Universal Weighted MSE Improvement of the Least-Squares Estimator

Yonina C. Eldar, *Senior Member, IEEE*

***Abstract*—Since the seminal work of Stein in the 1950s, there has been continuing research devoted to improving the total mean-squared error (MSE) of the least-squares (LS) estimator in the linear regression model. However, a drawback of these methods is that although they improve the total MSE, they do so at the expense of increasing the MSE of some of the individual signal components. Here we consider a framework for developing linear estimators that outperform the LS strategy over bounded norm signals, *under all weighted MSE measures*. This guarantees, for example, that both the total MSE and the MSE of each of the elements will be smaller than that resulting from the LS approach. We begin by deriving an easily verifiable condition on a linear method that ensures LS domination for every weighted MSE. We then suggest a minimax estimator that minimizes the worst-case MSE over all weighting matrices and bounded norm signals subject to the universal weighted MSE domination constraint.**

***Index Terms*—Admissible estimators, dominating estimators, linear estimation, weighted minimax MSE estimation.**

I. Introduction

LINEAR regression, or estimation in linear models, has been studied extensively since the pioneering work of Gauss on least-squares (LS) fitting [1]. One of the reasons for the broad interest in this problem is its applicability to a wide host of applications in diverse areas ranging from communication and economics to seismology and control.

The celebrated LS method is aimed at estimating a deterministic parameter vector $\mathbf{x}$ from noisy observations $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$ where $\mathbf{H}$ is a known model matrix and $\mathbf{n}$ is a perturbation vector. While typically in an estimation context the goal is to construct an estimate $\hat{\mathbf{x}}$ that is close in some sense to $\mathbf{x}$, the LS design criterion is the data error $\|\hat{\mathbf{y}} - \mathbf{y}\|^2$ between the estimated data $\hat{\mathbf{y}} = \mathbf{H}\hat{\mathbf{x}}$ and $\mathbf{y}$. Evidently, this approach is deterministic in nature: the objective is deterministic, and no prior statistical information is assumed on $\mathbf{x}$ or $\mathbf{n}$. Nonetheless, if the covariance of the noise is known, then it can be incorporated into the data error in the form of a weighting matrix, such that the resulting weighted LS estimate minimizes the variance among all unbiased methods. In the past 30 years attempts have been made to develop linear estimators that may be biased but closer to the true parameter $\mathbf{x}$ [2]–[7]. By now it is well established that even though unbiasedness may be appealing intuitively, it does not necessarily lead to a small estimation error $\hat{\mathbf{x}} - \mathbf{x}$ [8].

An alternative approach to account for the noise covariance is to define a statistical objective which directly measures the estimation error $\hat{\mathbf{x}} - \mathbf{x}$. A common design criterion is the total mean-squared error (MSE) given by $E\{\|\hat{\mathbf{x}} - \mathbf{x}\|^2\}$. Unfortunately, since $\mathbf{x}$ is deterministic, this measure depends in general on $\mathbf{x}$ and therefore cannot be minimized. One way to eliminate the signal dependency is by restricting attention to linear unbiased methods, resulting in the LS design. A different strategy is to assume that $\mathbf{x}$ is norm-bounded, and then minimize the worst-case MSE. This leads to the minimax trace MSE (MXTM) method, which was first suggested in [9] and then later extended in [10], [11]. A nice feature of this approach is that the total MSE of the MXTM estimator can be shown to be smaller than that of the LS method, *for all* values of $\mathbf{x}$ whose norm is smaller than the given bound [11]–[13]. Thus, the MXTM strategy dominates LS in the total MSE sense.

The concept of domination leads to a partial ordering among methods [14]. An estimator $\hat{\mathbf{x}}_1$ whose total MSE is no larger than that of a different estimate $\hat{\mathbf{x}}_2$ for all values of $\mathbf{x}$ on a given set and strictly smaller for some $\mathbf{x}$ is said to dominate $\hat{\mathbf{x}}_2$ on this set. An estimate $\hat{\mathbf{x}}_1$ is *admissible* if it is not dominated by any other strategy. The theory of LS domination has been well developed since the seminal work of Stein and James [15], [16], in which they showed that it is possible to construct a nonlinear estimator dominating the LS approach in a total MSE sense. Various modifications of the James–Stein method have since been developed that are applicable to the general linear model considered here [17]–[21].

One of the known shortcomings of the James–Stein concept is that it reduces the total MSE at the expense of an increase in the individual component MSEs. In the simplest setting in which $\mathbf{H} = \mathbf{I}$ and the noise is white with variance equal to 1, the MSE of an element of $\mathbf{x}$ can be as large as $m/4$, where m is the signal dimension [22], [23]. Consequently, although the total MSE may be small, specific elements may be severely miss-estimated. This drawback was formulated nicely by Lehmann [14]: "No one wants his or her blood test or Pap smear subjected to the possibility of large errors in order to improve a laboratory's average performance."

Componentwise MSE is an example of a weighted MSE measure where different weights are given to the individual signal elements to be estimated. A desirable property we may wish our estimator to possess is that it has "good" performance with different choices of weighting. For example, we may want our estimate to have low total MSE while still maintaining small componentwise MSE. Therefore, we consider a broader notion

Manuscript received May 28, 2006; revised October 1, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Simon J. Godsill. This work was supported in part by the Israel Science Foundation and by the Glasberg-Klein Research Fund.

The author is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: yonina@ee.technion.ac.il).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2007.913158

of domination: our goal is to characterize and design estimators that dominate the LS *for every possible choice of weighted MSE*.

The notion of local weighted-MSE superiority has been investigating previously in the statistical literature, where domination is required only for specific values of $\mathbf{x}$ (see, e.g., [24], [25], and references therein). However, since $\mathbf{x}$ is not known, the fact that one estimator may be better than another for some $\mathbf{x}$ does not help us select between estimators. Here, we focus on domination for all feasible values of $\mathbf{x}$ and, in contrast to previous approaches, we develop conditions that are independent of $\mathbf{x}$. This is a much stronger and more useful notion of superiority as it allows to decisively choose between strategies.

Unfortunately, it is impossible to dominate the LS method componentwise over the entire space [14]. Instead, several strategies have been proposed that dominate LS in the total MSE sense, and have better componentwise behavior than the James-Stein approach [23], [26]. However, as we show in this paper, if we restrict our attention to norm-bounded signals $\|\mathbf{x}\|_{\mathbf{T}} \leq L$, then we can design *linear* estimates that dominate LS simultaneously *under all weighted MSE measures*. Mathematically, this requires that the MSE matrix of our estimate $\hat{\mathbf{x}}$ is smaller or equal (in a matrix sense) than the MSE matrix of LS. Focusing on *linear* estimates, we derive an easily verifiable necessary and sufficient condition such that $\hat{\mathbf{x}}$ dominates LS in a matrix sense for all $\|\mathbf{x}\|_{\mathbf{T}} \leq L$. As we show, there is a large class of methods with this property. An important question is how to select a "good" strategy from all dominating possibilities. To this end, we suggest a minimax matrix MSE (MXMM) method that minimizes the worst-case weighted MSE among all weighting matrices and feasible vectors $\mathbf{x}$ subject to the domination constraint. The MXMM solution dominates LS under all weighted MSE criteria, and at the same time has small worst-case MSE. As we show, this approach has the additional desirable property that it is admissible in a weighted MSE sense, meaning that there is no other linear estimator with smaller MSE matrix.

We begin in Section II by describing our problem and the shortcomings of the MXTM method. A necessary and sufficient domination condition in the matrix sense is derived in Section III. In Section IV we develop the MXMM estimate and show that it can be found as a solution to a semidefinite programming problem (SDP) [27], [28]. We then consider, in Section V, a broad class of settings in which a more explicit solution can be found which depends on a single parameter. In Section VII, we compare our approach with the MXTM and LS strategies.

II. MSE Matrix Domination of Least-Squares

We denote vectors in $\mathbb{C}^m$ by boldface lowercase letters and matrices in $\mathbb{C}^{n\times m}$ by boldface uppercase letters. The ith element of a vector $\mathbf{y}$ is represented by $\mathbf{y}_i$ and the iith element of a matrix $\mathbf{Q}$ by $[\mathbf{Q}]_{ii}$. The identity matrix of appropriate dimension is written as $\mathbf{I}$, $(\cdot)^*$ is the Hermitian conjugate of the corresponding matrix, $\widehat{(\cdot)}$ is an estimated vector or matrix, and $\operatorname{diag}(\delta_1,\ldots,\delta_m)$ is an $m\times m$ diagonal matrix with diagonal elements δ_i. The vector $\mathbf{e}^i$ has 1 in the ith component and 0 everywhere else. For two Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A}\succ\mathbf{B}$ ($\mathbf{A}\succeq\mathbf{B}$) means that $\mathbf{A}-\mathbf{B}$ is positive definite (semidefinite). The largest and smallest eigenvalues of a Hermitian matrix $\mathbf{A}$ are denoted $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$, respectively. The weighted norm of a vector $\mathbf{x}$ is defined as $\|\mathbf{x}\|_{\mathbf{T}}^2 = \mathbf{x}^*\mathbf{T}\mathbf{x}$.

A. Estimation Problem

We treat the problem of estimating a deterministic parameter vector $\mathbf{x}\in\mathbb{C}^m$ from observations $\mathbf{y}\in\mathbb{C}^n$ which are related through the linear model

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}. \tag{1}$$

Here $\mathbf{H}$ is a known $n\times m$ model matrix with full rank m, and $\mathbf{n}$ is a zero-mean random vector with covariance $\mathbf{C}\succ\mathbf{0}$. We assume that the weighted norm of $\mathbf{x}$ is bounded so that $\|\mathbf{x}\|_{\mathbf{T}} \leq L$ for some $\mathbf{T}\succ\mathbf{0}$ and scalar $L>0$. This constraint is used in many different statistical methods (see, e.g., [4], [9], [29]). In practice, if L is unknown, then we can estimate it from the data using the LS estimator [21]; an example is given in Section VII.

We restrict our attention to linear estimators of $\mathbf{x}$ of the form $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$ for some $m\times n$ matrix $\mathbf{G}$. A popular measure of estimator performance is the total MSE defined by

$$\begin{aligned}\mathrm{MSE}(\hat{\mathbf{x}}) &= E\{\|\hat{\mathbf{x}}-\mathbf{x}\|^2\} = \sum_{i=1}^{m} E\{|\hat{\mathbf{x}}_i-\mathbf{x}_i|^2\}\\ &= \mathrm{Tr}(\mathbf{M}(\hat{\mathbf{x}}))\end{aligned} \tag{2}$$

where $\mathbf{M}(\hat{\mathbf{x}})$, or $\mathbf{M}(\mathbf{G})$, is the MSE matrix

$$\mathbf{M}(\hat{\mathbf{x}}) = E\{(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^*\} = E\{(\mathbf{G}\mathbf{y}-\mathbf{x})(\mathbf{G}\mathbf{y}-\mathbf{x})^*\}. \tag{3}$$

Using the model (1), it is easy to show that

$$\mathbf{M}(\mathbf{G}) = (\mathbf{I}-\mathbf{G}\mathbf{H})\mathbf{x}\mathbf{x}^*(\mathbf{I}-\mathbf{G}\mathbf{H})^* + \mathbf{G}\mathbf{C}\mathbf{G}^*. \tag{4}$$

More generally, we may consider a weighted total MSE

$$\mathrm{MSEW}(\hat{\mathbf{x}}) = E\{(\hat{\mathbf{x}}-\mathbf{x})^*\mathbf{W}(\hat{\mathbf{x}}-\mathbf{x})\} = \mathrm{Tr}(\mathbf{W}\mathbf{M}(\hat{\mathbf{x}})) \tag{5}$$

for some weighting matrix $\mathbf{W}\succeq\mathbf{0}$ so that different weights are assigned to the individual errors. For example, choosing $\mathbf{W} = \mathbf{e}^i\mathbf{e}^{i^*}$ results in $\mathrm{MSEW}(\hat{\mathbf{x}}) = E\{|\hat{\mathbf{x}}_i-\mathbf{x}_i|^2\}$, i.e., the MSE of the ith component.

For a given choice of $\mathbf{W}$, a possible design criterion is to minimize the weighted MSE (5). Unfortunately, this measure depends in general on $\mathbf{x}$, which is unknown, and therefore cannot be minimized. The dependency of the MSE on $\mathbf{x}$ can be eliminated by requiring that $\mathbf{G}\mathbf{H} = \mathbf{I}$, or equivalently, restricting attention to unbiased estimators. When $\mathbf{W} = \mathbf{I}$, minimizing the resulting MSE leads to the celebrated LS estimator

$$\hat{\mathbf{x}}_{\mathrm{LS}} = (\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}^{-1}\mathbf{y} = \mathbf{G}_{\mathrm{LS}}\mathbf{y}. \tag{6}$$

However, this does not mean that the residual MSE is small. In fact, it is well known that the MSE of the LS method can be large in many estimation problems.

To directly control the MSE, a minimax total MSE (MXTM) approach was suggested in [10], in which the worst-case total MSE is minimized over $\|\mathbf{x}\|_{\mathbf{T}} \leq L$. It was then shown in [12] that the MXTM strategy dominates LS in terms of total MSE, meaning that its total MSE is smaller than that of LS *for all* *feasible values of* $\mathbf{x}$. Furthermore, this estimator is total MSE admissible, namely, there is no other linear method with smaller total MSE for all $\mathbf{x}$ [11]. Although the MXTM estimator has smaller total MSE, the MSE of an individual component may be larger than that resulting from the LS method. To illustrate this point, suppose that $\mathbf{T} = \mathbf{I}$. In this case the MXTM estimator is given by

$$\hat{\mathbf{x}}_{\mathrm{MX}} = \frac{L^2}{L^2 + \mathrm{Tr}(\mathbf{Q})} \hat{\mathbf{x}}_{\mathrm{LS}} \tag{7}$$

where we denoted

$$\mathbf{Q} = (\mathbf{H}^* \mathbf{C}^{-1} \mathbf{H})^{-1}. \tag{8}$$

The MSE of the ith component can be computed by substituting $\mathbf{W} = \mathbf{e}^i \mathbf{e}^{i*}$ into (5) which together with (4) yields

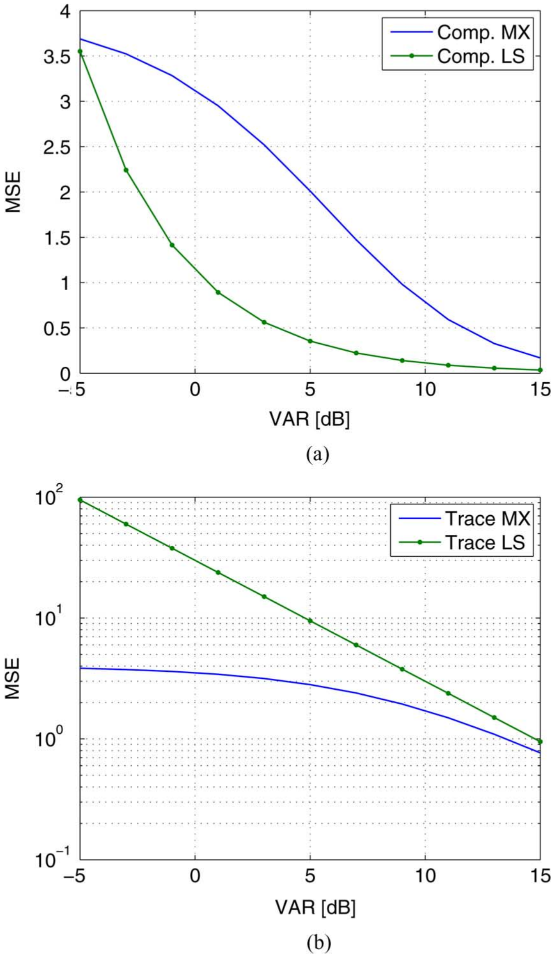
$$E\{|[\hat{\mathbf{x}}_{\mathrm{MX}} - \mathbf{x}]_i|^2\} = \frac{\mathrm{Tr}^2(\mathbf{Q})|\mathbf{x}_i|^2 + L^4 [\mathbf{Q}]_{ii}}{(L^2 + \mathrm{Tr}(\mathbf{Q}))^2}. \tag{9}$$

The largest value of (9) over $\|\mathbf{x}\| \leq L$ is obtained when $\mathbf{x}_i = L\mathbf{e}^i$. In comparison, the MSE of the ith component using the LS approach is

$$E\{|[\hat{\mathbf{x}}_{\mathrm{LS}} - \mathbf{x}]_i|^2\} = [\mathbf{Q}]_{ii} \tag{10}$$

which is independent of $\mathbf{x}$. The total MSE of the MXTM and LS methods can be obtained from (9) and (10) respectively, by summing over i.

In Fig. 1, we compare the MSE of the LS with the worst-case MSE resulting from the MXTM approach for $L = 2$, white noise and a random choice of $\mathbf{H}$, with $n = 8, m = 5$. In Fig. 1(a) we plot the MSE in estimating the first component, as a function of the noise variance (in dB). As can be seen from the figure, the component MSE of the MXTM estimator can be higher than that of the LS approach. In this particular example, 3 of the components have behavior similar to that of Fig. 1(a), while the other two components have a very large MSE using the LS strategy and a substantially smaller MSE with the MXTM approach. In Fig. 1(b) we plot the total MSE of the two methods. As expected, the total MSE of the MXTM strategy is always smaller than that of LS.



(a)

(b)

Fig. 1. MSE in estimating $\mathbf{x}$ as a function of the noise variance using the MXTM (7) and LS estimators: (a) MSE of the first component; (b) total MSE. For the MXTM estimator, the worst-case MSE over $\|\mathbf{x}\| \leq L$ is plotted in each case.

B. Matrix Domination

Fig. 1 illustrates that minimizing the total MSE may be insufficient when in addition we would like each of the components to have small MSE, or when a more general weighted total MSE is of interest. To ensure LS domination for a weighted MSE, $\hat{\mathbf{x}}$ must have the property that

$$E\{(\hat{\mathbf{x}} - \mathbf{x})^* \mathbf{W} (\hat{\mathbf{x}} - \mathbf{x})\} \leq E\{(\hat{\mathbf{x}}_{\mathrm{LS}} - \mathbf{x})^* \mathbf{W} (\hat{\mathbf{x}}_{\mathrm{LS}} - \mathbf{x})\} \tag{11}$$

for all $\|\mathbf{x}\|_{\mathbf{T}} \leq L$. Since different choices of $\mathbf{W}$ may be considered simultaneously, for example we may want small total MSE and low componentwise MSE, we require that (11) holds for all choices of $\mathbf{W}$. This leads to the following definition.

Definition 1: For two linear estimators $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ we say that $\hat{\mathbf{x}}_1$ dominates $\hat{\mathbf{x}}_2$ in a weighted MSE sense if

$$\mathrm{MSEW}(\hat{\mathbf{x}}_1) \leq \mathrm{MSEW}(\hat{\mathbf{x}}_2), \quad \text{for all } \|\mathbf{x}\|_{\mathbf{T}} \leq L, \quad \mathbf{W} \succeq \mathbf{0} \tag{12}$$

where $\mathrm{MSEW}(\hat{\mathbf{x}})$ is defined by (5), and for each $\mathbf{W} \succ \mathbf{0}$,

$$\mathrm{MSEW}(\hat{\mathbf{x}}_1) < \mathrm{MSEW}(\hat{\mathbf{x}}_2), \quad \text{for some } \|\mathbf{x}\|_{\mathbf{T}} \leq L. \tag{13}$$

It is easy to show that (12) and (13) translate into a simple condition on the MSE matrices of $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$:

Proposition 1: A linear estimate $\hat{\mathbf{x}}_1$ dominates a linear estimate $\hat{\mathbf{x}}_2$ in the weighted MSE sense if and only if

$$\mathbf{M}(\hat{\mathbf{x}}_1) \preceq \mathbf{M}(\hat{\mathbf{x}}_2) \quad \text{for all } \|\mathbf{x}\|_{\mathbf{T}} \leq L \tag{14}$$

and[1] $\mathbf{M}(\hat{\mathbf{x}}_1) \neq \mathbf{M}(\hat{\mathbf{x}}_2)$, where $\mathbf{M}(\hat{\mathbf{x}})$ is defined by (4).

Note that we require (13) to hold for $\mathbf{W} \succ 0$ and not all $\mathbf{W} \succeq 0$. This is because the later requirement cannot be satisfied for $\mathbf{W} = 0$ and is therefore too strong.

Proof: Suppose first that (14) is satisfied and $\mathbf{M}(\hat{\mathbf{x}}_1) \neq \mathbf{M}(\hat{\mathbf{x}}_2)$. Then $\mathrm{Tr}(\mathbf{W}\mathbf{M}(\hat{\mathbf{x}}_1)) \leq \mathrm{Tr}(\mathbf{W}\mathbf{M}(\hat{\mathbf{x}}_2))$ for all $\mathbf{W} \succeq \mathbf{0}$, which together with (5) proves (12). To show that strict inequality holds when $\mathbf{W} \succ \mathbf{0}$ for some $\mathbf{x}$, suppose to the contrary

[1] By the matrix inequality we mean that we do not have equality for all $\mathbf{x}$, although we may have equality for some $\mathbf{x}$.

that for some $\mathbf{W} \succ \mathbf{0}$ and each $\|\mathbf{x}\|_{\mathbf{T}} \leq L$, $\mathrm{Tr}(\mathbf{W}\mathbf{M}(\hat{\mathbf{x}}_1)) = \mathrm{Tr}(\mathbf{W}\mathbf{M}(\hat{\mathbf{x}}_2))$, or

$$\mathrm{Tr}(\mathbf{W}\mathbf{A}) = 0, \quad \text{for all } \|\mathbf{x}\|_{\mathbf{T}} \leq L \tag{15}$$

where

$$\mathbf{A} = \mathbf{M}(\hat{\mathbf{x}}_1) - \mathbf{M}(\hat{\mathbf{x}}_2) \preceq \mathbf{0}. \tag{16}$$

Since $\mathbf{W} \succ \mathbf{0}$, (15) and (16) together imply that $\mathbf{A} = \mathbf{0}$, or $\mathbf{M}(\hat{\mathbf{x}}_1) = \mathbf{M}(\hat{\mathbf{x}}_2)$ for all $\|\mathbf{x}\|_{\mathbf{T}} \leq L$.

It remains to show that if $\mathbf{M}(\hat{\mathbf{x}}_1) = \mathbf{M}(\hat{\mathbf{x}}_2)$ for all $\|\mathbf{x}\|_{\mathbf{T}} \leq L$ then $\mathbf{M}(\hat{\mathbf{x}}_1) = \mathbf{M}(\hat{\mathbf{x}}_2)$ everywhere. Choosing $\mathbf{x} = 0$ implies that $\mathbf{G}_1\mathbf{C}\mathbf{G}_1^* = \mathbf{G}_2\mathbf{C}\mathbf{G}_2^*$. For any other choice of $\mathbf{x}$, the equality then means that for all $\|\mathbf{x}\|_{\mathbf{T}} \leq L$,

$$(\mathbf{I} - \mathbf{G}_1\mathbf{H})\mathbf{x}\mathbf{x}^*(\mathbf{I} - \mathbf{G}_1\mathbf{H})^* = (\mathbf{I} - \mathbf{G}_2\mathbf{H})\mathbf{x}\mathbf{x}^*(\mathbf{I} - \mathbf{G}_2\mathbf{H})^*. \tag{17}$$

Now, let $\|\mathbf{x}\|_{\mathbf{T}} > L$ be arbitrary. Then we can define $\mathbf{x}' = L\mathbf{x}/\|\mathbf{x}\|_{\mathbf{T}}$ which satisfies $\|\mathbf{x}'\|_{\mathbf{T}} \leq L$ so that for this choice of $\mathbf{x}'$, (17) must hold. But this also implies that (17) is true for $\mathbf{x}$ which means that $\mathbf{M}(\hat{\mathbf{x}}_1) = \mathbf{M}(\hat{\mathbf{x}}_2)$ everywhere.

Next, let (12) hold for all $\mathbf{W} \succeq \mathbf{0}$. Choosing $\mathbf{W} = \mathbf{w}\mathbf{w}^*$ for an arbitrary $\mathbf{w} \neq \mathbf{0}$ we have that

$$\mathbf{w}^*\mathbf{M}(\hat{\mathbf{x}}_1)\mathbf{w} \leq \mathbf{w}^*\mathbf{M}(\hat{\mathbf{x}}_2)\mathbf{w} \text{ for all } \|\mathbf{x}\|_{\mathbf{T}} \leq L, \mathbf{w} \tag{18}$$

which proves (14). Furthermore, from (13), for all $\mathbf{W} \succ \mathbf{0}$,

$$\mathrm{Tr}(\mathbf{W}\mathbf{M}(\hat{\mathbf{x}}_1)) < \mathrm{Tr}(\mathbf{W}\mathbf{M}(\hat{\mathbf{x}}_2)) \quad \text{for some } \mathbf{x} \tag{19}$$

so that $\mathbf{M}(\hat{\mathbf{x}}_1) \neq \mathbf{M}(\hat{\mathbf{x}}_2)$. ■

Proposition 1 implies that weighted MSE domination is equivalent to *matrix domination*: the MSE matrix of $\hat{\mathbf{x}}_1$ must be no larger in the matrix sense than that of $\hat{\mathbf{x}}_2$.

The connection between weighted MSE and matrix domination without requiring strict domination was proved in [30]. The additional requirement, which we add here, for strict domination for every $\mathbf{W} \succ 0$ results in the necessary and sufficient condition $\mathbf{M}(\hat{\mathbf{x}}_1) \neq \mathbf{M}(\hat{\mathbf{x}}_2)$.

In the special case in which $\hat{\mathbf{x}}_2 = \hat{\mathbf{x}}_{\mathrm{LS}}$, Proposition 1 can be further simplified by noting that $\mathbf{M}(\hat{\mathbf{x}}) = \mathbf{M}(\hat{\mathbf{x}}_{\mathrm{LS}})$ for all $\mathbf{x}$ if and only if $\hat{\mathbf{x}} = \hat{\mathbf{x}}_{\mathrm{LS}}$:

Proposition 2: A linear estimate $\hat{\mathbf{x}} \neq \hat{\mathbf{x}}_{\mathrm{LS}}$ dominates the LS estimate $\hat{\mathbf{x}}_{\mathrm{LS}}$ in the matrix sense if and only if

$$\mathbf{M}(\hat{\mathbf{x}}) \preceq (\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H})^{-1} \stackrel{\triangle}{=} \mathbf{Q} \quad \text{for all} \|\mathbf{x}\|_{\mathbf{T}} \leq L \tag{20}$$

where $\mathbf{M}(\hat{\mathbf{x}})$ is defined by (4).

Proof: We first note that substituting (6) into (4), $\mathbf{M}(\hat{\mathbf{x}}_{\mathrm{LS}}) = \mathbf{Q}$. The proof then follows from combining Proposition 1 with the fact that

$$\mathbf{M}(\hat{\mathbf{x}}) = \mathbf{M}(\hat{\mathbf{x}}_{\mathrm{LS}}) \quad \text{for all } \mathbf{x} \Rightarrow \hat{\mathbf{x}} = \hat{\mathbf{x}}_{\mathrm{LS}}. \tag{21}$$

To establish (21), suppose that $\mathbf{x} = \mathbf{0}$. The left-hand equality then implies $\mathbf{G}\mathbf{C}\mathbf{G}^* = \mathbf{G}_{\mathrm{LS}}\mathbf{C}\mathbf{G}_{\mathrm{LS}}^* = \mathbf{Q}$. Choosing an arbitrary $\mathbf{x} \neq \mathbf{0}$ we have $\mathbf{G}\mathbf{H} = \mathbf{I}$. Thus, $\mathbf{M}(\hat{\mathbf{x}}) = \mathbf{M}(\hat{\mathbf{x}}_{\mathrm{LS}})$ for all $\mathbf{x}$ only if

$$\mathbf{G}\mathbf{C}\mathbf{G}^* = \mathbf{G}_{\mathrm{LS}}\mathbf{C}\mathbf{G}_{\mathrm{LS}}^*, \quad \mathbf{G}\mathbf{H} = \mathbf{I}. \tag{22}$$

Now, recall that $\mathbf{G}_{\mathrm{LS}}$ minimizes the MSE among unbiased estimators, so that it is the solution to

$$\min_{\mathbf{G}}\{\mathrm{Tr}(\mathbf{G}\mathbf{C}\mathbf{G}^*) : \mathbf{G}\mathbf{H} = \mathbf{I}\}. \tag{23}$$

Since the problem (23) is strictly convex, the minimizer is unique, implying that

$$\mathbf{G}\mathbf{C}\mathbf{G}^* \succ \mathbf{G}_{\mathrm{LS}}\mathbf{C}\mathbf{G}_{\mathrm{LS}}^*, \quad \text{for all } \mathbf{G}\mathbf{H} = \mathbf{I}, \mathbf{G} \neq \mathbf{G}_{\mathrm{LS}} \tag{24}$$

which contradicts (22). ■

In Section III, we use Proposition 2 to derive an easily verifiable necessary and sufficient condition on $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$ to dominate $\hat{\mathbf{x}}_{\mathrm{LS}}$ in a matrix sense over all $\|\mathbf{x}\|_{\mathbf{T}} \leq L$.

C. Estimation Strategy

An important question is how to choose a "good" method among all the LS matrix-dominating possibilities. An obvious property we would like our approach to posses is that it is *admissible* in the matrix sense, namely that it is not matrix-dominated by any other linear strategy. In addition, we would like our estimate to have small weighted MSE for all choices of $\mathbf{W}$. To construct an admissible dominating method with good MSE performance we propose choosing an estimate that minimizes the worst-case weighted MSE over all $\mathbf{W} \succeq \mathbf{0}$ and $\|\mathbf{x}\|_{\mathbf{T}} \leq L$, subject to the matrix domination condition. In order to obtain a well-defined problem we need to constrain the norm of $\mathbf{W}$. This is because the weighted MSE $\mathrm{Tr}(\mathbf{M}(\hat{\mathbf{x}})\mathbf{W})$ can grow without bound if $\mathbf{W}$ is unbounded. Furthermore, minimizing $\mathrm{Tr}(\mathbf{M}(\hat{\mathbf{x}})\mathbf{W})$ is equivalent to minimizing $\alpha\mathrm{Tr}(\mathbf{M}(\hat{\mathbf{x}})\mathbf{W})$ for any $\alpha > 0$ so that the choice of scaling is immaterial. Therefore, we assume that $\mathbf{W} \preceq \mathbf{I}$, leading to the following optimization problem:

$$\begin{aligned} &\min_{\hat{\mathbf{x}}} \max_{\mathbf{x},\mathbf{W}}\{\mathrm{Tr}(\mathbf{W}\mathbf{M}(\hat{\mathbf{x}})) : \|\mathbf{x}\|_{\mathbf{T}} \leq L, \mathbf{0} \preceq \mathbf{W} \preceq \mathbf{I}\} \\ &\text{s.t. } \mathbf{M}(\hat{\mathbf{x}}) \preceq \mathbf{M}(\hat{\mathbf{x}}_{\mathrm{LS}}), \quad \text{for all } \|\mathbf{x}\|_{\mathbf{T}} \leq L. \end{aligned} \tag{25}$$

The solution $\hat{\mathbf{x}}$ is referred to as the minimax matrix-MSE (MXMM) estimate and is denoted by $\hat{\mathbf{x}}_{\mathrm{MXMM}}$.

In Section IV we show that $\hat{\mathbf{x}}_{\mathrm{MXMM}}$ is admissible, and derive a size m SDP formulation of (25). This allows to compute the solution efficiently using standard software packages. An explicit expression for $\hat{\mathbf{x}}_{\mathrm{MXMM}}$ is developed in Section V under the assumption that the matrices $\mathbf{T}$ and $\mathbf{Q}$ can be jointly diagonalized.

Note that we could have used any other constraint to restrict $\mathbf{W}$ to be bounded, for example, $\mathrm{Tr}(\mathbf{W}) \leq 1$. However, for this choice, it can be shown that the problem we end up with is not strictly convex, and therefore the solution is not unique. In Section IV we prove that the admissibility of $\hat{\mathbf{x}}_{\mathrm{MXMM}}$ is a direct consequence of the uniqueness of the solution to (25) so that it is important to restrict $\mathbf{W}$ in a manner that results in a strictly convex problem.

III. LS Matrix-Dominating Estimators

A problem that has been treated previously in the statistical literature is that of local MSE superiority, where matrix domination holds for a specific value of $\mathbf{x}$, i.e., $\mathbf{M}(\hat{\mathbf{x}}_1) \preceq \mathbf{M}(\hat{\mathbf{x}}_2)$ for some $\mathbf{x}$ (see e.g., [24], [25] and references therein). Our interest is in domination over all feasible values of $\mathbf{x}$ so that the condition for domination is independent of $\mathbf{x}$.

Theorem 1 below shows that the LS matrix-domination condition of Proposition 2 can be translated into the requirement that the largest eigenvalue of an appropriate matrix is non-positive.

Theorem 1: Let $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$ be a linear estimate of $\mathbf{x}$ in the model (1) and let $\mathbf{Q} = (\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H})^{-1}$. Then $\hat{\mathbf{x}} \neq \hat{\mathbf{x}}_{\text{LS}}$ dominates LS in the matrix sense for all $\|\mathbf{x}\|_{\mathbf{T}} \leq L$ if and only if

$$\lambda_{\max}(\mathbf{G}\mathbf{C}\mathbf{G}^* - \mathbf{Q} + L^2(\mathbf{I} - \mathbf{G}\mathbf{H})\mathbf{T}^{-1}(\mathbf{I} - \mathbf{G}\mathbf{H})^*) \leq 0.$$

Proof: From (4) and (20) matrix domination is equivalent to

$$\mathbf{B}\mathbf{x}\mathbf{x}^*\mathbf{B}^* + \mathbf{A} \preceq \mathbf{0}, \quad \forall \|\mathbf{x}\|_{\mathbf{T}} \leq L \tag{26}$$

where we defined $\mathbf{A} = \mathbf{G}\mathbf{C}\mathbf{G}^* - \mathbf{Q}$ and $\mathbf{B} = \mathbf{I} - \mathbf{G}\mathbf{H}$. In order for (26) to be satisfied we must have that

$$\max_{\|\mathbf{x}\|_{\mathbf{T}} \leq L} \mathbf{y}^*\mathbf{B}\mathbf{x}\mathbf{x}^*\mathbf{B}^*\mathbf{y} + \mathbf{y}^*\mathbf{A}\mathbf{y} \leq 0, \quad \forall \mathbf{y}. \tag{27}$$

Now

$$\begin{aligned}\max_{\|\mathbf{x}\|_{\mathbf{T}} \leq L} \mathbf{y}^*\mathbf{B}\mathbf{x}\mathbf{x}^*\mathbf{B}^*\mathbf{y} &= \max_{\|\mathbf{x}\| \leq L} \mathbf{x}^*\mathbf{T}^{-1/2}\mathbf{B}^*\mathbf{y}\mathbf{y}^*\mathbf{B}\mathbf{T}^{-1/2}\mathbf{x} \\ &= L^2\mathbf{y}^*\mathbf{B}\mathbf{T}^{-1}\mathbf{B}^*\mathbf{y}. \end{aligned} \tag{28}$$

Therefore, (26) is equivalent to

$$L^2\mathbf{y}^*\mathbf{B}\mathbf{T}^{-1}\mathbf{B}^*\mathbf{y} + \mathbf{y}^*\mathbf{A}\mathbf{y} \leq 0, \quad \forall \mathbf{y} \tag{29}$$

or $L^2\mathbf{B}\mathbf{T}^{-1}\mathbf{B}^* + \mathbf{A} \preceq \mathbf{0}$, which completes the proof. ■

A. Examples of Matrix Dominating Estimators

We now present some examples of Theorem 1. For simplicity, we assume that $\mathbf{T} = \mathbf{I}$.

A popular class of estimators for the linear regression model are the generalized shrinkage (GS) methods, which were first introduced by Obenchain [31]. Let $\mathbf{Q}$ have an eigendecomposition $\mathbf{Q} = \mathbf{V}\Sigma\mathbf{V}^*$ where $\mathbf{V}$ is a unitary matrix and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_m)$. Then the GS estimators have the form

$$\hat{\mathbf{x}} = \mathbf{V}\mathbf{D}\mathbf{V}^*\hat{\mathbf{x}}_{\text{LS}} \tag{30}$$

for some $\mathbf{D} = \text{diag}(d_1, \ldots, d_m)$ with $d_i \in \mathbb{R}$. This class is quite broad and includes many special cases that are commonly used in practice, such as the shrunken estimator [5], Tikhonov regularization [2], [4] and the principle component method [32].

We now use Theorem 1 to develop conditions on d_i such that the GS estimator of (30) dominates LS in the matrix sense. We will then apply these results to some special cases.

Corollary 2: The GS estimator (30) dominates LS in the matrix sense for all $\|\mathbf{x}\| \leq L$ if and only if

$$\frac{L^2 - \sigma_i}{L^2 + \sigma_i} \leq d_i \leq 1, \quad 1 \leq i \leq m. \tag{31}$$

Proof: We begin by noting that for the GS estimator

$$\mathbf{G} = \mathbf{V}\mathbf{D}\mathbf{V}^*\mathbf{Q}\mathbf{H}^*\mathbf{C}^{-1} = \mathbf{V}\mathbf{D}\Sigma\mathbf{V}^*\mathbf{H}^*\mathbf{C}^{-1}. \tag{32}$$

Therefore, the condition of Theorem 1 becomes

$$\lambda_{\max}(\mathbf{V}(\mathbf{D}^2 - \mathbf{I})\Sigma\mathbf{V}^* + L^2\mathbf{V}(\mathbf{I} - \mathbf{D})^2\mathbf{V}^*) \leq 0 \tag{33}$$

where we used the fact that $\mathbf{G}\mathbf{C}\mathbf{G}^* = \mathbf{V}\mathbf{D}^2\Sigma\mathbf{V}^*$ and $\mathbf{G}\mathbf{H} = \mathbf{V}\mathbf{D}\mathbf{V}^*$. Since for any diagonal matrix $\mathbf{B} = \text{diag}(b_1, \ldots, b_m)$, $\lambda_{\max}(\mathbf{V}\mathbf{B}\mathbf{V}^*) = \max_i b_i$, the condition (33) is equivalent to $\max_i f_i(d_i) \leq 0$ or $f_i(d_i) \leq 0, 1 \leq i \leq m$, where

$$\begin{aligned} f_i(d_i) &= (d_i^2 - 1)\sigma_i + L^2(1 - d_i)^2 \\ &= (d_i - 1)((L^2 + \sigma_i)d_i - L^2 + \sigma_i). \end{aligned} \tag{34}$$

Now, $f_i(d_i)$ is a quadratic convex function with zeros at $d_i^- = (L^2 - \sigma_i)/(L^2 + \sigma_i)$ and $d_i^+ = 1$. Therefore, $f_i(d_i) \leq 0$ for $d_i^- \leq d_i \leq d_i^+$, which proves (31). ■

We now consider some special cases of Corollary 2. In all the examples, $\mathbf{T} = \mathbf{I}$, so that domination results are for $\|\mathbf{x}\| \leq L$.

Example I: A popular estimation strategy for the model (1) is the shrunken estimator [5], which is a scaling of LS:

$$\hat{\mathbf{x}}_{\text{SH}} = \alpha\hat{\mathbf{x}}_{\text{LS}} = \alpha\mathbf{Q}\mathbf{H}^*\mathbf{C}^{-1}\mathbf{y} \tag{35}$$

with $\alpha \in \mathbb{R}$. Clearly $\hat{\mathbf{x}}_{\text{LS}}$ has the form (30) with $\mathbf{D} = \alpha\mathbf{I}$. From Corollary 2, it then follows that $\hat{\mathbf{x}}_{\text{SH}}$ dominates $\hat{\mathbf{x}}_{\text{LS}}$ in the matrix sense if and only if

$$\frac{L^2 - \lambda_{\min}(\mathbf{Q})}{L^2 + \lambda_{\min}(\mathbf{Q})} \leq \alpha < 1 \tag{36}$$

where we used the fact that

$$\max_i \left\{ \frac{L^2 - \sigma_i}{L^2 + \sigma_i} \right\} = \frac{L^2 - \lambda_{\min}(\mathbf{Q})}{L^2 + \lambda_{\min}(\mathbf{Q})}. \tag{37}$$

The MXTM estimator of (7) is a special case of $\hat{\mathbf{x}}_{\text{SH}}$ with $\alpha = L^2/(L^2 + \text{Tr}(\mathbf{Q}))$. This then implies that $\hat{\mathbf{x}}_{\text{MXMM}}$ dominates LS in the matrix sense if and only if

$$\lambda_{\min}(\mathbf{Q}) \geq \frac{L^2\text{Tr}(\mathbf{Q})}{2L^2 + \text{Tr}(\mathbf{Q})}. \tag{38}$$

When $m = 1$, i.e., estimation of a scalar, (38) is always satisfied. However, if $m \geq 2$, then (38) may not hold true, as illustrated for the specific choice $\mathbf{W} = \mathbf{e}^1\mathbf{e}^{1^*}$ in Fig. 1.

Example II: Another popular LS alternative is the Tikhonov regularization [2], [4]

$$\hat{\mathbf{x}}_{\text{TIK}} = (\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H} + \alpha\mathbf{I})^{-1}\mathbf{H}^*\mathbf{C}^{-1}\mathbf{y} \tag{39}$$

which is also of the form (30) with $d_i = 1/(1+\alpha\sigma_i)$. For this estimator, the condition of Corollary 2 becomes

$$\begin{aligned}&\alpha > 0 \text{ or } \alpha \le \frac{2}{L^2-\lambda}, \quad L^2 < \lambda;\\ &0 < \alpha \le \frac{2}{L^2-\lambda}, \quad L^2 \ge \lambda\end{aligned}$$

where $\lambda = \lambda_{\min}(\mathbf{Q})$. In particular, any $0 < \alpha \le 2/L^2$ will result in a LS matrix-dominating estimator regardless of the choice of $\mathbf{Q}$.

Example III: A final example is the principle component estimator, which has the form (30) with

$$d_i = \begin{cases} 1, & \sigma_i \le \gamma \\ 0, & \sigma_i > \gamma \end{cases} \tag{40}$$

where γ is a predefined threshold. This estimator will dominate the LS in a matrix sense if and only if $\sigma_i \ge L^2$ for every i such that $\sigma_i > \gamma$.

IV. Minimax Matrix-MSE Estimator

We now derive the MXMM estimator (25) which minimizes the worst-case weighted MSE over all choices of $\mathbf{x}$ and $\mathbf{W}$ while guaranteeing LS matrix-domination.

Using Theorem 1 we can express (25) in terms of $\mathbf{G}$ as

$$\begin{aligned}&\min_{\mathbf{G}} \max_{\mathbf{x},\mathbf{W}} \{\mathrm{Tr}(\mathbf{W}\mathbf{M}(\mathbf{G})) : \|\mathbf{x}\|_{\mathbf{T}} \le L, \mathbf{0} \preceq \mathbf{W} \preceq \mathbf{I}\}\\ &\text{s.t. } \lambda_{\max}(\Phi(\mathbf{G}) + L^2(\mathbf{I}-\mathbf{G}\mathbf{H})\mathbf{T}^{-1}(\mathbf{I}-\mathbf{G}\mathbf{H})^*) \le 0\end{aligned} \tag{41}$$

where $\mathbf{M}(\mathbf{G})$ is defined by (4) and for brevity we denoted

$$\Phi(\mathbf{G}) = \mathbf{G}\mathbf{C}\mathbf{G}^* - \mathbf{Q}. \tag{42}$$

Since $\mathbf{M}(\mathbf{G}) \succeq \mathbf{0}$, the inner maximization with respect to $\mathbf{W}$ is obtained when $\mathbf{W} = \mathbf{I}$, and (41) reduces to

$$\begin{aligned}&\min_{\mathbf{G}} \{\mathrm{Tr}(\mathbf{G}\mathbf{C}\mathbf{G}^*) + \max_{\|\mathbf{x}\|_{\mathbf{T}} \le L} \mathbf{x}^*(\mathbf{I}-\mathbf{G}\mathbf{H})^*(\mathbf{I}-\mathbf{G}\mathbf{H})\mathbf{x}\}\\ &\text{s.t. } \lambda_{\max}(\Phi(\mathbf{G}) + L^2(\mathbf{I}-\mathbf{G}\mathbf{H})\mathbf{T}^{-1}(\mathbf{I}-\mathbf{G}\mathbf{H})^*) \le 0.\end{aligned} \tag{43}$$

Note that the objective in (43) is the worst-case total MSE over $\|\mathbf{x}\|_{\mathbf{T}} \le L$. However, in contrast to the MXTM estimator of [10] that minimizes this objective, here we have an additional constraint that ensures LS matrix domination.

Theorem 3 below establishes that the MXMM estimator is admissible so that it is not matrix-dominated by any other linear strategy.

Theorem 3: Let $\hat{\mathbf{G}}$ be the solution to (43). Then

1. $\hat{\mathbf{G}}$ is unique;
2. $\hat{\mathbf{G}}$ is admissible in the matrix sense;
3. there exists a $\mathbf{G}$ dominating $\mathbf{G}_{\mathrm{LS}}$ in the matrix sense if and only if $\hat{\mathbf{G}} \ne \mathbf{G}_{\mathrm{LS}}$.

Proof: We prove each of the statements 1–3:

1. Uniqueness follows from strict convexity in $\mathbf{G}$ of the objective in (43) (because $\mathrm{Tr}(\mathbf{G}\mathbf{C}\mathbf{G}^*)$ is strictly convex).
2. Suppose there exists a $\mathbf{G}'$ with $\mathbf{M}(\mathbf{G}') \preceq \mathbf{M}(\hat{\mathbf{G}})$ for all $\|\mathbf{x}\|_{\mathbf{T}} \le L$. Then

$$\mathrm{Tr}(\mathbf{M}(\mathbf{G}')) \le \mathrm{Tr}(\mathbf{M}(\hat{\mathbf{G}})), \quad \text{for all } \|\mathbf{x}\|_{\mathbf{T}} \le L \tag{44}$$

and in addition $\mathbf{G}'$ satisfies the constraint in (43) because $\mathbf{M}(\mathbf{G}') \preceq \mathbf{M}(\mathbf{G}_{\mathrm{LS}})$. Since the objective in (43) is equal to $\mathrm{Tr}(\mathbf{M}(\mathbf{G}))$ and $\hat{\mathbf{G}}$ is the unique minimizer, (44) can hold true only if $\mathbf{G}' = \hat{\mathbf{G}}$ which implies that $\hat{\mathbf{G}}$ is admissible.

3. If $\hat{\mathbf{G}} \ne \mathbf{G}_{\mathrm{LS}}$ then clearly it dominates the LS strategy in the matrix sense since it satisfies the condition of Theorem 1. Conversely, if there exists a $\mathbf{G}$ dominating $\mathbf{G}_{\mathrm{LS}}$ then from (20), $\mathrm{Tr}(\mathbf{M}(\mathbf{G})) \le \mathrm{Tr}(\mathbf{M}(\mathbf{G}_{\mathrm{LS}}))$ for all $\mathbf{x}$. Suppose first that $\mathbf{G} = \hat{\mathbf{G}}$. Since $\hat{\mathbf{G}}$ is the unique minimizer of $\max_{\|\mathbf{x}\|_{\mathbf{T}} \le L} \mathrm{Tr}(\mathbf{M}(\mathbf{G}))$ subject to the domination constraint, we conclude that

$$\max_{\mathbf{x}\in\mathcal{T}} \mathrm{Tr}(\mathbf{M}(\hat{\mathbf{G}})) < \max_{\mathbf{x}\in\mathcal{T}} \mathrm{Tr}(\mathbf{M}(\mathbf{G}_{\mathrm{LS}})) \tag{45}$$

where $\mathcal{T} = \{\mathbf{x} : \|\mathbf{x}\|_{\mathbf{T}} \le L\}$, and $\hat{\mathbf{G}} \ne \mathbf{G}_{\mathrm{LS}}$. If $\mathbf{G} \ne \hat{\mathbf{G}}$, then uniqueness of $\hat{\mathbf{G}}$ implies that

$$\max_{\mathbf{x}\in\mathcal{T}} \mathrm{Tr}(\mathbf{M}(\hat{\mathbf{G}})) < \max_{\mathbf{x}\in\mathcal{T}} \mathrm{Tr}(\mathbf{M}(\mathbf{G})) \le \max_{\mathbf{x}\in\mathcal{T}} \mathrm{Tr}(\mathbf{M}(\mathbf{G}_{\mathrm{LS}})) \tag{46}$$

and again $\hat{\mathbf{G}} \ne \mathbf{G}_{\mathrm{LS}}$.

A. SDP Formulation

Our goal now is to formulate $\hat{\mathbf{x}}_{\mathrm{MXMM}}$ as a solution to an SDP, which is the problem of minimizing a linear function subject to linear matrix inequalities (LMIs). A key element in deriving the LMI representation is Schur's Lemma:

Lemma 1 [33, p. 28]: Let

$$\mathbf{S} = \begin{bmatrix} \mathbf{X} & \mathbf{Y}^* \\ \mathbf{Y} & \mathbf{Z} \end{bmatrix}$$

be a Hermitian matrix with $\mathbf{Z} \succ \mathbf{0}$. Then $\mathbf{S} \succeq \mathbf{0}$ if and only if $\mathbf{X} - \mathbf{Y}^*\mathbf{Z}^{-1}\mathbf{Y} \succeq \mathbf{0}$.

Using the relation

$$\max_{\mathbf{x}^*\mathbf{T}\mathbf{x} \le L^2} \mathbf{x}^*\mathbf{Z}\mathbf{x} = L^2\lambda_{\max}(\mathbf{Z}\mathbf{T}^{-1}) = \min_{\lambda}\{L^2\lambda : \mathbf{Z} \preceq \lambda\mathbf{T}\} \tag{47}$$

for any $\mathbf{Z} \succeq \mathbf{0}$, (43) is equivalent to

$$\begin{aligned}&\min_{\mathbf{G},\lambda} \{\mathrm{Tr}(\mathbf{G}\mathbf{C}\mathbf{G}^*) + L^2\lambda\}\\ &\text{s.t. } (\mathbf{I}-\mathbf{G}\mathbf{H})^*(\mathbf{I}-\mathbf{G}\mathbf{H}) \preceq \lambda\mathbf{T};\\ &\qquad \lambda_{\max}(\Phi(\mathbf{G}) + L^2(\mathbf{I}-\mathbf{G}\mathbf{H})\mathbf{T}^{-1}(\mathbf{I}-\mathbf{G}\mathbf{H})^*) \le 0.\end{aligned} \tag{48}$$

Lemma 2 below asserts that the optimal $\mathbf{G}$ has the form $\hat{\mathbf{G}} = \mathbf{K}(\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}^{-1}$ for an $m \times m$ matrix $\mathbf{K}$, which reduces the dimensionality of the problem when $m < n$. The proof of the Lemma is similar to that of [11] [Lemma 1], and is therefore omitted.

Lemma 2: Let the $m \times n$ matrix $\hat{\mathbf{G}}$ be the solution to (48). Then

$$\hat{\mathbf{G}} = \mathbf{K}(\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}^{-1} = \mathbf{K}\mathbf{Q}\mathbf{H}^*\mathbf{C}^{-1} \tag{49}$$

where $\mathbf{K}$ is the $m \times m$ matrix that is the solution to

$$\begin{aligned}&\min_{\mathbf{K},\lambda}\{\mathrm{Tr}(\mathbf{K}\mathbf{Q}\mathbf{K}^*) + L^2\lambda\}\\ &\text{s.t. } (\mathbf{I}-\mathbf{K})^*(\mathbf{I}-\mathbf{K}) \preceq \lambda\mathbf{T}\\ &\qquad \mathbf{K}\mathbf{Q}\mathbf{K}^* + L^2(\mathbf{I}-\mathbf{K})\mathbf{T}^{-1}(\mathbf{I}-\mathbf{K})^* \preceq \mathbf{Q}.\end{aligned} \tag{50}$$

Our goal now is to convert (50) into a convex SDP so that the solution can be computed efficiently. Defining $\mathbf{X} = \mathbf{K}\mathbf{Q}\mathbf{K}^*$, (50) becomes

$$\begin{aligned}&\min_{\mathbf{K},\mathbf{X},\lambda}\{\mathrm{Tr}(\mathbf{X}) + L^2\lambda\}\\ &\text{s.t. } (\mathbf{I}-\mathbf{K})^*(\mathbf{I}-\mathbf{K}) \preceq \lambda\mathbf{T}\\ &\qquad \mathbf{X} + L^2(\mathbf{I}-\mathbf{K})\mathbf{T}^{-1}(\mathbf{I}-\mathbf{K})^* \preceq \mathbf{Q}\\ &\qquad \mathbf{X} = \mathbf{K}\mathbf{Q}\mathbf{K}^*.\end{aligned} \tag{51}$$

The objective in (51) is linear, and the first two constraints can be converted into LMIs using Schur's Lemma (Lemma 1). The last constraint however is nonconvex. Nonetheless, replacing this equality with the convex constraint $\mathbf{X} \succeq \mathbf{K}\mathbf{Q}\mathbf{K}^*$ does not change the solution. To see this, suppose that the solutions $\hat{\mathbf{X}}$ and $\hat{\mathbf{K}}$ to the relaxed problem satisfy $\hat{\mathbf{X}} \succeq \hat{\mathbf{K}}\mathbf{Q}\hat{\mathbf{K}}^*$ but $\hat{\mathbf{X}} \neq \hat{\mathbf{K}}\mathbf{Q}\hat{\mathbf{K}}^*$. Then $\mathbf{X}' = \hat{\mathbf{K}}\mathbf{Q}\hat{\mathbf{K}}^*$ obeys the constraints in (51) and $\mathrm{Tr}(\mathbf{X}') < \mathrm{Tr}(\hat{\mathbf{X}})$ (here we used the fact that for a matrix $\mathbf{A} \succeq \mathbf{0}$, $\mathrm{Tr}(\mathbf{A}) = 0$ if and only if $\mathbf{A} = \mathbf{0}$) so that $\hat{\mathbf{X}}$ cannot be optimal. Applying Lemma 1 to the resulting convex constraints leads to the following theorem.

Theorem 4: Let $\mathbf{x}$ denote the deterministic unknown parameters in the model $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$, where $\mathbf{H}$ is a known $n \times m$ matrix with rank m, and $\mathbf{n}$ is a zero-mean random vector with covariance $\mathbf{C} \succ \mathbf{0}$. Let $\mathbf{Q} = (\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H})^{-1}$ and denote by $\mathbf{M}(\mathbf{x})$ of (4) the MSE matrix. Then the MXMM estimator which is the solution to

$$\begin{aligned}&\min_{\hat{\mathbf{x}}}\max_{\mathbf{x},\mathbf{W}}\{\mathrm{Tr}(\mathbf{W}\mathbf{M}(\hat{\mathbf{x}})) : \|\mathbf{x}\|_{\mathbf{T}} \le L, \mathbf{0} \preceq \mathbf{W} \preceq \mathbf{I}\}\\ &\text{s.t. } \mathbf{M}(\hat{\mathbf{x}}) \preceq \mathbf{M}(\hat{\mathbf{x}}_{\mathrm{LS}}), \quad \text{for all } \|\mathbf{x}\|_{\mathbf{T}} \le L\end{aligned}$$

is

$$\hat{\mathbf{x}}_{\mathrm{MXMM}} = \mathbf{K}(\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}^{-1}\mathbf{y}$$

where the $m \times m$ matrix $\mathbf{K}$ is a solution to the SDP

$$\begin{aligned}&\min_{\mathbf{K},\mathbf{X},\lambda}\{\mathrm{Tr}(\mathbf{X}) + L^2\lambda\}\\ &\text{s.t. } \begin{bmatrix}\lambda\mathbf{T} & \mathbf{I}-\mathbf{K}\\ (\mathbf{I}-\mathbf{K})^* & \mathbf{I}\end{bmatrix} \succeq \mathbf{0}\\ &\qquad \begin{bmatrix}\mathbf{Q}-\mathbf{X} & \mathbf{I}-\mathbf{K}\\ (\mathbf{I}-\mathbf{K})^* & (1/L^2)\mathbf{T}\end{bmatrix} \succeq \mathbf{0}\\ &\qquad \begin{bmatrix}\mathbf{X} & \mathbf{K}\\ \mathbf{K}^* & \mathbf{Q}^{-1}\end{bmatrix} \succeq \mathbf{0}.\end{aligned} \tag{52}$$

V. Commuting Matrices

We now develop an explicit expression for the MXMM estimate when $\mathbf{T}$ and $\mathbf{Q}$ have the same eigenvector matrix. Thus, if $\mathbf{Q}$ has an eigendecomposition $\mathbf{Q} = \mathbf{V}\Sigma\mathbf{V}^*$ where $\mathbf{V}$ is a unitary matrix and $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_m)$, then $\mathbf{T} = \mathbf{V}\Lambda\mathbf{V}^*$ for some $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_m)$.

Theorem 5: Consider the setting of Theorem 4. Let $\mathbf{Q} = \mathbf{V}\Sigma\mathbf{V}^*$ where $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_m) \succ \mathbf{0}$ and let $\mathbf{T} = \mathbf{V}\Lambda\mathbf{V}^*$ where $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_m) \succ \mathbf{0}$. Then

$$\hat{\mathbf{x}}_{\mathrm{MXMM}} = \mathbf{V}\mathbf{D}\mathbf{V}^*(\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}^{-1}\mathbf{y} \tag{53}$$

where $\mathbf{D} = \mathrm{diag}(d_1, \dots, d_m)$ with

$$d_i = \begin{cases}1 - \sqrt{\beta_0\lambda_i}, & 1 - \sqrt{\beta_0\lambda_i} \ge \alpha_i\\ \alpha_i, & \text{otherwise.}\end{cases} \tag{54}$$

Here

$$\alpha_i = \left[\frac{L^2 - \sigma_i\lambda_i}{L^2 + \sigma_i\lambda_i}\right]_+, \quad 1 \le i \le m \tag{55}$$

with $[a]_+ = a$ if $a \ge 0$ and 0 otherwise, and $\beta_0 \ge 0$ is the unique value of β satisfying $\mathcal{G}(\beta_+) < 0$ and $\mathcal{G}(\beta_-) > 0$ where β_- and β_+ are the values to the right and left of β,

$$\mathcal{G}(\beta) = \sum_{i=1}^{m} \lambda_i\tilde{\mu}_i(\beta) - L^2, \tag{56}$$

and for $1 \le i \le m$,

$$\tilde{\mu}_i(\beta) = \begin{cases}\sigma_i\left(\frac{1}{\sqrt{\beta\lambda_i}} - 1\right), & 1 - \sqrt{\beta\lambda_i} \ge \alpha_i\\ 0, & 1 - \sqrt{\beta\lambda_i} < \alpha_i.\end{cases} \tag{57}$$

Before proving the theorem we discuss how to find β_0. It is easy to see that $0 \le \beta_0 \le \beta_{\mathrm{TH}}$, where

$$\beta_{\mathrm{TH}} = \max_{1 \le i \le m}\left\{\frac{(1-\alpha_i)^2}{\lambda_i}\right\} \tag{58}$$

since for $\beta > \beta_{\mathrm{TH}}$, we have $\tilde{\mu}_i(\beta) = 0, 1 \le i \le m$. We also note that $\mathcal{G}(\beta)$ is a strictly monotonically decreasing function with $\mathcal{G}(\beta) \to \infty$ for $\beta \to 0$ and $\mathcal{G}(\beta) = -L^2$ when $\beta > \beta_{\mathrm{TH}}$. Furthermore, $\mathcal{G}(\beta)$ is continuous at all points $\beta \neq (1-\alpha_i)^2/\lambda_i$. Therefore, there is a unique value β such that $\mathcal{G}(\beta_+) < 0$ and $\mathcal{G}(\beta_-) > 0$ which can be found by using a bisection algorithm on the interval $[0, \beta_{\mathrm{TH}}]$.

Proof: From Lemma 2 the optimal $\mathbf{G}$ has the form (49). Since $\mathbf{V}$ is invertible, we can always express $\mathbf{K}$ of (49) in the form

$$\mathbf{K} = \mathbf{V}\mathbf{D}\mathbf{V}^* \tag{59}$$

for some $m \times m$ matrix $\mathbf{D}$. Next, we show that the optimal $\mathbf{D}$ is a diagonal matrix.

Using representation (59) of $\mathbf{K}$ together with $\mathbf{Q} = \mathbf{V}\Sigma\mathbf{V}^*$, $\mathbf{T} = \mathbf{V}\Lambda\mathbf{V}^*$ and the fact that for any matrix $\mathbf{A}$, $\mathbf{V}\mathbf{A}\mathbf{V}^* \preceq \mathbf{0}$ if and only if $\mathbf{A} \preceq \mathbf{0}$, problem (50) can be written as

$$\begin{aligned}&\min_{\mathbf{D},\lambda}\{\mathrm{Tr}(\mathbf{D}\Sigma\mathbf{D}^*) + L^2\lambda\}\\ &\text{s.t. } (\mathbf{I}-\mathbf{D})^*(\mathbf{I}-\mathbf{D}) \preceq \lambda\Lambda\\ &\qquad \mathbf{D}\Sigma\mathbf{D}^* + L^2(\mathbf{I}-\mathbf{D})\Lambda^{-1}(\mathbf{I}-\mathbf{D})^* \preceq \Sigma.\end{aligned} \tag{60}$$

Let $\mathbf{J}$ be any diagonal matrix with diagonal elements equal to ± 1. If $\mathbf{D}$ satisfies the constraints (60), then so does $\mathbf{J}\mathbf{D}\mathbf{J}$. This follows from the fact that $\mathbf{J}^*\mathbf{J} = \mathbf{J}^2 = \mathbf{I}$ and for any diagonal matrix Σ, $\mathbf{J}\Sigma\mathbf{J} = \Sigma$. Furthermore, the objective value is the same for $\mathbf{D}$ and $\mathbf{JDJ}$. Since (60) is a strictly convex problem in $\mathbf{D}$, the solution is unique, which implies that the optimal value satisfies $\mathbf{D} = \mathbf{JDJ}$ for any $\mathbf{J}$. This can hold true only if $\mathbf{D}$ is a diagonal matrix.

Substituting $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_m)$ into (60), our problem becomes

$$\begin{aligned}&\min_{d_i,\lambda} \left\{ \sum_{i=1}^m d_i^2\sigma_i + L^2\lambda \right\}\\ &\text{s.t. } (1-d_i)^2 \le \lambda\lambda_i, \quad 1 \le i \le m\\ &\qquad d_i^2\sigma_i\lambda_i + L^2(1-d_i)^2 \le \sigma_i\lambda_i, \quad 1 \le i \le m. \end{aligned} \tag{61}$$

We can immediately verify that (61) is a strictly feasible, convex problem (to establish strict feasibility choose λ large enough and $d_i = L^2/(\sigma_i\lambda_i + L^2)$ which minimizes the left-hand side of the second constraint). Therefore, its solution can be determined by solving the dual problem. The Lagrangian associated with (61) is

$$\begin{aligned}\mathcal{L} = \sum_{i=1}^m d_i^2\sigma_i + L^2\lambda + \sum_{i=1}^m \mu_i((1-d_i)^2 - \lambda\lambda_i)\\ + \sum_{i=1}^m \delta_i((d_i^2-1)\sigma_i\lambda_i + L^2(1-d_i)^2).\end{aligned} \tag{62}$$

Differentiating $\mathcal{L}$ with respect to λ and equating to 0,

$$\sum_{i=1}^m \mu_i\lambda_i = L^2. \tag{63}$$

Minimizing with respect to d_i results in

$$d_i = \frac{\mu_i + L^2\delta_i}{\mu_i + L^2\delta_i + \sigma_i(1+\lambda_i\delta_i)}. \tag{64}$$

Substituting (63) and (64) into (62), the Lagrangian becomes

$$\mathcal{L} = \sum_{i=1}^m \left\{ \frac{\sigma_i(1+\delta_i\lambda_i)(\mu_i + L^2\delta_i)}{\mu_i + L^2\delta_i + \sigma_i(1+\lambda_i\delta_i)} - \delta_i\sigma_i\lambda_i \right\} \tag{65}$$

and the dual problem is

$$\begin{aligned}&\max_{\mu_i,\delta_i} \sum_{i=1}^m \left\{ \frac{\sigma_i(1+\delta_i\lambda_i)(\mu_i + L^2\delta_i)}{\mu_i + L^2\delta_i + \sigma_i(1+\lambda_i\delta_i)} - \delta_i\sigma_i\lambda_i \right\}\\ &\text{s.t. } \mu_i \ge 0, \delta_i \ge 0, \quad 1 \le i \le m\\ &\qquad \sum_{i=1}^m \mu_i\lambda_i = L^2. \end{aligned} \tag{66}$$

The dual optimal values are given in the following lemma.

Lemma 3: Let

$$\mu_i(\beta) = \begin{cases} \sigma_i\left(\frac{1}{\sqrt{\beta\lambda_i}} - 1\right), & 1 - \sqrt{\beta\lambda_i} > \alpha_i\\ 0, & 1 - \sqrt{\beta\lambda_i} < \alpha_i\\ x_i, & 1 - \sqrt{\beta\lambda_i} = \alpha_i \end{cases} \tag{67}$$

where x_i is an arbitrary number satisfying $0 \le x_i \le \sigma_i\alpha_i/(1-\alpha_i)$, and let

$$\delta_i(\beta) = \begin{cases} \frac{L^2 - \sigma_i\lambda_i}{\lambda_i(L^2+\sigma_i\lambda_i)}, & 1 - \sqrt{\beta\lambda_i} > \alpha_i\\ 0, & 1 - \sqrt{\beta\lambda_i} < \alpha_i\\ \frac{L^2 - \sigma_i\lambda_i - 2\lambda_i x_i}{\lambda_i(L^2+\sigma_i\lambda_i)}, & 1 - \sqrt{\beta\lambda_i} = \alpha_i. \end{cases} \tag{68}$$

Then the solution to (66) is $\mu_i(\beta_0)$ and $\delta_i(\beta_0)$ where β_0 is the unique root of

$$\mathcal{T}(\beta) = \sum_{i=1}^m \lambda_i\mu_i(\beta) - L^2 \tag{69}$$

with x_i chosen if necessary such that $\mathcal{T}(\beta)$ has a root. ■

Proof: See Appendix I.

At the end of the Proof of Lemma 3 we show that $\mathcal{T}(\beta)$ is monotonically decreasing, continuous at all points $\beta \ne (1-\alpha_i)^2/\lambda_i$ and that x_i in (67) can be chosen such that $\mathcal{T}(\beta)$ has a unique root. Choosing $x_i = \sigma_i\alpha_i/(1-\alpha_i)$, $\mu_i(\beta)$ of (67) is equal to $\tilde{\mu}_i(\beta)$ of (57). It then follows that there is a unique β satisfying $\mathcal{G}(\beta_+) < 0$ and $\mathcal{G}(\beta_-) > 0$ where $\mathcal{G}(\beta)$ is given by (56), and this value is equal to the unique root of $\mathcal{T}(\beta)$. Thus, the optimal β_0 given by Lemma 3 is equal to that defined by the theorem statement.

To complete the proof of the theorem we use the relationship (64) between d_i and μ_i, δ_i given by Lemma 3, which results in (54). ■

A. Comparison Between the MXMM and MXTM Methods

It is interesting to compare between the MXMM and MXTM approaches. For simplicity, we focus here on the case in which $\mathbf{T}$ and $\mathbf{Q}$ can be jointly diagonalized.

The MXTM estimator under this assumption is derived in [10] and has the same form as $\hat{\mathbf{x}}_{\mathrm{MXMM}}$ of (53), where $\mathbf{D} = \mathrm{diag}(\tilde{d}_1, \ldots, \tilde{d}_m)$ with

$$\tilde{d}_i = \begin{cases} 1 - \sqrt{\zeta_0\lambda_i}, & 1 - \sqrt{\zeta_0\lambda_i} \ge 0\\ 0. & \text{otherwise.} \end{cases} \tag{70}$$

Here ζ_0 is the unique value satisfying $\sum_{i=1}^m \eta_i(\zeta_0)\lambda_i = L^2$ with

$$\eta_i(\zeta) = \begin{cases} \sigma_i\left(\frac{1}{\sqrt{\zeta\lambda_i}} - 1\right), & 1 - \sqrt{\zeta\lambda_i} > 0\\ 0, & 1 - \sqrt{\zeta\lambda_i} \le 0. \end{cases} \tag{71}$$

Let the eigenvalues of $\mathbf{T}$ be sorted in decreasing order such that $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_m$ (note that this will change the order of the eigenvectors in $\mathbf{V}$ which in turn will permute the values of σ_i). With this ordering,

$$\sqrt{\zeta_0} = \frac{\sum_{i=k+1}^m \sqrt{\lambda_i}\sigma_i}{L^2 + \sum_{i=k+1}^m \lambda_i\sigma_i} \tag{72}$$

where k is the smallest index such that $0 \le k \le m-1$ and $\sqrt{\zeta_0\lambda_{k+1}} < 1$.

Comparing with the MXMM estimate of Theorem 5 leads to the following result.

Theorem 6: Consider the problem of Theorem 5. Let $\mathbf{V}, \Sigma$ and Λ be ordered such that $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_m$. Then the MXMM and MXTM estimators both have the form (53) with $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_m)$ for the MXMM estimate and $\mathbf{D} = \mathrm{diag}(\tilde{d}_1, \ldots, \tilde{d}_m)$ for the MXTM estimate where

$$d_i \ge \tilde{d}_i, \quad 1 \le i \le m. \tag{73}$$

Furthermore, the estimators coincide if

$$\begin{aligned} &1-\sqrt{\zeta_0\lambda_i}\geq \alpha_i, && k+1\leq i\leq m;\\ &\alpha_i=0, && 1\leq i\leq k \end{aligned} \tag{74}$$

where α_i is defined by (55) and ζ_0 is given by (72) with $0\leq k\leq m-1$ the smallest index such that $\sqrt{\zeta_0\lambda_{k+1}}<1$. In particular, if $L^2\leq \sigma_i\lambda_i$, $1\leq i\leq m$, then the MXMM and MXTM methods are equivalent.

Proof: See Appendix II. ■

Note that both the MXMM and MXTM estimators are generalized shrinkage estimators of the form (30) with shrinkage factors d_i and $\tilde{d}_i$ satisfying (73). Evidently, the shrinkage of the eigenvalues in the MXTM estimate is larger than that of the MXMM method. Thus, larger shrinkage can decrease the total MSE at the expense of increasing the MSE of some components.

B. The Case $\mathbf{T}=\mathbf{I}$

Using the results of Theorem 5 we now treat the special case in which $\mathbf{T}=\mathbf{I}$.

Corollary 7: Consider the setting of Theorem 5 with $\mathbf{T}=\mathbf{I}$. Let the eigendecomposition of $\mathbf{Q}$ be given by $\mathbf{Q}=\mathbf{V}\Sigma\mathbf{V}^*$ where $\Sigma=\operatorname{diag}(\sigma_1,\dots,\sigma_m)$ with σ_i sorted in decreasing order: $\sigma_1\geq\sigma_2\geq\cdots\geq\sigma_m$. Then the optimal values of d_i are given as follows. Let $1\leq k\leq m$ be the largest value satisfying

$$\gamma_k\triangleq\frac{L^2}{L^2+\sum_{i=1}^k\sigma_i}\geq\alpha_k \tag{75}$$

where α_i is defined by (55). If

$$\gamma_k<\alpha_i,\quad \text{for } i>k \tag{76}$$

then $1-\sqrt{\beta}=\gamma_k$ and

$$d_i=\begin{cases}\frac{L^2}{L^2+\sum_{i=1}^k\sigma_i}, & 1\leq i\leq k\\ \alpha_i, & i\geq k+1.\end{cases} \tag{77}$$

Otherwise

$$d_i=\begin{cases}\alpha_{k+1}, & 1\leq i\leq k+1\\ \alpha_i, & i>k+1.\end{cases} \tag{78}$$

Proof: See Appendix III. ■

Note that when $\mathbf{T}=\mathbf{I}$, k in (72) is 0 and

$$\sqrt{\zeta_0}=\frac{\sum_{i=1}^m\sigma_i}{L^2+\sum_{i=1}^m\sigma_i}. \tag{79}$$

From Theorem 6, it then follows that the MXMM and MXTM methods coincide if

$$\gamma_m=\frac{L^2}{L^2+\sum_{i=1}^m\sigma_i}\geq\alpha_i,\quad 1\leq i\leq m \tag{80}$$

or equivalently, using the ordering (105), $\gamma_m\geq\alpha_m$. Substituting the expressions for γ_m and α_m this condition becomes

$$L^2\leq\frac{\sigma_m\sum_{i=1}^m\sigma_i}{\sum_{i=1}^m\sigma_i-2\sigma_m}. \tag{81}$$

An interesting special case is when $\mathbf{Q}=a\mathbf{I}$ for some $a>0$. Since $\alpha_k=(L^2-a)/(L^2+a)$ is independent of k, γ_k of (75) cannot satisfy (76). Therefore, either $k=m$, in which case the MXMM estimate is equal to the MXTM approach, or $d_i=\alpha$. Now, $k=m$ if (81) is satisfied, resulting in

$$a\geq\left(1-\frac{2}{m}\right)L^2\triangleq a_0. \tag{82}$$

Thus, the MXMM estimate can be written in this case as

$$\hat{\mathbf{x}}_{\mathrm{MXMM}}=\begin{cases}a\frac{L^2}{L^2+ma}\mathbf{H}^*\mathbf{C}^{-1}y, & a\geq a_0\\ a\frac{L^2-a}{L^2+a}\mathbf{H}^*\mathbf{C}^{-1}y, & a\leq a_0.\end{cases} \tag{83}$$

VI. Examples

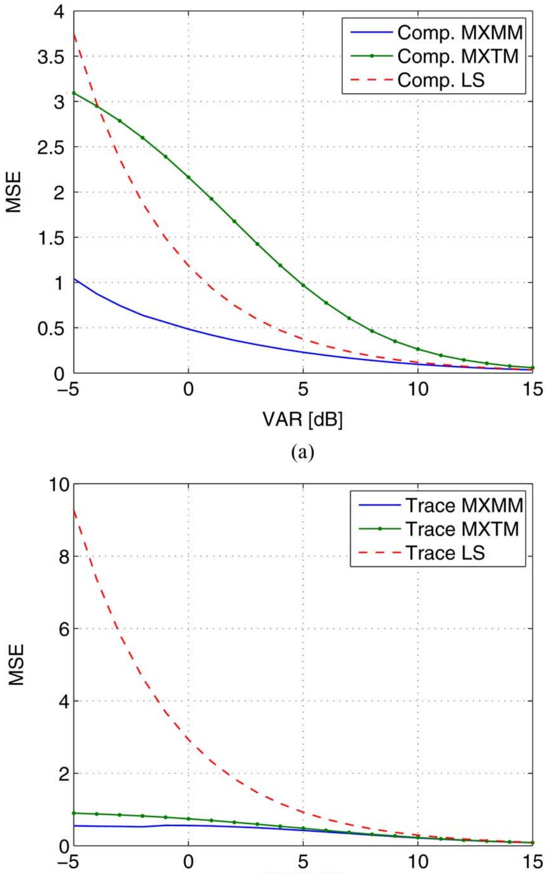
In this section, we compare the MSE performance of the MXTM, the proposed MXMM, and the LS methods. We consider two measures of MSE: Trace MSE and the MSE of the 1st component.

In all the examples we assume that $L=\|\mathbf{x}\|$. In practice, the norm of $\mathbf{x}$ may not be known exactly. Instead, we may have a bound on the norm that can replace the true norm value. Alternatively, as suggested and studied in [21], we can replace L by the norm of the LS estimate: $L=\|\mathbf{x}_{\mathrm{LS}}\|$ which corresponds to the choice $\mathbf{T}=\mathbf{I}$. As another approach, we may choose $L^2=\mathbf{x}_{\mathrm{LS}}^*(\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{x}_{\mathrm{LS}}$ which corresponds to $\mathbf{T}=(\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H})^{-1}$ (see [21] for details).

Example I: In the first example, we generate a random model matrix $\mathbf{H}$ with $n=7, m=5$ and a random vector $\mathbf{x}$. The noise is assumed to be white, $\mathbf{T}=\mathbf{I}$ and $L=\|\mathbf{x}\|$. In Fig. 2, we plot the MSE as a function of the noise variance (in dB) for the MXMM, MXTM and LS estimators. In this example, $L^2\approx 5$. The MSE of the first component is plotted in Fig. 2(a), and the trace MSE divided by m in Fig. 2(b). Interestingly, the trace MSE of the MXMM and MXTM methods are very similar, while the MSE of the 1st component is much lower using the MXMM approach. Note that the MXTM estimator is only guaranteed to have smaller total MSE for the worst-case $\mathbf{x}$, so that it is possible, as we see in the figure, to achieve lower total MSE with the MXMM strategy for other choices of $\mathbf{x}$. It is also evident from the figures that the MXMM method dominates LS in terms of both trace and componentwise MSE, while the MXTM approach dominates LS only in trace MSE. In Fig. 3, we plot the MSE of the third component. Here we see that the MXMM and MXTM approaches lead to comparable performance.

In Fig. 4, we repeat the simulations leading to Fig. 2(a), but now instead of using $L=\|\mathbf{x}\|$, we estimate L from the data as the norm of the LS method. Evidently, even though the value used is now not a true bound on the signal norm, since $\|\mathbf{x}_{\mathrm{LS}}\|$ can be smaller than $\|\mathbf{x}\|$, still the MXMM approach leads to improved performance.

The behavior in Figs. 2 and 3 seems to be representative of the performance in random models. In simulations we observed that often the trace behavior of the MXMM and MXTM methods is similar. In contrast, the componentwise performance of MXMM is typically much better for some of the components, while for others the behavior of both the MXMM and



(a)

(b)

Fig. 2. MSE in estimating $\mathbf{x}$ as a function of the noise variance using the MXTM, MXMM and LS estimators: (a) MSE of the first component; (b) total MSE.

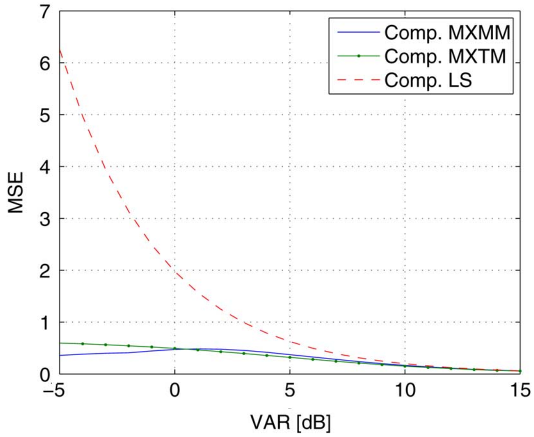


Fig. 3. MSE of the third component when estimating $\mathbf{x}$ as a function of the noise variance using the MXTM, MXMM, and LS estimators.

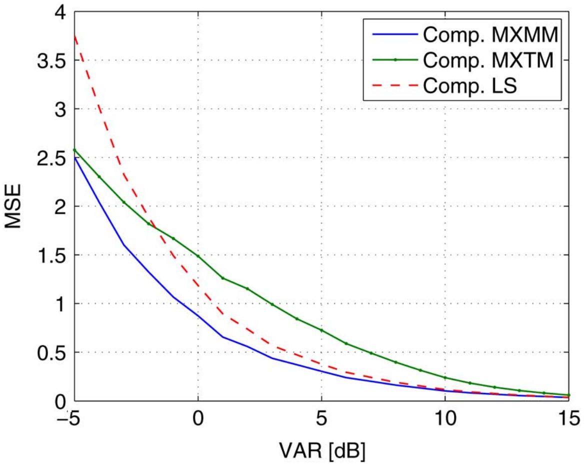


Fig. 4. MSE in the first component when estimating $\mathbf{x}$ as a function of the noise variance using the MXTM, MXMM, and LS estimators with estimated bound $L = \|\mathbf{x}_{\mathrm{LS}}\|$.

MXTM estimators is comparable, so that overall the MXMM leads to better componentwise behavior. Thus, it seems like the MXMM approach can substantially decrease the weighted MSE with only a small increase in the trace MSE with respect to the MXTM estimator.

Example II: This class of examples is taken from the Regularization Tools [34] for Matlab. All the problems in this toolbox are discretized versions of the Fredholm integral equation of the first kind:

$$g(s) = \int_a^b K(s,t) f(t)\, dt \tag{84}$$

where $K(s,t)$ is the kernel and $f(t)$ is the solution for a given $g(s)$. The problem is to estimate $f(t)$ from noisy samples of $g(s)$. Using a midpoint rule with n points, (84) reduces to an $n \times n$ linear system $\mathbf{y}_T = \mathbf{H}\mathbf{x}$. The functions in this toolbox differ in $K(t,s)$ and $f(s)$. Below we consider two choices. In both cases $n = 256$, the observations are $\mathbf{y} = \mathbf{y}_T + \mathbf{w}$ where $\mathbf{w}$ is a white Gaussian noise vector with standard deviation $\sigma = 0.1$, $L = \|\mathbf{x}\|$ and we use a weighting $\mathbf{T} = (\mathbf{H}^*\mathbf{H})^{-1.5}$. This choice of $\mathbf{T}$ reflects the fact that components of $\mathbf{x}$ corresponding to small eigenvalues of $\mathbf{H}^*\mathbf{H}$ should receive a smaller weight than components corresponding to large eigenvalues.

First we implement the function $\mathsf{Shaw(n)}$ which corresponds to the kernel

$$K(s,t) = (\cos(s) + \cos(t)) \frac{\sin^2(\pi(\sin(s) + \sin(t)))}{\pi^2(\sin(s) + \sin(t))^2} \tag{85}$$

with integration over $[-\pi/2, \pi/2]$. The output of the function is the matrix $\mathbf{H}$ and the true vector $\mathbf{x}$ (which represents $f(t)$). The original signal along with the estimates using the MXMM and MXTM methods are plotted in Fig. 5. The LS estimate is not given since the results are very poor.

In Fig. 6 we show the results using the function $\mathsf{Phillips(n)}$ corresponding to the kernel

$$K(s,t) = \begin{cases} 1 + \cos((s-t)\pi/3), & |s-t| < 3 \\ 0, & |s-t| \geq 3 \end{cases} \tag{86}$$

with integration over $[-6, 6]$. Here again the estimate using the LS approach is poor and is therefore omitted.

In both figures we see that the MXMM method provides a better approximation of the original signal. This can be determined visually and from the resulting total MSEs, which are summarized in Table I.

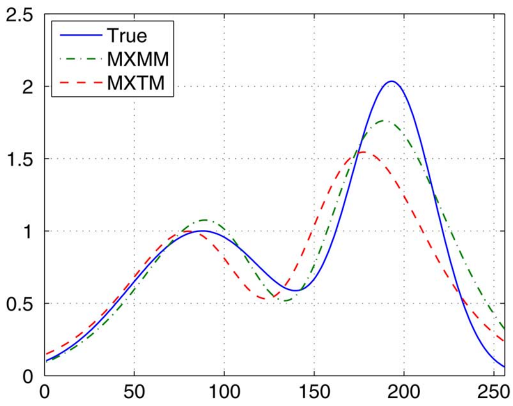


Fig. 5. True signal and its estimates using the MXMM and MXTM method for the Shaw problem.

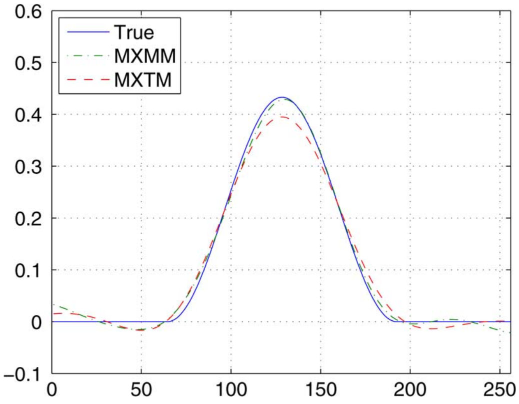


Fig. 6. True signal and its estimates using the MXMM and MXTM method for the Phillips problem.

TABLE I
TOTAL MSE

	shaw	phillips
MXTM	4.295	0.289
MXMM	2.405	0.175

APPENDIX I PROOF OF LEMMA 3

To solve (66), we form the Lagrangian

$$\mathcal{L} = \sum_{i=1}^{m} \left(\delta_i \sigma_i \lambda_i - \frac{\sigma_i(1+\delta_i\lambda_i)(\mu_i + L^2\delta_i)}{\mu_i + L^2\delta_i + \sigma_i(1+\lambda_i\delta_i)} \right) - \sum_{i=1}^{m} \gamma_i \mu_i - \sum_{i=1}^{m} \xi_i \delta_i + \beta \sum_{i=1}^{m} \mu_i \lambda_i. \quad (87)$$

Differentiating with respect to μ_i, δ_i and equating to 0,

$$\frac{\sigma_i^2(1+\delta_i\lambda_i)^2}{(\mu_i + L^2\delta_i + \sigma_i(1+\lambda_i\delta_i))^2} + \gamma_i = \beta\lambda_i \quad (88)$$

$$\frac{(\mu_i + L^2\delta_i)^2\lambda_i + \sigma_i L^2(1+\delta_i\lambda_i)^2}{(\mu_i + L^2\delta_i + \sigma_i(1+\lambda_i\delta_i))^2} + \epsilon_i = \lambda_i, \quad (89)$$

for $1 \le i \le m$, where $\epsilon_i = \xi_i/\sigma_i$. In addition, the Lagrange multipliers γ_i, ϵ_i must be nonnegative and satisfy the complementary slackness conditions

$$\gamma_i \mu_i = 0, \quad \gamma_i \ge 0, \quad 1 \le i \le m \quad (90)$$

$$\epsilon_i \delta_i = 0, \quad \epsilon_i \ge 0, \quad 1 \le i \le m. \quad (91)$$

Suppose first that $\beta\lambda_i > 1$. Since the first expression in (88) is always smaller or equal to 1, we must have $\gamma_i > 0$ which implies from (90) that $\mu_i = 0$. Substituting into (89)

$$\frac{L^4\delta_i^2\lambda_i + \sigma_i L^2(1+\lambda_i\delta_i)^2}{(L^2\delta_i + \sigma_i(1+\lambda_i\delta_i))^2} + \epsilon_i = \lambda_i. \quad (92)$$

If $\delta_i = 0$, then to satisfy (92) with some $\epsilon_i \ge 0$ we must have $\sigma_i\lambda_i \ge L^2$. Otherwise $\delta_i > 0$, in which case from (91), $\epsilon_i = 0$. Substituting into (92)

$$\delta_i = \frac{L^2 - \sigma_i\lambda_i}{\lambda_i(L^2 + \sigma_i\lambda_i)} \quad (93)$$

which satisfies $\delta_i > 0$ as long as $\sigma_i\lambda_i < L^2$. Thus, we conclude that for $1 \le i \le m$,

$$\beta\lambda_i > 1 \Rightarrow \begin{cases} \mu_i = 0 \\ \delta_i = \begin{cases} \frac{L^2 - \sigma_i\lambda_i}{\lambda_i(L^2+\sigma_i\lambda_i)}, & \sigma_i\lambda_i < L^2 \\ 0, & \sigma_i\lambda_i \ge L^2. \end{cases} \end{cases} \quad (94)$$

Next, we consider the setting $\beta\lambda_i \le 1$. We first note that from (88), $\beta \ge 0$. Furthermore, either $\gamma_i = 0$ or $\epsilon_i = 0$. To see this, suppose to the contrary that $\gamma_i > 0$ and $\epsilon_i > 0$. Then from (90) and (91), $\mu_i = \delta_i = 0$. Substituting into (88),

$$\gamma_i = \beta\lambda_i - 1 \le 0 \quad (95)$$

since $\beta\lambda_i \le 1$, which contradicts the assumption $\gamma_i > 0$.

Suppose first that $\gamma_i > 0$. Then $\mu_i = \epsilon_i = 0$ and δ_i is given by (93). This solution is valid only if $\delta_i \ge 0$, and (88) is satisfied for some $\gamma_i > 0$, $\beta\lambda_i \le 1$. The first condition is equivalent to $\sigma_i\lambda_i \le L^2$. The second constraint translates into

$$1 \ge \sqrt{\beta\lambda_i} > \frac{2\sigma_i\lambda_i}{L^2 + \sigma_i\lambda_i} \triangleq t_i. \quad (96)$$

Note that in particular, (96) implies $\sigma_i\lambda_i \le L^2$ since for $\sigma_i\lambda_i > L^2$, $2\sigma_i\lambda_i > L^2 + \sigma_i\lambda_i$. Thus, if (96) holds, then δ_i is given by (93) and $\mu_i = 0$.

Next, let $\epsilon_i > 0$, which implies $\delta_i = \gamma_i = 0$. From (88),

$$\mu_i = \sigma_i \left(\frac{1}{\sqrt{\beta\lambda_i}} - 1 \right). \quad (97)$$

The solution (97) is always nonnegative since $\beta\lambda_i \le 1$. To satisfy (89) for some $\epsilon_i > 0$ we must have $\sigma_i\lambda_i > L^2 - 2\mu_i\lambda_i$. Substituting μ_i from (97), the condition becomes $\sqrt{\beta\lambda_i} < t_i$ where t_i is defined by (96). Thus, (97) is valid if $\beta\lambda_i \le 1$ and $\sqrt{\beta\lambda_i} < t_i$.

Finally, suppose that $\gamma_i = \epsilon_i = 0$. By simple algebraic manipulations, it can be shown that (88) and (89) are satisfied in this case only if $\sqrt{\beta\lambda_i} = t_i$. Then

$$\mu_i = \frac{1}{2\lambda_i}(1+\delta_i\lambda_i)(L^2 - \lambda_i\sigma_i) - L^2\delta_i \triangleq x_i \quad (98)$$

and $\delta_i \geq 0$ is any value that ensures $\mu_i \geq 0$ with μ_i given by (98), or

$$0 \leq \delta_i \leq \frac{L^2 - \sigma_i \lambda_i}{\lambda_i(L^2 + \sigma_i \lambda_i)}. \tag{99}$$

Since $t_i = \sqrt{\beta \lambda_i} \leq 1$, $L^2 \geq \sigma_i \lambda_i$ and the bound in (99) is nonnegative. Substituting the constraints (99) into (98) we obtain that

$$0 \leq x_i \leq \frac{L^2 - \sigma_i \lambda_i}{2\lambda_i} = \frac{\sigma_i(1 - t_i)}{t_i}. \tag{100}$$

The relation (98) then implies that

$$\delta_i = \frac{L^2 - \sigma_i \lambda_i - 2\lambda_i x_i}{\lambda_i(L^2 + \sigma_i \lambda_i)}. \tag{101}$$

Summarizing our discussion so far, we have shown that

$$\mu_i(\beta) = \begin{cases} \sigma_i \left(\frac{1}{\sqrt{\beta \lambda_i}} - 1 \right), & \sqrt{\beta \lambda_i} < t_i, \sqrt{\beta \lambda_i} \leq 1 \\ x_i, & \sqrt{\beta \lambda_i} = t_i, \sqrt{\beta \lambda_i} \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{102}$$

where x_i satisfies (100), and

$$\delta_i(\beta) = \begin{cases} \frac{L^2 - \sigma_i \lambda_i}{\lambda_i(L^2 + \sigma_i \lambda_i)}, & \sqrt{\beta \lambda_i} > t_i, t_i < 1 \\ \frac{L^2 - \sigma_i \lambda_i - 2\lambda_i x_i}{\lambda_i(L^2 + \sigma_i \lambda_i)}, & \sqrt{\beta \lambda_i} = t_i, \sqrt{\beta \lambda_i} \leq 1 \\ 0, & \text{otherwise.} \end{cases} \tag{103}$$

Since $t_i \leq 1$ is equivalent to $L^2 \geq \sigma_i \lambda_i$, when $t_i \leq 1$, α_i of (55) can be written as $\alpha_i = 1 - t_i$. The condition $\sqrt{\beta \lambda_i} < (=)t_i, \sqrt{\beta \lambda_i} \leq 1$ is therefore equivalent to $1 - \sqrt{\beta \lambda_i} > (=)\alpha_i$. When $t_i > 1$, $\mu_i(\beta) = 0$ if $\sqrt{\beta \lambda_i} > 1$ which is equivalent to $1 - \sqrt{\beta \lambda_i} < 0 = \alpha_i$ since $t_i > 1$ implies $\alpha_i = 0$. If $\sqrt{\beta \lambda_i} = t_i \leq 1$, then the upper bound in (100) becomes $\sigma_i \alpha_i (1 - \alpha_i)$. On the other hand, when $\sqrt{\beta \lambda_i} = t_i \geq 1$, $\mu_i(\beta) = 0$ which is consistent with (67) since $\alpha_i = 0$ and the upper bound on x_i is 0. Thus, (102) and (103) are equivalent to (67) and (68).

To complete the proof it remains to determine the optimal value of β. This can be accomplished by enforcing the last constraint in (66) which implies that β must be a root of $\mathcal{T}(\beta)$ given by (69). Since $\mathcal{T}(\beta)$ is monotonically decreasing on $0 < \beta \leq \beta_{\text{TH}}$ where β_{TH} is defined by (58), the root is unique. To show that there is always a value β such that $\mathcal{T}(\beta) = 0$ note that $\mathcal{T}(\beta)$ is continuous for all $\beta \neq t_i^2/\lambda_i$. At these points, x_i can be chosen such that $\mathcal{T}(\beta)$ can take on any value between $\mathcal{T}(\beta_-)$ and $\mathcal{T}(\beta_+)$ where β_- and β_+ are the values of β to the right and left of the point of discontinuity. This follows from the fact that for $\sqrt{\beta \lambda_i} = t_i, 0 \leq \mu_i(\beta) \leq (L^2 - \lambda_i \sigma_i)/(2\lambda_i) = \sigma_i(1/t_i - 1)$. In addition, $\mathcal{T}(\beta) > 0$ for $\beta \to 0$, and $\mathcal{T}(\beta) = -L^2$ for $\beta \geq \beta_{\text{TH}}$. Therefore, either $\mathcal{T}(\beta) = 0$ for a value β for which $\mathcal{T}(\beta)$ is continuous, or that we can choose x_i at the discontinuity points such that $\mathcal{T}(\beta) = 0$.

Appendix II
Proof of Theorem 6

To prove (73) we note that $d_i(\beta)$ and $\tilde{d}_i(\zeta)$ are both monotonically decreasing until they reach a constant value: α_i for $d_i(\beta)$ where $\alpha_i \geq 0$, and 0 for $\tilde{d}_i(\zeta)$. From the definitions (57), (71) of $\tilde{\mu}_i, \eta_i$ it follows that they are both monotonically decreasing and that for a given β, $\eta_i(\beta) \geq \tilde{\mu}_i(\beta)$. Now, at the optimal values of ζ and β, $\sum_i \lambda_i \eta_i(\zeta_0) = \sum_i \lambda_i \tilde{\mu}_i(\beta_0) = L^2$. Therefore, the optimal choices of ζ and β satisfy $\zeta_0 \geq \beta_0$, from which we conclude that $\tilde{d}_i \leq d_i$.

To prove the second part, suppose that (74) is satisfied. By definition, $\sum_{i=1}^m \lambda_i \eta_i(\zeta_0) = L^2$ with

$$\eta_i(\zeta_0) = \begin{cases} \sigma_i \left(\frac{1}{\sqrt{\zeta_0 \lambda_i}} - 1 \right), & k+1 \leq i \leq m \\ 0, & 1 \leq i \leq k. \end{cases} \tag{104}$$

Now, from (57) and (74), $\tilde{\mu}_i(\zeta_0) = \eta_i(\zeta_0)$. This follows from the fact that for $1 \leq i \leq k$, $\sqrt{\zeta_0 \lambda_i} \geq 1$ and $\alpha_i = 0$. Thus, $\sum_{i=1}^m \lambda_i \tilde{\mu}_i(\zeta_0) = L^2$ and ζ_0 is optimal for the MXMM estimator. Substituting ζ_0 into (54) and using $\alpha_i = 0, 1 \leq i \leq k$ we have $d_i = \tilde{d}_i$ of (70).

Finally, if $L^2 \leq \sigma_i, 1 \leq i \leq m$, then $\alpha_i = 0, 1 \leq i \leq m$. Since by definition, $1 - \sqrt{\zeta_0 \lambda_i} > 0, k+1 \leq i \leq m$, in this case (74) is satisfied.

Appendix III
Proof of Corollary 7

Since the eigenvalues σ_i of are sorted in decreasing order and $\lambda_i = 1, 1 \leq i \leq m$, it is easy to see that the threshold values α_i of (55) increase with i:

$$\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_m. \tag{105}$$

Therefore, if for some value k, $1 - \sqrt{\beta} \geq \alpha_k$, then $1 - \sqrt{\beta} \geq \alpha_i$ for all $1 \leq i \leq k$.

As indicated after the statement of Theorem 5, we can find β by first checking whether $\mathcal{G}(\beta)$ has a root. To this end we need to determine $\tilde{\mu}_i(\beta)$ of (57). Let k be the largest index such that $1 - \sqrt{\beta} \geq \alpha_k$. Then

$$\tilde{\mu}_i(\beta) = \begin{cases} \sigma_i \left(\frac{1}{\sqrt{\beta}} - 1 \right), & 1 \leq i \leq k \\ 0, & i > k. \end{cases} \tag{106}$$

Thus, β is a root of $\mathcal{G}(\beta)$ if $\sum_{i=1}^m \tilde{\mu}_i(\beta) = L^2$, or

$$1 - \sqrt{\beta} = \frac{L^2}{L^2 + \sum_{i=1}^k \sigma_i} = \gamma_k. \tag{107}$$

The solution (107) is valid only if for this choice

$$\begin{aligned} \gamma_k &\geq \alpha_i, \quad 1 \leq i \leq k \\ \gamma_k &< \alpha_i, \quad i > k. \end{aligned} \tag{108}$$

Since γ_k is monotonically decreasing in k, $\gamma_j < \gamma_k$ if $j < k$. Therefore, using the ordering (105), instead of having to check (108) for all k, it is sufficient to consider the largest value k for which $\gamma_k \geq \alpha_k$, which is equivalent to (76). Substituting (107) into (54) leads to (77).

If (76) is not satisfied, then the optimal β is of the form $1 - \sqrt{\beta} = \alpha_\kappa$ for κ such that $\mathcal{G}(\beta_-) > 0$ and $\mathcal{G}(\beta_+) > 0$. We now show that $\kappa = k+1$ so that $\sqrt{\beta_0} = 1 - \alpha_{k+1}$.

Since k is the largest index for which $\gamma_k \geq \alpha_k$, we have that $\gamma_{k+1} < \alpha_{k+1}$. Therefore

$$\begin{aligned}\sum_{i=1}^{m} \tilde{\mu}_i((\beta_0)_-) &\geq \sum_{i=1}^{m} \tilde{\mu}_i(\beta_0) \\ &= \frac{\alpha_{k+1} \sum_{i=1}^{k+1} \sigma_i}{1-\alpha_{k+1}} \\ &> \frac{\gamma_{k+1}}{1-\gamma_{k+1}} \sum_{i=1}^{k+1} \sigma_i = L^2 \end{aligned} \tag{109}$$

and $\mathcal{G}((\beta_0)_-) > 0$. Next, we note that $\gamma_k \geq \alpha_i$ for some $i > k$. Using the ordering (105), this implies that $\gamma_k \geq \alpha_{k+1}$. Thus,

$$\sum_{i=1}^{m} \tilde{\mu}_i((\beta_0)_+) < \frac{\alpha_{k+1} \sum_{i=1}^{k} \sigma_i}{1-\alpha_{k+1}} \leq \frac{\gamma_k}{1-\gamma_k} \sum_{i=1}^{k} \sigma_i = L^2 \tag{110}$$

and $\mathcal{G}((\beta_0)_+) < 0$, completing the proof.

Acknowledgment

The author would like to thank the reviewers for their careful comments which improved the presentation in several places.

References

[1] K. F. Gauss, Theoria Combinationis Obsercationunt Erronbus Minimis Obnoxiae 1821.

[2] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, Feb. 1970.

[3] D. W. Marquardt, "Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation," *Technometrics*, vol. 12, no. 3, pp. 592–612, Aug. 1970.

[4] A. N. Tikhonov and V. Y. Arsenin, *Solution of Ill-Posed Problems*. Washington, DC: V. H. Winston, 1977.

[5] L. S. Mayer and T. A. Willke, "On biased estimation in linear models," *Technometrics*, vol. 15, pp. 497–508, Aug. 1973.

[6] Y. C. Eldar and A. V. Oppenheim, "Covariance shaping least-squares estimation," *IEEE Trans. Signal Process.*, vol. 51, no. 3, pp. 686–697, Mar. 2003.

[7] Y. C. Eldar, A. Ben-Tal, and A. Nemirovski, "Linear minimax regret estimation of deterministic parameters with bounded data uncertainties," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2177–2188, Aug. 2004.

[8] B. Efron, "Biased versus unbiased estimation," *Adv. Math.*, vol. 16, pp. 259–277, 1975.

[9] M. S. Pinsker, "Optimal filtering of square-integrable signals in Gaussian noise," *Problems Inf. Trans.*, vol. 16, pp. 120–133, 1980.

[10] 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.

[11] 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.

[12] Z. Ben-Haim and Y. C. Eldar, "Maximum set estimators with bounded estimation error," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3172–3182, Aug. 2005.

[13] Y. C. Eldar, "Uniformly improving the Cramér-Rao bound and maximum-likelihood estimation," *IEEE Trans. Signal Process.*, vol. 54, no. 8, pp. 2943–2956, Aug. 2006.

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

[15] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," in *Proc. 3rd Berkeley Symp. Math. Statist. Prob.*, Univ. California Press, Berkeley, CA, 1956, vol. 1, pp. 197–206.

[16] W. James and C. Stein, "Estimation of quadratic loss," in *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, Univ. California Press, Berkeley, 1961, vol. 1, pp. 361–379, .

[17] W. E. Strawderman, "Proper Bayes minimax estimators of multivariate normal mean," *Ann. Math. Statist.*, vol. 42, pp. 385–388, 1971.

[18] K. Alam, "A family of admissible minimax estimators of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 1, pp. 517–525, 1973.

[19] 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.

[20] Z. Ben-Haim and Y. C. Eldar, "Blind minimax estimators: Improving on least squares estimation," presented at the IEEE Workshop on Statistical Signal Processing (SSP'05), Bordeaux, France, Jul. 2005.

[21] Z. Ben-Haim and Y. C. Eldar, "Blind minimax estimation," *IEEE Trans. Inf. Theory*, vol. 53, no. 9, pp. 3145–3157, Sep. 2007.

[22] A. J. Baranchik, "Multiple regression and estimation of the mean of a multivariate normal distribution," Stanford Univ., Stanford, CA, Tech. Rep. 51, 1964.

[23] B. Efron and C. Morris, "Limiting the risk of Bayes and empirical Bayes estimators—Part II: The empirical Bayes case," *J. Amer. Stat. Assoc.*, vol. 67, no. 337, pp. 130–139, Mar. 1972.

[24] T. Teräsvirta, "Superiority comparisons of homogeneous linear estimators," *Commun. Statist.—Theor. Meth.*, vol. 11, no. 14, pp. 1595–1601, 1982.

[25] G. Trenkler and H. Toutenburg, "Mean squared error matrix comparisons between biased estimators—An overview of recent results," *Stat. Papers*, vol. 31, pp. 165–179, 1990.

[26] K. Alam and A. Mitra, "Component risk in multiparameter estimation," *Ann. Inst. Statist. Math.*, pp. 339–410, 1986.

[27] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Rev.*, vol. 38, no. 1, pp. 40–95, Mar. 1996.

[28] Y. Nesterov and A. Nemirovski, *Interior-Point Polynomial Algorithms in Convex Programming*. Philadelphia, PA: SIAM, 1994.

[29] K. Hoffmann, "Admissibility of linear estimation with respect to restricted parameter sets," *Math. Oper. Statist. Ser. Statist.*, vol. 8, pp. 425–438, 1977.

[30] C. M. Theobald, "Generalizations of mean square error applied to ridge regression," *J. R. Statist. Soc. B*, vol. 36, pp. 103–106, 1974.

[31] R. L. Obenchain, "Ridge analysis following a preliminary test of the shrunken hypothesis," *Technometrics*, vol. 17, no. 4, pp. 431–441, Nov. 1975.

[32] W. F. Massy, "Principle component regression in exploratory statistical research," *J. Amer. Statist. Assoc.*, vol. 60, pp. 234–256, 1965.

[33] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*. Philadelphia, PA: SIAM, 1994.

[34] P. C. Hansen, "Regularization tools, a Matlab package for analysis of discrete regularization problems," *Numer. Algorithms*, vol. 6, pp. 1–35, 1994.

Yonina C. Eldar (S'98–M'02–SM'07) received the B.Sc. degree in physics and the B.Sc. degree in electrical engineering both from Tel-Aviv University (TAU), Tel-Aviv, Israel, in 1995 and 1996, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 2001.

From January 2002 to July 2002, she was a Postdoctoral Fellow at the Digital Signal Processing Group at MIT. She is currently an Associate Professor in the Department of Electrical Engineering at the Technion—Israel Institute of Technology, Haifa, Israel. She is also a Research Affiliate with the Research Laboratory of Electronics at MIT. Her research interests are in the general areas of signal processing, statistical signal processing, and computational biology.

Dr. Eldar was in the program for outstanding students at TAU from 1992 to 1996. In 1998, she held the Rosenblith Fellowship for study in electrical engineering at MIT, and in 2000, she held an IBM Research Fellowship. From 2002 to 2005, she was a Horev Fellow of the Leaders in Science and Technology program at the Technion and an Alon Fellow. In 2004, she was awarded the Wolf Foundation Krill Prize for Excellence in Scientific Research, in 2005 the Andre and Bella Meyer Lectureship, in 2007 the Henry Taub Prize for Excellence in Research, and in 2008 the Hershel Rich Innovation Award. She is a member of the IEEE Signal Processing Theory and Methods technical committee, an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING the *EURASIP Journal of Signal Processing*, and the *SIAM Journal on Matrix Analysis and Applications*, and on the Editorial Board of *Foundations and Trends in Signal Processing*.