MMSE Whitening and Subspace Whitening

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

Abstract—**This correspondence develops a linear whitening transformation that minimizes the mean-squared error (MSE) between the original and whitened data, i.e.,one that results in a white output that is as close as possible to the input, in an MSE sense. When the covariance matrix of the data is not invertible, the whitening transformation is designed to optimally whiten the data on a subspace in which it is contained. The optimal whitening transformation is developed both for the case of finite-length data vectors and infinite-length signals.**

Index Terms—**Mean-squared error (MSE) whitening, subspace whitening, whitening.**

I. INTRODUCTION

Data whitening arises in a variety of contexts in which it is useful to either decorrelate a data sequence prior to subsequent processing, or to control the spectral shape after processing. Examples in which data whitening has been used to advantage include enhancing direction of arrival algorithms by prewhitening [1], [2], and improving probability of correct detection in multisignature systems [3], [4] and multiuser wireless communication systems [5] by prewhitening.

Whitening of a random sequence parallels closely the concept of orthogonalization of a set of vectors. Specifically, orthogonalizing a set of vectors involves mapping the set of vectors to a new set of vectors through a linear transformation so that the inner product between any two vectors in the set is zero. Similarly, whitening a zero-mean random sequence involves mapping the sequence to a new sequence through a linear transformation so that the expectation of the product of any two elements in the sequence is zero. Since the expectation of the product of two random variables is an inner product, the mathematics associated with whitening of a random sequence parallels the mathematics associated with orthogonalizing a set of vectors.

Any whitening transformation cascaded with a linear unitary transformation will result in a different whitening transformation, so that the linear transformation that whitens a data vector or infinite length signal is not unique. While in some applications of whitening certain conditions might be imposed on the whitening transformation such as causality or symmetry, there have been no general assertions of optimality for various choices of a linear whitening transformation.

Recently, the concept of least-squares (LS) orthogonalization has been introduced [6]–[8] in which an orthogonal set of vectors is constructed from a given set of vectors in such a way that the orthogonal vectors are as close as possible in a LS sense to the given set of vectors. LS orthogonalization was originally motivated by a detection problem

in quantum mechanics [6], and later applied to the design of optimal frames [9], [10].

Paralleling the concept of LS orthogonalization, in this paper we develop an optimal linear whitening transformation. Our criterion for optimality is motivated by the fact that, in general, whitening a data vector or signal introduces distortion to the values of the data relative to the unwhitened data. In certain applications of whitening, it may be desirable to whiten the data while minimizing this distortion. Therefore, in this correspondence we propose choosing a linear whitening transformation that minimizes the mean-squared error (MSE) between the original and whitened data, i.e., that results in a white output that is as close as possible to the input in an MSE sense. We refer to such a whitening transformation as a minimum MSE (MMSE) whitening transformation. Extensions of this concept to other forms of covariance shaping are considered in [4], [11].

Applications of MMSE whitening and subspace whitening to matched-filter detection, multiuser detection, and LS estimation are considered in [3], [5], [12]–[14]. The essential idea in the detection applications is to improve the detection performance by optimally whitening the output of conventional receivers prior to detection using an MMSE or subspace MMSE whitening transformation. As we show by simulations in [3] and analytically in [5], in many cases this approach can, in fact, lead to improved detection performance.

To illustrate the use of MMSE whitening and subspace whitening in more detail, we consider here an application of these ideas to LS estimation. This application is developed and explored in more detail in [12], [13]. Specifically, we consider the problem of estimating a set of unknown deterministic parameters x observed through a known linear transformation H and corrupted by additive noise. The traditional LS estimator chooses as its estimate the parameters that minimize the LS error between the observed data and the estimated noise free data that would be obtained with the estimated parameters. Depending on H, large errors in x might result in small errors in the output in which case the resulting LS estimate can be a poor estimate of the parameters x. In the infinite-dimensional case, this problem corresponds to linear LS deconvolution of noisy data which is well known to be highly sensitive to additive noise, when the signal-to-noise ratio (SNR) is low or moderate or the dynamic range of the spectrum of the linear transformation is high. Exploiting the results developed in this paper for linear MMSE whitening, in Section V we apply MMSE whitening to the LS estimator so that we control the spectral shape and the dynamic range of the estimation error. This leads to a new estimator which we refer to here as the whitened LS (WTLS) estimator. This estimator is a special case of the more general covariance shaping LS estimator, developed in [12]. As we show, regardless of the value of x there is always a threshold SNR, below which the WTLS estimator yields a lower MSE than the LS estimator.

In Section II, we derive the linear MMSE whitening transformation for a finite-length data vector with positive definite covariance matrix. In Section III, we consider optimal whitening for the case in which the covariance matrix is not positive definite, i.e., is not invertible. In this case, whitening and optimal whitening are restricted to the subspace in which the random vector is contained with probability 1. In Section IV, we consider optimal whitening of infinite-length stationary data, i.e., stationary random processes, both in the case of positive definite and positive semidefinite covariance functions. Section V considers the application of MMSE whitening and subspace whitening to LS estimation.

Throughout the correspondence, we denote vectors in $\mathcal{R}^m$ (m arbitrary) by boldface lower case letters, and matrices in $\mathcal{R}^{m \times m}$ by boldface upper case letters. $\boldsymbol{P}_{\mathcal{V}}$ denotes the orthogonal projection operator

onto the subspace $\mathcal{V}$ and $\boldsymbol{I}_m$ denotes the $m \times m$ identity matrix. The adjoint of a transformation is denoted by $(\cdot)^*$, the *Moore–Penrose pseudoinverse* [15] is denoted by $(\cdot)^\dagger$, and $\hat{(\cdot)}$ denotes an optimal vector or transformation. The cross covariance of random variables a and b is denoted by $\mathrm{cov}(a, b)$, and $E(\cdot)$ denotes expectation.

II. OPTIMAL WHITENING TRANSFORMATION

Let $\boldsymbol{a} \in \mathcal{R}^m$ denote a zero-mean[1] random vector with positive definite covariance matrix $\boldsymbol{C}_a$. We wish to whiten[2] the vector $\boldsymbol{a}$ using a whitening transformation $\boldsymbol{W}$ to obtain the random vector $\boldsymbol{b} = \boldsymbol{W}\boldsymbol{a}$, where the covariance matrix of $\boldsymbol{b}$ is given by $\boldsymbol{C}_b = c^2 \boldsymbol{I}_m$ for some $c > 0$. Thus, we seek a transformation $\boldsymbol{W}$ such that

$$\boldsymbol{C}_b = \boldsymbol{W}\boldsymbol{C}_a\boldsymbol{W}^* = c^2 \boldsymbol{I}_m \tag{1}$$

for some $c > 0$. We refer to any $\boldsymbol{W}$ satisfying (1) as a whitening transformation.

Given a covariance matrix $\boldsymbol{C}_a$, there are many ways to choose a whitening transformation $\boldsymbol{W}$ satisfying (1), for example, using the eigen decomposition or Cholesky factorization of $\boldsymbol{C}_a$ [16]. Although there are an unlimited number of whitening transformations satisfying (1), no general assertion of optimality is known for the output $\boldsymbol{b} = \boldsymbol{W}\boldsymbol{a}$ of these different transformations. In particular, the white random vector $\boldsymbol{b} = \boldsymbol{W}\boldsymbol{a}$ may not be "close" to the input vector $\boldsymbol{a}$. If the vector $\boldsymbol{b}$ undergoes some noninvertible processing, or is used as an estimator of some unknown parameters represented by the data $\boldsymbol{a}$, then we may wish to choose the whitening transformation such that $\boldsymbol{b}$ is close to $\boldsymbol{a}$ in some sense. This can be particularly important in applications in which $\boldsymbol{b}$ is the input to a detector, so that we may wish to whiten $\boldsymbol{a}$ prior to detection, but at the same time minimize the distortion to $\boldsymbol{a}$ by choosing $\boldsymbol{W}$ so that $\boldsymbol{b}$ is close to $\boldsymbol{a}$. Applications of this type have been recently investigated in various contexts including matched-filter detection [3], [4] and multiuser detection [5]. We, therefore, propose a whitening transformation that is optimal in the sense that it results in a random vector $\boldsymbol{b}$ that is as close as possible to $\boldsymbol{a}$ in MSE. Specifically, among all possible whitening transformations we seek the one that minimizes the total MSE given by

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

subject to (1), where a_k and b_k are the kth components of $\boldsymbol{a}$ and $\boldsymbol{b}$, respectively. We may wish to constraint the constant c in (1), or may choose c such that the total MSE is minimized.

The MMSE whitening transformation is given by the following theorem.

Theorem 1 (MMSE Whitening Transformation): Let $\boldsymbol{a} \in \mathcal{R}^m$ be a random vector with positive definite covariance matrix $\boldsymbol{C}_a = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^*$, where $\boldsymbol{D}$ is a diagonal matrix with diagonal elements $d_k > 0$ and $\boldsymbol{V}$ is a unitary matrix. Let $\hat{\boldsymbol{W}}$ be the optimal whitening transformation that minimizes the MSE defined by (2), between the input $\boldsymbol{a}$ and the output $\boldsymbol{b} = \boldsymbol{W}\boldsymbol{a}$ with covariance $\boldsymbol{C}_b = c^2 \boldsymbol{I}_m$ where $c > 0$. Then

$$\hat{\boldsymbol{W}} = \alpha \boldsymbol{V}\boldsymbol{D}^{-1/2}\boldsymbol{V}^* = \alpha \boldsymbol{C}_a^{-1/2}$$

where

1) if c is specified then $\alpha = c$;

---

2) if c is chosen to minimize the MSE then

$$\alpha = \frac{1}{m} \sum_{k=1}^{m} \sqrt{d_k}.$$

Proof: Let $\boldsymbol{C}_a$ have an eigen decomposition $\boldsymbol{C}_a = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^*$, where $\boldsymbol{V}$ is a unitary matrix and $\boldsymbol{D}$ is a diagonal matrix with diagonal elements $d_k > 0$. Define $\overline{\boldsymbol{a}} = \boldsymbol{V}^*\boldsymbol{a}$ and $\overline{\boldsymbol{b}} = \boldsymbol{V}^*\boldsymbol{b}$. Then we have immediately that the covariance of $\overline{\boldsymbol{a}}$ is $\boldsymbol{V}^*\boldsymbol{C}_a\boldsymbol{V} = \boldsymbol{D}$ so that the elements of $\overline{\boldsymbol{a}}$ are uncorrelated: $\mathrm{cov}(\overline{a}_k, \overline{a}_l) = d_k \delta_{kl}$, where $\overline{a}_k$ denotes the kth component of $\overline{\boldsymbol{a}}$. Furthermore, since $\boldsymbol{V}^*$ is unitary and $\boldsymbol{C}_b = c^2 \boldsymbol{I}_m$, the covariance of $\overline{\boldsymbol{b}}$ is $c^2 \boldsymbol{I}_m$, and the MSE defined by (2) between $\boldsymbol{a}$ and $\boldsymbol{b}$ is equal to the MSE between $\overline{\boldsymbol{a}}$ and $\overline{\boldsymbol{b}}$. With $\overline{\hat{\boldsymbol{W}}}$ and $\hat{\boldsymbol{W}}$ denoting the optimal whitening transformations in the new and original coordinate systems, respectively, it is straightforward to show that

$$\hat{\boldsymbol{W}} = \boldsymbol{V}\overline{\hat{\boldsymbol{W}}}\boldsymbol{V}^*. \tag{3}$$

To determine $\overline{\hat{\boldsymbol{W}}}$, we express $\varepsilon_{\mathrm{MSE}}$ of (2) as

$$\varepsilon_{\mathrm{MSE}} = \sum_{k=1}^{m} E\left((\overline{a}_k - \overline{b}_k)^2\right) = \sum_{k=1}^{m} d_k + mc^2 - 2\sum_{k=1}^{m} E(\overline{a}_k\overline{b}_k) \tag{4}$$

where $d_k = E(\overline{a}_k^2)$ are the eigenvalues of $\boldsymbol{C}_a$, and $\overline{b}_k$ denotes the kth component of $\overline{\boldsymbol{b}}$. From the Cauchy–Schwarz inequality

$$E(\overline{a}_k\overline{b}_k) \le |E(\overline{a}_k\overline{b}_k)| \le \left(E\left(\overline{a}_k^2\right)E\left(\overline{b}_k^2\right)\right)^{1/2} = c^2 E\left(\overline{a}_k^2\right) \tag{5}$$

with equality if and only if $\overline{b}_k = (c/\sqrt{d_k})\overline{a}_k$ with probability one (w.p. 1). Note, that $\overline{b}_k$ can always be chosen proportional to $\overline{a}_k$ since the variables $\overline{a}_k$ are uncorrelated.

We, therefore, conclude that if the constant c in (4) is specified, then $\overline{\hat{\boldsymbol{W}}} = c\boldsymbol{D}^{-1/2}$, and from (3)

$$\hat{\boldsymbol{W}} = c\boldsymbol{V}\boldsymbol{D}^{-1/2}\boldsymbol{V}^* = c\boldsymbol{C}_a^{-1/2}. \tag{6}$$

Alternatively, we may choose to further minimize (4) with respect to c. Substituting $\overline{b}_k = (c/\sqrt{d_k})\overline{a}_k$ back into (4), we choose c to minimize

$$mc^2 - 2c\sum_{k=1}^{m} \sqrt{d_k}. \tag{7}$$

The optimal value of c, denoted by $\hat{c}$, is therefore given by

$$\hat{c} = \frac{1}{m} \sum_{k=1}^{m} \sqrt{d_k} \tag{8}$$

and the optimal whitening transformation is

$$\hat{\boldsymbol{W}} = \hat{c}\boldsymbol{V}\boldsymbol{D}^{-1/2}\boldsymbol{V}^* = \hat{c}\boldsymbol{C}_a^{-1/2}. \tag{9}$$

$\square$

It is interesting to note that the MMSE whitening transformation has the additional property that it is the unique *symmetric* whitening transformation [17] (up to a possible minus sign). It is also proportional to the Mahalanobis transformation, that is frequently used in signal processing applications incorporating whitening (see, e.g., [18], [1], [2]).

III. OPTIMAL SUBSPACE WHITENING

A. Subspace Whitening

Suppose now that $\boldsymbol{a}$ is a zero-mean random vector in $\mathcal{R}^m$ with noninvertible covariance matrix $\boldsymbol{C}_a$, where $\mathrm{rank}(\boldsymbol{C}_a) = n < m$, and let $\mathcal{V} \subset \mathcal{R}^m$ denote the range space of $\boldsymbol{C}_a$. It then follows that for any

$v \in \mathcal{V}^{\perp}$, $C_a v = 0$, so that $v^* a = 0$ w.p. 1 for any realization of a, and, consequently, any realization of the random vector a lies in $\mathcal{V}$ w.p. 1.[3] In this case, there is no whitening transformation W such that $W C_a W^* = c^2 I_m$. Instead, we propose whitening a on the space $\mathcal{V}$ in which it is contained, which we refer to as *subspace whitening*.

Let b denote the output of a subspace whitening transformation of a. Since $a \in \mathcal{V}$, we require that $b \in \mathcal{V}$. In addition, we require that the representation of b in terms of some orthonormal basis for $\mathcal{V}$ is white, which implies that the representation in terms of any orthonormal basis for $\mathcal{V}$ is white. We now translate these conditions to conditions on the covariance C_b of b.

Denote by $C_a = V D V^*$ the eigen decomposiiton of C_a, where V is a unitary matrix with orthonormal columns v_k where the first n columns $\{v_k, 1 \le k \le n\}$ span the range space of C_a, and D is a diagonal matrix with diagonal elements d_k where $d_k > 0$, $1 \le k \le n$ and $d_k = 0$, $n + 1 \le k \le m$. It then follows that b lies in $\mathcal{V}$ if and only if the null space of C_b contains $\mathcal{V}^{\perp}$, i.e., if and only if

$$C_b v_k = 0, \qquad n + 1 \le k \le m. \tag{10}$$

Next, let V_1 denote the matrix of columns $\{v_k, 1 \le k \le n\}$ that form a basis for $\mathcal{V}$. Then the representation of b in this basis for $\mathcal{V}$ is $b_v = V_1^* b$, $b_v \in \mathcal{R}^n$, and the covariance of b_v is $V_1^* C_b V_1$. Therefore, b_v is white on $\mathcal{V}$ if and only if C_b satisfies

$$V_1^* C_b V_1 = c^2 I_n \tag{11}$$

for some $c > 0$. Combining (10) and (11) we conclude that b is white on $\mathcal{V}$ if and only if

$$C_b = c^2 P_{\mathcal{V}} = c^2 V \tilde{I} V^* \tag{12}$$

where $P_{\mathcal{V}}$ is the orthogonal projection operator onto $\mathcal{V}$ and

$$\tilde{I} = \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix}. \tag{13}$$

B. MMSE Subspace Whitening Transformation

We now seek a subspace whitening transformation W_s such that the vector $b = W_s a$ is white on the range space $\mathcal{V}$ of C_a, and is as close as possible to a in the MSE sense. Thus, we seek the transformation that minimizes (2) subject to

$$C_b = W_s C_a W_s^* = c^2 V \tilde{I} V^* \tag{14}$$

where $\tilde{I}$ is given by (13).

The MMSE subspace whitening transformation, denoted by $\hat{W}_s$, is derived in the Appendix in an analogous manner to the derivation of the MMSE whitening transformation of Section II, and is summarized in the following theorem.

Theorem 2 (MMSE Subspace Whitening): Let $a \in \mathcal{R}^m$ be a random vector with covariance matrix $C_a = V D V^*$ with $\text{rank}(C_a) = n < m$, where D is a diagonal matrix with diagonal elements $d_k > 0$, $1 \le k \le n$ and $d_k = 0$, $n + 1 \le k \le m$, and V is a unitary matrix. Let $\mathcal{V}$ denote the range space of C_a spanned by the first n columns of V. Let $\hat{W}_s$ be any subspace whitening transformation that minimizes the MSE defined by (2), between the input a and the output b with covariance $C_b = c^2 P_{\mathcal{V}} = c^2 V \tilde{I} V^*$, where $\tilde{I}$ is given by (13) and $c > 0$. Then

1) $\hat{W}_s$ is not unique;

2) $\hat{W}_s = \alpha_s V (D^{1/2})^{\dagger} V^* = \alpha_s (C_a^{1/2})^{\dagger}$ is an optimal subspace whitening transformation where
   a) if c is specified then $\alpha_s = c$,
   b) if c is chosen to minimize the MSE then

$$\alpha_s = \frac{1}{n} \sum_{k=1}^{n} \sqrt{d_k};$$

3) define $W_s^{\mathcal{V}} = \hat{W}_s P_{\mathcal{V}}$ where $P_{\mathcal{V}}$ is an orthogonal projection onto $\mathcal{V}$ and $\hat{W}_s$ is any optimal subspace whitening transformation; then
   a) $W_s^{\mathcal{V}}$ is unique, and is given by

$$W_s^{\mathcal{V}} = \alpha_s V (D^{1/2})^{\dagger} V^* = \alpha_s (C_a^{1/2})^{\dagger},$$

   b) $\hat{W}_s a = W_s^{\mathcal{V}} a$ w.p. 1,
   c) $b = \hat{W}_s a$ is unique w.p. 1.

Part 3 of Theorem 2 states the intuitively reasonable fact, which also follows from the proof in the Appendix, that $\hat{W}_s$ is uniquely specified on $\mathcal{V}$ but can be arbitrary on $\mathcal{V}^{\perp}$. However, since the input a to the whitening transformation lies in $\mathcal{V}$ w.p. 1, the choice of $\hat{W}_s$ on $\mathcal{V}^{\perp}$ does not affect the output b (w.p. 1).

IV. OPTIMAL WHITENING OF STATIONARY RANDOM PROCESSES

We now consider optimal whitening and subspace whitening of a stationary random process.

A. MMSE Whitening

Let $a[n]$ be a zero-mean stationary random process with positive spectrum $S_a(\omega)$. Suppose we wish to whiten $a[n]$, i.e., find a linear time-invariant (LTI) whitening filter with frequency response $W(\omega)$, such that the spectrum $S_b(\omega)$ of the filter output $b[n] = a[n] * w[n]$ is $S_b(\omega) = c^2$ for some $c > 0$. Since [18]

$$S_b(\omega) = S_a(\omega) |W(\omega)|^2 \tag{15}$$

$W(\omega)$ is a whitening filter if and only if

$$|W(\omega)|^2 = c^2 S_a^{-1}(\omega). \tag{16}$$

From all possible whitening filters satisfying (16) we seek the filter that minimizes

$$\varepsilon_{\text{MSE}} = E\left((a[n] - b[n])^2 \right) \tag{17}$$

so that $b[n]$ is as close as possible to $a[n]$ in an MSE sense.

The MMSE whitening filter is given by the following theorem.

Theorem 3 (MMSE Whitening Filter): Let $a[n]$ denote a random process with positive spectrum $S_a(\omega)$. Let $\hat{W}(\omega)$ be the frequency response of the optimal whitening filter with impulse response $\hat{w}[n]$ that minimizes the MSE defined by (17), between the input $a[n]$ and the output $b[n] = a[n] * w[n]$ with spectrum $S_b(\omega) = c^2$ for $c > 0$. Then

$$\hat{W}(\omega) = \sigma S_a(\omega)^{-1/2}$$

where

1) if c is specified then $\sigma = c$;
2) if c is chosen to minimize the MSE then

$$\sigma = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_a^{1/2}(\omega) \, d\omega.$$

Note that the MMSE whitening filter is the unique zero-phase filter that satisfies (16).

Proof: Expanding (17) we have

$$\varepsilon_{\mathrm{MSE}} = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_a(\omega)\, d\omega + c^2 - \frac{1}{\pi} \int_{-\pi}^{\pi} S_{ab}(\omega)\, d\omega \qquad (18)$$

where $S_{ab}(\omega)$ is the cross spectrum between $a[n]$ and $b[n]$ and is given by [18]

$$S_{ab}(\omega) = W^*(\omega) S_a(\omega). \qquad (19)$$

It, therefore, follows that minimizing (18) with respect to $b[n]$ is equivalent to maximizing

$$\varepsilon = \int_{-\pi}^{\pi} S_{ab}(\omega)\, d\omega = \int_{-\pi}^{\pi} c^2 W^{-1}(\omega)\, d\omega \qquad (20)$$

where we used (16). Now

$$\varepsilon = \int_{-\pi}^{\pi} c^2 W^{-1}(\omega)\, d\omega \le \int_{-\pi}^{\pi} c^2 |W^{-1}(\omega)|\, d\omega$$

$$= \int_{-\pi}^{\pi} c S_a^{1/2}(\omega)\, d\omega \qquad (21)$$

with equality if and only if

$$W^{-1}(\omega) = |W^{-1}(\omega)| = \frac{1}{c} S_a^{1/2}(\omega) \qquad (22)$$

or, equivalently

$$W(\omega) = c S_a^{-1/2}(\omega). \qquad (23)$$

We conclude that if c is specified, then the MMSE whitening filter is given by

$$\hat{W}(\omega) = c S_a^{-1/2}(\omega). \qquad (24)$$

We may further wish to minimize the MSE with respect to c. Substituting $\hat{b}[n] = \hat{w}[n] * a[n]$ into (18) where $\hat{w}[n]$ is the inverse Fourier transform of $\hat{W}(\omega)$ and minimizing with respect to c, the optimal value of c is

$$\hat{c} = E\left(a[n](a[n] * \hat{w}[n]) \right) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_a^{1/2}(\omega)\, d\omega \qquad (25)$$

and the MMSE whitening filter is $\hat{W}(\omega) = \hat{c} S_a^{-1/2}(\omega)$. $\qquad \square$

The MMSE whitening filter given by Theorem 3 is reminiscent of the MMSE whitening transformation given by Theorem 1. The optimal whitening transformation is proportional to the inverse square root of the input covariance matrix, and is symmetric. Similarly, the Fourier transform of the optimal whitening filter is proportional to the inverse square root of the input spectral density function, and has zero phase.

B. MMSE Subspace Whitening

When the correlation function of $a[n]$ is not positive definite so that $S_a(\omega) = 0$ for some ω, in analogy to the finite-dimensional case, we propose whitening $a[n]$ on the subspace to which it is confined, which is equivalent to whitening $a[n]$ over the frequency intervals for which $S_a(\omega) \ne 0$. Thus, the subspace whitening filter satisfies

$$|W(\omega)|^2 = \begin{cases} c^2 S_a^{-1}(\omega), & \omega \text{ such that } S_a(\omega) \ne 0 \\ \text{arbitrary}, & \omega \text{ such that } S_a(\omega) = 0. \end{cases} \qquad (26)$$

The frequency response of the MMSE subspace whitening filter is given by Theorem 3 at frequencies for which $S_a(\omega) \ne 0$, and is arbitrary otherwise.

V. APPLICATION TO LS ESTIMATION

We now consider an application of MMSE whitening to the problem of estimating the unknown deterministic parameters $\boldsymbol{x}$ in the linear model

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \qquad (27)$$

where $\boldsymbol{H}$ is a known $n \times m$ matrix and $\boldsymbol{n}$ is a zero-mean random vector with covariance $\boldsymbol{C}_n$. For simplicity of exposition, we assume that $\boldsymbol{C}_n = \sigma^2 \boldsymbol{I}_n$. This application is developed in more detail in [12], [13].

Many signal processing estimation problems can be represented by the linear model (27), and consequently, this problem has been studied extensively in the literature (see, e.g., [19], [18]). A common approach to estimating the parameters $\boldsymbol{x}$ is to restrict the estimator to be linear and unbiased, and then seek the estimator of this form that minimizes the variance [19]. The optimal estimator is the well-known LS estimator

$$\hat{\boldsymbol{x}}_{\mathrm{LS}} = (\boldsymbol{H}^*\boldsymbol{H})^\dagger \boldsymbol{H}^*\boldsymbol{y} \qquad (28)$$

which also minimizes the LS error defined by

$$\varepsilon_{\mathrm{LS}} = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 = (\boldsymbol{y} - \hat{\boldsymbol{y}})^*(\boldsymbol{y} - \hat{\boldsymbol{y}}) \qquad (29)$$

where $\hat{\boldsymbol{y}} = \boldsymbol{H}\hat{\boldsymbol{x}}$.

The LS estimator seeks the estimate of $\boldsymbol{x}$ that results in an estimated data vector $\hat{\boldsymbol{y}}$ that is as close as possible to the original data vector $\boldsymbol{y}$. However, in an estimation context, typically we are more interested in minimizing the error between $\boldsymbol{x}$ and the estimate of $\boldsymbol{x}$. In many cases, the data vector $\boldsymbol{y}$ is not very sensitive to changes in $\boldsymbol{x}$, so that a large error in estimating $\boldsymbol{x}$ may translate into a small error in estimating the data vector $\boldsymbol{y}$, in which case $\hat{\boldsymbol{x}}_{\mathrm{LS}}$ may be a poor estimate of $\boldsymbol{x}$. This effect is especially predominant at low to moderate SNR, where the data vector $\boldsymbol{y}$ is typically affected more by the noise than by changes in $\boldsymbol{x}$; the exact SNR range will depend on the properties of the model matrix $\boldsymbol{H}$. To improve the performance of the LS estimator at low to moderate SNR, we propose a modification of the LS estimate based on the concept of MMSE whitening, which we refer to as the WTLS estimator.

Since $\hat{\boldsymbol{x}}_{\mathrm{LS}} = \boldsymbol{x} + \tilde{\boldsymbol{n}}$ where $\tilde{\boldsymbol{n}} = (\boldsymbol{H}^*\boldsymbol{H})^{-1}\boldsymbol{H}^*\boldsymbol{n}$, the covariance of the noise component $\tilde{\boldsymbol{n}}$ in $\hat{\boldsymbol{x}}_{\mathrm{LS}}$ is equal to the covariance of $\hat{\boldsymbol{x}}_{\mathrm{LS}}$, which is given by $\sigma^2(\boldsymbol{H}^*\boldsymbol{H})^{-1}$. Evidently, the estimation error resulting from the LS estimator can have a large variance and a covariance structure with a large dynamic range. These properties of the estimation error tend to limit the performance of the LS estimator. Therefore, to improve the performance of the LS estimator, we propose whitening the noise component in the estimator $\hat{\boldsymbol{x}}_{\mathrm{LS}}$ on the space in which it is contained, so that we control the dynamic range and spectral shape of the covariance of the estimation error. From Theorem 1, it follows that the optimal whitening transformation is proportional to $\hat{\boldsymbol{W}} = (\boldsymbol{H}^*\boldsymbol{H})^{1/2}$ so that the WTLS estimator, denoted by $\hat{\boldsymbol{x}}_{\mathrm{WTLS}}$, has the form

$$\hat{\boldsymbol{x}}_{\mathrm{WTLS}} = \beta(\boldsymbol{H}^*\boldsymbol{H})^{1/2}\hat{\boldsymbol{x}}_{\mathrm{LS}} = \beta(\boldsymbol{H}^*\boldsymbol{H})^{\dagger/2}\boldsymbol{H}^*\boldsymbol{y} \qquad (30)$$

for some constant β, where we introduce the notation $(\cdot)^{\dagger/2} = ((\cdot)^{1/2})^\dagger$.

The scaling β is chosen to minimize the total variance of the data error $\boldsymbol{y} - \hat{\boldsymbol{y}}$, where the data estimate $\hat{\boldsymbol{y}}$ is given by

$$\hat{\boldsymbol{y}} = \boldsymbol{H}\hat{\boldsymbol{x}}_{\mathrm{WTLS}} = \beta\boldsymbol{H}(\boldsymbol{H}^*\boldsymbol{H})^{\dagger/2}\boldsymbol{H}^*\boldsymbol{y}. \qquad (31)$$

Using (31) we may express this covariance as

$$E\left((\boldsymbol{y}' - \hat{\boldsymbol{y}}')^*(\boldsymbol{y}' - \hat{\boldsymbol{y}}') \right)$$

$$= E\left(\boldsymbol{y}'^*(\boldsymbol{I}_n - \beta\boldsymbol{H}(\boldsymbol{H}^*\boldsymbol{H})^{\dagger/2}\boldsymbol{H}^*)^2 \boldsymbol{y}' \right)$$

$$= \sigma^2 \mathrm{Tr}\left((\boldsymbol{I}_n - \beta\boldsymbol{H}(\boldsymbol{H}^*\boldsymbol{H})^{\dagger/2}\boldsymbol{H}^*)^2 \right)$$

$$= \sigma^2 \left(n + \beta^2 \mathrm{Tr}(\boldsymbol{H}^*\boldsymbol{H}) - 2\beta \mathrm{Tr}((\boldsymbol{H}^*\boldsymbol{H})^{1/2}) \right) \qquad (32)$$
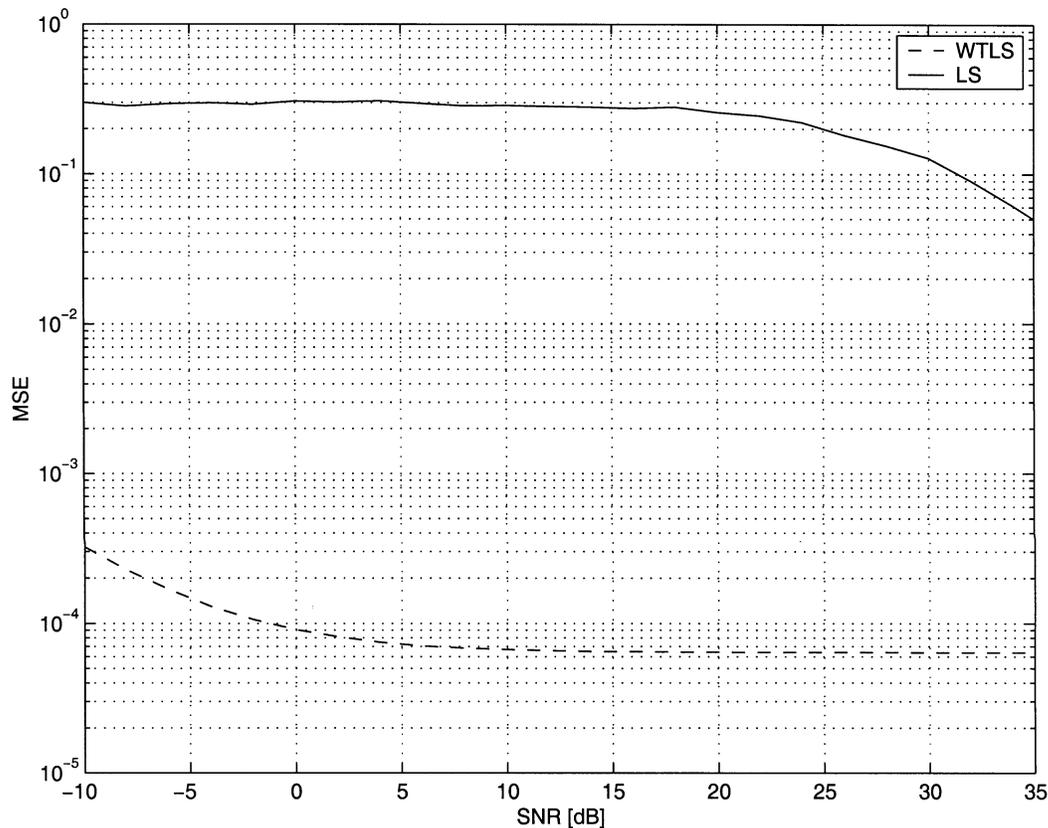
Fig. 1.   MSE in estimating a set of AR parameters using the LS estimator and the WTLS estimator.

where $\boldsymbol{y}' = \boldsymbol{y} - E(\boldsymbol{y})$, and $\hat{\boldsymbol{y}}' = \hat{\boldsymbol{y}} - E(\hat{\boldsymbol{y}})$. Minimizing (32) with respect to β, the optimal value of β is given by

$$\beta = \frac{\operatorname{Tr}\left((\boldsymbol{H}^*\boldsymbol{H})^{1/2}\right)}{\operatorname{Tr}(\bar{\boldsymbol{H}}^*\bar{\boldsymbol{H}})}. \tag{33}$$

Analysis of the MSE of the WTLS estimator [12] demonstrates that over a wide range of SNR, this estimator results in a lower MSE than the traditional LS estimator, for all values of the unknown parameters. Specifically, let $\zeta = \|\boldsymbol{x}\|^2/(\sigma^2 m)$ denote the SNR per dimension. Then with $\{\lambda_k, \ 1 \le k \le m\}$ denoting the eigenvalues of $\boldsymbol{H}^*\boldsymbol{H}$, where we assume that $\boldsymbol{H}$ has full rank, and denoting

$$\alpha = \frac{\sum\limits_{k=1}^{m} \lambda_k^{1/2}}{\sum\limits_{k=1}^{m} \lambda_k}$$

the MSE of the WTLS estimator is less than or equal to the MSE of the LS estimator for $\zeta \le \zeta_{\mathrm{WC}}$, where

$$\zeta_{\mathrm{WC}} = \frac{(1/m)\sum\limits_{k=1}^{m} \lambda_k^{-1} - \alpha^2}{\left|\alpha\lambda_\gamma^{1/2} - 1\right|^2} \tag{34}$$

and

$$\gamma = \arg\max \left|\alpha\lambda_k^{1/2} - 1\right|^2. \tag{35}$$

Note that ζ_{WC} is a worst case bound. In practice, the WTLS estimator will outperform the LS estimator for higher values of SNR than ζ_{WC}.

Examples presented in [12], [4] indicate that in a variety of applications ζ_{WC} can be quite large.

In Fig. 1, we illustrate the performance advantage in using the WTLS estimator with one simulation from [12]. In this figure, we plot the MSE in estimating a set of autoregressive (AR) parameters in an autoregressive moving average (ARMA) model contaminated by white noise, using both the WTLS and the LS estimators from 20 noisy observations of the channel, averaged over 2000 noise realizations, as a function of $-10\log\sigma^2$, where σ^2 is the noise variance. As can be seen from the figure, in this example, the WTLS estimator significantly outperforms the LS estimator. Further simulations presented in [4], [12] strongly suggest that the WTLS estimator can significantly decrease the MSE of the estimation error over the LS estimator for a wide range of SNR values.

APPENDIX
SUBSPACE MMSE WHITENING

Let $\bar{\boldsymbol{a}} = \boldsymbol{V}^*\boldsymbol{a}$ and $\bar{\boldsymbol{b}} = \boldsymbol{V}^*\boldsymbol{b}$, where $\boldsymbol{b}$ is white on $\mathcal{V}$ so that $\boldsymbol{b}$ has covariance $\boldsymbol{C}_b = c^2\boldsymbol{V}\tilde{\boldsymbol{I}}\boldsymbol{V}^*$. The covariance of $\bar{\boldsymbol{a}}$ is then $\boldsymbol{V}^*\boldsymbol{C}_a\boldsymbol{V} = \boldsymbol{D}$, and the covariance of $\bar{\boldsymbol{b}}$ is $\boldsymbol{V}^*\boldsymbol{C}_b\boldsymbol{V} = c^2\tilde{\boldsymbol{I}}$. As in MMSE whitening, instead of seeking a subspace whitening transformation that minimizes the MSE between $\boldsymbol{a}$ and $\boldsymbol{b}$, we may seek a transformation $\hat{\bar{\boldsymbol{W}}}_s$ such that the vector $\bar{\boldsymbol{b}} = \hat{\bar{\boldsymbol{W}}}_s\bar{\boldsymbol{a}}$ is as close as possible to $\bar{\boldsymbol{a}}$, and such that $\bar{\boldsymbol{b}}$ has covariance $c^2\tilde{\boldsymbol{I}}$. From (3) it then follows that $\hat{\boldsymbol{W}}_s = \boldsymbol{V}\hat{\bar{\boldsymbol{W}}}_s\boldsymbol{V}^*$, where $\hat{\boldsymbol{W}}_s$ is the optimal subspace whitening transformation that minimizes the MSE between $\boldsymbol{a}$ and $\boldsymbol{b}$.

Using the Cauchy–Schwarz inequality, it follows that $\hat{\bar{\boldsymbol{W}}}_s$ is such that $\bar{b}_k = c\bar{a}_k/\sqrt{d_k}$ for $1 \le k \le n$. Since the covariance of $\bar{\boldsymbol{b}}$ must be equal to $c^2\tilde{\boldsymbol{I}}$, $\hat{\bar{\boldsymbol{W}}}_s$ must also be chosen so that $\operatorname{var}(\bar{b}_k) = 0$ for $n+1 \le k \le m$. Now, the covariance of $\bar{\boldsymbol{a}}$ is $\boldsymbol{D}$, where the kth diagonal

element of $\boldsymbol{D}$ is equal to 0 for $n + 1 \leq k \leq m$. Consequently, $\overline{a}_k = 0$ w.p. 1 for $n + 1 \leq k \leq m$. Therefore, we conclude that $\hat{\overline{W}}_s$ is block diagonal. The upper left $n \times n$ block is a diagonal matrix, with diagonal elements $c / \sqrt{d_k}$; the lower right block is arbitrary, since $\overline{b}_k = \overline{a}_k = 0$ regardless of the choice of this block. We, therefore, choose $\hat{\overline{W}}_s$ to be a diagonal matrix with the first n diagonal elements equal to $c / \sqrt{d_k}$ and the remaining diagonal elements equal to 0. Thus, $\hat{\overline{W}}_s = c (\boldsymbol{D}^{1/2})^{\dagger}$, and

$$\hat{\boldsymbol{W}}_s = c \boldsymbol{V} (\boldsymbol{D}^{1/2})^{\dagger} \boldsymbol{V}^* = c (\boldsymbol{C}_a^{1/2})^{\dagger}. \tag{36}$$

If we choose to minimize the MSE with respect to c as well, then it is straightforward to show that the optimal value of c is given by

$$\alpha_s = \frac{1}{n} \sum_{k=1}^{n} \sqrt{d_k}.$$

References

[1] E. M. Friel and K. M. Pasala, "Direction finding with compensation for a near field scatterer," in *Proc. Int. Symp. Antennas and Propagation Society*, 1995, pp. 106–109.

[2] R. J. Piechocki, N. Canagarajah, and J. P. McGeehan, "Improving the direction-of-arrival resolution via double code filtering in WCDMA," in *Proc. 1st Int. Conf. 3G Mobile Communication Technologies*, Mar. 2000, pp. 204–207.

[3] Y. C. Eldar, A. V. Oppenheim, and D. Egnor, "Orthogonal and projected orthogonal matched filter detection," *Signal Processing*, submitted for publication.

[4] Y. C. Eldar, "Quantum signal processing," Ph.D. dissertation, MIT, Cambridge, MA, Dec. 2001. also available at http://allegro.mit.edu/ dspg/publications/TechRep/index.html.

[5] Y. C. Eldar and A. M. Chan, "An optimal whitening approach to linear multiuser detection," *IEEE Trans. Inform. Theory*, to be published.

[6] Y. C. Eldar and G. D. Forney, Jr., "On quantum detection and the square-root measurement," *IEEE Trans. Inform. Theory*, vol. 47, pp. 858–872, Mar. 2001.

[7] Y. C. Eldar, "Least-squares inner product shaping," *Linear Alg. Appl.*, vol. 348, pp. 153–174, May 2002.

[8] ——, "Least-squares orthogonalization using semidefinite programming," *Linear Alg. Appl.*, submitted for publication.

[9] Y. C. Eldar and G. D. Forney, Jr., "Optimal tight frames and quantum measurement," *IEEE Trans. Inform. Theory*, vol. 48, pp. 599–610, Mar. 2002.

[10] Y. C. Eldar and H. Bölcskei, "Geometrically uniform frames," *IEEE Trans. Inform. Theory*, vol. 49, pp. 993–1006, Apr. 2003.

[11] Y. C. Eldar, "Minimum mean-squared error covariance shaping," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, vol. 6, Hong Kong, Apr. 2003, pp. 713–716.

[12] Y. C. Eldar and A. V. Oppenheim, "Covariance shaping least-squares estimation," *IEEE Trans. Signal Processing*, vol. 51, pp. 696–697, Mar. 2003.

[13] Y. C. Eldar, "Covariance shaping approach to linear least-squares estimation," in *Proc. Asilomar Conf. Signals, Systems, and Computers*, Nov. 2002.

[14] Y. C. Eldar and A. V. Oppenheim, "Quantum signal processing," *IEEE Signal Processing Mag.*, pp. 12–32, Nov. 2002.

[15] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.

[16] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[17] J. J. Atick and A. N. Redlich, "Convergent algorithm for sensory receptive field development," *Neurak Comp.*, vol. 5, pp. 45–60, 1993.

[18] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[19] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ: Prentice-Hall, 1993.

Deviation Bounds for Wavelet Shrinkage

Dawei Hong and Jean-Camille Birget

Abstract—We analyze the wavelet shrinkage algorithm of Donoho and Johnstone in order to assess the quality of the reconstruction of a signal obtained from noisy samples. We give a deviation estimate for the maximum squared error (and, consequently, for the average squared error), under the assumption that the signal comes from a Hölder class, and the noise samples are independent, of zero mean, and bounded. Our main technique is Talagrand's isoperimetric theorem. Our result shows a better behavior of the wavelet shrinkage.

Index Terms—Deviation bound, maximum squared error, wavelet shrinkage.

I. Introduction

We address the classical problem of the reconstruction of signal samples from noisy samples. We consider an original signal of bounded duration $f : t \in [0, 1] \rightarrow f(t) \in \mathrm{R}$. We also have additive noise $e : [0, 1] \rightarrow \mathrm{R}$. Thus, the observed *noisy signal* at time t is $y(t) = f(t) + e(t)$.

We sample the noisy signal at n uniformly spaced instants and we denote the sample values by

$$y_i = f_i + e_i = f\left(\frac{i}{n}\right) + e\left(\frac{i}{n}\right) \qquad (\text{for } 1 \leq i \leq n).$$

Our goal is to recover a good approximation of the original signal samples $(f_1, \ldots, f_n)$ from the noisy signal samples $(y_1, \ldots, y_n)$. For this to be possible we need some assumptions that distinguish the signal from the noise.

• The original signal f has a certain degree of "smoothness," i.e., f belongs to a Hölder class $\Lambda^{\alpha}(M)$ for some $\alpha > 0$ and $M > 0$.

• The noise is "random," i.e., $(e_1, \ldots, e_n)$ consists of n independent random variables.

The *Hölder classes* are defined as follows:

For $0 < \alpha \leq 1$, $\quad \Lambda^{\alpha}(M) = \{ h \in \mathrm{R}^{[0, 1]} :$
$(\forall x_1, x_2 \in [0, 1]), |h(x_1) - h(x_2)| \leq M |x_1 - x_2|^{\alpha} \}$.

For $1 < \alpha$, $\quad \Lambda^{\alpha}(M) = \{ h \in \mathrm{R}^{[0, 1]} : (\forall x \in [0, 1])$
$|h'(x)| \leq M, h^{\lfloor \alpha \rfloor}$ exists, and $(\forall x_1, x_2 \in [0, 1])$
$|h^{\lfloor \alpha \rfloor}(x_1) - h^{\lfloor \alpha \rfloor}(x_2)| \leq M |x_1 - x_2|^{\alpha - \lfloor \alpha \rfloor} \}$.

Let $(\tilde{y}_1, \ldots, \tilde{y}_n)$ be an approximation of $(f_1, \ldots, f_n)$, obtained from $(y_1, \ldots, y_n)$. Most commonly, the closeness of this approximation is measured by $\frac{1}{n} \sum_{i=1}^{n} (\tilde{y}_i - f_i)^2$ or by the expectation $\mathrm{E}[\frac{1}{n} \sum_{i=1}^{n} (\tilde{y}_i - f_i)^2]$.

The wavelet shrinkage algorithm of Donoho and Johnstone [6], [7] is a very efficient tool for finding good estimates $\tilde{y}$. In outline, the algorithm works as follows: