Generic Filtering of Infinite Sets of Stochastic Signals

Anatoli Torokhti and Phil Howlett

Abstract—A theory for optimal filtering of infinite sets of random signals is presented. There are several new distinctive features of the proposed approach. First, a *single* optimal filter for processing any signal from a given *infinite* signal set is provided. Second, the filter is presented in the special form of a sum with p terms where each term is represented as a combination of three operations. Each operation is a special stage of the filtering aimed at facilitating the associated numerical work. Third, an iterative scheme is implemented into the filter structure to provide an improvement in the filter performance at each step of the scheme. The final step of the scheme concerns signal compression and decompression. This step is based on the solution of a new rank-constrained matrix approximation problem. The solution to the matrix problem is described in this paper. A rigorous error analysis is given for the new filter.

Keywords—Optimal filtering, data compression, stochastic signals.

I. Introduction

A. Motivation

In this paper, extensions of known approaches to optimal filtering based on the Wiener idea[1] are considered. A theory for a new nonlinear filter which processes *infinite* sets of random signals is presented. The filter is constructed via an iterative scheme that provides a signal processing improvement with each step. The filter provides simultaneous signal filtering and compression and the subsequent decompression (reconstruction).

There has been significant attention in the literature to filters that process finite sets of random signals but it seems that a filter which is able to process *infinite* sets of random signals has not been developed. The filter presented in this paper is designed specifically to process infinite sets of random signals. For the case of *finite* sets of random signals, we show that our filter leads to a lower computational load and better accuracy than the known filters; the improved accuracy is due to the special iteration procedure incorporated into the filter structure (see Section III-E2).

There are three motivations for the proposed method which we now describe.

1) First motivation: infinite sets of signals: Most of the literature on Wiener-like filtering provides an optimal filter for an *individual* input signal given by a *finite* random vector[2]. This means that if we wish to transform an infinite set $Y = \{\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \ldots, \mathbf{y}_{(N)}, \ldots\}$ of input vector signal into an infinite set $X = \{\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(N)}, \ldots\}$ of output vector signals using a Wiener-like approach then we have to find a set of corresponding Wiener filters $\{\mathcal{F}_{(1)}, \mathcal{F}_{(2)}, \ldots, \mathcal{F}_{(N)}, \ldots\}$ so that each representative $\mathcal{F}_i$ of the filter set relates to a representative $\mathbf{y}_{(i)}$ of the signal-vector set Y. Therefore such a filter cannot be applied if X and Y are infinite sets of signals. Moreover, in some situations, a *recognizer* must be used that will determine to which of the filters $\{\mathcal{F}_{(1)}, \mathcal{F}_{(2)}, \ldots, \mathcal{F}_{(N)}, \ldots\}$ each component from Y should be directed.

Note that even in the case when Y and X are finite sets, $Y = \{\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \ldots, \mathbf{y}_{(N)}\}$ and $X = \{\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(N)}\}$, and then Y and X can be represented as finite vectors, the Wiener approach leads to computation of large covariance matrices. Indeed, if each $\mathbf{y}_i$ has n components and each $\mathbf{x}_i$ has m components then the Wiener approach leads to computation of a product of an $mN \times nN$ matrix and an $nN \times nN$ matrix and computation of an $nN \times nN$ pseudo-inverse matrix [14]. This requires $O(2mn^2N^3)$ and $O(22n^3N^3)$ flops, respectively [5]. If m, n and N are sufficiently large then the computational work associated with this approach becomes unreasonably hard.

To avoid such drawbacks, we here study an approach that allows us to use *only one* filter to process any signal from the infinite set Y.

The first question we address in the paper is as follows. Let X and Y be *infinite* sets of signals. How should we construct a *single* optimal filter $\mathcal{F} : Y \to X$ which can be applied to each pair of signals $(\mathbf{x}, \mathbf{y}) \in X \times Y$ and which, moreover, transforms each $\mathbf{y}$ to a corresponding $\mathbf{x}$ with associated minimal error?

Surprisingly, perhaps, the answer is based firstly, on a dual representation of signal $\mathbf{x}$ in different spaces and secondly, on the use of the special norm (3) in the statement of the problem. The dual representation means that $\mathbf{x}$ is considered as a single signal in one representation, and on the other hand, as an infinite set of signals in the other, original, representation. A detailed explanation is given in Section VI. Examples of different special cases of the norm (3) used in our statement of the problem are presented in Section VI.

The answer for the first question is provided in Sections II-A, II-C, in Theorems 3 and 4, and in Section III-E. The special norm is given by (3) below.

2) Second motivation: improvement in the filter performance: The performance of filters used for data filtering, compression and subsequent reconstruction, is characterized by the accuracy, the compression ratio and the related computational load. The Karhunen-Loève filter (KLF) [10], [11],

Anatoli Torokhti and Phil Howlett are with Centre for Industrial and Applied Mathematics, School of Mathematics and Statistics, University of South Australia, Mawson Lakes, SA 5095, Australia..

Manuscript received November 19, 2008.

[1] Some references on Wiener-like filtering can be found in [6], [9], [13], [14], [15].

[2] We say a random vector $\mathbf{x}$ is finite if each realization $\mathbf{x} = \mathbf{x}(\omega)$ has a finite number of scalar components.

[14][3] is known to be the *optimal* filter that minimizes the reconstruction error over the class of all *linear* data compression-reconstruction filters. The KLF model is based on on the solution of a rank-constrained matrix approximation problem. Nevertheless, it may happen that the accuracy and compression ratio associated with the KLF are still not satisfactory in some circumstances.

The second question we address in this paper is as follows. Is there a filter that will have greater accuracy and better compression ratio then the KLF? An obvious but not constructive answer is that such a filter must be *nonlinear*. We propose a constructive determination of a nonlinear filter in the form of a sum with p terms given by (1) below, where each term is represented as a combination of three operations $\mathcal{G}_k$, $\mathcal{H}_k$ and $\mathcal{P}_k$ for each $k = 1, \ldots, p$. The operator $\mathcal{P}_k$ is a non-linear operator that allows us to incorporate additional information, the operator $\mathcal{H}_k$ is a generalized Gram-Schmidt orthogonalization that decouples the additional information and the operator $\mathcal{G}_k$ minimizes the estimation error. The orthogonalization is used primarily to reduce the associated numerical load.

The prime idea is to determine $\mathcal{G}_k$ from an iterative scheme aimed at improving the filter performance with each step. At the final step of the scheme, $\mathcal{G}_k$ is determined subject to the prescribed rank restrictions. The scheme is described in Section II-C. Moreover, due to the orthogonal structure created by the $\mathcal{H}_k$ it turns out that $\mathcal{G}_k$ is determined by solution of a *separate* minimization problem for each $k = 1, \ldots, p$. See Theorems 3 and 4 in this regard. The nonlinear operators $\mathcal{P}_k$ imply *non-linearity* of the filter. In the case $\mathcal{P}_k = I$ for each $k = 1, \ldots, p$, where I is the identity mapping, the filter is linear and the method presented in Sections II-C and III-B below becomes degenerate. Examples of possible nonlinear operators $\mathcal{P}_k$ are given in Section III-C.

The operations $\mathcal{P}_k$ and $\mathcal{H}_k$ are auxiliary operations related to finding $\mathcal{G}_k$, and they are described in Sections II-C and III-C, and Lemma 2, respectively.

3) Third motivation: generalized rank-constrained matrix approximation: Data compression filters are often based on the solution of a best rank-constrained matrix approximation problem. This is best explained for a linear filter. In this simplest case, the linear data compression filter (LDCF) consists of the two parts, compressor and de-compressor. Mathematically, the two parts are represented by an $m \times r$ matrix $F_{_C}$ and an $r \times n$ matrix $F_{_D}$ where $r < \min\{m, n\}$. The LDCF itself is represented as an $m \times n$ matrix $F = F_{_C} F_{_B}$. Thus, to find an optimal LDCF in the sense of minimizing its cost function, one can represent F as $F = F_{_D} F_{_C}$ and deal with two unknowns, $F_{_C}$ and $F_{_D}$. On the other hand the matrix F such that $F = F_{_D} F_{_C}$ can equivalently be written as F subject to $\operatorname{rank} F \le r$. Therefore, if finding an optimal LDCF is formulated as the problem of determining the matrix F that minimizes the cost function subject to $\operatorname{rank} F \le r$ then we deal with the only one unknown. The latter is more computationally preferable.

It has been shown in [4] that for such problems the solution given in [4] should be used. This solution has been obtained for the case of a minimal norm matrix. In Section III-A below, we extend the solution to a more general case and exploit it in our filter derivation.

B. Contribution

We provide a theory for the new filter which processes infinite sets of random signals. The filter is presented in the form (1) and is based on a new iterative scheme (Section II-C) that ensures the filter accuracy is improved with each step. The final step provides signal compression and decompression (Section III-D). This step is based on the solution of a new rank-constrained matrix approximation problem (Section III-A) following the methodology used in [4]. A rigorous error analysis for the filter is also given (Section III-B).

II. Method description and statement of the problem

A. Filter structure

Let (Ω, Σ, μ) be a probability space, where $\Omega = \{\omega\}$ is the set of outcomes, Σ a σ–field of measurable subsets in Ω and $\mu : \Sigma \to [0, 1]$ an associated probability measure on Σ with $\mu(\Omega) = 1$.

Let $\mathbf{x} = \{\mathbf{x}(\cdot, \alpha) \in L^2(\Omega, \mathbb{R}^m) \mid \alpha \in K\}$ where K is a measurable set [7] in some appropriate σ–field with a measure $\lambda(\alpha)$. Thus $\mathbf{x} \in L^2(\Omega \times K, \mathbb{R}^m)$. We observe an interesting duality in the representation of $\mathbf{x}$ in that $\mathbf{x}$ is a single signal in the space $L^2(\Omega \times K, \mathbb{R}^m)$ and is an infinite set of signals $\{\mathbf{x}(\cdot, \alpha) \in L^2(\Omega, \mathbb{R}^m) \mid \alpha \in K\}$ in the space $L^2(\Omega, \mathbb{R}^m)$. In similar fashion we write $\mathbf{y} = \{\mathbf{y}(\cdot, \alpha) \in L^2(\Omega, \mathbb{R}^m) \mid \alpha \in K\}$ and $\mathbf{y} \in L^2(\Omega \times K, \mathbb{R}^m)$. We interpret $\mathbf{x}$ as a signal that should be estimated and $\mathbf{y}$ as an observed signal that can be used as an input to the filter $\mathcal{F}_p$ studied below. In an intuitive way $\mathbf{y}$ can be regarded as a noise-corrupted version of $\mathbf{x}$. For example, $\mathbf{y}$ can be interpreted as $\mathbf{y} = \mathbf{x} + \mathbf{n}$ where $\mathbf{n}$ is white noise. In this paper, we do not restrict ourselves to this simplest version of $\mathbf{y}$ and assume that the dependence of $\mathbf{y}$ on $\mathbf{x}$ and $\mathbf{n}$ is arbitrary.

Let $\mathbf{x}_1 = \mathbf{y}$ and let $\mathbf{x}_1, \ldots, \mathbf{x}_p$ be a sequence of estimates of $\mathbf{x}$ obtained by the method described in Sections II-C, III-A and III-B below. The filter $\mathcal{F}_p$ is presented in the form

$$\mathcal{F}_p(\mathbf{y}) = \sum_{k=1}^{p} \mathcal{G}_k \mathcal{H}_k \mathcal{P}_k(\mathbf{x}_1, \ldots, \mathbf{x}_k), \tag{1}$$

where $\mathcal{P}_k : L^2((\Omega \times K)^k, \mathbb{R}^n) \to L^2(\Omega \times K, \mathbb{R}^n)$, $\mathcal{H}_k : L^2(\Omega \times K, \mathbb{R}^n) \to L^2(\Omega \times K, \mathbb{R}^n)$ and $\mathcal{G}_k : L^2(\Omega \times K, \mathbb{R}^n) \to L^2(\Omega \times K, \mathbb{R}^m)$.

The purpose of the components $\mathcal{G}_k$, $\mathcal{H}_k$ and $\mathcal{P}_k$ has been described above in Section I-A2. The components $\mathcal{P}_k$ are defined in Sections II-C and III-C, and components the $\mathcal{G}_k$ and $\mathcal{H}_k$ are determined in Section III-B.

B. Basic ideas

The input to the filter $\mathcal{F}_p$ is a stochastic signal – a random vector $\mathbf{y} \in L^2(\Omega \times K, \mathbb{R}^m)$. We wish to filter and compress $\mathbf{y}$, and to reconstruct a compressed vector that is close to $\mathbf{x}$.

[3] The Karhunen-Loève filter is often called the rank-reduced Wiener filter [11]. The analytic procedure is described by statisticians as Principal Component Analysis [8].

The first idea is to represent $\mathbf{y}$ as a generalized sequence of signals

$$\mathbf{y} = \{\mathbf{y}(\cdot,\alpha) \in L^2(\Omega,\mathbb{R}^m) \mid \alpha \in K\}$$

where (typically) $K \subset \mathbb{R}$ and to use the special norm (3) in the statement of the problems in Sections II-C and II-D. The filter processes each $\mathbf{y}(\cdot,\alpha)$ separately but the filter components are determined from the entire sets $\mathbf{x}$ and $\mathbf{y}$.

The filter is best illustrated for a particular case when the sets $\mathbf{x}$ and $\mathbf{y}$ are finite (as in Example 4 of Section VI), and we have, in (1), $p = 1$, $\mathcal{H}_1$ and $\mathcal{P}_1$ are each the identity mapping, and $\mathcal{G}_1$ is determined from problem (4). In such a case, we set $\mathbf{y} = \{\mathbf{y}(\cdot,t_k) \in L^2(\Omega,\mathbb{R}^m) \mid t_k \in \mathbb{R} \ \forall \ k = 1,\ldots,N\}$ and $\mathbf{x} = \{\mathbf{x}(\cdot,t_k) \in L^2(\Omega,\mathbb{R}^m) \mid t_k \in \mathbb{R} \ \forall \ k = 1,\ldots,N\}$. The estimate $\widetilde{\mathbf{x}}$ of $\mathbf{x}$ by the proposed filter is given by

$$\widetilde{\mathbf{x}} = [\widetilde{\mathbf{x}}(\cdot,t_1),\ldots,\widetilde{\mathbf{x}}(\cdot,t_N)] = \widetilde{\mathcal{G}}_1[\mathbf{y}(\cdot,t_1),\ldots,\mathbf{y}(\cdot,t_N)]$$

where $\widetilde{\mathcal{G}}_1$ is provided by Theorem 3 of Section III-B and, for each $k = 1,\ldots,N$, the estimate $\widetilde{\mathbf{x}}(\cdot,t_k) = \widetilde{\mathcal{G}}_1\mathbf{y}(\cdot,t_k)$ is determined separately.

The second idea is based on the following observation. *It is natural to expect that the filter performance would be better if the input to $\mathcal{F}_p$ was $\mathbf{x}$, and not $\mathbf{y}$.* Indeed, in this case, no transformation of $\mathbf{y}$ to $\mathbf{x}$ need be done and in particular, no noise filtering is necessary. The filter would provide only compression and subsequent decompression. This means that the error associated with such a filter performance will be linked only to compression and decompression, and will not be increased by an error associated with the transformation of $\mathbf{y}$ to $\mathbf{x}$. Hence, in such a case, the overall error will be less.

Therefore, our next idea is to construct an iterative scheme such that, at each step, the scheme allows us:

(a) to find a new input for $\mathcal{F}_p$ which is closer to $\mathbf{x}$ than those determined in preceding steps,

(b) to determine operators $\mathcal{G}_1,\ldots,\mathcal{G}_k$ that minimize the associated error at each step, and

(c) to define operators $\mathcal{H}_1,\ldots,\mathcal{H}_k$ so that the desired operators $\mathcal{G}_1,\ldots,\mathcal{G}_k$ are determined from a numerically simple scheme (see Section III-E1 for more detail).

This idea is realized below in Sections II-C and III-B in such a way that $\mathcal{F}_p$ is determined from *a sequence* of associated error minimization problems.

C. Basic device

Let the model for the filter $\mathcal{F}_p$ be given by (1). The proposed device for determining operators $\mathcal{H}_k$, $\mathcal{G}_k$ and $\mathcal{P}_k$ in (1) is as follows. The operators $\mathcal{H}_k$ in (1) are defined by Lemma 2 of Section III-B below. The recommended procedure for determining the operators $\mathcal{G}_k$ and $\mathcal{P}_k$ consists of p successive steps based on the following idea. If $\mathbf{x}_1 = \mathbf{y}$ and the subsequent estimates of $\mathbf{x}$ denoted by $\mathbf{x}_2,\ldots,\mathbf{x}_{k-1}$ have been determined from successive steps then for the kth step, we define

$$\mathbf{x}_k = \sum_{j=1}^{k-1} \widetilde{\mathcal{G}}_j \mathcal{H}_j(\mathbf{y}_1,\ldots,\mathbf{y}_j) \tag{2}$$

where $\mathbf{y}_j = \mathcal{P}_j(\mathbf{x}_1,\ldots,\mathbf{x}_j)$ and $\mathcal{P}_j$ is conceptually defined in (1) and where $\widetilde{\mathcal{G}}_1$, ..., $\widetilde{\mathcal{G}}_k$ are determined from the proposed error minimization criteria. We propose the following scheme.

First, we introduce the norm $\|\cdot\|_{K,\Omega}$ defined by

$$\|\mathbf{x}\|^2_{K,\Omega} = \frac{1}{\lambda(K)} \iint_{\Omega\times K} \|\mathbf{x}(\omega,\alpha)\|_2^2 d(\mu(\omega),\lambda(\alpha)), \tag{3}$$

where $\lambda(K) = \int_K d\lambda(\alpha)$ is the measure of K and where $\|\mathbf{x}(\omega,\alpha)\|_2$ is the Euclidean norm of $\mathbf{x}(\omega,\alpha) \in \mathbb{R}^m$. More explanation is given in Section VI.

The proposed scheme consists of a sequence of steps.
Step 1 *The initial step.* For $k = 1$, let $\mathbf{x}_1 := \mathbf{y}$ and let $\mathcal{P}_1 = I$ and $\mathcal{H}_1 = I$ so that $\mathbf{y}_1 := \mathbf{x}_1$ and find $\widetilde{\mathcal{G}}_1$ that satisfies

$$\|\mathbf{x} - \widetilde{\mathcal{G}}_1(\mathbf{y}_1)\|^2_{K,\Omega} = \min_{\mathcal{G}_1} \|\mathbf{x} - \mathcal{G}_1(\mathbf{y}_1)\|^2_{K,\Omega}. \tag{4}$$

Set $\mathbf{x}_2 = \widetilde{\mathcal{G}}_1(\mathbf{x}_1)$.
Step $\ell-1$ *for* $\ell = 3,\ldots,p$. *The filtering steps.* Let $\mathbf{x}_2,\ldots,\mathbf{x}_{\ell-1}$ be known from the preceding $\ell-2$ steps. For $k = \ell-1$ in (2) define $\mathcal{P}_{\ell-1}$ so that $\mathbf{y}_{\ell-1} = \mathcal{P}_{\ell-1}(\mathbf{x}_1,\ldots,\mathbf{x}_{\ell-1})$ and find $\widetilde{\mathcal{G}}_1$, $\widetilde{\mathcal{G}}_2$, ..., $\widetilde{\mathcal{G}}_{\ell-1}$ that satisfy

$$\begin{aligned}&\|\mathbf{x} - \sum_{j=1}^{\ell-1} \widetilde{\mathcal{G}}_j \mathcal{H}_j(\mathbf{y}_1,\ldots,\mathbf{y}_j)]\|^2_{K,\Omega}\\ &\qquad = \min_{\mathcal{G}_1,\ldots,\mathcal{G}_{\ell-1}} \|\mathbf{x} - \sum_{j=1}^{\ell-1} \mathcal{G}_j \mathcal{H}_j(\mathbf{y}_1,\ldots,\mathbf{y}_j)\|^2_{K,\Omega}.\end{aligned} \tag{5}$$

Set $\mathbf{x}_\ell = \sum_{j=1}^{\ell-1} \widetilde{\mathcal{G}}_j \mathcal{H}_j(\mathbf{y}_1,\ldots,\mathbf{y}_j)$.
Step p. *The reconstruction step.* At the final step, for $k = p$ in (2) define $\mathcal{P}_p$ so that $\mathbf{y}_p = \mathcal{P}_p(\mathbf{x}_1,\ldots,\mathbf{x}_p)$ and find $\widehat{\mathcal{G}}_1$, $\widehat{\mathcal{G}}_2$, ..., $\widehat{\mathcal{G}}_p$ that satisfy

$$\begin{aligned}&\|\mathbf{x} - \sum_{j=1}^{p} \widehat{\mathcal{G}}_j \mathcal{H}_j(\mathbf{y}_1,\ldots,\mathbf{y}_j)]\|^2_{K,\Omega}\\ &\qquad = \min_{\mathcal{G}_1,\ldots,\mathcal{G}_p} \|\mathbf{x} - \sum_{j=1}^{p} \mathcal{G}_j \mathcal{H}_j(\mathbf{y}_1,\ldots,\mathbf{y}_j)\|^2_{K,\Omega}\end{aligned} \tag{6}$$

subject to the rank constraints

$$\sum_{j=1}^{p} \text{rank}\, \widehat{\mathcal{G}}_j \le r \le \min\{m,n\} \quad \text{and} \quad \text{rank}\, \widehat{\mathcal{G}}_j \le r_j \tag{7}$$

with $\displaystyle\sum_{j=1}^{p} r_j = r.$

We set $\mathbf{x}_{p+1} = \sum_{i=1}^{p} \widehat{\mathcal{G}}_i \mathcal{H}_i(\mathbf{y}_1,\ldots,\mathbf{y}_i)$.

Due to condition (7), this step provides data compression–reconstruction as it is shown in Section III-D below.

D. Statement of the problem

The problem we solve follows from the basic structure presented above. Let $\mathbf{x}$ and $\mathbf{y}$ be random signals where $\mathbf{x}$ is an unobservable signal and $\mathbf{y}$ is an observable input signal. Both $\mathbf{x}$ and $\mathbf{y}$ are analytically unknown and the only available information is given by covariance matrices formed from $\mathbf{x}$ and $\mathbf{y}$. We wish to find a filter of the form (1) with $\mathcal{G}_1,\ldots,\mathcal{G}_p$ determined from the error minimization problems (4)–(7). The filter must satisfy the requirements (a), (b) and (c) of the Section II-B.

The main differences of the problem given by (1), (4)–(7) from known problems are as follows. First, we are looking for the optimal filter that minimizes the associated error for any signal from an *infinite* signal set. This is achieved due to the new norm given by (3). The second major difference is the special structure of the filter (1)–(7). Each step of the suggested procedure represents the solution of a best approximation problem that minimizes an associated error. If the final stage *Step* p is omitted the scheme performs filtering only. If *Step* p is included then the scheme provides simultaneous signal filtering and compression and the subsequent reconstruction.

Solutions to problems (4)–(7) are given below under assumptions that certain covariance matrices are known. The assumptions are specified in Theorems 3 and 4 below. It is important to note that for new signals constructed during the phase of iterative improvement the required covariance matrices can be constructed from the original data.

E. Advantages of the filter

The advantages of the proposed filter are discussed in Section III-E.

III. Main results

The solution to the problem (6)–(7) is based on the results presented in Section III-A. The solution itself is given in Theorem 4 below.

A. Generic low-rank matrix approximation problem

Let $\mathbb{C}^{m\times n}$ be the set of $m \times n$ complex valued matrices, and denote by $\mathcal{R}(m,n,k) \subseteq \mathbb{C}^{m\times n}$ the variety of all $m \times n$ matrices of rank k at most. Fix $A = [a_{ij}]_{i,j=1}^{m,n} \in \mathbb{C}^{m\times n}$. Then $A^* \in \mathbb{C}^{n\times m}$ is the conjugate transpose of A. Let the SVD of A be given by

$$A = U_A \Sigma_A V_A^*, \tag{8}$$

where $U_A \in \mathbb{C}^{m\times m}$ and $V_A \in \mathbb{C}^{n\times n}$ are unitary matrices and

$$\Sigma_A := \text{diag}(\sigma_1(A), \ldots, \sigma_{\min(m,n)}(A)) \in \mathbb{C}^{m\times n}$$

is a generalized diagonal matrix with singular values $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq 0$ on the main diagonal. We write

$$A^\dagger = V_A \Sigma_A^\dagger U_A^*$$

for the Moore-Penrose pseudo-inverse [1] of the matrix A. In this definition

$$\Sigma_A^\dagger := \text{diag}(\sigma_1(A)^\dagger, \ldots, \sigma_{\min(m,n)}(A)^\dagger) \in \mathbb{C}^{n\times m}$$

where we use the notation

$$\sigma_j(A)^\dagger = \begin{cases} \sigma_j(A)^{-1} & \text{if } \sigma_j(A) \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Let $U_A = [u_1 \; u_2 \; \ldots u_m]$ and $V_A = [v_1 \; v_2 \; \ldots v_n]$ be the representations of U and V in terms of their columns. Let $r = \text{rank}\, A$ and write $U_A = [U_{A1}\, U_{A2}]$ where $U_{A1} \in \mathbb{C}^{m\times r}$ and $U_{A2} \in \mathbb{C}^{m\times(m-r)}$ and $V_A = [V_{A1}\, V_{A2}]$ where $V_{A1} \in \mathbb{C}^{n\times r}$ and $V_{A2} \in \mathbb{C}^{n\times(n-r)}$. Then

$$P_{A,L} := \sum_{i=1}^{r} u_i u_i^* = U_{A1}U_{A1}^* \text{ and } P_{A,R} := \sum_{i=1}^{r} v_i v_i^* = V_{A1}V_{A1}^* \tag{9}$$

are the orthogonal projections on the ranges of A and A^*, respectively. We note that $\Sigma_A = U_A^* A V_A$ and we will use the notation

$$\begin{aligned}\Sigma_A &= \begin{bmatrix} U_{A1}^* \\ U_{A2}^* \end{bmatrix} A \begin{bmatrix} V_{A1} & V_{A2} \end{bmatrix} \\ &= \begin{bmatrix} U_{A1}^* A V_{A1} & U_{A1}^* a V_{A2} \\ U_{A2}^* A V_{A1} & U_{A2}^* A V_{A2} \end{bmatrix} = \begin{bmatrix} A_1 & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}.\end{aligned}$$

Define

$$A_k := (A)_k := \sum_{i=1}^{k} \sigma_i(A) u_i v_i^* = U_{Ak}\Sigma_{Ak}V_{Ak}^* \in \mathbb{C}^{m\times n} \tag{10}$$

for $k = 1, \ldots, \text{rank}\, A - 1$, where

$$\begin{aligned} U_{Ak} &= [u_1 \; u_2 \; \ldots u_k], \;\; \Sigma_{Ak} = \text{diag}(\sigma_1(A), \ldots, \sigma_k(A)) \\ &\text{and} \quad V_{Ak} = [v_1 \; v_2 \; \ldots v_k]. \end{aligned} \tag{11}$$

For $k \geq \text{rank}\, A$, we note that $A_k := A \; = (A)_{\text{rank}\, A}$. For $1 \leq k < \text{rank}\, A$, the matrix A_k is uniquely defined if and only if $\sigma_k(A) > \sigma_{k+1}(A)$.

Lemma 1: Let $A \in \mathbb{C}^{m\times n}$. If $R \in \mathbb{C}^{m\times m}$ and $S \in \mathbb{C}^{n\times n}$ are unitary matrices then $R(A)_k S^* = (RAS^*)_k$.

Proof. If $A = U_A\Sigma_A V_A^*$ is the singular value decomposition for A then $RAS^* = (RU_A)\Sigma_A(SV_A)^*$ is the singular value decomposition of RAS^*. Thus

$$\begin{aligned}(RAS^*)_k &= \sum_{i=1}^{k} \sigma_i(A)(Ru_i)(Sv_i)^* = R\left[\sum_{i=1}^{k} \sigma_i(A) u_i v_i^*\right] S^* \\ &= R(A)_k S^*. \qquad \blacksquare\end{aligned}$$

Henceforth $\|\cdot\|_F$ denotes the Frobenius norm. Below, we provide generalizations of the classical minimization problem due to Eckart and Young [3]. First, we present the result obtained in [4].

Theorem 1: [4] Let matrices $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{m\times p}$ and $C \in \mathbb{C}^{q\times n}$ be given. Then

$$X = B^\dagger (P_{B,L} A P_{C,R})_k C^\dagger \tag{12}$$

is a solution to the minimization problem

$$\min_{X\in\mathcal{R}(p,q,k)} ||A - BXC||_F, \tag{13}$$

having the minimal $||X||_F$. This solution is unique if and only if either

$$k \geq \text{rank}\; (P_{B,L}AP_{C,R})$$

or

$$1 \leq k < \text{rank}\; (P_{B,L}AP_{C,R})$$
$$\text{and} \quad \sigma_k(P_{B,L}AP_{C,R}) > \sigma_{k+1}(P_{B,L}AP_{C,R}).$$

We will explain this result as briefly as possible and refer the reader to the original article by Friedland and Torokhti [4] for more details. Let $s = \text{rank}\, B$ and $t = \text{rank}\, C$. Using the notation introduced above we write $U_B = [U_{B1}\, U_{B2}]$ where $U_{B1} \in \mathbb{C}^{m\times s}$ and $U_{B2} \in \mathbb{C}^{m\times(m-s)}$ and $V_C = [V_{C1}\, V_{C2}]$ where $V_{C1} \in \mathbb{C}^{n\times t}$ and $V_{C2} \in \mathbb{C}^{n\times(n-t)}$. Thus we have

$$\begin{aligned}\Sigma_B &= \begin{bmatrix} U_{B1}^* \\ U_{B2}^* \end{bmatrix} A \begin{bmatrix} V_{B1} & V_{B2} \end{bmatrix} \\ &= \begin{bmatrix} U_{B1}^* B V_{B1} & U_{B1}^* B V_{B2} \\ U_{B2}^* B V_{B1} & U_{B2}^* B V_{B2} \end{bmatrix} = \begin{bmatrix} B_1 & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}\Sigma_C &= \begin{bmatrix} U_{C1}^* \\ U_{C2}^* \end{bmatrix} A \begin{bmatrix} V_{C1} & V_{C2} \end{bmatrix} \\ &= \begin{bmatrix} U_{C1}^* C V_{C1} & U_{C1}^* C V_{C2} \\ U_{C2}^* C V_{C1} & U_{C2}^* C V_{C2} \end{bmatrix} = \begin{bmatrix} C_1 & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}\end{aligned}$$

We will write

$$\widetilde{A} = U_B^* A V_C = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where $A_{11} = U_{B1}^* A V_{C1} \in \mathbb{C}^{s\times t}$, $A_{12} = U_{B1}^* A V_{C2} \in \mathbb{C}^{s\times(n-t)}$, $A_{21} = U_{B2}^* A V_{C1} \in \mathbb{C}^{(m-s)\times t}$ and $A_{22} = U_{B2}^* A V_{C2} \in \mathbb{C}^{(m-s)\times(n-t)}$. The idea behind our explanation of Theorem 1 is simple. The Frobenius norm is not changed when we multiply on the left or the right by a unitary matrix. Thus we have

$$\begin{aligned}\|A - BXC\|_F &= \|U_B^*(A - U_B \Sigma_B V_B^* X U_C \Sigma_C V_C^*) V_C\|_F \\ &= \|\widetilde{A} - \Sigma_B \widehat{X} \Sigma_C\|_F\end{aligned}$$

where

$$\widehat{X} = V_B^* X U_C = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}.$$

Since

$$\Sigma_B = \begin{bmatrix} B_1 & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} \quad \text{and} \quad \Sigma_C = \begin{bmatrix} C_1 & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}$$

it follows that

$$\begin{aligned}\|A - BXC\|_F &= \|A_{11} - B_1 X_{11} C_1\|_F \\ &+ \|A_{12}\|_F + \|A_{21}\|_F + \|A_{22}\|_F\end{aligned}$$

from which it follows that the solution $X = X_0$ to problem (13) is given by $X_{11} = B_1^{-1}(A_{11})_k C_1^{-1}$, $X_{12} = \mathbb{O}$, $X_{21} = \mathbb{O}$ and $X_{22} = \mathbb{O}$. We will use the following argument on a number of occasions. We observe $B_1^{-1} = V_{B1}^* B^\dagger U_{B1}$ and $C_1^{-1} = V_{C1}^* C^\dagger U_{C1}$ and hence deduce

$$\begin{aligned} & X_{11} = B_1^{-1}(A_{11})_k C_1^{-1} \\ &= \begin{bmatrix} B_1^{-1} & \mathbb{O} \end{bmatrix} \left(\begin{bmatrix} I_s & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} I_t & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} \right)_k \\ & \times \begin{bmatrix} C_1^{-1} \\ \mathbb{O} \end{bmatrix} \\ &= V_{B1}^* B^\dagger \begin{bmatrix} U_{B1} & U_{B2} \end{bmatrix} \left(\begin{bmatrix} U_{B1}^* \\ \mathbb{O} \end{bmatrix} A \begin{bmatrix} V_{C1} & \mathbb{O} \end{bmatrix} \right)_k \\ & \times \begin{bmatrix} V_{C1}^* \\ V_{C2}^* \end{bmatrix} C^\dagger U_{C1} \\ &= V_{B1}^* B^\dagger \left(U_{B1} U_{B1}^* A V_{C1} V_{C1}^* \right)_k C^\dagger U_{C1} \\ &= V_{B1}^* B^\dagger (P_{B,L} A P_{C,R})_k C^\dagger U_{C1}.\end{aligned}$$

We know that $V_{B1} V_{B1}^* = B^\dagger B$ and $U_{C1} U_{C1}^* = C C^\dagger$ and so

$$\begin{aligned} V_{B1} X_{11} U_{C1}^* &= P_{B,R} B^\dagger (P_{B,L} A P_{C,R})_k C^\dagger P_{C,L} \\ &= B^\dagger (P_{B,L} A P_{C,R})_k C^\dagger.\end{aligned}$$

Therefore

$$\begin{aligned} X_0 &= \begin{bmatrix} V_{B1} & V_{B2} \end{bmatrix} \begin{bmatrix} B_1^{-1}(A_{11})_k C_1^{-1} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} \begin{bmatrix} U_{C1}^* \\ U_{C2}^* \end{bmatrix} \\ &= B^\dagger (P_{B,L} A P_{C,R})_k C^\dagger.\end{aligned}$$

To find a general solution to problem (13) without the constraint of the minimal $||X||_F$ we must have

$$X = V_B \widehat{X} U_C^* \quad \text{with} \quad \widehat{X} = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}, \qquad (14)$$

where $X_{11} = V_{B1}^* B^\dagger (P_{B,L} A P_{C,R})_k C^\dagger U_{C1}$ and X_{12}, X_{21} and X_{22} are chosen in such a way that $\widehat{X} \in \mathcal{R}(p, q, k)$. In Theorem 2 below, we show how to choose X_{12}, X_{21} and X_{22} (see (21)) to satisfy the condition $\widehat{X} \in \mathcal{R}(p, q, k)$.

Theorem 2: If the requirement that $\|X\|_F$ should be minimized is omitted then the solution to problem (13) is not unique. In this case the general solution to problem (13) can be written in the form

$$X = B^\dagger (P_{B,L} A P_{C,R})_k C^\dagger + K. \qquad (15)$$

where

$$K = \begin{bmatrix} P_{B,R} & I - P_{B,R} \end{bmatrix} \begin{bmatrix} \mathbb{O} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \begin{bmatrix} P_{C,L} \\ I - P_{C,L} \end{bmatrix}, \qquad (16)$$

$$\begin{aligned} & K_{12} = V_{B1} X_{11} Q U_{C2}^*, \quad K_{21} = V_{B2} P X_{11} U_{C1}^*, \qquad (17) \\ & K_{22} = V_{B2} P X_{11} Q U_{C2}^*, \ X_{11} = V_{B1}^* B^\dagger (P_{B,L} A P_{C,R})_k C^\dagger U_{C1},\end{aligned}$$

and where $P \in \mathbb{C}^{(p-s)\times s}$ and $Q \in \mathbb{C}^{t\times(q-t)}$ are arbitrary.

Proof. To preserve $\operatorname{rank} \widehat{X} = \operatorname{rank} X_{11}$ we should choose X_{12}, X_{21} and X_{22} from (14) in a compatible form. To be specific there must exist matrices $P \in \mathbb{C}^{(p-s)\times s}$ and $Q \in \mathbb{C}^{t\times(q-t)}$ such that

$$\begin{bmatrix} I & \mathbb{O} \\ -P & I \end{bmatrix} \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \begin{bmatrix} I & -Q \\ \mathbb{O} & I \end{bmatrix} = \begin{bmatrix} X_{11} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix}$$

from which it follows that

$$X_{12} = X_{11} Q, \quad X_{21} = P X_{11} \quad \text{and} \quad X_{22} = P X_{11} Q \quad (18)$$

where $X_{11} = V_{B1}^* B^\dagger (P_{B,L} A P_{C,R})_k C^\dagger U_{C1}$. Since

$$X = V_B \widehat{X} U_C^* = V_B \begin{bmatrix} X_{11} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{bmatrix} U_C^* + V_B \begin{bmatrix} \mathbb{O} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} U_C^*$$

we have

$$\begin{aligned} X &= B^\dagger (P_{B,L} A P_{C,R})_k C^\dagger + V_{B1} X_{12} U_{C2}^* \\ &+ V_{B2} X_{21} U_{C1}^* + V_{B2} X_{22} U_{C2}^*.\end{aligned}$$

We note that $V_{B1} = V_{B1} V_{B1}^* V_{B1} = P_{B,R} V_{B1}$, $U_{C1}^* = U_{C1}^* U_{C1} U_{C1}^* = U_{C1}^* P_{C,L}$, $V_{B2} = V_{B2} V_{B2}^* V_{B2} = (I_p -$

$P_{B,R})V_{B2}$ and $U^*_{C2} = U^*_{C2}U_{C2}U^*_{C2} = U^*_{C2}(I_q - P_{C,L})$ and hence

$$\begin{aligned} X &= B^\dagger(P_{B,L}AP_{C,R})_kC^\dagger + P_{B,R}V_{B1}X_{11}QU^*_{C2}(I_q - P_{C,L}) \\ &\quad +(I_p - P_{B,R})V_{B2}PX_{11}U^*_{C1}P_{C,L} \\ &\quad\quad +(I_p - P_{B,R})V_{B2}PX_{11}QU^*_{C2}(I_q - P_{C,L}) \\ &= B^\dagger(P_{B,L}AP_{C,R})_kC^\dagger + P_{B,R}K_{12}(I_q - P_{C,L}) \\ &\quad +(I_p - P_{B,R})K_{21}P_{C,L} + (I_p - P_{B,R})K_{22}(I_q - P_{C,L}) \end{aligned}$$

where $K_{12} = V_{B1}X_{11}QU^*_{C2}$, $K_{21} = V_{B2}PX_{11}U^*_{C1}$, $K_{22} = V_{B2}PX_{11}QU^*_{C2}$ and $X_{11} = V^*_{B1}B^\dagger(P_{B,L}AP_{C,R})_kC^\dagger U_{C1}$. ■

The following Corollary will be used in the next Section.

Corollary 1: Let $p = m$, $q = n$ and $B = I$ and define $\widetilde{A} = AV_C = [A_1\, A_2]$ where $A_1 = AV_{C1} \in \mathbb{C}^{m\times t}$ and $A_2 = AV_{C2} \in \mathbb{C}^{m\times(n-t)}$. The general solution to the problem

$$\min_{X\in\mathcal{R}(m,n,k)} ||A - XC||_F \tag{19}$$

is given by

$$X = (AP_{C,R})_kC^\dagger + (AP_{C,R})_kC^\dagger U_{C1}QU^*_{C2}(I_q - P_{C,L}). \tag{20}$$

where $Q \in \mathbb{C}^{t\times(n-t)}$ is arbitrary.

Proof Let $\widehat{X} = XU_C = X[U_{C1}\, U_{C2}] = [X_1\, X_2]$. Since

$$\|A - XC\|_F = \|\widetilde{A} - \widehat{X}\Sigma_C\|_F = \|A_1 - X_1C_1\|_F + \|A_2\|_F$$

it follows that the solution to problem () with minimum value for $\|X\|_F$ is given by $X = (A_1P_{C,R})_kCF^\dagger$. Now consider the problem without the constraint requiring a minimum value for $\|X\|_F$. To preserve $\mathrm{rank}\,\widehat{X} = \mathrm{rank}\,X_1$ we should choose X_2 in a compatible form. To be specific there must exist a matrix $Q \in \mathbb{C}^{t\times(q-t)}$ such that

$$\left[\begin{array}{cc} X_1 & X_2 \end{array}\right]\left[\begin{array}{cc} I & -Q \\ \mathbb{O} & I \end{array}\right] = \left[\begin{array}{cc} X_1 & \mathbb{O} \end{array}\right]$$

from which it follows that

$$X_2 - X_1Q = \mathbb{O}. \tag{21}$$

By adapting the general argument used in the previous theorem we can see that $X_1 = (AP_{C,R})_kC^\dagger U_{C1}$ and $X_2 = (AP_{C,R})_kC^\dagger U_{C1}Q$. Therefore

$$X = \widehat{X}U^*_C = X_1U^*_{C1} + X_2U^*_{C2}$$

and so $X = (AP_{C,R})_kC^\dagger + (AP_{C,R})_kC^\dagger U_{C1}QU^*_{C2}$. We note that $U^*_{C2} = U^*_{C2}U_{C2}U^*_{C2} = U^*_{C2}(I_q - P_{C,L})$ and hence

$$X = (AP_{C,R})_kC^\dagger + (AP_{C,R})_kC^\dagger U_{C1}QU^*_{C2}(I_q - P_{C,L}).$$

■

Remark 1: The Eckart-Young theorem *[3]* follows from Theorem 2 as a particular case when $p = m$, $q = n$, $B = I_m$ and $C = I_n$.

B. *Solution of the problems* (4)–(5) *and* (6)–(7)

Let $\mathbf{x} = \{\mathbf{x}(\cdot,\alpha) \in L^2(\Omega,\mathbb{R}^m) \mid \alpha \in K\} \in L^2(\Omega\times K,\mathbb{R}^m)$ and $\mathbf{y} = \{\mathbf{y}(\cdot,\alpha) \in L^2(\Omega,\mathbb{R}^m) \mid \alpha \in K\} \in L^2(\Omega\times K,\mathbb{R}^m)$ where K is a measurable set with finite measure $\lambda(K)$ in some sigma field of measurable sets [7].

We write $\mathbf{x} = (x^{(1)},\ldots,x^{(m)})^T$ and $\mathbf{y} = (y^{(1)},\ldots,y^{(n)})^T$ where $x^{(i)} \in L^2(\Omega\times K,\mathbb{R})$ for $i = 1,\ldots,m$ and $y^{(j)} \in L^2(\Omega\times K,\mathbb{R})$ for $j = 1,\ \ldots,n$ are real valued random variables. Let

$$\langle x^{(i)},y^{(j)}\rangle = \frac{1}{\lambda(K)}\iint_{\Omega\times K} x^{(i)}(\omega,\alpha)y^{(j)}(\omega,\alpha)\, d(\mu(\omega),\lambda(\alpha)) \tag{22}$$

and

$$\langle y^{(i)},y^{(j)}\rangle = \frac{1}{\lambda(K)}\iint_{\Omega\times K} y^{(i)}(\omega,\alpha)y^{(j)}(\omega,\alpha)\, d(\mu(\omega),\lambda(\alpha)) \tag{23}$$

and define covariance matrices

$$R[\mathbf{x}\mathbf{y}^T] = \left[\langle x^{(i)},y^{(j)}\rangle\right] \in \mathbb{R}^{m\times n} \tag{24}$$

and

$$R[\mathbf{y}\mathbf{y}^T] = \left[\langle y^{(i)},y^{(j)}\rangle\right] \in \mathbb{R}^{n\times n}. \tag{25}$$

If $M \in \mathbb{R}^{m\times n}$ then we define a corresponding operator $\mathcal{M} : L^2(\Omega\times K,\mathbb{R}^n) \to L^2(\Omega\times K,\mathbb{R}^m)$ by the formula

$$[\mathcal{M}(\mathbf{y})](\omega,\alpha) = M[\mathbf{y}(\omega,\alpha)]. \tag{26}$$

Henceforth all such operators will be denoted by a calligraphic letter.

We wish

(i) to give explicit solutions to the problems (4)–(5) and (6)–(7), and

(ii) to show that the error associated with the proposed method decreases as the number k of steps (4)–(7) increases.

First, we give a method for the determination of the operators $\mathcal{H}_1,\ldots,\mathcal{H}_k$ in the filter model (1).

Definition 1: The random vectors $\{\mathbf{v}_i\}_{i=1,\ldots,k} \subset L^2(\Omega\times K,\mathbb{R}^n)$ are called pairwise orthogonal if

$$R[\mathbf{v}_i\mathbf{v}_j^T] = \mathbb{O} \in \mathbb{R}^{n\times n} \tag{27}$$

for $i \neq j$ with $i,j = 1,\ldots,k$.

Lemma 2: Let $\mathbf{y}_i = \mathcal{P}_i(\mathbf{x}_1,\ldots,\mathbf{x}_i)$ and $\mathbf{v}_i = \mathcal{H}_i(\mathbf{y}_1,\ldots,\mathbf{y}_i)$ for each $i = 1,2,\ldots,k$ where the linear operators $\mathcal{H}_i : L^2((\Omega\times K)^i,\mathbb{R}^n) \mapsto L^2(\Omega\times K,\mathbb{R}^n)$ are defined inductively by the generalized Gram-Schmidt algorithm

$$\mathbf{v}_1 = \mathbf{y}_1 \quad\text{and}\quad \mathbf{v}_i = \mathbf{y}_i - \sum_{\ell=1}^{i-1}\mathcal{L}_{i\ell}(\mathbf{v}_\ell) \tag{28}$$

with $\mathcal{L}_{i\ell} : L^2(\Omega\times K,\mathbb{R}^n) \to L^2(\Omega\times K,\mathbb{R}^n)$ defined by $\mathcal{L}_{i\ell}(\mathbf{v}_\ell) = L_{i\ell}\mathbf{v}_\ell$ where

$$L_{i\ell} = R[\mathbf{y}_i\mathbf{v}_\ell^T]R[\mathbf{v}_\ell\mathbf{v}_\ell^T]^\dagger + M_{i\ell}(I - R[\mathbf{v}_\ell\mathbf{v}_\ell^T]R[\mathbf{v}_\ell\mathbf{v}_\ell^T]^\dagger) \tag{29}$$

and $M_{i\ell} \in \mathbb{R}^{n\times n}$ is arbitrary for each $i = 2,\ldots,k$. Then the vectors $\mathbf{v}_1,\mathbf{v}_2\ldots,\mathbf{v}_k$ are pairwise orthogonal in $L^2(\Omega\times K,\mathbb{R}^n)$.

Proof. The proof follows directly from (28), (29) and Definition 1 on the basis of the relation $R[\mathbf{y}_i\mathbf{v}_\ell^T] = R[\mathbf{y}_i\mathbf{v}_\ell^T]R[\mathbf{v}_\ell\mathbf{v}_\ell^T]^\dagger R[\mathbf{v}_\ell\mathbf{v}_\ell^T]$. We refer the reader to Torokhti and Howlett [14], p. 168 for additional information. ■

Remark 2: In practice it is usual to take $M_{i\ell} = \mathbb{O}$.

Remark 3: The Gram-Schmidt formula can be rearranged to give

$$\mathbf{y}_1 = \mathbf{v}_1 \quad \text{and} \quad \mathbf{y}_i = \mathbf{v}_i + \sum_{\ell=1}^{i-1} L_{i\ell}\mathbf{v}_\ell$$

for each $i = 2, \ldots, k$. Thus we can write

$$\mathbf{y}^{(i)} = (\mathbf{I}_{im} + \mathbf{L}_i)\mathbf{v}^{(i)}$$

$$\Leftrightarrow \quad \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \vdots \\ \mathbf{y}_i \end{bmatrix} = \begin{bmatrix} \mathbf{I}_m & \mathbb{O} & \mathbb{O} & \cdots & \mathbb{O} \\ \mathbf{L}_{21} & \mathbf{I}_m & \mathbb{O} & \cdots & \mathbb{O} \\ \mathbf{L}_{31} & \mathbf{L}_{32} & \mathbf{I}_m & \cdots & \mathbb{O} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{L}_{i1} & \mathbf{L}_{i2} & \mathbf{L}_{i3} & \cdots & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \\ \vdots \\ \mathbf{v}_i \end{bmatrix}$$

from which it follows easily that

$$\mathbf{v}^{(i)} = (\mathbf{I}_{im} + \mathbf{L}_i)^{-1}\mathbf{y}^{(i)}$$

$$= \begin{bmatrix} \mathbf{I}_m & \mathbb{O} & \mathbb{O} & \cdots & \mathbb{O} \\ \mathbf{H}_{21} & \mathbf{I}_m & \mathbb{O} & \cdots & \mathbb{O} \\ \mathbf{H}_{31} & \mathbf{H}_{32} & \mathbf{I}_m & \cdots & \mathbb{O} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_{i1} & \mathbf{H}_{i2} & \mathbf{H}_{i3} & \cdots & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \vdots \\ \mathbf{y}_i \end{bmatrix}$$

is well defined. This formula shows clearly that $\mathbf{v}_i = \mathcal{H}_i(\mathbf{y}_1, \ldots, \mathbf{y}_i) = \mathbf{H}_{i1}\mathbf{y}_1 + \cdots + \mathbf{H}_{i(i-1)}\mathbf{y}_{i-1} + \mathbf{y}_i$.

The matrices $R[\mathbf{x}\mathbf{v}_j^T]$, $R[\mathbf{y}_i\mathbf{v}_j^T]$ and $R[\mathbf{v}_j\mathbf{v}_j^T]$ are assumed known. Estimation of these matrices is discussed in Section 5.3 of the book by Torokhti and Howlett [14].

Theorem 3: For $k = 1, \ldots, p-1$ the solution $\widetilde{G}_j$ to problems (4)–(5) is

$$\widetilde{G}_j = R[\mathbf{x}\mathbf{v}_j^T]R^\dagger[\mathbf{v}_j\mathbf{v}_j^T] + M_j(I - R[\mathbf{v}_j\mathbf{v}_j^T]R^\dagger[\mathbf{v}_j\mathbf{v}_j^T]) \quad (30)$$

where M_j is arbitrary for each $j = 1, \ldots, k-1$ and $\mathbf{v}_j = \mathcal{H}_j(\mathbf{y}_1, \ldots, \mathbf{y}_j)$ has been determined for each $j = 1, 2, \ldots, k$ by Lemma 2. The associated error is

$$\|\mathbf{x} - \mathbf{x}_k\|^2_{K,\Omega} = \text{tr}\left\{ R[\mathbf{x}\mathbf{x}^T] - \sum_{j=1}^{k-1} R[\mathbf{x}\mathbf{v}_j^T]R[\mathbf{v}_j\mathbf{v}_j^T]^\dagger R[\mathbf{v}_j\mathbf{x}^T] \right\}.$$

Proof We write $\delta_k = \|\mathbf{x} - \mathbf{x}_k\|^2_{K,\Omega}$. We have $\|\mathbf{x}\|^2_{K,\Omega} = \text{tr}\{R[\mathbf{x}\mathbf{x}^T]\}$. Therefore, for $k = 1, \ldots, p$,

$$\begin{aligned} \delta_k &= \text{tr}\left\{ R\left[\left(\mathbf{x} - \sum_{j=1}^{k-1} \mathcal{G}_j(\mathbf{v}_j)\right) \left(\mathbf{x} - \sum_{j=1}^{k-1} \mathcal{G}_j(\mathbf{v}_j)\right)^T \right] \right\} \\ &= \text{tr}\left\{ R[\mathbf{x}\mathbf{x}^T] - \sum_{j=1}^{k-1} \left(R[\mathbf{x}\mathbf{v}_j^T]G_j^T + G_j R[\mathbf{v}_j\mathbf{x}^T] \right) \right. \\ &\quad \left. + R\left[\left(\sum_{j=1}^{k-1} \mathcal{G}_j(\mathbf{v}_j)\right) \left(\sum_{j=1}^{k-1} \mathcal{G}_j(\mathbf{v}_j)\right)^T \right] \right\}. \end{aligned} \quad (31)$$

In (31),

$$R\left[\left(\sum_{j=1}^{k-1} \mathcal{G}_j(\mathbf{v}_j)\right) \left(\sum_{j=1}^{k-1} \mathcal{G}_j(\mathbf{v}_j)\right)^T \right] = \sum_{j=1}^{k} G_j R[\mathbf{v}_j\mathbf{v}_j^T]G_j^T$$

owing to the orthogonality of vectors $\mathbf{v}_1, \mathbf{v}_2 \ldots, \mathbf{v}_k$. Thus,

$$\begin{aligned} \delta_k &= \|R^{1/2}[\mathbf{x}\mathbf{x}^T]\|_F^2 \\ &\quad - \sum_{j=1}^{k-1} \|R[\mathbf{x}\mathbf{v}_j^T](R[\mathbf{v}_j\mathbf{v}_j^T]^{1/2})^\dagger\|_F^2 + \sum_{j=1}^{k-1} J_j(G_j) \end{aligned} \quad (32)$$

where

$$J_j(G_j) = \|A_j - G_jC_j\|_F^2. \quad (33)$$

Because the terms $J_1(G_1)$, $\ldots$, $J_k(G_k)$ in the sum $\sum_{j=1}^k J_j(G_j)$ are determined independently we have

$$\min_{G_1,\ldots,G_k} \sum_{j=1}^{k} J_j(G_j) = \sum_{j=1}^{k} \min_{G_j} J_j(G_j). \quad (34)$$

Therefore, for $k = 1, \ldots, p-1$, the minimum in (4)–(5) is achieved if

$$A_j - G_jC_j = \mathbb{O} \;\Leftrightarrow\; (G_jR[\mathbf{v}_j\mathbf{v}_j^T] - R[\mathbf{x}\mathbf{v}_j^T])(R[\mathbf{v}_j\mathbf{v}_j^T]^{1/2})^\dagger = \mathbb{O}$$

and if we multiply on the right by $(R[\mathbf{v}_j\mathbf{v}_j^T]^{1/2})^\dagger R[\mathbf{v}_j\mathbf{v}_j^T]$ we can see that

$$G_jR[\mathbf{v}_j\mathbf{v}_j^T] - R[\mathbf{x}\mathbf{v}_j^T] = \mathbb{O}.$$

According to [1] a general solution to this equation is given by (30) if and only if $R[\mathbf{x}\mathbf{v}_j^T] = R[\mathbf{x}\mathbf{v}_j^T]R[\mathbf{v}_j\mathbf{v}_j^T]^\dagger R[\mathbf{v}_j\mathbf{v}_j^T]$. This is clearly true. Therefore the optimal value $\widetilde{G_j}$ is given by

$$\widetilde{G_j} = R[\mathbf{x}\mathbf{v}_j^T]R[\mathbf{v}_j\mathbf{v}_j^T]^\dagger + M_j(I - R[\mathbf{x}\mathbf{v}_j^T]R[\mathbf{v}_j\mathbf{v}_j^T]^\dagger)$$

where M_j is arbitrary for all $j = 1, 2, \ldots, k-1$. ■

Remark 4: We remind the reader that there is no rank restriction on the matrix operators during the filtration procedure. Thus Theorem 3 solves a matrix optimization problem with no rank restriction.

Define

$$A_j = R[\mathbf{x}\mathbf{v}_j^T]R[\mathbf{v}_j\mathbf{v}_j^T]^{1/2\dagger} \quad \text{and} \quad C_j = R[\mathbf{v}_j\mathbf{v}_j^T]^{1/2}.$$

and let

$$A_j = U_{A_j}\Sigma_{A_j}V_{A_j}^T \quad \text{and} \quad C_j = U_{C_j}\Sigma_{C_j}V_{C_j}^T \quad (35)$$

be the singular value decompositions for A_j and C_j respectively. We use a similar notation to that used in (8). To derive a solution to the problem (6)–(7) define

$$U_{C_j} = [U_{C_j1}\ U_{C_j2}] \quad \text{and} \quad V_{C_j} = [V_{C_j1}\ V_{C_j2}]$$

and let $\widetilde{A}_j = A_jV_{C_j}$ and $D_j = \text{diag}(\sigma_1(C_j), \ldots, \sigma_{t_j}(C_j)) \in \mathbb{C}^{t_j \times t_j}$ with $t_j = \text{rank}\, C_j$ and

$$K_j = (\widetilde{A}_{j1})_{r_j} D_j^{-1} Q_j U_{j2}^T (I - C_jC_j^\dagger)$$

where Q_j is arbitrary. The singular values of matrix A_j are denoted by $\sigma_k(A_j)$. The matrix

$$(A_j)_{r_j} = U_{A_jr_j}\Sigma_{A_jr_j}V_{A_jr_j}^T \quad (36)$$

is determined similarly to $(A)_k$ given by (10) with the replacement of A and k by A_1 and r_1, respectively.

Theorem 4: The solution to problem (6)–(7) is given by

$$\widehat{G}_j = (A_j)_{r_j} C_j^\dagger + K_j \tag{37}$$

and the associated error is

$$\|\mathbf{x} - \mathbf{x}_{p+1}\|^2_{K,\Omega} = \|R^{1/2}[\mathbf{x}\mathbf{x}^T]\|^2 - \sum_{j=1}^{p}\sum_{k=1}^{r_j}[\sigma_k(A_j)]^2. \tag{38}$$

Proof. We write $\epsilon = \|\mathbf{x} - \mathbf{x}_{p+1}\|^2_{K,\Omega}$. By an argument analogous to that which established (32) and (34) we have

$$\begin{aligned}\epsilon &= \|R^{1/2}[\mathbf{x}\mathbf{x}^T]\|_F^2 \\ &\quad - \sum_{j=1}^{p}\|R[\mathbf{x}\mathbf{v}_j^T](R[\mathbf{v}_j\mathbf{v}_j^T])^{1/2\dagger}\|_F^2 + \sum_{j=1}^{p} J_j(G_j)\end{aligned} \tag{39}$$

and

$$\begin{aligned}&\min_{\text{rank } G_j,\le r_j \forall j} \sum_{j=1}^{p}\|A_j - G_jC_j\|_F^2 \\ &\qquad = \sum_{j=1}^{p} \min_{\text{rank } G_j \le r_j} \|A_j - G_jC_j\|_F^2.\end{aligned} \tag{40}$$

By Corollary 1 the solution to

$$\min_{\text{rank } G_j \le r_j} \|A_j - G_jC_j\|_F^2$$

is attained when

$$G_j = (A_jP_{C_j,R})_{r_j}V_{C_j}^T C_j^\dagger + K_j = (A_j)_{r_j}C_j^\dagger + K_j = \widehat{G}_j.$$

We refer to [4] for more details. If $\widehat{G}_j$ is substituted instead of G_j in (39), then we have

$$\begin{aligned}\epsilon &= \|R^{1/2}[\mathbf{x}\mathbf{x}^T]\|_F^2 - \sum_{j=1}^{p}\|R[\mathbf{x}\mathbf{v}_j^T](R[\mathbf{v}_j\mathbf{v}_j^T])^{1/2}\dagger\|_F^2 \\ &\qquad + \sum_{j=1}^{p}\|A_j - [(A_j)_{r_j}C_j^\dagger + K_j]C_j\|_F^2 \\ &= \|R^{1/2}[\mathbf{x}\mathbf{x}^T]\|_F^2 - \sum_{j=1}^{p}\|A_j\|_F^2 + \sum_{j=1}^{p}\|A_j - (A_j)_{r_j}\|_F^2 \qquad (41) \\ &= \|R^{1/2}[\mathbf{x}\mathbf{x}^T]\|_F^2 - \sum_{j=1}^{p}\sum_{k=1}^{r_j}[\sigma_k(A_j)]^2 \qquad (42)\end{aligned}$$

since $(A_j)_{r_j} = (A_j)_{r_j}C_j^\dagger C_j$ by [14]. ■

Remark 5: In (38) the number p is the number of steps in the procedure described by (4)– (7). It follows from Theorem 4 that the error given by (38) decreases as p increases.

Remark 6: It follows from Theorems 3 and 4 that the components $\widetilde{G}_j$ and $\widehat{G}_j$ of the proposed filter are computed separately and require computation of $m\times n$ and $n\times n$ matrices for $\widetilde{G}_j$ and $m \times r_j$ and $r_j \times n$ matrices for $\widehat{G}_j$. To the best of our knowledge, a filter that is able to process infinite signal sets is not previously known. Therefore we compare the computational loads needed for our proposed filters with those for known filters in the case of finite sets of signals only. In this case most known nonlinear filters *[14]* require computation of much larger matrices than those mentioned above. The procedures presented in Theorems 3 and 4 are advantageous from a numerical point of view because they require much less computational work.

C. *Choice of operators $\mathcal{P}_k$ for (1)–(7)*

The purpose of using the operators $\mathcal{P}_k$ in our filter has been discussed in Section I-A2. Here, we illustrate the choice of $\mathcal{P}_k$ with some possible particular suggestions.

For instance, $\mathcal{P}_k$ could be chosen using the Hadamard product as in [12], [13] or in the form

$$[\mathcal{P}_k(\mathbf{x}_1,\ldots,\mathbf{x}_k)](\omega,\alpha) = \frac{1}{\gamma}\sum_{j=1}^{k}\gamma_j\mathbf{x}_j(\omega,\alpha)$$

where $\gamma = \sum_{j=1}^k \gamma_j$ and $\gamma_j \in \mathbb{R}$ is a constant.

Another possible choice for $\mathcal{P}_k$ is a time shifting operator similar to that used in [9]. For the case considered in Example 4 of Section VI, we define

$$\begin{aligned}\mathcal{P}_k(\mathbf{x}_1,\ldots,\mathbf{x}_k) &= \mathcal{P}_k[\mathbf{x}_k(\cdot,t_1),\ldots,\mathbf{x}_k(\cdot,t_N)] \\ &= [\mathbf{x}_k(\cdot,t_1-\Delta_1),\ldots,\mathbf{x}_k(\cdot,t_N-\Delta_N)]\end{aligned}$$

with $\Delta_k \in \mathbb{R}$ for all $k = 1,\ldots,p$.

Alternatively $\mathcal{P}_k$ could be chosen as a k-linear operator. In particular, $\mathcal{P}_k$ could be a k-linear integral operator. We cite [2] in this regard.

D. *Device for data compression and reconstruction*

Let us denote $D_j^{(1)} = U_{A_jr_j}\Sigma_{A_jr_j}$ and $D_j^{(2)} = V_{A_jr_j}^T C_j^\dagger$. See (35) and (36) in this regard. In (37) the matrix K_j is arbitrary. Let us assume that $K_j = \mathbb{O}$. Then the filter (1) based on solutions (30) and (37) to problems (4)–(5) and (6)–(7), respectively, is given by

$$F(y) = \sum_{j=1}^{p} D_j^{(1)}D_j^{(2)}v_j$$

where $v_j = H_jx_j$. Note that $D_j^{(1)} \in \mathbb{R}^{m\times r_j}$ and $D_j^{(2)} \in \mathbb{R}^{r_j\times n}$. Thus $D_j^{(2)}$ provides compression of a vector v_j with n components to a vector $D_j^{(2)}v_j$ with r_j components while $D_j^{(1)}$ provides decompression (reconstruction) from a vector with r_j components to a vector with m components. The compression ratio is given by

$$c = \frac{r}{\min\{m,n\}}, \quad \text{where } r = r_1 + \ldots + r_p. \tag{43}$$

We reiterate that m and n are the numbers of components in signals $\mathbf{x}$ and $\mathbf{y}$, respectively, and r is the number of components in the compressed data. By the condition (7), $r \le \min\{m,n\}$, i.e. the compressed data should be 'shorter' than $\mathbf{x}$ and $\mathbf{y}$. See also Section 1 in this regard. In Section 4 below, we consider examples of our filters with compression ratio $c = 25/116$ and $c = 9/116$.

E. Advantages associated with filtering based on scheme (4)–(7)

1) General advantages: The proposed filter consists of the two parts. The first one is the filtering of a random signal $\mathbf{y}$ based on the concatenation of solutions to $p-1$ unconstrained error minimization problems given by (4)–(5). This filter is important in its own right since it can be considered as a stand alone filter aimed at optimal filtering of stochastic signals.

The second part is based on the solution of the constrained error minimization problem given by (6)–(7). As a result, while this part is still filtering the input, it also compresses and decompresses the input. The latter procedure is realized via the solution of the problem (6)–(7) and is described in Section III-B.

The overall advantages of our proposed approach are the ability to determine a *single* optimal filter to process an *infinite* set of signals and the progressive decrease in the filtering error with an increase in the number of steps (4)–(7).

The mathematical advantages include simplification of the numerical procedure for determining the filter components $\widehat{\mathcal{G}}_1, \ldots, \widehat{\mathcal{G}}_k$ in (4)–(6) which we address in Remark 6 above, the rigorous theoretical justification of the results given in Sections III-A and III-B and the effective procedure for signal compression and decompression presented in Section III-D.

In the following Section, we consider other specific advantages by comparison with known filters.

2) Specific advantage: comparison with the known filters: As we have mentioned in Section I-A1, to the best of our knowledge, the previously known filters can process only finite sets of random signals. Our filter processes infinite sets of random signals. Therefore, a comparison between the known filters and the proposed one can only be done for the particular case when each set $\mathbf{x}$ and $\mathbf{y}$ (see Section II-A) consists of one signal only. In such a case, the results obtained above are presented in terms of the norm (46), $\|\cdot\|^2_\Omega$ (see Section VI below). The norm (46) is a particular case of the norm (3), $\|\cdot\|^2_{X,\Omega}$.

The errors associated with the Karhunen-Loève filter (KLF) [14] and the Wiener filter [14] follow from Theorems above if $p=1$, $k=1$ and the norm $\|\cdot\|^2_{X,\Omega}$ is replaced with the norm $\|\cdot\|^2_\Omega$. The errors presented in Theorems above are less than those for the KLF and Wiener filter if $p=2,3,\ldots$ and $k=2,3,\ldots$. Thus our filter provides improved accuracy when compared to the KLF and Wiener filter.

The compression ratio of the KLF is $\eta/\min\{m,n\}$ where η is the number of components in the compressed signal. Thus, the compression ratio (43) of our filter is better than that of the KLF if $r_1+\ldots+r_p<\eta$.

A comparison with other known filters [14] can be done in a similar way and leads to similar conclusions.

IV. Numerical examples and simulations

The following elementary Example 1 illustrates the proposed filter performance in the case of infinite sets of signals.

Example 1. Let

$$\mathbf{x}=\{\mathbf{x}(\omega,t)=[\omega t^2,\omega^2 t]^T \mid \omega\in[0,1], t\in[0,1]\}$$

and

$$\mathbf{y}=\{\mathbf{y}(\omega,t)=[0.8\omega t^2,1.2\omega^2 t]^T \mid \omega\in[0,1], t\in[0,1]\}.$$

That is $\mathbf{x}$ and $\mathbf{y}$ are represented by infinite sets of signals. For such signals, the norm (3) is reduced to the particular case given by (48). Accordingly, matrices $R[\mathbf{x}\mathbf{y}^T]$ and $R[\mathbf{y}\mathbf{y}^T]$ given by (22)–(24) are determined in terms of the norm (48). For instance, the entries of $R[\mathbf{x}\mathbf{y}^T]$ are

$$\langle\mathbf{x}^{(i)},\mathbf{y}^{(j)}\rangle=\int_0^1\int_0^1\mathbf{x}^{(i)}(\omega,t)\mathbf{y}^{(j)}(\omega,t)d\omega dt.$$

Therefore

$$R[\mathbf{x}\mathbf{y}^T]=\begin{bmatrix}0.053 & 0.075\\ 0.050 & 0.080\end{bmatrix},\quad R[\mathbf{y}\mathbf{y}^T]=\begin{bmatrix}0.011 & 0.060\\ 0.060 & 0.096\end{bmatrix}$$

$$\text{and}\quad R[\mathbf{y}\mathbf{y}^T]^\dagger=\begin{bmatrix}-37.7358 & 23.5849\\ 23.5849 & -4.3239\end{bmatrix}.$$

Let us first consider the simplest case of our filter presented by (1) with $p=1$, $\mathcal{H}_1$ and $\mathcal{P}_1$ given by the identity, and $G_1=\widetilde{G}_1$ where $\widetilde{G}_1$ is determined by (30) with $M_1=\mathbb{O}$. Then for any $\omega\in[0,1]$ and $t\in[0,1]$, the estimate of signals $\mathbf{x}(\omega,t)$ by the considered particular case of our filter is given by

$$\tilde{\mathbf{x}}(\omega,t)=R[\mathbf{x}\mathbf{y}^T]R[\mathbf{y}\mathbf{y}^T]^\dagger[0.8\omega t^2,1.2\omega^2 t]^T. \qquad (44)$$

Note that (44) does not coincide with an estimate by the generalized Wiener filter which can process a fixed random signal-vector [14], but is not able to process infinite sets of signals. Differences between (44) and an estimate by the generalized Wiener filter [14] are matrices $R[\mathbf{x}\mathbf{y}^T]$ and $R[\mathbf{y}\mathbf{y}^T]^\dagger$ determined in terms of the norm (47) used in (44), and the term $[0.8\omega t^2,1.2\omega^2 t]^T$.

Table 1 contains magnitudes of the error $\Delta_1=\|\mathbf{x}(\omega,t)-\tilde{\mathbf{x}}(\omega,t)\|^2_2$ with respect to some particular values of ω and t.

Next, let us now consider other simplest case of the proposed filter when in (1), as before, $p=1$, $\mathcal{H}_1$ and $\mathcal{P}_1$ are the identity, but $G_1=\widehat{G}_1$ where $\widehat{G}_1$ is determined by (37) with $r_1=1$ and $K_1=\mathbb{O}$. Then for any $\omega\in[0,1]$ and $t\in[0,1]$, the estimate of signals $\mathbf{x}(\omega,t)$ by this particular case of our filter is given by

$$\hat{\mathbf{x}}(\omega,t)=(A_1)_{r_1}C_1^\dagger[0.8\omega t^2,1.2\omega^2 t]^T \qquad (45)$$

where $A_1=R[\mathbf{x}\mathbf{y}^T]R[\mathbf{y}\mathbf{y}^T]^{1/2\dagger}$, $C_1=R[\mathbf{y}\mathbf{y}^T]^{1/2}$ and $(A_1)_{r_1}$ is determined by (36). Compression and decompression is realized by $(A_1)_{r_1}C_1^\dagger$ as those described in Section III-D above. For the same values of ω and t as those in Table 1, the error $\Delta_2=\|\mathbf{x}(\omega,t)-\hat{\mathbf{x}}(\omega,t)\|^2_2$ is worse than the error Δ_1 by approximately 20%. This is because $\hat{\mathbf{x}}(\omega,t)$ is determined with the truncated SVD given by (36).

The estimate $\hat{\mathbf{x}}(\omega,t)$ does not coincide with an estimate by the KLF [14] for reasons which are similar to those described above for $\tilde{\mathbf{x}}(\omega,t)$ when making comparison with the generalized Wiener filter.

Of course, in many instances, matrices $R[\mathbf{x}\mathbf{y}^T]$ and $R[\mathbf{y}\mathbf{y}^T]$ are unknown and should be estimated. Some estimation methods can be found in [14].

Example 2. Here, we consider a case where $\mathbf{x}$ and $\mathbf{y}$ are represented by finite signal sets and illustrate advantages of the proposed approach over the known methods. This case has been discussed briefly in Section I-A1.

Let $\mathbf{x} = \{\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(N)}\}$ and $\mathbf{y} = \{\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \cdots, \mathbf{y}_{(N)}\}$, where $N = 8$ and $\mathbf{x}_{(j)}, \mathbf{y}_{(j)} \in L^2(\Omega, \mathbb{R}^n)$ with $n = 116$ for each $j = 1, \ldots, 8$. Random vectors $\mathbf{y}_{(j)}$ and $\mathbf{x}_{(j)}$ are interpreted as an input of the filter and its 'idealistic' output, respectively. That is $\mathbf{x}_{(j)}$ is the signal that should be estimated by the filter. In this example, $\mathbf{x}_{(j)}$ and $\mathbf{y}_{(j)}$ are simulated as digital images presented by 116×256 matrices $X_{(j)}$ and $Y_{(j)}$, respectively. The columns of matrices $X_{(j)}$ and $Y_{(j)}$ represent a realization of signals $\mathbf{x}_{(j)}$ and $\mathbf{y}_{(j)}$, respectively.

Each picture $X_{(j)}$ in the sequence $X_{(1)}, \ldots, X_{(8)}$ has been taken at a certain time t_j with $j = 1, \ldots, 8$. Images $Y_{(1)}, \ldots, Y_{(8)}$ have been simulated from $X_{(1)}, \ldots, X_{(8)}$ in the form $Y_{(j)} = X_{(j)} \bullet \texttt{rand}_{(j)}$ for each $j = 1, \ldots, 8$. Here, $\bullet$ means the Hadamard product and $\texttt{rand}_{(j)}$ is a 116×256 matrix whose elements are uniformly distributed in the interval $(0, 1)$. Images $X_{(1)}, \ldots, X_{(8)}$ are presented in Fig. 1. Image $Y_{(3)}$ is given in Fig. 2 (a) as an example of images $Y_{(1)}, \ldots, Y_{(8)}$. Other images $Y_{(j)}$ with $j = 1, \ldots, 8$, $j \neq 3$ are similar.

For finite signal sets $\mathbf{x}$ and $\mathbf{y}$ considered in this example, the norm (47) is used. As a consequence, for $j = 1, 2$, matrix $R[\mathbf{x}\mathbf{v}_j^T]$ in (30) and (37) is presented by

$$R[\mathbf{x}\mathbf{v}_j^T] = \{r_{ik}\}_{i,q=1}^{116,256}$$

with

$$r_{iq} = \frac{1}{8}\sum_{k=1}^{8} \int_\Omega \mathbf{x}^{(i)}(\omega, t_k)\mathbf{v}_j^{(q)}(\omega, t_k) d\mu(\omega).$$

The matrix $R[\mathbf{v}_j\mathbf{v}_j^T]$ is determined similarly.

First, the simplest case of our filter defined from (4) by Theorem 3 with $p = 1$ has been applied to the considered signal sets. Then $R[\mathbf{x}\mathbf{v}_1^T] = R[\mathbf{x}\mathbf{y}^T]$ and $R[\mathbf{v}_1\mathbf{v}_1^T] = R[\mathbf{y}\mathbf{y}^T]$. In (1), G_1 is determined as $\widetilde{G}_1$ by (30). An estimate of $X_{(3)}$ by this filter is denoted by $\widetilde{X}_{1,(3)}$ and the filter itself is denoted by $\widetilde{\mathcal{F}}_1$.

To illustrate the performance of the proposed filters associated with different steps of the scheme (5)-(7), their related versions have also been applied to the given signal sets. Estimates of $X_{(3)}$ are denoted as follows:

- $\widetilde{X}_{2,(3)}$ is the estimate by $\widetilde{\mathcal{F}}_2$ defined by (5) and Theorem 3 with $p = 3$;
- $\widehat{X}_{2,(3)}$ is the estimate by $\widehat{\mathcal{F}}_2$ defined by (6)-(7) and Theorem 4 with $p = 2$, $r_1 = 10$ and $r_2 = 15$, with the compression ratio $c = 25/116$; and
- $\bar{X}_{2,(3)}$ is the estimate by $\bar{\mathcal{F}}_2$ defined by (6)-(7) and Theorem 4 with $p = 2$, $r_1 = 4$ and $r_2 = 5$, with the compression ratio $c = 9/116$.

We point out again that by the proposed method, *the same* filter is applied to each pair of signals $Y_{(j)}$ and $X_{(j)}$ for $j = 1, \ldots, 8$. That is the form presented by (1), (4)–(7) and Theorems 3, 4 is invariant with respect to different pairs $(Y_{(j)}, X_{(j)})$. In particular the matrices determined by (30) and (37) remain the same regardless of which pair $(Y_{(j)}, X_{(j)})$ is processed.

The known filters based on the Wiener filtering approach [14] are specifically constructed for *each pair* $(Y_{(j)}, X_{(j)})$. As a result, such filters incur much greater computational load. Therefore, by comparison with known filters, this important difference should be borne in mind. We denote by $\widetilde{X}_{2,(3)}$ and $X_{KLF,(3)}$ the estimate of $X_{(3)}$ by the Wiener filter and by the KLF of rank=25, respectively. The compression ratio of the KLF with rank=25 is $c = 25/116$.

In Fig. 2, estimates of signal $X_{(3)}$ by different filters are presented. Numerical results related to estimation of o signal $X_{(3)}$ are given in Tables 2 and 3. Results associated with estimates of other signals $X_{(j)}$ with $j = 1, \ldots, 8$, $j \neq 3$ are similar.

In particular, while the error associated with the filter $\widetilde{\mathcal{F}}_1$ is $\|X_{(3)} - \widetilde{X}_{1,(3)}\|^2 = 0.9893e + 07$, the error $\|X_{(3)} - \widetilde{X}_{2,(3)}\|^2$ associated with the filter $\widetilde{\mathcal{F}}_2$ is significantly lesser. This numerical illustration relates to Remark 5.

Next, we observe, that for the same compression ratio, $c = 25/116$, the error $\|X_{(3)} - \widehat{X}_{2,(3)}\|^2$ associated with the filter $\widehat{\mathcal{F}}_2$ is four times less than the error $\|X_{(3)} - X_{KLF,(3)}\|^2$ associated with the KLF.

It is interesting to compare compression ratios of those filters in the case when associated error are similar. In particular, it follows from the first and third columns of Table 3, that while the errors associated with the filter $\bar{\mathcal{F}}_2$ and the KLF are almost the same, the compression ratio of the filter $\bar{\mathcal{F}}_2$ (i.e. $c = 9/116$) is more than two times better that the KLF compression ratio (which is $c = 25/116$).

V. Conclusion

We have provided the theory for a new approach to filtering, compression and decompression for infinite sets of stochastic signals. Distinctive features of the proposed approach are as follows. While we consider processing infinite signal sets, the proposed filter is nevertheless fixed for all signal pairs. The filter is nonlinear and consists of p steps. Each step contains three components represented by three consecutive operations. The various operations are determined at each step by an iterative scheme designed to improve filter performance as the number of steps increases. Signal compression and decompression is provided by the final step of the scheme. The final step is based on a new method [4] for the best rank-constrained matrix approximation. The error associated with our filter decreases when the number of steps in the iterative scheme increases.

VI. Appendix A: Particular cases of the norm (3)

Here, we consider some particular norms that lead to the norm (3) used in our statement of the problem in Sections II-C–II-D.

Example 3. Let $\mathbf{x} \in L^2(\Omega, \mathbb{R}^m)$. Then we set

$$\|\mathbf{x}\|_\Omega^2 = \int_\Omega \|\mathbf{x}(\omega)\|_2^2 d\mu(\omega). \tag{46}$$

Note, that most of results related to Wiener-like optimal filtering have been obtained using the norm (46). Some relevant references can be found in [14].

Example 4. Let $\mathbf{x} = \{\mathbf{x}(\cdot, t_k) \in L^2(\Omega, \mathbb{R}^m) \mid t_k \in \mathbb{R} \ \forall\ k = 1, \ldots, N\}$. Thus, $\mathbf{x} \in L^2(\Omega \times [t_1, \ldots, t_N], \mathbb{R}^m)$. In other words, in the space $L^2(\Omega\times[t_1,\ldots,t_N],\mathbb{R}^m)$, $\mathbf{x}$ is a fixed signal, but in the space $L^2(\Omega,\mathbb{R}^m)$, $\mathbf{x}$ is represented by the set of signals $\{\mathbf{x}(\cdot,t_k)\in L^2(\Omega,\mathbb{R}^m) \mid t_k\in\mathbb{R}\ \forall\ k=1,\ldots,N\}$.

Let us put

$$\|\mathbf{x}\|^2_{[t_1,\ldots,t_N],\Omega}=\frac{1}{N}\sum_{k=1}^{N}\int_\Omega \|\mathbf{x}(\omega,t_k)\|_2^2 d\mu(\omega). \tag{47}$$

We note that $\|\mathbf{x}\|^2_{[t_1,\ldots,t_N],\Omega}$ can be represented as

$$\|\mathbf{x}\|^2_{[t_1,\ldots,t_N],\Omega}=\frac{1}{N}\sum_{k=1}^{N}\|\mathbf{x}(\cdot,t_k)\|^2_\Omega,$$

where

$$\|\mathbf{x}(\cdot,t_k)\|^2_\Omega=\int_\Omega \|\mathbf{x}(\omega,t_k)\|_2^2 d\mu(\omega).$$

The norm (46) follows from (47) if the set $\mathbf{x}$ consists of a single signal, $\mathbf{x}(\cdot,t_k)$, with t_k fixed.

Example 5. Let $\mathbf{x}=\{\mathbf{x}(\cdot,t)\in L^2(\Omega,\mathbb{R}^m)\mid t\in[a,b]\subset\mathbb{R}\}$. Thus $\mathbf{x}\in L^2(\Omega\times[a,b],\mathbb{R}^m)$. We set

$$\|\mathbf{x}\|^2_{[a,b],\Omega}=\frac{1}{b-a}\int_a^b\int_\Omega \|\mathbf{x}(\omega,t)\|_2^2 d\mu(\omega)dt. \tag{48}$$

Similarly to Example 4,

$$\|\mathbf{x}\|^2_{[a,b],\Omega}=\frac{1}{b-a}\int_a^b \|\mathbf{x}(\cdot,t)\|^2_\Omega dt.$$

The norm (46) follows from (48) if the set $\mathbf{x}$ consists of a single signal, $\mathbf{x}(\cdot,t)$, with t fixed.

Example 6. Let $\mathbf{x}=\{\mathbf{x}(\cdot,\tau)\in L^2(\Omega,\mathbb{R}^m)\mid \tau\in C^q\subset\mathbb{R}^q\}$ where C^q is a q-dimensional cube in $\mathbb{R}^q$. We put

$$\begin{aligned}\|\mathbf{x}\|^2_{C^q,\Omega}&=\frac{1}{V}\int_{C^q}\int_\Omega \|\mathbf{x}(\omega,\tau)\|_2^2 d\mu(\omega)d\tau\\&=\frac{1}{V}\int_{C^q}\|\mathbf{x}(\cdot,\tau)\|^2_\Omega d\tau,\end{aligned} \tag{49}$$

where $V=\int_{C^q}d\tau$. Then $\|\mathbf{x}\|^2_\Omega$ follows from $\|\mathbf{x}\|^2_{C^q,\Omega}$ if the set $\mathbf{x}$ consists of a single signal, $\mathbf{x}(\cdot,\tau)$, with τ fixed.

The norms given by (46)–(49) are generalized in the following way. Let $\mathbf{x}=\{\mathbf{x}(\cdot,\alpha)\in L^2(\Omega,\mathbb{R}^m)\mid\alpha\in K\}$ where K is a measurable set [7] with a measure $\lambda(\alpha)$. Thus, $\mathbf{x}$ is a single signal in the space $L^2(\Omega\times K,\mathbb{R}^m)$ and is the infinite set of signals $\{\mathbf{x}(\cdot,\alpha)\in L^2(\Omega,\mathbb{R}^m)\mid\alpha\in K\}$ in the space $L^2(\Omega,\mathbb{R}^m)$.

Let us set the norm by

$$\|\mathbf{x}\|^2_{K,\Omega}=\frac{1}{\lambda(K)}\int_K\int_\Omega \|\mathbf{x}(\omega,\alpha)\|_2^2 d\mu(\omega)d\lambda(\alpha), \tag{50}$$

where $\lambda(K)=\int_K d\lambda(\mathbf{x})$. Then the norms given by (46)–(49) follow from $\|\mathbf{x}\|^2_{K,\Omega}$ for a suitable choice of K and $\lambda(\alpha)$. In particular, $\|\mathbf{x}\|^2_\Omega$ follows from $\|\mathbf{x}\|^2_{K,\Omega}$ if X consists of a single signal only, $\mathbf{x}(\cdot,\alpha)$, with α fixed.

References

[1] T. L. Boullion and P. L. Odell, *Generized Inverse Matrices,* John Willey & Sons, Inc., New York, 1972.

[2] N. Dunford and J. T. Schwartz, *Linear Operators, Part 1, General Theory,* Wiley Classics Library, Wiley, New York, 1988.

[3] C. Eckart and G. Young, The Approximation of One Matrix by Another of Lower Rank, *Psychometrika,* **1**, 211-218, 1936.

[4] S. Friedland and A. P. Torokhti, Generalized rank-constrained matrix approximations, *SIAM J. Matrix Anal. Appl.,* **29**, issue 2, pp. 656-659, 2007.

[5] G.H. Golub and C.F. Van Loan, *Matrix Computation*, Johns Hopkins Univ. Press, 3rd Ed., 1996.

[6] S. Haykin, *Adaptive Filter Theory,* Prentice–Hall, Englewood Cliffs, N. J., 1991.

[7] V. Hutson and J.S. Pym, *Applications of Functional Analysis and Operator Theory,* Academic Press, London, 1980.

[8] I.T. Jolliffe, *"Principal Component Analysis,"* Springer Verlag, New York, 1986.

[9] J. Manton and Y. Hua, Convolutive reduced rank Wiener filtering, *Proc. of ICASSP'01,* **6**, pp. 4001-4004, 2001.

[10] V. J. Mathews and G. L. Sicuranza, *Polynomial Signal Processing,* J. Wiley & Sons, 2001.

[11] L.L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis,*New York: Addison-Wesley Publishing Co., 1990.

[12] A. Torokhti and P. Howlett, Optimal fixed rank transform of the second degree, *IEEE Trans. on Circuits and Systems. Part II, Analog & Digital Signal Processing,* **48**, 309–315, 2001.

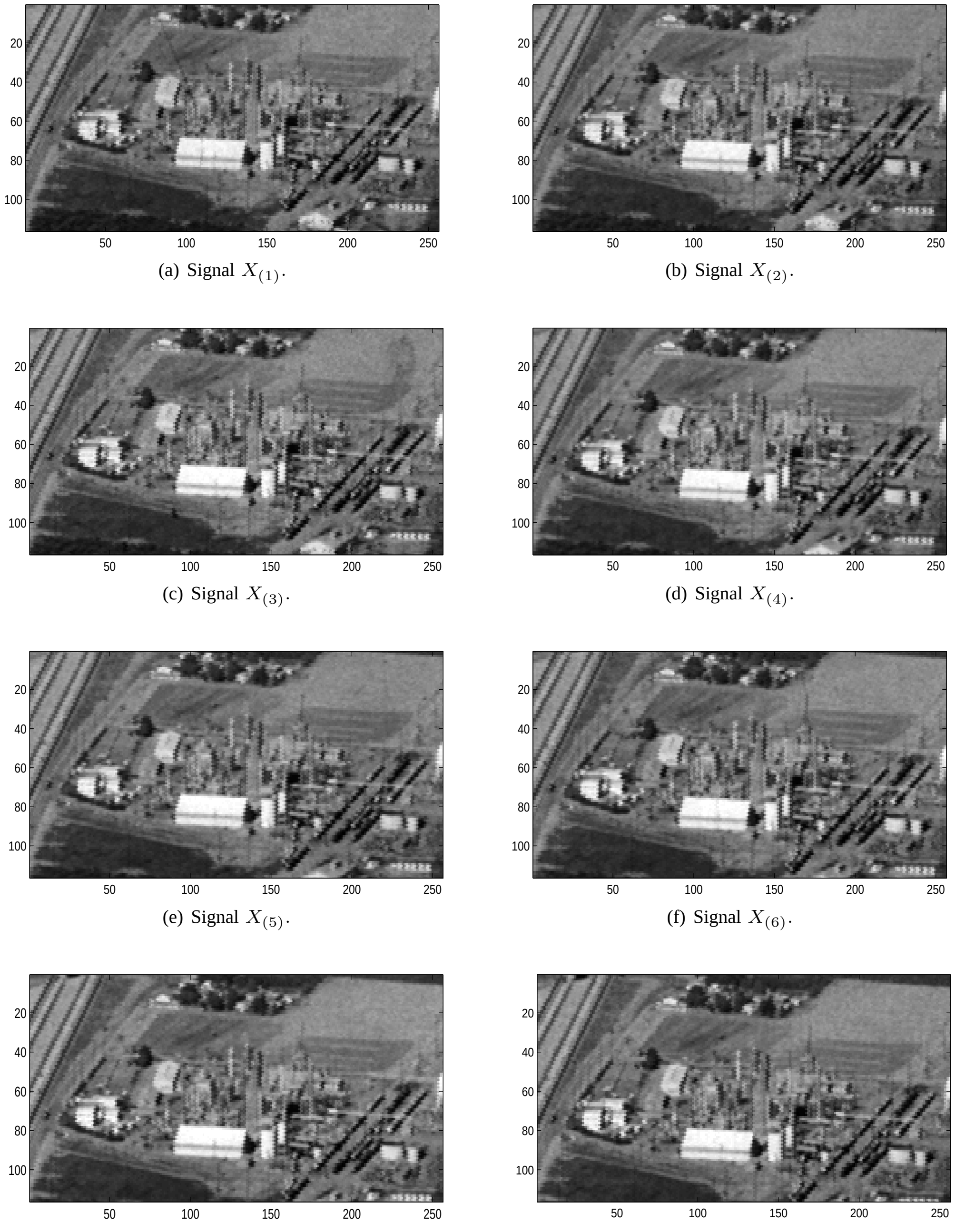
[13] A. Torokhti and P. Howlett, Method of recurrent best estimators of second degree for optimal filtering of random signals, *Signal Processing,* **83**, 5, 1013 - 1024, 2003.

[14] A. Torokhti and P. Howlett, *Computational Methods for Modelling of Nonlinear Systems,* Elsevier, 2007.

[15] N. Wiener, *The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications,* Academic Press, New York, 1949.

Anatoli Torokhti Anatoli Torokhti is an Associate Professor at the School of Mathematics and Statistics, University of South Australia, Mawson Lakes, SA, Australia. His research interests are in the area of mathematical modelling of nonlinear systems, mathematical signal processing, operator approxiamtion, mathematical ststaistics, and numerical methods for differential and integral equations.

Phil Howlett Phil Howlett is Professor of Industrial and Applied Mathematics at the University of South Australia. Professor Howlett is currently Chair of ANZIAM, the Australia and New Zealand Industrial and Applied Mathematics D ivision of the Australian Mathematical Society. Dr Howlett is the leader of the Scheduling and Control Group at the University of South Australia. The Group has invented several patented devices related to the optimal control of trains and has also worked extensively on solar-powered racing cars and electric vehicles. Dr Howlett is interested in a wide spectrum of applied mathematics including estimation of random signals, stochastic and deterministic control, operator approximation in Banach spaces and singular perturbations. Phil Howlett played Australian Rules Football for the South Adelaide Football Club in the South Australian National Football League from 1965 to 1974.



(a) Signal $X_{(1)}$.

(b) Signal $X_{(2)}$.

(c) Signal $X_{(3)}$.

(d) Signal $X_{(4)}$.

(e) Signal $X_{(5)}$.

(f) Signal $X_{(6)}$.

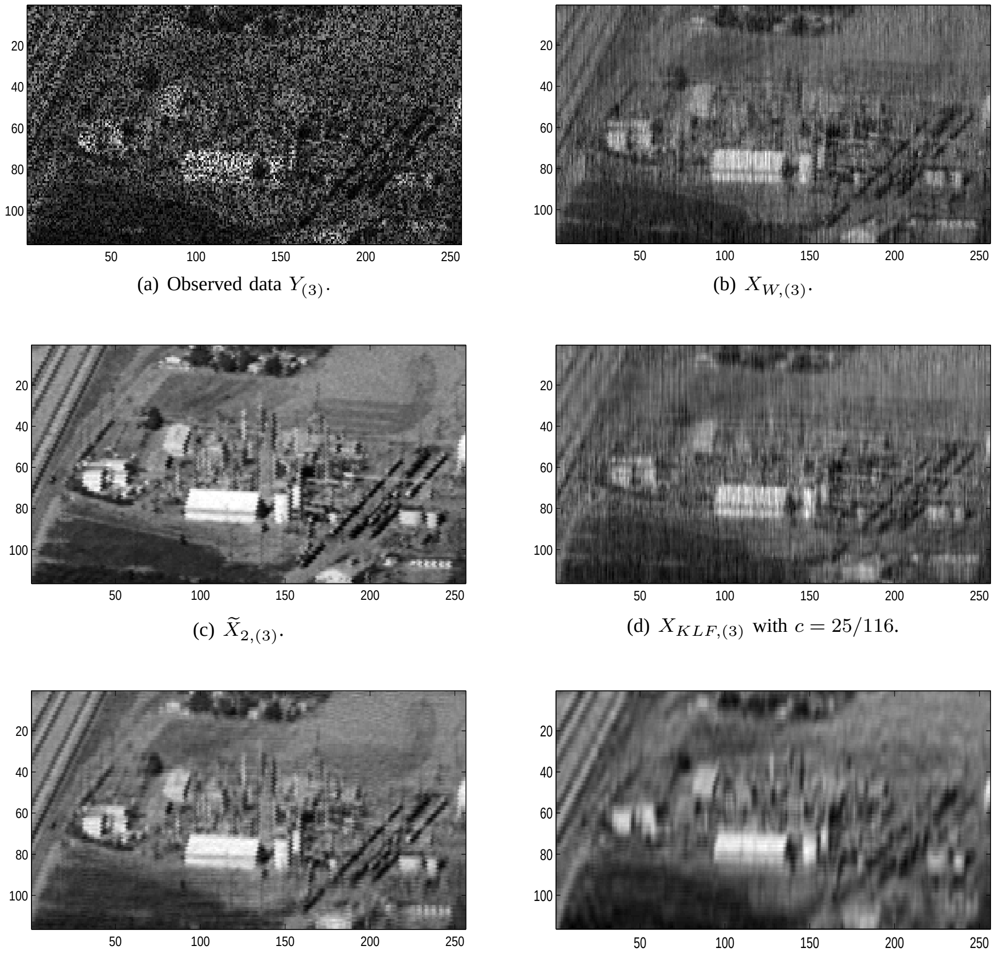
(g) Signal $X_{(7)}$.

(h) Signal $X_{(8)}$.

Fig. 1. Signals $X_1, \ldots, X_8$ to be estimated from observed data.

Table I
Magnitudes of error Δ_1 for some particular values of ω and t.

Δ_1	0.86×10^{-4}	1.16×10^{-2}	1.13×10^{-4}	1.35×10^{-2}	2.90×10^{-3}
ω	0.5	0.4	0.4	0.2	0.9
t	0.5	0.7	0.3	0.8	0.9



(a) Observed data $Y_{(3)}$.

(b) $X_{W,(3)}$.

(c) $\widetilde{X}_{2,(3)}$.

(d) $X_{KLF,(3)}$ with $c = 25/116$.

(e) $\widehat{X}_{2,(3)}$ with $c = 25/116$.

(f) $\bar{X}_{2,(3)}$ with $c = 9/116$.

Fig. 2. Illustration to signal $X_{(3)}$ estimation by different filters. Here, $X_{W,(3)}$ is the estimate by the Wiener filter, $\widetilde{X}_{2,(3)}$ is the estimate that follows from (5) with $p = 3$, $X_{KLF,(3)}$ is the estimate by KLF with $c = 25/116$, $\widehat{X}_{2,(3)}$ is the estimate that follows from (6)-(7) with $c = 25/116$, and $\bar{X}_{2,(3)}$ is the estimate that follows from (6)-(7) with $c = 9/116$.

Table II
Error estimations of signal $X_{(3)}$.
Here, $X_{W,(3)}$ is the estimate by the Wiener filter,
$\widetilde{X}_{2,(3)}$ is the estimate (5) with $p = 3$.

$\|X_{(3)} - X_{W,(3)}\|^2$	$\|X_{(3)} - \widetilde{X}_{2,(3)}\|^2$
$0.7555e + 07$	$1.5089e - 07$

Table III
Error estimations of signal $X_{(3)}$ by the KLT and (5) with $p = 3$,
for some different compression ratios c.

$\|X_{(3)} - X_{KLF,(3)}\|^2$ $c = 25/116$	$\|X_{(3)} - \widehat{X}_{2,(3)}\|^2$ $c = 25/116$	$\|X_{(3)} - \bar{X}_{2,(3)}\|^2$ $c = 9/116$
$8.8739e + 06$	$2.1730e + 06$	$8.2173e + 06$