_{\mathrm{LS}}^*\boldsymbol{Q}^b\hat{\boldsymbol{x}}_{\mathrm{LS}} > 0$. Substituting this value of L^2 into Proposition 1 of [22] yields

$$\begin{aligned}\hat{\boldsymbol{x}}_{\mathrm{EBM}} &= \boldsymbol{V}\operatorname{diag}\left(\underbrace{0,\ldots,0}_{k},\underbrace{1,\ldots,1}_{m-k}\right)\boldsymbol{V}^*\left(\boldsymbol{I}-\alpha\boldsymbol{Q}^{b/2}\right)\hat{\boldsymbol{x}}_{\mathrm{LS}}\\ &= \boldsymbol{V}\operatorname{diag}\left(\underbrace{0,\ldots,0}_{k},1-\alpha\sigma_{k+1}^{b/2},\ldots,1-\alpha\sigma_m^{b/2}\right)\boldsymbol{V}^*\hat{\boldsymbol{x}}_{\mathrm{LS}}.\end{aligned} \quad (22)$$

From (21), it follows that $1 - \alpha\sigma_i^{b/2} < 0$ for all $i \leq k$, and therefore (22) can be written as (19). □

We note that, as long as $\|\hat{\boldsymbol{x}}_{\rm LS}\|^2_{\boldsymbol{Q}^b} > 0$, it is always possible to find a value k which satisfies (21). In particular, for $k = m-1$, we have

$$\alpha = \frac{\sigma_m^{b/2-1}}{\|\hat{\boldsymbol{x}}_{\rm LS}\|^2_{\boldsymbol{Q}^b} + \sigma_m^{b-1}} < \frac{\sigma_m^{b/2-1}}{\sigma_m^{b-1}} \tag{23}$$

which satisfies the requirement (21).

While the closed form of the EBME appears somewhat more intimidating than that of the SBME, the computational complexities of the two estimators are comparable. The major difference is the calculation of the value k, for which m divisions are required. Like the SBME, the EBME also dominates the LS estimator under suitable conditions, as shown in the following theorem.

Theorem 2: Let $\hat{\boldsymbol{x}}_{\rm EBM}$ be the EBME (19) and suppose that

$$\mathrm{Tr}(\boldsymbol{Q}^{b/2-1}) > 4\lambda_{\max}(\boldsymbol{Q}^{b/2-1}) \tag{24}$$

where $\lambda_{\max}(\boldsymbol{Q}^{b/2-1})$ is the largest eigenvalue of $\boldsymbol{Q}^{b/2-1}$ and $\boldsymbol{Q} = \boldsymbol{H}^*\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H}$. Then, $\hat{\boldsymbol{x}}_{\rm EBM}$ strictly dominates the LS estimator.

Note that by substituting $b = 0$, this result can be used to demonstrate the dominance of the SBME over LS estimation (Theorem 1). However, the method of proof here is different, and the proof of Theorem 1 will also be used in Section V.

Also note that the dominance condition (24) is satisfied by many reasonable estimation 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 other parameters; in this case, the LS method is admissible or nearly so.

In order to prove Theorem 2, we observe that the form (19) of the EBME is similar to Baranchik's positive-part modification [5], [24] of the James–Stein estimator. Baranchik proposed using a shrinkage factor of 0 whenever the James–Stein technique contains negative shrinkage, and showed that the resulting method dominates the James–Stein estimator. Although the EBME is not a shrinkage technique, it resembles Baranchik's modification, since each negative diagonal component in (19) is replaced with zero. The following proposition shows that the MSE can be reduced by eliminating this negative shrinkage.

Proposition 3: Let $\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^*$ be the eigenvalue decomposition of $\boldsymbol{Q} = \boldsymbol{H}^*\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H}$, and let $b \in \mathbb{R}$ be a constant. Suppose $\hat{\boldsymbol{x}}$ is an estimator of the form $\hat{\boldsymbol{x}} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^*\hat{\boldsymbol{x}}_{\rm LS}$, where $\boldsymbol{D}$ is a diagonal matrix, whose diagonal elements d_i are functions of the random variable $\|\hat{\boldsymbol{x}}_{\rm LS}\|^2_{\boldsymbol{Q}^b}$. Suppose at least one of the elements d_i is negative with nonzero probability. Then, $\hat{\boldsymbol{x}}$ is dominated by the (generalized) positive-part estimator

$$\hat{\boldsymbol{x}}_+ = \boldsymbol{V}\boldsymbol{D}_+\boldsymbol{V}^*\hat{\boldsymbol{x}}_{\rm LS} \tag{25}$$

where $\boldsymbol{D}_+$ is a diagonal matrix with diagonal elements $d_{i+} = \max(0, d_i)$.

Proof: Our proof follows that of Baranchik [24]. We will show that $\mathrm{MSE}\,(\hat{\boldsymbol{x}}) - \mathrm{MSE}\,(\hat{\boldsymbol{x}}_+)$ is nonnegative for all $\boldsymbol{x}$, and positive for any value of $\boldsymbol{x}$ whose elements are all nonzero.

$$\begin{aligned}\mathrm{MSE}\,(\hat{\boldsymbol{x}}) - \mathrm{MSE}\,(\hat{\boldsymbol{x}}_+) &= E\{\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|^2\} - E\{\|\hat{\boldsymbol{x}}_+-\boldsymbol{x}\|^2\} \\ &= E\{\|\hat{\boldsymbol{x}}\|^2 - \|\hat{\boldsymbol{x}}_+\|^2\} - 2E\{\hat{\boldsymbol{x}}^*\boldsymbol{x} - \hat{\boldsymbol{x}}_+^*\boldsymbol{x}\} \\ &= E\{\hat{\boldsymbol{x}}_{\rm LS}^*\boldsymbol{V}(\boldsymbol{D}^2 - \boldsymbol{D}_+^2)\boldsymbol{V}^*\hat{\boldsymbol{x}}_{\rm LS}\} \\ &\quad - 2E\{\hat{\boldsymbol{x}}_{\rm LS}^*\boldsymbol{V}(\boldsymbol{D} - \boldsymbol{D}_+)\boldsymbol{V}^*\boldsymbol{x}\}. \end{aligned} \tag{26}$$

Since $d_i^2 - d_{i+}^2 \geq 0$ for all i, the first term in (26) is nonnegative. Hence, to prove the proposition, it suffices to show that $E\{\hat{\boldsymbol{x}}_{\rm LS}^*\boldsymbol{V}(\boldsymbol{D} - \boldsymbol{D}_+)\boldsymbol{V}^*\boldsymbol{x}\}$ is nonpositive for all $\boldsymbol{x}$, and negative for values $\boldsymbol{x}$ with nonzero elements.

To this end, define $\boldsymbol{z} = \boldsymbol{V}^*\boldsymbol{x}$ and $\hat{\boldsymbol{z}} = \boldsymbol{V}^*\hat{\boldsymbol{x}}_{\rm LS}$. We note that $\hat{\boldsymbol{z}} \sim \mathcal{N}_m(\boldsymbol{z}, \boldsymbol{\Sigma}^{-1})$, so that the elements of $\hat{\boldsymbol{z}}$ are statistically independent. To calculate $E\{\hat{\boldsymbol{x}}_{\rm LS}^*\boldsymbol{V}(\boldsymbol{D}-\boldsymbol{D}_+)\boldsymbol{V}^*\boldsymbol{x}\}$, we condition on $\|\hat{\boldsymbol{x}}_{\rm LS}\|^2_{\boldsymbol{Q}^b}$, obtaining

$$\begin{aligned} &E\{\hat{\boldsymbol{x}}_{\rm LS}^*\boldsymbol{V}(\boldsymbol{D} - \boldsymbol{D}_+)\boldsymbol{V}^*\boldsymbol{x}\} \\ &\quad = E\{E\{\hat{\boldsymbol{z}}^*(\boldsymbol{D} - \boldsymbol{D}_+)\boldsymbol{z} \,|\, \hat{\boldsymbol{z}}^*\boldsymbol{\Sigma}^b\hat{\boldsymbol{z}}\}\} \\ &\quad = E\left\{\sum_{i=1}^m (d_i - d_{i+})E\{\hat{z}_i z_i \,|\, \hat{\boldsymbol{z}}^*\boldsymbol{\Sigma}^b\hat{\boldsymbol{z}}\}\right\} \end{aligned} \tag{27}$$

where we used the fact that $\|\hat{\boldsymbol{x}}_{\rm LS}\|^2_{\boldsymbol{Q}^b} = \hat{\boldsymbol{z}}^*\boldsymbol{\Sigma}^b\hat{\boldsymbol{z}}$, and that d_i and d_{i+} are deterministic when conditioned on $\|\hat{\boldsymbol{x}}_{\rm LS}\|^2_{\boldsymbol{Q}^b}$. For each i, we further condition on $|\hat{z}_i|$, to obtain

$$\begin{aligned} &E\{\hat{\boldsymbol{x}}_{\rm LS}^*\boldsymbol{V}(\boldsymbol{D} - \boldsymbol{D}_+)\boldsymbol{V}^*\boldsymbol{x}\} \\ &\quad = E\left\{\sum_{i=1}^m (d_i - d_{i+})E\{\hat{z}_i z_i \,\Big|\, \hat{\boldsymbol{z}}^*\boldsymbol{\Sigma}^b\hat{\boldsymbol{z}}, |\hat{z}_i|\}\right\} \\ &\quad = E\left\{\sum_{i=1}^m (d_i - d_{i+})|\hat{z}_i z_i|E\{\mathrm{sgn}(\hat{z}_i z_i) \,\Big|\, \hat{\boldsymbol{z}}^*\boldsymbol{\Sigma}^b\hat{\boldsymbol{z}}, |\hat{z}_i|\}\right\}. \end{aligned} \tag{28}$$

Given $|\hat{z}_i|$, we have that either $\hat{z}_i = |\hat{z}_i|\mathrm{sgn}(z_i)$ or that $\hat{z}_i = -|\hat{z}_i|\mathrm{sgn}(z_i)$. It is evident from the probability density function (pdf) of $\hat{z}_i$ that the latter option has lower probability, i.e.,

$$\begin{aligned} &\Pr\{\mathrm{sgn}(\hat{z}_i) = \mathrm{sgn}(z_i) \,|\, \hat{\boldsymbol{z}}^*\boldsymbol{\Sigma}^b\hat{\boldsymbol{z}}, \Big|\hat{z}_i|\} \\ &\qquad > \Pr\{\mathrm{sgn}(\hat{z}_i) \neq \mathrm{sgn}(z_i) \,\Big|\, \hat{\boldsymbol{z}}^*\boldsymbol{\Sigma}^b\hat{\boldsymbol{z}}, |\hat{z}_i|\}. \end{aligned} \tag{29}$$

It follows that $E\{\mathrm{sgn}(\hat{z}_i z_i) \,\big|\, \hat{\boldsymbol{z}}^*\boldsymbol{\Sigma}^b\hat{\boldsymbol{z}}, |\hat{z}_i|\} \geq 0$, with strict inequality for $z_i \neq 0$. Therefore, all terms in (28) are nonnegative, except for $(d_i - d_{i+})$, which is nonpositive. As a result, (28) (and hence (26)) is nonpositive for all $\boldsymbol{x}$, so that the MSE of $\hat{\boldsymbol{x}}_+$ is never higher than that of $\hat{\boldsymbol{x}}$.

We must also show that, for some $\boldsymbol{x}$, (28) is strictly negative. To this end, we choose $\boldsymbol{x}$ for which all elements are nonzero; as a result, all terms in (28) are strictly positive with probability 1, except for $(d_i - d_{i+})$. The latter term is negative when $d_i < 0$ and zero otherwise. Since d_i is negative with nonzero probability for at least one value of i, we conclude that for the chosen value of $\boldsymbol{x}$, (28) is strictly negative, completing the proof of Proposition 3. □

This generalization of the concept of a positive part estimator is now used to prove Theorem 2.

Proof of Theorem 2: Clearly, the EBME (19) is the positive part of the estimator

$$\begin{aligned} \hat{\boldsymbol{x}}_0 &= \boldsymbol{V}\mathrm{diag}\left(1 - \alpha\sigma_1^{b/2}, \ldots, 1 - \alpha\sigma_m^{b/2}\right)\boldsymbol{V}^*\hat{\boldsymbol{x}}_{\rm LS} \\ &= (\boldsymbol{I} - \alpha\boldsymbol{Q}^{b/2})\hat{\boldsymbol{x}}_{\rm LS}. \end{aligned} \tag{30}$$

Therefore, it suffices to show that $\hat{\boldsymbol{x}}_0$ dominates the LS estimator, and the theorem follows using Proposition 3.

The MSE of $\hat{\boldsymbol{x}}_0$ is given by

$$\begin{aligned}
&E\{||\boldsymbol{x}-\hat{\boldsymbol{x}}_{\rm LS}+\alpha\boldsymbol{Q}^{b/2}\hat{\boldsymbol{x}}_{\rm LS}||^2\}\\
&\quad=E\left\{\left\|\boldsymbol{x}-\hat{\boldsymbol{x}}_{\rm LS}+\frac{r_1\boldsymbol{Q}^{b/2}\hat{\boldsymbol{x}}_{\rm LS}}{||\hat{\boldsymbol{x}}_{\rm LS}||^2_{\boldsymbol{Q}^b}+r_2}\right\|^2\right\}\\
&\quad=\epsilon_0+E\left\{\frac{r_1^2||\hat{\boldsymbol{x}}_{\rm LS}||^2_{\boldsymbol{Q}^b}}{\left(||\hat{\boldsymbol{x}}_{\rm LS}||^2_{\boldsymbol{Q}^b}+r_2\right)^2}\right\}\\
&\qquad+2E\left\{\frac{r_1(\boldsymbol{x}-\hat{\boldsymbol{x}}_{\rm LS})^*\boldsymbol{Q}^{b/2}\hat{\boldsymbol{x}}_{\rm LS}}{||\hat{\boldsymbol{x}}_{\rm LS}||^2_{\boldsymbol{Q}^b}+r_2}\right\}. \qquad (31)
\end{aligned}$$

To analyze this expression, we define

$$\begin{aligned}
\boldsymbol{v}&\triangleq\boldsymbol{V}^*\boldsymbol{Q}^{1/2}\boldsymbol{x}\\
\hat{\boldsymbol{v}}&\triangleq\boldsymbol{V}^*\boldsymbol{Q}^{1/2}\hat{\boldsymbol{x}}_{\rm LS}. \qquad (32)
\end{aligned}$$

Using this notation, the third term in (31) becomes

$$\begin{aligned}
A_3&\triangleq E\left\{\frac{r_1(\boldsymbol{x}-\hat{\boldsymbol{x}}_{\rm LS})^*\boldsymbol{Q}^{b/2}\hat{\boldsymbol{x}}_{\rm LS}}{||\hat{\boldsymbol{x}}_{\rm LS}||^2_{\boldsymbol{Q}^b}+r_2}\right\}\\
&=E\left\{\frac{r_1(\boldsymbol{x}-\hat{\boldsymbol{x}}_{\rm LS})^*\boldsymbol{Q}^{1/2}\boldsymbol{V}\boldsymbol{V}^*\boldsymbol{Q}^{b/2-1}\boldsymbol{V}\boldsymbol{V}^*\boldsymbol{Q}^{1/2}\hat{\boldsymbol{x}}_{\rm LS}}{\hat{\boldsymbol{x}}^*_{\rm LS}\boldsymbol{Q}^{1/2}\boldsymbol{V}\boldsymbol{V}^*\boldsymbol{Q}^{b-1}\boldsymbol{V}\boldsymbol{V}^*\boldsymbol{Q}^{1/2}\hat{\boldsymbol{x}}_{\rm LS}+r_2}\right\}\\
&=E\left\{\frac{r_1(\boldsymbol{v}-\hat{\boldsymbol{v}})^*\boldsymbol{\Sigma}^{b/2-1}\hat{\boldsymbol{v}}}{\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{b-1}\hat{\boldsymbol{v}}+r_2}\right\}\\
&=\sum_{i=1}^m\sigma_i^{b/2-1}E\left\{\frac{r_1(v_i-\hat{v}_i)\hat{v}_i}{\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{b-1}\hat{\boldsymbol{v}}+r_2}\right\}. \qquad (33)
\end{aligned}$$

Next, define

$$g_i(\hat{\boldsymbol{v}})\triangleq\frac{r_1\hat{v}_i}{\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{b-1}\hat{\boldsymbol{v}}+r_2}. \qquad (34)$$

Note that r_1 and r_2 are implicitly dependent on k, which in turn depends on $\hat{\boldsymbol{v}}$. Thus, $g_i(\hat{\boldsymbol{v}})$ is discontinuous for some values of $\hat{\boldsymbol{v}}$, namely, those values for which $\alpha=\sigma_i^{b/2}$. However, these values of $\hat{\boldsymbol{v}}$ occur with probability zero; for all other values, k (and hence r_1 and r_2) are constant for sufficiently small changes in $\hat{\boldsymbol{v}}$. Thus

$$\frac{\partial g_i}{\partial\hat{v}_i}=\frac{r_1}{\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{b-1}\hat{\boldsymbol{v}}+r_2}-\frac{2r_1\sigma_i^{b-1}\hat{v}_i^2}{(\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{b-1}\hat{\boldsymbol{v}}+r_2)^2}\quad\text{w.p. 1} \qquad (35)$$

and $E\{|\partial g_i/\partial\hat{v}_j|\}<\infty$ for all i,j. Furthermore, observe that $\hat{\boldsymbol{v}}\sim\mathcal{N}_m(\boldsymbol{v},\boldsymbol{I})$. We can therefore apply Lemma 1 to g_i. This yields

$$\begin{aligned}
&E\left\{\frac{r_1\hat{v}_i(v_i-\hat{v}_i)}{\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{b-1}\hat{\boldsymbol{v}}+r_2}\right\}\\
&\quad=E\left\{-\frac{r_1}{\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{b-1}\hat{\boldsymbol{v}}+r_2}+2\frac{r_1\sigma_i^{b-1}\hat{v}_i^2}{(\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{b-1}\hat{\boldsymbol{v}}+r_2)^2}\right\}. \qquad (36)
\end{aligned}$$

Substituting into (33), we obtain

$$\begin{aligned}
A_3&=\sum_{i=1}^m\sigma_i^{b/2-1}E\left\{\frac{2r_1\sigma_i^{b-1}\hat{v}_i^2}{(\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{b-1}\hat{\boldsymbol{v}}+r_2)^2}-\frac{r_1}{\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{b-1}\hat{\boldsymbol{v}}+r_2}\right\}\\
&=E\left\{\frac{2r_1\sum_{i=1}^m\sigma_i^{3b/2-2}\hat{v}_i^2}{(\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{b-1}\hat{\boldsymbol{v}}+r_2)^2}-\frac{r_1\sum_{i=1}^m\sigma_i^{b/2-1}}{\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{b-1}\hat{\boldsymbol{v}}+r_2}\right\}\\
&=E\left\{\frac{2r_1\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{3b/2-2}\hat{\boldsymbol{v}}}{(\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{b-1}\hat{\boldsymbol{v}}+r_2)^2}-\frac{r_1\mathrm{Tr}(\boldsymbol{\Sigma}^{b/2-1})}{\hat{\boldsymbol{v}}^*\boldsymbol{\Sigma}^{b-1}\hat{\boldsymbol{v}}+r_2}\right\}. \qquad (37)
\end{aligned}$$

Using the definition (32) of $\hat{\boldsymbol{v}}$, A_3 may be written as

$$E\left\{\frac{r_1}{||\hat{\boldsymbol{x}}_{\rm LS}||^2_{\boldsymbol{Q}^b}+r_2}\left[\frac{2||\hat{\boldsymbol{x}}_{\rm LS}||^2_{\boldsymbol{Q}^{3b/2-1}}}{||\hat{\boldsymbol{x}}_{\rm LS}||^2_{\boldsymbol{Q}^b}+r_2}-\mathrm{Tr}(\boldsymbol{Q}^{b/2-1})\right]\right\}. \qquad (38)$$

Note that

$$\begin{aligned}
\frac{||\hat{\boldsymbol{x}}_{\rm LS}||^2_{\boldsymbol{Q}^{3b/2-1}}}{||\hat{\boldsymbol{x}}_{\rm LS}||^2_{\boldsymbol{Q}^b}+r_2}&<\frac{||\hat{\boldsymbol{x}}_{\rm LS}||^2_{\boldsymbol{Q}^{3b/2-1}}}{||\hat{\boldsymbol{x}}_{\rm LS}||^2_{\boldsymbol{Q}^b}}\\
&=\frac{(\boldsymbol{Q}^{b/2}\hat{\boldsymbol{x}}_{\rm LS})^*\boldsymbol{Q}^{b/2-1}(\boldsymbol{Q}^{b/2}\hat{\boldsymbol{x}}_{\rm LS})}{(\boldsymbol{Q}^{b/2}\hat{\boldsymbol{x}}_{\rm LS})^*(\boldsymbol{Q}^{b/2}\hat{\boldsymbol{x}}_{\rm LS})}\\
&\le\lambda_{\max}(\boldsymbol{Q}^{b/2-1}). \qquad (39)
\end{aligned}$$

Thus

$$A_3<E\left\{\frac{r_1}{||\hat{\boldsymbol{x}}_{\rm LS}||^2_{\boldsymbol{Q}^b}+r_2}\left[2\lambda_{\max}(\boldsymbol{Q}^{b/2-1})-\mathrm{Tr}(\boldsymbol{Q}^{b/2-1})\right]\right\}. \qquad (40)$$

Substituting back into (31), we have

$$\begin{aligned}
\mathrm{MSE}<\epsilon_0+E\Bigg\{&\frac{r_1}{||\hat{\boldsymbol{x}}_{\rm LS}||^2_{\boldsymbol{Q}^b}+r_2}\\
&\cdot\left[r_1+4\lambda_{\max}(\boldsymbol{Q}^{b/2-1})-2\mathrm{Tr}(\boldsymbol{Q}^{b/2-1})\right]\Bigg\} \qquad (41)
\end{aligned}$$

and using the fact that $r_1\le\mathrm{Tr}(\boldsymbol{\Sigma}^{b/2-1})=\mathrm{Tr}(\boldsymbol{Q}^{b/2-1})$, we conclude that the MSE is bounded by

$$\epsilon_0+E\left\{\frac{r_1}{||\hat{\boldsymbol{x}}_{\rm LS}||^2_{\boldsymbol{Q}^b}+r_2}\left[4\lambda_{\max}(\boldsymbol{Q}^{b/2-1})-\mathrm{Tr}(\boldsymbol{Q}^{b/2-1})\right]\right\}. \qquad (42)$$

Thus, if $\mathrm{Tr}(\boldsymbol{Q}^{b/2-1})>4\lambda_{\max}(\boldsymbol{Q}^{b/2-1})$, then $\mathrm{MSE}<\epsilon_0$, proving that the EBME dominates the LS estimator. □

Thus far, we have presented two examples of BMEs which dominate the LS method under suitable conditions. Both approaches are extensions of Thompson's technique to the non-i.i.d. case. In the next section, we demonstrate that other BMEs extend different LS-dominating techniques, namely Stein's estimator and Baranchik's positive-part improvement.

V. Relation to Stein-Type Estimation

In Section III, the SBME (7) was constructed by using $L^2=||\hat{\boldsymbol{x}}_{\rm LS}||^2$ as an estimate of $||\boldsymbol{x}||^2$. However, the fact that shrinkage techniques such as the SBME dominate LS indicates that $\hat{\boldsymbol{x}}_{\rm LS}$ is in fact an overestimate of $\boldsymbol{x}$. It is arguably more accurate to use

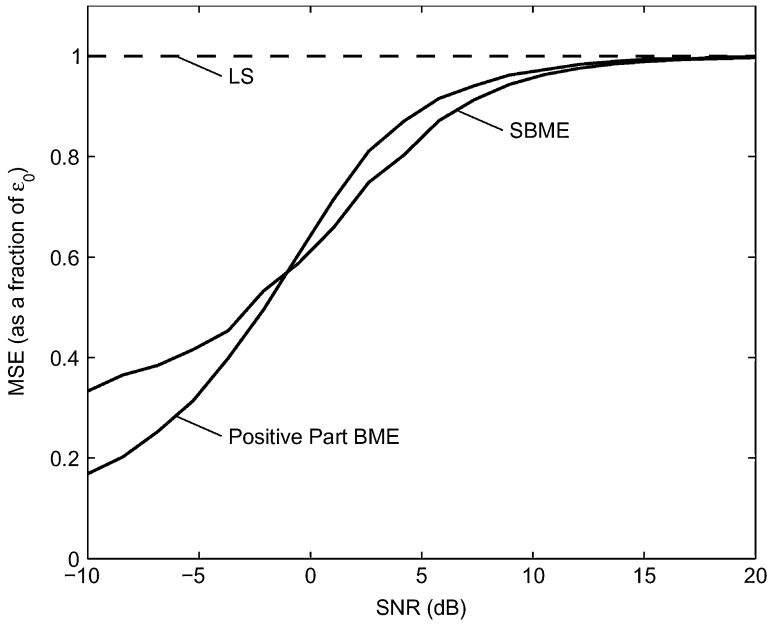

Fig. 3. Comparison between the positive part approach and the SBME. The positive part method results in stronger shrinkage, which improves performance for low SNR at the expense of high SNR.

a smaller value than $||\hat{\boldsymbol{x}}_{\rm LS}||^2$ to estimate $||\boldsymbol{x}||^2$. In particular, it is readily shown that

$$E\{||\hat{\boldsymbol{x}}_{\rm LS}||^2\} = ||\boldsymbol{x}||^2 + \epsilon_0. \tag{43}$$

Hence, one may opt to use

$$L^2 = ||\hat{\boldsymbol{x}}_{\rm LS}||^2 - \epsilon_0 \tag{44}$$

as an estimate of $||\boldsymbol{x}||^2$. It is important to note that such a value of L^2 cannot be used with the linear minimax method, since L^2 is negative with nonzero probability; a parameter set with negative radius is undefined. However, substituting (44) into a minimax technique, as per the blind minimax approach, can still lead to well-defined estimators. In particular, substituting (44) into the spherical minimax method (6) yields the "balanced" BME

$$\hat{\boldsymbol{x}}_{\rm BBM} = \left(1 - \frac{\epsilon_0}{||\hat{\boldsymbol{x}}_{\rm LS}||^2}\right)\hat{\boldsymbol{x}}_{\rm LS}. \tag{45}$$

A striking property of the balanced BME is that it reduces to Stein's estimator [6] in the i.i.d. case. Both techniques are well-defined unless $\hat{\boldsymbol{x}}_{\rm LS} = \boldsymbol{0}$, an event which has zero probability. Furthermore, the balanced BME extends Stein's method, in that it continues to dominate LS for the non-i.i.d. case, under suitable conditions. This is shown by the following theorem.

Theorem 3: Suppose $\epsilon_0/\epsilon_{\max} > 4$, where ϵ_0 is given by (3), $\epsilon_{\max}$ is the largest eigenvalue of $\boldsymbol{Q}^{-1}$, and $\boldsymbol{Q}$ is given by (4). Then, the balanced BME (45) strictly dominates the LS estimator.

Proof: The theorem follows by substituting $c = 0$ in Proposition 1. □

A well-known drawback of Stein's approach is that it sometimes causes negative shrinkage, i.e., the shrinkage factor in (45) is negative with nonzero probability. This is known to increase the MSE [24]. From the blind minimax perspective, this negative shrinkage is a result of the fact that L^2 can become negative. Thus, it is natural to replace (44) with

$$L^2 = (||\hat{\boldsymbol{x}}_{\rm LS}||^2 - \epsilon_0)_+ \tag{46}$$

where $(a)_+ = \max(a, 0)$. Substituting this value of L^2 into the spherical minimax estimator yields the "positive-part BME," given by

$$\hat{\boldsymbol{x}}_{\rm PBM} = \left(\frac{(||\hat{\boldsymbol{x}}_{\rm LS}||^2 - \epsilon_0)_+}{(||\hat{\boldsymbol{x}}_{\rm LS}||^2 - \epsilon_0)_+ + \epsilon_0}\right)\hat{\boldsymbol{x}}_{\rm LS}. \tag{47}$$

Note that when $||\hat{\boldsymbol{x}}_{\rm LS}||^2 - \epsilon_0 < 0$, the estimator $\hat{\boldsymbol{x}}_{\rm PBM}$ equals $\boldsymbol{0}$; in all other cases, $\hat{\boldsymbol{x}}_{\rm PBM} = \hat{\boldsymbol{x}}_{\rm BBM}$. Thus, (47) may be written as

$$\hat{\boldsymbol{x}}_{\rm PBM} = \left(1 - \frac{\epsilon_0}{||\hat{\boldsymbol{x}}_{\rm LS}||^2}\right)_+ \hat{\boldsymbol{x}}_{\rm LS}. \tag{48}$$

In other words, $\hat{\boldsymbol{x}}_{\rm PBM}$ is the positive part of the balanced BME. Specifically, in the i.i.d. case, $\hat{\boldsymbol{x}}_{\rm PBM}$ is the positive-part correction of Stein's estimator. In the i.i.d. case, Baranchik [24] demonstrated that $\hat{\boldsymbol{x}}_{\rm PBM}$ dominates $\hat{\boldsymbol{x}}_{\rm BBM}$. An interesting question for further research is whether the dominance property holds in the non-i.i.d. case as well.

The "balanced" method presented in this section for estimating the parameter set radius results in a value (44) of L^2 which is smaller than that of the SBME. As a result, the balanced approach causes more shrinkage towards the origin. This tends to improve performance for low signal-to-noise ratio (SNR) at the expense of performance degradation for high SNR. In particular, $\hat{\boldsymbol{x}}_{\rm PBM}$ has a positive probability of yielding an estimate of $\boldsymbol{0}$. 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 Fig. 3, the positive part estimator $\hat{\boldsymbol{x}}_{\rm PBM}$ is compared with the SBME of Section III. The problem setting of this simulation

is identical to that of Fig. 5(a), which will be described in detail in Section VII. 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.

VI. Comparison With LS Regularization

Independently of the development of Stein-type estimators, many researchers became aware of deficiencies of the LS approach for solving ill-conditioned problems. A variety of alternatives were proposed as a result. These substitutes were generally not required to dominate the LS estimator; rather, they were intended to improve estimation quality in specific scenarios. Of these approaches, the most common is Tikhonov regularization [25], also referred to as ridge regression [26].

Tikhonov regularization is intended for ill-posed problems, i.e., problems in which $\boldsymbol{H}^*\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H}$ is nearly singular. The matrix $\boldsymbol{Q} = \boldsymbol{H}^*\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H}$ is guaranteed to be positive-definite (and hence invertible), since we assume that $\boldsymbol{H}$ is full-rank and $\boldsymbol{C}_{\boldsymbol{w}}$ is positive-definite. However, $\boldsymbol{Q}$ may contain eigenvalues which are very close to zero. In these cases, the LS estimator (which depends on the term $\boldsymbol{Q}^{-1}$) causes severe amplification of measurement noise. In effect, an ill-posed setting is one in which the SNR of at least one parameter is extremely low; as we have seen, the LS approach results in overestimation in such conditions. Regularization techniques attempt to mitigate this problem by improving the conditioning of the matrix $\boldsymbol{Q}$.

Tikhonov regularization may be justified in a Bayesian setting, as follows. Suppose that the parameter vector $\boldsymbol{x}$ is known to be distributed normally, independently of the noise $\boldsymbol{w}$, with zero mean and a covariance matrix $\boldsymbol{C}_{\boldsymbol{x}}$. The minimum MSE estimator of $\boldsymbol{x}$ given $\boldsymbol{y}$ is then the Wiener filter [1], [27]

$$\hat{\boldsymbol{x}} = \left(\boldsymbol{H}^*\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H} + \boldsymbol{C}_{\boldsymbol{w}}^{-1}\right)^{-1}\boldsymbol{H}^*\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{y}. \tag{49}$$

In practice, $\boldsymbol{x}$ is a deterministic parameter, and thus does not have a covariance matrix. However, by replacing $\boldsymbol{C}_{\boldsymbol{w}}^{-1}$ with an appropriately chosen regularization matrix, the (generalized) Tikhonov estimator is obtained.

There are several methods for empirically selecting a regularization matrix $\boldsymbol{C}_{\boldsymbol{w}}^{-1}$. If nothing is known about the parameter $\boldsymbol{x}$, one possibility is to choose $\boldsymbol{C}_{\boldsymbol{w}} = \sigma_x^2\boldsymbol{I}$, where σ_x^2 is to be estimated from $\boldsymbol{y}$. Optimally, one would like to use the average value of x_i^2 as an approximation of the variance σ_x^2. However, since $\boldsymbol{x}$ is unknown, this is not possible. Instead, σ_x^2 can be estimated as $\sum \hat{x}_{\mathrm{LS},i}^2/m$, which is an approximation of the desired quantity $\sum x_i^2/m$. This results in the estimator

$$\hat{\boldsymbol{x}}_{\mathrm{T}}^{(1)} = \left(\boldsymbol{H}^*\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H} + \frac{m}{||\hat{\boldsymbol{x}}_{\mathrm{LS}}||^2}\boldsymbol{I}\right)^{-1}\boldsymbol{H}^*\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{y}. \tag{50}$$

This derivation is based on an empirical Bayes approach [28], in which the elements of $\boldsymbol{x}$ are assumed to be i.i.d. An alternative is to assume instead that the variance of $\boldsymbol{x}$ is proportional to the variance of the noise $\boldsymbol{w}$, which implies $\boldsymbol{C}_{\boldsymbol{w}} = \alpha\boldsymbol{Q}^{-1}$. In analogy to the previous derivation, one may then estimate α as $m/||\hat{\boldsymbol{x}}_{\mathrm{LS}}||_{\boldsymbol{Q}}^2$. Substituting into (49) results in the shrinkage estimator

$$\hat{\boldsymbol{x}}_{\mathrm{T}}^{(2)} = \frac{||\hat{\boldsymbol{x}}_{\mathrm{LS}}||_{\boldsymbol{Q}}^2}{m + ||\hat{\boldsymbol{x}}_{\mathrm{LS}}||_{\boldsymbol{Q}}^2}\hat{\boldsymbol{x}}_{\mathrm{LS}}. \tag{51}$$

Unfortunately, the Tikhonov estimators $\hat{\boldsymbol{x}}_{\mathrm{T}}^{(1)}$ and $\hat{\boldsymbol{x}}_{\mathrm{T}}^{(2)}$ do not dominate LS; like the original Tikhonov regularization, they perform poorly at high SNR values. To illustrate this, we performed a simulation in which the MSE of the LS method was compared to that of $\hat{\boldsymbol{x}}_{\mathrm{T}}^{(1)}$ and $\hat{\boldsymbol{x}}_{\mathrm{T}}^{(2)}$. In this example, 15 parameters were estimated using 15 independent measurements, with $\boldsymbol{H} = \boldsymbol{I}$. The noise variance of five of the measurements was 100 times larger than the noise variance of the remaining measurements. The parameter vector was chosen in the direction of a high-variance measurement, and its magnitude was varied to obtain different SNR values. Here and in the remainder of the paper, we define the SNR as

$$\mathrm{SNR} = \frac{||\boldsymbol{x}||^2}{E\{||\boldsymbol{w}||^2\}} = \frac{||\boldsymbol{x}||^2}{\mathrm{Tr}(\boldsymbol{C}_{\boldsymbol{w}})}. \tag{52}$$

For comparison, the MSE of the LS and blind minimax techniques were also calculated.

The results are displayed in Fig. 4. It is evident from this figure that the Tikhonov regularization is inadequate at high SNR, as it performs worse than the LS estimator. Both Tikhonov approaches converge to the LS approach at infinite SNR, but consistently obtain higher MSE than the LS method for SNR values above 5 dB. This makes them unattractive candidates for replacing the LS technique.

VII. Numerical Results

Estimator performance depends on a variety of operating conditions, including the effective dimension, the SNR, the eigenvalues of $\boldsymbol{Q} = \boldsymbol{H}^*\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H}$, and the value of the unknown parameter vector $\boldsymbol{x}$. Several computer simulations were implemented to test the effect of these conditions on performance of the SBME and EBME. In these tests, a value of $b = -1$ was used for the parameter set (18) of the EBME. The simulations were also used to compare the BMEs with Bock's estimator [13], which is the most commonly used extended Stein estimator [16], [17]. Like Stein's results, Bock's approach consists of a shrinkage estimator, given by

$$\hat{\boldsymbol{x}}_{\mathrm{Bock}} = \left(1 - \frac{\epsilon_0/\epsilon_{\max} - 2}{||\hat{\boldsymbol{x}}_{\mathrm{LS}}||_{\boldsymbol{Q}}^2}\right)\hat{\boldsymbol{x}}_{\mathrm{LS}}. \tag{53}$$

The theorems of Sections III and IV ensure that the BMEs achieve lower MSE than the LS estimator, but do not guarantee that this improvement is substantial. To measure this performance gain, we first chose a typical scenario, in which the number of parameters m and the number of measurements n were both 15. The system matrix $\boldsymbol{H}$ was chosen as $\boldsymbol{I}$, and the noise covariance $\boldsymbol{C}_{\boldsymbol{w}}$ was

$$\boldsymbol{C}_{\boldsymbol{w}} = \sigma^2\mathrm{diag}(1,1,1,1,0.5,0.2,0.2,0.2,0.2,0.1,0.1,0.1,0.1,0.05,0.05) \tag{54}$$

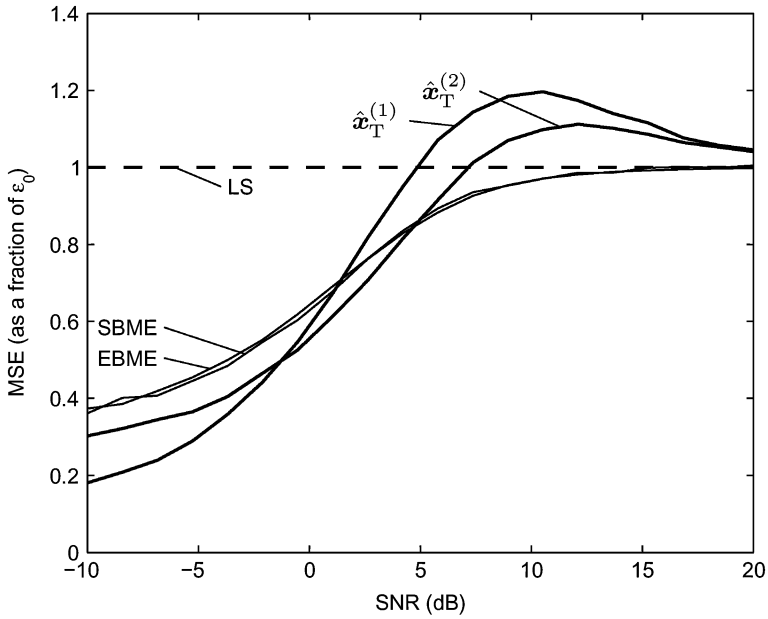

Fig. 4. Comparison between Tikhonov regularization, LS, and BME. The Tikhonov estimators $\hat{\boldsymbol{x}}_{\mathrm{T}}$ are seen to perform worse than the LS estimator at high SNR, whereas the BMEs dominate the LS method.

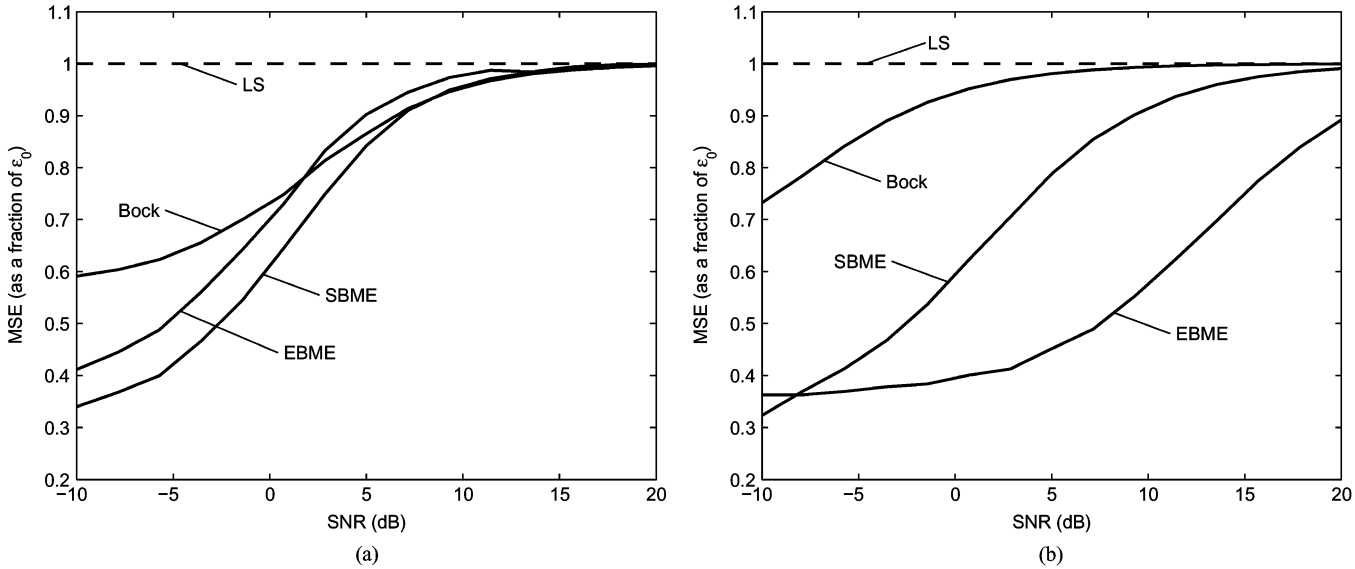

Fig. 5. MSE versus SNR for a typical operating condition: effective dimension 5.8, $m = n = 15$. (a) Parameter vector $\boldsymbol{x}$ in direction of maximum noise. (b) Parameter vector $\boldsymbol{x}$ in direction of minimum noise.

resulting in an effective dimension of 5.8. Here σ^2 was selected to achieve the desired SNR (52). To illustrate the dependence on the value of the parameter vector $\boldsymbol{x}$, two different settings were tested. In Fig. 5(a), $\boldsymbol{x}$ is chosen in the direction of the maximum eigenvector of $\boldsymbol{Q}^{-1}$, while in Fig. 5(b), $\boldsymbol{x}$ 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 Fig. 5 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 be optimal [1], 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, another simulation was performed, in which a large set of parameter values $\boldsymbol{x}$ were generated for different SNRs. For each estimator, and for each SNR, the lowest and highest MSE were determined, resulting in a measure of the performance range for each estimator. This performance range is displayed in Fig. 6 for two different choices of $\boldsymbol{C_w}$, which are indicated in the figure caption. One may observe that both BMEs outperform Bock's estimator under nearly all circumstances. It is also interesting to note that

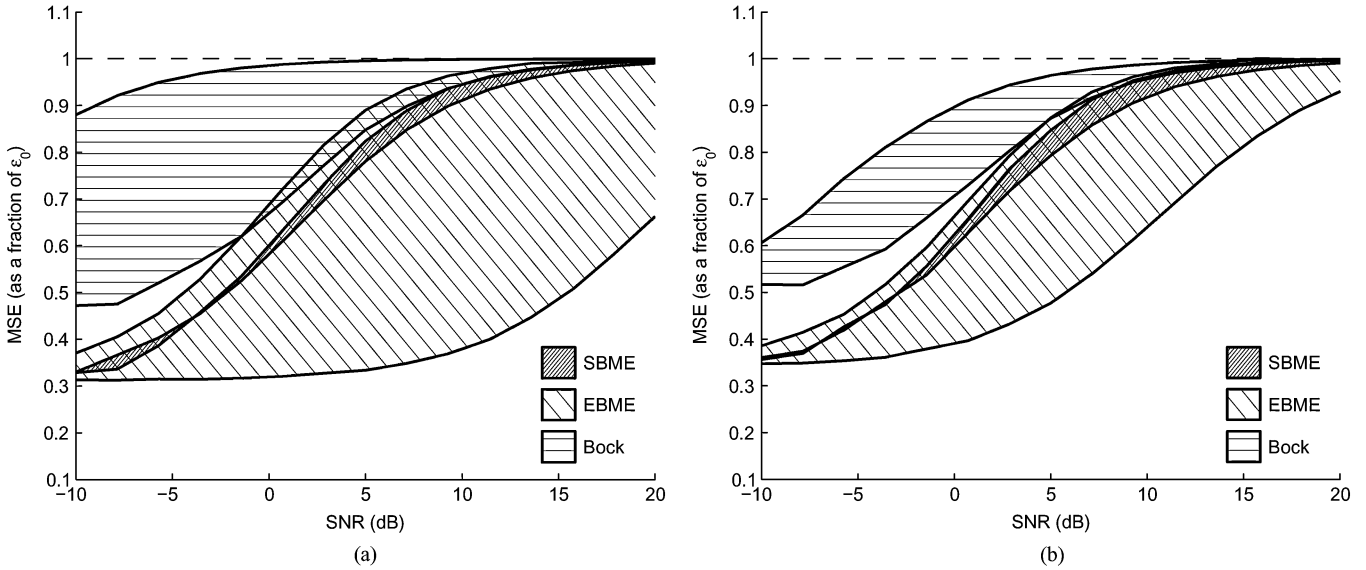

Fig. 6. Range of possible MSE values obtained for different values of $\boldsymbol{x}$, as a function of SNR. $\boldsymbol{H} = \boldsymbol{I}$ for both parts. (a) $m = n = 15$, with eigenvalues of $\boldsymbol{C_w}$ distributed uniformly between 1 and 0.01, resulting in an effective dimension of 7.6. (b) $m = n = 10$, with $\boldsymbol{C_w}$ containing five eigenvalues of 1 and five eigenvalues of 0.1, resulting in an effective dimension of 5.5.

while the MSE of the EBME is highly dependent on the value of the parameter value $\boldsymbol{x}$, 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 the parameter vector.

It is insightful to compare the performance of the SBME and EBME in Figs. 5 and 6. While the worst case performance of the two blind minimax techniques is similar, the EBME performs considerably better for some values of $\boldsymbol{x}$. 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 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 $\boldsymbol{x}$ is orthogonal to the noisiest measurements, whence the selective shrinkage is most effective; its performance gain is less substantial when $\boldsymbol{x}$ 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 $\boldsymbol{Q}$ is ill-conditioned, i.e., when some eigenvalues are much larger than others. This is because the shrinkage in Bock's method (53) is a function of $1/\|\hat{\boldsymbol{x}}_{\mathrm{LS}}\|^2_{\boldsymbol{Q}}$. As a result, when $\hat{\boldsymbol{x}}_{\mathrm{LS}}$ contains a significant component in the direction of a large eigenvalue of $\boldsymbol{Q}$, shrinkage becomes negligible. Yet, in this case, shrinkage is still desirable for the remaining eigenvalues. This effect is demonstrated in Fig. 7, which plots the performance of the various approaches for matrices $\boldsymbol{Q}$ having condition numbers between 1 and 1000. Here, ten parameters and ten measurements are used, $\boldsymbol{H} = \boldsymbol{I}$, and the noise covariance is chosen such that the first five eigenvalues equal 1 and the remaining five eigenvalues equal a value v, which is chosen to obtain the desired condition number. For each condition number, a large set of values $\boldsymbol{x}$ are chosen such that the SNR is 0 dB; as in Fig. 6, 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, despite the fact that shrinkage can still improve performance, as indicated by the performance of the SBME. 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.

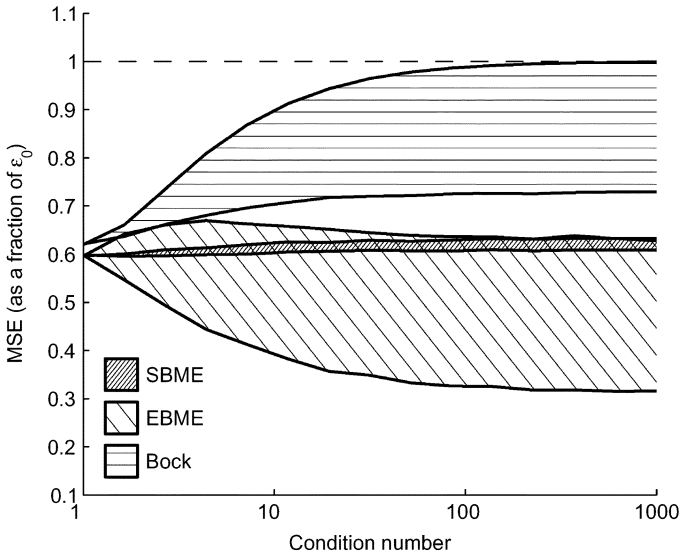

Fig. 7. Range of possible MSE values obtained for different values of $\boldsymbol{x}$, as a function of the condition number of $\boldsymbol{Q}$. SNR = 0 dB, $m = n = 10$.

VIII. Discussion

The blind minimax approach is a general technique for using minimax estimators in situations for which no parameter set is known. We considered an application of this concept to the

Gaussian linear regression model. Two novel estimators were proposed: a technique based on a spherical parameter set, and one based on an ellipsoidal parameter set. In Sections III and IV, these approaches were shown to dominate the LS method. Under fairly weak conditions, in any application which makes use of the LS estimator, the MSE performance can be improved by using a BME instead. Furthermore, in Section V, we demonstrated that Stein's approach, as well as its positive part modification, can be derived and generalized using the blind minimax framework.

It can readily be shown that the dominance condition of the SBME (Theorem 1) is weaker than the dominance condition of the EBME (Theorem 2), i.e., the conditions for SBME dominance hold whenever the conditions for EBME dominance hold. The dominance condition of Bock's estimator [13] is still weaker.[1] This would seem to indicate that Bock's estimator is superior to the proposed estimators. Yet the results of Section VII demonstrate that the opposite is true: the BMEs usually outperform Bock's estimator. This is true in particular for ill-conditioned problems, for which the LS estimator is notoriously inaccurate; for such problems, Bock's approach dominates the LS method by a negligible margin, whereas the BMEs achieve a significant performance gain. Thus, while dominance theorems are useful in providing sufficient conditions for improving on the LS estimator, they are ill-suited for comparing LS-dominating estimators. This conclusion is noteworthy since estimators are sometimes chosen by maximizing the range of conditions for which dominance is guaranteed. It seems that other analytical tools are required for comparing LS-dominating estimators. For example, it may be possible to prove that BMEs dominate Bock's estimator, for some problem settings.

The choice between the different BMEs is application dependent. As demonstrated in Section VII, the SBME reliably achieves constant performance for a variety of values of $\boldsymbol{x}$, although the typical performance of the EBME is superior. The EBME is particularly well adapted to ill-posed problems, in which some measurements are much more noisy than others. In such cases, the use of a single shrinkage factor for all measurements is clearly suboptimal. As a result, scalar shrinkage methods such as the SBME and Bock's technique often result in little improvement over the LS estimator, while the EBME is capable of selectively shrinking the noisy measurements, thus improving performance.

The use of a componentwise shrinkage technique such as the EBME may be useful in additional contexts as well. In some applications, MSE minimization is only a nominal goal which approximates some other error criterion. In these cases, a shrinkage estimator has no impact on the actual objective. For example, if the vector $\boldsymbol{x}$ is an image which is to be reconstructed, its subjective quality is not affected by multiplying the entire estimate by a scalar. Likewise, in a binary receiver, the sign of $\boldsymbol{x}$ must be determined, but the sign does not change when the estimate is shrunk. In such applications, the SBME (and Bock's estimator) have no effect on the final result, whereas the EBME can be used to improve performance.

[1]A simple change to the SBME (adding -2 to the numerator) changes its dominance condition to that of Bock's estimator, without significantly affecting its performance. However, we have been unable to derive this modification using the blind minimax approach, and thus prefer the simpler form of the SBME used in the paper.

IX. Conclusion

In this paper, we presented the blind minimax strategy, whereby one uses linear minimax estimators whose parameter set is itself estimated 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 method. Hence, in any such problem, the proposed estimators can be used in place of the LS approach, with a guaranteed performance gain. Apart from being useful in and of themselves, the proposed techniques support the underlying concept of blind minimax estimation. This concept can be applied to many other problems, such as estimation with uncertain system matrices, estimation with non-Gaussian noise, and sequential estimation. Use of the blind minimax approach in such problems remains a topic for further study.

Stein's discovery of LS-dominating estimators, half a century ago, shocked the statistics community, and his results are still rarely used in practice. It is our hope that the blind minimax concept will provide additional support for such estimators, both by supplying an intuitive understanding of Stein's phenomenon, and by providing a wide class of powerful new estimators.

Acknowledgment

The authors are grateful to the anonymous reviewers for their careful reading of the paper, which helped clarify and improve several results.

References

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[2] K. F. Gauss, "Theoria combinationis obsercationunt erronbus minimis obnoxiae," 1821.

[3] A.-M. Legendre, "Nouvelles méthodes pour la détermination des orbites des cometes," 1806.

[4] J. P. Romano and A. F. Siegel, *Counterexamples in Probability and Statistics*. Monterey, CA: Wadsworth & Brooks, 1985.

[5] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York: Springer, 1998.

[6] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate distribution," in *Proc. 3rd Berkeley Symp. Mathematical Statistics and Probability*, Berkeley, CA, 1956, vol. 1, pp. 197–206.

[7] A. Cohen, "All admissible linear estimators of the mean vector," *Ann. Math. Statist.*, vol. 37, no. 2, pp. 458–463, Apr. 1966.

[8] Y. C. Eldar, "Comparing between estimation approaches: Admissible and dominating linear estimators," *IEEE Trans. Signal Process.*, vol. 54, no. 5, pp. 1689–1702, May 2006.

[9] Y. Maruyama, "A unified and broadened class of admissible minimax estimators of a multivariate normal mean," *J. Multivariate Anal.*, vol. 64, pp. 196–205, 1998.

[10] Y. Maruyama, "Minimax admissible estimation of a multivariate normal mean and improvement upon the James-Stein estimator," Ph.D. dissertation, Univ. Tokyo, Tokyo, Japan, 2000.

[11] W. James and C. Stein, "Estimation with quadratic loss," in *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, Berkeley, CA, 1961, vol. 1, pp. 311–319.

[12] J. R. Thompson, "Some shrinkage techniques for estimating the mean," *J. Amer. Statist. Assoc.*, vol. 63, no. 321, pp. 113–122, Mar. 1968.
[13] M. E. Bock, "Minimax estimators of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 3, no. 1, pp. 209–218, Jan. 1975.
[14] B. Efron and C. Morris, "Stein's estimation rule and its competitors: An empirical Bayes approach," *J. Amer. Statist. Assoc.*, vol. 68, pp. 117–130, 1973.
[15] J. O. Berger, "Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss," *Ann. Statist.*, vol. 4, no. 1, pp. 223–226, Jan. 1976.
[16] J. H. Manton, V. Krishnamurthy, and H. V. Poor, "James-Stein state filtering algorithms," *IEEE Trans. Signal Process.*, vol. 46, no. 9, pp. 2431–2447, Sep. 1998.
[17] E. Greenberg and C. E. Webster, Jr., *Advanced Econometrics*, 2nd ed. New York: Wiley, 1983.
[18] B. Efron and C. Morris, "Combining possibly related estimation problems," *J. Roy. Statist. Soc. B*, vol. 35, no. 3, pp. 379–421, 1973.
[19] Z. Ben-Haim and Y. C. Eldar, "Minimax estimators dominating the least-squares estimator," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP 2005)*, Philadelphia, PA, Mar. 2005, vol. IV, pp. 53–56.
[20] Z. Ben-Haim and Y. C. Eldar, "Blind minimax estimators: Improving on least-squares estimation," in *Proc. IEEE Workshop on Statistical Signal Processing (SSP 2005)*, Bordeaux, France, Jul. 2005, pp. 545–550.
[21] M. S. Pinsker, "Optimal filtering of square-integrable signals in Gaussian noise," *Probl. Inf. Transm.*, vol. 16, pp. 120–133, 1980.
[22] Y. C. Eldar, A. Ben-Tal, and A. Nemirovski, "Robust mean-squared error estimation in the presence of model uncertainties," *IEEE Trans. Signal Process.*, vol. 53, no. 1, pp. 168–181, Jan. 2005.
[23] A. J. Baranchik, "A family of minimax estimators of the mean of a multivariate normal distribution," *Ann. Math. Statist.*, vol. 41, no. 2, Apr. 1970.
[24] A. J. Baranchik, "Multiple regression and estimation of the mean of a multivariate normal distribution," Dept. Statistics, Stanford Uni., Stanford, CA, Tech. Rep. 51, 1964.
[25] A. N. Tichonov and V. Y. Arsenin, *Solution of Ill-Posed Problems*. Washington, DC: Winston, 1977.
[26] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, 1970.
[27] N. Wiener, *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*. New York: Wiley, 1949.
[28] H. Robbins, "The empirical bayes approach to statistical decision problems," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 1–20, Mar. 1964.