On the Estimation and Compression of Distributed Correlated Signals with Incomplete Observations

Hendra I. Nurdin*, Ravi R. Mazumdar†, and Arunabha Bagchi*
{h.i.nurdin,a.bagchi}@math.utwente.nl, mazum@ecn.purdue.edu

*Department of Applied Mathematics, University of Twente,
P.O. Box 217, 7500 AE, Enschede, The Netherlands

†School of Electrical and Computer Engineering, Purdue University,
West Lafayette, IN 47907-1285, USA

Abstract

In this paper we study the problem of optimal compression and signal reconstruction 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 which are correlated. In particular this leads to the study of finding the optimal Karhunen-Loève basis based on the censored observations. We give a precise characterization of the necessary conditions with or without side information. We also provide new insights into the structure of the problem. In particular, we show that a recently proposed scheme provides estimates that satisfy only necessary conditions for optimality and hence can be sub-optimal.

AMS Subject Classification: 94A08, 94A12, 68P30

Keywords: Distributed signal processing, signal estimation and compression, Karhunen-Loéve transform

1 Introduction

With the advent of wide area sensor networks with a large number of spatially distributed sensors, the issue of compression, and reconstruction of signals from incomplete observations is coming into importance. More concretely, consider a 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.

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 compressed version of its measurement to a central computer/decoder which then uses these compressed and distributed measurements to estimate the true value of actual distributed signal. Our main concern is the issue of how the signals should be compressed at the sensors so that one may produce an *optimal linear estimate* of the actual distributed signal at the central computer/decoder.

This problem has been introduced in the information theory context by Wyner and Ziv in [11]. More recently, Gastpar, Dragotti, and Vetterli [5, 6, 4] have considered the natural setting of [11] in the context of compression and reconstruction of second order 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)[3, 8, 2]. To address the distributed problem, Gastpar et al[5] introduce the concepts of partial, conditional, and combined partial-conditional KLT based on the encoder/decoder information in the case of only one encoder and one decoder. They then generalize their results to the multiple encoder/decoder case and present an algorithm which they term the distributed Karhunen-Loève transform (DKLT). They provide numerical results to show convergence.

In this paper we re-visit these ideas and put them in a new light by the introduction of an appropriate Hilbert space framework (see [1, 9]). This allows us to pose the distributed compression and estimation problem more precisely and exposes the underlying geometric structure very clearly. The framework we offer here give more insight into the optimal choice of transformation matrices and provides us with a characterization of the necessary conditions for optimality in the multiple encoder/decoder case. We then prove the convergence of the DKLT in [5] to a solution that satisfies the necessary conditions. However, we show that the conditions are not sufficient to guarantee convergence to a true optimum and hence the DKLT can be sub-optimal.

The organization of this paper is as follows: In Section 2 we recall some basic facts from linear estimation that will be used in the sequel, and define some operators of interest. In Section 3 we discuss the single encoder-decoder case to show the basic structure of the problem. In Section 4 we consider the general multiple encoder-decoder problem and derive necessary conditions for construction of an optimal linear estimate at the decoder. We then provide an explicit proof of the convergence of the DKLT algorithm in [5] and show why it can be sub-optimal. In Section 5 we offer some concluding remarks.

2 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. $\Sigma_Z = \mathbb{E}\left[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^\top\right]$) and the covariance matrix between X and Y (i.e. $\mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^\top\right]$) as Σ_{XY}. Note that by vector we mean a column vector. All vectors and matrices are assumed to have real elements.

Definition 2.1 *For any $M \times N$ matrix A and for any $m < M$, $R(m, A)$ is defined as the matrix consisting of the first m rows of A.*

Definition 2.2 *Let A be an $n \times n$ symmetric non-negative definite matrix. An $m \times n$ matrix C is said to majorly diagonalize A if*

$$CAC^\top = \mathrm{Diag}(\lambda_1, \lambda_2, ..., \lambda_m)$$

and

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m \geq 0$$

Definition 2.3 *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 $S\Sigma_X S^\top$ majorly diagonalizes Σ_X is called a transposed eigenmatrix of Σ_X.*

Definition 2.4 *The set of all transposed eigenmatrices of a covariance matrix A is denoted by $\mathcal{T}(A)$.*

Remark 2.1 *Throughout the paper, the words "random variable" will often be abbreviated as r.v. and "random variables" as r.vs.*

Let $\mathcal{H}$ denote the set of second order scalar r.vs. (all r.vs. X satisfying $\mathrm{Var}(X) < \infty$) 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[1, 10, 9] and that given a pair of r.vs. $(X, Y) \in \mathcal{H}_0 \times \mathcal{H}_0$ with $\mathrm{Var}(X) > 0$ then the best linear mean square estimate $\hat{f}(\cdot)$ of Y given X is

$$\begin{aligned}\hat{f}(X) &= P[Y|X] \\ &= \mathbb{E}[YX]\,\mathrm{Var}(X)^{-1}X\end{aligned}$$

The r.v. $P[Y|X]$ is simply the unique projection of Y onto the subspace spanned by X. It follows that the minimum MSE is given by:

$$\begin{aligned}MSE &= \mathrm{Var}(Y - \hat{f}(X)) \\ &= \mathrm{Var}(Y) - \mathbb{E}[YX]\,\mathrm{Var}(X)^{-1}\mathbb{E}[YX]\end{aligned}$$

Definition 2.5 *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.*

Notation 2.1 *Throughout this paper, for any second order random vector X, we denote* $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^{\top}(X - \mathbb{E}[X])] = \mathrm{Tr}(Cov(X))$.

Definition 2.6 *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 it is denoted by $X \perp Y$.*

For any two zero mean second order random vectors Y and X with $\Sigma_X > 0$, the best linear mean square estimate $\hat{f}(\cdot)$ of Y given X is[1, 10, 9]

$$\begin{aligned}\hat{f}(X) &= P[Y|X] \\ &= \Sigma_{YX}\Sigma_X^{-1}X\end{aligned}$$

where the zero mean second order random vector $P[Y|X]$ has elements which are the projection of the corresponding elements of Y onto the linear subspace spanned by the elements of X. Furthermore, the mean square estimation error is given by the formula:

$$\begin{aligned}MSE &= \mathrm{Var}(Y - \hat{f}(X)) \\ &= \mathrm{Tr}\left(\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}\right)\end{aligned}$$

Remark 2.2 *Note that if $U = Y - P[Y|X]$ then $U \perp X$, meaning that every component of U is orthogonal to every component of X.*

Definition 2.7 *Let $P[Y|X] = AX$. Then A is denoted as $A_{Y|X}$.*

A useful concept associated with the theory of zero mean second order random vectors is that of the so-called Karhunen-Loève transform (KLT)[3, 8, 2], also known as principal component analysis (PCA)[7]. Given a zero mean second order random vector X of length n, and a positive integer $m < rank\,(\Sigma_X) \leq n$, 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 $\mathcal{H}_0$. We have already mentioned a property of the KLT in the introduction, but this property can be interpreted in 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 a projection of X onto that subspace. The subspace spanned by the elements of a 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 (see [3]). In other words, the elements of an m-dimensional KLT span an optimal m-dimensional subspace. This is an important fact and will be used in the proofs of some of our results.

Definition 2.8 *For $n \leq \dim(X)$, $\mathcal{M}_{KLT}(n, X) = \{A \in \mathbb{R}^{n \times \dim(X)} : \exists S \in \mathcal{T}(\Sigma_X)$ s.t. $A = R(n, S)\}$. Any element of $\mathcal{M}_{KLT}(n, X)$ is called an $n-$dimensional Karhunen Loève transform (KLT) matrix of X.*

Definition 2.9 *For $n \leq \dim(X)$, $KLT(n, X) = \{Y$ is a zero mean second order random vector: $\exists A \in \mathcal{M}_{KLT}(n, X)$ s.t. $Y = AX\}$. Any element of $KLT(n, X)$ is said to be an $n-$dimensional KLT of X.*

3 Single Encoder Scenarios

Let $X = (X_1, X_2, ..., X_N)^\top$ be the random vector being sensed where X has a known covariance matrix $\Sigma_X > 0$. The *encoder* senses a portion of X which we denote as $X_s = (X_1, X_2, ..., X_M)^\top$ with $M < N$. The section of X not being sensed, called the *hidden part*, is denoted as $X_{s^c} = (X_{M+1}, X_{M+2}, ..., X_N)^\top$. Note that $X = (X_s^\top, X_{s^c}^\top)^\top$ and obviously $\Sigma_X = \begin{bmatrix} \Sigma_{X_s} & \Sigma_{X_s X_{s^c}} \\ \Sigma_{X_s X_{s^c}}^\top & \Sigma_{X_{s^c}} \end{bmatrix}$. The encoder's function is to code the data vector X_s (of length M) into a smaller vector Z_s (of length $m < M$). The information from the encoder (i.e. Z_s) is then sent to a decoder which uses it to construct $\hat{X} = (\hat{X}_1, \hat{X}_2, ..., \hat{X}_N)^\top = (\hat{X}_s^\top, \hat{X}_{s^c}^\top)^\top$ as an estimate of X.

The problem we consider is how to construct Z_s such that $\hat{X}$ is optimal in a mean square sense under various scenarios.

Remark 3.1 *Throughout the paper we assume that $\mathbb{E}[X] = 0$ and $\Sigma_X > 0$. 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. Hence there will be no loss of generality.*

We shall consider the cases which are referred to as partial KLT, conditional KLT and combined partial-conditional KLT in [5], and show that, from a purely mathematical viewpoint, they can be subsumed into a single estimation framework.

3.1 Single encoder with no side information

In this particular scenario, the encoder senses X_s and produces Z_s. The decoder receives Z_s and produces $\hat{X}$, without further additional information. We formally state the problem below:

Problem 3.1 *Given a zero mean second order random vector X with $\Sigma_X > 0$, find a second order random vector Z_s of length $m < M$ with $\Sigma_{Z_s} > 0$ and linearly related to X_s (i.e. $Z_s = DX_s$ for some full row rank matrix $D \in \mathbb{R}^{m \times M}$) such that the quantity:*

$$\mathrm{Var}\left(P\left[X|X_s\right] - P\left[X|Z_s\right]\right)$$

is minimized.

Remark 3.2 *It can easily be seen that if there is a solution to Problem 3.1 then it is not unique. This is because if Z_s is a solution then AZ_s is also a solution since both random vectors span the same subspace of $\mathcal{H}_0$.*

We now start to provide a solution to the above problem. The optimal linear mean square estimate of X given X_s, which we denote by Y_s, is given by

$$\begin{aligned} Y_s &= P\left[X|X_s\right] \\ &= \begin{bmatrix} I_{M\times M} \\ A_{X_{s^c}|X_s} \end{bmatrix} X_s \end{aligned}$$

and we may write X as

$$\begin{aligned} X &= Y_s + V \\ &= \begin{bmatrix} I_{M\times M} \\ A_{X_{s^c}|X_s} \end{bmatrix} X_s + V \end{aligned}$$

where $V = X - P\left[X|X_s\right]$ and $V \perp X_s$. The encoder knows what X_s is, but for optimal linear reconstruction of $P\left[X|X_s\right]$ at the decoder, it should encode Y_s (since $P\left[X|X_s\right] = Y_s$) into an $m-$dimensional vector. Thus we may choose Z_s to be a standard $m-$dimensional KLT of Y_s, which is a linear transformation of Y_s. More concretely, we have the following:

Theorem 3.1 *Problem 3.1 has base solutions of the form:*

$$Z_{sb} = \bar{C}X_s$$

where $\bar{C} = R(m, C)$ and $C = R(M, S)A_{X|X_s}$ for some $S \in \mathcal{T}(\Sigma_{P[X|X_s]})$. Every other solution is given by:

$$Z_s = AZ_{sb}$$

where A is any arbitrary invertible $m \times m$ matrix.

The corresponding optimal estimate of $P\left[X|X_s\right]$ is given by:

$$\hat{X}_{opt} = S^\top \begin{bmatrix} \bar{C}X_s \\ 0_{N-m} \end{bmatrix} = P\left[X|Z_s\right]$$

and the approximation error incurred is:

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

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

Proof. We continue the argument prior to the statement of the theorem. Note that by the definition of S, C and $\bar{C}$, $Z_s = \bar{C}X_s$ is just a standard Karhunen Loéve transform of Y_s. From this we can optimally reconstruct $P[X|X_s]$ (in a mean square sense) at the decoder as[3, 8, 2]:

$$\hat{X}_{opt} = S^\top \begin{bmatrix} \bar{C}X_s \\ 0_{N-m} \end{bmatrix}$$

All that remains to be done is to verify that

$$\hat{X}_{opt} = P[X|Z_{sb}]$$

where $Z_{sb} = \bar{C}X_s$. To this end, note that:

$$SP\left[X|\bar{C}X_s\right] = S\Sigma_{XX_s}\bar{C}^\top(\bar{C}\Sigma_{X_s}\bar{C}^\top)^{-1}\bar{C}X_s$$

and that $\bar{C}\Sigma_{X_s}\bar{C}^\top$ is diagonal with positive entries (since $Z_{sb} = \bar{C}X_s$ has non-zero mutually orthogonal elements). Thus we write:

$$\begin{aligned}\bar{C}\Sigma_{X_s}\bar{C}^\top &= \mathrm{Diag}(d_1, \ldots, d_m)^{-1} \\ &= \mathrm{Diag}(\frac{1}{d_1}, \ldots, \frac{1}{d_m})\end{aligned}$$

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

$$P\left[X|\bar{C}X_s\right] = \Sigma_{XX_s}\bar{C}^\top \mathrm{Diag}(\frac{1}{d_1}, \ldots, \frac{1}{d_m})\bar{C}X_s$$

Let $\tilde{A} = A_{X|X_s} = \Sigma_{XX_s}\Sigma_{X_s}^{-1}$. 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:

$$\begin{aligned}SP\left[X|\bar{C}X_s\right] &= S\left(\Sigma_{XX_s}\Sigma_{X_s}^{-1}\right)\Sigma_{X_s}\bar{C}^\top \mathrm{Diag}(\frac{1}{d_1}, \ldots, \frac{1}{d_m})\bar{C}X_s \\ &= S\left(\Sigma_{XX_s}\Sigma_{X_s}^{-1}\right)\Sigma_{X_s}\left(\begin{bmatrix} I_{m\times m} & 0_{m\times(N-m)} \end{bmatrix} S\tilde{A}\right)^\top \mathrm{Diag}(\frac{1}{d_1}, \ldots, \frac{1}{d_m})\bar{C}X_s \\ &= \left(S\tilde{A}\Sigma_{X_s}\tilde{A}^\top S^\top\right)\begin{bmatrix} I_{m\times m} \\ 0_{(N-m)\times m} \end{bmatrix} \mathrm{Diag}(\frac{1}{d_1}, \ldots, \frac{1}{d_m})\bar{C}X_s \\ &= \left(S\Sigma_{E[X|X_s]}S^\top\right)\begin{bmatrix} I_{m\times m} \\ 0_{(N-m)\times m} \end{bmatrix} \mathrm{Diag}(\frac{1}{d_1}, \ldots, \frac{1}{d_m})\bar{C}X_s \\ &= \mathrm{Diag}(d_1, d_2, \ldots, d_M, \underbrace{0, \ldots, 0}_{N-M \text{ times}})\begin{bmatrix} I_{m\times m} \\ 0_{(N-m)\times m} \end{bmatrix} \mathrm{Diag}(\frac{1}{d_1}, \ldots, \frac{1}{d_m})\bar{C}X_s \\ &= \begin{bmatrix} \bar{C}X_s \\ 0_{(N-m)\times m} \end{bmatrix}\end{aligned}$$

Thus

$$SP\left[X|\bar{C}X_s\right] = \begin{bmatrix} \bar{C}X_s \\ 0_{(N-m)\times m} \end{bmatrix}$$

or equivalently

$$\begin{aligned}P\left[X|\bar{C}X_s\right] &= S^{-1}\begin{bmatrix} \bar{C}X_s \\ 0_{(N-m)\times m} \end{bmatrix} \\ &= S^\top\begin{bmatrix} \bar{C}X_s \\ 0_{(N-m)\times m} \end{bmatrix}\end{aligned}$$

Hence

$$\hat{X}_{opt} = P[X|Z_{sb}]$$

Finally, $Z_s = AZ_{sb}$ for any $m \times m$ invertible matrix A is also a solution since

$$P[X|AZ_{sb}] = P[X|Z_{sb}]$$

and in this case one simply constructs $\hat{X}_{opt}$ as:

$$\hat{X}_{opt} = S^{\top} \begin{bmatrix} A^{-1} & 0_{m\times(N-m)} \\ 0_{(N-m)\times m} & 0_{(N-m)\times(N-m)} \end{bmatrix} \begin{bmatrix} Z_s \\ 0_{N-m} \end{bmatrix}$$

We have now exhausted all possible linear solutions since at the key step of optimal linear compression of Y_s the only choices are precisely $\bar{C}Y_s \in KLT(m, Y_s)$ or $A\bar{C}Y_s$ for any $m \times m$ invertible matrix $A \neq I_{m\times m}$. This is because these choices correspond precisely to all zero mean second order random vectors whose elements span the same optimal subspace as the elements of $\bar{C}Y_s$ (see Section 2 on the KLT). ■

Remark 3.3 *Note that the term* $\sum_{i=m+1}^{M} \lambda_i$ *is the additional error due to compression of* X_s *into* Z_s *while* $Var(V)$ *is the estimation error when* X_s *is known perfectly.*

Remark 3.4 *One can check that the solution proposed in the theorem coincides with the solution obtained by the idea of partial KLT introduced in [5].*

3.2 Single encoder scenario with side information

The encoder can only sense X_s as in the previous scenario. However, in the present setup, the decoder has access to side information $Y_{s^c} = HX_{s^c}$, where H is a full row rank matrix. The aim once again is to determine Z_s such that X can be reconstructed optimally at the decoder.

Remark 3.5 *By the properties of* H *it follows that* $\Sigma_{Y_{s^c}} > 0$ *if* $\Sigma_{X_{s^c}} > 0$.

This scenario was treated in [5] by the introduction of the conditional KLT for the case where H has $N - M$ rows and the combined partial-conditional KLT for the case where H has less than $N - M$ rows. In this section we show that there is a single concept which allows us to treat this scenario for all cases (without having to split the analysis by considering two special cases). First we give a formal definition of the problem:

Problem 3.2 *Given a zero mean second order random vector* X *with* $\Sigma_X > 0$ *and side information vector* Y_{s^c}*, find a second order random vector* Z_s *of length* $m < M$ *with* $\Sigma_{Z_s} > 0$ *and linearly dependent on* X_s *(i.e.* $Z_s = DX_s$ *for some full row rank matrix* $D \in \mathbb{R}^{m\times M}$*) such that the quantity*

$$\mathrm{Var}\left(P\left[X|X_s, Y_{s^c}\right] - P\left[X|Z_s, Y_{s^c}\right]\right)$$

is minimized.

Remark 3.6 *Similar to the case of no side information (Remark 3.2), if there is a solution to Problem 3.2 then it is not unique.*

As in the previous section, we write

$$\begin{aligned} X &= P\left[X|X_s, Y_{s^c}\right] + V \\ &= AX_s + BY_{s^c} + V \end{aligned}$$

for some matrices A and B and a r.v. V which is orthogonal to the space spanned by X_s and Y_{s^c}. Define $U = X_s - P[X_s|Y_{s^c}] = X_s - \Sigma_{X_sY_{s^c}}\Sigma_{Y_{s^c}}^{-1}Y_{s^c}$, we call U *the innovation of* X_s. Note that $U \perp Y_{s^c}$. Then we may write:

$$X = \tilde{A}U + \tilde{B}Y_{s^c} + V$$

where $\tilde{A} = A$ and $\tilde{B} = B + A\Sigma_{X_sY_{s^c}}\Sigma_{Y_{s^c}}^{-1}$. The key observation is that since $U \perp Y_{s^c}$ and Y_{s^c} is known, all that remains is to compress $\tilde{A}U$ optimally in the mean square sense. To see this, let W_s denote an m-dimensional zero mean second order random vector with $\Sigma_{W_s} > 0$ and $W_s \perp Y_{s^c}$ (thus W_s does not repeat "linear estimation information" already carried by Y_{s^c}). Then the best linear estimate of $P\left[X|X_s, Y_{s^c}\right]$ given W_s and Y_{s^c} is clearly:

$$\begin{aligned} &P\left[P\left[X|X_s, Y_{s^c}\right]|W_s, Y_{s^c}\right] \\ &= P\left[\tilde{A}U + \tilde{B}Y_{s^c}|W_s, Y_{s^c}\right] \\ &= P\left[\tilde{A}U|W_s, Y_{s^c}\right] + P\left[\tilde{B}Y_{s^c}|W_s, Y_{s^c}\right] \\ &= P\left[\tilde{A}U|W_s\right] + \tilde{B}Y_{s^c} \text{ since } U \perp Y_{s^c} \text{ and } W_s \perp Y_{s^c}. \end{aligned}$$

and to minimize the quantity

$$\begin{aligned} &\mathrm{Var}\left(P\left[X|X_s, Y_{s^c}\right] - P\left[P\left[X|X_s, Y_{s^c}\right]|W_s, Y_{s^c}\right]\right) \\ &= \mathrm{Var}\left(\tilde{A}U - P\left[\tilde{A}U|W_s\right]\right) \end{aligned}$$

we simply choose W_s to be an m-dimensional KLT of $\tilde{A}U$ (see Section 2). Furthermore, with this choice of W_s it automatically follows that $V \perp \mathrm{span}\{W_s, Y_s\}$ due to W_s being a linear transformation of U. To this end, let $S \in \mathcal{T}\left(\Sigma_{\tilde{A}U}\right)$ 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 3.1 *Let A be a nonnegative symmetric $m \times m$ matrix. If T is an $n \times m$ $(n > m)$ matrix which majorly diagonalizes A then the $n - m$ smallest eigenvalues of $TAT^\top$ are zero.*

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

Corollary 3.1 *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 $B\Sigma_X B^\top$ and $\bar{S} = R(m, S)$ then*

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

and $\bar{S}$ majorly diagonalizes $B\Sigma_X B^\top$.

Proof. By the previous lemma $SB\Sigma_X B^\top S^\top = (SB)\Sigma_X(SB)^\top$ 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 $E[X] = 0$ these constants are actually zero. Thus we may write

$$SBX = \begin{bmatrix} \bar{S}BX \\ 0_{(n-m)\times m} \end{bmatrix}$$

From the above it is clear that $\bar{S}$ majorly diagonalizes $B\Sigma_X B^\top$. ■

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 $\bar{C} = R(m, C)$ then $\bar{C}U \in KLT(m, \tilde{A}U)$ and it is a linear transformation on U. Now recall that:

$$U = X_s - \Sigma_{X_s Y_{s^c}} \Sigma_{Y_{s^c}}^{-1} Y_{s^c}$$

hence

$$\bar{C}U = \bar{C}X_s - \bar{C}\Sigma_{X_s Y_{s^c}} \Sigma_{Y_{s^c}}^{-1} Y_{s^c}$$

Since the second term on the right of the equality can be computed at the decoder (because Y_{s^c} is known), the encoder only needs to send the remaining $m-$dimensional vector $Z_s = \bar{C}X_s$ so that $\bar{C}U$ can be reconstructed exactly. Once the decoder receives Z_s, $\tilde{A}U$ can be constructed approximately as $\tilde{U}$, which is defined as:

$$\begin{aligned} \tilde{U} &= S^{-1} \begin{bmatrix} \bar{C}U \\ 0_{N-m} \end{bmatrix} \\ &= S^\top \begin{bmatrix} \bar{C}U \\ 0_{N-m} \end{bmatrix} \end{aligned}$$

It then follows that the optimal mean square estimate of $P[X|X_s, Y_s]$ is:

$$\begin{aligned} \hat{X}_{opt} &= \tilde{U} + P[X|Y_{s^c}] \\ &= \tilde{U} + A_{X|Y_{s^c}} Y_{s^c} \end{aligned}$$

since $Y_{s^c} \perp U$ and therefore $Y_{s^c} \perp \tilde{U}$. The approximation error incurred is:

$$\begin{aligned} E\left[\left\|X - \hat{X}_{opt}\right\|^2\right] &= \mathrm{Var}(\tilde{A}U + A_{X|Y_{s^c}} Y_{s^c} + V - \tilde{U} - A_{X|Y_{s^c}} Y_{s^c}) \\ &= \mathrm{Var}(\tilde{A}U - \tilde{U}) + \mathrm{Var}(V) \\ &= \mathrm{Var}(\tilde{A}U - S^\top \begin{bmatrix} \bar{C}U \\ 0_{N-m} \end{bmatrix}) + \mathrm{Var}(V) \\ &= \mathrm{Var}(S\tilde{A}U - \begin{bmatrix} \bar{C}U \\ 0_{N-m} \end{bmatrix}) + \mathrm{Var}(V) \\ &= \mathrm{Var}(\begin{bmatrix} CU \\ 0_{N-M} \end{bmatrix} - \begin{bmatrix} \bar{C}U \\ 0_{N-m} \end{bmatrix}) + \mathrm{Var}(V) \\ &= \sum_{i=m+1}^{M} \lambda_i + \mathrm{Var}(V) \end{aligned}$$

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

Theorem 3.2 *Let U be the innovation of X_s. Then Problem 3.2 has base solutions of the form:*

$$Z_{sb} = \bar{C} X_s$$

where $\bar{C} = R(m, C)$ and $C = R(M, S) \cdot A_{X|U}$ for some $S \in \mathcal{M}_{KLT}(P[X|U])$. Every other solution Z_s is given by:

$$Z_s = A Z_{sb}$$

where A is any arbitrary invertible $m \times m$ matrix.

The corresponding optimal estimate of $P[X|X_s, Y_{s^c}]$ is given by:

$$\hat{X}_{opt} = S^\top \begin{bmatrix} \bar{C}U \\ 0_{N-m} \end{bmatrix} + A_{X|Y_{s^c}} Y_{s^c} = P[X|Z_s, Y_{s^c}]$$

and the approximation error incurred is:

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

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

Before proceeding to complete the proof of the theorem we first introduce the following definition:

Definition 3.1 *An $m \times M$ matrix $\bar{C}$ in Theorem 3.2 corresponding to a particular $S \in \mathcal{M}_{KLT}(\Sigma_{P[X|U]})$ is defined as the optimal m-dimensional transformation matrix associated with S and is denoted as $\bar{C} = OTM(m, S)$. If no reference is made to S then we call $\bar{C}$ simply as an m-dimensional optimal transformation matrix (m-OTM).*

Proof. This proof continues the argument prior to the statement of the theorem. First of all, since $U \perp Y_{s^c}$, $V \perp U$, it is clear that $\tilde{A}U = P[X|U]$. Next we show an important relation between an $m-$OTM $\bar{C}$ and $\hat{X}_{opt}$. Observe that in a similar way to proving that

$$P\left[X|\bar{C}X_s\right] = S^\top \begin{bmatrix} \bar{C}X_s \\ 0_{(N-m)\times m} \end{bmatrix}$$

in Theorem 3.1, but replacing X_s with U everywhere and using the corresponding definition of $\bar{C}$, we can prove that:

$$P\left[X|\bar{C}U\right] = S^\top \begin{bmatrix} \bar{C}U \\ 0_{(N-m)\times m} \end{bmatrix}$$

Hence

$$\begin{aligned} \hat{X}_{opt} &= S^\top \begin{bmatrix} \bar{C}U \\ 0_{(N-m)\times m} \end{bmatrix} + A_{X|Y_{s^c}} Y_{s^c} \\ &= P\left[X|\bar{C}U\right] + P[X|Y_{s^c}] \\ &= P\left[X|\bar{C}U, Y_{s^c}\right] \end{aligned}$$

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

$$\begin{aligned}\hat{X}_{opt} &= P\left[X|\bar{C}U, Y_{s^c}\right] \\ &= P\left[X|\bar{C}X_s, Y_{s^c}\right] \\ &= P\left[X|Z_{sb}, Y_{s^c}\right]\end{aligned}$$

Finally, $Z_s = AZ_{sb}$ for any $m \times m$ invertible matrix A is also a solution since

$$P\left[X|AZ_{sb}, Y_{s^c}\right] = P\left[X|Z_{sb}, Y_{s^c}\right]$$

and in this case one simply constructs $\tilde{U}$ as:

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

We have now exhausted all possible linear solutions since at the key step of optimal linear compression of $\tilde{A}U$ the only choices are precisely $\bar{C}U \in KLT(m, \tilde{A}U)$ or $A\bar{C}U$ for any $m \times m$ invertible matrix $A \neq I_{m\times m}$. This is because these choices correspond precisely to all zero mean second order random vectors whose elements span the same optimal subspace as the elements of $\bar{C}U$ (see Section 2 on the KLT). ■

Remark 3.7 *Note that the term* $\sum_{i=m+1}^{M} \lambda_i$ *is the additional error due to compression of* X_s *into* Z_s *while* $Var(V)$ *is the estimation error when* X_s *is also known perfectly (besides* Y_{s^c}*).*

Remark 3.8 *From the above theorem it readily follows that the innovation approach encompassess both the conditional and partial-conditional approach of [5] without making a distinction between the two.*

We may reformulate Problem 3.2 as the following equivalent problem:

Problem 3.3 *Let* $\mathcal{D}_{m\times M}$ *denote the space of* $m \times M$ $(m < M)$ *full row rank matrices. For any* $D \in \mathcal{D}_{m\times M}$*, define* $\hat{X}_D(X_s, Y_{s^c}) = P\left[X|DX_s, Y_{s^c}\right]$*. Given a zero mean second order random vector* X *with* $\Sigma_X > 0$*, find an element* $D \in \mathcal{D}_{m\times M}$ *which minimizes*

$$\mathrm{Var}\left(P\left[X|X_s, Y_{s^c}\right] - \hat{X}_D(X_s, Y_{s^c})\right)$$

Thus it is clear that the solutions to the above problem are the matrices $A\bar{C}$ given in Theorem 3.2 and this is formally stated in the next lemma.

Lemma 3.2 *The solutions of Problem 3.3 are precisely the matrices* $A\bar{C}$ *defined in Theorem 3.2.*

Remark 3.9 *The equivalent formulation will be particularly relevant for studying the multiple encoders problem in the section to follow (Problem 4.1).*

3.3 Unifying Theorem for Single Encoder Systems

If we take the convention that

1. $P[X|Z] = 0$ if $Z = 0$.
2. $P[X|Y, Z] = P[X|Y]$ if $Z = 0$.

then we may trivially combine Theorems 3.1 and 3.2 as one theorem that solves both Problems 3.1 and 3.2. We give this theorem below:

Theorem 3.3 *Let the innovation U be defined as:*

$$U = X_s - P[X_s|Y_{s^c}].$$

Then Problems 3.1 (where $Y_{s^c} = 0$) and 3.2 (where $Y_{s^c} \neq 0$) have base solutions of the form:

$$Z_{sb} = \bar{C}X_s$$

where $\bar{C} = R(m, C)$ and $C = R(M, S) \cdot A_{X|U}$ for some $S \in \mathfrak{M}_{KLT}(P[X|U])$. Every other solution Z_s is given by:

$$Z_s = AZ_{sb}$$

where A is any arbitrary invertible $m \times m$ matrix.

The corresponding optimal estimate of $P[X|X_s, Y_{s^c}]$ is given by:

$$\hat{X}_{opt} = S^\top \begin{bmatrix} \bar{C}U \\ 0_{N-m} \end{bmatrix} + A_{X|Y_{s^c}} Y_{s^c} = P[X|Z_s, Y_{s^c}]$$

and the approximation error incurred is:

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

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

Proof. Follows directly from Theorems 3.1 and 3.2 and the convention outlined at the beginning of this section. ■

The unifying theorem shows there is no intrinsic difference between the partial, conditional and combined partial-conditional approach of [5]. They are all special cases of the estimation framework that we have developed here. We end this section with some numerical examples.

Example 3.1 *We use Example 3 in [5]. 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 side information is just $X_{s,2}$*, i.e.* $H = I_{2\times 2}$*. We would like to produce a* 1*-dimensional approximation of* X*. Using Theorem 3.2 we get the following optimal (base) transformation matrices:*

$$C = \begin{bmatrix} 0.9447 & -0.3280 \\ 0.3280 & 0.9447 \end{bmatrix}$$
$$\bar{C} = \begin{bmatrix} 0.9447 & -0.3280 \end{bmatrix}$$

which is the same as the matrices reported in [5] except for the difference in sign, however as stated in Theorem 3.2 this difference is inconsequential. The optimal MSE that is computed is:

$$MSE_{opt} = 0.0264$$

which agrees with the value reported in [5].

Example 3.2 *Let* $X = (X_1, X_2, X_3, X_4, X_5, X_6)$ *with*

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

$X_{s,1} = (X_1, X_2, X_3)$ *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 side information is $Y_s = HX_{s,2}$ *with* $H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$*. We would like to produce a* $2-$*dimensional approximation of* X*. Again using Theorem 3.2 we get the following (base) 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:

$$MSE_{opt} = 2.0202$$

4 Multiple encoder scenarios

In this section we formulate the general distributed approximation problem with n-encoders ($n > 1$) and connect it with previous work that has been done on this problem in [5]. To this end, let $E_1, E_2, \ldots, E_n$ be n encoders which sense the vectors $X_{s,1}, X_{s,2}, \ldots, X_{s,n}$, respectively. Let $\dim(X_{s,i}) = M_i$. Then $X_s = \left(X_{s,1}^\top, X_{s,2}^\top, \ldots, X_{s,n}^\top\right)^\top$. Let the hidden part be X_{s^c} and the side information Y_{s^c} be defined

as before and let the output of $E_1, E_2, \ldots, E_n$ be denoted by $Z_{s,1}, Z_{s,2}, \ldots, Z_{s,n}$, respectively, with $\dim(Z_{s,i}) = m_i < M_i$. For mathematical tractability, let us 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 4.1).

In the spirit of Problem 3.3 for the single encoder case, we formulate the following optimization problem:

Problem 4.1 *Let $\mathfrak{I} = \{1, 2, \ldots, n\}$. For any $i \in \mathfrak{I}$, 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 $(D_1, D_2, \ldots, D_n) \in \mathcal{D}_{m_1 \times M_1} \times \mathcal{D}_{m_2 \times M_2} \times \ldots \times \mathcal{D}_{m_n \times M_n}$ define:*

$$\hat{X}_{D_1, D_2, \ldots, D_n}(X_{s,1}, X_{s,2}, \ldots, X_{s,n}, Y_{s^c}) = P[X | D_1 X_{s,1}, D_2 X_{s,2}, \ldots, D_n X_{s,n}, Y_{s^c}]$$

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

$$Err\,(D_1, D_2, \ldots, D_n) = \mathrm{Var}(P\,[X | X_s, Y_{s^c}] - \hat{X}_{D_1, D_2, \ldots, D_n}(X_{s,1}, X_{s,2}, \ldots, X_{s,n}, Y_{s^c}))$$

A solution to the above problem is called an optimal linear solution to the n-encoders distributed approximation problem.

Remark 4.1 *In the formulation of Problem 4.1, we have explicitly assumed that there is side information Y_{s^c} 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_{s^c} wherever it is found.*

An intuitive approach to solve Problem 4.1 is to set $D_2, \ldots, D_n$ arbitrarily, then proceeding to minimize Err one matrix at a time starting from D_1 and then $D_2, \ldots, D_n$ and starting over from D_1 until Err becomes relatively constant (i.e. the iteration has converged). This is the basically the idea proposed in [5], by an algorithm called the DKLT (for distributed Karhunen-Loève transform) algorithm. However, it was proposed without an explicit formulation of the optimization problem as we have done here. As we show in this section, the explicit problem formulation is particularly useful since it allows us to better understand the multiple encoders scenario. We also show that an execution of the DKLT algorithm can result in a sub-optimal solution. Before continuing, we first describe the DKLT algorithm:

Algorithm 4.1 (DKLT)

1. *Choose $(D_{2,0}, D_{3,0}, \ldots, D_{n,0})$ arbitrarily from $\mathcal{D}_{m_2 \times M_2} \times \mathcal{D}_{m_3 \times M_3} \times \ldots \times \mathcal{D}_{m_n \times M_n}$ and let $Z_{s,i}^{(0)} = D_{i,0} X_{s,i}$ for $\forall i \in \mathfrak{I} \backslash \{1\}$.*
2. *Set $k = 1$.*
3. *Let $i = (k-1) \bmod n + 1$. Choose an m_i-OTM $D_{i,k}^*$ for E_i (see Definition 3.1) by regarding the collection of vectors $\{Z_{s,j}^{(k-1)}, j \in \mathfrak{I} \backslash \{i\}\}$ and Y_{s^c} as side information (using Theorem 3.2). Set*

$$D_{i,k} = D_{i,k}^*$$
$$D_{j,k} = D_{j,k-1} \quad \forall j \in \mathfrak{I} \backslash \{i\}$$

and

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

4. *Repeat the proceduce of step 3 sequentially for $k = 2, 3, \ldots,$ 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.*

Going back to Problem 4.1, we can state the following result:

Theorem 4.1 (Necessity) *If $D_1^*, \ldots, D_n^*$ is a solution to Problem 4.1, then necessarily each encoder E_i, $i = 1, 2, \ldots, n$, must be linearly optimal as a single encoder system with side information $\left\{\left\{D_j^* X_{s,j},\ j \in \mathcal{I}\backslash\{i\}\right\}, Y_{s^c}\right\}$.*

Proof. Since $(D_1^*, \ldots, D_i^*, \ldots, D_n^*)$ is a solution to Problem 4.1, it is clear that:

$$Err\left(D_1^*, D_2^*, D_3^* \ldots, D_n^*\right) \leq Err\left(D_1, D_2^*, D_3^* \ldots, D_n^*\right) \quad \forall D_1 \in \mathcal{D}_{m_1 \times M_1}$$
$$Err\left(D_1^*, D_2^*, D_3^* \ldots, D_n^*\right) \leq Err\left(D_1^*, \ldots, D_{i-1}^*, D_i, D_{i+1}^*, \ldots, D_n^*\right) \quad \forall D_i \in \mathcal{D}_{m_i \times M_i}, i \in \mathcal{I}\backslash\{1, n\}$$
$$Err\left(D_1^*, \ldots, D_i^*, \ldots, D_n^*\right) \leq Err\left(D_1^*, \ldots, D_{n-1}^*, D_n\right) \quad \forall D_n \in \mathcal{D}_{m_n \times M_n}$$

implying that

$$\left.\begin{aligned} D_1^* &= \underset{D_1 \in \mathcal{D}_{m_1 \times M_1}}{\arg\min}\ Err\left(D_1, D_2^*, D_3^* \ldots, D_n^*\right) \\ D_i^* &= \underset{D_i \in \mathcal{D}_{m_i \times M_i}}{\arg\min}\ Err\left(D_1^*, \ldots, D_{i-1}^*, D_i, D_{i+1}^*, \ldots, D_n^*\right) \quad \forall i \in \mathcal{I}\backslash\{1, n\} \\ D_n^* &= \underset{D_n \in \mathcal{D}_{m_n \times M_n}}{\arg\min}\ Err\left(D_1^*, \ldots, D_{n-1}^*, D_n\right) \end{aligned}\right\} \tag{4.1}$$

By the relation

$$\begin{aligned} Err\left(D_1, \ldots, D_n\right) = {} & \mathrm{Var}\left(P\left[X|X_s, Y_{s^c}\right] - P\left[X|X_{s,i}, \left\{D_j X_{s,j}, j \in \mathcal{I}\backslash\{i\}\right\}, Y_{s^c}\right]\right) + \\ & \mathrm{Var}\left(P\left[X|X_{s,i}, \left\{D_j X_{s,j}, j \in \mathcal{I}\backslash\{i\}\right\}, Y_{s^c}\right] - P\left[X|D_i X_{s,i}, \left\{D_j X_{s,j},\ j \in \mathcal{I}\backslash\{i\}\right\}, Y_{s^c}\right]\right) \quad i \in \mathcal{I}, \end{aligned} \tag{4.2}$$

Problem 3.3, and Lemma 3.2, it follows from (4.1) that E_i $(i = 1, 2, \ldots, n)$ must be a linearly optimal single encoder having $\left\{D_j^* X_{s,j}, j \in \mathcal{I}\backslash\{i\}\right\}$ and Y_{s^c} as side information. ■

Now it is quite clear that the DKLT algorithm is a natural approach for obtaining transformation matrices satisfying the conditions of Theorem 4.1. However, we cannot conclude that a solution obtained by this method is globally optimal since the conditions are only necessary, i.e. the DKLT algorithm merely provides us with one set of transformation matrices which satisfy the necessary conditions. *Sufficiency of the conditions of Theorem 4.1 have not been established.* This is not easy since it is a nonlinear optimization problem in operator space with no readily usable convexity property. Thus, there is a gap in the assertion made in [5] of the optimality of a solution obtained with the DKLT algorithm. Additionally, in light of Theorem 3.3 and Lemma 3.2, we can give a direct proof of the convergence of the DKLT algorithm. In the remaining parts of this paper we provide this proof and show that the DKLT algorithm does not guarantee global optimality via an example.

Lemma 4.1 (Convergence of the DKLT algorithm) *At consecutive iterations of Algorithm 4.1, the estimation error cannot increase, i.e.*

$$Err(D_{1,k+1}, \ldots, D_{n,k+1}) \leq Err\left(D_{1,k}, \ldots, D_{n,k}\right)$$

for all $k \geq 0$. Furthermore, the algorithm has converged at iteration k (i.e. $Err\left(D_{1,m}, \ldots, D_{n,m}\right) = Err\left(D_{1,k}, \ldots, D_{n,k}\right)$ $\forall m > k$) if and only if $D_{1,k}, \ldots, D_{n,k}$ satisfy the conditions of Theorem 4.1. 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.

Proof. 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 $\{Z_{s,j}^{(k-1)}, j \in \mathfrak{I}\backslash\{i\}\}$ along with Y_{s^c} as side information for the encoder E_i. Since $D_{i,k} = D_{i,k}^*$ by Theorem 3.2 (with $\hat{X}_{k,opt}$ corresponding to $\hat{X}_{opt}$ in the theorem) we have:

$$\hat{X}_{k,opt} = P\left[X | D_{i,k} X_{s,i}, \{Z_{s,j}^{(k-1)}, j \in \mathfrak{I}\backslash\{i\}\}, Y_{s^c}\right]$$

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

$$\hat{X}_{k+1,opt} = P\left[X | D_{r,k+1} X_{s,r}, \{Z_{s,j}^{(k)}, j \in \mathfrak{I}\backslash\{r\}\}, Y_{s^c}\right]$$

By equation (4.2) and Lemma 3.2, 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 a solution to Problem 3.3 with E_r being the associated encoder. In other words,

$$\mathrm{Var}(P[X|X_s, Y_{s^c}] - \hat{X}_{k+1,opt}) \leq \mathrm{Var}(P[X|X_s, Y_{s^c}] - \hat{X}_{k,opt})$$

However, since by definition

$$\hat{X}_{k,opt} = \hat{X}_{D_{1,k},\ldots,D_{n,k}}(X_{s,1}, \ldots, X_{s,n}, Y_{s^c})$$

and

$$\hat{X}_{k+1,opt} = \hat{X}_{D_{1,k+1},\ldots,D_{n,k+1}}(X_{s,1}, \ldots, X_{s,n}, Y_{s^c})$$

we conclude that

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

for all $k \geq 0$. If $D_{1,k}, \ldots, D_{n,k}$ satisfy the conditions of Theorem 4.1 then each encoder E_i $(i = 1, \ldots, n)$ is optimal as a single encoder system with side information $\{\{D_j X_{s,j},\ j \in \mathfrak{I}\backslash\{i\}\}, Y_{s^c}\}$. This implies that no further sequential change of the transformation matrices can yield a lower estimation error. Hence we may set $D_{i,m} = D_{i,k}$ for $i = 1, 2, \ldots, n$ and $\forall m > k$, and we have that $Err\,(D_{1,m}, \ldots, D_{n,m}) = Err\,(D_{1,k}, \ldots, D_{n,k})$ $\forall m > k$. Conversely, if $D_{1,k}, \ldots, D_{n,k}$ do not satisfy the conditions of Theorem 4.1 then at least one encoder, excluding the encoder 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 encoders to be encountered in iterations $k+1, k+2, \ldots, k+n-1$, i.e. $\exists m \in \{1, \ldots, n-1\}$ such that $Err(D_{1,k+m}, \ldots, D_{n,k+m}) < Err\,(D_{1,k}, \ldots, D_{n,k})$. 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.

■

As we had argued earlier, based on Problem 4.1, it seems clear that in general, the necessary conditions need not be sufficient for optimality. The following example affirms this fact:

Example 4.1 *Let X be as given in Example 3.2. Encoder 1 senses $X_{s,1}$ while encoder 2 senses $X_{s,2}$. We would like to produce an optimal linear approximation of X under the constraint that each encoder may only send a $2-$dimensional vector.*

Let us first apply the DKLT algorithm by setting

$$D_{2,0} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{4.3}$$

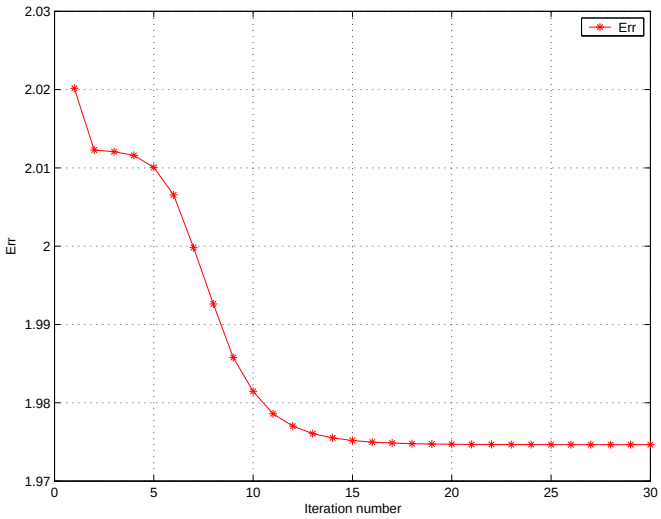

Figure 1: Convergence of the DKLT algorithm with $D_{2,0}$ as given in (4.3)

The result of the applying the DKLT algorithm is shown in Fig. 1. 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 3.2 (after 30 iterations) is:

$$\mathrm{Var}\left(X - \hat{X}_{D_1^*, D_2^*}(X_{s,1}, X_{s,2})\right) = 1.9746$$

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

$$D_{2,0} = \begin{bmatrix} 2 & 3 & 0 \\ 0 & -4 & 3 \end{bmatrix} \tag{4.4}$$

The result of the applying the DKLT algorithm with this new initial condition is shown in Fig. 2. 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:

$$\mathrm{Var}\left(X - \hat{X}_{D_1^*, D_2^*}(X_{s,1}, X_{s,2})\right) = 1.7567$$

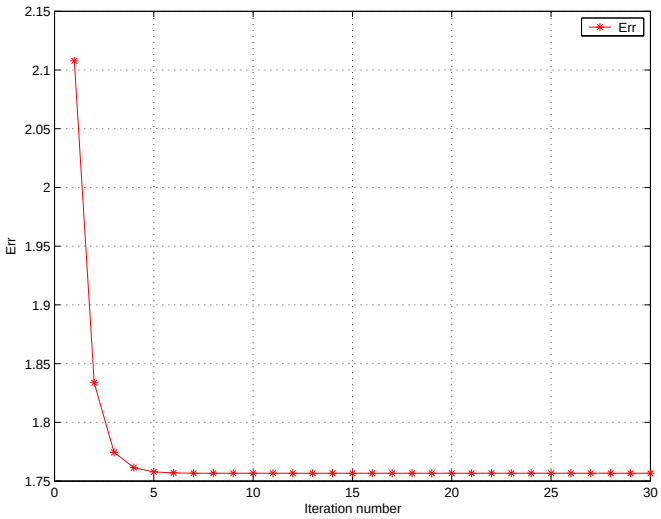

Figure 2: Convergence of the DKLT algorithm with $D_{2,0}$ as given in (4.4)

Thus with the DKLT algorithm, with different initial conditions, one can arrive at different points satisfying the necessary conditions of Theorem 4.1, but which result in different estimation errors. In this example starting at

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

results in a lower estimation error than starting at:

$$D_{2,0} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Therefore the DKLT algorithm will not necessarily give a global optimal solution to Problem 4.1. This example demonstrates that the necessary conditions need not be sufficient to ensure global optimality.

The main significance of the formulation of Problem 4.1 is that it gives insight into what the DKLT algorithm accomplishes and how it does not guarantee global optimality. The explicit formulation of the objective function opens the possibility for finding or developing other optimization algorithms, instead of the DKLT, which may be able to guarantee global or close to global minimality.

5 Concluding Remarks

In this paper we have shown in an explicit manner the geometric structure associated with the multiple encoder-decoder problem in the estimation of correlated second order r.v's based in incomplete observations by the different encoders and decoders. In the linear context this leads to a nice geometric interpretation in terms of the innovations and results in a nice decoupling property for encoder-decoder pairs. However, the conditions are only necessary and the derivation of sufficient conditions is extremely difficult. These results help us to better understand the results shown in [5] and in particular that the there is no intrinsic difference between the conditional, partial, and conditional-partial KLTs as defined in [5]. This geometric framework shows us why the DKLT algorithm proposed therein can be sub-optimal.

In future work we will pursue an optimal procedure based on the global optimization formulation given in the paper and what simplifications can be obtained in the large number of encoder context. A more general problem is one of combining estimation and LMS techniques to derive a pure distributed signal estimation where only the covariances of the observed r.v's are known (and the cross covariances are unknown) and the idea is to combine covariance and state estimation within an adaptive framework. These topics will be presented elsewhere.

References

[1] A. Bagchi. *Optimal Control of Stochastic Systems.* Prentice-Hall, 1993.

[2] R.J. Clarke. *Transform Coding of Images.* Microelectronics and Signal Processing. Academic Press, 1985.

[3] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, 2 edition, 1990.

[4] M. Gastpar, P-L. Dragotti, and M. Vetterli. The distributed Karhunen-Loeve transform. In *Proc. of 2002 IEEE International Workshop on Multimedia Signal Processing.* IEEE Signal Processing Society, December 2002.

[5] M. Gastpar, P-L. Dragotti, and M. Vetterli. *The Distributed Karhunen-Loève Transform.* EPFL I&C Technical Report IC/2003/12. EPFL-Lausanne, Switzerland, 2003.

[6] M. Gastpar, P-L. Dragotti, and M. Vetterli. The distributed, partial and conditional Karhunen-Loeve transforms. In *Proc. Data Compression Conference.* IEEE Computer Computer Society, March 2003.

[7] J.E. Jackson. *A User's Guide to Principal Components.* Wiley-Interscience, 1991.

[8] A.K. Jain. *Fundamentals of Digital Image Processing.* Prentice-Hall, 1989.

[9] T. Kailath, A.H. Sayed, and B. Hassibi. *Linear Estimation.* Prentice-Hall, 2000.

[10] E. Wong and B. Hajek. *Stochastic Processes in Engineering Systems.* Springer-Verlag, 1985.

[11] A.D. Wyner and J. Ziv. The rate distortion function for source coding with side information at the decoder. *IEEE Trans. on Inform. Th.*, IT-22(1):1–88, 1976.