IMPROVEMENT OF LEAST-SQUARES UNDER ARBITRARY WEIGHTED MSE

Yonina C. Eldar

Department of Electrical Engineering
Technion–Israel Institute of Technology, Haifa, Israel
Email: yonina@ee.technion.ac.il.

ABSTRACT

The seminal work of Stein in the 1950's ignited a large body of research devoted to improving the total mean-squared error (MSE) of the least-squares (LS) estimator. A drawback of these methods is that they improve the total MSE 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 LS over bounded norm signals, *under all weighted MSE measures*. We first derive 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*—** Weighted mean-squared error (MSE), minimax MSE, domination, admissibility, component MSE.

1. INTRODUCTION

Linear regression has been studied extensively since the pioneering work of Gauss on least-squares (LS) fitting. The celebrated LS method is aimed at estimating a deterministic vector $\mathbf{x} \in \mathbb{C}^m$ from noisy observations $\mathbf{y} \in \mathbb{C}^n$ which are related through

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

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 $\mathbf{y}$. If the noise covariance 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. Even though unbiasedness may be appealing intuitively, it does not necessarily lead to a small estimation error $\hat{\mathbf{x}} - \mathbf{x}$. Thus, many attempts have been made to develop linear estimators that may be biased but closer to the true $\mathbf{x}$ [1, 2, 3, 4].

An alternative approach is to define a statistical objective which directly measures the estimation error $\hat{\mathbf{x}} - \mathbf{x}$. A common design criterion is the total MSE given by $E\{\|\hat{\mathbf{x}} - \mathbf{x}\|^2\}$. Unfortunately, since $\mathbf{x}$ is deterministic, this measure depends on $\mathbf{x}$. One way to eliminate the signal dependency is to assume that $\mathbf{x}$ is norm-bounded, and then minimize the worst-case MSE leading to the minimax trace MSE (MXTM) method [5, 6]. A nice feature of this approach is that the MXTM strategy dominates LS in the total MSE sense, so that its total MSE is smaller than that of LS, *for all* bounded values of $\mathbf{x}$ [6].

The concept of domination leads to a partial ordering among methods [7]. An estimator $\hat{\mathbf{x}}_1$ whose 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 $\hat{\mathbf{x}}_2$. An estimate is *admissible* if it is not dominated by any other strategy. The theory of LS domination is well developed since the seminal work of James and Stein [8], in which they constructed a nonlinear estimator dominating LS in a total MSE sense. A shortcoming of the James-Stein concept is that it reduces the total MSE at the expense of an increase in the individual component MSEs [9], so that specific elements may be severely miss-estimated. Component-wise MSE is an example of a weighted MSE measure where different weights are given to the individual signal elements. A desirable property we may wish our estimator to possess is that it has "good" performance under different choices of weighting. Therefore, we consider a broader notion of domination: we characterize and design estimators that dominate the LS *for every choice of weighted MSE*. Mathematically, this requires that the MSE matrix of $\hat{\mathbf{x}}$ is smaller or equal (in a matrix sense) than that of the LS method.

In Section 2, we derive an easily verifiable necessary and sufficient condition such that a linear estimator dominates LS in a matrix sense for all norm-bounded vectors $\mathbf{x}$. As we show, there is a large class of estimators with this property. An important question is how to select a "good" strategy among all the dominating possibilities. To this end, we suggest in Section 3 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. As we show, this approach has the additional desirable property that it is admissible in a weighted MSE sense. To evaluate the MXMM estimate we first show that it can be formulated as a solution to a semidefinite programming problem (SDP) [10]. We then consider, in Section 4, a broad class of settings in which a more explicit solution can be found. In Section 5 we compare our approach with the MXTM and LS strategies.

2. 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 weighted norm of $\mathbf{x}$ is defined as $\|\mathbf{x}\|_{\mathbf{T}}^2 = \mathbf{x}^*\mathbf{T}\mathbf{x}$, $\mathbf{y}_i$ is the ith element of $\mathbf{y}$ and $\mathrm{diag}(\delta_1, \ldots, \delta_m)$ is an $m \times m$ diagonal matrix with diagonal elements δ_i. 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 eigenvalue of a Hermitian matrix $\mathbf{A}$ is denoted $\lambda_{\max}(\mathbf{A})$. The trace and Hermitian conjugate are written as $\mathrm{Tr}(\mathbf{A})$ and $\mathbf{A}^*$, respectively.

We consider the linear regression model (1) where $\mathbf{H}$ is a known $n \times m$ matrix with rank m, and $\mathbf{n}$ is a zero-mean random vector with covariance $\mathbf{C} \succ \mathbf{0}$. We estimate $\mathbf{x}$ from $\mathbf{y}$ using a linear estimator of the form $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}$ for an $m \times n$ matrix $\mathbf{G}$, where we assume that $\|\mathbf{x}\|_{\mathbf{T}} \leq L$ for some $\mathbf{T} \succ \mathbf{0}$ and scalar $L > 0$.

This work was supported by the Israel Science Foundation.

A popular measure of estimator performance is the total MSE

$$E\left\{\|\hat{\mathbf{x}}-\mathbf{x}\|^2\right\}=\sum_{i=1}^{m}E\{|\hat{\mathbf{x}}_i-\mathbf{x}_i|^2\}=\mathrm{Tr}(\mathbf{M}(\hat{\mathbf{x}})), \quad (2)$$

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

$$\begin{aligned}\mathbf{M}(\hat{\mathbf{x}}) &= E\left\{(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^*\right\}\\ &= (\mathbf{I}-\mathbf{GH})\mathbf{x}\mathbf{x}^*(\mathbf{I}-\mathbf{GH})^*+\mathbf{G}\mathbf{C}\mathbf{G}^*. \quad (3)\end{aligned}$$

More generally, we may consider a weighted total MSE

$$\mathrm{MSEW}(\hat{\mathbf{x}})=E\left\{(\hat{\mathbf{x}}-\mathbf{x})^*\mathbf{W}(\hat{\mathbf{x}}-\mathbf{x})\right\}=\mathrm{Tr}(\mathbf{W}\mathbf{M}(\hat{\mathbf{x}})), \quad (4)$$

for some weighting matrix $\mathbf{W}\succeq\mathbf{0}$ so that different weights are assigned to the individual errors. For example, choosing $\mathbf{W}=\mathbf{e}^i\mathbf{e}^{i*}$, where $\mathbf{e}^i$ has 1 in the ith component and 0 otherwise, results in the MSE of the ith component $\mathrm{MSEW}(\hat{\mathbf{x}})=E\{|\hat{\mathbf{x}}_i-\mathbf{x}_i|^2\}$.

For a given choice of $\mathbf{W}$, a possible design criterion is to minimize the weighted MSE (4). Unfortunately, this measure depends in general on $\mathbf{x}$, which is unknown, and therefore cannot be minimized. The dependency on $\mathbf{x}$ can be eliminated by requiring that $\mathbf{GH}=\mathbf{I}$, or restricting attention to unbiased estimators. When $\mathbf{W}=\mathbf{I}$, minimizing the resulting MSE leads to the LS estimator

$$\hat{\mathbf{x}}_{\mathrm{LS}}=(\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}^{-1}\mathbf{y}=\mathbf{G}_{\mathrm{LS}}\mathbf{y}. \quad (5)$$

However, this does not mean that the MSE is small. It is well-known that the MSE of LS can be large in many problems.

To directly control the MSE, a minimax total MSE (MXTM) approach was suggested in [5], in which the worst-case total MSE is minimized over $\|\mathbf{x}\|_{\mathbf{T}}\leq L$. It was then shown in [6] that the MXTM strategy is admissible and dominates LS in terms of total MSE. Nonetheless, the MSE of an individual component may be larger than that resulting from LS. To illustrate this point, in Fig. 1 we compare the MSE of the LS with that resulting from the MXTM approach for $L=2$, white noise and a random choice of $\mathbf{H}$, with $n=8,m=5$. For the MXTM estimator, the worst MSE over $\|\mathbf{x}\|\leq L$ is plotted in each case. Fig. 1(a) considers 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 LS. In Fig. 1(b) we plot the total MSE of the two methods. As expected, the total MSE of the MXTM strategy is smaller than that of LS.

Figure 1 illustrates that minimizing the total MSE may be insufficient when a weighted MSE is of interest. To ensure LS domination for all weighted MSE, the MSE matrix of $\hat{\mathbf{x}}$ must satisfy:

$$\mathbf{M}(\hat{\mathbf{x}})\preceq\mathbf{M}(\hat{\mathbf{x}}_{\mathrm{LS}})=(\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H})^{-1}\triangleq\mathbf{Q}, \quad \forall\|\mathbf{x}\|_{\mathbf{T}}\leq L, \quad (6)$$

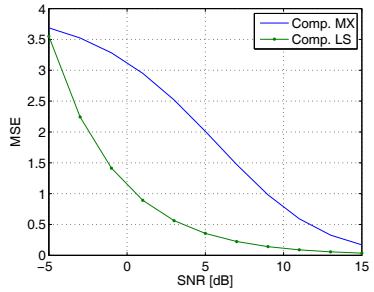
where $\mathbf{M}(\hat{\mathbf{x}})$ is defined by (3). An estimator $\hat{\mathbf{x}}$ with this property is said to matrix-dominate LS. An explicit condition for LS matrix-domination is given in the following theorem.

Theorem 1. *Let $\hat{\mathbf{x}}=\mathbf{G}\mathbf{y}$ be an estimate of $\mathbf{x}$ in the model (1). Then $\hat{\mathbf{x}}$ matrix-dominates LS for all $\|\mathbf{x}\|_{\mathbf{T}}\leq L$ if and only if*

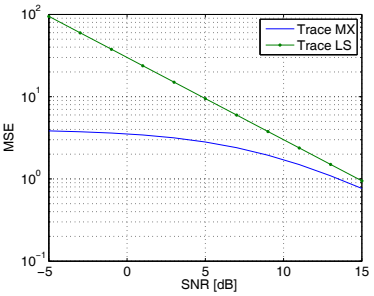
$$\lambda_{\max}\left(\mathbf{G}\mathbf{C}\mathbf{G}^*-\mathbf{Q}+L^2(\mathbf{I}-\mathbf{GH})\mathbf{T}^{-1}(\mathbf{I}-\mathbf{GH})^*\right)\leq 0.$$

Proof. From (3) and (6) matrix domination is equivalent to

$$\mathbf{B}\mathbf{x}\mathbf{x}^*\mathbf{B}^*+\mathbf{A}\preceq\mathbf{0}, \quad \forall\|\mathbf{x}\|_{\mathbf{T}}\leq L, \quad (7)$$



(a)



(b)

Fig. 1. MSE as a function of the noise variance using the MXTM and LS estimators (a) MSE of the first component (b) total MSE.

where we defined $\mathbf{A}=\mathbf{G}\mathbf{C}\mathbf{G}^*-\mathbf{Q}$ and $\mathbf{B}=\mathbf{I}-\mathbf{GH}$. In order for (7) to be satisfied we must have that

$$\max_{\|\mathbf{x}\|_{\mathbf{T}}\leq L}\mathbf{y}^*\mathbf{B}\mathbf{x}\mathbf{x}^*\mathbf{B}^*\mathbf{y}+\mathbf{y}^*\mathbf{A}\mathbf{y}\leq 0, \quad \forall\mathbf{y}. \quad (8)$$

Using the Cauchy-Scwartz inequality,

$$\max_{\|\mathbf{x}\|_{\mathbf{T}}\leq L}\mathbf{y}^*\mathbf{B}\mathbf{x}\mathbf{x}^*\mathbf{B}^*\mathbf{y}=L^2\mathbf{y}^*\mathbf{B}\mathbf{T}^{-1}\mathbf{B}^*\mathbf{y}. \quad (9)$$

Therefore, (7) is equivalent to

$$L^2\mathbf{y}^*\mathbf{B}\mathbf{T}^{-1}\mathbf{B}^*\mathbf{y}+\mathbf{y}^*\mathbf{A}\mathbf{y}\leq 0, \quad \forall\mathbf{y}, \quad (10)$$

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

3. MINIMAX MATRIX-MSE ESTIMATOR

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 this end we propose 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. Since 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 $\alpha>0$ we assume that $\mathbf{W}\preceq\mathbf{I}$, leading to the following optimization problem:

$$\begin{aligned}&\min_{\hat{\mathbf{x}}}\max_{\mathbf{x},\mathbf{W}}\left\{\mathrm{Tr}(\mathbf{W}\mathbf{M}(\hat{\mathbf{x}})):\|\mathbf{x}\|_{\mathbf{T}}\leq L,\mathbf{0}\preceq\mathbf{W}\preceq\mathbf{I}\right\}\\ &\text{s.t.}\quad \mathbf{M}(\hat{\mathbf{x}})\preceq\mathbf{M}(\hat{\mathbf{x}}_{\mathrm{LS}}), \quad \text{for all } \|\mathbf{x}\|_{\mathbf{T}}\leq L. \quad (11)\end{aligned}$$

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

Since $\mathbf{M}(\hat{\mathbf{x}}) \succeq \mathbf{0}$, the inner maximization with respect to $\mathbf{W}$ is obtained when $\mathbf{W} = \mathbf{I}$. Using Theorem 1, (11) reduces to

$$\min_{\mathbf{G}} \left\{ \mathrm{Tr}(\mathbf{G}\mathbf{C}\mathbf{G}^*) + \max_{\|\mathbf{x}\|_{\mathbf{T}} \le L} \mathbf{x}^*(\mathbf{I}-\mathbf{G}\mathbf{H})^*(\mathbf{I}-\mathbf{G}\mathbf{H})\mathbf{x} \right\}$$
$$\text{s.t. } \lambda_{\max}\left(\Phi(\mathbf{G}) + L^2(\mathbf{I}-\mathbf{G}\mathbf{H})\mathbf{T}^{-1}(\mathbf{I}-\mathbf{G}\mathbf{H})^*\right) \le 0, \quad (12)$$

where we denoted $\Phi(\mathbf{G}) = \mathbf{G}\mathbf{C}\mathbf{G}^* - \mathbf{Q}$. The following theorem establishes some important properties of the MXMM estimator:

Theorem 2. *Let $\widehat{\mathbf{G}}$ be the solution to (12). Then $\widehat{\mathbf{G}}$ is unique, and is admissible in the matrix sense.*

3.1. SDP Formulation

To evaluate $\hat{\mathbf{x}}_{\mathrm{MXMM}}$ we now formulate it as a solution to an SDP, which is the problem of minimizing a linear function subject to linear matrix inequalities (LMIs). Using the relation

$$\max_{\mathbf{x}^*\mathbf{T}\mathbf{x} \le L^2} \mathbf{x}^*\mathbf{Z}\mathbf{x} = L^2\lambda_{\max}(\mathbf{Z}\mathbf{T}^{-1}) = \min_{\lambda} \left\{ L^2\lambda : \mathbf{Z} \preceq \lambda\mathbf{T} \right\} \quad (13)$$

for any $\mathbf{Z} \succeq \mathbf{0}$, (12) is equivalent to

$$\min_{\mathbf{G}} \left\{ \mathrm{Tr}\left(\mathbf{G}\mathbf{C}\mathbf{G}^*\right) + L^2\lambda \right\}$$
$$\text{s.t. } (\mathbf{I}-\mathbf{G}\mathbf{H})^*(\mathbf{I}-\mathbf{G}\mathbf{H}) \preceq \lambda\mathbf{T};$$
$$\lambda_{\max}\left(\Phi(\mathbf{G}) + L^2(\mathbf{I}-\mathbf{G}\mathbf{H})\mathbf{T}^{-1}(\mathbf{I}-\mathbf{G}\mathbf{H})^*\right) \le 0. \quad (14)$$

Lemma 1 reduces the dimensionality of (14) when $m < n$.

Lemma 1. *Let the $m \times n$ matrix $\widehat{\mathbf{G}}$ be the solution to (14). Then*

$$\widehat{\mathbf{G}} = \mathbf{K}(\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}^{-1} = \mathbf{K}\mathbf{Q}\mathbf{H}^*\mathbf{C}^{-1}, \quad (15)$$

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

$$\min_{\mathbf{K},\lambda} \{Tr(\mathbf{K}\mathbf{Q}\mathbf{K}^*) + L^2\lambda\}$$
$$s.t. \quad (\mathbf{I}-\mathbf{K})^*(\mathbf{I}-\mathbf{K}) \preceq \lambda\mathbf{T}$$
$$\mathbf{K}\mathbf{Q}\mathbf{K}^* + L^2(\mathbf{I}-\mathbf{K})\mathbf{T}^{-1}(\mathbf{I}-\mathbf{K})^* \preceq \mathbf{Q}. \quad (16)$$

Our goal now is to convert (16) into an SDP so that the solution can be computed efficiently. To this end, let $\mathbf{X} = \mathbf{K}\mathbf{Q}\mathbf{K}^*$ and add this equality as a third constraint in (16). The objective in (16) is then linear in $\mathbf{X}$ and λ, and the first two constraints can be converted into LMIs using Schur's complement. The additional constraint however is nonconvex. Nonetheless, replacing this equality with $\mathbf{X} \succeq \mathbf{K}\mathbf{Q}\mathbf{K}^*$ does not change the solution. To see this, suppose that the solutions $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{K}}$ to the relaxed problem satisfy $\widehat{\mathbf{X}} \succeq \widehat{\mathbf{K}}\mathbf{Q}\widehat{\mathbf{K}}^*$ but $\widehat{\mathbf{X}} \neq \widehat{\mathbf{K}}\mathbf{Q}\widehat{\mathbf{K}}^*$. Then $\mathbf{X}' = \widehat{\mathbf{K}}\mathbf{Q}\widehat{\mathbf{K}}^*$ obeys the constraints and $\mathrm{Tr}(\mathbf{X}') < \mathrm{Tr}(\widehat{\mathbf{X}})$ (since for a matrix $\mathbf{A} \succeq \mathbf{0}$, $\mathrm{Tr}(\mathbf{A}) = 0$ only if $\mathbf{A} = \mathbf{0}$) so that $\widehat{\mathbf{X}}$ cannot be optimal. Applying Schur's complement to the resulting constraints leads to the following theorem.

Theorem 3. *The MXMM estimator of (11) is given by*

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

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

$$\min_{\mathbf{K},\mathbf{X},\lambda} \{Tr(\mathbf{X}) + L^2\lambda\}$$
$$s.t. \quad \begin{bmatrix} \lambda\mathbf{T} & \mathbf{I}-\mathbf{K} \\ (\mathbf{I}-\mathbf{K})^* & \mathbf{I} \end{bmatrix} \succeq \mathbf{0}$$
$$\begin{bmatrix} \mathbf{Q}-\mathbf{X} & \mathbf{I}-\mathbf{K} \\ (\mathbf{I}-\mathbf{K})^* & (1/L^2)\mathbf{T} \end{bmatrix} \succeq \mathbf{0}$$
$$\begin{bmatrix} \mathbf{X} & \mathbf{K} \\ \mathbf{K}^* & \mathbf{Q}^{-1} \end{bmatrix} \succeq \mathbf{0}, \quad (17)$$

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

4. COMMUTING MATRICES

An explicit expression for the MXMM estimate when $\mathbf{T}$ and $\mathbf{Q}$ have the same eigenvector matrix $\mathbf{V}$ is given below.

Theorem 4. *Consider the setting of Theorem 3. Let $\mathbf{Q} = \mathbf{V}\Sigma\mathbf{V}^*$ where $\mathbf{V}$ is a unitary matrix, $\Sigma = diag(\sigma_1, \ldots, \sigma_m) \succ \mathbf{0}$ and let $\mathbf{T} = \mathbf{V}\Lambda\mathbf{V}^*$ where $\Lambda = diag(\lambda_1, \ldots, \lambda_m) \succ \mathbf{0}$. Then*

$$\hat{\mathbf{x}}_{\mathrm{MXMM}} = \mathbf{V}\mathbf{D}\mathbf{V}^*(\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}^{-1}\mathbf{y}, \quad (18)$$

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

$$d_i = \begin{cases} 1-\sqrt{\beta_0\lambda_i}, & 1-\sqrt{\beta_0\lambda_i} \ge \alpha_i; \\ \alpha_i, & otherwise. \end{cases} \quad (19)$$

Here

$$\alpha_i = \left[\frac{L^2-\sigma_i\lambda_i}{L^2+\sigma_i\lambda_i}\right]_+, \quad 1 \le i \le m, \quad (20)$$

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

$$\mathcal{G}(\beta) = \sum_{i=1}^{m} \lambda_i\tilde{\mu}_i(\beta) - L^2, \quad (21)$$

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

$$\tilde{\mu}_i(\beta) = \begin{cases} \sigma_i\left(\frac{1}{\sqrt{\beta\lambda_i}} - 1\right), & 1-\sqrt{\beta\lambda_i} \ge \alpha_i; \\ 0, & 1-\sqrt{\beta\lambda_i} < \alpha_i. \end{cases} \quad (22)$$

In order to find β_0, note that $0 \le \beta_0 \le \beta_{\mathrm{TH}}$ where

$$\beta_{\mathrm{TH}} = \max_{1 \le i \le m} \left\{ \frac{(1-\alpha_i)^2}{\lambda_i} \right\}, \quad (23)$$

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 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 $[0, \beta_{\mathrm{TH}}]$.

4.1. Comparison between the MXMM and MXTM Methods

It is interesting to compare the MXMM and MXTM methods in the jointly diagonalizable case. The MXTM estimator under this assumption is derived in [5] and has the same form as $\hat{\mathbf{x}}_{\mathrm{MXMM}}$ of (18),where $\mathbf{D} = \mathrm{diag}(\tilde{d}_1, \ldots, \tilde{d}_m)$ with

$$\tilde{d}_i = \begin{cases} 1-\sqrt{\zeta\lambda_i}, & 1-\sqrt{\zeta\lambda_i} \ge 0; \\ 0. & \text{otherwise.} \end{cases} \quad (24)$$

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

$$\eta_i(\zeta) = \begin{cases} \sigma_i\left(\frac{1}{\sqrt{\zeta\lambda_i}} - 1\right), & 1-\sqrt{\zeta\lambda_i} > 0; \\ 0, & 1-\sqrt{\zeta\lambda_i} \le 0. \end{cases} \quad (25)$$

If the eigenvalues of $\mathbf{T}$ are sorted in decreasing order, then

$$\sqrt{\zeta} = \frac{\sum_{i=k+1}^{m} \sqrt{\lambda_i}\sigma_i}{L^2 + \sum_{i=k+1}^{m} \lambda_i \sigma_i}, \tag{26}$$

where k is the smallest index such that $\sqrt{\zeta\lambda_{k+1}} < 1$.

Comparing with Theorem 4 leads to the following result.

Theorem 5. *Consider the problem of Theorem 4. Let $\mathbf{V}, \Sigma$ and Λ be ordered such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$. Then the MXMM and MXTM estimators both have the form (18) with $\mathbf{D} = diag(d_1, \ldots, d_m)$ for the MXMM estimate and $\mathbf{D} = diag(\tilde{d}_1, \ldots, \tilde{d}_m)$ for the MXTM estimate where*

$$d_i \geq \tilde{d}_i, \quad 1 \leq i \leq m. \tag{27}$$

Furthermore, the estimators coincide if

$$\begin{aligned} &1 - \sqrt{\zeta\lambda_i} \geq \alpha_i, \quad && k+1 \leq i \leq m; \\ &\alpha_i = 0, && 1 \leq i \leq k, \end{aligned} \tag{28}$$

where ζ is given by (26) and k is the smallest index such that $0 \leq k \leq m-1$ and $\sqrt{\zeta\lambda_{k+1}} < 1$. In particular, if $L^2 \leq \sigma_i\lambda_i$, $1 \leq i \leq m$, then the MXMM and MXTM methods are equivalent.

We conclude that the shrinkage of the MXTM estimate is larger than that of the MXMM method. Evidently, larger shrinkage can decrease the total MSE at the expense of increasing the MSE of some components.

5. EXAMPLES

We now 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.

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 1st 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 component-wise MSE, while the MXTM approach dominates LS only in the trace MSE sense. The behavior in Fig. 2 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 are similar, while the component-wise performance of MXMM is typically much better. Thus, it seems like the MXMM approach can substantially decrease the weighted MSE with only a small increase in the trace MSE.

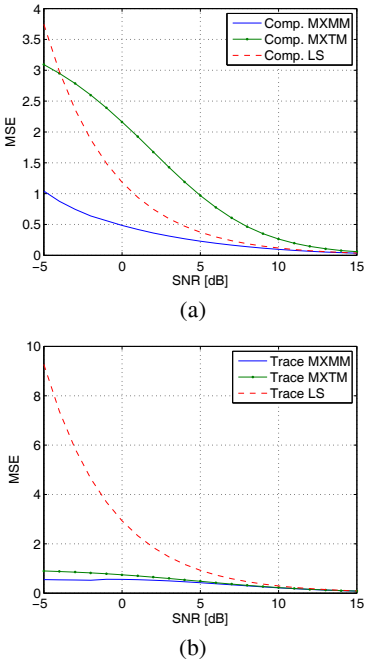


Fig. 2. MSE in estimating $\mathbf{x}$ as a function of the noise variance using the MXTM, MXMM and LS estimators (a) MSE of the first component (b) total MSE.

6. REFERENCES

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

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

[3] Y. C. Eldar and A. V. Oppenheim, "Covariance shaping least-squares estimation," *IEEE Trans. Signal Processing*, vol. 51, pp. 686–697, Mar. 2003.

[4] Y. C. Eldar, A. Ben-Tal, and A. Nemirovski, "Linear minimax regret estimation of deterministic parameters with bounded data uncertainties," *IEEE Trans. Signal Processing*, vol. 52, pp. 2177–2188, Aug. 2004.

[5] Y. C. Eldar, A. Ben-Tal, and A. Nemirovski, "Robust mean-squared error estimation in the presence of model uncertainties," *IEEE Trans. Signal Processing*, vol. 53, pp. 168–181, Jan. 2005.

[6] Y. C. Eldar, "Comparing between estimation approaches: Admissible and dominating linear estimators," *IEEE Trans. on Signal Processing*, vol. 54, pp. 1689–1702, May 2006.

[7] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, New York, NY: Springer-Verlag, Inc., second edition, 1998.

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

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

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