Covariance Shaping Least-Squares Estimation

Yonina C. Eldar, *Member, IEEE,* and Alan V. Oppenheim, *Fellow, IEEE*

***Abstract*—A new linear estimator is proposed, which we refer to as the covariance shaping least-squares (CSLS) estimator, for estimating a set of unknown deterministic parameters x observed through a known linear transformation H and corrupted by additive noise. The CSLS estimator is a biased estimator directed at improving the performance of the traditional least-squares (LS) estimator by choosing the estimate of x to minimize the (weighted) total error variance in the observations subject to a constraint on the covariance of the estimation error so that we control the dynamic range and spectral shape of the covariance of the estimation error.**

The CSLS estimator presented in this paper is shown to achieve the Cramér-Rao lower bound for biased estimators. Furthermore, analysis of the mean-squared error (MSE) of both the CSLS estimator and the LS estimator demonstrates that the covariance of the estimation error can be chosen such that there is a threshold SNR below which the CSLS estimator yields a lower MSE than the LS estimator for all values of x.

As we show, some of the well-known modifications of the LS estimator can be formulated as CSLS estimators. This allows us to interpret these estimators as the estimators that minimize the total error variance in the observations, among all linear estimators with the same covariance.

***Index Terms*—Biased estimation, covariance shaping, estimation, least squares, MMSE.**

I. Introduction

A generic estimation problem that has been studied extensively in the literature is that of estimating the unknown deterministic parameters $\mathbf{x}$ observed through a known linear transformation $\mathbf{H}$ and corrupted by zero-mean additive noise $\mathbf{w}$. A common approach to estimating the parameters $\mathbf{x}$ is to restrict the estimator to be linear in the data $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$ and then to find the linear estimate of $\mathbf{x}$ that results in an estimated data vector $\hat{\mathbf{y}}$ that is as close as possible to the given data vector $\mathbf{y}$ in a least-squares (LS) sense so that $\hat{\mathbf{y}}$ is chosen to minimize the total squared error in the observations.

The LS method is widely employed in diverse fields, both as an estimation criterion and as a method for parametric modeling of data (see e.g., [1]–[4]). Numerous extensions of the LS method have been previously proposed in the literature. The total LS method, first proposed by Golub and Van Loan in [5] (see also [6]), assumes that the model matrix $\mathbf{H}$ may not be known exactly and seeks the parameters $\mathbf{x}$ and the minimum perturbation to the model matrix that minimize the LS error. The extended LS method proposed by Yeredor in [7] seeks the parameters and some presumed underlying data that together minimize a weighted combination of model errors and measurement errors. In both of these extensions, it is assumed that the data model does not hold perfectly, either due to errors in $\mathbf{H}$ or errors in the data $\mathbf{y}$.

In our method, we assume that the data model holds i.e., $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$ with $\mathbf{H}$ and $\mathbf{y}$ known exactly, and our objective is to minimize the error between $\mathbf{x}$ and the estimate of $\mathbf{x}$. It is well known that among all possible unbiased linear estimators, the LS estimator minimizes the variance [2]. However, this does not imply that the resulting variance or mean-squared error (MSE) is small, where the MSE of an estimator is the sum of the variance and the squared norm of the bias. In particular, in many cases, the data vector $\mathbf{y}$ is not very sensitive to changes in $\mathbf{x}$ so that a large error in estimating $\mathbf{x}$ may translate into a small error in estimating the data vector $\mathbf{y}$, in which case, the LS estimate may result in a poor estimate of $\mathbf{x}$. This effect is especially predominant at low to moderate signal-to-noise ratio (SNR), where the data vector $\mathbf{y}$ is typically affected more by the noise than by changes in $\mathbf{x}$; the exact SNR range will depend on the properties of the model matrix $\mathbf{H}$. A difficulty often encountered in this estimation problem is that the error in the estimation can have a large variance and a covariance structure with a very high dynamic range.

Various modifications of the LS estimator for the case in which the data model is assumed to hold perfectly have been proposed [8]. In [9], Stein showed that the LS estimator for the mean vector in a multivariate Gaussian distribution with dimension greater than 2 is "inadmissible," i.e., for certain parameter values, other estimators exist with lower MSE. An explicit (nonlinear) estimator with this property, which is referred to as the James–Stein estimator, was later proposed and analyzed in [10]. This work appears to have been the starting point for the study of alternatives to LS estimators. Among the more prominent alternatives are the ridge estimator [11] (also known as Tikhonov regularization [12]) and the shrunken estimator [13].

To improve the performance of the LS estimator at low to moderate SNR, we propose a modification of the LS estimate, in which we choose the estimator of $\mathbf{x}$ to minimize the total error variance in the observations $\mathbf{y}$, subject to a constraint on the covariance of the error in the estimate of $\mathbf{x}$. The resulting estimator of $\mathbf{x}$ is derived in Section III, and is referred to as the covariance shaping LS (CSLS) estimator. In Section IV, we show that both the ridge estimator and the shrunken estimator can be formulated as CSLS estimators.

Manuscript received October 3, 2001; revised October 18, 2002. This work was supported in part through collaborative participation in the Advanced Sensors Collaborative Technology Alliance (CTA) sponsored by the U.S. Army Research Laboratory under Cooperative Agreement DAAD19-01-2-0008. The work of Y. C. Eldar was supported by a Horev Fellowship through the Taub Foundation. The associate editor coordinating the review of this paper and approving it for publication was Prof. Xioadong Wang.

Y. C. Eldar was with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA. She is now is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa, 32000, Israel (e-mail: yonina@ee.technion.ac.il).

A. V. Oppenheim is with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: avo@mit.edu).

Digital Object Identifier 10.1109/TSP.2002.808125

In Section V, we show that the CSLS estimator has a property analogous to the property of the LS estimator. Specifically, it is shown to achieve the Cramér-Rao lower bound (CRLB) for biased estimators [2], [14], [15] when the noise is Gaussian. This implies that for Gaussian noise, there is no linear or nonlinear estimator with a smaller variance, or MSE, and the same bias as the CSLS estimator.

In Section VI, we analyze the MSE in estimating $\mathbf{x}$ of both the CSLS estimator and the LS estimator and show that the covariance of the estimation error can be chosen so that there is a threshold SNR, below which the CSLS estimator yields a lower MSE than the LS estimator, for all values of $\mathbf{x}$. The simulations presented in Section IX strongly suggest that the CSLS estimator can significantly decrease the MSE of the estimation error in $\mathbf{x}$ over the LS estimator for a wide range of SNR values.

In Section VII, we show that the CSLS estimator can alternatively be expressed as an LS estimator followed by a weighted minimum mean-squared error (WMMSE) shaping transformation [18] that optimally shapes the covariance of the LS estimate of $\mathbf{x}$. The WMMSE covariance shaping transformation minimizes the weighted MSE between the original vector and the transformed vector, i.e., results in a vector with a specified covariance matrix that is closest in a weighted MSE sense to the original vector. The WMMSE covariance shaping problem is an extension of the minimum MSE (MMSE) whitening problem [16], [17], in which the transformed vector is constrained to be white, and the transformation is chosen to minimize the (unweighted) MSE between the original vector and the white vector.

Several applications of CSLS estimation are discussed in Section IX. The first application is to estimation of the parameters in an ARMA model. We show that the CSLS estimator can significantly decrease the MSE in estimating both the AR and the MA parameters over a wide range of SNRs. As a second application, the CSLS estimator is applied to the problem of estimating the amplitudes of complex exponentials with known frequencies and damping factor in additive noise.

II. Least-Squares Estimation

We denote vectors in $\mathbb{C}^m$ (m arbitrary) by boldface lowercase letters and matrices in $\mathbb{C}^{m\times m}$ by boldface uppercase letters. $\mathbf{I}_m$ denotes the $m \times m$ identity matrix. The adjoint of a transformation is denoted by $(\cdot)^*$, and $\hat{(\cdot)}$ denotes an optimal vector or transformation. The squared norm of the vector $\mathbf{x}$ is denoted by $\|\mathbf{x}\|^2 = \mathbf{x}^*\mathbf{x}$. A prime attached to a random variable or vector denotes the variable or vector with the mean subtracted, e.g., $\mathbf{a}' = \mathbf{a} - E(\mathbf{a})$.

We consider the class of estimation problems represented by the linear model

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

where $\mathbf{x}$ is a deterministic vector of unknown parameters, $\mathbf{H}$ is a known $n \times m$ matrix, and $\mathbf{w}$ is a zero-mean random vector with covariance $\mathbf{C}_w$. For simplicity of exposition, we assume that $\mathbf{H}$ has rank m; the results extend in a straightforward way to the case in which the rank of $\mathbf{H}$ is less than m [18].

The (weighted) least-squares (LS) estimate of $\mathbf{x}$, which is denoted $\hat{\mathbf{x}}_{\mathrm{LS}}$, is chosen such that $\hat{\mathbf{y}} = \mathbf{H}\hat{\mathbf{x}}_{\mathrm{LS}} = \mathbf{H}\mathbf{G}\mathbf{y}$ is as close as possible to $\mathbf{y}$ in a (weighted) LS sense so that $\hat{\mathbf{y}}$ minimizes the total squared error in the observations. Thus, the LS estimate $\hat{\mathbf{x}}_{\mathrm{LS}} = \mathbf{G}\mathbf{y}$ is chosen to minimize the total squared error

$$\varepsilon_{\mathrm{LS}} = (\mathbf{y} - \mathbf{H}\mathbf{G}\mathbf{y})^*\mathbf{W}(\mathbf{y} - \mathbf{H}\mathbf{G}\mathbf{y}) \tag{2}$$

where $\mathbf{W}$ is an arbitrary positive definite weighting matrix. If we choose $\mathbf{W} = \mathbf{C}_w^{-1}$, then the LS estimate is given by

$$\hat{\mathbf{x}}_{\mathrm{LS}} = \left(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y}. \tag{3}$$

The Gauss–Markov theorem [2] states that with $\mathbf{W} = \mathbf{C}_w^{-1}$, the LS estimator is the best linear unbiased estimator (BLUE) of $\mathbf{x}$, i.e., it minimizes the total variance defined by

$$V(\hat{\mathbf{x}}) = E((\hat{\mathbf{x}} - E(\hat{\mathbf{x}}))^*(\hat{\mathbf{x}} - E(\hat{\mathbf{x}}))) \tag{4}$$

from all linear *unbiased* estimators. Furthermore, if $\mathbf{w}$ is a zero-mean Gaussian random vector, then the LS estimator (with optimal weighting) is also the minimum variance unbiased estimator, i.e., it minimizes the variance from all linear and nonlinear *unbiased* estimators.

The LS estimator has a variety of optimality properties in the class of unbiased estimators. However, an unbiased estimator does not necessarily lead to minimum MSE, where the MSE of an estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ is defined by

$$\begin{aligned}\mathrm{MSE}(\hat{\mathbf{x}}) &= E(\|\hat{\mathbf{x}} - \mathbf{x}\|^2) = \mathrm{Tr}\left(E((\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^*)\right)\\ &= V(\hat{\mathbf{x}}) + \|B(\hat{\mathbf{x}})\|^2.\end{aligned} \tag{5}$$

Here, $B(\hat{\mathbf{x}}) = E(\hat{\mathbf{x}}) - \mathbf{x}$ denotes the bias of the estimator $\hat{\mathbf{x}}$. As we will show, in many cases, the CSLS estimator, which we develop in Section III, can result in lower MSE than the LS estimator by allowing for a bias. In Section V, we also show that the CSLS estimator has a property analogous to the LS estimator. Namely, for Gaussian noise, it is the estimator among all linear and nonlinear estimators that minimizes the variance subject to a constraint on the bias.

III. Covariance Shaping Least-Squares Estimation

Since the MSE depends explicitly on the unknown parameters $\mathbf{x}$, we cannot choose an estimate to directly minimize the MSE. A common approach is to restrict the estimator to be linear and unbiased and then seek the estimator of this form that minimizes the variance or the MSE, which leads to the BLUE or the LS estimator. In our development, the estimator is not constrained to be unbiased. Our approach for choosing the estimator is motivated by the observation that, in many cases, the data vector $\mathbf{y}$ is not very sensitive to changes in $\mathbf{x}$ so that a large error in estimating $\mathbf{x}$ may translate into a small error in estimating the data vector $\mathbf{y}$, in which case, $\hat{\mathbf{x}}_{\mathrm{LS}}$ may result in a poor estimate of $\mathbf{x}$. In the high SNR limit, where $\sigma^2 \to 0$, $\hat{\mathbf{x}}_{\mathrm{LS}} \to \mathbf{x}$ so that the LS estimate converges to the true parameters $\mathbf{x}$, regardless of the model matrix $\mathbf{H}$. The CSLS estimator is directed at improving the performance of the LS estimator at low to moderate SNR by choosing the estimate of $\mathbf{x}$ to minimize the total error variance in $\mathbf{y}$ subject to a constraint on the covariance of the error in the

estimate of $\mathbf{x}$ so that we control the dynamic range and spectral shape of the covariance of the estimation error.

The CSLS estimate of $\mathbf{x}$, which is denoted $\hat{\mathbf{x}}_{\mathrm{CSLS}}$, is chosen to minimize the total variance of the weighted error between $\hat{\mathbf{y}} = \mathbf{H}\hat{\mathbf{x}}_{\mathrm{CSLS}} = \mathbf{HGy}$ and $\mathbf{y}$, subject to the constraint that the covariance of the error in the estimate $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ is proportional to a given covariance matrix $\mathbf{R}$. From (1), it follows that the covariance of $\mathbf{y}$ is equal to $\mathbf{C}_w$ so that the covariance of $\hat{\mathbf{x}}_{\mathrm{CSLS}}$, which is equal to the covariance of the error in the estimate $\hat{\mathbf{x}}_{\mathrm{CSLS}}$, is given by $\mathbf{GC}_w\mathbf{G}^*$. Thus, $\hat{\mathbf{x}}_{\mathrm{CSLS}} = \mathbf{Gy}$ is chosen to minimize

$$\varepsilon_{\mathrm{CSLS}} = E\left((\mathbf{y}' - \mathbf{HGy}')^*\mathbf{C}_w^{-1}(\mathbf{y}' - \mathbf{HGy}')\right) \tag{6}$$

subject to

$$\mathbf{GC}_w\mathbf{G}^* = c^2\mathbf{R} \tag{7}$$

where $\mathbf{y}' = \mathbf{y} - E(\mathbf{y})$, $\mathbf{R}$ is a given covariance matrix, and $c > 0$ is a constant that is either specified in advance or chosen to minimize the error (6).

This minimization problem is a special case of the general *weighted MMSE (WMMSE) shaping problem*. Specifically, the problem of (6) and (7) can be restated as the problem of finding the transformation $\mathbf{W}$ to minimize

$$E\left((\mathbf{a} - \mathbf{b})^*\mathbf{C}_a^{-1}(\mathbf{a} - \mathbf{b})\right) \tag{8}$$

where $\mathbf{b} = \mathbf{Wa}$, subject to

$$\mathbf{C}_b = \mathbf{WC}_a\mathbf{W}^* = c^2\mathbf{Q} \tag{9}$$

with $\mathbf{a} = \mathbf{y}'$, $\mathbf{C}_a = \mathbf{C}_w$, $\mathbf{W} = \mathbf{HG}$, and $\mathbf{Q} = \mathbf{HRH}^*$.

A. WMMSE Shaping

In this section, we consider the WMMSE shaping problem of (8) with weighting matrix $\mathbf{C}_a^{-1}$. The more general case of arbitrary weighting is considered in [18]. Let $\mathbf{a}$ denote a zero-mean random vector with positive-definite covariance matrix $\mathbf{C}_a$, and let $\mathbf{b} = \mathbf{Wa}$. We seek the transformation $\mathbf{W}$ that minimizes (8) subject to (9), where $\mathbf{Q}$ is a given covariance matrix that is not assumed to be invertible, and $c > 0$ is a constant that is either specified or chosen to minimize the error (8).

Denoting by $\overline{\mathbf{a}} = \mathbf{C}_a^{-1/2}\mathbf{a}$ and $\overline{\mathbf{b}} = \mathbf{C}_a^{-1/2}\mathbf{b}$, we may rewrite (8) as

$$E\left((\mathbf{a} - \mathbf{b})^*\mathbf{C}_a^{-1}(\mathbf{a} - \mathbf{b})\right) = E\left((\overline{\mathbf{a}} - \overline{\mathbf{b}})^*(\overline{\mathbf{a}} - \overline{\mathbf{b}})\right) \tag{10}$$

where the covariance matrix of $\overline{\mathbf{a}}$ is equal to $\mathbf{I}_m$, and the covariance matrix of $\overline{\mathbf{b}}$ is

$$\mathbf{C}_{\overline{b}} = c^2\mathbf{C}_a^{-1/2}\mathbf{QC}_a^{-1/2}. \tag{11}$$

Thus, minimizing (8) subject to (9) is equivalent to finding the transformation

$$\overline{\mathbf{W}} = \mathbf{C}_a^{-1/2}\mathbf{WC}_a^{1/2} \tag{12}$$

such that the random vector $\overline{\mathbf{b}} = \overline{\mathbf{W}}\overline{\mathbf{a}}$ has covariance given by (11) and is closest in an MSE sense to the random vector $\overline{\mathbf{a}}$ with covariance $\mathbf{I}_m$.

This problem is very similar to the MMSE whitening problem considered in [16]. Using the method used in [16] and [18] to derive the MMSE whitening transformation, it is straightforward to show (see Appendix A) that the minimizing $\overline{\mathbf{W}}$ is given by

$$\widehat{\overline{\mathbf{W}}} = \mathbf{C}_{\overline{b}}^{1/2} = c\left(\mathbf{C}_a^{-1/2}\mathbf{QC}_a^{-1/2}\right)^{1/2}. \tag{13}$$

From (12), we then have that the optimal value of $\mathbf{W}$ is

$$\widehat{\mathbf{W}} = c\mathbf{C}_a^{1/2}\left(\mathbf{C}_a^{-1/2}\mathbf{QC}_a^{-1/2}\right)^{1/2}\mathbf{C}_a^{-1/2}. \tag{14}$$

Using (62) (see Appendix B), we may express $\widehat{\mathbf{W}}$ as

$$\widehat{\mathbf{W}} = c\left(\mathbf{QC}_a^{-1}\right)^{1/2}. \tag{15}$$

We may further wish to choose c such that (8) is minimized. Substituting $\widehat{\mathbf{W}}$ back into (8), and minimizing with respect to c, the optimal value of c, which is denoted by $\hat{c}$, is given by

$$\hat{c} = \frac{\mathrm{Tr}\left(\left(\mathbf{QC}_a^{-1}\right)^{1/2}\right)}{\mathrm{Tr}\left(\mathbf{QC}_a^{-1}\right)}. \tag{16}$$

If the scaling c in (9) is fixed and $\mathbf{R} = \mathbf{I}_m$, then the WMMSE whitening transformation is equal to the MMSE whitening transformation derived in [16]; however, the optimal scaling values are different in both cases.

The results above are summarized in the following theorem.

Theorem 1 (WMMSE Covariance Shaping): Let $\mathbf{a} \in \mathbb{C}^m$ be a random vector with positive-definite covariance matrix $\mathbf{C}_a$. Let $\widehat{\mathbf{W}}$ be the optimal covariance shaping transformation that minimizes the weighted MSE defined by (8), between the input $\mathbf{a}$ and the output $\mathbf{b} = \mathbf{Wa}$ with covariance $\mathbf{C}_b = c^2\mathbf{Q}$, where $\mathbf{Q}$ is a given covariance matrix, and $c > 0$. Then

$$\widehat{\mathbf{W}} = \beta\left(\mathbf{QC}_a^{-1}\right)^{1/2}$$

where we have the following.

1) If c is specified, then $\beta = c$.
2) If c is chosen to minimize the weighted MSE, then $\beta = \hat{c}$ given by (16).

B. CSLS Estimator

In the problem of (6), $\mathbf{a} = \mathbf{y}'$, $\mathbf{C}_a = \mathbf{C}_w$, $\mathbf{W} = \mathbf{HG}$, and $\mathbf{Q} = \mathbf{HRH}^*$. Denoting by $\widetilde{\mathbf{G}} = (1/c)\mathbf{G}$, we then have from Theorem 1 that the optimal value of $\widetilde{\mathbf{G}}$, which is denoted $\widehat{\widetilde{\mathbf{G}}}$, satisfies

$$\mathbf{H}\widehat{\widetilde{\mathbf{G}}} = \left(\mathbf{HRH}^*\mathbf{C}_w^{-1}\right)^{1/2}. \tag{17}$$

Using straightforward matrix manipulations, we show in Appendix B that

$$\begin{aligned}\widehat{\widetilde{\mathbf{G}}} &= \mathbf{R}\left(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{HR}\right)^{-1/2}\mathbf{H}^*\mathbf{C}_w^{-1} \\ &= \left(\mathbf{RH}^*\mathbf{C}_w^{-1}\mathbf{H}\right)^{-1/2}\mathbf{RH}^*\mathbf{C}_w^{-1}.\end{aligned} \tag{18}$$

If the scaling c in (7) is specified, then the CSLS estimator is given by

$$\hat{\mathbf{x}}_{\mathrm{CSLS}} = c\widehat{\widetilde{\mathbf{G}}}\mathbf{y}. \tag{19}$$

If c is chosen to minimize $\varepsilon_{\mathrm{CSLS}}$, then

$$\hat{\mathbf{x}}_{\mathrm{CSLS}} = \hat{c}\hat{\widehat{\mathbf{G}}}\mathbf{y} \tag{20}$$

where from Theorem 1

$$\hat{c} = \frac{\mathrm{Tr}\left(\left(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\right)^{1/2}\right)}{\mathrm{Tr}\left(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\right)}. \tag{21}$$

Note from (19) and (20) that $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ is a biased estimator of $\mathbf{x}$ so that when $\sigma^2 \to 0$, $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ does not converge to $\mathbf{x}$. At very high SNR, we therefore expect the LS estimator to perform better than the CSLS estimator. The advantage of the CSLS is at low to moderate SNR, where we reduce the MSE of the estimator by allowing for a biased estimator. Indeed, as we show in Section VI, for many choices of $\mathbf{R}$, regardless of the value of $\mathbf{x}$, there is always a threshold SNR, so that for SNR values below this threshold, the CSLS estimator yields a lower MSE than the LS estimator. As we show in [18], in applications, this threshold value can be pretty large.

Since the covariance of the LS estimate is given from (3) by $(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}$ and the covariance of the CSLS estimate is proportional to $c^2\mathbf{R}$, it follows immediately that $\hat{\mathbf{x}}_{\mathrm{LS}}$ can be equal to $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ only if $(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}$ is proportional to $\mathbf{R}$. In fact, using the CSLS estimator of (20), we have that $\hat{\mathbf{x}}_{\mathrm{LS}} = \hat{\mathbf{x}}_{\mathrm{CSLS}}$ if and only if $(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1} = d^2\mathbf{R}$ for some $d > 0$. Indeed, if $(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1} = d^2\mathbf{R}$, then $\hat{\mathbf{x}}_{\mathrm{LS}} = d^2\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y}$, and

$$\hat{\mathbf{x}}_{\mathrm{CSLS}} = \hat{c}(\mathbf{R}(d^2\mathbf{R})^{-1})^{-1/2}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y} = \hat{c}d\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y}. \tag{22}$$

From (21), $\hat{c} = d^{-1}/d^{-2} = d$ so that for any choice of d, $\hat{\mathbf{x}}_{\mathrm{CSLS}} = \hat{\mathbf{x}}_{\mathrm{LS}}$.

Finally, we note that the CSLS estimator of (20) is invariant to an overall gain in $\mathbf{C}_w$. Thus, if $\mathbf{C}_w = \sigma^2\mathbf{C}$ for some covariance matrix $\mathbf{C}$, then the CSLS estimator does not depend on σ. This property does not hold in the case in which c is chosen as a constant, independent of σ. In this case, the CSLS estimator depends explicitly on σ, which therefore must be known. Alternatively, if we let $c = \sigma$, then the CSLS estimator will not depend on σ, which might be unknown. We conclude that in the case in which the variance is unknown, we must either use $c = \hat{c}$ of (21) or $c = \sigma$.

The CSLS estimator is summarized in the following theorem.

Theorem 2 (CSLS Estimator): Let $\mathbf{x}$ denote the deterministic unknown parameters in the model $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$, where $\mathbf{H}$ is a known $n \times m$ matrix with rank m, and $\mathbf{w}$ is a zero-mean random vector with covariance $\mathbf{C}_w$. Let $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ denote the covariance shaping least-squares estimator of $\mathbf{x}$ that minimizes the error (6) subject to (7) for some $c > 0$. Then

$$\begin{aligned}\hat{\mathbf{x}}_{\mathrm{CSLS}} &= \beta\mathbf{R}(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\mathbf{R})^{-1/2}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y} \\ &= \beta(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1/2}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y}\end{aligned}$$

where we have the following.

1) If c is specified, then $\beta = c$.
2) If c is chosen to minimize the error, then $\beta = \hat{c}$ given by (21).

Furthermore, with $\beta = \hat{c}$, the least-squares estimate $\hat{\mathbf{x}}_{\mathrm{LS}}$ is equal to $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ if and only if $\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H} = d^2\mathbf{R}$ for some $d > 0$.

IV. Connection With Other Modifications of Least-Squares

In this section, we compare the CSLS estimator with the ridge estimator proposed by Hoerl and Kennard [11] and Tikhonov [12], as well as with the shrunken estimator proposed by Mayer and Willke [13]. In Section IX, we discuss a performance comparison in the context of a specific application.

The ridge estimator for the linear model (1), which is denoted by $\hat{\mathbf{x}}_R$, is defined by

$$\hat{\mathbf{x}}_R = \left(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H} + \delta\mathbf{I}_m\right)^{-1}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y} \tag{23}$$

where δ is a regularization parameter. It can be shown that $\hat{\mathbf{x}}_R$ minimizes the LS error subject to a constraint on the norm of $\hat{\mathbf{x}}_R$. Thus, for all estimators with fixed norm, $\hat{\mathbf{x}}_R$ given by (23) minimizes the LS error, where δ is chosen to satisfy the norm constraint.

To show that $\hat{\mathbf{x}}_R$ is equal to a CSLS estimator with an appropriate choice of $\mathbf{R}$, let $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ be the CSLS estimator with covariance $\mathbf{R}_R$, where $\mathbf{R}_R$ is the covariance of the estimate $\hat{\mathbf{x}}_R$ and is given by $\mathbf{R}_R = (\mathbf{I}_m + \delta(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{I}_m)^{-1}(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H} + \delta\mathbf{I}_m)^{-1}$. By direct substitution of $\mathbf{R}_R$ into the expression for $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ from Theorem 2, $\hat{\mathbf{x}}_{\mathrm{CSLS}} = \hat{\mathbf{x}}_R$. Based on this connection between the ridge estimator and the CSLS estimator, we may interpret the ridge estimator as the estimator that minimizes the error $\varepsilon_{\mathrm{CSLS}}$ given by (6) from all estimators with covariance $\mathbf{R}_R$.

The shrunken estimator for the linear model (1), which is denoted by $\hat{\mathbf{x}}_S$, is a scaled version of the LS estimator and is defined by

$$\hat{\mathbf{x}}_S = \kappa\hat{\mathbf{x}}_{\mathrm{LS}} = \kappa\left(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y} \tag{24}$$

where κ is a regularization parameter. A stochastically (nonlinear) shrunken estimator is a shrunken estimator in which κ is a function of the data $\mathbf{y}$, an example of which is the well-known James–Stein estimator [10].

The shrunken estimator $\hat{\mathbf{x}}_S$ can be formulated as a CSLS estimator where the covariance of $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ is chosen to be equal to the covariance of $\hat{\mathbf{x}}_S$ given by $\mathbf{R}_S = \kappa^2(\mathbf{H}^*\mathbf{C}_w\mathbf{H})^{-1}$. Substituting $\mathbf{R}_S$ into the expression for $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ from Theorem 2, we indeed have that $\hat{\mathbf{x}}_{\mathrm{CSLS}} = \hat{\mathbf{x}}_S$. Thus, we may interpret $\hat{\mathbf{x}}_S$ as the estimator that minimizes the error $\varepsilon_{\mathrm{CSLS}}$ of (6) from all estimators with covariance $\mathbf{R}_S$.

In summary, some of the more popular alternatives to the LS estimator under the model (1) can be interpreted within the framework of CSLS estimators. This provides additional insight and further optimality properties of these estimators. However, the CSLS estimator is more general since we are not constrained to a specific choice of covariance $\mathbf{R}$. By choosing $\mathbf{R}$ to "best" shape the estimator covariance in some sense, we can improve the performance of the estimator over these LS alternatives.

As a final note, suppose we are given an arbitrary linear estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ for which the covariance of the error is $\mathbf{C}_x$. Then, we can compute the CSLS estimate $\hat{\mathbf{x}}_{\mathrm{CSLS}}$ with $\mathbf{R} = \mathbf{C}_x$. If $\hat{\mathbf{x}}_{\mathrm{CSLS}} = \hat{\mathbf{x}}$, then the estimate $\hat{\mathbf{x}}$ has the additional property that from all estimators with covariance $\mathbf{C}_x$, it minimizes the (weighted) total error variance in the observations. If, on the

other hand, $\hat{\mathbf{x}}_{\mathrm{CSLS}} \neq \hat{\mathbf{x}}$, then we can always improve the total error variance of the estimate without altering its covariance by using $\hat{\mathbf{x}}_{\mathrm{CSLS}}$. Therefore, an estimate with covariance $\mathbf{C}_x$ is said to be consistent with the total error variance criterion if it minimizes this criterion from all estimators with covariance $\mathbf{C}_x$, in which case, it is equal to the CSLS estimate with $\mathbf{R} = \mathbf{C}_x$.

V. Cramér–Rao Lower Bound

The variance of an unbiased estimator $\hat{\mathbf{x}}$ of the unknown parameters $\mathbf{x}$ can be bounded by the CRLB [2], [14]. A similar bound is also given for the variance of a biased estimator, which is known as the biased CRLB [15]. Specifically, suppose we want to estimate a set of unknown deterministic parameters $\mathbf{x}$ from some given observations $\mathbf{y}$. Let $p(\mathbf{y}, \mathbf{x})$ denote the probability density function of the observations $\mathbf{y}$, which is characterized by $\mathbf{x}$. It is assumed that $p(\mathbf{y}, \mathbf{x})$ satisfies the regularity condition $E(\partial p(\mathbf{y}, \mathbf{x})/\partial \mathbf{x}) = 0$. Then, for any estimator $\hat{\mathbf{x}}$ of $\mathbf{x}$ with bias $B(\mathbf{x})$, the covariance of the estimator must satisfy

$$E((\hat{\mathbf{x}} - E(\hat{\mathbf{x}}))(\hat{\mathbf{x}} - E(\hat{\mathbf{x}}))^*) \geq \left(\mathbf{I}_m + \frac{\partial B(\mathbf{x})}{\partial \mathbf{x}}\right) J^{-1}(\mathbf{x}) \left(\mathbf{I}_m + \frac{\partial B(\mathbf{x})}{\partial \mathbf{x}}\right)^* \quad (25)$$

where $J(\mathbf{x})$ is the Fisher information matrix defined by

$$J(\mathbf{x}) = -E\left(\frac{\partial^2 \log p(\mathbf{y}, \mathbf{x})}{\partial \mathbf{x}^2}\right). \quad (26)$$

For the CSLS estimator, the bias is given by

$$B(\hat{\mathbf{x}}_{\mathrm{CSLS}}) = \left(\beta \left(\mathbf{R}\mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{H}\right)^{1/2} - \mathbf{I}_m\right) \mathbf{x} \quad (27)$$

and

$$\frac{\partial B(\hat{\mathbf{x}}_{\mathrm{CSLS}})}{\partial \mathbf{x}} = \beta \left(\left(\mathbf{R}\mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{H}\right)^{1/2}\right) - \mathbf{I}_m. \quad (28)$$

We now show that if the noise $\mathbf{w}$ in (1) is Gaussian with zero-mean and covariance $\mathbf{C}_w$, then the CSLS estimator achieves the CRLB for biased estimators $\hat{\mathbf{x}}$ with bias $B(\hat{\mathbf{x}})$ given by (27).

For the linear model of (1) with Gaussian noise, the Fisher information matrix is [2] $J(\mathbf{x}) = \mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{H}$. The CRLB on the variance of any estimator with bias $B(\hat{\mathbf{x}}_{\mathrm{CSLS}})$ is therefore given by

$$\begin{aligned} &E\left((\hat{\mathbf{x}} - E(\hat{\mathbf{x}}))(\hat{\mathbf{x}} - E(\hat{\mathbf{x}}))^*\right) \\ &\geq \beta^2 \left(\mathbf{R}\mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{H}\right)^{1/2} \left(\mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{H}\right)^{-1} \left(\mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{H}\mathbf{R}\right)^{1/2} \\ &= \beta^2 \left(\mathbf{R}\mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{H}\right)^{1/2} \left(\mathbf{R}\mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{H}\right)^{1/2} \left(\mathbf{H}^* \mathbf{C}_w^{-1} \mathbf{H}\right)^{-1} \\ &= \beta^2 \mathbf{R}. \end{aligned} \quad (29)$$

Now, for the CSLS estimator, $E(\hat{\mathbf{x}}_{\mathrm{CSLS}} - E(\hat{\mathbf{x}}_{\mathrm{CSLS}}))(\hat{\mathbf{x}}_{\mathrm{CSLS}} - E(\hat{\mathbf{x}}_{\mathrm{CSLS}}))^*) = \beta^2 \mathbf{R}$ so that the CRLB is achieved. Thus, from all estimators with bias given by (27) for some β and $\mathbf{R}$, the CSLS estimator minimizes the variance.

VI. Mean-Squared Error Performance

In Section V, we showed that the CSLS estimator minimizes the MSE among all estimators with a particular bias. While it would be desirable to analyze the MSE of the CSLS estimator for more general forms of bias, we cannot directly evaluate the MSE of the CSLS estimator since the bias and, consequently, the MSE depend explicitly on the unknown parameters $\mathbf{x}$. To gain some additional insight into the performance of the CSLS estimator, in this section, we instead compare its MSE with the MSE of the LS estimator. Our analysis indicates that there are many cases in which the CSLS estimator performs better than the LS estimator in a MSE sense for all values of the unknown parameters $\mathbf{x}$.

The approach we take in this section is to directly compare the MSE of the CSLS and the LS estimators and show that for a variety of choices of the output covariance $\mathbf{R}$, there is a threshold SNR such that for SNR values below this threshold, the CSLS estimator yields a lower MSE than the LS estimator for all values of $\mathbf{x}$. In our analysis, we assume that $\mathbf{C}_w = \sigma^2 \mathbf{C}$, where the diagonal elements of $\mathbf{C}$ are all equal to 1 so that the variance of each of the noise components of $\mathbf{C}_w$ is σ^2. To ensure that the estimator does not depend on σ, which may not be known, we let the scaling of the CSLS estimator be $\beta = \sigma$ or $\beta = \hat{c}$, which is given by (21).

The detailed analysis related to this discussion is carried out in Appendix C. Here, we focus on the interpretation of the results developed in the Appendix.

A. Fixed Scaling

We first consider the case in which $\beta = \sigma$. The MSE of the CSLS estimator is then given by

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) = \|((\mathbf{R}\mathbf{B})^{1/2} - \mathbf{I}_m)\mathbf{x}\|^2 + \sigma^2 \mathrm{Tr}(\mathbf{R}) \quad (30)$$

where $\mathbf{B} = \mathbf{H}^* \mathbf{C}^{-1} \mathbf{H}$. The first term in (30) is the squared norm of the bias of the estimate $\hat{\mathbf{x}}_{\mathrm{CSLS}}$, and the second term in (30) is the total variance of $\hat{\mathbf{x}}_{\mathrm{CSLS}}$.

For large values of σ^2 in comparison with $\|\mathbf{x}\|^2$, the first term in (30) is negligible, and $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \approx \sigma^2 \mathrm{Tr}(\mathbf{R})$. Thus, at sufficiently low SNR, where the SNR is defined as $\|\mathbf{x}\|^2/\sigma^2$, both $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$ and $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}})$ are proportional to σ^2, where we can always choose $\mathbf{R}$ so that the proportionality constant $\mathrm{Tr}(\mathbf{R})$ of the CSLS estimator is smaller than the proportionality constant $\mathrm{Tr}(\mathbf{B}^{-1})$ of the LS estimator. At sufficiently high SNR, the second term in (30) can be considered negligible and as $\sigma \to 0$, $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}})$ converges to the constant $\|((\mathbf{R}\mathbf{B})^{1/2} - \mathbf{I}_m)\mathbf{x}\|^2$. From this qualitative analysis, it is clear that there is a threshold SNR that will depend in general on $\mathbf{x}$ below which, for appropriate choices of $\mathbf{R}$, the CSLS estimator outperforms the LS estimator.

In Appendix C, we show that if $\hat{\mathbf{x}}_{\mathrm{LS}} \neq \hat{\mathbf{x}}_{\mathrm{CSLS}}$, then $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \leq \mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$ if

$$\zeta \leq \frac{\mathrm{Tr}(\mathbf{B}^{-1}) - \mathrm{Tr}(\mathbf{R})}{\sigma_\gamma} \triangleq \tilde{\zeta}_{\mathrm{WC}} \quad (31)$$

where $\zeta = \|\mathbf{x}\|^2/(\sigma^2 m)$ denotes the SNR per component, $\mathbf{B} = \mathbf{H}^* \mathbf{C}^{-1} \mathbf{H}$, $\gamma = \arg\max \sigma_k$, and σ_k are the eigenvalues of $\mathbf{Q} = ((\mathbf{R}\mathbf{B})^{1/2} - \mathbf{I}_m)^*((\mathbf{R}\mathbf{B})^{1/2} - \mathbf{I}_m)$. The bound $\tilde{\zeta}_{\mathrm{WC}}$ given by (31) is a worst-case bound since it corresponds to

the worst possible choice of parameters, namely, when the unknown vector $\mathbf{x}$ is in the direction of the eigenvector of $\mathbf{Q}$ corresponding to the eigenvalue σ_γ. In practice, the CSLS estimator will outperform the LS estimator for higher values of SNR than $\tilde{\zeta}_{\rm WC}$.

Since we have freedom in designing $\mathbf{R}$, we may always choose $\mathbf{R}$ so that $\tilde{\zeta}_{\rm WC} > 0$. In this case, we are guaranteed that there is a range of SNR values for which the CSLS estimator leads to a lower MSE than the LS estimator for all choices of the unknown parameters $\mathbf{x}$.

For example, suppose we wish to design an estimator with covariance proportional to some given covariance matrix $\mathbf{Z}$ so that $\mathbf{R} = a\mathbf{Z}$ for some $a > 0$. If we choose $a < \mathrm{Tr}(\mathbf{B}^{-1})/\mathrm{Tr}(\mathbf{Z})$, then we are guaranteed that there is an SNR range for which the CSLS estimator will have a lower MSE than the LS estimator for all values of $\mathbf{x}$.

In specific applications, it may not be obvious how to choose a particular proportionality factor a. In such cases, we may prefer using the CSLS estimator with optimal scaling, which we discuss in Section VI-B.

B. Optimal Scaling

In cases in which there is no natural scaling, it may be preferable to use the CSLS estimator with optimal scaling. In this case, the scaling is a function of $\mathbf{R}$ cannot be chosen arbitrarily, so that in general, we can no longer guarantee that there is a positive SNR threshold, i.e., that there is always an SNR range over which the CSLS performs better than the LS estimator. However, as we show, in the special case in which $\mathbf{R} = \mathbf{I}_m$, there is always such an SNR range.

If $\beta = \hat{c}$ and $\mathbf{R} = \mathbf{I}_m$, then (see Appendix C)

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) = \|(\alpha\mathbf{B}^{1/2} - \mathbf{I}_m)\mathbf{x}\|^2 + m\alpha^2\sigma^2 \tag{32}$$

where

$$\alpha = \frac{\hat{c}}{\sigma} = \frac{\mathrm{Tr}(\mathbf{B}^{1/2})}{\mathrm{Tr}(\mathbf{B})} = \frac{\sum_{i=1}^{m}\lambda_i^{1/2}}{\sum_{i=1}^{m}\lambda_i} \tag{33}$$

and $\lambda_i, 1 \le i \le m$ denote the eigenvalues of $\mathbf{B} = \mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$.

For sufficiently large values of σ^2 in comparison with $\|\mathbf{x}\|^2$, we can consider the first term in (32) negligible and $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) \approx m\alpha^2\sigma^2$. Thus, at sufficiently low SNR, both $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm LS})$ and $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS})$ are proportional to σ^2 and, as we show in Appendix C, the proportionality constant $m\alpha^2$ of the CSLS estimator is smaller than the proportionality constant $\mathrm{Tr}(\mathbf{B}^{-1})$ of the LS estimator. At sufficiently high SNR, the second term in (30) can be considered negligible, and as $\sigma \to 0$, $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS})$ converges to the constant $\|(\alpha\mathbf{B}^{1/2} - \mathbf{I}_m)\mathbf{x}\|^2$. These trends in the behavior of the MSE can be seen in the simulations in Section IX.

In Appendix C, we show that if $\hat{\mathbf{x}}_{\rm LS} \neq \hat{\mathbf{x}}_{\rm CSLS}$, then $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) \le \mathrm{MSE}(\hat{\mathbf{x}}_{\rm LS})$ if

$$\zeta \le \frac{\left(\frac{1}{m}\right)\sum_{k=1}^{m}\lambda_k^{-1} - \alpha^2}{\left|\alpha\lambda_\gamma^{1/2} - 1\right|^2} \triangleq \zeta_{\rm WC} \tag{34}$$

where $\gamma = \arg\max|\alpha\lambda_k^{1/2} - 1|^2$. We also show that when $\hat{\mathbf{x}}_{\rm CSLS} \neq \hat{\mathbf{x}}_{\rm LS}$, $\zeta_{\rm WC} > 0$ so that there is always a range of SNR values for which $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) \le \mathrm{MSE}(\hat{\mathbf{x}}_{\rm LS})$.

The bound $\zeta_{\rm WC}$ given by (34) is a worst-case bound since it corresponds to the worst possible choice of parameters. In practice, the CSLS estimator will outperform the LS estimator for higher values of SNR than $\zeta_{\rm WC}$.

In a similar manner, we show that $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) \ge \mathrm{MSE}(\hat{\mathbf{x}}_{\rm LS})$ if

$$\zeta \ge \frac{\left(\frac{1}{m}\right)\sum_{k=1}^{m}\lambda_k^{-1} - \alpha^2}{\left|\alpha\lambda_\kappa^{1/2} - 1\right|^2} \triangleq \zeta_{\rm BC} \tag{35}$$

where $\kappa = \arg\min|\alpha\lambda_k^{1/2} - 1|^2$.

The performance analysis of the CSLS estimator $\hat{\mathbf{x}}_{\rm CSLS}$ with optimal scaling in the case in which $\mathbf{R} = \mathbf{I}_m$ and $\mathbf{C}_w = \sigma^2\mathbf{C}$ can be summarized as follows: Let $\zeta = \|\mathbf{x}\|^2/(\sigma^2 m)$ denote the SNR per component. Then, with $\{\lambda_k, 1 \le k \le m\}$ denoting the eigenvalues of $\mathbf{B} = \mathbf{H}^*\mathbf{C}^{-1}\mathbf{H}$, and α given by (33), we have the following.

1) $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) \le \mathrm{MSE}(\hat{\mathbf{x}}_{\rm LS})$ for $\zeta \le \zeta_{\rm WC}$, where $\zeta_{\rm WC} > 0$ is the worst-case bound given by (34).
2) $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) \ge \mathrm{MSE}(\hat{\mathbf{x}}_{\rm LS})$ for $\zeta \ge \zeta_{\rm BC}$, where $\zeta_{\rm BC}$ is the best-case bound given by (35).
3) $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS})$ may be smaller or larger than $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm LS})$ for $\zeta_{\rm WC} \le \zeta \le \zeta_{\rm BC}$, depending on the value of $\mathbf{x}$. Thus, the true threshold value in a particular application will be between $\zeta_{\rm WC}$ and $\zeta_{\rm BC}$.

In addition, we have the following.

1) If $\mathbf{x}$ is in the direction of the eigenvector of $\mathbf{B}$ corresponding to the eigenvalue λ_γ, then $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) \le \mathrm{MSE}(\hat{\mathbf{x}}_{\rm LS})$ for $\zeta \le \zeta_{\rm WC}$, and $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) \ge \mathrm{MSE}(\hat{\mathbf{x}}_{\rm LS})$ for $\zeta \ge \zeta_{\rm WC}$.
2) If $\mathbf{x}$ is in the direction of the eigenvector of $\mathbf{B}$ corresponding to the eigenvalue λ_κ, then $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) \le \mathrm{MSE}(\hat{\mathbf{x}}_{\rm LS})$ for $\zeta \le \zeta_{\rm BC}$, and $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) \ge \mathrm{MSE}(\hat{\mathbf{x}}_{\rm LS})$ for $\zeta \ge \zeta_{\rm BC}$.

In [18], we consider some examples illustrating the threshold values for different matrices $\mathbf{B} = \mathbf{H}^*\mathbf{C}^{-1}\mathbf{H}$, where $\mathbf{C}_w = \sigma^2\mathbf{C}$. These examples indicate that in a variety of applications, the threshold values are pretty large, as can also be seen from the simulations in Section IX.

VII. LS Followed by WMMSE Shaping

The CSLS was derived to minimize the total variance in the data error subject to a constraint on the covariance of the estimator of $\mathbf{x}$. In this section, we show that the CSLS estimator can alternatively be expressed as a LS estimator followed by a weighted minimum mean-squared error (WMMSE) covariance shaping transformation.

Specifically, suppose we estimate the parameters $\mathbf{x}$ using the LS estimator $\hat{\mathbf{x}}_{\rm LS}$. Since $\hat{\mathbf{x}}_{\rm LS} = \mathbf{x} + \tilde{\mathbf{w}}$, where $\tilde{\mathbf{w}} = (\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{H}^*\mathbf{C}_w\mathbf{w}$, the covariance of the noise component $\tilde{\mathbf{w}}$ in $\hat{\mathbf{x}}_{\rm LS}$ is equal to the covariance of $\hat{\mathbf{x}}_{\rm LS}$, which is denoted $\mathbf{C}_{\hat{\mathbf{x}}_{\rm LS}}$ and is given by $\mathbf{C}_{\hat{\mathbf{x}}_{\rm LS}} = (\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}$. To improve the performance of the LS estimator, we consider

shaping the covariance of the noise component in the estimator $\hat{\mathbf{x}}_{\rm LS}$. Thus, we seek a transformation $\mathbf{W}$ such that the covariance matrix of $\hat{\mathbf{x}} = \mathbf{W}\hat{\mathbf{x}}_{\rm LS}$, which is denoted by $\mathbf{C}_{\hat{x}}$, satisfies

$$\mathbf{C}_{\hat{x}} = \mathbf{W}\mathbf{C}_{\hat{\mathbf{x}}_{\rm LS}}\mathbf{W}^* = c^2\mathbf{R} \tag{36}$$

for some $c > 0$. To minimize the distortion to the estimator $\hat{\mathbf{x}}_{\rm LS}$, from all possible transformations $\mathbf{W}$ satisfying (36), we choose the one that minimizes the weighted MSE

$$E\left((\hat{\mathbf{x}}'_{\rm LS} - \mathbf{W}\hat{\mathbf{x}}'_{\rm LS})^* \mathbf{C} (\hat{\mathbf{x}}'_{\rm LS} - \mathbf{W}\hat{\mathbf{x}}'_{\rm LS})\right) \tag{37}$$

where $\mathbf{C}$ is an arbitrary weighting matrix.

We now show that if we choose $\mathbf{C} = \mathbf{C}^{-1}_{\hat{\mathbf{x}}_{\rm LS}}$ in (37), then the resulting estimator $\hat{\mathbf{x}} = \mathbf{W}\hat{\mathbf{x}}_{\rm LS}$ is equal to $\hat{\mathbf{x}}_{\rm CSLS}$. Note that this choice of weighting matrix is reminiscent of the Gauss–Markov weighting in LS estimation [2]. The minimization problem of (37) with $\mathbf{C} = \mathbf{C}^{-1}_{\hat{\mathbf{x}}_{\rm LS}}$ is a special case of the general WMMSE shaping problem discussed in Section III-A with $\mathbf{a} = \hat{\mathbf{x}}'_{\rm LS}$, $\mathbf{C}_a = (\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}$, and $\mathbf{b} = \hat{\mathbf{x}}$. Thus, from Theorem 1

$$\begin{aligned}\hat{\mathbf{x}} =& \hat{c}\mathbf{R}\left(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\mathbf{R}\right)^{-1/2}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\hat{\mathbf{x}}_{\rm LS}\\ =& \hat{c}\mathbf{R}\left(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\mathbf{R}\right)^{-1/2}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y}\end{aligned} \tag{38}$$

and

$$\hat{c} = \frac{\mathrm{Tr}\left(\left(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\right)^{1/2}\right)}{\mathrm{Tr}\left(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\right)}. \tag{39}$$

Comparing (38) with $\hat{\mathbf{x}}_{\rm CSLS}$ given by Theorem 2, we conclude that $\hat{\mathbf{x}} = \hat{\mathbf{x}}_{\rm CSLS}$ so that the CSLS estimator can be determined by first finding the LS estimator $\hat{\mathbf{x}}_{\rm LS}$ and then optimally shaping its covariance.

In [18]–[20], a new linear multiuser receiver for synchronous code-division multiple-access (CDMA) systems, which is referred to as the orthogonal multiuser receiver, was proposed. In a CDMA system, the received signal $\mathbf{r}$ is modeled as $\mathbf{r} = \mathbf{S}\mathbf{A}\mathbf{b} + \mathbf{w}$, where $\mathbf{S}$ is a known matrix of signature vectors, $\mathbf{A}$ is a diagonal matrix of received amplitudes, $\mathbf{b}$ is the data vector, and $\mathbf{w}$ is a noise vector. Linear multiuser receivers consist of an estimator of $\mathbf{x} = \mathbf{A}\mathbf{b}$ followed by a channel decoder. The well-known decorrelator receiver [21] is based on a least-squares estimate of $\mathbf{x}$. The orthogonal multiuser receiver is designed to optimally whiten the output of the decorrelator receiver prior to detection and is therefore a special case of the CSLS estimator with $\mathbf{R} = \mathbf{I}_m$. Therefore, the performance properties of the CSLS estimator can now be used to establish optimality properties of the orthogonal multiuser receiver.

VIII. Matched Correlator Estimator Followed by MMSE Shaping

We now show that the CSLS estimator with fixed scaling can also be expressed as a matched correlator estimator followed by MMSE shaping. Consider estimating the parameters $\mathbf{x}$ using the transformation $\hat{\mathbf{x}}_{\rm MC} = \mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y}$. Then, the covariance of the noise component in $\hat{\mathbf{x}}_{\rm MC}$, which is equal to the covariance of $\hat{\mathbf{x}}_{\rm MC}$, is $\mathbf{C}_{\hat{\mathbf{x}}_{\rm MC}} = \mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}$. To improve the performance of $\hat{\mathbf{x}}_{\rm MC}$, we consider shaping its covariance so that we seek a transformation $\mathbf{T}$ such that the covariance matrix of $\hat{\mathbf{x}} = \mathbf{T}\hat{\mathbf{x}}_{\rm MC}$, which is denoted by $\mathbf{C}_{\hat{x}}$, satisfies

$$\mathbf{C}_{\hat{x}} = \mathbf{T}\mathbf{C}_{\hat{\mathbf{x}}_{\rm MC}}\mathbf{T}^* = c^2\mathbf{R} \tag{40}$$

where c is given. To minimize the distortion to the estimator $\hat{\mathbf{x}}_{\rm MC}$, from all possible transformations $\mathbf{T}$ satisfying (40), we choose the one that minimizes the MSE

$$E\left((\hat{\mathbf{x}}'_{\rm MC} - \mathbf{T}\hat{\mathbf{x}}'_{\rm MC})^* (\hat{\mathbf{x}}'_{\rm MC} - \mathbf{T}\hat{\mathbf{x}}'_{\rm MC})\right) \tag{41}$$

where $\hat{\mathbf{x}}'_{\rm MC} = \hat{\mathbf{x}}_{\rm MC} - E(\hat{\mathbf{x}}_{\rm MC})$.

This minimization problem is a special case of the general MMSE shaping problem considered in [18], from which it follows that

$$\begin{aligned}\hat{\mathbf{x}} =& c\left(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\right)^{-1/2}\mathbf{R}\hat{\mathbf{x}}_{\rm MC}\\ =& c\left(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\right)^{-1/2}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{y}.\end{aligned} \tag{42}$$

Comparing (42) with $\hat{\mathbf{x}}_{\rm CSLS}$ given by Theorem 2, we conclude that $\hat{\mathbf{x}} = \hat{\mathbf{x}}_{\rm CSLS}$ so that the CSLS estimator with fixed scaling can be determined by first finding the matched correlator estimator $\hat{\mathbf{x}}_{\rm MC}$ and then optimally shaping its covariance. The optimal scaling can be found by choosing c to minimize (6) with $\mathbf{G} = c(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1/2}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}$.

In [17] and [22], a modification of the well-known matched filter (MF) detector, which is referred to as the orthogonal MF detector, was proposed. The orthogonal MF detector is obtained by MMSE whitening of the MF output, which is equivalent to a matched correlator. Therefore, when scaled appropriately, the orthogonal MF implements a CSLS estimator with $\mathbf{R} = \mathbf{I}_m$. In [17] and [22], the orthogonal MF was proposed as an *ad-hoc* detector. The performance properties of the CSLS estimator can now be used to establish optimality properties of this detector.

IX. Applications

In this section we consider two applications of CSLS estimation. The first is to the estimation of parameters in an ARMA model. The second is to the estimation of amplitudes in exponential signal models.

A. System Identification

As one application of CSLS estimation, we consider the problem of estimating the parameters in an ARMA model, and compare the estimated parameters to those obtained by using the modified Yule-Walker equations in combination with Shanks' method [3], [23].

Suppose we are given a finite segment of noisy measurements of an ARMA signal $x[l]$, which is defined by

$$x[l] = \sum_{k=1}^{p} a_k x[l-k] + \sum_{k=0}^{q} b_k \delta[l-k] \tag{43}$$

for some coefficients a_k and b_k, where $q < p$. The coefficients a_k in (43) are the AR parameters of $x[l]$, and the coefficients b_k in (43) are the the MA parameters. The z-transform of $x[l]$ is

$$X(z) = \frac{b_0 + b_1 z^{-1} + \ldots + b_q z^{-q}}{1 + a_1 z^{-1} + \ldots + a_p z^{-p}} \triangleq B(z)H(z) \tag{44}$$

where $B(z)$ denotes the numerator polynomial, and $H(z)$ denotes the inverse of the denominator polynomial. The problem then is to estimate the AR and MA parameters from the data $y[0],\ldots,y[n-1]$, where

$$y[l]=x[l]+w[l]=\sum_{k=1}^{p}a_k x[l-k]+\sum_{k=0}^{q}b_k\delta[l-k]+w[l],\quad 0\le l\le n-1. \tag{45}$$

Here, $w[l]$ represents a combination of measurement noise and modeling error. In the simulations below, $w[l]$ is chosen as a zero-mean Gaussian noise process with variance σ^2.

Various methods exist for estimating the ARMA parameters based on different applications of LS estimation [3]. A popular method is to estimate the AR parameters using the modified Yule–Walker equations [3] and then use these estimates in combination with Shanks' method [22] to estimate the MA parameters. We use this method as a basis for comparison with our method.

From (45), it follows that

$$y[l]=\sum_{k=1}^{p}a_k x[l-k]+w[l],\quad q<l\le n-1. \tag{46}$$

We now use (46) to estimate the AR parameters a_k. Since we do not have access to the clean data $x[l-k]$, we estimate a_k by substituting $y[l-k]$ instead of $x[l-k]$ in (46). Then, with $\mathbf{a}$ denoting the vector with components $a_k, 1\le k\le p$, $\mathbf{y}$ denoting the data vector with components $y[l], p\le l\le n-1$, $\mathbf{w}$ denoting the vector with components $w[l], p\le l\le n-1$, and

$$\mathbf{H}_{\mathrm{AR}}=\begin{bmatrix} y[p-1] & y[p-2] & \cdots & y[0] \\ y[p] & y[p-1] & \cdots & y[1] \\ \vdots & \vdots & & \vdots \\ y[n-2] & y[n-3] & \cdots & y[n-p-1] \end{bmatrix} \tag{47}$$

we have that $\mathbf{y}\approx\mathbf{H}_{\mathrm{AR}}\mathbf{a}+\mathbf{w}$. The LS estimate of the AR parameters is then

$$\hat{\mathbf{a}}_{\mathrm{LS}}=(\mathbf{H}_{\mathrm{AR}}^{*}\mathbf{H}_{\mathrm{AR}})^{-1}\mathbf{H}_{\mathrm{AR}}^{*}\mathbf{y}. \tag{48}$$

From Theorem 2, the CSLS estimate of the AR parameters is

$$\hat{\mathbf{a}}_{\mathrm{CSLS}}=\hat{c}(\mathbf{R}\mathbf{H}_{\mathrm{AR}}^{*}\mathbf{H}_{\mathrm{AR}})^{-1/2}\mathbf{R}\mathbf{H}_{\mathrm{AR}}^{*}\mathbf{y} \tag{49}$$

where $\hat{c}$ is given by (21).

We now use these estimates of $\mathbf{a}$ to estimate the MA parameters using Shanks' method. Specifically, let $e[l]=y[l]-h[l]*b[l]$, where $h[l]$ is the impulse response of the filter with z-transform $H(z)$, which is computed using the estimates of the AR parameters, and $b[l]$ is the (unknown) impulse response of the filter with z-transform $B(z)$. Shanks proposed estimating the unknown sequence $b[l]$ by minimizing $\sum_{l=0}^{n-1}e^2[l]$. With $\mathbf{e}$ denoting the error vector with components $e[l], 0\le l\le n-1$, we have that $\mathbf{e}=\mathbf{y}-\mathbf{H}_{\mathrm{MA}}\mathbf{b}$, where $\mathbf{b}$ is the vector with components b_k, $1\le k\le q$, $\mathbf{y}$ is the data vector with components $y[l]$, $0\le l\le n-1$, and

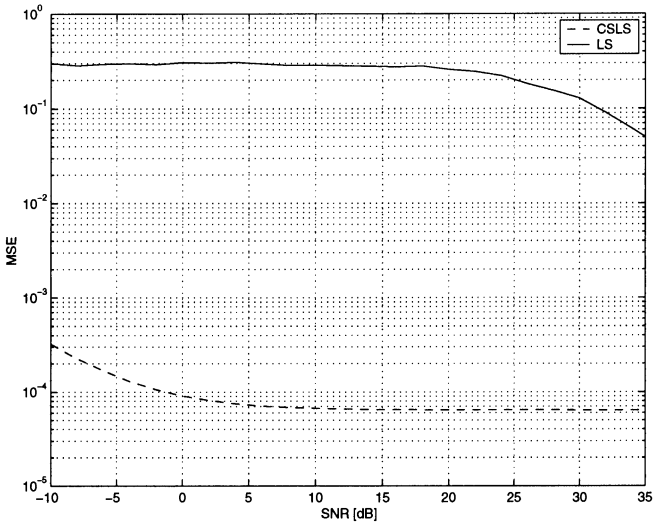

Fig. 1. Mean-squared error in estimating the AR parameters a_k given by (53) using the LS estimator (48) and the CSLS estimator (49).

$$\mathbf{H}_{\mathrm{MA}}=\begin{bmatrix} h[0] & 0 & \cdots & 0 \\ h[1] & h[0] & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ h[n-1] & h[n-2] & \cdots & h[n-q] \end{bmatrix} \tag{50}$$

so that Shanks' method reduces to a LS problem. The LS estimator of the MA parameters is then

$$\hat{\mathbf{b}}_{\mathrm{LS}}=(\mathbf{H}_{\mathrm{MA}}^{*}\mathbf{H}_{\mathrm{MA}})^{-1}\mathbf{H}_{\mathrm{MA}}^{*}\mathbf{y} \tag{51}$$

where $\mathbf{H}_{\mathrm{MA}}$ is computed using the LS estimate $\hat{\mathbf{a}}_{\mathrm{LS}}$ given by (48).

We can modify Shanks' estimator by using the CSLS estimator of the parameters $\mathbf{b}$, which leads to the estimator

$$\hat{\mathbf{b}}_{\mathrm{CSLS}}=\hat{c}\,(\mathbf{R}\mathbf{H}_{\mathrm{MA}}^{*}\mathbf{H}_{\mathrm{MA}})^{-1/2}\mathbf{R}\mathbf{H}_{\mathrm{MA}}^{*}\mathbf{y} \tag{52}$$

where $\hat{c}$ is given by (21), and now, $\mathbf{H}_{\mathrm{MA}}$ is computed using the CSLS estimate $\hat{\mathbf{a}}_{\mathrm{CSLS}}$ of (49).

To evaluate the performance of both estimators, we consider an example in which the ARMA parameters are given by

$$a_1=0.9,\quad a_2=0.6,\quad a_3=0.4,\quad b_0=1,\quad b_2=0.5 \tag{53}$$

and the matrix $\mathbf{R}$ is chosen as $\mathbf{R}=\mathbf{I}_n$.

In Fig. 1, we plot the MSE in estimating the AR parameters using $\hat{\mathbf{a}}_{\mathrm{CSLS}}$ and $\hat{\mathbf{a}}_{\mathrm{LS}}$ for $n=20$ averaged over 2000 noise realizations, as a function of $-10\log\sigma^2$ (to base 10), where σ^2 is the noise variance. As we expect, the MSE of the CSLS estimator decreases with σ^2 for low SNR and then converges to a constant in the high SNR limit. The MSE of the LS estimator decreases with σ^2 at a much slower rate. The experimental threshold is $\approx$ 65 dB so that for values of σ^2 greater than ≈ -65 dB, the CSLS estimator yields a lower MSE than the LS estimator.

We also compared the CSLS estimator with the shrunken estimator and the ridge estimator described in Section IV. Since both of these estimators depend on parameters that have to be

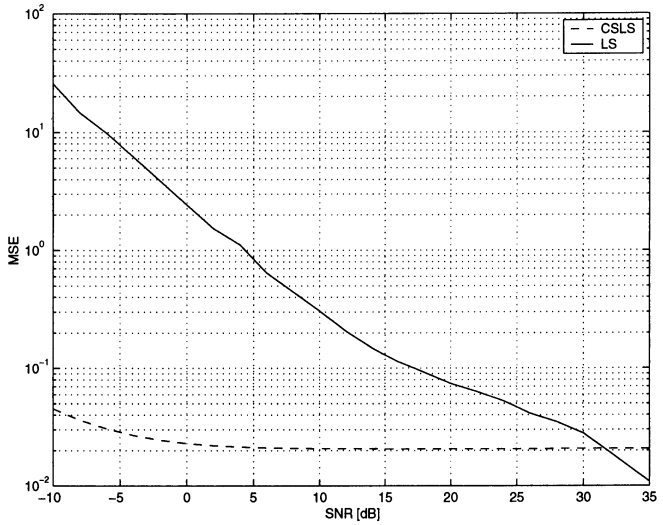

Fig. 2. Mean-squared error in estimating the MA parameters b_k given by (53) based on the estimated values of the AR parameters using the LS estimator (51) and the CSLS estimator (52).

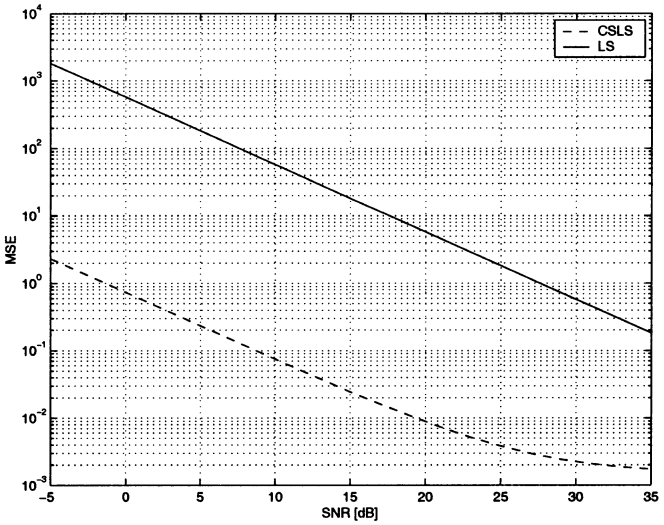

Fig. 3. Mean-squared error in estimating the amplitudes a_1 and a_2 in the model (54) using the LS estimator and the CSLS estimator. The parameter values are given by $s_1 = -0.6 + j2\pi(0.40)$, $s_2 = -0.6 + j2\pi(0.41)$, $n = 15$, and $a_1 = a_2 = 1$.

chosen, the performance of these estimators cannot be showed in a single figure. In our simulations with different choices of parameters, we observed that the CSLS estimator performs significantly better than both the shrunken estimator and the ridge estimator.

In Fig. 2, we plot the MSE in estimating the MA parameters using $\hat{\mathbf{b}}_{\mathrm{CSLS}}$ and $\hat{\mathbf{b}}_{\mathrm{LS}}$ for $n = 20$ averaged over 2000 noise realizations, as a function of $-10\log\sigma^2$. The experimental threshold is $\approx$ 32 dB. In this case, we observed that the CSLS estimator performs better than the shrunken estimator for all SNR. For SNR values up to roughly 25–30 dB (depending on the choice of regularization parameter), the CSLS estimator also performs better than the ridge estimator.

B. Exponential Signal Models

As a second application of the CSLS estimator, we consider the problem of estimating the amplitudes of two complex exponentials with known frequencies and damping factor in complex-valued additive white Gaussian noise. The data is thus given by

$$y[l] = a_1 e^{s_1 l} + a_2 e^{s_2 l} + w[l], \quad l = 0, 1, \ldots, n-1 \tag{54}$$

where $w[l]$ is a white complex Gaussian noise process with variance σ^2, and n is the number of data points.

Denoting by $\mathbf{y}$ the vector of components $y[l]$, we have that $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$, where $\mathbf{x}$ is the vector of components a_1 and a_2, $\mathbf{w}$ is the vector of components $w[l]$, and

$$\mathbf{H} = \begin{bmatrix} 1 & 1 \\ e^{s_1} & e^{s_2} \\ \vdots & \vdots \\ e^{s_2(n-1)} & e^{s_2(n-1)} \end{bmatrix}. \tag{55}$$

In Fig. 3, we plot the MSE in estimating the parameters a_1 and a_2 using the CSLS estimator and the LS estimator for the case in which $s_1 = -0.6 + j2\pi(0.40)$, $s_2 = -0.6 + j2\pi(0.41)$ and $n = 15$. The true parameter values are $a_1 = a_2 = 1$. For the noise variance range shown, the CSLS estimator performs better than the LS estimator. In this example, the experimental threshold variance is $\approx$ 56 dB so that for values of σ^2 greater than ≈ -56 dB, the CSLS estimator yields a lower MSE than the LS estimator.

APPENDIX A
MMSE SHAPING

In this appendix, we consider the problem of finding an optimal shaping transformation that results in a random vector $\mathbf{b} = \mathbf{W}\mathbf{a}$ with covariance $\mathbf{C}_b$ that is as close as possible to $\mathbf{a}$ in mean squared error, where the covariance of $\mathbf{a}$ is given by $\mathbf{C}_a = \mathbf{I}_m$. Specifically, among all possible shaping transformations, we seek the one that minimizes the total MSE given by

$$\varepsilon_{\mathrm{MSE}} = \sum_{k=1}^{m} E\left((a_k - b_k)^2\right) = E\left((\mathbf{a}-\mathbf{b})^*(\mathbf{a}-\mathbf{b})\right) \tag{56}$$

subject to

$$\mathbf{C}_b = \mathbf{W}\mathbf{C}_a\mathbf{W}^* = \mathbf{W}\mathbf{W}^* = c^2\mathbf{C}_b \tag{57}$$

where a_k and b_k are the kth components of $\mathbf{a}$ and $\mathbf{b}$, respectively.

Our approach to determining the shaping transformation that minimizes (56) is to perform a unitary change of coordinates $\mathbf{U}$ so that in the new coordinate system, $\mathbf{a}$ is mapped to $\overline{\mathbf{a}} = \mathbf{U}\mathbf{a}$, and $\mathbf{b}$ is mapped to $\overline{\mathbf{b}} = \mathbf{U}\mathbf{b}$, with the elements of $\overline{\mathbf{b}}$ uncorrelated. Since $\mathbf{U}$ is unitary and $\mathbf{C}_a = c^2\mathbf{I}_m$, the covariance matrix of $\overline{\mathbf{a}}$ is $\mathbf{C}_{\overline{a}} = \mathbf{I}_m$, and the MSE defined by (56) between $\mathbf{a}$ and $\mathbf{b}$ is equal to the MSE between $\overline{\mathbf{a}}$ and $\overline{\mathbf{b}}$.

Such a unitary transformation is provided by the eigendecomposition of $\mathbf{C}_b$. Specifically, suppose that $\mathbf{C}_b$ has an eigendecomposition $\mathbf{C}_b = \mathbf{V}\mathbf{D}\mathbf{V}^*$, where $\mathbf{V}$ is a unitary matrix, and $\mathbf{D}$ is a diagonal matrix with diagonal elements d_k. If we choose $\overline{\mathbf{b}} = \mathbf{V}^*\mathbf{b}$, then the covariance matrix of $\overline{\mathbf{b}}$ is $\mathbf{V}^*\mathbf{C}_b\mathbf{V} = \mathbf{D}$.

Thus, we may first solve the optimal shaping problem in the new coordinate system. Then, with $\widehat{\overline{\mathbf{W}}}$ and $\widehat{\mathbf{W}}$ denoting the optimal shaping transformations in the new and original coordinate systems, respectively, it is straightforward to show that

$$\widehat{\mathbf{W}} = \mathbf{U}^*\widehat{\overline{\mathbf{W}}}\mathbf{U}. \tag{58}$$

To determine $\widehat{\overline{\mathbf{W}}}$, we express $\varepsilon_{\rm MSE}$ of (56) as

$$\begin{aligned}\varepsilon_{\rm MSE} &= \sum_{k=1}^{m} E\left((\overline{a}_k - \overline{b}_k)^2\right) \\ &= \sum_{k=1}^{m} d_k + mc^2 - 2\sum_{k=1}^{m} E(\overline{a}_k\overline{b}_k)\end{aligned} \tag{59}$$

where $\overline{a}_k$ and $\overline{b}_k$ denote the kth components of $\overline{\mathbf{a}}$ and $\overline{\mathbf{b}}$, respectively, and $d_k = E\left(\overline{b}_k^2\right)$. From the Cauchy–Schwarz inequality

$$E(\overline{a}_k\overline{b}_k) \le \left|E(\overline{a}_k\overline{b}_k)\right| \le \left(E\left(\overline{a}_k^2\right)E\left(\overline{b}_k^2\right)\right)^{1/2} \tag{60}$$

with equality if and only if $\overline{b}_k = \gamma_k\overline{a}_k$ with probability one for some non-negative deterministic constant γ_k, in which case, we also have $E(\overline{b}_k^2) = \gamma_k^2 E(\overline{a}_k^2) = \gamma_k^2 = c^2 d_k$ so that $\gamma_k = c\sqrt{d_k}$. Note that $\overline{b}_k$ can always be chosen proportional to $\overline{a}_k$ since the variables $\overline{b}_k$ are uncorrelated. Thus, the optimal value of $\overline{b}_k$ is $\widehat{\overline{b}}_k = c\overline{a}_k\sqrt{d_k}$, and $\widehat{\overline{\mathbf{W}}} = c\mathbf{D}^{1/2}$. The optimal shaping transformation then follows from (58):

$$\widehat{\mathbf{W}} = c\mathbf{V}\mathbf{D}^{1/2}\mathbf{V}^* = c\mathbf{C}_b{}^{1/2}. \tag{61}$$

Appendix B
CSLS Estimator

From (17), the optimal value of $\widehat{\widehat{\mathbf{G}}}$ must satisfy

$$\mathbf{H}\widehat{\widehat{\mathbf{G}}} = \left(\mathbf{H}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\right)^{1/2}. \tag{62}$$

Multiplying both sides by $(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H})^{-1}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}$, we have that

$$\widehat{\widehat{\mathbf{G}}} = \left(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\right)^{-1}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\left(\mathbf{H}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\right)^{1/2}. \tag{63}$$

To simplify the expression for $\widehat{\widehat{\mathbf{G}}}$, we now prove the following matrix equality. Suppose that $\mathbf{S}$ is an $m\times n$ matrix of rank m, $\mathbf{T}$ is an $n\times m$ matrix of rank m and $\mathcal{R}(\mathbf{T}) \subseteq \mathcal{N}(\mathbf{S})^{\perp}$, where $\mathcal{R}(\cdot)$ and $\mathcal{N}(\cdot)$ denote the range space and null space, respectively. Then

$$\mathbf{S}(\mathbf{T}\mathbf{S})^{1/2} = (\mathbf{S}\mathbf{T})^{1/2}\mathbf{S}. \tag{64}$$

To prove (64), we first verify that

$$(\mathbf{T}\mathbf{S})^{1/2} = \mathbf{S}^{\dagger}(\mathbf{S}\mathbf{T})^{1/2}\mathbf{S} \tag{65}$$

where $\mathbf{S}^{\dagger}$ denotes the Moore–Penrose pseudo inverse of $\mathbf{S}$. Indeed

$$\mathbf{S}^{\dagger}(\mathbf{S}\mathbf{T})^{1/2}\mathbf{S}\mathbf{S}^{\dagger}(\mathbf{S}\mathbf{T})^{1/2}\mathbf{S} = \mathbf{S}^{\dagger}\mathbf{S}\mathbf{T}\mathbf{S} = \mathbf{T}\mathbf{S} \tag{66}$$

since from the properties of the pseudo inverse and the fact that $\mathbf{S}$ has full row rank, $\mathbf{S}\mathbf{S}^{\dagger} = \mathbf{I}_m$, and $\mathbf{S}^{\dagger}\mathbf{S}$ is an orthogonal projection onto $\mathcal{N}(\mathbf{S})^{\perp}$, where by our assumption, $\mathcal{R}(\mathbf{T}) \subseteq \mathcal{N}(\mathbf{S})^{\perp}$. Multiplying both sides of (65) on the right by $\mathbf{S}$ establishes (64).

Now, using (64) with $\mathbf{S} = \mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1/2}$ and $\mathbf{T} = \mathbf{C}_w^{-1/2}\mathbf{H}$, we may simplify (63) as

$$\begin{aligned}\widehat{\widehat{\mathbf{G}}} &= \left(\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\right)^{-1/2}\mathbf{R}\mathbf{H}^*\mathbf{C}_w^{-1} \\ &= \mathbf{R}\left(\mathbf{H}^*\mathbf{C}_w^{-1}\mathbf{H}\mathbf{R}\right)^{-1/2}\mathbf{H}^*\mathbf{C}_w^{-1}.\end{aligned} \tag{67}$$

Appendix C
Performance Analysis

To compare the performance of the LS and CSLS estimators, we evaluate the MSE of the estimators, where the MSE of an estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ is given by (5). In the following, we use $\mathbf{B}$ to denote the matrix product $\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H}$.

From (1) and (3), $\hat{\mathbf{x}}_{\rm LS} - \mathbf{x} = \mathbf{B}^{-1}\mathbf{H}^*\mathbf{C}^{-1}\mathbf{w}$, so that

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\rm LS}) = \sigma^2\mathrm{Tr}(\mathbf{B}^{-1}). \tag{68}$$

From Theorem 2, $\hat{\mathbf{x}}_{\rm CSLS} - \mathbf{x} = ((\beta/\sigma)(\mathbf{R}\mathbf{B})^{1/2} - \mathbf{I}_m)\mathbf{x} + (\beta/\sigma)(\mathbf{R}\mathbf{B})^{-1/2}\mathbf{R}\mathbf{H}^*\mathbf{C}^{-1}\mathbf{w}$ so that

$$\begin{aligned}&\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) \\ &= \left\|\left(\left(\frac{\beta}{\sigma}\right)(\mathbf{R}\mathbf{B})^{1/2} - \mathbf{I}_m\right)\mathbf{x}\right\|^2 + \beta^2\mathrm{Tr}\left((\mathbf{R}\mathbf{B})^{-1/2}\mathbf{R}(\mathbf{R}\mathbf{B})^{1/2}\right) \\ &= \left\|\left(\left(\frac{\beta}{\sigma}\right)(\mathbf{R}\mathbf{B})^{1/2} - \mathbf{I}_m\right)\mathbf{x}\right\|^2 + \beta^2\mathrm{Tr}(\mathbf{R}).\end{aligned} \tag{69}$$

A. Fixed Scaling

We first consider the case in which $\beta = \sigma$. Then, from (69)

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) = \|(\mathbf{R}\mathbf{B})^{1/2} - \mathbf{I}_m)\mathbf{x}\|^2 + \sigma^2\mathrm{Tr}(\mathbf{R}) \tag{70}$$

and

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) \le \sigma_\gamma\|\mathbf{x}\|^2 + \sigma^2\mathrm{Tr}(\mathbf{R}) \tag{71}$$

where $\gamma = \arg\max\sigma_k$, and σ_k are the eigenvalues of $\mathbf{Q} = ((\mathbf{R}\mathbf{B})^{1/2} - \mathbf{I}_m)^*((\mathbf{R}\mathbf{B})^{1/2} - \mathbf{I}_m)$. We have equality in (69) only in the event in which $\mathbf{x}$ is in the direction of the eigenvector of $\mathbf{Q}$ corresponding to the eigenvalue σ_γ.

Let $\zeta = \|\mathbf{x}\|^2/(\sigma^2 m)$ denote the SNR per component. Then, combining (68) and (71), we have that $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) \le \mathrm{MSE}(\hat{\mathbf{x}}_{\rm LS})$ if $\sigma_\gamma\zeta + \mathrm{Tr}(\mathbf{R}) \le \mathrm{Tr}(\mathbf{B}^{-1})$. Since $\sigma_\gamma = 0$ only if $\sigma_k = 0$ for all k, which implies that $\mathbf{R} = \mathbf{B}^{-1}$ and $\hat{\mathbf{x}}_{\rm LS} = \hat{\mathbf{x}}_{\rm CSLS}$, it follows that if $\hat{\mathbf{x}}_{\rm LS} \ne \hat{\mathbf{x}}_{\rm CSLS}$, then $\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) \le \mathrm{MSE}(\hat{\mathbf{x}}_{\rm LS})$ if

$$\zeta \le \frac{\mathrm{Tr}(\mathbf{B}^{-1}) - \mathrm{Tr}(\mathbf{R})}{\sigma_\gamma} \triangleq \tilde{\zeta}_{\rm WC}. \tag{72}$$

Note that $\tilde{\zeta}_{\rm WC}$ is a worst-case bound since it corresponds to the worst possible choice of parameters, namely, when the unknown vector $\mathbf{x}$ is in the direction of the eigenvector of $\mathbf{Q}$ corresponding to the eigenvalue σ_γ.

B. Optimal Scaling

Suppose now that $\beta = \hat{c}$ given by (21), and that $\mathbf{R} = \mathbf{I}_m$. Then

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\rm CSLS}) = \|(\alpha\mathbf{B}^{1/2} - \mathbf{I}_m)\mathbf{x}\|^2 + m\alpha^2\sigma^2 \tag{73}$$

where

$$\alpha = \frac{\hat{c}}{\sigma} = \frac{\mathrm{Tr}(\mathbf{B}^{1/2})}{\mathrm{Tr}(\mathbf{B})} = \frac{\sum_{i=1}^{m} \lambda_i^{1/2}}{\sum_{i=1}^{m} \lambda_i} \tag{74}$$

and $\lambda_i, 1 \le i \le m$ denote the eigenvalues of $\mathbf{B}$.

From (32), we have that

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) = \left|\alpha\lambda_\gamma^{1/2} - 1\right|^2 \|\mathbf{x}\|^2 + m\alpha^2\sigma^2 \tag{75}$$

where

$$\gamma = \arg\max \left|\alpha\lambda_k^{1/2} - 1\right|^2 \tag{76}$$

with α given by (74). We have equality in (75) only in the event in which $\mathbf{x}$ is in the direction of the eigenvector of $\mathbf{B}$ corresponding to the eigenvalue λ_γ.

Combining (68) and (75), we have that $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \le \mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$ if

$$\left|\alpha\lambda_\gamma^{1/2} - 1\right|^2 \zeta + \alpha^2 \le \frac{1}{m}\sum_{k=1}^{m} \lambda_k^{-1}. \tag{77}$$

The expression $|\alpha\lambda_\gamma^{1/2} - 1|$ is equal to zero only in the case in which $\lambda_k^{1/2} = 1/\alpha$ for all k so that $\mathbf{B} = (1/\alpha^2)\mathbf{I}_m$. From Theorem 2, it then follows that $|\alpha\lambda_\gamma^{1/2} - 1| = 0$ if and only if $\hat{\mathbf{x}}_{\mathrm{LS}} = \hat{\mathbf{x}}_{\mathrm{CSLS}}$. If $\hat{\mathbf{x}}_{\mathrm{LS}} \ne \hat{\mathbf{x}}_{\mathrm{CSLS}}$, then we have that $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \le \mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$ if

$$\zeta \le \frac{\left(\frac{1}{m}\right)\sum_{k=1}^{m} \lambda_k^{-1} - \alpha^2}{\left|\alpha\lambda_\gamma^{1/2} - 1\right|^2} \triangleq \zeta_{\mathrm{WC}}. \tag{78}$$

Note that ζ_{WC} is a worst case bound since it corresponds to the worst possible choice of parameters, namely, when the unknown vector $\mathbf{x}$ is in the direction of the eigenvector of $\mathbf{B}$ corresponding to the eigenvalue λ_γ.

We now show that when $\hat{\mathbf{x}}_{\mathrm{CSLS}} \ne \hat{\mathbf{x}}_{\mathrm{LS}}$, $\zeta_{\mathrm{WC}} > 0$ so that there is always a range of SNR values for which $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \le \mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$. To this end we need to prove that

$$\frac{1}{m}\sum_{k=1}^{m} \lambda_k^{-1} \ge \alpha^2 \tag{79}$$

or, equivalently

$$\frac{1}{m}\sum_{k=1}^{m} \lambda_k^{-1} \left(\sum_{k=1}^{m} \lambda_k\right)^2 \ge \left(\sum_{k=1}^{m} \lambda_k^{1/2}\right)^2 \tag{80}$$

with equality if and only if $\hat{\mathbf{x}}_{\mathrm{CSLS}} = \hat{\mathbf{x}}_{\mathrm{LS}}$. Using the inequality

$$\left(\sum_{k=1}^{m} f_i g_i\right)^2 \le \sum_{k=1}^{m} f_i^2 \sum g_i^2 \tag{81}$$

we have that

$$\frac{1}{m}\sum_{k=1}^{m} \lambda_k^{-1} \sum_{k=1}^{m} \lambda_k \ge \frac{1}{m}\left(\sum_{k=1}^{m} \lambda_k^{-1/2}\lambda_k^{1/2}\right)^2 = m. \tag{82}$$

Furthermore

$$\sum_{k=1}^{m} \lambda_k = \frac{1}{m}\sum_{k=1}^{m} 1 \sum_{k=1}^{m} \lambda_k \ge \frac{1}{m}\left(\sum_{k=1}^{m} \lambda_k^{1/2}\right)^2. \tag{83}$$

Combining (82) with (83) proves the inequality (80).

We have equality in (81) if and only if $f_i = a g_i$ for all i and some constant a. Thus, we have equality in (82) if and only if $\lambda_k^{-1/2} = a\lambda_k^{1/2}$, which implies that all the eigenvalues λ_k are equal, so that $\mathbf{B}$ is proportional to $\mathbf{I}_m$, and from Theorem 2, $\hat{\mathbf{x}}_{\mathrm{CSLS}} = \hat{\mathbf{x}}_{\mathrm{LS}}$. Under the same condition, we have equality in (83). We therefore conclude that when $\hat{\mathbf{x}}_{\mathrm{CSLS}} \ne \hat{\mathbf{x}}_{\mathrm{LS}}$, there is always a range of SNR for which $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \le \mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$.

In a similar manner, we can show that

$$\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \ge |\alpha\lambda_\kappa^{1/2} - 1|^2 \|\mathbf{x}\|^2 + m\alpha^2\sigma^2 \tag{84}$$

where

$$\kappa = \arg\min |\alpha\lambda_k^{1/2} - 1|^2 \tag{85}$$

with equality in (84) only if $\mathbf{x}$ is in the direction of the eigenvector of $\mathbf{B}$ corresponding to the eigenvalue λ_κ. Thus, $\mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{CSLS}}) \ge \mathrm{MSE}(\hat{\mathbf{x}}_{\mathrm{LS}})$ if

$$\zeta \ge \frac{\left(\frac{1}{m}\right)\sum_{k=1}^{m} \lambda_k^{-1} - \alpha^2}{|\alpha\lambda_\kappa^{1/2} - 1|^2} \triangleq \zeta_{\mathrm{BC}}. \tag{86}$$

ACKNOWLEDGMENT

The authors would like to thank Prof. G. Verghese for referring the authors to various modified LS estimators and for many other helpful discussions.

REFERENCES

[1] T. Kailath , *Lectures on Linear Least-Squares Estimation*, New York: Springer, 1976.
[2] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ: Prentice-Hall, 1993.
[3] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1992.
[4] J. A. Cadzow, "Signal processing via least squares modeling," *IEEE Acoust., Speech, Signal Processing Mag.*, pp. 12–31, Oct. 1990.
[5] G. H. Comb and C. F. Van Loan, "An analysis of the total least-squares problem," *SIAM J. Numer. Anal.*, vol. 17, no. 4, pp. 883–893, 1979.
[6] S. Van Huffel and J. Vandewalle, *The Total Least-Squares Problem: Computational Aspects and Analysis*. Philadelphia, PA: SIAM, 1991, vol. 9, Frontiers in Applied Mathematics.
[7] A. Yeredor, "The extended least-squares criterion: Minimization algorithms and applications," *IEEE Trans. Signal Processing*, vol. 49, pp. 74–86, Jan. 1991.
[8] M. H. J. Gruber, *Regression Estimators: A Comparative Study*. San Diego, CA: Academic, 1990.
[9] C. M. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," in *Proc. 3rd Berkeley Symp. Math. Stat. Prob.*, vol. 1, pp. 197–206.
[10] W. James and C. M. Stein, "Estimation with quadratic loss," in *Proc. 4th Berkeley Symp. Math. Stat. Prob.*, vol. 1, pp. 361–379.
[11] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometr.*, vol. 12, pp. 55–67, Feb. 1970.
[12] A. N. Tikhonov and V. Y. Arsenin, *Solution of III-Posed Problems*. Washington, DC: V. H. Winston, 1977.
[13] L. S. Mayer and T. A. Willke, "On biased estimation in linear models," *Technometr.*, vol. 15, pp. 497–508, Aug. 1973.
[14] C. R. Rao, "Minimum variance and the estimation of several parameters," *Proc. Cambridge Phil. Soc.*, pp. 280–283, 1946.

[15] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*. New York: Wiley, 1968.
[16] Y. C. Eldar and A. V. Oppenheim, "MMSE whitening and subspace whitening," *IEEE Trans. Inform. Theory*, to be published.
[17] ——, "Orthogonal matched filter detection," in *Proc. Int. Conf Acoust., Speech, Signal Process.*, Salt Lake, UT, May 2001.
[18] Y. C. Eldar, "Quantum signal processing," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, 2001.
[19] Y. C. Eldar and A. V. Oppenheim, "Orthogonal multiuser detection," *Signal Process.*, vol. 82, pp. 321–325, 2002.
[20] Y. C. Eldar and A. M. Chan, "An optimal whitening approach to linear multiuser detection," *IEEE Trans. Inform. Theory*, Jan. 2002, submitted for publication.
[21] R. Lupas and S. Verdu, "Linear multiuser detectors for synchronous code-division multiple-access channels," *IEEE Trans. Inform. Theory*, vol. 35, pp. 123–136, Jan. 1989.
[22] Y. C. Eldar, A. V. Oppenheim, and D. Egnor, "Orthogonal and projected orthogonal matched filter detection," *Signal Process.*, June 2002, submitted for publication.
[23] J. L. Shanks, "Recursion filters for digital processing," *Geophys.*, vol. 32, no. 1, pp. 33–51, Feb. 1967.

Yonina C. Eldar (S'98–M'03) received the B.Sc. degree in physics in 1995 and the B.Sc. degree in electrical engineering in 1996, both from Tel-Aviv University (TAU), Tel-Aviv, Israel, and the Ph.D. degree in electrical engineering and computer science in 2001 from the Massachusetts Institute of Technology (MIT), Cambridge.

From January 2002 to July 2002, she was a Postdoctoral fellow at the Digital Signal Processing Group, MIT. She is currently a Senior Lecturer in the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa, Israel. From 1992 to 1996, she was in the program for outstanding students in TAU. She has served as a teaching assistant for classes in linear systems, digital signal processing, and statistical signal processing. Her current research interests are in the general areas of signal processing and quantum detection.

Dr. Eldar held the Rosenblith Fellowship for study in Electrical Engineering at MIT in 1998, and in 2000, she held an IBM Research Fellowship. She is currently a Horev Fellow of the Leaders in Science and Technology Program at the Technion.

Alan V. Oppenheim (F'77) received the S.B. and S.M. degrees in 1961 and the Sc.D. degree in 1964, all in electrical engineering, from the Massachusetts Institute of Technology, Cambridge. He also received an honorary doctorate degree from Tel-Aviv University, Tel-Aviv, Israel, in 1995.

In 1964, he joined the faculty at MIT, where he currently is Ford Professor of Engineering and a Mac Vicar Faculty Fellow. Since 1967, he has also been affiliated with MIT Lincoln Laboratory and, since 1977, with the Woods Hole Oceanographic Institution, Woods Hole, MA. His research interests are in the general area of signal processing and its applications. He is coauthor of the widely used textbooks *Discrete-Time Signal Processing* and *Signals and Systems*. He is also editor of several advanced books on signal processing.

Dr. Oppenheim is a member of the National Academy of Engineering and a member of Sigma Xi and Eta Kappa Nu. He has been a Guggenheim Fellow and a Sackler Fellow at Tel Aviv University. He has also received a number of awards for outstanding research and teaching including the IEEE Education Medal, the IEEE Centennial Award, the Society Award, the Technical Achievement Award, and the Senior Award of the IEEE Acoustics, Speech, and Signal Processing Society. He has also received a number of awards at MIT for excellence in teaching, including the Bose Award and the Everett Moore Baker Award.