

Reduced-Dimension Linear Transform Coding of Distributed Correlated Signals With Incomplete Observations

Hendra I. Nurdin, *Member, IEEE*, Ravi R. Mazumdar, *Fellow, IEEE*, and Arunabha Bagchi

Abstract—We study the problem of optimal reduced-dimension linear transform coding and reconstruction of a signal based on distributed correlated observations of the signal. In the mean square estimation context this involves finding the optimal signal representation based on multiple incomplete or only partial observations that are correlated. In particular, this leads to the study of finding the optimal Karhunen–Loève basis based on the censored observations. The problem has been considered previously by Gastpar, Dragotti, and Vetterli in the context of jointly Gaussian random variables based on using conditional covariances. In this paper, we derive the estimation results in the more general setting of second-order random variables with arbitrary distributions, using entirely different techniques based on the idea of innovations. We explicitly solve the single transform coder case, give a characterization of optimality in the multiple distributed transform coders scenario and provide additional insights into the structure of the problem.

Index Terms—Distributed signal processing, innovations, Karhunen–Loève transform, optimal linear estimation.

I. INTRODUCTION

WITH the advent of wide area sensor networks with a large number of spatially distributed sensors, the issue of transform coding, compression and reconstruction of correlated signals from incomplete observations is becoming increasingly important. More concretely, consider the situation of spatially distributed sensors that can only sense part of a given signal. The sensors are autonomous and have a limited energy supply. Furthermore, communication between sensors should be minimized to reduce expenses, except to relay information to some cluster node where the information is reconstructed from all the sensor observations.

Manuscript received June 14, 2007; revised July 01, 2008. Current version published May 20, 2009. This work was supported in part by Grants from the NSF through the ANI program while the second author was at Purdue and in by the NSERC through the Discovery Grant program. The material in this paper was presented in part at 16th Conference on Mathematics Theory of Networking Systems (MTNS 2004), Leuven, Belgium, July 2004.

H. I. Nurdin is with the Department of Information Engineering, College of Engineering and Computer Science, The Australian National University, Canberra, ACT 0200, Australia (e-mail: Hendra.Nurdin@anu.edu.au).

R. R. Mazumdar is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: mazum@ece.uwaterloo.ca).

A. Bagchi is with the Systems, Signals and Control group, Department of Applied Mathematics, University of Twente, Enschede, The Netherlands (e-mail: a.bagchi@math.utwente.nl).

Communicated by H. Bölcskei, Associate Editor for Detection and Estimation.

Color version of Figure 1 in this paper is available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TIT.2009.2018349

In this paper, we consider the problem where several groups of sensors are used to measure the correlated components of a distributed signal, but in which the groups of sensors cannot communicate with one another. Each group of sensors sends a reduced-dimension representation of its measurement to a central computer/reconstructor which then uses these distributed representations of the measurements to estimate the true value of the actual distributed signal. Our main concern is the issue of how the signals should be represented at the sensors so that one may produce an optimal linear estimate of the actual distributed signal at the central computer/reconstructor.

The problem has been considered by Gastpar, Dragotti, and Vetterli [1]–[4] in the context of compression and reconstruction of jointly Gaussian signals. It is well known that in the mean square distortion context, the Karhunen–Loève transform (KLT), which allows us to obtain the eigenvectors of the most significant eigenvalues of the covariance, is optimal from the point of view of compression (representing a signal in terms of the energy constraint) [5]–[7]. To address the distributed problem, Gastpar *et al.* initially introduced the concepts of partial, conditional, and combined partial-conditional KLT for the case of a single transform coder and reconstructor in [2]. Then they generalize their results to the multiple transform coder/reconstructor case and present an algorithm which they term the distributed Karhunen–Loève transform (DKLT).

The purpose of the present paper is to take a fresh look at these ideas and put them in a new light by the introduction of an appropriate Hilbert space framework (see [8], [9]). This has two advantages. First, we show that the estimation results obtained in [1], [2] for jointly Gaussian random variables are actually valid in the more general and important setting of second-order random variables with *arbitrary distributions* if we restrict our attention to the case where both the transform coding of the signal and its reconstruction are *linear*. Note that this is an important distinction from [1], [2] in which the Gaussian assumption on the signal leads to a linear reduced-dimension transform as the mean square optimal transformation over all possible reduced-dimension (linear and nonlinear) transforms, thus it is not required to restrict the class of transforms to be linear and also allows them to treat the rate-constrained distributed quantization problem for the signal. As discussed in more detail below, the motivation for restricting our attention to linear transform coders and reconstructors is to formulate a *tractable* problem under the *minimal assumptions* of the paper. Also, much of statistical signal theory in practice is based on linear estimators

and, in the absence of Gaussianity, the linear estimators studied in the paper are a particularly attractive option.

Let X be the random vector signal to be estimated, X_s the part of X that is sensed by sensors, X_{sc} the part of X that are not observed and Y an additional (random) side information that is available for reconstruction, and let T denote the transpose of a matrix. Then our paper assumes almost nothing on the signals to be transformed other than the following.

Assumptions 1: The signal $X = (X_s^T, X_{sc}^T)^T$ and the side information Y have finite second-order moments, the joint covariance matrix of X and Y are known with the covariance matrix of X_s being positive definite, and the elements of Y are linearly independent of the elements of X_s .

Let $\dim(X)$ denote the length of X and suppose that $X_s = (X_{s,1}^T, X_{s,2}^T, \dots, X_{s,n}^T)^T$, with $X_{s,i}$ taking values in \mathbb{R}^{M_i} for $i = 1, \dots, n$, and Y takes values in \mathbb{R}^r , and let $m_i < M_i$ for $i = 1, \dots, n$. Then the most general type of distributed reduced-dimension transform coding and reconstruction problem that one could consider in this setting is:

Problem 2: (Optimal Reduced-Dimension Transform Coding and Reconstruction Problem) Under Assumptions 1, find a pair of measurable functions $f : \mathbb{R}^{M_1} \times \mathbb{R}^{M_2} \times \dots \times \mathbb{R}^{M_n} \rightarrow \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \dots \times \mathbb{R}^{m_n}$ (the reduced-dimension transform) and $g : \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \dots \times \mathbb{R}^{m_n} \times \mathbb{R}^r \rightarrow \mathbb{R}^{\dim(X)}$ (the reconstructor) that minimizes the cost $C(f, g)$ given by

$$C(f, g) = \mathbb{E}[\|X - g(f(X_{s,1}, X_{s,2}, \dots, X_{s,n}), Y)\|^2].$$

In the above, \mathbb{E} denotes the expectation. Note that in the formulation of the above problem neither f nor g are restricted to be linear. To the best of our knowledge there is no known analytical or algorithmic solution to Problem 2. This remains true even if the problem is slightly modified, for instance by restricting f to be a linear map. One of the factors that makes this problem difficult is the minimality of Assumptions 1 with which to work with. For instance, in this setting there is no general expression for $\mathbb{E}[X|\sigma(Y)]$, the conditional expectation of X given the sigma algebra $\sigma(Y)$ generated by Y , which is of course the unique best mean square estimate of X in the space of all $\sigma(Y)$ -measurable square integrable random vectors. Therefore, to obtain a tractable problem, in the paper we study a specialized problem in which f and g are both restricted to be linear. Moreover, we emphasize that even under the much stronger assumption that X and Y are jointly Gaussian as considered in [1], a general optimal solution to distributed reduced-dimension transform coding and reconstruction is not known for $n > 1$ (the multiple transform coder case).

Since this paper assumes no knowledge of the probability distribution of the signal, it focuses only on the problem of distributed estimation of signals and does not consider the problem of rate-constrained quantization of the signal. The estimation results of the paper are derived independently of the distribution by using an approach based on the idea of *innovations*, as opposed to the conditional covariance approach of [1], [2] that does depend on the joint Gaussianity of the random variables of interest. Secondly, it allows us to pose the distributed reduced-dimension transform coding and estimation problem more precisely and exposes the underlying geometric

structure very clearly. In this general setting, we derive a unifying theorem for the single transform coder scenario with and without side information (Theorem 14), analyze the multiple transform coder scenario and independently prove convergence of the DKLT algorithm (Lemma 25).

The organization of this paper is as follows. In Section II, we recall some basic facts from linear estimation that will be used in the sequel, and define some operators of interest. In Section III, we discuss the single transform coder-reconstructor case to show the basic structure of the problem. In Section IV, we consider the general multiple transform coders-reconstructor problem, derive necessary conditions for construction of an optimal linear estimate at the reconstructor and study convergence of the DKLT algorithm for signals with finite second moments. In order to focus on the main ideas, proofs for results of the paper are all collected in the Appendix. Finally, in Section V, we offer some concluding remarks.

II. PRELIMINARIES AND BASIC THEORY

In the following, we denote the covariance matrix of a random variable Z (which may be scalar or vector valued) as Σ_Z (i.e., $\Sigma_Z = \mathbb{E}[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^T]$) and the covariance matrix between X and Y (i.e., $\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T]$) as Σ_{XY} . Note that by vector we mean a column vector. All vectors and matrices considered in this paper will have real elements.

Let us now introduce a number of definitions that will be used throughout this paper.

Definition 3: For any $M \times N$ matrix A and for any $m \leq M$, $R(m, A)$ is defined as the matrix consisting of the first m rows of A .

Definition 4: Let A be an $n \times n$ symmetric nonnegative definite matrix. An $m \times n$ matrix C is said to majorly diagonalize A if $CAC^T = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$ and $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq 0$.

Definition 5: Let X be an N -dimensional random vector having finite variance and let Σ_X be the covariance matrix of X . Then a unitary $N \times N$ matrix S such that $S\Sigma_X S^T$ majorly diagonalizes Σ_X is called a transposed eigenmatrix of Σ_X .

Definition 6: The set of all transposed eigenmatrices of a covariance matrix Σ is denoted by $\mathcal{T}(\Sigma)$.

Let \mathcal{H} denote the set of second-order scalar random variables (r.vs), i.e., all r.vs. X satisfying $\text{var}(X) < \infty$, and let \mathcal{H}_0 denote the set of elements of \mathcal{H} of zero mean. It is well known that \mathcal{H}_0 is a Hilbert space [9], [10], [8] and that given a collection of linearly independent r.vs. $X, Y_1, Y_2, \dots, Y_n \in \mathcal{H}_0$ then the best linear mean square estimate $\hat{f}(\cdot)$ of X given Y_1, \dots, Y_n is

$$\hat{f}(Y) = \pi[X|Y] = \mathbb{E}[XY^T] \Sigma_Y^{-1} Y$$

where $Y = (Y_1, Y_2, \dots, Y_n)^T$. The r.v. $\pi[X|Y]$ is simply the unique orthogonal projection of X onto the subspace spanned by Y_1, Y_2, \dots, Y_n . It follows that the mean square error (MSE) is given by

$$\begin{aligned} MSE &= \text{var}(X - \hat{f}(Y)) \\ &= \text{var}(X) - \mathbb{E}[XY^T] \Sigma_Y^{-1} \mathbb{E}[YX^T] \end{aligned}$$

\mathcal{H}_0 is an infinite-dimensional Hilbert space. However, any finite collection of linearly independent elements of \mathcal{H}_0 forms a finite-dimensional subspace that can be visualized (up to dimension 3), or thought of, as vectors on a corresponding Euclidean space (for a more detailed discussion see, e.g., [9, Sec. 3.3]). For example, if X_1, X_2, \dots, X_n are n linearly independent elements of \mathcal{H}_0 then they span (by forming linear combinations of X_1, X_2, \dots, X_n) an n -dimensional vector space $H^n = \text{span}\{X_1, \dots, X_n\}$ that is isomorphic to an n -dimensional Euclidean space. Each element X_i can be represented as a vector in \mathbb{R}^n with an inner product $\langle X_i, X_j \rangle$ between two elements X_i and X_j given by $\langle X_i, X_j \rangle = \mathbb{E}[X_i X_j]$; the inner product can then be extended to any element of H^n by linearity. Any element of H^n is represented as a vector in \mathbb{R}^n formed by a linear combination of the respective Euclidean representations of X_1, X_2, \dots, X_n and the length of the Euclidean representation is the square root of the variance of the particular element of \mathcal{H}_0 .

A finite-length vector r.v. X with elements belonging to \mathcal{H} , is called a second-order vector r.v. or second-order random vector. For any second-order random vector X , we denote $\text{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^T (X - \mathbb{E}[X])] = \text{tr}(\Sigma_X)$ (here, tr denotes the trace of a matrix). If X, Y are two second-order random vectors and $\Sigma_{XY} = 0$ (a zero matrix of the corresponding size) then we say that X and Y are *uncorrelated* or *orthogonal* and denote this as $X \perp Y$.

For any two zero mean second-order random vectors $Y = (Y_1, Y_2, \dots, Y_p)^T$ and $X = (X_1, X_2, \dots, X_n)^T$ with $\Sigma_Y > 0$, the best linear mean square estimate $\hat{f}(\cdot)$ of X given Y is [9], [10], [8]

$$\hat{f}(Y) = \pi[X|Y] = \Sigma_{XY} \Sigma_Y^{-1} Y$$

where the zero mean second-order random vector $\pi[X|Y]$ has elements that are the projections of the corresponding elements of X onto the linear subspace spanned by the elements of Y . Furthermore, the mean square estimation error is given by the formula

$$MSE = \text{var}(X - \hat{f}(Y)) = \text{tr}(\Sigma_X - \Sigma_{XY} \Sigma_Y^{-1} \Sigma_{YX}).$$

Note that if $U = X - \pi[X|Y]$ then $U \perp Y$, meaning that every component of U is orthogonal to every component of Y . If $\Sigma_Y \geq 0$ but is nonzero and singular then projections can still be defined as follows. Since Σ_Y is singular and nonzero, some elements of Y , say $Y_{i_1}, Y_{i_2}, \dots, Y_{i_r}$ ($i_k \in \{1, \dots, p\}$ and $i_{k+1} > i_k$), are linearly dependent on other elements of Y . Denote the remaining (linearly independent) elements of Y as $Y_{(R)} = (Y_{j_1}, Y_{j_2}, \dots, Y_{j_q})^T$ with $q = p - r$ and $j_{k+1} > j_k$. Then the projection $\pi[X|Y]$ is defined as $\pi[X|Y] = \pi[X|Y_{(R)}]$, which will depend only on the elements of $Y_{(R)}$. In this case we thus have that $\pi[X|Y] = \Sigma_{XY_{(R)}} \Sigma_{Y_{(R)}}^{-1} Y_{(R)}$ and this can always be written as $\pi[X|Y] = AY$ with A having zero columns corresponding to elements of Y which are not in $Y_{(R)}$ and its remaining columns being given by the corresponding columns of $\Sigma_{XY_{(R)}} \Sigma_{Y_{(R)}}^{-1}$.

Definition 7: Let $\pi[X|Y] = AY$. Then A is denoted as $\pi_{X|Y}$.

Remark 8: If X and Y are jointly Gaussian then $\pi[X|Y]$ coincides with the conditional expectation of X given F^Y , where F^Y is the σ -field generated by Y . As is well known, the conditional expectation is the minimum mean square estimate (MMSE) while the projection in general corresponds to the minimum linear mean square estimate.

A useful concept associated with the theory of zero mean second-order random vectors is that of the so-called Karhunen–Loève transform (KLT) [5]–[7], also known as principal component analysis (PCA) [11]. Given a zero mean second-order random vector X of length n , and a positive integer $m < \text{rank}(\Sigma_X) \leq n$, a zero mean second-order random vector Z of length m is said to be a **(standard) m -dimensional KLT of X** if it can be written as $Z = R(m, S)X$ for some $S \in \mathcal{T}(\Sigma_X)$. The elements of Z are mutually orthogonal and they span an m -dimensional subspace of \mathcal{H}_0 . By this we mean that $\{Y \in \mathcal{H}_0 \mid Y = p^T Z, p \in \mathbb{R}^m\}$ is an m -dimensional subspace of \mathcal{H}_0 . We have already mentioned an optimality property of the KLT in the introduction, but this property can be interpreted in a way which will be particularly useful for our purpose. This interpretation is as follows. Given any m -dimensional subspace of \mathcal{H}_0 , one has the orthogonal projection of X onto that subspace. The subspace spanned by the elements of an m -dimensional KLT of X has the special property that when X is projected onto that subspace then the mean square difference between X and the projection is minimum among all projections of X to all possible m -dimensional subspaces of \mathcal{H}_0 (see, e.g., [5]). Then we say that the elements of an m -dimensional KLT span an *optimal* m -dimensional subspace of X . This is an important fact and will be used in the proofs of some of our results.

Definition 9: For $m \leq \text{dim}(X)$, $\mathcal{M}_{KLT}(m, X) = \{A \in \mathbb{R}^{m \times \text{dim}(X)} : \exists S \in \mathcal{T}(\Sigma_X) \text{ s.t. } A = R(m, S)\}$. Any element of $\mathcal{M}_{KLT}(m, X)$ is called an m -dimensional Karhunen–Loève transform (KLT) matrix of X .

Definition 10: For $m \leq \text{dim}(X)$, $KLT(m, X) = \{Y \text{ is a zero mean second-order random vector: } \exists A \in \mathcal{M}_{KLT}(m, X) \text{ s.t. } Y = AX\}$. Any element of $KLT(m, X)$ is said to be an m -dimensional KLT of X .

III. SINGLE TRANSFORM CODER SCENARIOS

Let $X = (X_1, X_2, \dots, X_N)^T$ be the random vector being sensed where X has a known covariance matrix Σ_X . A *transform coder* senses a portion of X which we denote as $X_s = (X_1, X_2, \dots, X_M)^T$ with $M \leq N$. The section of X not being sensed, called the *hidden part*, is denoted as $X_{s^c} = (X_{M+1}, X_{M+2}, \dots, X_N)^T$. Note that $X = (X_s^T, X_{s^c}^T)^T$ and obviously

$$\Sigma_X = \begin{bmatrix} \Sigma_{X_s} & \Sigma_{X_s X_{s^c}} \\ \Sigma_{X_{s^c} X_s}^T & \Sigma_{X_{s^c}} \end{bmatrix}.$$

More generally, besides the observable part X_s of X , the transform coder may also have access to side information Y :

Definition 11: The side information Y is a second-order random vector possibly correlated with X_{s^c} (i.e., $\Sigma_{X_{s^c} Y} \neq 0$)

and whose elements are linearly independent of the elements of X_s .

The transform coder's function is to transform the data vector X_s (of length M) into a smaller vector Z_s (of length $m < M$). The information from the transform coder (i.e., Z_s) as well as the side information Y are both sent to a reconstructor which uses them to construct a random vector $\hat{X} = (\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N)^T = (\hat{X}_s^T, \hat{X}_{s^c}^T)^T$ as an estimate of X .

Remark 12: Throughout this section we assume that $\mathbb{E}[X] = 0$ and $\Sigma_{X_s} > 0$. However, the results here also apply to the case where $\mathbb{E}[X] \neq 0$ by applying them to the zero mean random vector $X_0 = X - \mathbb{E}[X]$ instead of X . Similarly, if $Y \neq 0$ (i.e., there is side information available) we assume that $\mathbb{E}(Y) = 0$.

This section focuses on the problem of how to choose Z_s as a reduced-dimension linear transform of X_s and how to construct \hat{X} optimally as a linear function of Z_s , in the mean square sense. By reduced-dimension linear transform of X_s , we mean that Z_s is of the form $Z_s = FX_s$ for some matrix $F \in \mathbb{R}^{m \times M}$. Introduce the shorthand notation $\pi[X|Y_1, Y_2, \dots, Y_n]$ for the projection of X onto the subspace spanned by the elements of Y_1, Y_2, \dots, Y_n . Before stating the main result of this section on single transform coders, let us first establish the following conventions for random vectors X, Y, Z :

- 1) $\pi[X|Z] = 0$ if $Z = 0$;
- 2) $\pi[X|Y, Z] = \pi[X|Y]$ if $Z = 0$.

Definition 13: The innovation U of X_s is $U = X_s - \pi[X_s|Y] = X_s - \pi_{X_s|Y}Y$.

U is the part of X_s that is uncorrelated with Y , therefore it cannot be linearly estimated using Y ; this is the reason it is referred to as the ‘‘innovation’’ of X_s . The innovation U plays a key role in determining an optimal reduced-dimension linear transform of X_s , in the sense of the following theorem:

Theorem 14: For a random vector Z , let the MSE be

$$MSE = \text{var}(\pi[X|X_s, Y] - \pi[X|Z, Y]). \quad (1)$$

Let $Z_{sb} = \bar{C}X_s$, with $\bar{C} = R(m, C)$, $C = R(M, S) \cdot \pi_{X|U}$ and where S is a transposed eigenmatrix of $\Sigma_{\pi[X|U]}$ (i.e., $S \in \mathcal{T}(\Sigma_{\pi[X|U]})$). Then

$$Z_{sb} = \underset{Z=FX_s, F \in \mathbb{R}^{m \times M}}{\text{arg min}} \text{var}(\pi[X|X_s, Y] - \pi[X|Z, Y]).$$

That is, Z_{sb} minimizes the MSE (1) over all random vectors Z of the form $Z = FX_s$ with $F \in \mathbb{R}^{m \times M}$.

Remark 15: We shall refer to the matrix \bar{C} defined in Theorem 14 as an optimal transformation matrix for the single transform coder.

For a proof of the theorem, see Appendix A. The main idea of the construction of Z_{sb} is the following. First, note that the pair (U, Y) contains the same information about X as the pair (X_s, Y) since they span the same subspace. However, since Y is already available at the reconstructor, the optimal linear strategy would be to send a random vector Z_s to the reconstructor that makes it possible to construct a mean square optimal linear estimate of $\pi[X|U]$, which is a linear transform of U and thus

uncorrelated with Y , at the reconstructor; it turns out that the vector $Z_s = \bar{C}X_s$ satisfies this requirement. Note that since S is a transposed eigenmatrix of $\Sigma_{\pi[X|U]}$ and $\pi[X|U] = \pi_{X|U}U$, by definition C majorly diagonalizes Σ_U and by inspection $\bar{C}U$ is an m -dimensional KLT of $\pi[X|U]$. Letting $v = \bar{C}U = Z_s - \bar{C}\pi[X_s|Y]$, we see that v can be reconstructed at the reconstructor once Z_s is received. However, since v is an m -dimensional KLT of $\pi[X|U]$, we see that $\pi[X|U]$ can be optimally estimated in the mean square by an appropriate linear function of v and this is exactly what is needed to achieve an optimal linear strategy. The optimal linear estimate of X corresponding Z_{sb} is then given by the following corollary.

Corollary 16: The corresponding optimal linear estimate $\hat{X}_{\text{opt}}(Z_{sb})$ of $\pi[X|X_s, Y]$ as a function of Z_{sb} and Y is given by:

$$\hat{X}_{\text{opt}}(Z_{sb}) = S^T \begin{bmatrix} \bar{C}U \\ 0_{N-m} \end{bmatrix} + \pi_{X|Y}Y = \pi[X|Z_{sb}, Y]$$

and the approximation error incurred is

$$\mathbb{E} \left[\left\| X - \hat{X}_{\text{opt}}(Z_{sb}) \right\|^2 \right] = \sum_{i=m+1}^M \lambda_i + \text{var}(V)$$

where $V = X - \pi[X|X_s, Y]$, $V \perp \text{span}\{X_s, Y\}$, while λ_i , $i \geq m + 1$ are the $M - m$ smallest eigenvalues of $\Sigma_{\pi[X|U]}$ after $N - M$ zero eigenvalues have been discarded.

For a proof of the corollary, see Appendix B. Note that the term $\sum_{i=m+1}^M \lambda_i$ in the corollary is the additional error due to the reduced-dimension transformation of X_s into Z_{sb} while $\text{var}(V)$ is the estimation error when X_s is also known perfectly (besides Y).

Continuing the discussion preceding the corollary, we have seen that using v the reconstructor can furnish a mean square optimal linear estimate $\pi[\widehat{X}|U]$ of $\pi[X|U]$ as a linear transform of U , as given in the corollary. Combining this with the information provided by Y (exploiting the property that $\pi[\widehat{X}|U] \perp Y$) the reconstructor can construct a mean square optimal linear estimate of X as $\hat{X}(Z_s) = \pi[X|\pi[\widehat{X}|U], Y] = \pi[X|\pi[\widehat{X}|U]] + \pi[X|Y]$. A visual representation of this geometric construction (cf. Section II) is depicted in Fig. 1 for the simplest case where $N = 3$, $M = 2$, $m = 1$ and $X = (X_{s,1}, X_{s,2}, X_{s^c})^T$, where $X_{s,1}$, $X_{s,2}$ and X_{s^c} are correlated second-order random variables, and there is side information Y available that is correlated with X_{s^c} and linearly independent of the elements of X . In order to visualize this simplest case, a four-dimensional Euclidean space is required (since $X_{s,1}, X_{s,2}, X_{s,3}$ and Y are linearly independent) with w_1, w_2, w_3, w_4 as its axes.

Theorem 14 unifies the various single transform coder scenarios that were first analyzed in [2], [3], and later as special cases in [1], under the stronger assumption that all signals are jointly Gaussian, namely the scenarios referred to as partial KLT (when only part of X is available to the transform coder and no side information is available, $Y = 0$), conditional KLT (when all of X is available to the transform coder and there is side information) and partial-conditional KLT (when only part of X is available to the transform coder, but there is also side information). Our theorem holds under the weaker assumption that

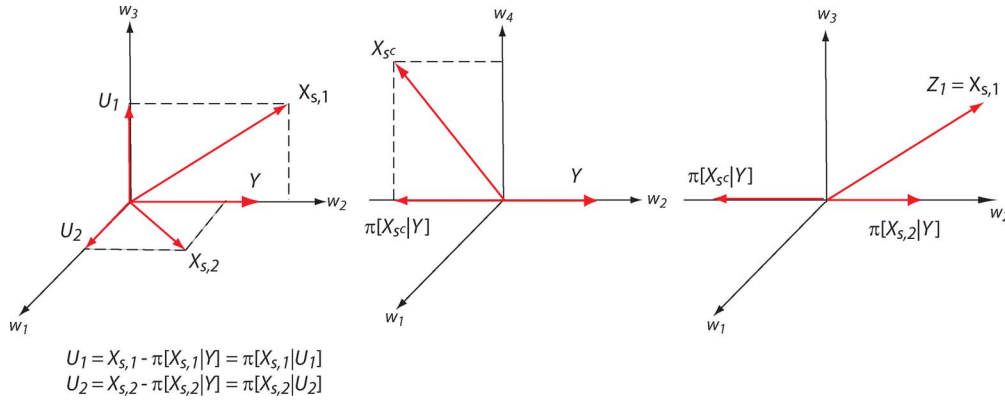


Fig. 1. Geometric visualization of an optimal reduced-dimension transform Z_s for $N = 3, M = 2$ and $m = 1, X = (X_{s,1}, X_{s,2}, X_{s,c})^T$ and side information Y . w_1, w_2, w_3, w_4 are the axes of the four-dimensional space \mathbb{R}^4 on which the elements of the random vectors can be represented. In this instance, $U_1 \perp U_2, X_{s,c} \perp \text{span}\{U_1, U_2\}$ and $\text{var}(U_1) > \text{var}(U_2)$, hence $\bar{C} = (1, 0, 0)$, $Z_s = X_{s,1}$ and the optimal linear estimate of X is $\hat{X}(Z_s) = \pi[X|Z_s, Y] = (X_{s,1}, \pi[X_{s,2}|Y], \pi[X_{s,c}|Y])^T$ as shown in the right hand most figure.

all signals have finite second-order moments, which contains jointly Gaussian signals as a special case, and is obtained by new proofs based on the idea of exploiting the innovation U as residual information.

Finally, notice that if the random vector Z_{sb} is a linear transform of X_s that minimizes the MSE (1) then so does any other linear transform Z_s of the form $Z_s = AZ_{sb}$ for some full rank $m \times m$ matrix A . This follows from the fact that the subspace spanned by Z_s and Y coincides with the subspace spanned by Z_{sb} and Y . In fact, all reduced-dimension linear transforms of X_s that minimize the MSE is of this form, as stated in the following corollary:

Corollary 17: Any random vector Z_s that minimizes (1) in the sense of Theorem 14 is of the form $Z_s = AZ_{sb}$ for some full rank $m \times m$ matrix A , where Z_{sb} is as defined in the theorem.

For a proof of the corollary, see Appendix C. We conclude this section on single transform coders by looking at some numerical examples.

Example 18: We use [2, Ex. 3]. Let $X = (X_1, X_2, X_3, X_4)$ with

$$\Sigma_X = \begin{bmatrix} 0.1 & 0 & 0.1 & 0.1 \\ 0 & 0.1 & 0.25 & 0 \\ 0.1 & 0.25 & 1 & 0.25 \\ 0.1 & 0 & 0.25 & 1 \end{bmatrix}$$

$X_{s,1} = (X_1, X_2)$ and $X_{s,2} = (X_3, X_4)$. $\Sigma_X > 0$ and has the positive eigenvalues

$$\{0.0244, 0.0900, 0.7914, 1.2942\}.$$

The sensed part of X is $X_{s,1}$ and the side information is just $X_{s,2}$, i.e., $Y = X_{s,2}$. We would like to produce a one-dimensional approximation of X . Using Theorem 14 we get the following optimal transformation matrices:

$$C = \begin{bmatrix} 0.9447 & -0.3280 \\ 0.3280 & 0.9447 \end{bmatrix}; \bar{C} = [0.9447 \quad -0.3280]$$

which is the same as the matrices reported in [2] except for the difference in sign, however as stated in Corollary 17 this difference is inconsequential. The optimal MSE that is computed is $MSE_{\text{opt}} = 0.0264$, that agrees with the value reported in [2].

Example 19: Let $X = (X_1, X_2, X_3, X_4, X_5, X_6)$ with

$$\Sigma_X = \begin{bmatrix} 1.3446 & -0.2448 & -0.2160 & -0.0114 & & \\ -0.2448 & 2.7902 & -0.6461 & 0.3244 & & \\ -0.2160 & -0.6461 & 2.0656 & -0.3812 & & \\ -0.0114 & 0.3244 & -0.3812 & 1.2775 & & \\ -0.6456 & 0.1772 & 0.2969 & -0.3034 & & \\ 0.7368 & -1.2198 & -0.4943 & 0.0331 & & \\ & & & & -0.6456 & 0.7368 \\ & & & & 0.1772 & -1.2198 \\ & & & & 0.2969 & -0.4943 \\ & & & & -0.3034 & 0.0331 \\ & & & & 1.4644 & -0.4189 \\ & & & & -0.4189 & 2.0577 \end{bmatrix}$$

$X_{s,1} = (X_1, X_2, X_3)$ and $X_{s,2} = (X_4, X_5, X_6)$. $\Sigma_X > 0$ and has the positive eigenvalues

$$\{0.5, 0.8, 1.2, 1.5, 3.0, 4.0\}.$$

The sensed part of X is $X_{s,1}$ and the side information is $Y = HX_{s,2}$ with $H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$. We would like to produce a two-dimensional approximation of X . Again using Theorem 14 we get the following optimal matrices:

$$C = \begin{bmatrix} 0.3027 & -1.0746 & 0.0864 \\ -0.4673 & 0.1893 & 1.0041 \\ 0.9330 & 0.2207 & 0.3189 \end{bmatrix}$$

$$\bar{C} = \begin{bmatrix} 0.3027 & -1.0746 & 0.0864 \\ -0.4673 & 0.1893 & 1.0041 \end{bmatrix}.$$

The optimal MSE that is computed is $MSE_{\text{opt}} = 2.0202$.

IV. MULTIPLE TRANSFORM CODER SCENARIOS

In this section we formulate the general distributed approximation problem with n -transform coders ($n > 1$) and connect it with previous work that has been done on this problem in [1]. To this end, let E_1, E_2, \dots, E_n be n transform coders that sense the vectors $X_{s,1}, X_{s,2}, \dots, X_{s,n}$, respectively. Let $\dim(X_{s,i}) = M_i$. Then $X_s = (X_{s,1}^T, X_{s,2}^T, \dots, X_{s,n}^T)^T$. Let the hidden part be $X_{s,c}$ and the side information Y be defined as before and let the output of E_1, E_2, \dots, E_n be denoted by $Z_{s,1}, Z_{s,2}, \dots, Z_{s,n}$, respectively, with $\dim(Z_{s,i}) = m_i < M_i$.

As elaborated in the introduction, for mathematical tractability, we assume that $Z_{s,i}$ is linearly related to $X_{s,i}$ and focus on the issue of finding an optimal linear solution (what is meant by optimal will be made clear in the formulation of Problem 21).

Remark 20: Following Section III, we assume that $\Sigma_{X_s} > 0$.

In the spirit of the problem solved in Theorem 14 for the single transform coder case, we formulate the following multiple transform coders estimation problem.

Problem 21: Let $\mathcal{I} = \{1, 2, \dots, n\}$. For any $i \in \mathcal{I}$, let $\mathcal{D}_{m_i \times M_i}$ denote the space of $m_i \times M_i$ ($m_i < M_i$) full row rank matrices. For any $(D_1, D_2, \dots, D_n) \in \mathcal{D}_{m_1 \times M_1} \times \mathcal{D}_{m_2 \times M_2} \times \dots \times \mathcal{D}_{m_n \times M_n}$ define:

$$\hat{X}_{D_1, D_2, \dots, D_n}(X_{s,1}, X_{s,2}, \dots, X_{s,n}, Y) = \pi[X|D_1 X_{s,1}, D_2 X_{s,2}, \dots, D_n X_{s,n}, Y].$$

Find $(D_1, D_2, \dots, D_n) \in \mathcal{D}_{m_1 \times M_1} \times \mathcal{D}_{m_2 \times M_2} \times \dots \times \mathcal{D}_{m_n \times M_n}$ that minimizes the **estimation error** defined by

$$\begin{aligned} \text{Err}(D_1, D_2, \dots, D_n) &= \text{var}(\pi[X|X_s, Y] - \\ &\quad \hat{X}_{D_1, D_2, \dots, D_n}(X_{s,1}, X_{s,2}, \dots, X_{s,n}, Y)). \end{aligned}$$

A solution to the above problem will be called an optimal linear solution to the n -transform coders distributed approximation problem.

Remark 22: In the formulation of Problem 21, we have explicitly assumed that there is side information Y available, we shall keep this assumption in our treatment of the problem. However, the case of no side information can be treated in an analogous manner simply by dropping the term Y wherever it is found.

An intuitive approach to solve Problem 21 is to set D_2, \dots, D_n arbitrarily, then proceeding to minimize Err one matrix at a time starting from D_1 and then D_2, \dots, D_n and starting over from D_1 until Err becomes relatively constant (i.e., the iteration has almost converged). This is the idea proposed in [1], by an algorithm called the DKLT (for distributed Karhunen-Loève transform) algorithm for jointly Gaussian signals. The algorithm was first introduced in [2], [3] without an explicit formulation of the associated optimization problem as we have done here and in the earlier works [1], [12]. The explicit problem formulation is particularly useful since it allows us to better understand the multiple transform coders scenario. Before stating the main results of the section, let us first describe the DKLT algorithm:

Algorithm 23 (DKLT):

- 1) Choose $(D_{2,0}, D_{3,0}, \dots, D_{n,0})$ arbitrarily from $\mathcal{D}_{m_2 \times M_2} \times \mathcal{D}_{m_3 \times M_3} \times \dots \times \mathcal{D}_{m_n \times M_n}$ and let $Z_{s,i}^{(0)} = D_{i,0} X_{s,i}$ for $\forall i \in \mathcal{I} \setminus \{1\}$.
- 2) Set $k = 1$.
- 3) Let $i = (k-1) \bmod n + 1$ and regard the collection of vectors $\{Z_{s,j}^{(k-1)}, j \in \mathcal{I} \setminus \{i\}\}$ and Y as side information for the transform coder E_i . If $D_{i,k}$ is already an optimal transformation matrix for E_i with the given side information

(cf. Corollary 17) keep it fixed, otherwise choose an optimal transformation matrix $D_{i,k}^*$ for E_i according to Theorem 14 (or Corollary 17) and set

$$D_{i,k} = D_{i,k}^*.$$

Then set

$$D_{j,k} = D_{j,k-1} \quad \forall j \in \mathcal{I} \setminus \{i\}$$

and

$$Z_{s,j}^{(k)} = D_{j,k} X_{s,j} \quad \forall j \in \mathcal{I}.$$

- 4) Repeat the procedure of step 3 sequentially for $k = 2, 3, \dots$, until the iterated transformation matrices remain constant after some iteration (the subscript k denotes iteration number) or if the transformation matrices are judged as no longer changing significantly.

The main contribution of this section is to show that based on the development of Section III, results on the DKLT algorithm obtained in [1] for jointly Gaussian signals can also be extended to distributed linear transform coding and reconstruction of signals with unknown probability distributions, but with finite second-order moments. Returning to Problem 21, we have the following characterization of optimal linear solutions:

Theorem 24 (Necessity): If D_1^*, \dots, D_n^* is a solution to Problem 21, then necessarily each transform coder E_i , $i = 1, 2, \dots, n$, must be linearly optimal as a single transform coder system with side information $\{\{D_j^* X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y\}$.

For a proof of the theorem, see Appendix D. The theorem is not analogous to [1, Corollary 8], but may be viewed as either a consequence of Theorem 14, or [1, Theorem 2] for the jointly Gaussian case. The key point here is the connection made in Theorem 24 between Problem 14 and the DKLT algorithm. The theorem explicitly states that if (D_1^*, \dots, D_n^*) is any solution to Problem 21 then necessarily D_i^* is an optimal transformation matrix for E_i , respectively. On the other hand, [1, Corollary 8] is a restatement of [1, Theorem 2] for the multiple transform coders scenario, stating how D_j can be chosen optimally given that all other transformation matrices $\{D_i, i \in \mathcal{I} \setminus \{j\}\}$, D_j are *fixed*; it serves as a precursor for the DKLT algorithm [1, Algorithm 1] in which round-robin optimizations of the D_j 's are performed.

Theorem 24 is quite intuitive because if there is a single transform coder that is not linearly optimal then one can change its transformation matrix, while keeping the transformation matrices of all remaining transform coders fixed, to lower the overall mean square error. It makes it clear that the DKLT algorithm is an obvious approach for obtaining transformation matrices satisfying the conditions of Theorem 24. However, the DKLT algorithm merely provides us with one set of transformation matrices that satisfy the necessary conditions for optimality. In general, the conditions of Theorem 24 are *not* sufficient for optimality. Establishing sufficiency is not easy since it is a nonlinear optimization problem in operator space with no readily usable convexity property. Despite this, in light of Theorem 14, we may however give an analogous

proof of the *local* convergence of the DKLT algorithm that generalizes [1, Theorem 12] to second-order random vectors and furthermore shows that convergence must be to a point satisfying the conditions of Theorem 24. Toward the end of this section, we will state the convergence lemma and present some numerical examples, but first we briefly digress with some remarks regarding the content of Theorem 24 in the context of rate-constrained quantization.

Although we do not treat the problem of distributed rate-constrained quantization, we point out that if the probability distributions of the signals are known (which we do not assume here) then the conclusion of Theorem 24 for distributed linear approximation does not hold in general when one considers the optimal distributed rate-constrained quantizer problem. That is, each optimal distributed rate-constrained quantizer will not in general be a singly optimal side information quantizer. However, there can be special cases where the principle of simultaneous singly optimal side information quantizers may lead to suboptimal yet simple and useful rate-constrained quantizers, such as in the case of jointly Gaussian signals as shown in [1].

Lemma 25: (Convergence of the DKLT algorithm for second-order random vectors) At consecutive iterations of Algorithm 23, the estimation error cannot increase, i.e.

$$\text{Err}(D_{1,k+1}, \dots, D_{n,k+1}) \leq \text{Err}(D_{1,k}, \dots, D_{n,k})$$

for all $k \geq 0$. Furthermore, the algorithm has converged at iteration k (i.e.,

$$\text{Err}(D_{1,m}, \dots, D_{n,m}) = \text{Err}(D_{1,k}, \dots, D_{n,k}), \forall m > k$$

if and only if $D_{1,k}, \dots, D_{n,k}$ satisfy the conditions of Theorem 24. In particular, if convergence has not been achieved at iteration k then a decrease in the estimation error always follows in the next $n - 1$ iterations, ensuring the convergence of the DKLT algorithm.

A proof of the lemma is given in Appendix E. As discussed above, based on Problem 21 it is clear that in general the necessary conditions need not be sufficient for optimality. The following example affirms this fact.

Example 26: Let X be as given in Example 19. Transform coder 1 senses $X_{s,1}$ while transform coder 2 senses $X_{s,2}$. We would like to produce an optimal linear approximation of X under the constraint that each transform coder may only send a two-dimensional vector.

Let us first apply the DKLT algorithm by setting

$$D_{2,0} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \quad (2)$$

At convergence the transformation matrices obtained are the following:

$$D_1^* = \begin{bmatrix} -0.2670 & 0.9004 & -0.4632 \\ -0.8597 & 0.2906 & 0.6917 \end{bmatrix}$$

$$D_2^* = \begin{bmatrix} -0.6661 & 0.7489 & 0.0753 \\ -0.4824 & -0.3828 & -0.8368 \end{bmatrix}$$

and the approximation error that is computed based on Theorem 14 (after 30 iterations) is

$$\text{var}(X - \hat{X}_{D_1^*, D_2^*}(X_{s,1}, X_{s,2})) = 1.9746.$$

Now, let us apply the DKLT once again but this time with a different initial condition. Thus, let

$$D_{2,0} = \begin{bmatrix} 2 & 3 & 0 \\ 0 & -4 & 3 \end{bmatrix}. \quad (3)$$

At convergence the transformation matrices obtained are the following:

$$D_1^* = \begin{bmatrix} -0.0603 & 0.7269 & -0.6849 \\ -0.5863 & 0.5463 & 0.6430 \end{bmatrix}$$

$$D_2^* = \begin{bmatrix} 0.0543 & -0.6715 & 0.9847 \\ 0.7606 & -0.4784 & -0.4607 \end{bmatrix}$$

and the approximation error that is computed is

$$\text{var}(X - \hat{X}_{D_1^*, D_2^*}(X_{s,1}, X_{s,2})) = 1.7567.$$

Thus with the DKLT algorithm, with different initial conditions, one can arrive at different points satisfying the necessary conditions of Theorem 24, but which result in different estimation errors. In this example, starting at (3) results in a lower estimation error than starting at (2).

The significance of the formulation of Problem 21 is that it gives us insight into what the DKLT algorithm accomplishes and that it does not in general guarantee global optimality. This observation was explicitly pointed out in [12] and subsequently in [1]. The explicit formulation of the objective function opens the possibility for finding or developing other optimization algorithms, besides the DKLT algorithm.

V. CONCLUDING REMARKS

In this paper we have shown in an explicit manner the geometric structure associated with the multiple transform coders-reconstructor problem in the estimation of correlated second-order random variables based on incomplete observations by the different transform coders and the reconstructor. In particular, our work extends the results of Gastpar, Dragotti and Vetterli [1]–[4] to the more general and important case of second-order random vectors. In the linear context this leads to a nice geometric interpretation in terms of the innovations and results in a nice decoupling property for transform coder-reconstructor pairs. However, the conditions are only necessary and the derivation of sufficient conditions is extremely difficult. Our geometric formulation of the distributed estimation problems suggests that it may be possible to develop other algorithms, besides the DKLT, for solving the problem. This could be the basis for some future investigations.

An important problem that has not been addressed in the literature is the fact of the assumption of the knowledge of the covariance structure of the observed correlated observations. In practice, one can expect that observations that are not far spatially are correlated but the covariance structure might be unknown. In that case one would need to estimate the covariance structure and then perform the KLT on the reduced structure for

which the procedure of the DKLT is not needed. These and other issues will be pursued elsewhere.

APPENDIX

PROOFS OF THEOREMS, LEMMATA, AND COROLLARIES

A. Proof of Theorem 14

We begin by writing

$$X = \pi[X|X_s, Y] + V = AX_s + BY + V$$

for some matrices A and B and a r.v. V that is orthogonal to the space spanned by X_s and Y . Then we may write: $X = \tilde{A}U + \tilde{B}Y + V$ where $\tilde{A} = A$ and $\tilde{B} = B + A\pi_{X_s|Y}$. The key observation is that since $U \perp Y$ and Y is known, all that remains is to find a reduced-dimension linear transform of $\tilde{A}U$ optimally in the mean square sense. To see this, suppose that W_s is an m -dimensional zero mean second-order random vector with $\Sigma_{W_s} > 0$ and $W_s \perp Y$ (thus W_s does not repeat “linear estimation information” already carried by Y). Then the best linear estimate of $\pi[X|X_s, Y]$ given W_s and Y is clearly

$$\begin{aligned} & \pi[\pi[X|X_s, Y]|W_s, Y] \\ &= \pi[\tilde{A}U + \tilde{B}Y|W_s, Y] \\ &= \pi[\tilde{A}U|W_s, Y] + \pi[\tilde{B}Y|W_s, Y] \\ &= \pi[\tilde{A}U|W_s] + \tilde{B}Y \text{ since } U \perp Y \text{ and } W_s \perp Y. \end{aligned}$$

and to minimize the quantity

$$\begin{aligned} \text{var}(\pi[X|X_s, Y] - \pi[\pi[X|X_s, Y]|W_s, Y]) \\ = \text{var}(\tilde{A}U - \pi[\tilde{A}U|W_s]) \end{aligned}$$

W_s must be chosen to be an m -dimensional KLT of $\tilde{A}U$ (see the discussion in Section II). Furthermore, with this choice of W_s it automatically follows that $V \perp \text{span}\{W_s, Y\}$ due to W_s being a linear transformation of U . To this end, let $S \in \mathcal{T}(\Sigma_{\tilde{A}U})$ then S majorly diagonalizes $\Sigma_{\tilde{A}U}$. Let us also define $C' = S\tilde{A}$, obviously C' majorly diagonalizes Σ_U . However, we have the following.

Lemma 27: Let A be a nonnegative symmetric $m \times m$ matrix. If K is an $n \times m$ ($n > m$) matrix which majorly diagonalizes A then the $n - m$ smallest eigenvalues of $KA K^T$ are zero.

Proof: The result follows from the fact that $\dim(\ker(K^T)) \geq n - m$. \square

Corollary 28: Let Σ_X be the covariance matrix of an m -dimensional zero mean random vector X and let B be an arbitrary $n \times m$ ($n > m$) matrix. If S is an $n \times n$ matrix which majorly diagonalizes $B\Sigma_X B^T$ and $\tilde{S} = R(m, S)$ then

$$SBX = \begin{bmatrix} \tilde{S}BX \\ 0_{n-m} \end{bmatrix}$$

and \tilde{S} majorly diagonalizes $B\Sigma_X B^T$.

Proof: By the previous lemma $SB\Sigma_X B^T S^T = (SB)\Sigma_X(SB)^T$ is diagonal with zeros on the lower $n - m$ diagonal. This implies that the lower $n - m$ elements of SBX are merely deterministic constants. Furthermore since $\mathbb{E}[X] = 0$ these constants are actually zero. Thus we may write

$$SBX = \begin{bmatrix} \tilde{S}BX \\ 0_{n-m} \end{bmatrix}.$$

From the above it is clear that \tilde{S} majorly diagonalizes $B\Sigma_X B^T$. \square

Thus we may write

$$C'U = \begin{bmatrix} CU \\ 0_{N-M} \end{bmatrix}$$

where $C = R(M, C')$ and it is obvious that $CU \in KLT(M, \tilde{A}U)$. If $\tilde{C} = R(m, C)$ then $\tilde{C}U \in KLT(m, \tilde{A}U)$ and it is a linear transformation on U . Now recall that $U = X_s - \pi_{X_s|Y}Y$, hence $\tilde{C}U = \tilde{C}X_s - \tilde{C}\pi_{X_s|Y}Y$. Since the second term on the right of the equality can be computed at the reconstructor (because Y is known), the transform coder only needs to send the remaining m -dimensional vector $Z_{sb} = \tilde{C}X_s$ so that $\tilde{C}U$ can be reconstructed exactly.

B. Proof of Corollary 16

For this proof we continue the arguments from Appendix A. Once the reconstructor receives Z_{sb} , $\tilde{A}U$ can be constructed approximately as \tilde{U} , which is defined as

$$\tilde{U} = S^{-1} \begin{bmatrix} \tilde{C}U \\ 0_{N-m} \end{bmatrix} = S^T \begin{bmatrix} \tilde{C}U \\ 0_{N-m} \end{bmatrix}.$$

It then follows that the mean square optimal linear estimate of $\pi[X|X_s, Y]$ is

$$\hat{X}_{\text{opt}}(Z_{sb}) = \tilde{U} + \pi[X|Y] = \tilde{U} + \pi_{X|Y}Y$$

since $Y \perp U$ and therefore $Y \perp \tilde{U}$. The approximation error incurred is

$$\begin{aligned} & E \left[\left\| X - \hat{X}_{\text{opt}}(Z_{sb}) \right\|^2 \right] \\ &= \text{var}(\tilde{A}U + \pi_{X|Y}Y + V - \tilde{U} - \pi_{X|Y}Y) \\ &= \text{var}(\tilde{A}U - \tilde{U}) + \text{var}(V) \\ &= \text{var}(\tilde{A}U - S^T \begin{bmatrix} \tilde{C}U \\ 0_{N-m} \end{bmatrix}) + \text{var}(V) \\ &= \text{var}(S\tilde{A}U - \begin{bmatrix} \tilde{C}U \\ 0_{N-m} \end{bmatrix}) + \text{var}(V) \\ &= \text{var}(\begin{bmatrix} CU \\ 0_{N-M} \end{bmatrix} - \begin{bmatrix} \tilde{C}U \\ 0_{N-m} \end{bmatrix}) + \text{var}(V) \\ &= \sum_{i=m+1}^M \lambda_i + \text{var}(V) \end{aligned}$$

where $\lambda_i, i \geq m + 1$ are the $M - m$ smallest eigenvalues of $\Sigma_{\tilde{A}U}$ after $N - M$ zero eigenvalues have been discarded.

Now, since $U \perp Y$ and $V \perp U$, it is clear that $\tilde{A}U = \pi[X|U]$. Next we show an important relation between \tilde{C} and \hat{X}_{opt} . First

observe that $S\pi[X|\bar{C}U] = S\Sigma_{XU}\bar{C}^T(\bar{C}\Sigma_U\bar{C}^T)^{-1}\bar{C}U$ and that $\bar{C}\Sigma_U\bar{C}^T$ is diagonal with positive entries (since $v = \bar{C}U$ has non-zero mutually orthogonal elements). Therefore

$$\bar{C}\Sigma_U\bar{C}^T = \text{diag}(d_1, \dots, d_m)^{-1} = \text{diag}\left(\frac{1}{d_1}, \dots, \frac{1}{d_m}\right)$$

where $d_1, \dots, d_m > 0$. It follows that

$$\pi[X|\bar{C}U] = \Sigma_{XU}\bar{C}^T \text{diag}\left(\frac{1}{d_1}, \dots, \frac{1}{d_m}\right)\bar{C}U.$$

Let $\tilde{A} = \pi_{X|U} = \Sigma_{XU}\Sigma_U^{-1}$. Since $\bar{C} = R(m, S)\tilde{A} = \begin{bmatrix} I_{m \times m} & 0_{m \times (N-m)} \end{bmatrix} S\tilde{A}$ we have

$$\begin{aligned} S\pi[X|\bar{C}U] &= S(\Sigma_{XU}\Sigma_U^{-1})\Sigma_U\bar{C}^T \text{diag}\left(\frac{1}{d_1}, \dots, \frac{1}{d_m}\right)\bar{C}U \\ &= S(\Sigma_{XU}\Sigma_U^{-1})\Sigma_U \left(\begin{bmatrix} I_{m \times m} & 0_{m \times (N-m)} \end{bmatrix} S\tilde{A} \right)^T \\ &\quad \text{diag}\left(\frac{1}{d_1}, \dots, \frac{1}{d_m}\right)\bar{C}U \\ &= \left(S\tilde{A}\Sigma_U\tilde{A}^T S^T \right) \begin{bmatrix} I_{m \times m} \\ 0_{(N-m) \times m} \end{bmatrix} \text{diag}\left(\frac{1}{d_1}, \dots, \frac{1}{d_m}\right)\bar{C}U \\ &= \left(S\Sigma_{\pi[X|U]}S^T \right) \begin{bmatrix} I_{m \times m} \\ 0_{(N-m) \times m} \end{bmatrix} \text{diag}\left(\frac{1}{d_1}, \dots, \frac{1}{d_m}\right)\bar{C}U \\ &= \text{diag}\left(d_1, d_2, \dots, d_m, \underbrace{0, \dots, 0}_{N-m \text{ times}}\right) \begin{bmatrix} I_{m \times m} \\ 0_{(N-m) \times m} \end{bmatrix} \\ &\quad \text{diag}\left(\frac{1}{d_1}, \dots, \frac{1}{d_m}\right)\bar{C}U \\ &= \begin{bmatrix} \bar{C}U \\ 0_{(N-m) \times m} \end{bmatrix}. \end{aligned}$$

Thus

$$\pi[X|\bar{C}U] = S^{-1} \begin{bmatrix} \bar{C}U \\ 0_{N-m} \end{bmatrix} = S^T \begin{bmatrix} \bar{C}U \\ 0_{N-m} \end{bmatrix}$$

and we immediately have that $\tilde{U} = \pi[X|\bar{C}U]$. Hence

$$\begin{aligned} \hat{X}_{\text{opt}}(Z_{sb}) &= S^T \begin{bmatrix} \bar{C}U \\ 0_{N-m} \end{bmatrix} + \pi_{X|Y}Y \\ &= \pi[X|\bar{C}U] + \pi[X|Y] \\ &= \pi[X|\bar{C}U, Y]. \end{aligned}$$

It is clear that there is a bijective linear relation between $(\bar{C}X_s, Y)$ and $(\bar{C}U, Y)$ (i.e., $(\bar{C}X_s, Y)$ can be retrieved from $(\bar{C}U, Y)$ and vice-versa) and that they both span the same subspace of \mathcal{H}_0 . Hence we have the desired result

$$\hat{X}_{\text{opt}}(Z_{sb}) = \pi[X|\bar{C}U, Y] = \pi[X|\bar{C}X_s, Y] = \pi[X|Z_{sb}, Y]. \quad \square$$

C. Proof of Corollary 17

Continuing the arguments from Appendices A and B, we observe that $Z_s = AZ_{sb}$ for any $m \times m$ invertible matrix A also minimizes the MSE (1) since $\pi[X|AZ_{sb}, Y] = \pi[X|Z_{sb}, Y]$ and in this case one simply constructs the random vector \tilde{U} as

$$\tilde{U} = S^T \begin{bmatrix} A^{-1} & 0_{m \times (N-m)} \\ 0_{(N-m) \times m} & 0_{(N-m) \times (N-m)} \end{bmatrix} \begin{bmatrix} A\bar{C}U \\ 0_{N-m} \end{bmatrix}.$$

We have now exhausted all possible linear solutions since at the key step of optimal reduced-dimension linear transformation of $\hat{A}U$ the only choices are precisely $\bar{C}U \in KLT(m, \hat{A}U)$ or $A\bar{C}U$ for any $m \times m$ invertible matrix $A \neq I_{m \times m}$. These choices correspond precisely to all zero mean second-order random vectors whose elements span the same subspace of \mathcal{H}_0 as the elements of $\bar{C}U$. \square

D. Proof of Theorem 24

Since $(D_1^*, \dots, D_i^*, \dots, D_n^*)$ is a solution to Problem 21, it is clear that

$$\begin{aligned} \text{Err}(D_1^*, D_2^*, D_3^*, \dots, D_n^*) \\ \leq \text{Err}(D_1^*, \dots, D_{i-1}^*, D_i, D_{i+1}^*, \dots, D_n^*), \\ \forall D_i \in \mathcal{D}_{m_i} \times M_i, i \in \mathcal{I} \end{aligned}$$

implying that

$$D_i^* = \arg \min_{D_i \in \mathcal{D}_{m_i} \times M_i} \text{Err}(D_1^*, \dots, D_{i-1}^*, D_i, D_{i+1}^*, \dots, D_n^*), \quad \forall i \in \mathcal{I}. \quad (4)$$

Let us note that since $X - \pi[X|X_s, Y]$ is orthogonal to $\pi[X|X_s, Y]$, it is also orthogonal to both $\pi[X|X_{s,i}, \{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y]$ and

$$\pi[X|D_i X_{s,i}, \{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y],$$

$\forall i \in \mathcal{I}$. On a similar note

$$X - \pi[X|X_{s,i}, \{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y]$$

is orthogonal to both $\pi[X|X_{s,i}, \{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y]$ and

$$\pi[X|D_i X_{s,i}, \{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y],$$

$\forall i \in \mathcal{I}$. Consequently, first we have that

$$\begin{aligned} \text{Err}(D_1, \dots, D_n) \\ = \text{var}(X - \pi[X|D_i X_{s,i}, \{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y]) - \\ \text{var}(X - \pi[X|X_s, Y]). \end{aligned} \quad (5)$$

Second, we have

$$\begin{aligned} \text{var}(X - \pi[X|D_i X_{s,i}, \{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y]) \\ = \text{var}(X - \pi[X|X_{s,i}, \{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y]) + \\ \text{var}(\pi[X|X_{s,i}, \{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y] - \\ \pi[X|D_i X_{s,i}, \{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y]) \\ = \text{var}(X - \pi[X|X_s, Y]) + \text{var}(\pi[X|X_s, Y] - \\ \pi[X|X_{s,i}, \{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y]) + \\ \text{var}(\pi[X|X_{s,i}, \{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y] - \\ \pi[X|D_i X_{s,i}, \{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y]). \end{aligned} \quad (6)$$

Therefore, from (5) and (6), we conclude that

$$\begin{aligned} \text{Err}(D_1, \dots, D_n) \\ = \text{var}(\pi[X|X_s, Y] - \\ \pi[X|X_{s,i}, \{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y]) + \end{aligned}$$

$$\begin{aligned} & \text{var}(\pi[X|X_{s,i}, \{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y] - \\ & \pi[X|D_i X_{s,i}, \{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y]) \quad i \in \mathcal{I}. \end{aligned} \quad (7)$$

Now, noting that the first term on the right-hand side of (7) is independent of D_i , it follows from Theorem 14 and(4) that E_i ($i = 1, 2, \dots, n$) must be a linearly optimal single transform coder having $\{D_j^* X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}$ and Y as side information. \square

E. Proof of Lemma 25

Let us consider some iteration step $k \geq 1$ and let $i = (k - 1) \bmod n + 1$. Let us also regard the collection of random vectors $\{Z_{s,j}^{(k-1)}, j \in \mathcal{I} \setminus \{i\}\}$ along with Y as side information for the transform coder E_i . Since $D_{i,k} = D_{i,k}^*$ by Theorem 14 (with $\hat{X}_{k,opt}(D_{i,k} X_{s,i})$ corresponding to $\hat{X}_{opt}(Z_s)$ in the theorem with the substitutions $Z_s \rightarrow D_{i,k} X_{s,i}$ and $Y \rightarrow \{\{Z_{s,j}^{(k-1)}, j \in \mathcal{I} \setminus \{i\}\}, Y\}$) we have

$$\hat{X}_{k,opt}(D_{i,k} X_{s,i}) = \pi \left[X|D_{i,k} X_{s,i}, \{Z_{s,j}^{(k-1)}, j \in \mathcal{I} \setminus \{i\}\}, Y \right].$$

Next let $r = k \bmod n + 1$. Since $D_{r,k+1} = D_{r,k+1}^*$, we analogously have

$$\begin{aligned} & \hat{X}_{k+1,opt}(D_{r,k+1} X_{s,r}) \\ & = \pi \left[X|D_{r,k+1} X_{s,r}, \{Z_{s,j}^{(k)}, j \in \mathcal{I} \setminus \{r\}\}, Y \right]. \end{aligned}$$

By (7) and Theorem 14, changing the transformation matrix of E_r from $D_{r,k}$ to $D_{r,k+1}^*$ at iteration $k+1$ while keeping all other matrices fixed (in particular, $D_{i,k+1} = D_{i,k}^*$) cannot result in a higher estimation error since $D_{r,k+1}^*$ is an optimal transformation matrix (cf. Remark 15) with E_r being the associated transform coder. In other words

$$\begin{aligned} & \text{var}(\pi[X|X_s, Y] - \hat{X}_{k+1,opt}(D_{r,k+1} X_{s,r})) \\ & \leq \text{var}(\pi[X|X_s, Y] - \hat{X}_{k,opt}(D_{i,k} X_{s,i})). \end{aligned}$$

However, since by definition

$$\begin{aligned} & \hat{X}_{k,opt}(D_{i,k} X_{s,i}) \\ & = \hat{X}_{D_{1,k}, \dots, D_{n,k}}(X_{s,1}, \dots, X_{s,n}, Y) \end{aligned}$$

and

$$\begin{aligned} & \hat{X}_{k+1,opt}(D_{r,k+1} X_{s,r}) \\ & = \hat{X}_{D_{1,k+1}, \dots, D_{n,k+1}}(X_{s,1}, \dots, X_{s,n}, Y) \end{aligned}$$

we conclude that

$$\text{Err}(D_{1,k+1}, \dots, D_{n,k+1}) \leq \text{Err}(D_{1,k}, \dots, D_{n,k}) \quad \forall k \geq 0.$$

Suppose that $D_{1,k}, \dots, D_{n,k}$ satisfy the conditions of Theorem 24 then each transform coder E_i ($i = 1, \dots, n$) is optimal as a single transform coder system with $\{\{D_j X_{s,j}, j \in \mathcal{I} \setminus \{i\}\}, Y\}$ as side information. This implies that no further sequential change of the transformation matrices can yield a lower estimation error (since any transformation matrices that are already optimal are kept fixed). Hence we may

set $D_{i,m} = D_{i,k}$ for $i = 1, 2, \dots, n$ and $\forall m > k$, and we have that $\text{Err}(D_{1,m}, \dots, D_{n,m}) = \text{Err}(D_{1,k}, \dots, D_{n,k}) \quad \forall m > k$. Conversely, if $D_{1,k}, \dots, D_{n,k}$ do not satisfy the conditions of Theorem 24 then at least one transform coder, excluding the transform coder that had just been optimized at step k , is not optimal. Thus we may reduce the estimation error by optimizing the first of those sub-optimal transform coders to be encountered in iterations $k + 1, k + 2, \dots, k + n - 1$, i.e., $\exists m \in \{1, \dots, n - 1\}$ such that $\text{Err}(D_{1,k+m}, \dots, D_{n,k+m}) < \text{Err}(D_{1,k}, \dots, D_{n,k})$. Therefore the algorithm has not converged at step k and a decrease in the estimation error always follows in $n - 1$ iterations after k . Finally, since Err is bounded from below by 0, it is clear that the decreasing property of Err whenever convergence has not been achieved guarantees that the DKLT algorithm always converges. \square

ACKNOWLEDGMENT

The authors thank the referees and the Associate Editor for their detailed comments that have helped to improve the quality of this paper.

REFERENCES

- [1] M. Gastpar, P. L. Dragotti, and M. Vetterli, "The distributed Karhunen-Loève transform," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5177–5196, 2006.
- [2] M. Gastpar, P.-L. Dragotti, and M. Vetterli, The Distributed Karhunen-Loève Transform EPFL-Lausanne, Switzerland, 2003, EPFL I&C Tech. Rep. IC/2003/12.
- [3] M. Gastpar, P.-L. Dragotti, and M. Vetterli, "The distributed, partial and conditional Karhunen-Loeve transforms," in *Proc. Data Compress. Conf. IEEE Comput. Soc.*, Mar. 2003.
- [4] M. Gastpar, P.-L. Dragotti, and M. Vetterli, "The distributed Karhunen-Loeve transform," in *Proc. 2002 IEEE Int. Workshop Multimed. Signal Process. IEEE Signal Process. Soc.*, Dec. 2002.
- [5] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.
- [6] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
- [7] R. J. Clarke, *Transform Coding of Images*. New York: Academic, 1985, Microelectron. Signal Process..
- [8] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Upper Saddle River, NJ: Prentice-Hall, 2000.
- [9] A. Bagchi, *Optimal Control of Stochastic Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
- [10] E. Wong and B. Hajek, *Stochastic Processes in Engineering Systems*. New York: Springer-Verlag, 1985.
- [11] J. E. Jackson, *A User's Guide to Principal Components*. New York: Wiley-Intersci., 1991.
- [12] H. I. Nurdin, R. R. Mazumdar, and A. Bagchi, "On the estimation and compression of distributed correlated signals with incomplete observations," in *Proc. 16th Conf. Math. Theory Netw. Syst. (MTNS 2004)*, Leuven, Belgium, Jul. 2004.

Hendra I. Nurdin (S'01–M'07) received the Sarjana Teknik (bachelor) degree in electrical engineering from the Institut Teknologi Bandung, Indonesia, in 1999, the M.Sc. degree in engineering mathematics from the University of Twente, Enschede, The Netherlands, in 2002, and the Ph.D. degree in engineering and information science from the Australian National University, Canberra, in 2007.

During 2007–2008, he was a Research Fellow with the Department of Engineering of the Australian National University. He is currently the recipient of an Australian Research Council (ARC) Discovery Project research grant and an ARC Australian Postdoctoral (APD) Fellow with the Department of Information Engineering of the Australian National University. His research interests include control and realization of quantum systems, classical control, stochastic approximation and modeling, and distributed sensor networks.

Ravi R. Mazumdar (M'83–SM'94–F'05) was born in Bangalore, India. He received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Bombay, in 1977, the M.Sc. DIC degree in control systems from Imperial College, London, U.K., in 1978, and the Ph.D. degree in systems science from the University of California, Los Angeles (UCLA), in 1983.

He is currently a University Research Chair Professor of Electrical and Computer Engineering (ECE) with the University of Waterloo, Waterloo, ON, Canada, and an Adjunct Professor of Electrical and Computer Engineering with Purdue University, West Lafayette, IN. He has served on the faculties of Columbia University, New York, and INRS-Telecommunications, Montreal, QC, Canada. He held a Chair in Operational Research and Stochastic Systems with the Department of Mathematics, University of Essex, Colchester, U.K., and from 1999 to 2005 was a Professor of Electrical and Computer engineering with Purdue University. He has held visiting positions and sabbatical leaves at UCLA; the University of Twente, The Netherlands; the Indian Institute of Science, Bangalore; and the Ecole Nationale Supérieure des Télécommunications, Paris, France. His research interests are in applied probability, stochastic analysis, optimization, and game theory with applications to wireless and wireline networks, traffic engineering, filtering theory, and mathematical finance.

Dr. Mazumdar is a Fellow of the Royal Statistical Society. He is a member of the working groups WG6.3 and 7.1 of the IFIP and a member of SIAM and the IMS. He is a recipient of the IEEE INFOCOM 2006 Best Paper Award and was runner-up for the Best Paper at INFOCOM 1998.

Arunabha Bagchi was born in Calcutta, India, in September 1947. He received the M.Sc. degree in applied mathematics from Calcutta University in 1969 and the M.S. and Ph.D. degrees in engineering from the University of California, Los Angeles (UCLA), in 1970 and 1974, respectively.

Since 1974, he has been with the University of Twente, Enschede, The Netherlands, where he is currently a Professor of Applied Mathematics (Chair in Stochastic Systems and Signals) and Professor of Finance and Accounting (Chair in Financial Engineering and Risk Management). He is the founder and head of the Financial Engineering Laboratory (FELab) of the University of Twente. His current research interest mainly lies in particle filtering, distributed sensor network, and financial engineering. He is the author of *Optimal Control of Stochastic Systems* (New York: Prentice-Hall Int., 1993) and *Stackelberg Differential Games in Economic Models* (LCCIS 64, Springer-Verlag, 1984).

Dr. Bagchi has been Associate Editor of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL and of *Automatica*. He was a Fulbright Scholar during 1998–1999 and has been a Visiting Professor with UCLA, SUNY Stony Brook, Northeastern University, and the Indian Statistical Institute.