Optimal Reduced-Rank Estimation and Filtering

Yingbo Hua, *Senior Member, IEEE*, Maziar Nikpour, and Petre Stoica, *Fellow, IEEE*

***Abstract*—This paper provides a unified view of, and a further insight into, a class of optimal reduced-rank estimators and filters. An alternating power (AP) method for computing the optimal reduced-rank estimators and filters is derived and analyzed. The AP method is a generalization of the conventional power method for subspace computation, which is shown to be globally and exponentially convergent under weak conditions. When the rank reduction is relatively large, the AP method is computationally more efficient than the conventional methods. The AP method is useful for adaptive computation of the canonical components of a desired reduced-rank estimate, which in turn facilitates the detection of a time-varying rank. The study shown in this paper is particularly useful for applications that involve a large number of sources and a large number of receivers, where rank reduction is either inherent in the multivariate system or required to reduce the model complexity and/or the computational load.**

***Index Terms*—Alternating power method, power method, rank estimation, rank reduction, reduced-rank channel equalization, reduced-rank channel identification, reduced-rank estimators, reduced-rank filters, reduced-rank maximum likelihood estimation, reduced-rank multilayer neural network, reduced-rank Wiener filter, SVD.**

I. Introduction

In RECENT research of wireless communications and telephone networks [2], the following multivariate linear regression model has attracted considerable attention:

$$\mathbf{y}(k) = \mathbf{T}\mathbf{x}(k) + \mathbf{e}(k). \tag{1}$$

Here,

$\mathbf{y}(k) \in C^n$	channel output;
$\mathbf{x}(k) \in C^m$	channel input;
$\mathbf{T} \in C^{n\times m}$	channel matrix;
$\mathbf{e}(k) \in C^n$	channel noise vector.

The channel output vector can be the output of multiple receivers and/or slide-windowed sequences of the original output signals. The effects of multipath signals and cross-signal interference can be described or represented by the internal structure of the channel matrix. In general, the channel matrix has a possibly reduced-rank $r < \min(m, n)$. Once the channel transfer function represented by $\mathbf{T}$ is estimated, efficient ways are available to estimate jointly the angles and delays of multipath signals [18]. As discussed in [1], the model (1) is also applicable to a range of other applications where multiple sensors and multiple transmitters are employed. For such reduced-rank problems, reduced-rank estimators or filters are required for estimating the channel matrix and/or the channel input.

Indeed, reduced-rank estimation and filtering are important for a wide range of signal processing applications where data or model reduction, robustness against noise or model errors, or high computational efficiency is desired. Fundamental results on optimal reduced-rank estimators and filters include the work by Brillinger [17], the reduced-rank Wiener filter (RRWF) by Scharf [3], [11]–[13], and the reduced-rank maximum likelihood estimation (RRMLE) by Stoica–Viberg [1]. Other examples of the reduced-rank estimators and filters include the reduced-rank multilayer neural network (RRMNN) by Diamantara–Kung [6], the relative Karhunen–Loeve transform (RKLT) by Yamashita–Ogawa [4], and the generalized Karhunen–Loeve transform (GKLT) by Hua–Liu [5]. In Section II, we provide a unified view of, and a further insight into, these optimal reduced rank estimators and filters.

A fundamental tool for reduced-rank estimation and filtering is the singular value decomposition (SVD) [7]. Indeed, most (if not all) reduced-rank techniques known so far can be expressed in terms of SVD or its related eigenvalue decomposition (EVD) or subspace decomposition (SSD). This reality has driven the search for fast algorithms for computing the SVD, EVD, and SSD and their adaptive forms. Some of the early research work in this direction was done by Tufts, among others [16]. More recent results can be found in [14], [15], [23], and the references therein. However, for many reduced-rank estimators and filters, the SSD, EVD, or SVD is only an intermediate part of a more complex process. A fast algorithm for SSD, EVD, or SVD alone may not be sufficient to make the whole process computationally efficient. In Section III, we show an alternating power (AP) method for computing the reduced-rank estimators and filters. If the rank reduction is relatively large, the AP method is much more efficient in computation than the conventional methods that require additional computations before and after some fast SSD, EVD, or SVD is employed. The AP method is a generalization of the power method [7], [10], [15] for computing the principal components of a given matrix. As a computational tool, the AP method is also related to the back propagation (BP) method [6] for linear multilayer neural network learning where a rank reduction is implemented via reduced number of inner neurons. However, the AP method and the BP method are based on different computational principles. The AP method is a generalization of an iterative quadratic minimum distance (IQMD) ap-

Manuscript received July 21, 1999; revised September 15, 2000. This work was supported in part by the Australian Research Council Large Grant, the Cooperative Research Centre for Sensor Signal and Information Processing, and the Senior Individual Grant Program of the Swedish Foundation for Strategic Research. The associate editor coordinating the review of this paper and approving it for publication was Prof. Dimitrios Hatzinakos.

Y. Hua was with the Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong. He is now with the Department of Electrical Engineering, University of California, Riverside, CA 92521 USA (e-mail: eeyhua@ee.ust.hk).

M. Nikpour is with the Department of Electrical and Electronic Engineering, University of Melbourne, Victoria, Australia (e-mail: mazn@ee.mu.oz.au).

P. Stoica is with the Department of Systems and Control, Uppsala University, Uppsala, Sweden (e-mail: ps@syscon.uu.se).

Publisher Item Identifier S 1053-587X(01)01410-6.

proach [8], and the BP method is based on the gradient descent searching. The AP method converges much faster than the BP method as the latter requires the use of a very small step size. If the step size is not small enough, the BP method diverges. (If the inverse of a Hessian matrix were used, the computation at each iteration would be significantly increased.) In Section III-A, we derive the AP method. In Section III-B, we establish the global and exponential convergence property of the AP method. In Section III-C, we demonstrate how the AP method can be used to compute the canonical components of reduced-rank estimators and filters. In Section III-D, the detection of time-varying ranks is discussed. In Section III-E, some issues of adaptive computation are addressed. Simulation results are provided in Section IV.

II. Review of Reduced-Rank Estimation and Filtering

A. Main Framework

A unified view of several optimal reduced-rank estimators and filters is shown next. Consider the two random (complex[1]) processes $\mathbf{x}(k) \in C^m$ and $\mathbf{y}(k) \in C^n$, which may or may not satisfy the model (1). Let $\hat{\mathbf{y}}(k) = \hat{\mathbf{T}}\mathbf{x}(k)$ be a reduced-rank estimate[2] of $\mathbf{y}(k)$ from $\mathbf{x}(k)$ for some matrix $\hat{\mathbf{T}} \in C^{n\times m}$ and rank $r < \min(m, n)$. The correlation matrix of the error vector $\mathbf{z}(k) = \mathbf{y}(k) - \hat{\mathbf{y}}(k)$ can be expressed as follows:

$$\begin{aligned}\mathbf{C_{zz}} &= E\{\mathbf{z}(k)\mathbf{z}(k)^H\} \\ &= \mathbf{C_{yy}} - \mathbf{C_{yx}}\hat{\mathbf{T}}^H - \hat{\mathbf{T}}\mathbf{C}_{\mathbf{yx}}^H + \hat{\mathbf{T}}\mathbf{C_{xx}}\hat{\mathbf{T}}^H \end{aligned} \tag{2}$$

where

$E\{\,\}$	either ensemble average or time-averaging over a finite set of data (i.e., both definitions are valid in this paper);
H	conjugate transpose;
$\mathbf{C_{yy}} = E\{\mathbf{y}(k)\mathbf{y}(k)^H\}$	auto-correlation matrix of $\mathbf{y}(k)$;
$\mathbf{C_{xx}} = E\{\mathbf{x}(k)\mathbf{x}(k)^H\}$	auto-correlation matrix of $\mathbf{x}(k)$;
$\mathbf{C_{yx}} = E\{\mathbf{y}(k)\mathbf{x}(k)^H\}$	cross-correlation matrix between $\mathbf{y}(k)$ and $\mathbf{x}(k)$.

We assume that $\mathbf{C_{xx}}$ and $\mathbf{C_{yy}}$ are nonsingular. The optimum choice of the filtering matrix $\hat{\mathbf{T}}$ depends on the measure applied to $\mathbf{C_{zz}}$. There are three common measures:

$$\begin{cases} J_{\rm tr} = {\rm tr}\{\mathbf{C_{zz}}\} \\ J_{w-{\rm tr}} = {\rm tr}\{\mathbf{W}^{-1}\mathbf{C_{zz}}\} \\ J_{\rm det} = \det\{\mathbf{C_{zz}}\} \end{cases}$$

where tr{ } denotes trace, det{ } determinant, and $\mathbf{W}$ is a nonsingular weighting matrix. The minimizers of the three measures are different in general. It is obvious that the minimizer of $J_{\rm tr}$ is a special case of that of $J_{w-{\rm tr}}$, but it is not obvious how the minimizers of $J_{w-{\rm tr}}$ and $J_{\rm det}$ are related to each other. Such a connection will be given below.

[1]Complex data are assumed in this paper unless specified otherwise.

[2]A constant offset vector may be used as in [18], but it is a trivial part of the process and, hence, omitted here.

The SVD of matrices will be used frequently. We denote the SVD of a matrix $\mathbf{R} \in C^{n\times m}$ as $\mathbf{R} = \sum_{i=1}^{\min(m,n)} \sigma_i \mathbf{u}_i \mathbf{v}_i^H$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)}$. We also define

$$\begin{aligned} \Sigma_1 &= {\rm diag}[\sigma_1 \quad \sigma_2 \quad \cdots \quad \sigma_r] \\ \Sigma_2 &= {\rm diag}[\sigma_{r+1} \quad \sigma_{r+2} \quad \cdots \quad \sigma_{\min(m,n)}] \\ \mathbf{U}_1 &= U_1(\mathbf{R}) = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_r] \\ \mathbf{U}_2 &= U_2(\mathbf{R}) = [\mathbf{u}_{r+1} \quad \mathbf{u}_{r+2} \quad \cdots \quad \mathbf{u}_{\min(m,n)}] \\ \mathbf{V}_1 &= V_1(\mathbf{R}) = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_r] \\ \mathbf{V}_2 &= V_2(\mathbf{R}) = [\mathbf{v}_{r+1} \quad \mathbf{v}_{r+2} \quad \cdots \quad \mathbf{v}_{\min(m,n)}] \end{aligned}$$

$$trun_r(\mathbf{R}) = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^H.$$

It is known from [17, Th. 10.2.4] that the minimizer of $J_{w-{\rm tr}}$ is given by

$$\hat{\mathbf{T}}_{w-{\rm tr}} = \mathbf{W}^{1/2}\mathbf{U}_{w-{\rm tr},1}\mathbf{U}_{w-{\rm tr},1}^H\mathbf{W}^{-1/2}\mathbf{C_{yx}}\mathbf{C_{xx}^{-1}} \tag{3a}$$

where $\mathbf{U}_{w-{\rm tr},1} = U_1(\mathbf{R}_{w-{\rm tr}})$, and $\mathbf{R}_{w-{\rm tr}} = \mathbf{W}^{-1/2}\mathbf{C_{yx}}\mathbf{C_{xx}}^{-1/2^H}$. The superscript "1/2" denotes the square root and the superscript "−1/2" the inverse square root. The square-root matrices are not required to be symmetric. Namely, $\mathbf{W} = \mathbf{W}^{1/2}\mathbf{W}^{1/2^H}$, and $\mathbf{C_{xx}} = \mathbf{C_{xx}^{1/2}}\mathbf{C_{xx}^{1/2^H}}$. It is easy to verify that an alternative form of (3a) is

$$\hat{\mathbf{T}}_{w-{\rm tr}} = \mathbf{C_{yx}}\mathbf{C_{xx}}^{-1/2^H}\mathbf{V}_{w-{\rm tr},1}\mathbf{V}_{w-{\rm tr},1}^H\mathbf{C_{xx}^{-1/2}} \tag{3b}$$

where $\mathbf{V}_{w-{\rm tr},1} = V_1(\mathbf{R}_{w-{\rm tr}})$. Clearly, the minimizer $\hat{\mathbf{T}}_{\rm tr}$ of $J_{\rm tr}$ is given by (3) with $\mathbf{W} = \mathbf{I}$ (the identity matrix). The matrix $\mathbf{R}_{w-{\rm tr}}$ with $\mathbf{W} = \mathbf{I}$ will be denoted by $\mathbf{R}_{\rm tr}$.

We now consider $J_{\rm det}$. One can verify using (2) that

$$J_{\rm det} = \det\{\mathbf{I} - \mathbf{R}_{\rm det}\mathbf{R}_{\rm det}^H + \mathbf{Z}\mathbf{Z}^H\}\det\{\mathbf{C_{yy}}\}$$

where $\mathbf{R}_{\rm det} = \mathbf{C_{yy}^{-1/2}}\mathbf{C_{yx}}\mathbf{C_{xx}}^{-1/2^H}$, and $\mathbf{Z} = \mathbf{R}_{\rm det} - \mathbf{C_{yy}^{-1/2}}\hat{\mathbf{T}}\mathbf{C_{xx}^{1/2}}$. It follows that $J_{\rm det}$ is proportional to $\det\{\mathbf{I} - \mathbf{R}_{\rm det}\mathbf{R}_{\rm det}^H + \mathbf{Z}\mathbf{Z}^H\}$. It is known (easy to prove) that all singular values of $\mathbf{R}_{\rm det}$ are no larger than one. If a singular value of $\mathbf{R}_{\rm det}$ equals one, we can show that the minimum of $J_{\rm det}$ is zero and is achieved by a wide range of minimizers. Therefore, we need to assume here that all singular values of $\mathbf{R}_{\rm det}$ are strictly less than one, which is satisfied in practice with probability one unless $\mathbf{y}(k)$ is a linear transform of $\mathbf{x}(k)$ or vice versa. With this assumption, $J_{\rm det}$ is equivalent to

$$\begin{aligned} &\det\left\{\mathbf{I} + (\mathbf{I} - \mathbf{R}_{\rm det}\mathbf{R}_{\rm det}^H)^{-1/2}\mathbf{Z}\mathbf{Z}^H(\mathbf{I} - \mathbf{R}_{\rm det}\mathbf{R}_{\rm det}^H)^{-1/2}\right\} \\ &\quad = \prod_i (1 + \varpi_i^2) \end{aligned}$$

where ϖ_i is the ith singular value of $(\mathbf{I} - \mathbf{R}_{\rm det}\mathbf{R}_{\rm det}^H)^{-1/2}\mathbf{Z}$. Here, the square root is assumed to be conjugate symmetric without loss of generality. Hence, minimizing $J_{\rm det}$ is equivalent

to minimizing all ϖ_i. The matrix $(\mathbf{I}-\mathbf{R}_{\rm det}\mathbf{R}_{\rm det}^H)^{-1/2}\mathbf{Z}$ can be rewritten as

$$(\mathbf{I}-\mathbf{R}_{\rm det}\mathbf{R}_{\rm det}^H)^{-1/2}\mathbf{R}_{\rm det}-(\mathbf{I}-\mathbf{R}_{\rm det}\mathbf{R}_{\rm det}^H)^{-1/2}\mathbf{C}_{\mathbf{yy}}^{-1/2}\hat{\mathbf{T}}\mathbf{C}_{\mathbf{xx}}^{1/2}.$$

It is known (see [17, Th. 3.7.4], for example) that given a fixed matrix $\mathbf{M}$ and a rank-r matrix $\mathbf{N}$ (of the same dimensions as $\mathbf{M}$), all the singular values of $\mathbf{M}-\mathbf{N}$ are minimized by $\mathbf{N}$ if $\mathbf{N}=trun_r(\mathbf{M})$. Therefore, all ϖ_i are minimized by the rank-r matrix $\hat{\mathbf{T}}_{\rm det}$ that satisfies the following:

$$\begin{aligned}&(\mathbf{I}-\mathbf{R}_{\rm det}\mathbf{R}_{\rm det}^H)^{-1/2}\mathbf{C}_{\mathbf{yy}}^{-1/2}\hat{\mathbf{T}}_{\rm det}\mathbf{C}_{\mathbf{xx}}^{1/2}\\&\quad=trun_r((\mathbf{I}-\mathbf{R}_{\rm det}\mathbf{R}_{\rm det}^H)^{-1/2}\mathbf{R}_{\rm det}).\end{aligned}$$

Since the eigenvectors of $(\mathbf{I}-\mathbf{R}_{\rm det}\mathbf{R}_{\rm det}^H)^{-1/2}$ are the same as the left singular vectors of $\mathbf{R}_{\rm det}$, one can verify that

$$\begin{aligned}&trun_r((\mathbf{I}-\mathbf{R}_{\rm det}\mathbf{R}_{\rm det}^H)^{-1/2}\mathbf{R}_{\rm det})\\&\quad=(\mathbf{I}-\mathbf{R}_{\rm det}\mathbf{R}_{\rm det}^H)^{-1/2}trun_r(\mathbf{R}_{\rm det}).\end{aligned}$$

The above two equations imply that $\mathbf{C}_{\mathbf{yy}}^{-1/2}\hat{\mathbf{T}}_{\rm det}\mathbf{C}_{\mathbf{xx}}^{1/2}=trun_r(\mathbf{R}_{\rm det})$, and hence, $\hat{\mathbf{T}}_{\rm det}=\mathbf{C}_{\mathbf{yy}}^{1/2}trun_r(\mathbf{R}_{\rm det})\mathbf{C}_{\mathbf{xx}}^{-1/2}$. It is then easy to verify that

$$\hat{\mathbf{T}}_{\rm det}=\mathbf{C}_{\mathbf{yy}}^{1/2}\mathbf{U}_{{\rm det},1}\mathbf{U}_{{\rm det},1}^H\mathbf{C}_{\mathbf{yy}}^{-1/2}\mathbf{C}_{\mathbf{yx}}\mathbf{C}_{\mathbf{xx}}^{-1} \tag{4a}$$

or, alternatively

$$\hat{\mathbf{T}}_{\rm det}=\mathbf{C}_{\mathbf{yx}}\mathbf{C}_{\mathbf{xx}}^{-1/2}\mathbf{V}_{{\rm det},1}\mathbf{V}_{{\rm det},1}^H\mathbf{C}_{\mathbf{xx}}^{-1/2} \tag{4b}$$

where $\mathbf{U}_{{\rm det},1}=U_1(\mathbf{R}_{\rm det})$ and $\mathbf{V}_{{\rm det},1}=V_1(\mathbf{R}_{\rm det})$. It is clear that (4) is a special case of (3), i.e., $\hat{\mathbf{T}}_{\rm det}=\hat{\mathbf{T}}_{w-{\rm tr}}$ with $\mathbf{W}=\mathbf{C}_{\mathbf{yy}}$.

Note that the three minimizers (reduced-rank estimators/filters) $\hat{\mathbf{T}}_{\rm tr}$, $\hat{\mathbf{T}}_{w-{\rm tr}}$, and $\hat{\mathbf{T}}_{\rm det}$ are dependent on the first r principal (left or right) singular vectors of the three "characteristic" matrices $\mathbf{R}_{\rm tr}$, $\mathbf{R}_{w-{\rm tr}}$, and $\mathbf{R}_{\rm det}$, respectively. The singular vectors of these matrices are called the canonical coordinates with respect to $J_{\rm tr}$, $J_{w-{\rm tr}}$ and $J_{\rm det}$, respectively. From now on, we will assume that the square roots used in (3a), (3b), (4a), and (4b) are conjugate symmetric for convenience, unless specified otherwise.

B. Relations to Some Existing Results

It is easy to verify from (3b) and (4b) that without rank reduction [i.e., $r=\min(m,n)$], the minimizer of each of the three measures is simply the well-known Wiener filter $\hat{\mathbf{T}}_{WF}=\mathbf{C}_{\mathbf{yx}}\mathbf{C}_{\mathbf{xx}}^{-1}$. The reduced-rank Wiener filter (RRWF) [3], [11] is simply $\hat{\mathbf{T}}_{\rm tr}$. The singular vectors of $\mathbf{R}_{\rm tr}=\mathbf{C}_{\mathbf{yx}}\mathbf{C}_{\mathbf{xx}}^{-1/2}$ are called the canonical coordinates of the RRWF [3], [11]. A newest version of the RRWF is shown in [12] and [13], where the singular vectors of $\mathbf{R}_{\rm det}=\mathbf{C}_{\mathbf{yy}}^{-1/2}\mathbf{C}_{\mathbf{yx}}\mathbf{C}_{\mathbf{xx}}^{-1/2}$ are used as the "full" canonical coordinates to form the RRWF. The "full" RRWF is simply $\hat{\mathbf{T}}_{\rm det}$ or $\hat{\mathbf{T}}_{w-{\rm tr}}$ with $\mathbf{W}=\mathbf{C}_{\mathbf{yy}}$. The matrix $\mathbf{R}_{\rm det}$ is called the coherence matrix in [12] and [13].

The multivariate linear regression model (1) is considered in [1], where $\mathbf{e}(k)$ is assumed to be uncorrelated with $\mathbf{x}(k)$, temporally white Gaussian, and of zero-mean and an unknown covariance matrix $\mathbf{C}_{\mathbf{ee}}$. As shown in [1], the reduced-rank maximum likelihood estimate (RRMLE) of $\mathbf{T}$ is simply $\hat{\mathbf{T}}_{\rm det}$. Note that the expression of (4b) was also obtained in [1], but for the real valued data, and a different approach was used there.

Another variation of the RRWF is available in [4] and [5] and referred to in [5] as the generalized Karhunen–Loeve transform (GKLT), where $\mathbf{C}_{\mathbf{xx}}^{-1/2}$ is allowed to be the pseudoinverse of the square root of the (possibly singular) matrix $\mathbf{C}_{\mathbf{xx}}$. The RRMNN [6] is a neural network version of the GKLT, where the generalized singular value decomposition (GSVD) is used in the presentation of the optimal reduced-rank transform.

We note that if $\mathbf{C}_{\mathbf{xx}}$ and $\mathbf{W}$ are possibly singular, then the rank-r minimizer of $J_{w-{\rm tr}}$ is not unique, but the minimizer with the minimum F-norm is still given by (3), except that the inverses should be interpreted as the pseudoinverses. Such a proof can be obtained by a simple modification of the proof for the GKLT [5]. In the rest of this paper, we only address the nonsingular case, i.e., $\mathbf{C}_{\mathbf{xx}}$ and $\mathbf{W}$ (and $\mathbf{C}_{\mathbf{yy}}$) are nonsingular.

C. FIR and IIR Filtering

The framework shown in Section II-A is also applicable to finite impulse response (FIR) channels. For example, if one is interested to identify the multi-input and multi-output (MIMO) FIR system $\mathbf{y}(k)=\sum_{l=0}^{L}\mathbf{T}(l)\mathbf{x}(k-l)$, one can construct the "expanded" vectors and matrix $\mathbf{x}'(k)$, $\mathbf{y}'(k)$, and $\mathbf{T}'$ from $\mathbf{x}(k)$, $\mathbf{y}(k)$, and $\mathbf{T}(l)$, respectively, such that $\mathbf{y}'(k)=\mathbf{T}'\mathbf{x}'(k)$ (see, for example, [19]–[21]). Provided that the rank of $\mathbf{T}'$ [not $\mathbf{T}(k)$] is of interest, the formulation shown in Section II-A clearly holds. One should note, however, that with the reconstructed model $\mathbf{y}'(k)=\mathbf{T}'\mathbf{x}'(k)$, the corresponding noise term in (1) is generally temporally correlated and has some well-defined structure. In this case, the RRMLE is unknown, and the statistical analyzes shown in [1], [17], and [22] are invalid. More research in this direction is desirable.

In some applications, one may be interested in a more general setup as follows. Assume a system $\mathbf{y}(k)=\mathbf{T}(k)*\mathbf{x}(k)$, where $\mathbf{T}(k)=\Phi(k)*\Psi(k)^H$, $\Phi(k)\in C^{n\times r}$, $\Psi(k)\in C^{m\times r}$, and the operator $*$ denotes convolution. The optimum reduced-rank filters $\Phi(k)$ and $\Psi(k)$ can be chosen such that one of the cost functions $J_{\rm tr}$, $J_{w-{\rm tr}}$, or $J_{\rm det}$ is minimized. An early study based on $J_{\rm tr}$ is available in [17], although the optimal causal filters remain an open research topic. This paper will not address this area further, but it is important to note that the AP method shown in Section III can be modified to compute the optimum filters $\Phi(k)$ and $\Psi(k)$ given in [17] through the spectral density functions of $\mathbf{x}(k)$ and $\mathbf{y}(k)$.

III. Computation of Reduced-Rank Estimators and Filters

The efficient computation of the reduced-rank estimators or filters may seem straightforward as there are indeed efficient algorithms to compute the principal singular vectors of any of the characteristic matrices $\mathbf{R}_{\rm tr}$, $\mathbf{R}_{w-{\rm tr}}$, and $\mathbf{R}_{\rm det}$. However, the major computational burden here is not just the SVD of a given matrix. Consider $\hat{\mathbf{T}}_{\rm det}$ in (4b), for example. A conventional method to compute $\hat{\mathbf{T}}_{\rm det}$ first requires the computation of $\mathbf{R}_{\rm det}=\mathbf{C}_{\mathbf{yy}}^{-1/2}\mathbf{C}_{\mathbf{yx}}\mathbf{C}_{\mathbf{xx}}^{-1/2}$, then the SVD (or the like) of $\mathbf{R}_{\rm det}$, and, finally, the product $\hat{\mathbf{T}}_{\rm det}=\mathbf{C}_{\mathbf{yx}}\mathbf{C}_{\mathbf{xx}}^{-1/2}\mathbf{V}_{{\rm det},1}\mathbf{V}_{{\rm det},1}^H\mathbf{C}_{\mathbf{xx}}^{-1/2}$. The computations

required before and after the SVD involves, in particular, square-root-inverses of large matrices, which alone requires more than $O(\min(m^2n, n^2m))$ flops.

The objective of this section is to present an efficient approach to computing the reduced-rank estimators and filters for the case where $r \ll \min(m, n)$. Without loss of generality, however, we will focus on the computation of $\hat{\mathbf{T}}_{\rm det}$. For simplicity of notation, we will use $\hat{\mathbf{T}}_{\rm opt} = \hat{\mathbf{T}}_{\rm det}, J_{\rm opt} = tr\{\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{C}_{\mathbf{zz}}\}$, and $\mathbf{R}_{\rm opt} = \mathbf{R}_{\rm det}$. It is clear that $\hat{\mathbf{T}}_{\rm opt}$ is the minimizer of $J_{\rm opt}$.

A. Alternating Power Method

We write

$$J_{\rm opt} = \text{tr}\{\mathbf{C}_{\mathbf{yy}}^{-1}(\mathbf{C}_{\mathbf{yy}} - \mathbf{C}_{\mathbf{yx}}\hat{\mathbf{T}}^H - \hat{\mathbf{T}}\mathbf{C}_{\mathbf{yx}}^H + \hat{\mathbf{T}}\mathbf{C}_{\mathbf{xx}}\hat{\mathbf{T}}^H)\}. \quad (5)$$

We can also write the rank-r matrix $\hat{\mathbf{T}}$ as $\mathbf{AB}^H$, where $\mathbf{A} \in C^{n\times r}$ and $\mathbf{B} \in C^{m\times r}$. One can then verify that

$$\begin{aligned} J_{\rm opt} =& \text{tr}(\mathbf{C}_{\mathbf{yy}}^{-1}(\mathbf{A}(\mathbf{B}^H\mathbf{C}_{\mathbf{xx}}\mathbf{B}) - \mathbf{C}_{\mathbf{yx}}\mathbf{B})(\mathbf{B}^H\mathbf{C}_{\mathbf{xx}}\mathbf{B})^{-1} \\ &\cdot (\mathbf{A}(\mathbf{B}^H\mathbf{C}_{\mathbf{xx}}\mathbf{B}) - \mathbf{C}_{\mathbf{yx}}\mathbf{B})^H) + f(\mathbf{B}) \\ =& \text{tr}(\mathbf{C}_{\mathbf{xx}}^{-1}(\mathbf{C}_{\mathbf{xx}}\mathbf{B}(\mathbf{A}^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{A}) - \mathbf{C}_{\mathbf{yx}}^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{A}) \\ &\cdot (\mathbf{A}^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{A})^{-1}(\mathbf{C}_{\mathbf{xx}}\mathbf{B}(\mathbf{A}^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{A}) \\ &- \mathbf{C}_{\mathbf{yx}}^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{A})^H) + g(\mathbf{A}) \quad (6) \end{aligned}$$

where

$$\begin{aligned} f(\mathbf{B}) =& n - \text{tr}(\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{C}_{\mathbf{yx}}\mathbf{B}(\mathbf{B}^H\mathbf{C}_{\mathbf{xx}}\mathbf{B})^{-1}\mathbf{B}^H\mathbf{C}_{\mathbf{yx}}^H) \\ g(\mathbf{A}) =& n - \text{tr}(\mathbf{C}_{\mathbf{xx}}^{-1}\mathbf{C}_{\mathbf{yx}}^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{A}(\mathbf{A}^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{A})^{-1}\mathbf{A}^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{C}_{\mathbf{yx}}). \end{aligned}$$

We now try to minimize $J_{\rm opt}$ with respect to $\mathbf{A}$ and $\mathbf{B}$, alternately. Let k denote the index of iteration. Given $\mathbf{B}(k)$, the new $\mathbf{A}(k+1)$ is obtained by minimizing (6) with respect to $\mathbf{A}$, and then, the new $\mathbf{B}(k+1)$ is obtained by minimizing (6) with respect to $\mathbf{B}$. A simple analysis of (6) shows that the above process leads to the following iterative equations:

$$\begin{cases} \mathbf{A}(k+1)(\mathbf{B}(k)^H\mathbf{C}_{\mathbf{xx}}\mathbf{B}(k)) - \mathbf{C}_{\mathbf{yx}}\mathbf{B}(k) = 0 \\ \mathbf{C}_{\mathbf{xx}}\mathbf{B}(k+1)(\mathbf{A}(k+1)^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{A}(k+1)) \\ \quad -\mathbf{C}_{\mathbf{yx}}^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{A}(k+1) = 0 \end{cases} \quad (7a)$$

or, equivalently

$$\begin{cases} \mathbf{A}(k+1) = \mathbf{C}_{\mathbf{yx}}\mathbf{B}(k)(\mathbf{B}(k)^H\mathbf{C}_{\mathbf{xx}}\mathbf{B}(k))^{-1} \\ \mathbf{B}(k+1) = \mathbf{C}_{\mathbf{xx}}^{-1}\mathbf{C}_{\mathbf{yx}}^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{A}(k+1) \\ \quad \cdot(\mathbf{A}(k+1)^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{A}(k+1))^{-1}. \end{cases} \quad (7b)$$

The above algorithm is a more general form of the iterative quadratic minimum distance (IQMD) method shown in [8]. This algorithm can be further generalized into the following AP method.

Batch Version of the AP Method:

$$\begin{cases} \mathbf{A}(k+1) = \mathbf{C}_{\mathbf{yx}}\mathbf{B}(k)\mathbf{G}(k) \\ \mathbf{B}(k+1) = \mathbf{C}_{\mathbf{xx}}^{-1}\mathbf{C}_{\mathbf{yx}}^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{A}(k+1) \\ \quad \cdot(\mathbf{A}(k+1)^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{A}(k+1))^{-1} \end{cases} \quad (8)$$

where $\mathbf{G}(k) \in C^{r\times r}$ is a nonsingular matrix. If $\mathbf{G}(k) = (\mathbf{B}(k)^H\mathbf{C}_{\mathbf{xx}}\mathbf{B}(k))^{-1}$, (8) becomes (7b). It will be shown in Section III-B that with a very wide range of choices of $\mathbf{G}(k)$ and a weak condition on the initial matrix $\mathbf{B}(0)$ and the singular values of $\mathbf{R}_{\rm opt}$, both $\mathbf{A}(k)$ and $\mathbf{B}(k)$ from (8) remain upper bounded for all k, and the product $\mathbf{A}(k)\mathbf{B}(k)^H$ converges to $\hat{\mathbf{T}}_{\rm opt}$ globally and exponentially.

Note that $range(\mathbf{A}) = range(\hat{\mathbf{T}}_{\rm opt})$ and that $range(\mathbf{B}) = range(\hat{\mathbf{T}}_{\rm opt}^H)$. The AP method updates the two subspaces alternately by matrix multiplications. The matrices left multiplied to $\mathbf{A}(k+1)$ and $\mathbf{B}(k)$ update their column spaces. The matrices right multiplied to $\mathbf{A}(k+1)$ and $\mathbf{B}(k)$ serve as "matrix scaling," which ensures that $\mathbf{A}(k)$ and $\mathbf{B}(k)$ are bounded for all k and that their product converges to the desired matrix.

Computationally, the AP method is attractive. Due to the smaller dimensions of $\mathbf{A}(k)$ and $\mathbf{B}(k)$, the AP method can be implemented using only $O(\max(m^2r, n^2r))$ flops at each iteration, where r can be much smaller than $\min(m, n)$ in practice. As shown in Section III-B, the number of iterations required in practice can be very small. Note that the inverses $\mathbf{C}_{\mathbf{xx}}^{-1}$ and $\mathbf{C}_{\mathbf{yy}}^{-1}$ do not need to be computed explicitly. Indeed, $\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{A}(k+1)$ can be obtained by solving the linear equation $\mathbf{C}_{\mathbf{yy}}\mathbf{A}'(k+1) = \mathbf{A}(k+1)$ for $\mathbf{A}'(k+1)$, which requires only $O(n^2r)$ flops [7]. The inverse $\mathbf{C}_{\mathbf{xx}}^{-1}$ should be similarly handled after the product of the matrices on the right side of $\mathbf{C}_{\mathbf{xx}}^{-1}$ is obtained. It can be shown that if the AP method is implemented adaptively with the approach in [10], the number of flops can be reduced further.

B. Global and Exponential Convergence

The following analysis establishes a global and exponential convergence property of (8). Although relatively lengthy, this analysis provides an important insight into the AP method. A different approach [6] can be used to show the fact that the cost function $J_{\rm opt}$ has only one global minimizer, and all but one stationary points of $J_{\rm opt}$ are saddle points. However, this fact is not sufficient to imply (although a good hint) that either (7) or (8) is globally convergent. We will need the following SVD of $\mathbf{R}_{\rm opt}$:

$$\mathbf{R}_{\rm opt} = \mathbf{U}_{\rm opt}\mathbf{\Sigma}_{\rm opt}\mathbf{V}_{\rm opt}^H$$

where $\mathbf{U}_{\rm opt} \in C^{n\times n}$ is the matrix of the left singular vectors; $\mathbf{V}_{\rm opt} \in C^{m\times m}$ is the matrix of the right singular vectors; and $\mathbf{\Sigma}_{\rm opt} = \text{diag}(\,\sigma_1 \quad \sigma_2 \quad \cdots \quad \sigma_{\min(m,n)}\,)$ is the $n \times m$ "diagonal" matrix of the singular values in descending order. It then follows that

$$\hat{\mathbf{T}}_{\rm opt} = \mathbf{C}_{\mathbf{yx}}\mathbf{C}_{\mathbf{xx}}^{-1/2}\mathbf{V}_{\rm opt,1}\mathbf{V}_{\rm opt,1}^H\mathbf{C}_{\mathbf{xx}}^{-1/2}$$

where $\mathbf{V}_{\rm opt,1} = V_1(\mathbf{R}_{\rm opt})$. With the assumption that $\mathbf{C}_{\mathbf{xx}}$ is nonsingular, one can express

$$\mathbf{B}(k) = \mathbf{C}_{\mathbf{xx}}^{-1/2}\mathbf{V}_{\rm opt}\mathbf{P}(k) \quad (9)$$

where $\mathbf{P}(k)$ is $m\times r$. It will be assumed that the top $r\times r$ submatrix of $\mathbf{P}(0)$ is nonsingular. This assumption is clearly satisfied with probability one by a randomly selected $\mathbf{B}(0)$, which means that the convergence proved in the sequel is "almost global." Substituting (9) into (8) yields

$$\mathbf{B}(k+1) = \mathbf{C}_{\mathbf{xx}}^{-1/2}\mathbf{V}_{\rm opt}\mathbf{P}(k+1) \quad (10)$$

where

$$\mathbf{P}(k+1) = \Lambda\mathbf{P}(k)(\mathbf{P}(k)^H\Lambda\mathbf{P}(k))^{-1}\mathbf{G}(k)^{-H} \tag{11}$$

with $\Lambda = \text{diag}\{\lambda_1 \quad \lambda_2 \quad \cdots \quad \lambda_{\min(m,n)}\} = \mathbf{\Sigma}_{\text{opt}}^H\mathbf{\Sigma}_{\text{opt}}$, the diagonal elements of which are in descending order. We assume that $\lambda_r > \lambda_{r+1}$ [which is a weak generic condition that is also required for (4)].[3] Denote the top $r \times r$ submatrix of $\mathbf{P}(k)$ by $\mathbf{P}_1(k)$ and the lower $(m-r)\times r$ submatrix of $\mathbf{P}(k)$ by $\mathbf{P}_2(k)$. Accordingly, let $\Lambda_1 = \text{diag}\{\lambda_1 \quad \lambda_2 \quad \cdots \quad \lambda_r\}$ and $\Lambda_2 = \text{diag}\{\lambda_{r+1} \quad \lambda_{r+2} \quad \cdots \quad \lambda_{\min(m,n)}\}$. It is clear from (11) that $\mathbf{P}_1(k)$ is nonsingular for any finite k. Then, (11) implies that

$$\mathbf{P}_2(k+1)\mathbf{P}_1(k+1)^{-1} = \Lambda_2\mathbf{P}_2(k)\mathbf{P}_1(k)^{-1}\Lambda_1^{-1} \tag{12}$$

and hence

$$\mathbf{P}_2(k+1)\mathbf{P}_1(k+1)^{-1} = \Lambda_2^{k+1}\mathbf{P}_2(0)\mathbf{P}_1(0)^{-1}\Lambda_1^{-(k+1)}. \tag{13}$$

Since $\lambda_r > \lambda_{r+1}$, (13) implies that $\mathbf{P}_2(k)\mathbf{P}_1(k)^{-1}$ converges to zero exponentially, and for large k, it is on the order of $\varepsilon(k) = (\lambda_{r+1}/\lambda_r)^k$. If $\mathbf{G}(k)$ is such that $\mathbf{P}_1(k)$ is upper bounded for all k, then the above means that $\mathbf{P}_2(k)$ converges to zero exponentially. Assuming this property (to be established later), (11) becomes, for large k

$$\mathbf{P}(k+1) = \begin{bmatrix}\mathbf{P}_1(k)^{-H}\mathbf{G}(k)^{-H}\\ \mathbf{0}\end{bmatrix} + O(\varepsilon(k)). \tag{14}$$

From (8)–(10), one has

$$\begin{aligned}&\mathbf{A}(k+1)\mathbf{B}(k+1)^H\\ &\quad = \mathbf{C}_{\mathbf{yx}}\mathbf{C}_{\mathbf{xx}}^{-1/2}\mathbf{V}_{\text{opt}}\mathbf{P}(k)\mathbf{G}(k)\mathbf{P}(k+1)^H\mathbf{V}_{\text{opt}}^H\mathbf{C}_{\mathbf{xx}}^{-1/2}.\end{aligned} \tag{15}$$

Then, (14) and (15) imply that for large k

$$\begin{aligned}&\mathbf{A}(k+1)\mathbf{B}(k+1)^H\\ &\quad = \mathbf{C}_{\mathbf{yx}}\mathbf{C}_{\mathbf{xx}}^{-1/2}\mathbf{V}_{\text{opt}_1}\mathbf{V}_{\text{opt}_1}^H\mathbf{C}_{\mathbf{xx}}^{-1/2} + O(\varepsilon(k))\end{aligned} \tag{16}$$

which means that $\mathbf{A}(k)\mathbf{B}(k)^H$ converges to $\hat{\mathbf{T}}_{\text{opt}}$ exponentially.

Note that when the norm of $\mathbf{P}_1(k)$ is upper bounded for all k, the norm of $\mathbf{B}(k)$ is also upper bounded [from (9)]. If, in addition, the norm of $\mathbf{G}(k)$ is upper bounded, then the norm of $\mathbf{A}(k)$ is also upper bounded [from (8)]. Therefore, the algorithm (8) indeed yields the desired solution provided that a right choice of $\mathbf{G}(k)$ is made.

There are an infinite number of ways of choosing $\mathbf{G}(k)$ to ensure that the norms of $\mathbf{G}(k)$ and $\mathbf{P}_1(k)$ are upper bounded for all k. Let us consider the simplest choice $\mathbf{G}(k) = c\mathbf{I}$ (with a constant norm) , where c is an arbitrary nonzero real number, and $\mathbf{I}$ is the identity matrix. (This is a case more difficult to analyze than that considered in [8].)

With the choice $\mathbf{G}(k) = c\mathbf{I}$, (11) becomes

$$\mathbf{P}(k+1) = \frac{1}{c}\Lambda\mathbf{P}(k)(\mathbf{P}(k)^H\Lambda\mathbf{P}(k))^{-1}. \tag{17}$$

[3]Otherwise, it can be shown that the rank-r minimizer of (2) is not unique and that the method (8) converges to a random solution within the space of all valid solutions.

It will be shown next that the norm of $\Lambda^{1/2}\mathbf{P}(2k)$ is nondecreasing and upper bounded for all k. This implies that the norm of $\mathbf{P}_1(2k)$ is upper bounded for all k. The behavior of the norm of $\Lambda^{1/2}\mathbf{P}(2k+1)$ is similar to the one of $\Lambda^{1/2}\mathbf{P}(2k)$, and hence, its analysis is omitted. A simple iteration of (17) gives

$$\mathbf{P}(2k+2) = \Lambda^2\mathbf{P}(2k)(\mathbf{P}(2k)^H\Lambda^3\mathbf{P}(2k))^{-1}\mathbf{P}(2k)^H\Lambda\mathbf{P}(2k). \tag{18}$$

Let $\mathbf{S}(l) = \Lambda^{1/2}\mathbf{P}(2l)$. Then, (18) becomes

$$\mathbf{S}(l+1) = \Lambda^2\mathbf{S}(l)(\mathbf{S}(l)^H\Lambda^2\mathbf{S}(l))^{-1}\mathbf{S}(l)^H\mathbf{S}(l). \tag{19}$$

It is clear from (19) that $\mathbf{S}(l)^H\mathbf{S}(l+1) = \mathbf{S}(l)^H\mathbf{S}(l)$. This equation, along with the fact that $\|\mathbf{S}(l)^H\mathbf{S}(l)\| = \|\mathbf{S}(l)\|^2$ and $\|\mathbf{S}(l)^H\mathbf{S}(l+1)\| \le \|\mathbf{S}(l)\| \cdot \|\mathbf{S}(l+1)\|$, implies that

$$\|\mathbf{S}(l+1)\| \ge \|\mathbf{S}(l)\| \tag{20}$$

which means that the 2-norm of $\mathbf{S}(l)$ is nondecreasing.

By iterating (19) from $l = 0$, one can verify that

$$\mathbf{S}(l) = \mathbf{H}(l)\mathbf{H}(l-1)\cdots\mathbf{H}(1)\mathbf{S}(0) \tag{21}$$

where

$$\mathbf{H}(l) = \Lambda^{2l}\mathbf{S}(0)(\mathbf{S}(0)^H\Lambda^{4l-2}\mathbf{S}(0))^{-1}\mathbf{S}(0)^H\Lambda^{2l-2}. \tag{22}$$

Applying the same partitions to Λ^{2l} and $\mathbf{S}(l)$ as for Λ and $\mathbf{P}(k)$, (22) becomes

$$\begin{aligned}\mathbf{H}(l) = &\begin{bmatrix}\Lambda_1^{2l}\mathbf{S}_1(0)\\ \Lambda_2^{2l}\mathbf{S}_2(0)\end{bmatrix}(\mathbf{S}_1(0)^H\Lambda_1^{4l-2}\mathbf{S}_1(0)\\ &+\mathbf{S}_2(0)^H\Lambda_2^{4l-2}\mathbf{S}_2(0))^{-1}\\ &\cdot(\mathbf{S}_1(0)^H\Lambda_1^{2l-2} + \mathbf{S}_2(0)^H\Lambda_2^{2l-2}).\end{aligned} \tag{23}$$

Then, applying the matrix identity $(\mathbf{C}+\mathbf{D})^{-1} = \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{D}(\mathbf{I}+\mathbf{C}^{-1}\mathbf{D})^{-1}\mathbf{C}^{-1}$ to the inverse matrix in (23), one can verify that for large l

$$\mathbf{H}(l) = \begin{bmatrix}\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} + O(\varepsilon^2(l)) \tag{24}$$

where $\mathbf{I}_r$ is the $r \times r$ identity matrix. It then follows from (21) that

$$\begin{aligned}\|\mathbf{S}(\infty)\| &\le \|\mathbf{S}(0)\| \prod_{i=1,\cdots,\infty}\|\mathbf{H}(i)\|\\ &= \|\mathbf{S}(0)\| \prod_{i=1,\cdots,\infty}(1+O(\varepsilon^2(i)))\\ &= \|\mathbf{S}(0)\| \exp\left(\sum_{i=1,\cdots,\infty}\log(1+O(\varepsilon^2(i)))\right)\\ &\le \|\mathbf{S}(0)\| \exp\left(\sum_{i=1,\cdots,\infty}O(\varepsilon^2(i))\right)\\ &= \|\mathbf{S}(0)\| \exp\left(O\left(\sum_{i=1,\cdots,\infty}\varepsilon^2(i)\right)\right)\\ &= \|\mathbf{S}(0)\| \exp\left(O\left(\frac{1}{1-(\lambda_{r+1}/\lambda_r)^2}\right)\right).\end{aligned} \tag{25}$$

This completes the proof of the fact that the norm of $\mathbf{P}_1(2k)$ (and, similarly, $\mathbf{P}_1(2k+1)$) is always upper bounded when $\mathbf{G}(k) = c\mathbf{I}$, where c is an arbitrary nonzero real number.

C. Computing the Canonical Components

It has been shown that under a weak condition, the algorithm (8) converges globally and exponentially to $\hat{\mathbf{T}}_{\rm opt}$ with a preselected rank r, which is now denoted by $\hat{\mathbf{T}}_{{\rm opt},r}$. Recall the expression $\hat{\mathbf{T}}_{{\rm opt},r} = \mathbf{C_{yx}}\mathbf{C_{xx}}^{-1/2}\mathbf{V}_{{\rm opt}_1}\mathbf{V}_{{\rm opt}_1}^H\mathbf{C_{xx}}^{-1/2}$. Let $\mathbf{Q}$ be a nonsingular matrix. We define the ith column of $\mathbf{C_{yx}}\mathbf{C_{xx}}^{-1/2}\mathbf{V}_{\rm opt1}\mathbf{Q}$ as the ith *left canonical vector* denoted by $\mathbf{a}_i$ and the ith column of $\mathbf{C_{xx}}^{-1/2}\mathbf{V}_{{\rm opt}_1}\mathbf{Q}^{-H}$ as the ith *right canonical vector* denoted by $\mathbf{b}_i$. We refer to $\mathbf{a}_i\mathbf{b}_i^H$ as the ith *canonical matrix*. It then follows that $\hat{\mathbf{T}}_{{\rm opt},r+1} = \hat{\mathbf{T}}_{{\rm opt},r} + \mathbf{a}_{r+1}\mathbf{b}_{r+1}^H = \sum_{i=1}^{r+1}\mathbf{a}_i\mathbf{b}_i^H$. Although the canonical vectors are obviously not unique, the canonical matrices are unique if the singular values of $\mathbf{R}_{\rm opt} = \mathbf{C_{xx}}^{-1/2}\mathbf{C_{yx}}^H\mathbf{C_{yy}}^{-1/2}$ are distinct. In fact, the ith canonical matrix is unique if and only if the ith singular value of $\mathbf{R}_{\rm opt}$ is distinct from the rest of the singular values.

Given $\mathbf{A}_r = \mathbf{C_{yx}}\mathbf{C_{xx}}^{-1/2}\mathbf{V}_{{\rm opt}_1}\mathbf{Q}$ and $\mathbf{B}_r^H = \mathbf{Q}^{-1}\mathbf{V}_{{\rm opt}_1}^H\mathbf{C_{xx}}^{-1/2}$, which consist of the first r canonical components of $\hat{\mathbf{T}}_{\rm opt}$, an AP algorithm for computing the $(r+1)$th pair of canonical vectors can be derived as follows. Recall the SVD of $\mathbf{R}_{\rm opt}$:

$$\mathbf{R}_{\rm opt} = \mathbf{V}_{{\rm opt}_1}\Sigma_{{\rm opt}_1}^H\mathbf{U}_{{\rm opt}_1}^H + \mathbf{V}_{{\rm opt}_2}\Sigma_{{\rm opt}_2}^H\mathbf{U}_{{\rm opt}_2}^H \tag{26}$$

where the first term on the right is associated with the first r canonical components, and the second term is associated with the rest. It follows that

$$\mathbf{C_{yy}}^{-1/2}\mathbf{A}_r = \mathbf{C_{yy}}^{-1/2}\mathbf{C_{yx}}\mathbf{C_{xx}}^{-1/2}\mathbf{V}_{{\rm opt}_1}\mathbf{Q} = \mathbf{U}_{{\rm opt}_1}\Sigma_{{\rm opt}_1}\mathbf{Q} \tag{27}$$

and hence

$$\begin{aligned}&\mathbf{C_{xx}}^{-1/2}\mathbf{C_{yx}}^H\mathbf{C_{yy}}^{-1/2}(\mathbf{I} - \mathbf{C_{yy}}^{-1/2}\mathbf{A}_r(\mathbf{A}_r^H\mathbf{C_{yy}}^{-1}\mathbf{A}_r)^{-1}\mathbf{A}_r^H\mathbf{C_{yy}}^{-1/2})\\ &\quad = \mathbf{V}_{\rm opt2}\Sigma_{\rm opt2}^H\mathbf{U}_{\rm opt2}^H.\end{aligned} \tag{28}$$

Note that (28) is a deflated version of $\mathbf{R}_{\rm opt}$. To retrieve the $(r+1)$th pair of canonical vectors, we now replace $\mathbf{R}_{\rm opt}$ inherent in (8) by its deflated version, which yields (with $\mathbf{G}(k) = \mathbf{I}$)

$$\begin{cases}\hat{\mathbf{a}}_{r+1}(k+1) = \mathbf{C_{yx}}\hat{\mathbf{b}}_{r+1}(k)\\ \hat{\mathbf{b}}_{r+1}(k+1) = \mathbf{C_{xx}}^{-1}\mathbf{C_{yx}}^H(\mathbf{C_{yy}}^{-1} - \mathbf{C_{yy}}^{-1}\mathbf{A}_r\\ \quad\cdot(\mathbf{A}_r^H\mathbf{C_{yy}}^{-1}\mathbf{A}_r)^{-1}\mathbf{A}_r^H\mathbf{C_{yy}}^{-1})\hat{\mathbf{a}}_{r+1}(k+1)\\ \quad\cdot(\hat{\mathbf{a}}_{r+1}(k+1)^H\mathbf{C_{yy}}^{-1}\hat{\mathbf{a}}_{r+1}(k+1))^{-1}\end{cases} \tag{29}$$

where $\hat{\mathbf{a}}_i(k)$ and $\hat{\mathbf{b}}_i(k)$ denote the estimates of $\mathbf{a}_i$ and $\mathbf{b}_i$ at iteration k, respectively. Following a proof similar to that for (8), one can verify that (29) yields the $(r+1)$th pair of canonical vectors at the rate of $(\lambda_{r+2}/\lambda_{r+1})^k = (\sigma_{r+2}/\sigma_{r+1})^{2k}$ for a random choice of $\mathbf{b}_{r+1}(0)$.

Since (8) yields $\mathbf{A}_r$ and $\mathbf{B}_r$ asymptotically, (29) and (8) can be run at the same time, with $\mathbf{A}_r$ in (29) replaced by $\mathbf{A}(k+1)$. Furthermore, (29) can be run successively for $r = 0, 1, \cdots, r_0 \le \min(m, n)$, for each given k, to retrieve all desired canonical vectors. This algorithm can be easily derived and, hence, only summarized as follows.

Canonical Component Version of the AP Method: At each iteration k, do the following for $r = 0, 1, 2, \cdots, r_0$:

$$\begin{cases}\hat{\mathbf{a}}_{r+1}(k+1) = \mathbf{C_{yx}}\hat{\mathbf{b}}_{r+1}(k)\\ \mathbf{c}_{r+1}(k+1) = \mathbf{C_{yy}}^{-1}\hat{\mathbf{a}}_{r+1}(k+1)\\ \mathbf{d}_{r+1}(k+1) = \mathbf{A}_r(k+1)^H\mathbf{c}_{r+1}(k+1)\\ \alpha_{r+1}(k+1) = \hat{\mathbf{a}}_{r+1}(k+1)^H\mathbf{c}_{r+1}(k+1)\\ \hat{\mathbf{b}}_{r+1}(k+1) = \mathbf{C_{xx}}^{-1}\mathbf{C_{yx}}^H\\ \quad\cdot(\mathbf{c}_{r+1}(k+1) - \mathbf{C_{yy}}^{-1}\mathbf{A}_r(k+1)\mathbf{S}_r(k+1)^{-1}\mathbf{d}_{r+1}(k+1))\\ \quad\cdot\dfrac{1}{\alpha_{r+1}(k+1)}\\ \mathbf{A}_{r+1}(k+1) = [\,\mathbf{A}_r(k+1)\quad\hat{\mathbf{a}}_{r+1}(k+1)\,]\\ \mathbf{S}_{r+1}(k+1) = \begin{bmatrix}\mathbf{S}_r(k+1) & \mathbf{d}_{r+1}(k+1)\\ \mathbf{d}_{r+1}(k+1)^H & \alpha_{r+1}(k+1)\end{bmatrix}\end{cases} \tag{30}$$

where $\mathbf{A}_0(k+1)$, $\mathbf{S}_0(k+1)$, and $\mathbf{d}_1(k+1)$ must be ignored. Efficient programming of the above algorithm requires some care. The operations should be carried out "from right to left," i.e., scalar-vector multiplication first and the vector-matrix multiplication second. Any matrix-matrix multiplication can and should be avoided. The standard partitioned-matrix inversion lemma (e.g., see [9]) should be applied to compute the inverse of $\mathbf{S}_{r+1}(k+1)$ recursively with respect to r, namely, as in (31), shown at the bottom of the page, where

$$\begin{aligned}&\mathbf{e}_{r+1}(k+1) = \mathbf{S}_r(k+1)^{-1}\mathbf{d}_{r+1}(k+1)\\ &\mu_r(k+1) = \alpha_{r+1}(k+1) - \mathbf{d}_{r+1}(k+1)^H\mathbf{e}_{r+1}(k+1).\end{aligned} \tag{32}$$

D. Rank Detection with the AP Method

In the case where the rank of the matrix $\mathbf{T}$ in (1) is unknown, the canonical components can be used to detect it. In a high SNR environment, the rank of $\mathbf{T}$ can be chosen to be the first r for which $\|\hat{\mathbf{T}}_{{\rm opt},r}\| \gg \|\sum_{i>r}\hat{\mathbf{a}}_i\hat{\mathbf{b}}_i^H\|$. In general, however, it may be difficult to attach a precise meaning to the condition "$\gg$," and the (simplified) generalized likelihood ratio test

$$\mathbf{S}_{r+1}(k+1)^{-1} = \begin{bmatrix}\mathbf{S}_r(k+1)^{-1} + \mathbf{e}_{r+1}(k+1)\mathbf{e}_{r+1}(k+1)^H/\mu_r(k+1) & -\mathbf{e}_{r+1}(k+1)/\mu_r(k+1)\\ -\mathbf{e}_{r+1}(k+1)^H/\mu_r(k+1) & 1/\mu_r(k+1)\end{bmatrix} \tag{31}$$

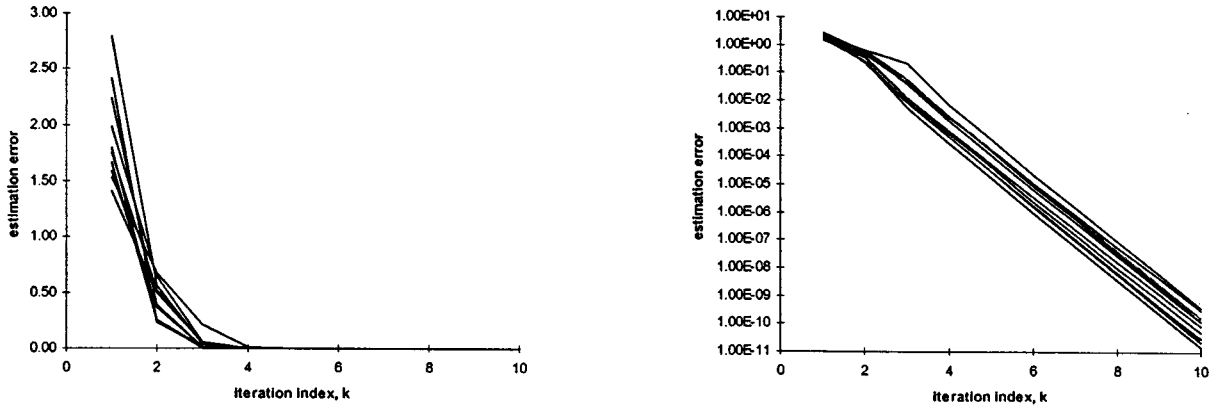

Fig. 1. Estimation errors of the batch version of the AP method with ten independent initializations. (a) Linear scale (b) Logarithmic scale.

(GLRT) developed in [1] can be used. The GLRT requires the availability of the singular values of $R_{\rm opt}$. The singular values can be easily computed once the canonical vectors are available. This is explained below. Upon convergence, the estimated canonical vectors $\hat{\mathbf{a}}_i$ and $\hat{\mathbf{b}}_i$ yield the unique canonical matrix $\hat{\mathbf{a}}_i\hat{\mathbf{b}}_i^H = \mathbf{a}_i\mathbf{b}_i^H$, and one can write $\hat{\mathbf{a}}_i(k) = \gamma(k)\mathbf{C}_{\mathbf{yx}}\mathbf{C}_{\mathbf{xx}}^{-1/2}\mathbf{v}_i$ and $\hat{\mathbf{b}}_i(k) = (1/\gamma(k)^H)\mathbf{C}_{\mathbf{xx}}^{-1/2}\mathbf{v}_i$, where $\gamma(k)$ is a complex scalar,[4] and $\mathbf{v}_i$ is the ith right singular vector of $\mathbf{R}_{\rm opt}$. It is easy to verify that $\hat{\mathbf{a}}_i^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{C}_{\mathbf{yx}}\hat{\mathbf{b}}_i = \sigma_i^2$. Therefore, the singular values can be updated as

$$\hat{\sigma}_i(k)^2 = \hat{\mathbf{a}}_i(k)^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{C}_{\mathbf{yx}}\hat{\mathbf{b}}_i(k). \tag{33}$$

The GLRT developed in [1] is summarized here for convenience. Define the test statistics

$$\varsigma_{\hat{r}}(k) = -N\sum_{i=\hat{r}+1}^{\min(n,m)}\ln(1-\hat{\sigma}_i(k)^2) \tag{34}$$

where N is the length of data (or the effective window length in on-line applications). Define the threshold $\beta_\alpha(\hat{r})$ to be such that the following condition is met:

$$prob\{\omega \geq \beta_\alpha(\hat{r})\} = \alpha \tag{35}$$

where α is a small positive number (much smaller than 1),[5] and ω is a chi-square distributed random variable with $(m-\hat{r})(n-\hat{r})$ degrees of freedom. Then, the rank at time k should be the first $\hat{r}$(starting from 1) that satisfies

$$\varsigma_{\hat{r}}(k) \leq \beta_\alpha(\hat{r}). \tag{36}$$

E. Adaptive Computation

Because of its recursive nature, the AP method can be easily applied to track the time-variations of a rank-reduced matrix $\mathbf{T}$ in (1). The idea is simply to allow the correlation matrices $\mathbf{C}_{\mathbf{xx}}$, $\mathbf{C}_{\mathbf{yy}}$, and $\mathbf{C}_{\mathbf{yx}}$ to be updated as new data become available during the iteration of the AP method. The inverses of the autocorrelation matrices can be efficiently obtained at each iteration by using the standard rank-one inverse update. For example, if $\mathbf{C}_{\mathbf{xx}}(k+1) = \delta\mathbf{C}_{\mathbf{xx}}(k) + \mathbf{x}(k+1)\mathbf{x}(k+1)^H$, where δ is a forgetting factor between (0, 1), then

$$\mathbf{C}_{\mathbf{xx}}(k+1)^{-1} = \frac{1}{\delta}\,\mathbf{C}_{\mathbf{xx}}(k)^{-1} - \mathbf{g}(k+1)\mathbf{g}(k+1)^H/\beta(k+1) \tag{37}$$

where $\mathbf{g}(k+1) = (1/\delta)\mathbf{C}_{\mathbf{xx}}(k)^{-1}\mathbf{x}(k+1)$, and $\beta(k+1) = \mathbf{x}(k+1)^H\mathbf{g}(k+1)+1$. With the updated correlation matrices and their inverses, one can easily update the canonical components by (30), the singular values by (33), and, hence, the rank of $\mathbf{T}$ by (34)–(36). In the context of on-line applications, the data length N shown in (34) should be replaced by the effective (asymptotical) window length $1/(1-\delta)$. If the rank of $\mathbf{T}$ increases or decreases by no more than one within an effective window, only the first $r+1$ pairs of the canonical vectors (as opposed to all canonical vectors) need to be tracked at any given time without losing track of the rank (here, r is the current estimate of the rank).

IV. Simulation Examples

To illustrate the performance of the AP method in the context of RRMLE of the matrix $\mathbf{T}$ in (1), $m = 10$ and $n = 20$ are chosen (i.e., $\mathbf{T}$ is 10×20), and $\mathbf{T}$ is constructed as follows:

$$\mathbf{T} = \sum_{i=1}^{r}\tilde{\mathbf{a}}_i\tilde{\mathbf{b}}_i^H \tag{38}$$

where $\tilde{\mathbf{a}}_i$ and $\tilde{\mathbf{b}}_i$ are randomly selected. The rank of this matrix $\mathbf{T}$ is r. Each element of the input signal $\mathbf{x}(k)$ is independently selected from $N(0, \sigma_x^2\mathbf{I})$, and each element of the noise $\mathbf{e}(k)$ is independently selected from $N(0, \sigma_e^2\mathbf{I})$. The SNR is defined as SNR $= 10\log_{10}E(\|\mathbf{T}\mathbf{x}(k)\|^2)/E(\|\mathbf{e}(k)\|^2) = 10\log_{10}\sigma_{\mathbf{x}}^2{\rm tr}(\mathbf{T}\mathbf{T}^H)/m\sigma_{\mathbf{e}}^2$. The SNR is 10 dB for all the cases shown below. For a wide range of SNR, a behavior similar to what is shown next has been observed, and hence, the cases for other SNRs will not be shown.

[4]It is easy to show that $\gamma(k+1) = 1/\gamma(k)^H$.

[5]$\alpha = 0.05$ is chosen in the simulations.

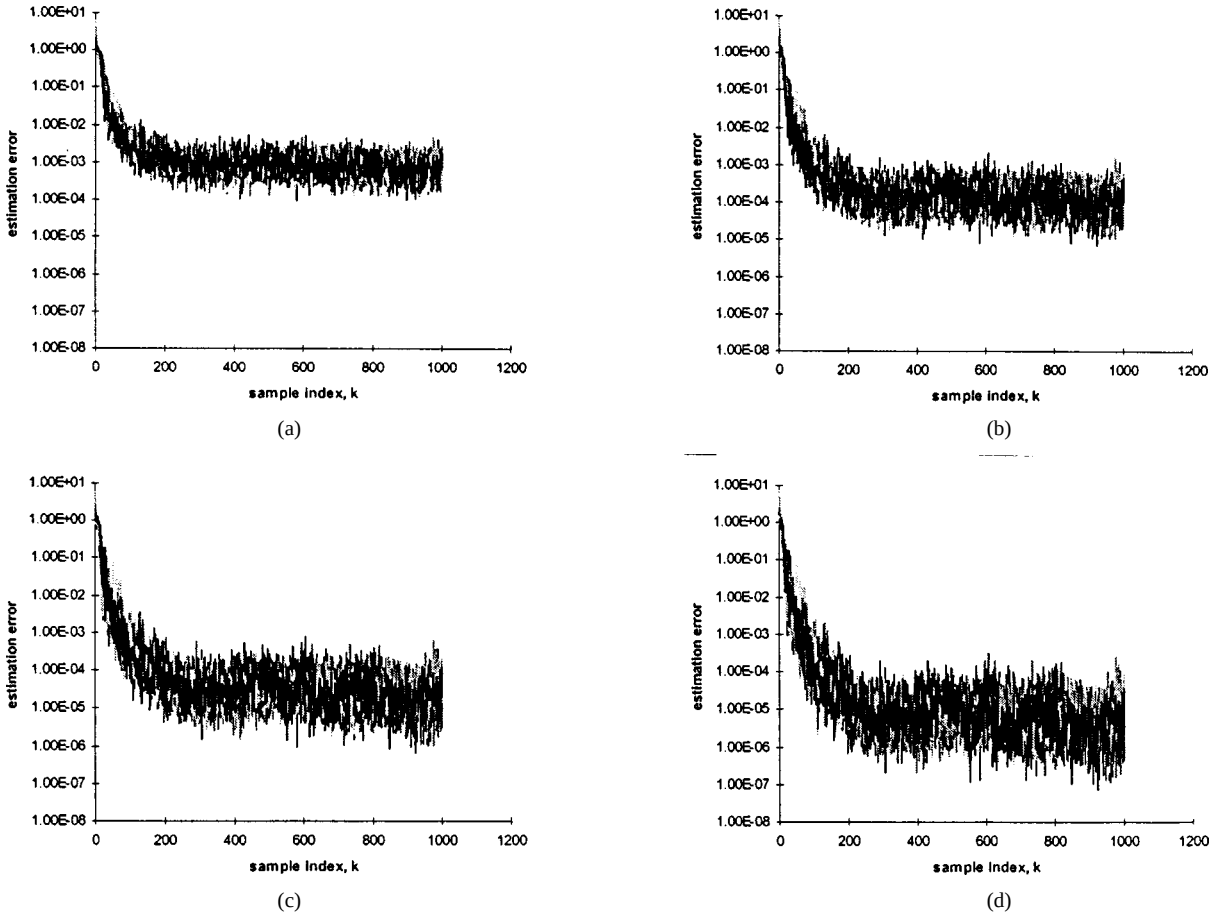

Fig. 2. Estimation errors of the adaptive batch version of the AP method with a varied number of iterations per sample index: (a) One iteration. (b) Two iterations. (c) Three iterations. (d) Four iterations.

As before, we denote the ith pair of the ideal canonical vectors by $\mathbf{a}_i$ and $\mathbf{b}_i$ and the corresponding estimated vectors by $\hat{\mathbf{a}}_i$ and $\hat{\mathbf{b}}_i$. The estimation error of the ith canonical matrix is defined as

$$j_i(k) = \frac{||\hat{\mathbf{a}}_i(k)\hat{\mathbf{b}}_i(k)^H - \mathbf{a}_i\mathbf{b}_i^H||}{||\mathbf{a}_i\mathbf{b}_i^H||} \tag{39}$$

which is also referred to as the "ith component error." We also define the rank-r group of the ideal left canonical vectors as $\mathbf{A}_r = [\mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_r]$ and the rank-r group of the ideal right canonical components as $\mathbf{B}_r = [\mathbf{b}_1 \quad \mathbf{b}_2 \quad \cdots \quad \mathbf{b}_r]$. The groups of the estimated canonical vectors are similarly defined, i.e., $\hat{\mathbf{A}}_r = [\hat{\mathbf{a}}_1 \quad \hat{\mathbf{a}}_2 \quad \cdots \quad \hat{\mathbf{a}}_r]$, and $\hat{\mathbf{B}}_r = [\hat{\mathbf{b}}_1 \quad \hat{\mathbf{b}}_2 \quad \cdots \quad \hat{\mathbf{b}}_r]$. The following is referred to as the "rank-r group error":

$$J_r(k) = \frac{||\hat{\mathbf{A}}_r(k)\hat{\mathbf{B}}_r(k)^H - \mathbf{A}_r\mathbf{B}_r^H||}{||\mathbf{A}_r\mathbf{B}_r^H||}. \tag{40}$$

Case 1—Batch AP Method: The rank of $\mathbf{T}$ is fixed at $r = 4$. The correlation matrices are computed from $N = 500$ samples and kept constant during the iteration of the AP method (8).[6] Fig. 1(a) shows the rank-4 group error versus the iteration index k, where ten independent initializations are used. This figure suggests that after the fourth iteration, the error is very small (not visible). Fig. 1(b) shows the logarithm-scale version of Fig. 1(a). The straight lines in this figure are consistent with the theoretical result that the AP method is exponentially convergent.

Case 2—Adaptive Batch AP Method: The assumptions used here are the same as for Case 1, except that the correlation matrices are updated using a new sample pair $\mathbf{x}(k)$ and $\mathbf{y}(k)$ after one or more iterations of the AP method. The forgetting factor used for updating the correlation matrices in this case (and other cases shown later) is 0.99. Fig. 2 (a)–(d) shows the rank-4 group errors versus the iteration index for ten independent initializations (and ten independent runs), where the number of iterations for each new sample pair is 1, 2, 3, and 4, respectively. It is clear that one can control the accuracy of the adaptive AP method by choosing the number of iterations: the more iterations, the more accurate the method.

Case 3: the adaptive canonical component AP method. The rank of the matrix $\mathbf{T}$ is varied after every 500 samples. The sequence of ranks is 2, 3, 4, and 3. The canonical component AP method (30) is used where the correlation matrices are updated at each iteration (i.e., one iteration for each new sample).

Fig. 3(a)–(f) shows the errors of the estimated components and the errors of the corresponding estimated singular values as functions of the sample index. It can be observed from Fig. 3(b)–(c) that the errors of the second and third components are relatively large at the sample indices around 1500. This is because the second and third singular values are very close to each other at those sample indices (see Fig. 5). It is clear from Fig. 3(d)–(f) that

[6]The gain matrix $\mathbf{G}(k)$ is the identity for all cases.

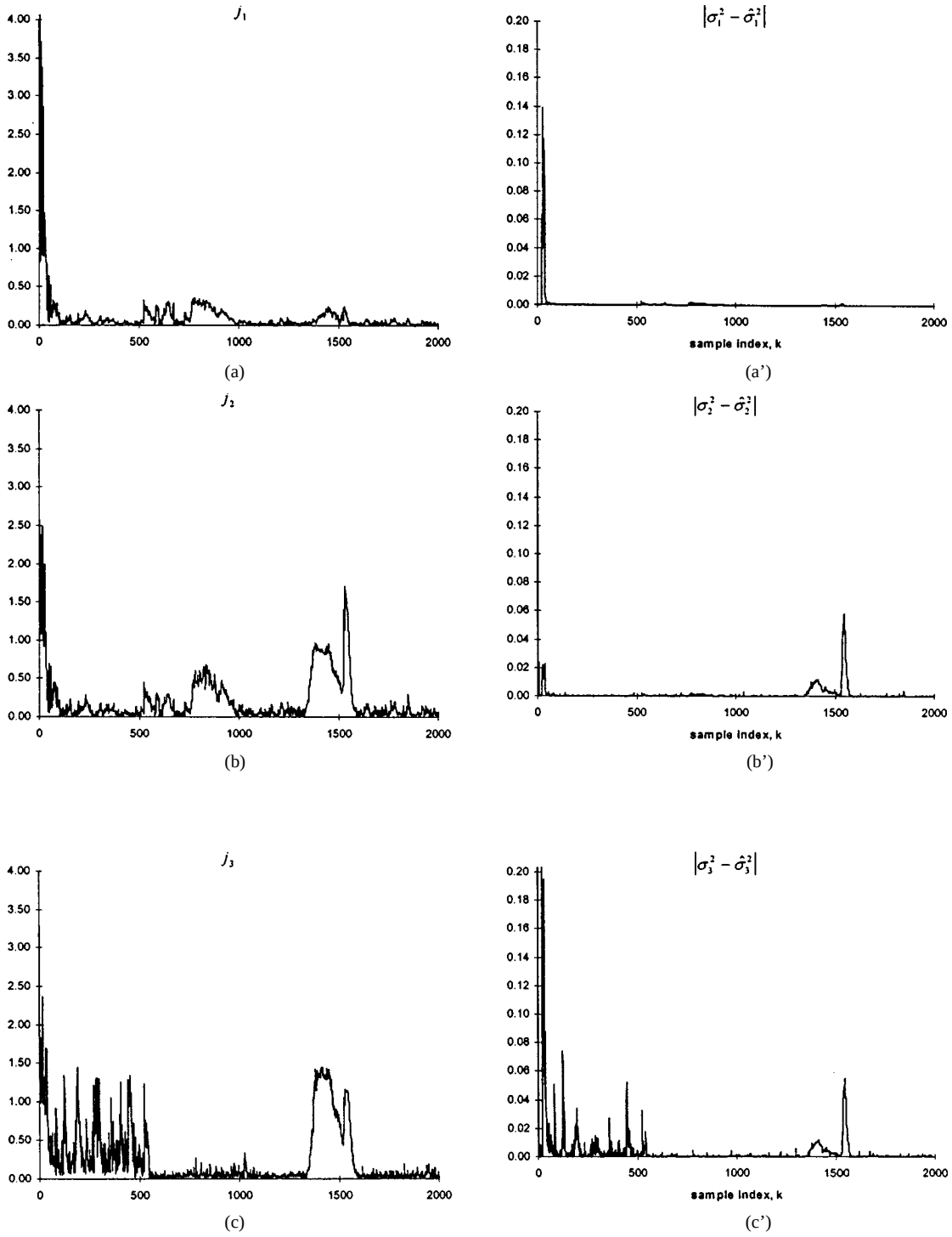

Fig. 3. Component errors: (a)–(a') First-order. (b)–(b') Second-order. (c)-(c') Third-order component errors.

the errors corresponding to the noise (nondominant) components are always large, which is consistent with the fact that the "noise singular values" are close to each other.

Fig. 4(a)–(f) shows the group errors of different ranks as functions of the sample index. Comparing Fig. 3(b) with Fig. 4(b), for example, suggests that errors of the estimated singular values are related to the errors of the corresponding group errors than not more corresponding (vector) component errors. This is a rather surprising phenomenon. However, an explanation is given in Appendix B.

Fig. 5 shows the ten estimated (nonzero) singular values as functions of the iteration (sample) index. A few critical singular values are marked in the figure. Fig. 6 shows the detected ranks based on the estimated singular values using (36). All ranks are correctly detected after some delays, as expected. The delays are somewhere between 100 and 120 samples. Clearly, these delays largely depend on the choice of the forgetting factor used in updating the correlation matrices. Indeed, the delays are about the same as the effective window length (which is 100) corresponding to the forgetting factor used (which is 0.99).

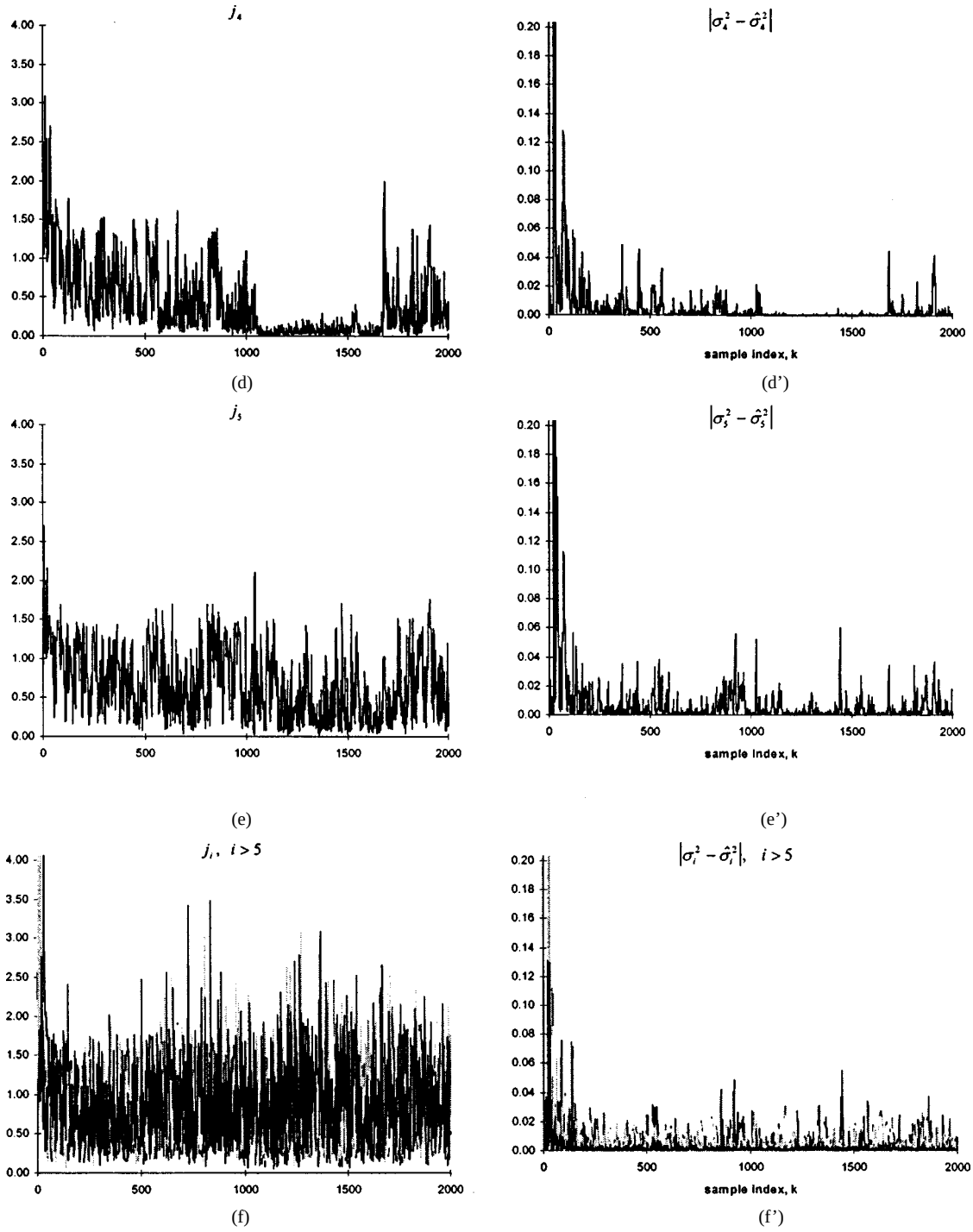

Fig. 3. *(Continued.)* Component errors. (d) (d') Fourth-order. (e) (e') Fifth-order. (f) (f') Sixth and higher order.

V. Conclusions

A class of optimal reduced rank estimators and filters have been reviewed in a unified way, and a further insight has been presented. The alternating power (AP) method developed in this paper is an efficient way for computing the rank-reduced estimators and filters when the rank reduction is relatively large. It is particularly useful for on-line adaptive applications. The global and exponential convergence property of the AP method is an important feature. The canonical components obtained by the AP method make the adaptive rank detection an easy task. A useful specialization of the AP method is shown in Appendix A.

Appendix A
Specialization of the AP Method

This section shows a specialized version of the AP method (8) for computing the optimum rank-r reduction of an arbitrary $m \times n$ matrix $\mathbf{X}$ of rank no less than r. Let the SVD of $\mathbf{X}$ be expressed as $\mathbf{X} = \mathbf{U}_1\Sigma_1\mathbf{V}_1^H + \mathbf{U}_2\Sigma_2\mathbf{V}_2^H$, where the first term contains the r principal components of the SVD. It is known [7], [17] that the optimum rank-r reduction (in F-norm) of $\mathbf{X}$ is given by $\mathbf{X}_r = \mathbf{U}_1\Sigma_1\mathbf{V}_1^H$. By setting $\mathbf{C_{yx}} = \mathbf{X}$ and $\mathbf{C_{xx}} = \mathbf{C_{yy}} = \mathbf{I}$ in the AP method (8), onecan construct the following algorithm for computing $\mathbf{X}_r$ (without computing the

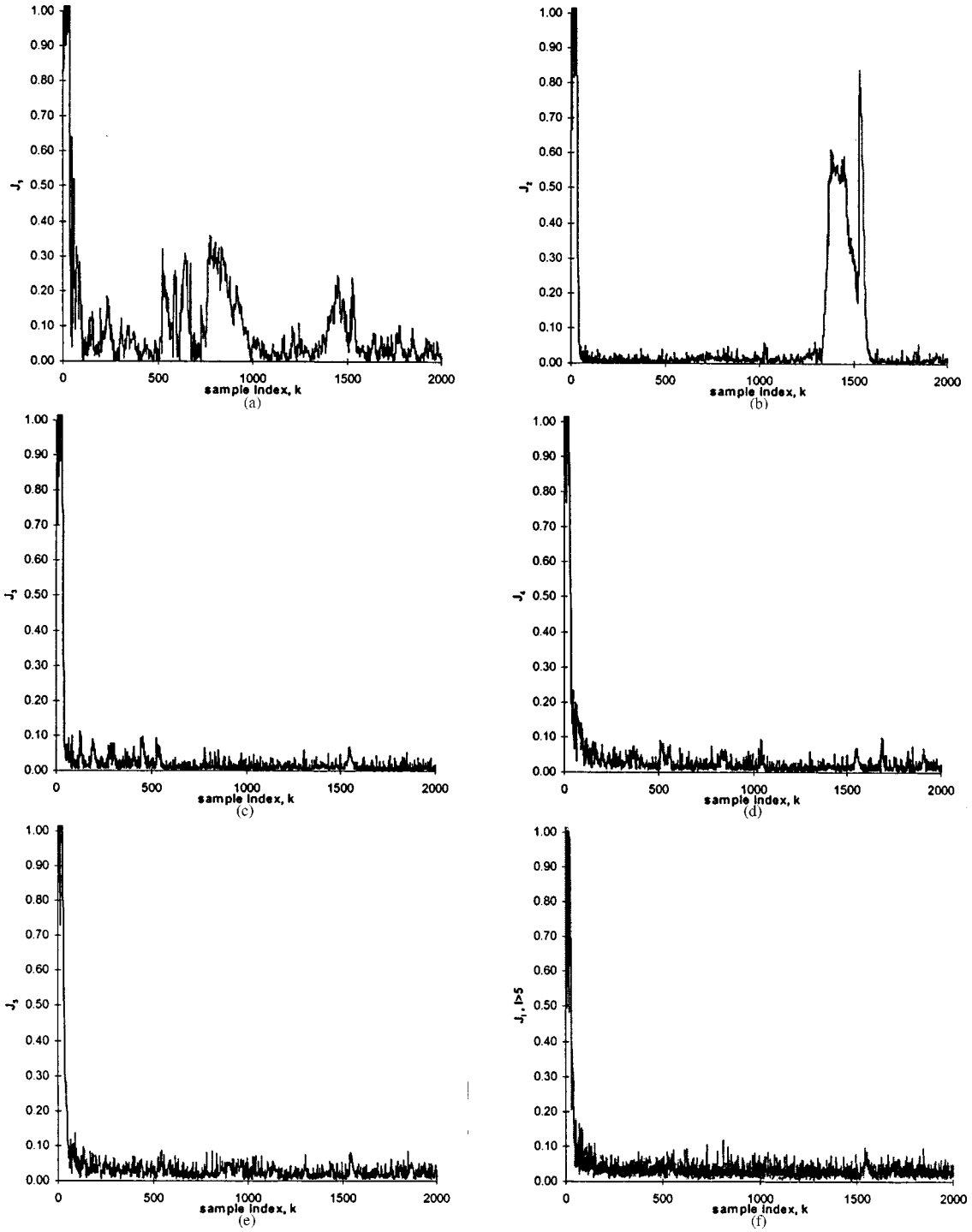

Fig. 4. Group errors: (a) Rank-1. (b) Rank-2. (c) Rank-3. (d) Rank-4. (e) Rank-5. (f) Rank-6 (and higher).

SVD explicitly):

$$\begin{cases} \mathbf{A}(k+1) = \mathbf{X}\mathbf{B}(k)\mathbf{G}(k) \\ \mathbf{B}(k+1) = \mathbf{X}^H\mathbf{A}(k+1)(\mathbf{A}(k+1)^H\mathbf{A}(k+1))^{-1} \end{cases} \tag{A.1}$$

where $\mathbf{A}$ and $\mathbf{B}$ are r-column matrices. Under a condition similar to that of (8), the product $\mathbf{A}(k+1)\mathbf{B}(k+1)^H$ from (A.1) globally and exponentially converges to $\mathbf{X}_r$. Provided that the desired rank is given and much smaller than $\min(m, n)$, (A.1) is an efficient algorithm for computing $\mathbf{X}_r$ as it requires $O(mnr)$ flops at each iteration. This algorithm can also be adopted for on-line applications where the matrix $\mathbf{X}$ varies.

APPENDIX B
GROUP ERROR VERSUS COMPONENT ERROR

This section shows that the accuracy of the estimated singular values based on (33) is at least as good[7] as the accuracy of a

[7]In terms of the order of errors.

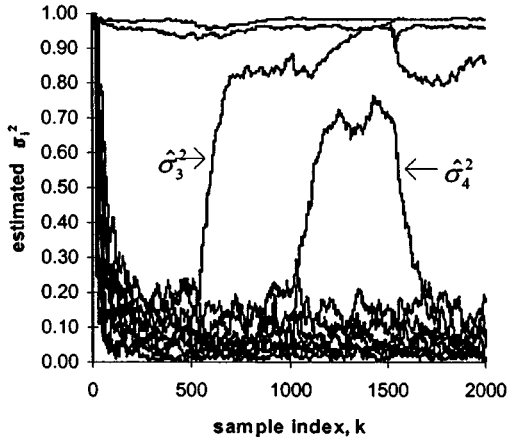

Fig. 5. Estimated singular values of the characteristic matrix by the adaptive component version of the AP method.

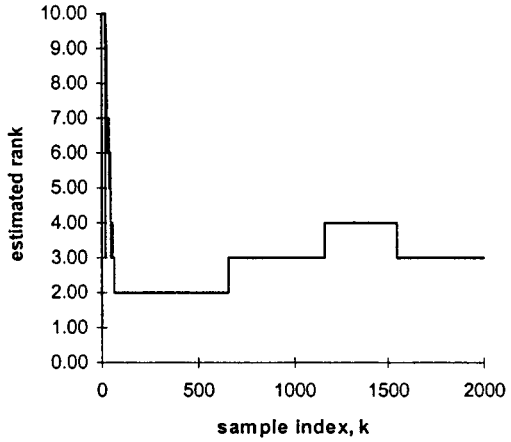

Fig. 6. Estimated rank values.

group of estimated canonical components, regardless of the accuracy of individual estimated canonical components. This is somehow counterintuitive and, hence, surprising at the first sight.

The component errors and the group errors are defined in (39) and (40), respectively. The convergence rate of the ith (with $i < r$) component error is asymptotically[8] governed by $(\sigma_{i+1}/\sigma_i)^{2k}$, which can be very slow if the noise level is low [see (C.3)]. On the other hand, it can be shown that the rank-r group error converges to zero at a rate asymptotically governed by $(\sigma_{r+1}/\sigma_r)^{2k}$, which is fast, provided that the noise level is not too high and the matrix $\mathbf{T}$ has a "well defined" rank r.

Assuming that the rank-r group error is in the order $O(\varepsilon)$ where ε is a small number dependent on the data length and the noise level, we can write

$$\hat{\mathbf{A}}_r\hat{\mathbf{B}}_r^H = \mathbf{A}_r\mathbf{B}_r^H + O(\varepsilon). \tag{B.1}$$

It then follows that

$$\begin{aligned}\hat{\mathbf{A}}_r &= \mathbf{A}_r\mathbf{Q}_a + O(\varepsilon)\\ \hat{\mathbf{B}}_r &= \mathbf{B}_r\mathbf{Q}_b + O(\varepsilon)\end{aligned} \tag{B.2}$$

where $\mathbf{Q}_a\mathbf{Q}_b^H = \mathbf{I}$, and hence, $\mathbf{Q}_b^H\mathbf{Q}_a = \mathbf{Q}_a^H\mathbf{Q}_b = \mathbf{I}$.

[8]Assuming that all the correlation matrices are kept constant. This assumption is made throughout this section.

According to (33), the estimated singular values are given by the diagonal elements of

$$\hat{\Lambda} = \hat{\mathbf{A}}_r^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{C}_{\mathbf{yx}}\hat{\mathbf{B}}_r \tag{B.3}$$

where the correlation matrices are assumed to be constant. Using (B.2) in (B.3) and some simple manipulation (recall $\mathbf{A}_r = \mathbf{C}_{\mathbf{yx}}\mathbf{C}_{\mathbf{xx}}^{-1/2}\mathbf{V}_{\mathrm{opt}_1}$ and $\mathbf{B}_r^H = \mathbf{V}_{\mathrm{opt}_1}^H\mathbf{C}_{\mathbf{xx}}^{-1/2}$) yields

$$\begin{aligned}\hat{\Lambda} &= \mathbf{Q}_a^H\mathbf{A}_r^H\mathbf{C}_{\mathbf{yy}}^{-1}\mathbf{C}_{\mathbf{yx}}\mathbf{B}_r\mathbf{Q}_b + O(\varepsilon)\\ &= \mathbf{Q}_a^H\Lambda\mathbf{Q}_b + O(\varepsilon)\end{aligned} \tag{B.4}$$

where $\Lambda = \mathrm{diag}(\,\sigma_1^2 \quad \sigma_2^2 \quad \cdots \quad \sigma_r^2\,)$. Applying the property $\mathbf{Q}_a^H\mathbf{Q}_b = \mathbf{I}$, one has $\hat{\sigma}_i^2 = \sigma_i^2 + O(\varepsilon)$. We have now shown that the accuracy of the first r estimated singular values is at least as good as that of the group (or span) of the first r pairs of estimated canonical components (matrices) despite possible large errors of individual canonical components (matrices).

Appendix C
Singular Values of $\mathbf{R}_{\mathrm{opt}}$

This section provides an explicit expression of the singular values of $\mathbf{R}_{\mathrm{opt}}$ for the multivariate linear regression model where the noise is spatially white, i.e., $\mathbf{C}_{\mathbf{ee}} = \sigma_{\mathbf{e}}^2\mathbf{I}$, where $\sigma_{\mathbf{e}}^2$ is the noise variance. With this assumption, we have

$$\begin{aligned}\mathbf{R}_{\mathrm{opt}} &= (\mathbf{T}\mathbf{C}_{\mathbf{xx}}\mathbf{T}^H + \sigma_{\mathbf{e}}^2\mathbf{I})^{-1/2}\mathbf{T}\mathbf{C}_{\mathbf{xx}}\mathbf{C}_{\mathbf{xx}}^{-1/2}\\ &= (\mathbf{T}\mathbf{C}_{\mathbf{xx}}\mathbf{T}^H + \sigma_{\mathbf{e}}^2\mathbf{I})^{-1/2}\mathbf{T}\mathbf{C}_{\mathbf{xx}}^{1/2}.\end{aligned} \tag{C.1}$$

Let the singular value decomposition of $\mathbf{T}\mathbf{C}_{\mathbf{xx}}^{1/2}$ be $\mathbf{E}\mathbf{S}\mathbf{F}^H$, where $\mathbf{S} = \mathrm{diag}(\,s_1 \quad s_2 \quad \cdots \quad s_{\min(m,\,n)}\,)$, and the diagonal elements are in descending order. It then follows that

$$\mathbf{R}_{\mathrm{opt}} = \mathbf{E}(\mathbf{S}^2 + \sigma_{\mathbf{e}}^2\mathbf{I})^{1/2}\mathbf{S}\mathbf{F}^H \tag{C.2}$$

and hence, the singular values of $\mathbf{R}_{\mathrm{opt}}$ are

$$\sigma_i = \frac{s_i}{\sqrt{s_i^2 + \sigma_{\mathbf{e}}^2}}. \tag{C.3}$$

As expected, all the singular values are between (0, 1). A useful observation from (C.3) is that when the noise is much weaker than the signal [associated with $\mathbf{x}(k)$], all the dominant singular values are close to one, and as the noise level increases, all the singular values are pulling toward zero. Note that although each of the first $r-1$ estimated canonical matrices is not reliable when SNR is very high, it does not affect (as much) any of the corresponding estimated singular values or $\hat{\mathbf{T}}_{\mathrm{opt}}$ of rank r [provided that r is the correct rank of the model (1)].

Acknowledgment

During the course of writing this paper, the first author had discussions with L. Scharf on several occasions. The generous time and deep insight shared by him contributed to the current form of this paper. The first author also acknowledges a discussion with M. Viberg at ICASSP'99, which attracted their attention to the work by Brillinger.

References

[1] P. Stoica and M. Viberg, "Maximum likelihood parameter and rank estimation in reduced-rank multivariate linear regressions," *IEEE Trans. Signal Processing*, vol. 44, pp. 3069–3078, Dec. 1996.

[2] A. J. Paulraj and C. B. Papadias, "Space-time processing for wireless communications," *IEEE Signal Processing Mag.*, vol. 14, pp. 49–83, Nov. 1997.
[3] L. L. Scharf, *Statistical Signal Processing—Detection, Estimation, and Time Series Analysis*. Reading, MA: Addison-Wesley, 1991, p. 330.
[4] Y. Yamashita and H. Ogawa, "Relative Karhunen–Loeve transform," *IEEE Trans. Signal Processing*, vol. 44, pp. 371–378, Feb. 1996.
[5] Y. Hua and W. Liu, "Generalized Karhunen–Loeve transform," *IEEE Signal Processing Lett.*, vol. 5, pp. 141–142, June 1998.
[6] K. I. Diamantara and S. Y. Kung, "Multilayer neural networks for reduced-rank approximation," *IEEE Trans. Neural Networks*, vol. 5, pp. 684–697, Sept. 1994.
[7] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1983.
[8] Y. Hua and M. Nikpour, "Computing the reduced rank wiener filter by IQMD," *IEEE Signal Processing Lett.*, vol. 6, pp. 240–242, Sept. 1999.
[9] D. Zwillinger, Ed., *Standard Mathematical Tables and Formulae*. Boca Raton, FL: CRC, 1996.
[10] Y. Hua, Y. Xiang, T. Chen, K. Abed-Meraim, and Y. Miao, "A new look at the power method for fast subspace tracking," *Digital Signal Process.*, vol. 9, no. 4, pp. 297–314, Oct. 1999.
[11] L. L. Scharf, "The SVD and reduced rank signal processing," *Signal Process.*, vol. 25, pp. 113–133, 1991.
[12] L. L. Scharf and J. K. Thomas, "Wiener filter in canonical coordinates for transform coding, filtering, and quantizing," *IEEE Trans. Signal Processing*, vol. 46, pp. 647–654, Mar. 1998.
[13] L. L. Scharf and C. T. Mullis, "Canonical coordinates and the geometry of inference, rate, and capacity," *IEEE Trans. Signal Processing*, vol. 48, pp. 824–831, Mar. 2000.
[14] Y. Miao and Y. Hua, "Fast subspace tracking and neural network learning by a novel information criterion," *IEEE Trans. Signal Processing*, vol. 46, pp. 1967–1979, July 1998.
[15] B. Yang, "Projection approximation subspace tracking," *IEEE Trans. Signal Processing*, vol. 43, pp. 95–107, Jan. 1995.
[16] D. W. Tufts and C. D. Melissinos, "Simple, effective computation of principal eigenvectors and their eigenvalues and application to high-resolution estimation of frequencies," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1046–1053, Oct. 1986.
[17] D. R. Brillinger, *Time Series: Data Analysis and Theory*. New York: Holt, Rinehart and Winston, 1975.
[18] M. C. Vanderveen, C. B. Papadias, and A. Paulraj, "Joint angle and delay estimation (JADE) for multipath signals arriving at an antenna array," *IEEE Commun. Lett.*, vol. 1, pp. 12–14, Jan. 1997.
[19] Y. Hua, "Fast maximum likelihood for blind identification of multiple FIR channels," *IEEE Trans. Signal Processing*, vol. 44, pp. 661–672, Mar. 1996.
[20] K. Abed-Meraim, W. Qiu, and Y. Hua, "Blind system identification," *Proc. IEEE*, vol. 85, pp. 1310–1322, Aug. 1997.
[21] L. Tong and S. Perreau, "Multichannel blind identification: From subspace to maximum likelihood methods," *Proc. IEEE*, vol. 86, pp. 1951–1968, Oct. 1998.
[22] A. J. Izenman, "Reduced-rank regression for the multivariate linear model," *J. Multivariate Anal.*, vol. 5, pp. 248–264, 1975.
[23] E. C. Real, D. W. Tufts, and J. W. Cooley, "Two algorithms for fast approximation subspace tracking," *IEEE Trans. Signal Processing*, vol. 47, p. 1936, July 1999.

Yingbo Hua (SM'92) was born in China in 1960. He received the B.E. degree from Nanjing Institute of Technology, Nanjing, China, in 1982 and the M.S. and Ph.D. degrees from Syracuse University, Syracuse, NY, in 1983 and 1988, respectively.

He worked as Teaching Assistant, Research Assistant, Summer Lecturer, and Research Fellow at Syracuse. In July 1989, he accepted an offer to join the University of Melbourne, Australia, where he was promoted from Lecturer to Senior Lecturer in 1992 and to Associate Professor and Reader in 1995. He was on Visiting Faculty with the Hong Kong University of Science and Technology from 1999 to 2000. He is now a Professor of Electrical Engineering with the University of California, Riverside. He is an author/co-author of over 70 journal articles, five book chapters, and 120 conference papers in the fundamental areas of estimation, detection, system identification, and fast algorithms, with applications in communications, remote sensing, and medical data analysis. He is a coeditor of the book *Signal Processing Advances in Wireless Communications* (Englewood Cliffs, NJ: Prentice-Hall, 2000).

Dr. Hua received a Chinese Government Scholarship for Overseas Study from 1983 to 1984 and Syracuse University Graduate Fellowship from 1985 to 1986. He served as Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 1994 to 1997 and curently serves as Associate Editor of the IEEE SIGNAL PROCESSING LETTERS. He is an Invited Reviewer for over 16 international journals. He is an Invited Speaker, Session Chair, and Organizer of many international conferences. He has been an Elected Member of the IEEE Signal Processing Society's Technical Committee for Sensor Array and Multichannel Signal Processing since 1998 and served on the committee for Underwater Acoustic Signal Processing from 1997 to 1998.

Maziar Nikpour was born in 1975. He received the B.E. degree (with honors) in electrical engineering and the B.Sc. degree in physics from the University of Melbourne, Australia, in 1997. He has since been pursuing the Ph.D. degree at the University of Melbourne, concentrating on reduced-rank estimation and filtering, subspace tracking, and fast algorithms.

Petre Stoica (SM'91–F'94) received the D.Sc. degree in automatic control from the Bucharest Polytechnic Institute (BPI), Bucharest, Romania, in 1979 and an Honorary Doctorate degree in science from Uppsala University (UU), Uppsala, Sweden, in 1993.

He is a Professor of System Modeling with the Department of Systems and Control at UU. Previously, he was a Professor of Signal Processing at BPI. He held longer visiting positions with the Eindhoven University of Technology, Eindhoven, The Netherlands, Chalmers University of Technology, Göteborg, Sweden (where he held a Jubilee Visiting Professorship), UU, University of Florida, Gainesville, and Stanford University, Stanford, CA. His main scientific interests are in the areas of system identification, time series analysis and prediction, statistical signal and array processing, spectral analysis, wireless communications, and radar signal processing. He has published seven books, ten book chapters, and some 450 papers in archival journals and conference records on these topics. The most recent book he co-authored is *Introduction to Spectral Analysis* (Englewood Cliffs, NJ: Prentice-Hall, 1997). Recently, he co-edited two books on signal processing advances in wireless and mobile communications (Englewood Cliffs, NJ: Prentice-Hall, 2000). He is on the editorial boards of five journals in the field: *Signal Processing*; *Journal of Forecasting*; *Circuits, Systems and Signal Processing*; *Multidimensional Systems and Signal Processing*; and *Digital Signal Processing: A Review Journal*. He was a guest co-editor for several special issues on system identification, signal processing, spectral analysis, and radar for some of the above journals and for the *Proceedings of the IEE*.

Dr. Stoica was co-recipient of the 1989 ASSP Society Senior Award for a paper on statistical aspects of array signal processing and recipient of the 1996 Technical Achievement Award of the IEEE Signal Processing Society for fundamental contributions to statistical signal processing with applications in time series analysis, system identification, and array processing. In 1998, he received a Senior Individual Grant Award from the Swedish Foundation for Strategic Research. He is also co-recipient of the 1998 EURASIP Best Paper Award for Signal Processing for a work on parameter estimation of exponential signals with time-varying amplitude, a 1999 IEEE Signal Processing Society Best Paper Award for a paper on parameter and rank estimation of reduced-rank regressions, a 2000 IEEE Third Millennium Medal, and the 2000 IEEE W.R.G. Baker Paper Prize Award for a work on maximum likelihood methods for radar. He was a Member of the international program committees of many topical conferences. From 1981 to 1986, he was a Director of the International Time Series Analysis and Forecasting Society, and he has been a Member of the IFAC Committee on Modeling, Identification, and Signal Processing since 1994. He also is a Honorary Member of the Romanian Academy and a Fellow of the Royal Statistical Society.