The Distributed Karhunen–Loève Transform

Michael Gastpar, *Member, IEEE*, Pier Luigi Dragotti, *Member, IEEE*, and Martin Vetterli, *Fellow, IEEE*

***Abstract*—The Karhunen–Loève transform (KLT) is a key element of many signal processing and communication tasks. Many recent applications involve *distributed* signal processing, where it is not generally possible to apply the KLT to the entire signal; rather, the KLT must be approximated in a distributed fashion. This paper investigates such distributed approaches to the KLT, where several distributed terminals observe disjoint subsets of a random vector.**

We introduce several versions of the distributed KLT. First, a *local* KLT is introduced, which is the optimal solution for a given terminal, assuming all else is fixed. This local KLT is different and in general improves upon the *marginal* KLT which simply ignores other terminals. Both optimal approximation and compression using this local KLT are derived. Two important special cases are studied in detail, namely, the *partial observation* KLT which has access to a subset of variables, but aims at reconstructing them all, and the *conditional* KLT which has access to side information at the decoder. We focus on the jointly Gaussian case, with known correlation structure, and on approximation and compression problems. Then, the distributed KLT is addressed by considering local KLTs in turn at the various terminals, leading to an iterative algorithm which is locally convergent, sometimes reaching a global optimum, depending on the overall correlation structure. For compression, it is shown that the classical distributed source coding techniques admit a natural transform coding interpretation, the transform being the distributed KLT. Examples throughout illustrate the performance of the proposed distributed KLT. This distributed transform has potential applications in sensor networks, distributed image databases, hyper-spectral imagery, and data fusion.

***Index Terms*—Distributed source coding, distributed transforms, rate–distortion function, principal components analysis, side information, transform coding.**

I. Introduction

THE approximation or compression of an observed signal is a central and widely studied problem in signal processing and communication. The Karhunen–Loève transform (KLT), also referred to as principal component analysis (PCA), [3]–[5], has always played a pivotal role in this context. Assume, for instance, that the observed signal is a random vector $\boldsymbol{x}$ with covariance matrix Σ_x and that the statistics of the source are known. Then to solve the approximation problem, one can apply the KLT to $\boldsymbol{x}$ to obtain uncorrelated components and the optimal linear least squares k-order approximation of the source is given by the k components corresponding to the k largest eigenvalues of Σ_x. In the case of compression, the uncorrelated components can be compressed independently and more rate can be allocated to the components related to the largest eigenvalues of Σ_x, according to a principle that is sometimes referred to as "reverse water-filling," see, e.g., [6, p. 349]. This compression process is widely known as transform coding and, if the input source is a jointly Gaussian source, it is possible to show that it is optimal [7]. For an excellent review on transform coding and a discussion of the optimality of the KLT in this context, we refer to the exposition in [8].

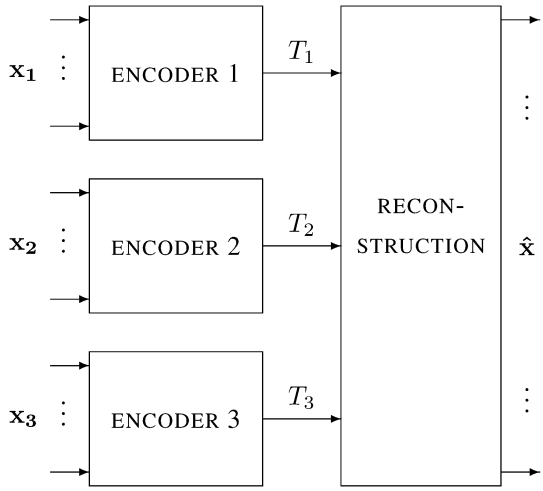

Fig. 1. The distributed KLT problem: Distributed compression of multiple correlated vector sources. Encoder ℓ provides a description T_ℓ of its observation. This paper investigates the case where T_ℓ is a k_ℓ-dimensional approximation of the observation, and the case where T_ℓ is a bit stream of R_ℓ bits per observation.

In the present work, which builds on [1] and [2], we investigate a related scenario where there are multiple sensors, each observing only a part of the vector $\boldsymbol{x}$, see Fig. 1. For the scope of the present paper, $\boldsymbol{x}$ is assumed to be a vector of jointly Gaussian random variables, even though some of our results are more general. The sensors transmit an approximation of the observed subvector to a fusion center and cannot communicate with each other. Thus, signal processing must be done in a distributed fashion and the full KLT cannot be applied to the data. Therefore, the original approximation and compression problems change significantly under these circumstances, and we

Manuscript received November 4, 2004; revised July 31, 2006. This work was supported in part by the National Science Foundation under award CCF-0347298 (CAREER) and by the Swiss National Science Foundation in the framework of the National Competence Center in Research for Mobile Information and Communication Systems (http://www.mics.org). The material in this paper was presented in part at the 2002 International Workshop on Multimedia Signal Processing, St. Thomas, U.S. Virgin Islands, December 2002, and at the 2003 IEEE Data Compression Conference, Snowbird, UT, March 2003.

M. Gastpar is with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA 94720-1770 USA (e-mail: gastpar@eecs.berkeley.edu).

P. L. Dragotti is with the Department of Electrical and Electronic Engineering, Imperial College, London, SW7 2BT, U.K. (e-mail: p.dragotti@imperial.ac.uk).

M. Vetterli is with the Institute of Communication Systems, Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland, and with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA 94720 USA (e-mail: martin.vetterli@epfl.ch).

Communicated by S. A. Savari, Associate Editor for Source Coding.

Digital Object Identifier 10.1109/TIT.2006.885449

show how to extend the concept of the KLT (or PCA) to such a distributed scenario.

In this paper, the usual, nondistributed scenario is termed *joint* KLT. We pose the following questions: given a distributed scenario, what is the best possible performance, and how can it be attained? A trivial upper bound to the distortion of a distributed scheme is given by the *marginal* KLT, where each terminal computes a KLT based only on its local statistics, ignoring the dependencies with other terminals. And an obvious lower bound is the performance of the unconstrained, joint KLT. Depending on the correlation structure and on the subsets that the various terminals observe, the two bounds will be close or far apart. The distributed KLT lives between these bounds, and aims at the best performance (in approximation error or distortion rate) under the constraint of a distributed scenario.

Of course, in the compression case, various distributed KLT problems are classical distributed compression problems. The "conditional" KLT, defined in Section IV-B, is the scenario where all but one of the terminals provide all of their observations to the reconstruction point, and the task is to design the source code for the remaining terminal. This is a Wyner–Ziv problem (compression with side information at the decoder [9]), and the distributed KLT is part of the general distributed source coding problem, for which bounds are known [10]. Our aim, in the compression part, is thus to give a *transform coding* perspective to these distributed coding problems. In addition to giving intuition and structure to the problem, this perspective has clear computational advantages, since lossy source compression is almost always transform based for complexity reasons.

Related work is twofold. First, as already pointed out earlier, the distributed KLT compression falls into the general area of distributed lossy source compression. This work goes back to the classic lossless result of Slepian and Wolf [11], and the lossy case with side information to Wyner and Ziv [9]. An early set of general upper and lower bounds on the performance have been found by Berger and Tung [10], [12], with many refinements for special cases, including the scenario where all but one of the sources are either not to be reconstructed, or encoded perfectly (see [13]–[15]). Moreover, certain conclusive results are available for the case of high resolution [16], for the CEO problem [17], [18], for the case with side information [19], and for certain special distortion measures (not including the mean-squared error) [15]. Finally, a recent result solves the two-source jointly Gaussian distributed source coding problem with respect to mean-squared error [20]. Second, a recent flurry of papers has looked at various facets of practical distributed source coding. This includes various schemes using channel codes for distributed source compression [21]–[29], including video compression [30]–[32]. Several authors looked at distributed transform coding, for example in the high-rate case and, in this regime, some optimality conditions for the transforms can be proved [33]–[35]. In [36], [37], a related transform is studied from the perspective of estimation theory, and in [38], a particle filtering view of our problem is investigated. In follow-up work, a study of the large block-size (as $N \to \infty$) case was done in [39], using the asymptotic eigenvalue distribution of Toeplitz matrices (see, e.g., [40]). Another study considers the generalized problem where the observations are noisy [41].

The paper is organized as follows: Section II reviews the classic approximation and compression problems for a random vector source with known second-order statistics, together with the standard results for the joint KLT. Then, the problem leading to the distributed KLT is formally stated.

Section III takes a terminal-by-terminal perspective, that is, all terminals but one are fixed, and we search for the optimal encoding procedure for the remaining terminal. This leads to the optimal *local* KLT, under the following two scenarios: In the first scenario, the remaining terminal needs to select a k-dimensional approximation of its observation; this is sometimes referred to as the *linear approximation* problem. In the second perspective, the remaining terminal needs to provide a representation of its observation using R bits per source sample; that is, we study the information-theoretic compression or *rate–distortion* problem.

In Section IV, two simple special cases are investigated in detail: on the one hand, there is the case when all but one terminal are cut off. This means that the remaining terminal must provide an approximation of its observation that permits the best reconstruction of the *entire* vector. We call this the *partial-observation* KLT. On the other hand, there is the case when all but one terminal provide their observations *entirely* and uncompressed to the decoder. This means that the remaining terminal can exploit this side information at the decoder in order to provide a more efficient description of its observation. While the linear approximation perspective of this problem appears to be new, the related rate–distortion (i.e., compression) problem is well known and has been solved by Wyner and Ziv [9]. In this sense, the present paper extends the result of Wyner and Ziv to the case of correlated vectors and this leads to the introduction of a new transform called the *conditional* KLT.

Section V addresses the distributed scenario by using the local KLT of Section III in turn at each terminal. This leads to an iterative algorithm for finding the *distributed KLT*. The approximation problem and the compression problem (under a sum–rate constraint) are studied. In both cases, the convergence of the iterative procedure to a local optimum or a saddle point is shown. The question of local versus global optima is explored through a set of examples in Section VI. Possible applications of the distributed KLT and topics of further research conclude the paper in Section IV.

II. The Distributed KLT Problem

The problem leading to the distributed KLT is illustrated in Fig. 1: There are L terminals (the figure illustrates the case $L = 3$), and each terminal samples a part of a vector of N jointly Gaussian real-valued random variables

$$\boldsymbol{x} \stackrel{\text{def}}{=} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}. \tag{1}$$

The Gaussian random vector $\boldsymbol{x}$ has zero mean[1] and covariance matrix Σ_x.

The standard (nondistributed) Karhunen–Loève transform arises as the solution of the approximation (or compression) problem illustrated in Fig. 1, but with all encoders *merged* into one overall encoder that observes the entire vector $\boldsymbol{x}$. The task of the encoder is to provide a description of the vector $\boldsymbol{x}$ in such a way as to permit the reconstruction point to produce an estimate $\hat{\boldsymbol{x}}$ that minimizes the mean-squared error

$$E[\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|^2] \tag{2}$$

where $E[\cdot]$ denotes the expectation.

The distributed version of the problem is interesting because the description provided by the encoder is *constrained*, i.e., the encoder cannot simply provide the reconstruction point with the entire observation. In this paper, we consider two types of approximations: *linear approximation* and *compression* (in a rate-distortion sense).

A. Linear Approximation

For the standard (nondistributed) KLT, the goal of the encoder is to provide a k-dimensional approximation of the vector $\boldsymbol{x}$. For a fixed k, the goal is to find the approximation space that minimizes the resulting mean-squared error. The matrix Σ_x is real, symmetric, and positive semidefinite, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$. It allows thus a diagonalization given by

$$\Sigma_x = Q_x \Lambda_x Q_x^T \tag{3}$$

where $Q_x \in \mathbb{R}^{N\times N}$ is a unitary matrix whose columns are the eigenvectors of the matrix Σ_x, ordered by decreasing eigenvalues, and $\Lambda_x \in \mathbb{R}^{N\times N}$ is diagonal, with entries $\lambda_1, \lambda_2, \ldots, \lambda_N$. The matrix Q_x^T is called the Karhunen–Loève transform for $\boldsymbol{x}$.[2] More specifically, in this paper, we will refer to Q_x^T as the *joint* KLT of the vector $\boldsymbol{x}$, by contrast to the *distributed* KLT developed in this paper. We denote the transformed version of $\boldsymbol{x}$ by

$$\boldsymbol{y} = Q_x^T \boldsymbol{x}. \tag{4}$$

Since Q_x is unitary, it follows that

$$\begin{aligned} E[\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|^2] &= E[\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|^2] \\ &= \sum_{m=1}^{N} E[|y_m - \hat{y}_m|^2] \end{aligned} \tag{5}$$

where $\hat{\boldsymbol{y}} = Q_x^T \hat{\boldsymbol{x}}$. The key insight is that the components of $\boldsymbol{y}$ are uncorrelated. Therefore, in terms of the components $(y_1, y_2, \ldots, y_N)$, a simple answer can be given. First, if the component y_m is retained, then clearly its corresponding estimate is $\hat{y}_m = y_m$. However, if y_m is *not* retained, then its corresponding estimate is $\hat{y}_m = 0$; none of the other components of the vector $\boldsymbol{y}$ contain anything relevant about y_m. The best k-dimensional approximation space is therefore easily found in terms of $\boldsymbol{y}$: Denote the set of the k indices corresponding to the retained components of $\boldsymbol{y}$ by $\mathcal{T}$. Then, the incurred distortion can be expressed as

$$E[\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|^2] = \sum_{m\in\mathcal{T}^c} \lambda_m \tag{6}$$

where $\mathcal{T}^c$ denotes the complement of $\mathcal{T}$ in the set $\{1, 2, \ldots, N\}$. Hence, the best k-dimensional approximation is given by the eigenvectors corresponding to the k largest eigenvalues.

In the distributed KLT scenario, there are L separate encoding terminals that cannot communicate with each other. The first terminal observes the first M_1 components of the vector $\boldsymbol{x}$, denoted by $\boldsymbol{x}_1 = (x_1, x_2, \ldots, x_{M_1})$, the second terminal the next M_2 components, denoted by $\boldsymbol{x}_2 = (x_{M_1+1}, x_{M_1+2}, \ldots, x_{M_1+M_2})$, and so on. Clearly, in that case, it is not possible to apply the KLT to the vector $\boldsymbol{x}$. Instead, each terminal individually provides a certain approximation of its samples to a central decoder. The goal of the central decoder is to produce an estimate $\hat{\boldsymbol{x}}$ in such a way as to minimize the mean-squared error $E[\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|^2]$.

Terminal ℓ provides a k_ℓ-dimensional approximation of its sampled vector, where $k_1, k_2, \ldots, k_L$ are fixed integers, $0 \leq k_\ell \leq M_\ell$, for $\ell = 1, 2, \ldots, L$. One approach would be for each encoder to apply a (standard) KLT to its observations, and provide the components corresponding to the largest k_ℓ eigenvalues. In this paper, we will refer to this approach as the *marginal* KLTs. It is easy to verify that the marginal KLT will lead to suboptimal performance in general. Hence, what are the best approximation spaces for the L terminals? This question can be answered directly in the following sense: Terminal ℓ applies a $k_\ell \times M_\ell$ matrix C_ℓ to its local observation, and provides this to the reconstruction point. Hence, the reconstruction point has access to

$$C\boldsymbol{x} \stackrel{\text{def}}{=} \begin{pmatrix} C_1 & 0 & \cdots & 0 \\ 0 & C_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & C_L \end{pmatrix} \boldsymbol{x}. \tag{7}$$

However, since we have assumed that $\boldsymbol{x}$ is Gaussian, it is well known that the best estimate of $\boldsymbol{x}$ based on $C\boldsymbol{x}$ is the linear minimum mean-squared error estimator, given by the standard formula (see, e.g., [42, Theorem 3.2.1])

$$\hat{\boldsymbol{x}} = \Sigma_x C^T (C\Sigma_x C^T)^{-1} (C\boldsymbol{x}) \tag{8}$$

and the corresponding mean-squared error distortion can be written as (see, e.g., [42, Theorem 3.2.1]

$$D = \operatorname{trace}(\Sigma_x - \Sigma_x C^T (C\Sigma_x C^T)^{-1} C\Sigma_x). \tag{9}$$

From this perspective, our problem can be stated as the minimization of this distortion over all block-diagonal matrices $C \in \mathbb{R}^{(\sum_{\ell=1}^{L} k_\ell)\times N}$, where the ℓth block is precisely C_ℓ. That is, C has the shape given in (7). A simple solution for the best such matrix C does not seem to exist in general. Instead, this paper provides an iterative algorithm that finds (locally) optimal solutions.

[1]The assumption that $\boldsymbol{x}$ has zero mean is not crucial, but it considerably simplifies the notation. Therefore, it is kept throughout the paper.

[2]The matrix Q_x always exists, and it is unique if $\lambda_m^2 \neq \lambda_n^2$, for $m \neq n$. Conversely, if the matrix Σ_x has repeated eigenvalues, then combinations of the respective eigenvectors are also eigenvectors, leading to nonunique approximation and compression. This is a technicality we do not get into any further.

B. Compression

In a second version of the problem, the encoder has to compress the observed vector in a rate-distortion sense. That is, the encoder gets to observe a long *sequence* $\{\boldsymbol{x}[i]\}_{i=1}^{n}$ of source vectors, where $\boldsymbol{x}[i] \in \mathbb{R}^N$ and i is the (discrete) time index. This sequence can be encoded jointly.

For the standard (nondistributed) problem, the encoder output is a binary sequence T that appears at rate R bits per input vector. The reconstruction point produces a sequence of estimates $\{\hat{\boldsymbol{x}}_T[i]\}_{i=1}^{n}$, and the goal is to minimize the *average* mean-squared error over the entire block of n source vectors, defined as

$$D_n(R) = \frac{1}{n}\sum_{i=1}^{n} E[\|\boldsymbol{x}[i] - \hat{\boldsymbol{x}}_T[i]\|^2]. \tag{10}$$

Our interest concerns specifically the limit as $n \to \infty$. It is well known that one architecture of an optimal encoder is to apply a KLT to the vector $\boldsymbol{x}[i]$ to obtain $\boldsymbol{y}[i]$ as in (4). Then, each component of $\boldsymbol{y}[i]$ can be encoded separately, using a rate-distortion optimal code. The rate allocation between the components of $\boldsymbol{y}[i]$ is determined by the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$, see [43] or [6, p. 348].

In the distributed KLT scenario illustrated in Fig. 1, there are multiple separate encoding terminals that cannot communicate with each other. Clearly, in that case, it is not possible to apply the KLT to the vector $\boldsymbol{x}$. Instead, each terminal individually provides a binary sequence T_ℓ that appears at rate R_ℓ bits per input vector, for $\ell = 1, 2, \ldots, L$. The reconstruction point produces a sequence of estimates $\{\hat{\boldsymbol{x}}_{T_1,T_2,\ldots,T_L}[i]\}_{i=1}^{n}$, and the task is to minimize the resulting average mean-squared error as in (10). As discussed generally in Section I, optimum tradeoffs between the rates $R_1, R_2, \ldots, R_L$ and the incurred distortion D, and hence, the optimum processing architecture, are mostly unknown to date, except in the case $L = 2$ (and $N = 2$) [20]. Instead, in this paper, we evaluate the best known achievable rate-distortion region [10], [44] in the considered transform coding framework, and show that they can be attained by an architecture where each terminal applies a suitable local transform C_ℓ, followed by *separate* encoding of each coefficient stream, for $\ell = 1, 2, \ldots, L$. It should also be pointed out that a brief treatment of the case of two jointly stationary Gaussian sources has been given in [12, Ch. 6, pp. 75–78].

III. Terminal-by-Terminal Perspective

This section investigates a local perspective of the problem described in Section II. Suppose that all terminals except terminal j have fixed descriptions $T_\ell, \ell \neq j, \ell = 1, 2, \ldots, L$, and the goal is to determine the optimal description T_j. As outlined above, we consider this problem in two different settings: in a linear approximation framework, and in a rate–distortion (i.e., compression) framework.

A. Linear Approximation

From the perspective of a selected terminal j, suppose that all other terminals have decided on (arbitrary) suitable approximations of their observations, and the question becomes to optimally choose the approximation to be provided by terminal j,

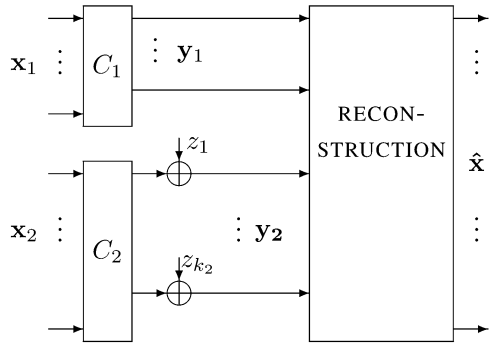

Fig. 2. For a *fixed* transform C_2 (and a fixed noise covariance matrix Σ_z), the *local KLT* finds the best transform C_1.

where we arbitrarily (and without loss of generality) set $j = 1$. Terminal 1 observes M components of the overall data, denoted by $\boldsymbol{x}_1$. The remaining $N - M$ components may be thought of as being merged into one terminal whose observations we denote, for short, by $\boldsymbol{x}_2$. In line with this, we can partition the covariance matrix of the entire vector $\boldsymbol{x}$ into four parts, according to

$$\Sigma_x = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_2 \end{pmatrix} \tag{11}$$

where $\Sigma_1 = E[\boldsymbol{x}_1\boldsymbol{x}_1^T], \Sigma_2 = E[\boldsymbol{x}_2\boldsymbol{x}_2^T],$, and $\Sigma_{12} = E[\boldsymbol{x}_1\boldsymbol{x}_2^T]$. The approximations provided by all other terminals can be expressed as

$$\boldsymbol{y}_2 = C_2\boldsymbol{x}_2 + \boldsymbol{z}_2 \tag{12}$$

where $C_2 \in \mathbb{R}^{k_2 \times (N-M)}$ is a fixed matrix, and $\boldsymbol{z}_2$ is a vector of k_2 jointly Gaussian random variables of mean zero and covariance matrix Σ_z, independent of $\boldsymbol{x}_2$. This is illustrated in Fig. 2. Note that Σ_z is not generally assumed to be of full rank; specifically, we are also interested in the case where $\boldsymbol{z}_2 \equiv 0$. The goal is for the remaining Terminal 1 to select a k-dimensional approximation of the observed vector $\boldsymbol{x}_1$, denoted by $\boldsymbol{y}_1 = C_1\boldsymbol{x}_1$, in such a way as to minimize the resulting overall distortion

$$E[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2]. \tag{13}$$

In order to formulate our solution of this problem, consider the matrix $A \in \mathbb{R}^{(N-M)\times(M+k_2)}$, defined as

$$A = \begin{pmatrix} \Sigma_{12}^T & \Sigma_2 C_2^T \end{pmatrix} \begin{pmatrix} \Sigma_1 & \Sigma_{12}C_2^T \\ C_2\Sigma_{12}^T & C_2\Sigma_2 C_2^T + \Sigma_z \end{pmatrix}^{-1} \tag{14}$$

and let A_1 consist of the first M columns of A, thus, $A_1 \in \mathbb{R}^{(N-M)\times M}$. Moreover, define the matrix $\Sigma_w \in \mathbb{R}^{N\times N}$, as follows:

$$\Sigma_w = \begin{pmatrix} I_M \\ A_1 \end{pmatrix} \left(\Sigma_1 - \Sigma_{12}C_2^T\left(C_2\Sigma_2 C_2^T + \Sigma_z\right)^{-1} C_2\Sigma_{12}^T\right) \cdot \begin{pmatrix} I_M & A_1^T \end{pmatrix}. \tag{15}$$

Note that this matrix has $\operatorname{rank}(\Sigma_w) \leq M$, and denote its eigendecomposition by

$$\Sigma_w = Q_w \operatorname{diag}(\lambda_{w,1}, \lambda_{w,2}, \ldots, \lambda_{w,N}) Q_w^T \tag{16}$$

where

$$\lambda_{w,1} \geq \lambda_{w,2} \geq \cdots \geq \lambda_{w,M} \geq \lambda_{w,M+1} = \cdots = \lambda_{w,N} = 0$$

are the (nonincreasingly ordered) eigenvalues of the matrix Σ_w. Note that since Σ_w is a covariance matrix, its eigenvectors (collected in the matrix Q_w) are real-valued.

Definition 1 (Local KLT): The local KLT of $\boldsymbol{x}_1$ with respect to $\boldsymbol{y}_2 = C_2\boldsymbol{x}_2 + \boldsymbol{z}_2$ is the matrix $C_1 \in \mathbb{R}^{M\times M}$ given by

$$C_1 = \left(Q_w^{(M)}\right)^T \begin{pmatrix} I_M \\ A_1 \end{pmatrix} \tag{17}$$

where $Q_w^{(M)}$ denotes the matrix consisting of the first M columns of Q_w, which is defined in (16), I_M denotes the M-dimensional identity matrix, and A_1 denotes the first M columns of A as defined in (14).

Remark (Nonuniqueness of the Local KLT): It is clear that the local KLT is not unique: it inherits the nonuniqueness of the first M columns of Q_w. More specifically, if the eigenvalues $\lambda_{w,1}, \lambda_{w,2}, \ldots, \lambda_{w,M}$ are all distinct, then the local KLT is unique. If, on the contrary, some of these eigenvalues are equal, then the local KLT is nonunique. However, *any* resulting local KLT is equivalent for the mean-squared error criterion considered in this paper.

In order to provide some intuition for the local KLT, we now establish the following two properties.

Lemma 1: The local KLT has the following properties:

i) $\operatorname{rank}(C_1) = M$, but C_1 is *not* unitary;
ii) the components of $C_1\boldsymbol{x}_1$ are conditionally uncorrelated, given $\boldsymbol{y}_2 = C_2\boldsymbol{x}_2 + \boldsymbol{z}_2$, and have mean zero and conditional variances given by $\lambda_{w,m}$, for $m = 1, 2, \ldots, M$, as in (16), conditioned on $\boldsymbol{y}_2$.

The proof of this lemma is given in Appendix I.

The motivation for defining the local KLT as in Definition 1 is that it provides the optimum k-dimensional approximation for the problem illustrated in Fig. 2. We record the following theorem.

Theorem 2 (Local KLT): The best k-dimensional linear approximation of $\boldsymbol{x}_1$ for a decoder that has access to $C_2\boldsymbol{x}_2 + \boldsymbol{z}_2$ is given by the first k components of the local KLT of $\boldsymbol{x}_1$ with respect to $\boldsymbol{y}_2 = C_2\boldsymbol{x}_2 + \boldsymbol{z}_2$, that is,

$$\boldsymbol{y}_1 = C_1^{(k)}\boldsymbol{x}_1 \tag{18}$$

where $C_1^{(k)}$ denotes the matrix consisting of the first k rows of C_1, as given in Definition 1, i.e.,

$$C_1 = \left(Q_w^{(M)}\right)^T \begin{pmatrix} I_M \\ A_1 \end{pmatrix} \tag{19}$$

and the resulting MSE distortion is given by

$$D = D_2 + \sum_{m=k+1}^{M} \lambda_{w,m} \tag{20}$$

where D_2 is given in (21) at the bottom of the page.

The proof of this theorem is given in Appendix I.

B. Compression

By analogy to Section III-A, consider again the perspective of a selected terminal j, and suppose that all other terminals have decided on (arbitrary) suitable approximations of their observations. In the compression scenario, as explained in Section II-B, we consider sequences $\{\boldsymbol{x}[i]\}_{i\geq 1}$ of source output vectors, where i is the (discrete) time index. We assume that $\{\boldsymbol{x}[i]\}_{i\geq 1}$ is a sequence of independent and identically distributed (i.i.d.) Gaussian random vectors with mean zero and covariance matrix Σ_x. Again, to keep notation simple, we reorder the components of $\boldsymbol{x}$ and denote the observation of the considered terminal j by $\boldsymbol{x}_1[i] = (x_1[i], x_2[i], \ldots, x_M[i])^T$. The remaining components of $\boldsymbol{x}$ will be denoted by

$$\boldsymbol{x}_2[i] = (x_{M+1}[i], x_{M+2}[i], \ldots, x_N[i])^T.$$

We study the problem where terminal j is allowed to use nR bits to encode a sequence of n consecutive observed vectors, denoted by $\{\boldsymbol{x}_1[i]\}_{i=1}^n$. As explained earlier, the key feature of the present *local* consideration is that it has already been decided how $\{\boldsymbol{x}_2[i]\}_{i=1}^n$ is to be compressed, and we are looking for the optimal compression of $\{\boldsymbol{x}_1[i]\}_{i=1}^n$. It is not immediately clear how this should be modeled. For the purpose of this paper, we use the following approach, to be justified to some (though limited) extent in the sequel. Specifically, the effect of compression of $\{\boldsymbol{x}_2[i]\}_{i=1}^n$ is captured by providing the decoder with a *noisy* sequence

$$\boldsymbol{y}_2[i] = C_2\boldsymbol{x}_2[i] + \boldsymbol{z}_2[i] \tag{22}$$

where $\{\boldsymbol{z}_2[i]\}_{i=1}^n$ is a sequence of i.i.d. Gaussian random vectors of mean zero and covariance matrix Σ_z. Again, $C_2 \in \mathbb{R}^{k_2\times(N-M)}$. This is illustrated in Fig. 3.

To justify the model considered in (22), one may think of the case where the remaining terminals have compressed $\{\boldsymbol{x}_2[i]\}_{i=1}^n$ using a rate-distortion optimal source code (but entirely ignoring $\boldsymbol{x}_1$). The effect of such a code (for a Gaussian source under mean-squared error) is equivalent to a linear transformation, followed by additive white Gaussian noise, see, e.g., [6, p. 345].

The distortion to be minimized can be expressed as

$$D_n = E\left[\frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}[i] - \hat{\boldsymbol{x}}[i]\|^2\right] \tag{23}$$

subject to the constraints that, i), the codeword produced by terminal 1 may only depend on $\{\boldsymbol{x}_1[i]\}_{i=1}^n$, and ii), the coding rate

$$D_2 = \operatorname{trace}\left(\Sigma_2 - \begin{pmatrix}\Sigma_{12}^T & \Sigma_2 C_2^T\end{pmatrix}\begin{pmatrix}\Sigma_1 & \Sigma_{12}C_2^T \\ C_2\Sigma_{12}^T & C_2\Sigma_2 C_2^T + \Sigma_z\end{pmatrix}^{-1}\begin{pmatrix}\Sigma_{12} \\ C_2\Sigma_2\end{pmatrix}\right). \tag{21}$$

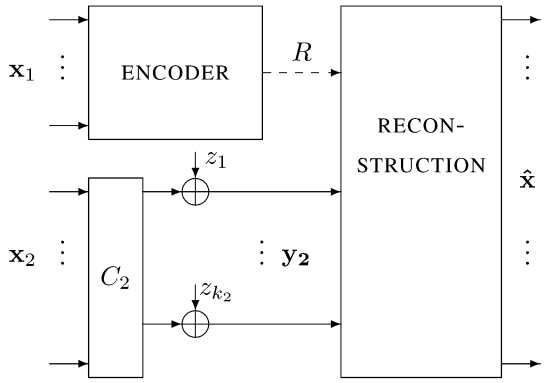

Fig. 3. Terminal 1's local perspective on the compression problem considered in this paper.

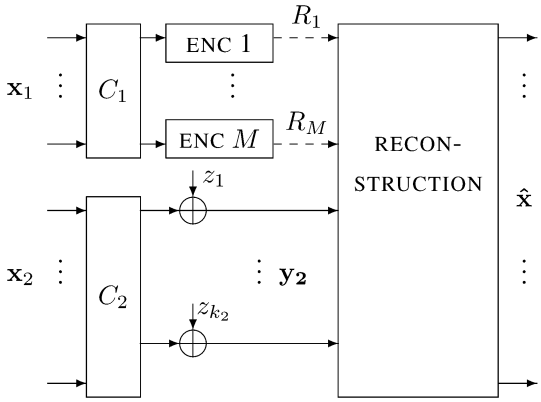

Fig. 4. An optimal architecture for the encoder in Fig. 3 consists in applying a local KLT to $\boldsymbol{x}_1$ and separately compressing each component in the transform domain.

used by terminal 1 may not exceed nR bits to encode the entire (length-n) source sequence. The reconstruction sequence $\hat{\boldsymbol{x}}[i]$ is generated based on the codeword provided by terminal 1 *and* the value of the side information sequence $\{\boldsymbol{y}_2[i]\}_{i=1}^n$.

In this paper, we study the performance in the (information-theoretic) limit, that is, as $n \to \infty$. With reference to (23), we denote $D \stackrel{\text{def}}{=} D_\infty$. The minimum rate $R_{\text{local}}(D)$ required to reconstruct the vector source $\{\boldsymbol{x}[i]\}_{i=1}^\infty$ at distortion level D is given by (details are given in Appendix II)

$$R_{\text{local}}(D) = \min I(\boldsymbol{x}_1; u \,|\, \boldsymbol{y}_2) \tag{24}$$

where the minimization is over all ("auxiliary") random vectors $\boldsymbol{u}$ for which $p(\boldsymbol{u}, \boldsymbol{x}_1, \boldsymbol{y}_2) = p(\boldsymbol{u}|\boldsymbol{x}_1)p(\boldsymbol{x}_1, \boldsymbol{y}_2)$ and for which

$$E[||\boldsymbol{x} - E[\boldsymbol{x} \,|\, \boldsymbol{u}, y_2]||^2] \le D. \tag{25}$$

The solution to this minimization problem can be characterized by complete analogy to our derivation in Section III-A, namely, in terms of the eigenvalues of the matrix Σ_w as defined in (15), as follows.

Theorem 3: The rate required to encode $\{\boldsymbol{x}_1[i]\}_{i\ge 1}$ for a decoder that has access to $\boldsymbol{y}_2[i] = C_2\boldsymbol{x}_2[i] + \boldsymbol{z}_2[i]$ is given by

$$R_{\text{local}}(D) = \sum_{m=1}^{M} \max\left\{\frac{1}{2}\log_2 \frac{\lambda_{w,m}}{d_m}, 0\right\}, \quad \text{for } D \ge D_2 \tag{26}$$

where D_2 is given in (21), and

$$d_m = \begin{cases} \theta, & \text{if } \theta < \lambda_{w,m} \\ \lambda_{w,m}, & \text{if } \theta \ge \lambda_{w,m} \end{cases} \tag{27}$$

where θ is chosen such that $\sum_{m=1}^M d_m + D_2 = D$. Note that it is not possible to attain a distortion $D < D_2$.

The proof of this theorem is an extension of [45] and is given in Appendix II.

This theorem establishes that the encoder in Fig. 3 can be broken into two stages: a *linear precoding* stage, consisting of applying the transform (matrix) C_1 (i.e., the local KLT) to $\boldsymbol{x}_1$ (with respect to $\boldsymbol{y}_2$), followed by *separate compression* of the components in the transform domain. This is illustrated in Fig. 4.

Moreover, in the proof of the theorem, it is also found that the auxiliary random vector $\boldsymbol{u}$ (with conditional distribution $p(\boldsymbol{u} \,|\, \boldsymbol{x}_1)$) that solves the minimization problem specified by (24)–(25) is jointly Gaussian with the source vector $\boldsymbol{x}$. To interpret this, consider Fig. 4: It is consistent to replace the encoder boxes (the boxes labeled ENC $1, \ldots,$ ENC M) by additive Gaussian noises, in the sense that the resulting overall distortion will be the same. Hence, we obtain an overall *symmetric* picture, in the sense that if terminals $2, 3, \ldots, L$ provide noisy observations, where the (additive) noise has a Gaussian distribution, then the optimum encoding for terminal 1 is found to *also* be characterized by providing noisy observations, where the noise has a Gaussian distribution. This perspective will prove to be useful in the distributed setting in Section III-B. Therefore, we record the following corollary.

Corollary 4: For a fixed rate R to be used by the encoder, let $k \le M$ be the largest integer satisfying

$$d_k \stackrel{\text{def}}{=} G_k(\lambda_w) 2^{-\frac{2R}{k}} \le \lambda_{w,m}, \quad \text{for } m = 1, 2, \ldots, k \tag{28}$$

where $G_k(\lambda_w) = \left(\prod_{m=1}^k \lambda_{w,m}\right)^{1/k}$ is the geometric mean of the k largest eigenvalues of the matrix Σ_w. Moreover, denote the first k rows of the local KLT of $\boldsymbol{x}_1$ with respect to $\boldsymbol{y}_2$ (see Definition 1) by $C_1^{(k)}$. Then, the auxiliary random vector $\boldsymbol{u}$ (with conditional distribution $p(\boldsymbol{u} \,|\, \boldsymbol{x}_1)$) that solves the minimization problem specified by (24)–(25) can be written as

$$\boldsymbol{u} = C_1^{(k)}\boldsymbol{x}_1 + \boldsymbol{z}_1 \tag{29}$$

where $\boldsymbol{z}_1$ is a Gaussian random vector with mean zero and diagonal covariance matrix, with diagonal entries

$$\zeta_m^2 = \frac{\lambda_{w,m} d_k}{\lambda_{w,m} - d_k} \tag{30}$$

for $m = 1, 2, \ldots, k$. The resulting overall distortion can be expressed as

$$D = kd_k + \sum_{m=k+1}^{M} \lambda_{w,m} + D_2$$

where D_2 is given in (21).

The proof of this corollary is given in Appendix II.

IV. Special Cases

A. The Partial-Observation KLT ($C_2 = 0$)

Consider the case when C_2 is the all-zero matrix, and hence, all other terminals are cut off. This is illustrated in Fig. 5. In this case, the matrix A in (14) simplifies to $A = \Sigma_{12}^T \Sigma_1^{-1}$. Moreover, $A_1 = A$, and hence, we find $\Sigma_w = \binom{I_M}{A} \Sigma_1 (I_M, A^T)$. For this case, the local KLT thus becomes particularly simple.

Definition 2 (Partial-Observation KLT): The *partial-observation KLT* of $\boldsymbol{x}_1$ with respect to $\boldsymbol{x}_2$ is the linear transform characterized by the matrix

$$C_p = Q^T \begin{pmatrix} I_M \\ \Sigma_{12}^T \Sigma_1^{-1} \end{pmatrix} \tag{31}$$

where Q is the unitary matrix for which

$$Q^T \binom{I_M}{\Sigma_{12}^T \Sigma_1^{-1}} \Sigma_1 (I_M, (\Sigma_{12}^T \Sigma_1^{-1})^T) Q$$

is diagonal. The transformed version of $\boldsymbol{x}_1$ will be denoted by $\boldsymbol{y}_1 = C_p \boldsymbol{x}_1$.

Remark 2 (Nonuniqueness of the Partial-Observation KLT): In line with Remark 1, it should be noted that the partial-observation KLT is not unique.

Hence, we get the following corollary to Theorem 2.

Corollary 5: The best k-dimensional linear approximation of $\boldsymbol{x}_1$ for a decoder that needs to reconstruct the entire vector $\boldsymbol{x} = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T)^T$ with respect to mean-squared error is given by the first k components of the partial-observation KLT of $\boldsymbol{x}_1$ with respect to $\boldsymbol{x}_2$.

By analogy, a similar corollary can be given for the case of Theorem 3.

It is intuitively clear that in the partial-observation (or *subsampling*) scenario of Fig. 5, it is suboptimal to simply apply the "marginal" KLT to the observations $\boldsymbol{x}_1$, acting as if $\boldsymbol{x}_2$ did not exist. To further illustrate this point, we consider simple examples.

Example 1 (Approximation): A toy example illustrating the basic issue is the following: Suppose that a Gaussian random vector $\boldsymbol{x}$ has mean zero and the following covariance matrix:

$$\Sigma_x = \begin{pmatrix} \sigma_1^2 & 0 & 0.1 & 0.1 \\ 0 & 0.1 & 0.25 & 0 \\ 0.1 & 0.25 & 1 & 0.25 \\ 0.1 & 0 & 0.25 & 1 \end{pmatrix} \tag{32}$$

Suppose that the first two components are observed by terminal 1, i.e., $M = 2$. The terminal is asked to provide a one-dimensional approximation. For $\sigma_1^2 = 0.11$, the marginal KLT is the identity matrix since the first two components are uncorrelated. Then, selecting the eigenvector corresponding to the largest eigenvalue of Σ_1 incurs a distortion of $D_{m\,\mathrm{klt}} \approx 1.9182$. By contrast, the partial-observation KLT is found to be

$$C_p \approx \begin{pmatrix} 1.1119 & 2.6353 \\ 1.1902 & -0.5524 \end{pmatrix} \tag{33}$$

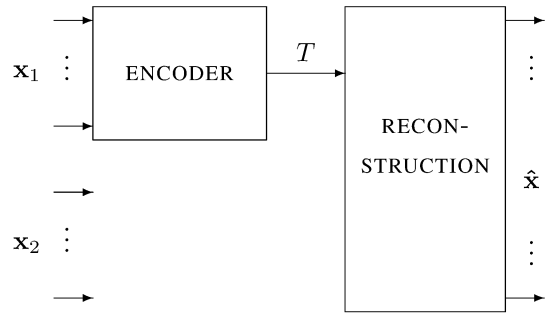

Fig. 5. Special case: The partial-observation (or "subsampling") KLT. The components of $\boldsymbol{x}_2$ are not observed, but need to be reconstructed..

which is substantially different from the marginal KLT. Retaining the first component of $C_p \boldsymbol{x}_1$, the resulting distortion is found to be $D_{p\,\mathrm{klt}} \approx 1.3795$, substantially smaller than the distortion incurred by the marginal KLT.

Example 2 (Compression): Consider again the covariance matrix given in Example 1, with $\sigma_1^2 = 0.11$. The first two components are sampled and can be encoded using a total rate R. The systematic error (as defined in (21)) for this example is $D_2 = 1.1932$. The rate–distortion tradeoff is shown in Fig. 6. The solid line is the rate–distortion function $R_1(D)$ (i.e., incorporating the partial-observation KLT). The dashed line is the performance for a compression scheme that ignores the hidden part when encoding. At decoding time, the hidden part is estimated optimally from the available information. The figure witnesses a clear advantage for the partial-observation KLT, illustrating the fact that the hidden part does alter the compression problem significantly. In the limit of low rates, as $R \to 0$, it is clear that both schemes have the same performance, since no information is transmitted. In the limit of high rates, both schemes end up encoding the observations perfectly, and again, the same distortion results.

Remark 3 (Best Sensor Placement): For given statistics Σ_x and desired distortion D, what is the best "placement" of M sensors? In other words, what choice of M components of $\boldsymbol{x}$ minimizes the rate required for the desired distortion D? The solution to this problem is given by Theorem 1: Compute $R_{\mathrm{local}}(D)$ for all subsets of M components of the vector $\boldsymbol{x}$. (For the asymmetric scenario of Example 2, it is easily verified that the best sensor placement is to sample the last two components. This is intuitively clear from the covariance matrix in (32): the last two components have by far the largest variances.)

B. The Conditional KLT ($C_2 = I_{N-M}$ and $k_2 = N - M$)

In this subsection, we study the scenario of Fig. 7: All other terminals provide the reconstruction point with their exact observations. In this case, the local KLT can be simplified. Specifically, since $\boldsymbol{y}_2 = \boldsymbol{x}_2$, we find that $A_1 = 0$. This implies that Σ_w (as in (15)) simplifies to

$$\Sigma_w = \Sigma_1 - \Sigma_{12} \Sigma_2^{-1} \Sigma_{12}^T \tag{34}$$

and the local KLT takes the following simple shape.

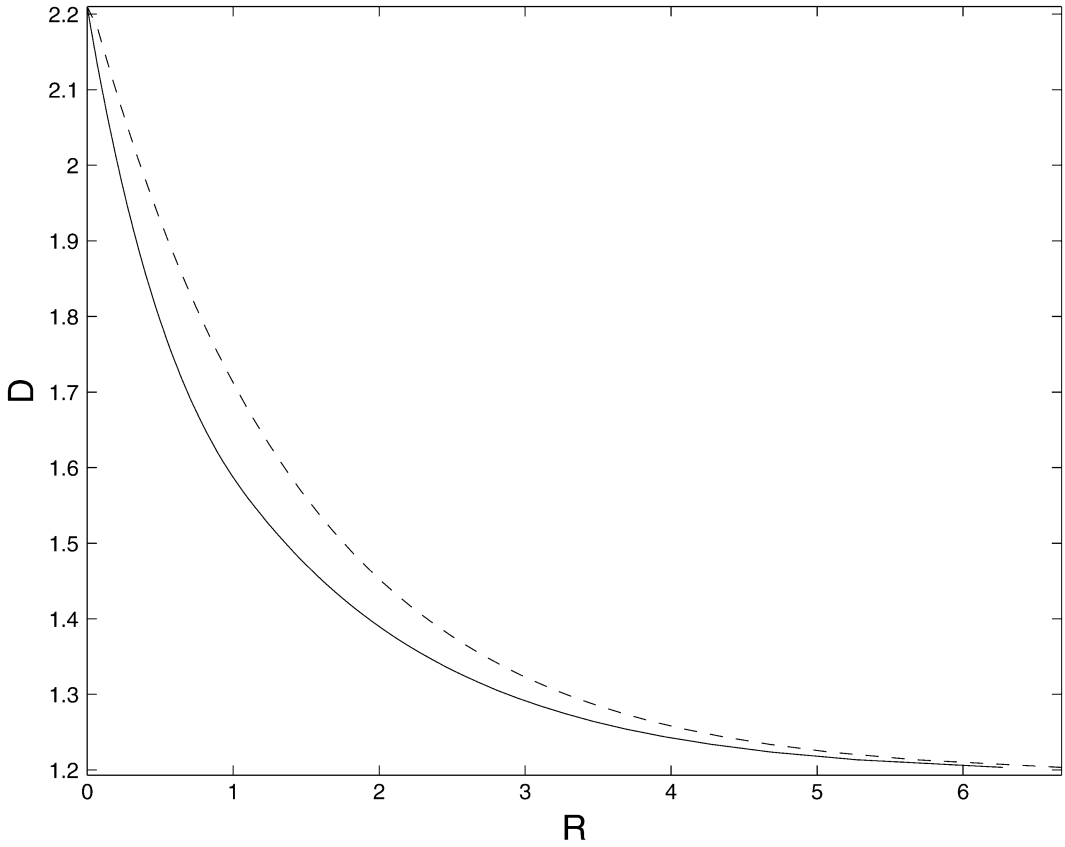

Fig. 6. The rate–distortion behavior for the partially observed process of Example 2. The solid line represents the (optimal) partial observation KLT, while the dashed line is the case of the marginal KLT. The rate is given in bits.

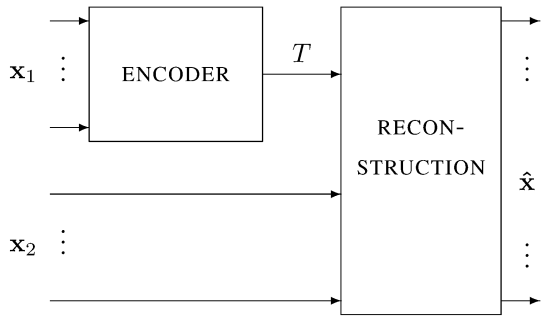

Fig. 7. Special case: The conditional KLT. The components $\boldsymbol{x}_2 = (x_{M+1}, \ldots, x_N)$ are not observed by the encoder, but they are available at reconstruction time.

Definition (Conditional KLT): The *conditional KLT* of $\boldsymbol{x}_1$ with respect to $\boldsymbol{x}_2$ is the linear transform characterized by the matrix $C_c \in \mathbb{R}^{M\times M}$ that satisfies

$$C_c \Sigma_w C_c^T = \operatorname{diag}(\lambda_{w,1}, \ldots, \lambda_{w,M}). \tag{35}$$

The transformed version of $\boldsymbol{x}_1$ will be denoted by $\boldsymbol{y}_1 = C_c \boldsymbol{x}_1$.

Remark 4 (Nonuniqueness of the Conditional KLT): In line with Remark 1, it should be noted that the conditional KLT is not unique.

Lemma 6: The conditional KLT C_c has the following properties:

1) C_c is an orthonormal transform;
2) the components of the vector $\boldsymbol{y}_1$ are conditionally uncorrelated given $\boldsymbol{x}_2$.

It is immediately clear by construction that C_c will be orthonormal. The second property was established in Lemma 1.

Hence, we get the following corollary to Theorem 2.

Corollary 7: The best k-dimensional linear approximation of $\boldsymbol{x}_1$ for a decoder that has access to $\boldsymbol{x}_2$ and needs to reconstruct the entire vector $\boldsymbol{x}$ with respect to mean-squared error is given by the first k components of the conditional KLT of $\boldsymbol{x}_1$ with respect to $\boldsymbol{x}_2$.

By analogy, a similar corollary can be given for the case of Theorem 3.

Example 3 (Approximation): A toy example illustrating the basic issue is the following: Suppose that a Gaussian random vector $\boldsymbol{x}$ has mean zero and the covariance matrix specified in (32). Suppose that the first two components are sampled by the terminal, i.e., $M = 2$. The terminal is asked to provide a one-dimensional approximation. For $\sigma_1^2 = 0.1$, applying the marginal (usual) KLT to the first two components is simple in this example: the first two components are uncorrelated, hence the KLT is the identity. Selecting the eigenvector corresponding to the

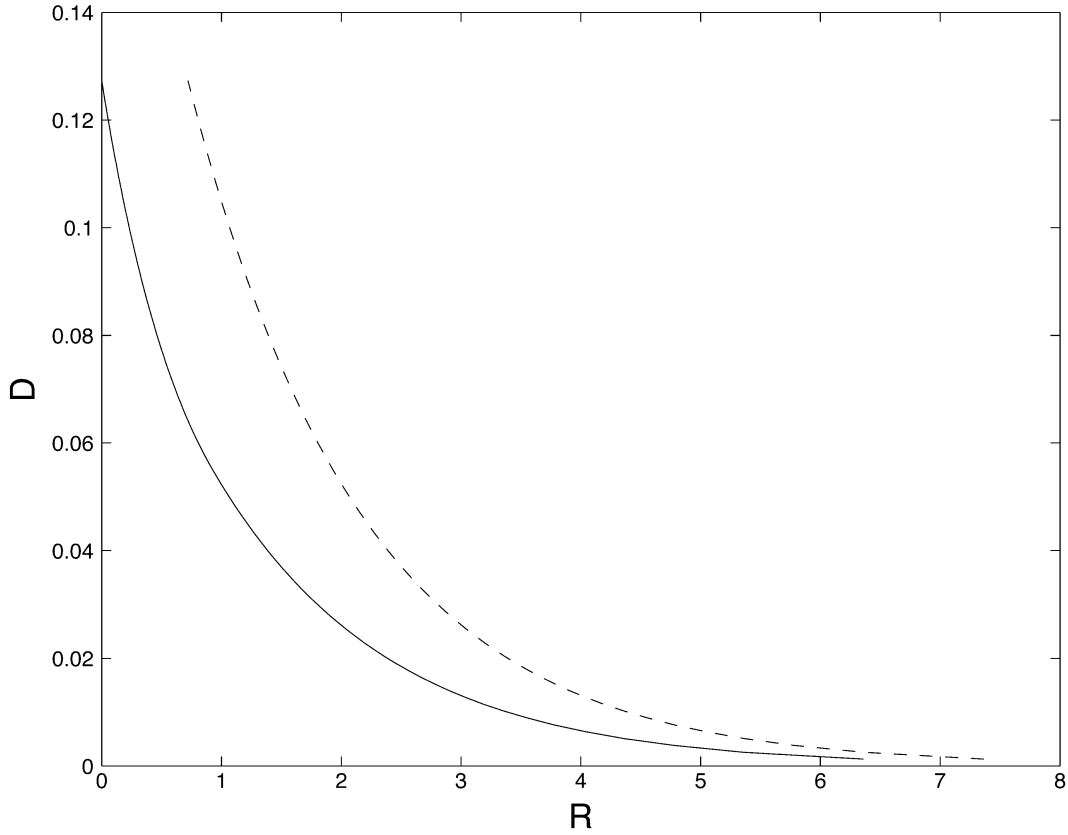

Fig. 8. Rate–distortion behavior for compression with side information in Example 4. The solid line is the performance of the conditional KLT, while the dashed line shows the one of the marginal KLT. The rate is given in bits.

larger eigenvalue of Σ_S incurs a distortion of $D_{\mathrm{klt}} \approx 0.0720$. Using the conditional KLT discussed in this section, and hence making the optimal choice, results in a distortion of $D_{c\,\mathrm{klt}} \approx 0.0264$, and the transform is

$$C_c \approx \begin{pmatrix} -0.9447 & 0.3280 \\ -0.3280 & -0.9447 \end{pmatrix}. \tag{36}$$

Example 4 (Compression): Consider again the covariance matrix specified in (32), with $\sigma_1^2 = 0.11$, and suppose that the first two components are sampled by the terminal, i.e., $M = 2$. The task is to encode these first two components using a rate R. The resulting rate–distortion function is the solid line in Fig. 8. For comparison, the dashed line shows the rate–distortion performance for a scheme that ignores the side information both at the encoder and at the decoder, in other words, a scheme that simply compresses (and reconstructs) the sampled vector $\boldsymbol{x}_1$, using a rate R. The performance of this scheme is hence determined by the rate–distortion function for Gaussian vectors under mean-squared error (see, e.g., [6, p. 348]). As expected, there is a significant difference between the two curves.

V. The Distributed KLT Algorithms

The local perspective derived in Section III and further explored in the previous section suggests an iterative approach to the problem of finding the best distributed approximation to the KLT: in turn, each terminal optimizes its local encoder, while all other encoders are held fixed (and known). This is an *off-line* calculation, where the covariance matrix Σ_x is kept fixed. Such a "round-robin" optimization, while natural, may or may not be optimal, depending on the shape of the cost function. That is, while each step is optimal (as in Theorems 2 and 3, respectively) the sequence of steps might lead to a locally stable point that is not a global optimum. This question is central to distributed optimization, and has in general no simple answer, and we will thus explore it through several examples in the sequel.

A. Linear Approximation

Let us now return to the problem illustrated in Fig. 1: There are L terminals. Each terminal observes a part $\boldsymbol{x}_\ell$ of the entire vector $\boldsymbol{x}$, for $\ell = 1, 2, \ldots, L$, as defined in Section II-A: The first terminal observes the first M_1 components of $\boldsymbol{x}$ and has to provide a k_1-dimensional approximation. The second terminal observes the next M_2 components of $\boldsymbol{x}$ and has to provide a k_2-dimensional approximation, and so on. The goal is to find the optimum approximation spaces for each terminal.

From Theorem 2, we know the best approximation space for a selected terminal, when all other terminals have fixed their approximation spaces. For further reference, we restate Theorem 2 in more general notation in the shape of the following corollary.

Corollary 8: For fixed $C_\ell, \ell = 1, 2, \ldots, L, \ell \neq j$, the best k_j-dimensional approximation that terminal j can provide is de-

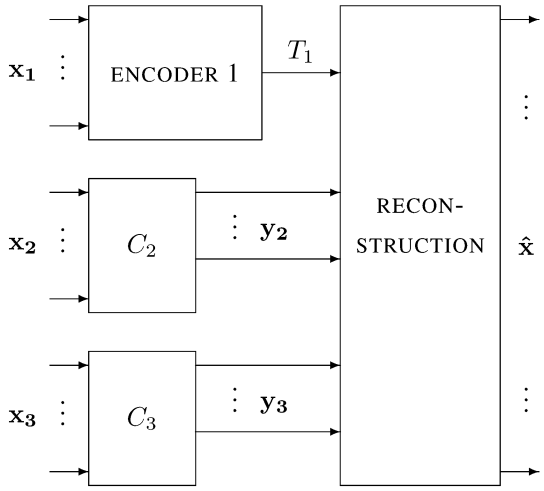

Fig. 9. An iteration of the distributed KLT algorithm: The transform matrices C_2 and C_3 are kept fixed while Encoder 1 is chosen optimally.

termined by the matrix C_j containing the first k_j rows of the local KLT of $\boldsymbol{x}_j$ with respect to

$$\boldsymbol{y}_j^c \stackrel{\text{def}}{=} \{C_\ell \boldsymbol{x}_\ell\}_{\ell=1,\ell\neq j}^{L}. \tag{37}$$

Proof: This corollary follows directly from Theorem 2. □

As pointed out in Section II-A, there does not seem to be a simple and direct solution to find the optimal approximation spaces $C_1, C_2, \ldots, C_L$, but Corollary 8 suggests an iterative procedure to find an approximate solution, as follows.

Algorithm 1 (Distributed KLT for Linear Approximation): Set $n = 0$ and initialize by selecting arbitrary matrices $C_\ell^{(n)} \in \mathbb{R}^{k_\ell \times M_\ell}$, for $\ell = 1, 2, \ldots, L$. Then,

1) Set $n = n + 1$, and for each $j, j = 1, 2, \ldots, L$:
   a) Set $C_j^{(n)}$ to be the best k_j-dimensional approximation space, as described in Corollary 8.

Terminate when the difference in the resulting mean- squared error distortion is smaller than a fixed tolerance value ϵ. □

Remark 5: As we illustrate in Section VI, in general, different initializations of the matrices $C_1, C_2, \ldots, C_L$ may lead to different outcomes of the algorithm. However, since the computations are done off-line, one may run Algorithm 1 many times, using techniques such as simulated annealing, and thus enhance the chance of finding the global optimum.

This algorithm is illustrated in Fig. 9. The figure shows one iteration of the algorithm: The transform matrices C_2 and C_3 are kept fixed while Encoder 1 is chosen optimally. By Corollary 8, the optimal choice of Encoder 1 is indeed composed of a transform matrix C_1^*, followed by an appropriate choice of k_1 components in the transform domain.

The key property of this algorithm is the following.

Theorem 9 (Convergence of the Distributed KLT Algorithm): Denote the transform matrices provided by Algorithm 1 after iteration n by $C_\ell^{(n)} \in \mathbb{R}^{k_\ell \times M_\ell}$, for $\ell = 1, 2, \ldots, L$. Denote the minimum mean-squared error estimate of $\boldsymbol{x}$ based on the observations $\{C_\ell^{(n)} \boldsymbol{x}_\ell\}_{\ell=1}^{L}$ by $\hat{\boldsymbol{x}}^{(n)}$. Then

$$E\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}^{(n)}\|^2\right] \geq E\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}^{(n+1)}\|^2\right] \tag{38}$$

i.e., the distortion is a nonincreasing function of the iteration number.

Proof: The theorem follows directly from Corollary 8: Suppose that in iteration n, the transform matrix C_j is being updated. That is, C_j is selected according to Corollary 8. Note that the corollary imposes *no restrictions* on C_j; in particular, the current value of the matrix C_j lies *inside* the optimization space. Therefore, the distortion cannot increase. □

This theorem implies that the distributed KLT algorithm will converge to a stable point that is either a saddle point or a local minimum, but it clearly cannot guarantee convergence to a globally optimal configuration of approximation spaces. Before we study the convergence behavior in more detail, let us illustrate our finding with a few simple examples.

Example 5: A toy example illustrating the basic issue is the following: Suppose that a Gaussian random vector $\boldsymbol{x}$ has mean zero and the covariance matrix specified in (32). Suppose that the first two components are sampled by terminal 1, i.e., $\boldsymbol{x}_1 = (x_1, x_2)$, and the last two components by terminal 2, i.e., $\boldsymbol{x}_2 = (x_3, x_4)$. Both terminals are asked to provide a one-dimensional approximation. For $\sigma_1^2 = 0.11$, if each terminal applies the marginal KLT to its observation a distortion of $D_{m\,\text{klt}} \approx 0.8207$ is incurred. Note that the KLTs are simple: terminal 1 applies the identity transform, and terminal 2 applies

$$C_{2,\text{marginal}} \approx \begin{pmatrix} 0.7071 & 0.7071 \\ -0.7071 & 0.7071 \end{pmatrix}. \tag{39}$$

Using the distributed KLT algorithm discussed in this section, and hence making the optimal choice, results in a distortion of $D_{d\,\text{klt}} \approx 0.3457$, and the transforms are

$$C_1 \approx \begin{pmatrix} 0.6968 & 2.6205 \\ -0.9820 & 0.1996 \end{pmatrix} \tag{40}$$

$$C_2 \approx \begin{pmatrix} 0.2385 & 0.9758 \\ 0.9717 & -0.2366 \end{pmatrix}. \tag{41}$$

The convergence of the distributed KLT algorithm, when C_2 is initially the identity matrix, is shown in Fig. 10. The figure shows the error in the "middle" of the nth iteration (i.e., after the nth update to C_1, but before the nth update to C_2), and at the end of the n iteration.

Finally, if the entire vector could be handled jointly and the goal is to find the best two-dimensional approximation, a distortion of $D_{\text{joint klt}} \approx 0.1243$ is feasible.

Example 6: Suppose Σ_x is a Toeplitz matrix with first row $(1, \rho, \rho^2, \ldots)$, $\boldsymbol{x}_1$ contains the odd-indexed components of $\boldsymbol{x}$, and $\boldsymbol{x}_2$ the even-indexed components. For $N = 40, M = 20$, $k_1 = k_2 = 10$, and $\rho = 0.7$, the marginal KLT, i.e., the standard KLT applied to each part separately, leads to a distortion $D_{m\,\text{klt}} \approx 8.3275$, while the distributed KLT gives $D_{d\,\text{klt}} \approx$

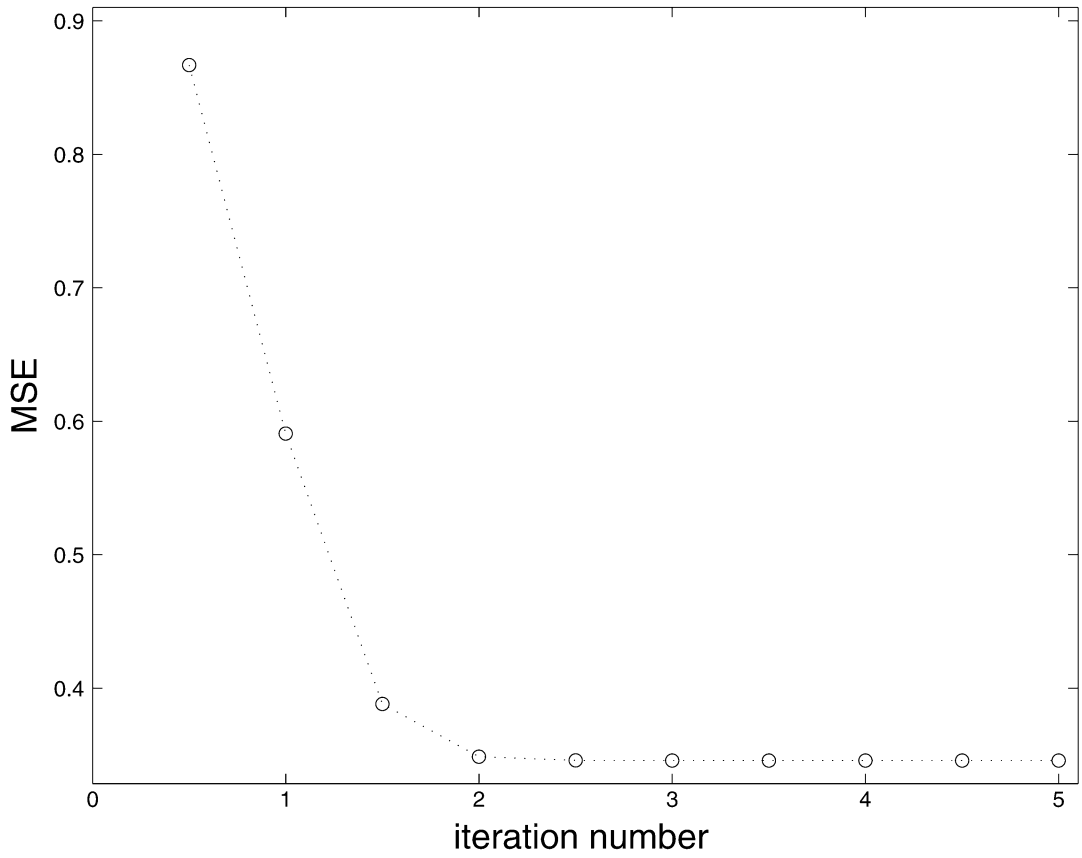

Fig. 10. Convergence for Example 5.

6.8464. Hence, even in this seemingly symmetric scenario, the distributed KLT is *substantially* different from the standard, i.e., joint KLT. For comparison, the full standard KLT, applied to the entire vector $\boldsymbol{x}$, would give $D \approx 4.5195$.

Example 7 (Computational Perspective): Applying the KLT to a vector of length N requires N^2 multiplications and N^2 additions. Usually, we are only interested in the first k elements, and hence, the computational cost becomes kN multiplications and kN additions. The distributed KLT provides a block-diagonal approximation of the KLT. Suppose the full vector of length N is decomposed into subvectors of lengths $M_1, M_2, \ldots, M_L$, and each subvector is approximated by a k_ℓ-dimensional vector, for $\ell = 1, 2, \ldots, L$. Then, applying the distributed KLT to the vector requires only $\sum_{\ell=1}^{L} k_\ell M_\ell$ multiplications, and an equal number of additions. To gain insight, suppose that $M_\ell = N/L$ and $k_\ell = k/L$ are both integer. Then, the distributed KLT requires kN/L multiplications, and an equal number of additions. For large L, i.e., for a highly distributed KLT, this clearly is considerably smaller than the kN multiplications and additions required by the full standard KLT. Hence, the distributed KLT permits a tradeoff between computational cost and approximation quality (measured in mean-squared error), though this need not be the optimum such tradeoff. The following numerical example illustrates this in more detail.

Let us consider the case $N = 60$ and $\sum_{\ell=1}^{L} k_\ell = 12$. Moreover, let Σ_x be Toeplitz with first row $(1, \rho, \rho^2, \ldots, \rho^{N-1})$, for $\rho = 0.9$. Then, we can study the behavior of the resulting mean-squared error as the M_ℓ are varied, and hence, the amount of "distributedness" is increased. In particular, we consider the case where all M_ℓ are equal. For example, we may select $L = 4$, and hence, $M_\ell = 15$ and $k_\ell = 3$ for $\ell = 1, 2, 3, 4$. This requires 180 multiplications. The resulting distortion is found to be $D_{d\,\mathrm{klt}} \approx 6.4965$. Fig. 11 graphically illustrates the outcome, comparing the marginal KLT to the distributed KLT, and revealing respective gains. Note that by selecting the covariance matrix Σ_x appropriately, these gains can be made large.

B. Compression

This subsection addresses the distributed compression problem, illustrated in Fig. 1. Terminal ℓ observes n source output vectors of length M_ℓ and individually provides a binary sequence T_ℓ that appears at rate R_ℓ bits per observed source vector, for $\ell = 1, 2, \ldots, L$. The reconstruction point produces a sequence $\{\hat{\boldsymbol{x}}_{T_1, T_2, \ldots, T_L}[i]\}_{i=1}^{n}$ such as to minimize the distortion

$$D_n(R_1, R_2, \ldots, R_L) = \frac{1}{n} \sum_{i=1}^{n} E[\|\boldsymbol{x}[i] - \hat{\boldsymbol{x}}_{T_1, T_2, \ldots, T_L}[i]\|^2]. \tag{42}$$

Again, the goal is to analyze the problem as $n \to \infty$. In particular, we will consider the case of a *fixed total* (sum) rate $\sum_{\ell=1}^{L} R_\ell = R_{\mathrm{tot}}$; the goal is to determine and achieve the smallest possible distortion.

Unfortunately, this question cannot be answered in a conclusive manner: the distributed compression problem illustrated in

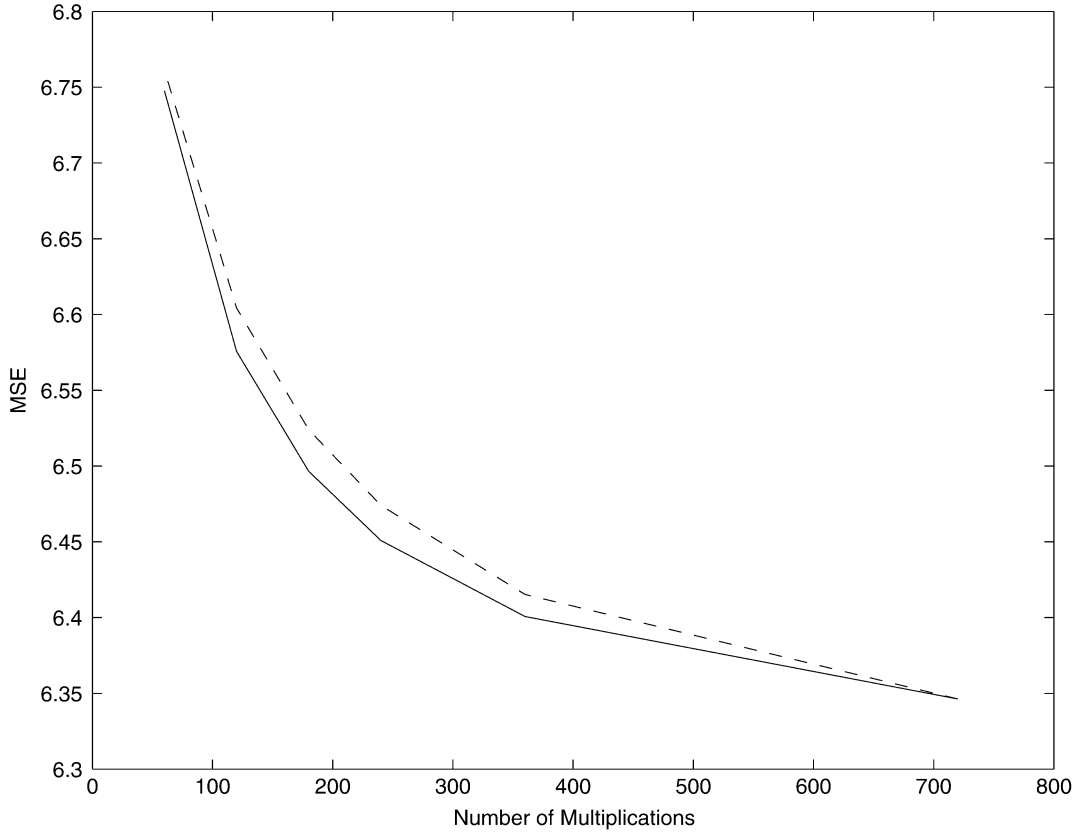

Fig. 11. The resulting distortion (y-axis) versus the number of multiplications (x-axis) for Example 7. The solid line is the distributed KLT, the dashed line is the marginal KLT.

Fig. 1 is an open problem to date (except in the jointly Gaussian two-source $(N = 2)$ case [20]). The best currently known achievable rate regions and outer bounds appear in [10], as discussed in somewhat more detail in Section II-B.

Instead, in this paper, we resort to a consideration of *achievable* rate–distortion behavior. We will proceed along similar lines as in Section III-B. There, we used the well-known information-theoretic result given by (24)–(25) and developed it for the vector case under consideration, establishing that an optimal architecture is for the remaining terminal to apply the *local* KLT and then separately compress each component in the transform domain. Here, by analogy, we will consider a well-known information-theoretic achievable rate region. However, *by contrast*, there is no proof that this region is the optimal region. The region can be described as follows.

Theorem 10 ([44], [12], [15]): At fixed total rate R_{tot}, any distortion

$$E[||\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_L)||^2] \tag{43}$$

is achievable, where the auxiliary random vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_L$ satisfy

$$p(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_L, \boldsymbol{x}) = p(\boldsymbol{u}_1|\boldsymbol{x}_1)p(\boldsymbol{u}_2|\boldsymbol{x}_2)\cdots p(\boldsymbol{u}_L|\boldsymbol{x}_L)p(\boldsymbol{x}) \tag{44}$$

and

$$I(\boldsymbol{x}; \boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_L) \leq R_{\text{tot}}. \tag{45}$$

This theorem follows in a straightforward manner from the work of Housewright and Omura (see [44]) and of Berger and Tung (see [12, Ch. 6, Sec. 6.1]). Details can be found in [15, Section IV].

The goal of our work is to determine the conditional distributions $p(\boldsymbol{u}_\ell|\boldsymbol{x}_\ell)$ of the auxiliary random vectors $\boldsymbol{u}_\ell$, for $\ell = 1, 2, \ldots, L$, and hence, the architecture of the coding scheme, such as to *minimize* the distortion. In this paper, we again take an iterative approach in which we *fix* all but one of these conditional distributions, leading to a set of auxiliary random vectors denoted by

$$\boldsymbol{u}_j^c \stackrel{\text{def}}{=} \{\boldsymbol{u}_\ell\}_{\ell=1, \ell\neq j}^{L} \tag{46}$$

and determine the optimal $\boldsymbol{u}_j$. More specifically, if we select $\boldsymbol{u}_j^c$ to be jointly Gaussian with $\boldsymbol{x}$, then we know from Theorem 3 that the overall situation will be *symmetric* in the sense that the remaining $\boldsymbol{u}_j$ should *also* be selected jointly Gaussian with $\boldsymbol{x}$. Therefore, we restate Theorem 3 in more general notation in the shape of the following corollary.

Corollary 11: Assume the following:

a) For $\ell = 1, 2, \ldots, L, \ell \neq j$, the conditional distributions $p(\boldsymbol{u}_\ell|\boldsymbol{x}_\ell)$ are Gaussian with mean zero. Hence, they can be characterized by $\boldsymbol{u}_\ell = C_\ell \boldsymbol{x}_\ell + \boldsymbol{z}_\ell$, where C_ℓ is a transform matrix and $\boldsymbol{z}_\ell$ is a random vector independent of $\boldsymbol{x}$, with zero mean and covariance matrix $\Sigma_{z,\ell}$. Define

$$\boldsymbol{u}_j^c \stackrel{\text{def}}{=} \{C_\ell \boldsymbol{x}_\ell + \boldsymbol{z}_\ell\}_{\ell=1, \ell\neq j}^{L} \tag{47}$$

and $R_0 \stackrel{\text{def}}{=} I(\boldsymbol{x}; \boldsymbol{u}_j^c)$.

b) A total rate budget of R_{tot} is imposed such that $R_0 \leq R_{\text{tot}}$.

Then, the optimal (distortion-minimizing) conditional distribution for the remaining terminal, $p(\boldsymbol{u}_j \,|\, \boldsymbol{x}_j)$, is also Gaussian (with mean zero), and is found by first determining the local KLT C of $\boldsymbol{x}_j$ with respect to $\boldsymbol{u}_j^c$. Let k be the largest integer satisfying

$$d_k \stackrel{\text{def}}{=} G_k(\lambda_w) 2^{-\frac{2(R_{\text{tot}}-R_0)}{k}} \leq \lambda_{w,m}, \quad \text{for } m = 1, 2, \ldots, k \tag{48}$$

where $G_k(\lambda_w) = \left(\prod_{m=1}^{k} \lambda_{w,m}\right)^{1/k}$ is the geometric mean of the largest k eigenvalues of the matrix Σ_w. Then, the conditional distribution $p(\boldsymbol{u}_j \,|\, \boldsymbol{x}_j)$ of the auxiliary random vector $\boldsymbol{u}_j$ minimizing (43) subject to (44) and (45) is characterized by $\boldsymbol{u}_j = C_j \boldsymbol{x}_j + \boldsymbol{z}_j$, where C_j contains the first k rows of C, and the Gaussian vector $\boldsymbol{z}_j$ has mean zero and covariance matrix

$$\Sigma_{z,j} = \operatorname{diag}\left(\left\{\frac{\lambda_{w,m} d_k}{\lambda_{w,m} - d_k}\right\}_{m=1}^{k}\right). \tag{49}$$

Remark 6: It is a simple matter to verify that the $\boldsymbol{u}_j$ constructed along the lines of Corollary 11 satisfies

$$I\left(\boldsymbol{x}_j; \boldsymbol{u}_j \,\middle|\, \boldsymbol{u}_j^c\right) = R_{\text{tot}} - R_0. \tag{50}$$

But then, the total rate of the source code (as in Theorem 10) is simply given by

$$I(\boldsymbol{x}; \boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_L) = I\left(\boldsymbol{x}; \boldsymbol{u}_j^c\right) + I\left(\boldsymbol{x}_j; \boldsymbol{u}_j \,\middle|\, \boldsymbol{u}_j^c\right) \tag{51}$$

$$= R_{\text{tot}}, \tag{52}$$

as desired. Here, the first equality follows by the chain rule of mutual information [6, Theorem 2.5.2] and the fact that $\boldsymbol{u}_j$ is conditionally independent of $\boldsymbol{x}_\ell$, for $\ell \neq j$, when conditioned on $\boldsymbol{x}_j$.

To obtain an iterative procedure based on Corollary 11, assume that *initially*, the auxiliary random vectors $\boldsymbol{u}_2, \boldsymbol{u}_3, \ldots, \boldsymbol{u}_L$, and $\boldsymbol{x}$ are jointly Gaussian random vectors. Hence, they can be parameterized by transforms $C_2, C_3, \ldots, C_L$ and additive Gaussian "compression" noises, as illustrated in Fig. 3.

Algorithm 2 (Distributed KLT for Compression): The first terminal observes the first M_1 components, the second terminal observes the next M_2 components, and so on. A total rate of R bits per source vector is available, to be shared by the L terminals. Initialize as follows:

a) Let $C_\ell^{(0)} = \mathbf{0}_\ell$, where $\mathbf{0}_\ell$ denotes a column vector of length M_ℓ with all zero entries, for $\ell = 1, 2, \ldots L$ and $\Sigma_{z,\ell}^{(0)} = 0$.

b) Select a suitable nonnegative, nondecreasing, discrete-indexed function (a "rate schedule") $R(i)$ satisfying $R(0) = 0, R(i) \leq R$, for all $i > 0$.

In the ith iteration $(i > 0)$

1) For each $j, j = 1, 2, \ldots, L$:
Fix the transform matrices $C_\ell^{(i-1)}$ and the covariance matrices $\Sigma_{z,\ell}^{(i-1)}$, for $\ell = 1, 2, \ldots, L, \ell \neq j$, set $R_{\text{tot}} = R(i)$ and find the best transform $\tilde{C}_j$ and the best covariance matrix $\tilde{\Sigma}_{z,j}$, as described in Corollary 11. Denote the resulting overall distortion by D_j.

2) Only retain the one updated matrix that enables the largest decrease in distortion. More formally, set $j^* = \arg\min_j D_j$ and update as follows:

$$\text{for } \ell = j^*: \qquad C_{j^*}^{(i)} = \tilde{C}_{j^*} \text{ and } \Sigma_{z,j^*}^{(i)} = \tilde{\Sigma}_{z,j^*} \tag{53}$$

$$\text{for } \ell \neq j^*: \qquad C_\ell^{(i)} = C_\ell^{(i-1)} \text{ and } \Sigma_{z,\ell}^{(i)} = \Sigma_{z,\ell}^{(i-1)} \tag{54}$$

□

Remark 7: This algorithm attempts to *simultaneously* find a rate allocation between the terminals *and* the corresponding source code such as to minimize the overall distortion. It is important to point out that this algorithm may yield entirely different solutions, both in terms of the rate allocation and in terms of the corresponding source code, depending on the "rate schedule" function $R(i)$ that is selected.

Remark 8: A greedy rate allocation strategy as used in Algorithm 2 is inherently "short-sighted" and may not lead to a globally optimal solution in general [46, Sec. 8.4, p.234].

By analogy to Algorithm 1, it is easy to show that this algorithm must converge to a stable point that can either be a saddle point or a local minimum. Clearly, one cannot generally expect the algorithm to converge to a global minimum. Instead, we again illustrate the behavior with a few examples.

Theorem 12 (Local Convergence): Denote the transform and covariance matrices provided by Algorithm 2 after iteration i by $C_\ell^{(i)} \in \mathbb{R}^{k_\ell \times M_i}$ and $\Sigma_{z,\ell}^{(i)}$, respectively, for $\ell = 1, 2, \ldots, L$. Let $\boldsymbol{z}_\ell^{(i)}$ be a Gaussian random vector, independent of $\boldsymbol{x}$, with mean zero and covariance matrix $\Sigma_{z,\ell}^{(i)}$. Denote the minimum mean-squared error estimate of $\boldsymbol{x}$ based on the observations $\{C_\ell^{(i)} \boldsymbol{x}_\ell + \boldsymbol{z}_\ell^{(i)}\}_{\ell=1}^{L}$ by $\hat{\boldsymbol{x}}^{(i)}$. Then

$$E\left[\left\|\boldsymbol{x} - \hat{\boldsymbol{x}}^{(i)}\right\|^2\right] \geq E\left[\left\|\boldsymbol{x} - \hat{\boldsymbol{x}}^{(i+1)}\right\|^2\right] \tag{55}$$

i.e., the distortion is a nonincreasing function of the iteration number.

Proof: The theorem follows directly from Corollary 11: Suppose that in iteration i, the transform matrix C_j, and the corresponding covariance matrix $\Sigma_{z,j}$ provide the largest decrease in mean-squared error, and thus, are being updated. That is, they are selected according to Corollary 11. Note that the only restriction imposed by the corollary is (45). However, since $R(i) \geq R(i-1)$, it is clear that the current values of C_j and $\Sigma_{z,j}$ are *inside* the optimization space. Therefore, the distortion cannot increase. □

Example 8: To illustrate Algorithm 2, reconsider the scenario of Example 6. Rather than providing a 10-dimensional linear approximation each, the two terminals now provide descriptions at a total rate of R bits. The outcomes of our numerical investigation are shown in Fig. 12. The solid line is the performance of the scheme following from the distributed KLT algorithm. That is, for each point on the solid line, the distributed KLT algorithm was run, using a random rate schedule. More specifically, the

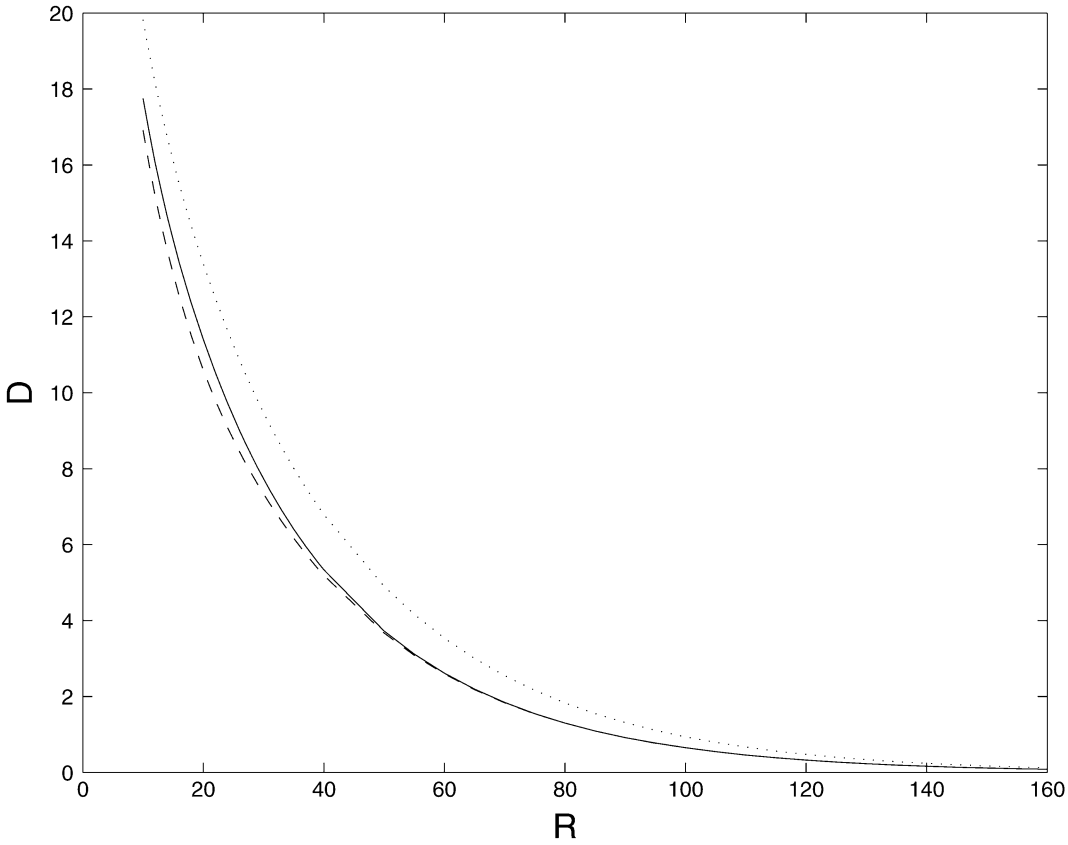

Fig. 12. Rate–distortion functions for Example 8: The solid line is the rate–distortion function for the distributed KLT algorithm (Algorithm 2), the dashed line the rate–distortion function for joint encoding of the entire vector, and the dotted line is the rate–distortion function for the case where each terminal individually compresses its observations, ignoring the other terminal's presence. The rate is given in bits.

function $R(i)$ in Algorithm 2 was made up of 100 random increments from $R(i=0)=0$ to $R(i=100)=R$. We also verified that for the example at hand, multiple runs with different random rate schedules led to the same rate–distortion points. For comparison, the dashed line is the rate–distortion function for *joint* encoding of the entire vector $\boldsymbol{x}$, and the dotted line is the performance of a scheme where no distributed coding is done, but instead each terminal individually (and optimally) compresses its observations, ignoring each other's presence.

VI. Illustration: Convergence of the Distributed KLT Algorithm

While the distributed KLT algorithm presented in Section V is an intuitively pleasing approach and can be shown to converge to a locally stable point which is either a local optimum or a saddle point (Theorem 9), convergence to a global optimum cannot be guaranteed in general. In this section, we illustrate this fact by two simple examples. In the first example, there is only one local minimum, which for that reason must also be a global minimum. Therefore, the proposed algorithm will converge to the globally optimum solution. In the second example, different local minima exist, and hence, the outcome of the proposed algorithm depends on the parameter settings and initializations.

Example 9 (Gauss–Markov Source): Consider the case $N=4$, $M_1=M_2=2$, and $k_1=k_2=1$. Let

$$\Sigma_x=\begin{pmatrix}1 & \rho & \rho^2 & \rho^3\\ \rho & 1 & \rho & \rho^2\\ \rho^2 & \rho & 1 & \rho\\ \rho^3 & \rho^2 & \rho & 1\end{pmatrix}. \tag{56}$$

In this simple example, the transform applied by terminal 1 is simply a vector of length 2. Since the performance is invariant under scaling, the transform can be parameterized by a single real number, characterizing the relationship between the two components. By analogy, the same holds for the transform applied by terminal 2. For example, we may parameterize the transforms by two angles, α and β, by which we mean that terminal 1 provides $y_1=x_1\cos\alpha+x_2\sin\alpha$, and terminal 2 provides $y_3=x_3\cos\beta+x_4\sin\beta$. For $\rho=0.9$, Fig. 13 shows the resulting error surface, as a function of α and β, revealing the obvious π-periodic structure and the fact that there is only one local minimum in each period. Consequently, any locally converging algorithm, and in particular, the suggested distributed KLT algorithm (Algorithm 1) will converge to the globally optimal solution. For the considered example, the distortion incurred by the distributed KLT is $D_{d\,\mathrm{klt}}\approx 0.1693$, while applying marginal KLTs at each terminal results in a distortion of

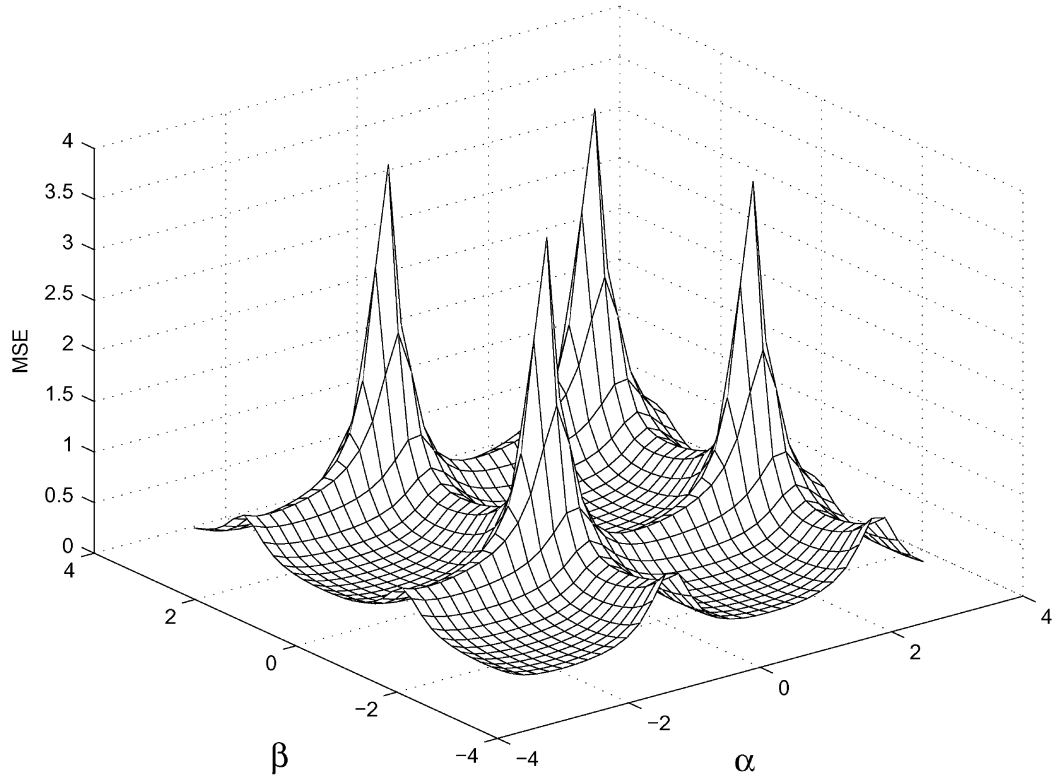

Fig. 13. The resulting distortion as a function of the distributed transforms (parameterized by α and β) for Example 9. All local minima are of the same depth, and hence, global minima.

$D_{m\,\mathrm{klt}} \approx 0.1714$. The joint KLT, having access to the entire vector simultaneously, incurs $D_{j\,\mathrm{klt}} \approx 0.1637$.

By contrast to the preceding example, the following example shows that a local optimum need not be a global optimum in the distributed KLT problem.

Example 10: Consider again the case $N = 4, M_1 = M_2 = 2$, and $k_1 = k_2 = 1$, but now, let

$$\Sigma_x = \begin{pmatrix} 2 & 0 & 1 & 0 \\ 0 & 2.3 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2.1 \end{pmatrix}. \tag{57}$$

By analogy to Example 9, we again plot the error surface as a function of α and β. The outcome is shown in Fig. 14, clearly revealing the existence of two different local minima. Obviously, *no* locally convergent algorithm can be guaranteed to find the global minimum (i.e., the smaller of the two local minima).

To get further insight, suppose that Algorithm 1 is initialized with $C_2 = (1\ 0)$, i.e., terminal 2 provides the component x_3. It is easy to see that Algorithm 1 will then select $C_1 = (0\ 1)$, i.e., terminal 1 provides the component x_2. Thereafter, Algorithm 1 does not change the matrices C_1 and C_2 any further. The resulting distortion is $D = 3.6$. However, if Algorithm 1 is initialized with $C_2 = (0\ 1)$, then the (locally) optimal choice is $C_1 = (1\ 0)$. Thereafter, Algorithm 1 does not change the matrices C_1 and C_2 any further. The resulting distortion is $D = 3.8$, illustrating that the algorithm converged to a local minimum that is not a global minimum.

VII. Conclusion

This paper derives a distributed version of the KLT: when the correlated data cannot be observed centrally, a scenario arising for example in sensor networks, it is impossible to apply the KLT to the entire data vector. Instead, we suppose that L independent agents each observe a separate part of the data, and have to locally process their part, providing a compressed version upstream. We consider two kinds of compressed versions: on the one hand, we consider linear approximation, where each agent provides a small-dimensional approximation of its observation, and on the other hand, we consider a rate–distortion framework where each agent provides a bit stream. Somewhere upstream sits a central data collector, wishing to reconstruct the *entire* underlying data vector at the smallest mean-squared error possible. The problem studied in this paper is to determine the optimum local operations to be executed by the independent agents.

Special cases are addressed for which explicit solutions can be given, including the *partial* and the *conditional* KLT. For the general case, the paper derives a locally convergent algorithm. For the Gauss–Markov example (Example 9), the algorithm converges to a global optimum. Generally, however, we show that the distributed KLT problem typically results in a nonconvex optimization problem, and hence, further investigations are necessary to determine the precise (global) convergence behavior.

As far as applications are concerned, many scenarios of current interest involve distributed compression. For example, data gathering in sensor networks involves distributed coding of correlated data, where the distributed KLT can play a role. Another

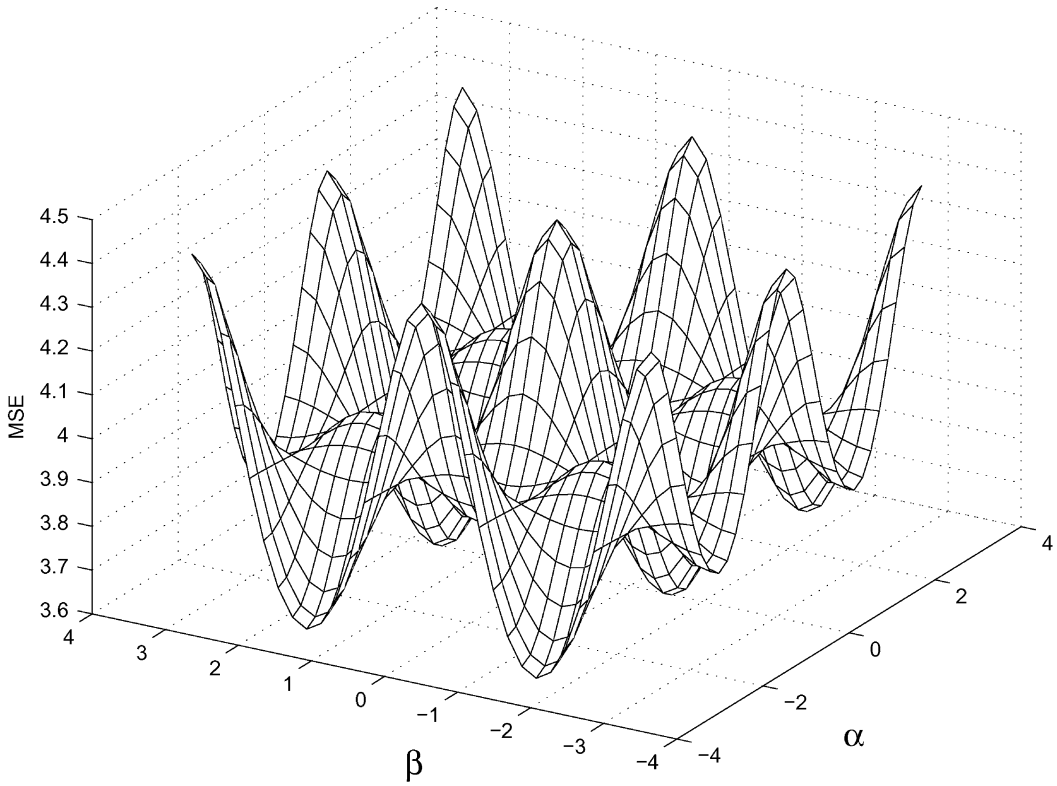

Fig. 14. The resulting distortion as a function of the distributed transforms (parameterized by α and β) for Example 10. As the figure reveals, there exist local minima that are not global minima. More precisely, there are two kinds of (periodically repeated) local minima. In the graph, the global minima can be seen most easily along the β-axis, and the local minima that are not global along the α-axis.

example is found in interactive communication schemes, e.g., two databases exchanging images with known cross correlation, which leads directly to a conditional KLT problem.

Appendix I

To establish Lemma 1 and Theorem 2, we start by noting that since $\boldsymbol{x}$ is assumed to be a vector of jointly Gaussian random variables,[3] and since $\boldsymbol{y}_2 = C_2\boldsymbol{x}_2 + \boldsymbol{z}_2$, where $\boldsymbol{z}_2$ is also Gaussian and independent of $\boldsymbol{x}$, there exist matrices A_1 and A_2 such that

$$\boldsymbol{x}_2 = A_1\boldsymbol{x}_1 + A_2\boldsymbol{y}_2 + \boldsymbol{v} \tag{58}$$

where the Gaussian random vector $\boldsymbol{v}$ is *independent* of $\boldsymbol{x}_1$ and of $\boldsymbol{y}_2$. It is straightforward to verify that the matrix $A = (A_1, A_2)$ is the one given in (14). Simultaneously, there exists a matrix B_2 such that we can write

$$\begin{pmatrix} I_M \\ A_1 \end{pmatrix} \boldsymbol{x}_1 = B_2\boldsymbol{y}_2 + \boldsymbol{w}, \tag{59}$$

where the random vector $\boldsymbol{w}$ is *independent* of $\boldsymbol{y}_2$. The matrix Σ_w given in (15) is the covariance matrix of the random vector $\boldsymbol{w}$. We will now use these relationships to simplify the distortion expression to be minimized.

Proof of Lemma 1: For Lemma 1, the fact that $\mathrm{rank}(C_1) = M$ follows from the fact that the columns of $Q_w^{(M)}$ are orthogonal and that the matrix $\binom{I_M}{A_1}$ has full (column) rank, trivially. Clearly, however, C_1 is not generally unitary. The second fact can be established by considering the random matrix

$$\begin{aligned}
&\mathrm{Cov}(C_1\boldsymbol{x}_1, C_1\boldsymbol{x}_1 \,|\, \boldsymbol{y}_2) \\
&= \mathrm{Cov}\left(\left(Q_w^{(M)}\right)^T \begin{pmatrix} I_M \\ A_1 \end{pmatrix} \boldsymbol{x}_1,\ \left(Q_w^{(M)}\right)^T \begin{pmatrix} I_M \\ A_1 \end{pmatrix} \boldsymbol{x}_1 \middle| \boldsymbol{y}_2 \right) \\
&= \mathrm{Cov}\left(\left(Q_w^{(M)}\right)^T (B_2\boldsymbol{y}_2 + \boldsymbol{w}),\ \left(Q_w^{(M)}\right)^T (B_2\boldsymbol{y}_2 + \boldsymbol{w}) \middle| \boldsymbol{y}_2 \right) \\
&= \mathrm{Cov}\left(\left(Q_w^{(M)}\right)^T \boldsymbol{w}, \left(Q_w^{(M)}\right)^T \boldsymbol{w} \middle| \boldsymbol{y}_2 \right)
\end{aligned} \tag{60}$$

and since $\boldsymbol{w}$ is independent of $\boldsymbol{y}_2$, we find

$$\mathrm{Cov}(C_1\boldsymbol{x}_1, C_1\boldsymbol{x}_1 | \boldsymbol{y}_2) = \left(Q_w^{(M)}\right)^T \Sigma_w Q_w^{(M)} \tag{61}$$

with probability one. Moreover, the matrix on the left-hand side (LHS) of (61) is diagonal by construction of $Q_w^{(M)}$, with diagonal entries $\lambda_{w,m}$, for $m = 1, 2, \ldots, M$, as in (16). Hence, the components of $C_1\boldsymbol{x}_1$ have conditional variances $\lambda_{w,m}$, for $m = 1, 2, \ldots, M$, conditioned on $\boldsymbol{y}_2$. □

Proof of Theorem 2: To establish Theorem 2, we assume that the encoder provides a linear approximation, i.e., the encoder provides[4]

$$\boldsymbol{y}_1 = C_1\boldsymbol{x}_1. \tag{62}$$

[3]The theorem can be proved more generally if one *imposes* the restriction of linear reconstruction, but this is beyond the scope of the present paper.

[4]Note that the same analysis applies to the case where the encoder is to provide a *noisy* linear approximation of the form $\boldsymbol{y}_1 = C_1\boldsymbol{x}_1 + \boldsymbol{z}_1$, where $\boldsymbol{z}_1$ is additional noise, independent of all components of the original vector $\boldsymbol{x}$.

Hence, the decoder has access to $\boldsymbol{y}_1 = C_1\boldsymbol{x}_1$ (or, along the same lines, a noisy version $C_1\boldsymbol{x}_1 + \boldsymbol{z}_1$) and to $\boldsymbol{y}_2 = C_2\boldsymbol{x}_2 + \boldsymbol{z}_2$, and needs to provide an estimate $\hat{\boldsymbol{x}}$ such as to minimize

$$E[\|\boldsymbol{x} - \hat{\boldsymbol{x}}^2]. \tag{63}$$

In the Gaussian case considered in this paper, it is well known that linear reconstruction is optimal (a straightforward consequence of, e.g., [47, Theorem 34.8]), that is, $\hat{\boldsymbol{x}}$ is an (arbitrary) linear[5] function of $\boldsymbol{y}_1 = C_1\boldsymbol{x}_1$ and of $\boldsymbol{y}_2 = C_2\boldsymbol{x}_2 + \boldsymbol{z}_2$. The matrix C_2 is fixed, and the goal is to determine the optimal matrix C_1. To this end, let us first trivially rewrite

$$E[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2] = E[\|\boldsymbol{x}_1 - \hat{\boldsymbol{x}}_1\|^2] + E[\|\boldsymbol{x}_2 - \hat{\boldsymbol{x}}_2\|^2] \tag{64}$$

where we use $\hat{\boldsymbol{x}}_1$ to denote the first M components of $\hat{\boldsymbol{x}}$, and $\hat{\boldsymbol{x}}_2$ to denote its remaining $N - M$ components. Furthermore

$$\begin{aligned} &E[\|\boldsymbol{x}_1 - \hat{\boldsymbol{x}}_1\|^2] + E[\|\boldsymbol{x}_2 - \hat{\boldsymbol{x}}_2\|^2] \\ &\quad = E[\|\boldsymbol{x}_1 - \hat{\boldsymbol{x}}_1\|^2] + E[\|A_1\boldsymbol{x}_1 + A_2\boldsymbol{y}_2 + \boldsymbol{v} - \hat{\boldsymbol{x}}_2\|^2] \end{aligned} \tag{65}$$

where we have used (58). Recall that $\boldsymbol{v}$ is independent of $\boldsymbol{x}_1$ and $\boldsymbol{y}_2$ (by construction of $\boldsymbol{v}$, see (58)). Therefore,

$$\hat{\boldsymbol{x}}_2 = A_1\hat{\boldsymbol{x}}_1 + A_2\boldsymbol{y}_2 \tag{66}$$

and thus,

$$\begin{aligned} &E[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2] \\ &\quad = E[\|\boldsymbol{x}_1 - \hat{\boldsymbol{x}}_1\|^2] + E[\|A_1\boldsymbol{x}_1 - A_1\hat{\boldsymbol{x}}_1\|^2] + E[\|\boldsymbol{v}\|^2] \end{aligned} \tag{67}$$

Finally, merging the two contributions that involve $\boldsymbol{x}_1$, we obtain

$$E[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2] = E\left[\left\|\binom{I_M}{A_1}\boldsymbol{x}_1 - \binom{I_M}{A_1}\hat{\boldsymbol{x}}_1\right\|^2\right] + E[\|\boldsymbol{v}\|^2]. \tag{68}$$

To proceed from here, we will rewrite the distortion as an iterated expectation, as follows:

$$\begin{aligned} E[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2] = E\left[E\left[\left\|\binom{I_M}{A_1}\boldsymbol{x}_1 - \binom{I_M}{A_1}\hat{\boldsymbol{x}}_1\right\|^2 \middle| \boldsymbol{y}_2\right]\right] \\ + E[\|\boldsymbol{v}\|^2]. \end{aligned} \tag{69}$$

Now, consider the random variable $E[\|\binom{I_M}{A_1}\boldsymbol{x}_1 - \binom{I_M}{A_1}\hat{\boldsymbol{x}}_1\|^2 \,|\, \boldsymbol{y}_2]$. For a specific instance $\boldsymbol{y}_2 = \boldsymbol{\xi}_2$, we can consider

$$E\left[\left\|\binom{I_M}{A_1}\boldsymbol{x}_1 - \binom{I_M}{A_1}\hat{\boldsymbol{x}}_1\right\|^2 \middle| \boldsymbol{y}_2 = \boldsymbol{\xi}_2\right]. \tag{70}$$

Next, since the columns of the matrix $Q_w^{(M)}$ are orthonormal, we can multiply both expressions inside the expectation in (70) by $(Q_w^{(M)})^T$ without changing the outcome. But then, using our definition of C_1, we can rewrite

$$\begin{aligned} E\left[\left\|\binom{I_M}{A_1}\boldsymbol{x}_1 - \binom{I_M}{A_1}\hat{\boldsymbol{x}}_1\right\|^2 \middle| \boldsymbol{y}_2 = \boldsymbol{\xi}_2\right] \\ = E[\|C_1\boldsymbol{x}_1 - C_1\hat{\boldsymbol{x}}_1\|^2 \,|\, \boldsymbol{y}_2 = \boldsymbol{\xi}_2] \end{aligned} \tag{71}$$

which we can now rewrite in terms of the components of the vector $\boldsymbol{y}_1 = C_1\boldsymbol{x}_1 = (y_1, y_2, \ldots, y_M)^T$ as

$$\begin{aligned} E\left[\left\|\binom{I_M}{A_1}\boldsymbol{x}_1 - \binom{I_M}{A_1}\hat{\boldsymbol{x}}_1\right\|^2 \middle| \boldsymbol{y}_2 = \boldsymbol{\xi}_2\right] \\ = \sum_{m=1}^{M} E[|y_m - \hat{y}_m|^2 \,|\, \boldsymbol{y}_2 = \boldsymbol{\xi}_2]. \end{aligned} \tag{72}$$

Lemma 1 ensures that the components $y_1, y_2, \ldots, y_m$ are conditionally uncorrelated given $\boldsymbol{y}_2 = \boldsymbol{\xi}_2$. Therefore, in terms of the components $y_1, y_2, \ldots, y_m$, it is a simple matter to determine the optimum k-dimensional approximation space. First, if the component y_m is retained, then clearly its corresponding estimate is $\hat{y}_m = y_m$. However, if y_m is *not* retained, then its corresponding estimate is $\hat{y}_m = 0$; none of the other components of the vector $\boldsymbol{y}$ contain anything relevant about y_m, conditioned on $\boldsymbol{y}_2$. The best k-dimensional approximation space is therefore easily found in terms of $\boldsymbol{y}_1$: Denote the set of the k indices corresponding to the retained components of $\boldsymbol{y}$ by $\mathcal{T}$. Then, the incurred distortion can be expressed as

$$E\left[\left\|\binom{I_M}{A_1}\boldsymbol{x}_1 - \binom{I_M}{A_1}\hat{\boldsymbol{x}}_1\right\|^2 \middle| \boldsymbol{y}_2 = \boldsymbol{\xi}_2\right] = \sum_{m\in\mathcal{T}^c} \lambda_{w,m} \tag{73}$$

where $\lambda_{w,m}$, for $m = 1, 2, \ldots, M$, denote the conditional variances of the components of $\boldsymbol{y}_1 = C_1\boldsymbol{x}_1$, conditioned on $\boldsymbol{y}_2 = \boldsymbol{\xi}_2$, as given in (16), and $\mathcal{T}^c$ denotes the complement of $\mathcal{T}$ in the set $\{1, 2, \ldots, M\}$. Therefore, the best k-dimensional approximation is given by the components y_m corresponding to the largest eigenvalues $\lambda_{w,m}$. It is important to note that the matrix C_1, and hence the resulting approximation space, *does not depend* on the particular realization of $\boldsymbol{y}_2$. Therefore, we can trivially evaluate the iterated expectation in (69) to obtain

$$E[\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|^2] = E\left[E\left[\left\|\binom{I_M}{A_1}\boldsymbol{x}_1 - \binom{I_M}{A_1}\hat{\boldsymbol{x}}_1\right\|^2 \middle| \boldsymbol{y}_2\right]\right] + E[\|\boldsymbol{v}\|^2] \tag{74}$$

$$= \sum_{m\in\mathcal{T}^c} \lambda_{w,m} + E[\|\boldsymbol{v}\|^2]. \tag{75}$$

Finally, to evaluate $E[\|\boldsymbol{v}\|^2]$, we need the covariance matrix Σ_v of the random vector $\boldsymbol{v}$ in (58). From standard covariance considerations, we find that Σ_v must satisfy

$$\begin{aligned} \Sigma_v = \Sigma_2 - \begin{pmatrix}\Sigma_{12}^T & \Sigma_2 C_2^T\end{pmatrix} \begin{pmatrix}\Sigma_1 & \Sigma_{12}C_2^T \\ C_2\Sigma_{12}^T & C_2\Sigma_2 C_2^T + \Sigma_z\end{pmatrix}^{-1} \\ \times \begin{pmatrix}\Sigma_{12} \\ C_2\Sigma_2\end{pmatrix}, \end{aligned} \tag{76}$$

and by definition, $E[\|\boldsymbol{v}\|^2] = \operatorname{trace}\Sigma_v$. □

APPENDIX II

Proof of Lemma 3: We establish the theorem in two parts, first showing that no lower rate can be hoped for, and then showing that there exists a coding technique that achieves this rate.

Proofs of similar statements can be found in the literature, see, e.g., [48]–[50].

[5]Recall, however, that beyond the case of Gaussian statistics, nonlinear estimation may perform better. We do not address this issue in the present paper.

Converse: To establish the converse, suppose that the value of the side information $\boldsymbol{y}_2$ is known *both* at the encoder and at the decoder. In this idealized scenario, denote the minimum rate required to achieve a distortion no larger than D by $R_{\rm both}(D)$. Clearly

$$R_{\rm local}(D) \geq R_{\rm both}(D) \tag{77}$$

since any code that works for the scenario of Theorem 3 also works in the idealized scenario.

The goal of the first part of the proof is to provide a lower bound to $R_{\rm both}(D)$, and hence, to $R_{\rm local}(D)$. The idealized scenario where both the encoder and the decoder know the side information $\boldsymbol{y}_2$ is sometimes referred to as the *conditional rate–distortion* function, and has been considered in [51]. Our scenario is slightly different from [51] in that the distortion criterion involves the side information. Let us define the following object:

$$R_c(D) = \min I(\boldsymbol{x}_1; \hat{\boldsymbol{x}} \,|\, \boldsymbol{y}_2) \tag{78}$$

where the minimum is over all $p(\hat{\boldsymbol{x}} \,|\, \boldsymbol{x}_1, \boldsymbol{y}_2)$ satisfying

$$E[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2] \leq D. \tag{79}$$

It follows straightforwardly from [6, Lemma 13.4.1] that $R_c(D)$ is a convex and nonincreasing function of D.

Consider any (block) code of length n for the observed source vector sequences $\{\boldsymbol{x}_1[i]\}_{i=1}^n$. If the rate per source (vector) sample is $R_{\rm both}(D)$, there are at most $2^{nR_{\rm both}(D)}$ different codewords. Along the lines of [6, eqs. (13.58)–(13.70)], it is easy to show that $R_{\rm both}(D) \geq R_c(D)$.

The next step is to evaluate $R_c(D)$ for the case at hand. First, we rewrite

$$\begin{pmatrix} I_M \\ A_1 \end{pmatrix} \boldsymbol{x}_1 = B_2 \boldsymbol{y}_2 + \boldsymbol{w} \tag{80}$$

which implies $I(\boldsymbol{x}_1; \hat{\boldsymbol{x}} \,|\, \boldsymbol{y}_2) = I(\boldsymbol{w}; \hat{\boldsymbol{x}} \,|\, \boldsymbol{y}_2)$. Similarly, for the distortion, we can write

$$\begin{aligned} E[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2] &= E\left[\left\|\binom{I_M}{A_1}\boldsymbol{x}_1 - \hat{\boldsymbol{x}}\right\|^2\right] \\ &= E[\|\boldsymbol{w} - \hat{\boldsymbol{x}}^2] + E[\|\boldsymbol{v}\|^2]. \end{aligned} \tag{81}$$

Next, consider $\tilde{\boldsymbol{w}} = Q_w^T \boldsymbol{w}$, where Q_w is the unitary matrix satisfying

$$Q_w^T \Sigma_w Q_w = \mathrm{diag}(\lambda_{w,1}, \ldots, \lambda_{w,M}, \underbrace{0, 0, 0, \ldots, 0}_{N-M \text{ zeros}}). \tag{82}$$

In terms of $\tilde{w}$, we can rewrite the minimization problem (78)–(79) as

$$\min I(\tilde{\boldsymbol{w}}; \hat{\boldsymbol{x}} \,|\, \boldsymbol{y}_2) \tag{83}$$

where $\tilde{\boldsymbol{w}}$ and $\boldsymbol{y}_2$ are independent Gaussian random vectors, and the minimization is over all $p(\hat{\boldsymbol{x}} \,|\, \boldsymbol{x}_1, \boldsymbol{y}_2)$ that satisfy

$$E[\|\tilde{\boldsymbol{w}} - \hat{\boldsymbol{x}}\|^2] \leq D. \tag{84}$$

Since $\tilde{\boldsymbol{w}}$ and $\boldsymbol{y}_2$ are independent, the minimizing $\hat{\boldsymbol{x}}$ will depend only on $\tilde{\boldsymbol{w}}$. In other words, the problem can be expressed as

$$\min I(\tilde{\boldsymbol{w}}; \hat{\boldsymbol{x}}) \tag{85}$$

where the minimization is over all $p(\hat{\boldsymbol{x}} \,|\, \tilde{\boldsymbol{w}})$ that satisfy

$$E[\|\tilde{\boldsymbol{w}} - \hat{\boldsymbol{x}}\|^2]. \tag{86}$$

But since the components of $\tilde{\boldsymbol{w}}$ are independent Gaussian random variables with mean zero and variances

$$\lambda_{w,1}, \ldots, \lambda_{w,M}, \underbrace{0, 0, 0, \ldots, 0}_{N-M \text{ zeros}}, \tag{87}$$

this is merely the problem of compressing independent Gaussian sources of different variances, whose solution is well known to be [6, p. 348]

$$R_c(D) = \min_{D_1, \ldots, D_M} \sum_{m=1}^{M} \max\left\{\frac{1}{2}\log_2 \frac{\lambda_{w,m}}{D_m}, 0\right\} \tag{88}$$

where $\lambda_{w,m}$ are the eigenvalues of the matrix Σ_w, and where the minimum is over all $D_1, \ldots, D_M$ satisfying $\sum_{m=1}^{M} D_m + E[\|\boldsymbol{v}\|^2] \leq D$.

Finally, evaluating the minimization over the $D_1, \ldots, D_M$, the following expression is obtained:

$$R_c(D) = \sum_{m=1}^{M} \max\left\{\frac{1}{2}\log_2 \frac{\lambda_{w,m}}{D_m}, 0\right\} \tag{89}$$

where

$$D_m = \begin{cases} \theta, & \text{if } \theta < \lambda_{w,m} \\ \lambda_{w,m}, & \text{if } \theta \geq \lambda_{w,m} \end{cases} \tag{90}$$

where θ is chosen such that $\sum_{m=1}^{M} D_m + E[\|\boldsymbol{v}\|^2] = D$.

Hence, we have shown that

$$R_{\rm local}(D) \geq R_c(D) \tag{91}$$

with $R_c(D)$ as in (89)–(90).

Achievability: The remaining part of the proof is to show that there exists a coding scheme for the scenario of Theorem 3 whose performance arbitrarily closely approaches $R_c(D)$. To do so, let the encoding device first apply the local KLT C_1 to $\boldsymbol{x}_1[i]$ (with respect to $\boldsymbol{y}_2[i]$), as in Definition 1. Note that since the source is memoryless, the transform C_1 does not depend on i. This yields a (transformed) sequence of vectors $\boldsymbol{y}_1[i] = C_1 \boldsymbol{x}_1[i]$. For each i, the components of the vector $\boldsymbol{y}_1[i]$ are conditionally independent given $\boldsymbol{y}_2[i]$. It now suffices to apply the result of Wyner and Ziv [9], [52] *separately* to each component of $\boldsymbol{y}_1[i]$. More precisely, for the "component sequence" $\{y_m[i]\}_{i=1}^{\infty}, m = 1, 2, \ldots, M$, with side information at the decoder $\{\boldsymbol{y}_2[i]\}_{i=1}^{\infty}$, it was found in [9], [52] that

$$R_m(D_m) = \max\left\{\frac{1}{2}\log_2 \frac{\lambda_{w,m}}{D_m}, 0\right\}. \tag{92}$$

Finally, we minimize the sum rate $\sum_{m=1}^{M} R_m$ subject to the constraint that $\sum_{m=1}^{M} D_m + E[\|\boldsymbol{v}\|^2] = D$. The solution is easily found to be (88). □

Proof of Corollary 4: Consider the proof of achievability for Theorem 3. After applying the local KLT to $\boldsymbol{x}_1$, yielding $\boldsymbol{y}_1 = C_1\boldsymbol{x}_1$, the components $y_1, y_2, \ldots, y_M$ of $\boldsymbol{y}_1$ are treated separately, incurring distortions of $D_1, D_2, \ldots, D_M$, respectively. That is, consider the single-source Wyner–Ziv problem (as in [9], [52]), the source being y_m, and the side information being the vector $\boldsymbol{y}_2$. The minimum rate can be characterized as

$$R_m(D_m) = \min I(y_m; u_m \,|\, \boldsymbol{y}_2) \tag{93}$$

where the minimization is over all distributions $p(u_m \,|\, y_m)p(y_m, \boldsymbol{y}_2)$ for which there exists a function $f_m(\cdot)$ such that

$$E[|y_m - f_m(u_m, \boldsymbol{y}_2)|^2] \le D_m. \tag{94}$$

Evaluating this reveals (92), but it also reveals that the minimizing $p(u_m \,|\, y_m)$ makes u_m and y_m jointly Gaussian random variables. Since u_m is only determined up to a scaling factor, this simply implies that we can express u_m as

$$u_m = y_m + z_m \tag{95}$$

where z_m is Gaussian and independent of y_m. It is easily verified that z_m has mean zero and variance σ_Z^2 satisfying

$$D_m = \frac{\lambda_{w,m}\sigma_Z^2}{\lambda_{w,m} + \sigma_Z^2}. \tag{96}$$

Finally, just like in the proof of achievability for Theorem 3, D_m must be chosen to minimize the resulting sum rate $\sum_{m=1}^{M} R_m(D_m)$, subject to the constraint

$$\sum_{m=1}^{M} D_m + E[\|\boldsymbol{v}\|^2] = D.$$

This minimization is carried out by analogy to (89)–(90), and (90) implies the rule for selecting k in Corollary 4. □

ACKNOWLEDGMENT

The authors are grateful to Olivier Roy for detailed comments on the manuscript, and to the three anonymous reviewers whose insightful comments helped to significantly improve this paper.

REFERENCES

[1] M. Gastpar, P. L. Dragotti, and M. Vetterli, "The distributed Karhunen-Loève transform," in *Proc. 2002 Int. Workshop on Multimedia Signal Processing*, St. Thomas, U.S. Virgin Islands, Dec. 2002, pp. 57–60.

[2] ——, "The distributed, partial, and conditional Karhunen-Loève transforms," in *Proc. IEEE 2003 Data Compression Conf. (DCC)*, Snowbird, UT, Mar. 2003, pp. 283–292.

[3] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, pp. 417–441, 498–520, 1933.

[4] K. Karhunen, "Über lineare Methoden in der Wahrscheinlichkeitsrechnung," *Ann. Acad. Sci. Fenn., Ser. A.1.: Math.-Phys.*, vol. 37, pp. 3–79, 1947.

[5] M. Loève, "Fonctions aléatoires du second ordre," in *Processus stochastiques et mouvements Browniens*, P. Lévy, Ed. Paris, France: Gauthier-Villars, 1948.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[7] J.-Y. Huang and P. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Commun. Syst.*, vol. COM-11, no. 9, pp. 289–296, Sep. 1963.

[8] V. K. Goyal, "Theoretical foundations of transform coding," *IEEE Signal Process.Mag.*, vol. 18, pp. 9–21, Sep. 2001.

[9] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 1–11, Jan. 1976.

[10] T. Berger, "Multiterminal source coding," *Lectures Presented at CISM Summer School on the Information Theory Approach to Communications*, Jul. 1977.

[11] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 4, pp. 471–480, Jul. 1973.

[12] S.-Y. Tung, "Multiterminal Source Coding," Ph.D. dissertation, Cornell Univ., Ithaca, NY, 1978, See also the Abstract in *IEEE Trans. Info. Theory*, vol. IT-21, no. 6, p. 787, Nov. 1978.

[13] A. H. Kaspi and T. Berger, "Rate-distortion for correlated source with partially separated encoders," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 6, pp. 828–840, Nov. 1982.

[14] T. Berger and R. W. Yeung, "Multiterminal source encoding with one distortion criterion," *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 228–236, Mar. 1989.

[15] Y. Oohama, "Gaussian multiterminal source coding," *IEEE Trans. Inf. Theory*, vol. 43, no. 7, pp. 1912–1923, Nov. 1997.

[16] R. Zamir and T. Berger, "Multiterminal source coding with high resolution," *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 106–117, Jan. 1999.

[17] Y. Oohama, "The rate-distortion function for the quadratic Gaussian CEO problem," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1057–1070, May 1998.

[18] ——, "Rate-distortion theory for Gaussian multiterminal source coding systems with several side informations at the decoder," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2577–2593, Jul. 2005.

[19] M. Gastpar, "The Wyner–Ziv problem with multiple sources," *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2762–2767, Nov. 2004.

[20] A. Wagner, S. Tavildar, and P. Viswanath, "The rate region of the quadratic Gaussian two-terminal source-coding problem," [Online]. Available: http://arXiv:cs.IT/0510095 2005

[21] S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): Design and construction," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 1999, pp. 158–167.

[22] S. Pradhan and K. Ramchandran, "Distributed source coding: Symmetric rates and applications to sensor network," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 2000, pp. 363–372.

[23] S. S. Pradhan, J. Kusuma, and K. Ramchandran, "Distributed compression in a dense microsensor network," *IEEE Signal Process. Mag.*, vol. 19, pp. 51–60, Mar. 2002.

[24] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (discus): Design and construction," *IEEE Trans. Inf. Theory*, vol. 49, no. 3, pp. 626–643, Mar. 2003.

[25] J. Garcia-Frias and Y. Zhao, "Compression of correlated binary sources using turbo codes," *IEEE Commun. Lett.*, vol. 5, no. 10, pp. 417–419, Oct. 2001.

[26] A. Aaron and B. Girod, "Compression with side information using turbo codes," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 2002, pp. 252–261.

[27] A. Liveris, Z. Xiong, and C. Georghiades, "A distributed source coding technique for correlated images using turbo codes," *IEEE Commun. Lett.*, vol. 6, no. 10, pp. 440–442, Oct. 2002.

[28] Y. Yang, V. Stankovic, Z. Xiong, and W. Zhao, "On multiterminal source code design," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 2005, pp. 43–52.

[29] V. Stankovic, A. D. Liveris, Z. Xiong, and C. N. Georghiades, "On code design for the Slepian–Wolf problem and lossless multiterminal networks," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1495–1507, Apr. 2006.

[30] R. Puri and K. Ramchandran, "PRISM: A new robust video coding architecture based on distributed compression principles," in *Proc. 40th Annu. Allerton Conf. Communication, Control and Computing*, Monticello, IL, Oct. 2002, vol. 40, pp. 586–595, Part 1.

[31] R. Puri and K. Ramchandran, "PRISM: A 'reversed' multimedia coding paradigm," in *Proc. IEEE Int. Conf. Image Processing* , Barcelona, Spain, Sep. 2003, vol. 1, pp. 617–620.

[32] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proc. IEEE*, vol. 93, no. 1, pp. 71–83, Jan. 2005.

[33] D. Rebollo-Monedero, A. Aaron, and B. Girod, "Transforms for high-rate distributed source coding," in *Proc. 37th Asilomar Conf. Signals, Systems, and Computers*, Thousand Oaks, CA, Nov. 2003, pp. 850–854.

[34] D. Rebollo-Monedero, S. Rane, A. Aaron, and B. Girod, "High-rate quantization and transform coding with side information at the decoder," *Signal Process.*, vol. 86, no. 11, pp. 3160–3179, Nov 2006, Special Section: Distributed Source Coding.

[35] M. Flierl and P. Vandergheynst, "Distributed coding of highly correlated image sequences with motion-compensated temporal wavelets," *EURASIP J. Appl. Signal Process.*, vol. 2006, 2006, Article ID 46747, 10 pp.

[36] K. Zhang, "Best Linear Unbiased Estimation Fusion with Constraints," Ph.D. dissertation, Univ. New Orleans, New Orleans, LA, 2003.

[37] K. S. Zhang, X. R. Li, P. Zhang, and H. F. Li, "Optimal linear estimation fusion part VI: Sensor data compression," in *Proc. Int. Conf. Information Fusion*, Cairns, Queensland, Australia, Jul. 2003, vol. 1, pp. 221–228.

[38] H. Nurdin, R. R. Mazumdar, and A. Bagchi, "On estimation and compression of distributed correlated signals with incomplete observations," in *Proc. Conf. Mathematical Theory of Networks and Systems*, Leuven, Belgium, Jul. 2004.

[39] O. Roy and M. Vetterli, "On the asymptotic distortion behavior of the distributed Karhunen-Loève transform," in *Proc. 43rd Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Sep. 2005, pp. 406–415.

[40] R. M. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matrices," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 6, pp. 725–730, Nov. 1972.

[41] I. D. Schizas, G. B. Giannakis, and Z.-Q. Luo, "Distributed estimation using reduced-dimensionality sensor observations," *IEEE Trans. Signal Process.*, to be published.

[42] T. Kailath, A. Sayed, and B. Hassibi, *Linear Estimation*. Upper Saddle River, NJ: Prentice-Hall, 2000.

[43] T. Berger, *Rate Distortion Theory: A Mathematical Basis For Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[44] K. B. Housewright, "Source Coding Studies For Multiterminal Systems," Ph.D. dissertation, Univ. California, Los Angeles, CA, 1977.

[45] S. S. Pradhan, On the rate-distortion function of Gaussian sources with memory in the presence of side information at the decoder Univ. Illinois at Urbana-Champaign, Project Rep., ECE 480, 1998.

[46] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston: Kluwer Academic, 1992.

[47] P. Billingsley, *Probability and Measure*, 3rd ed. New York: Wiley, 1995.

[48] H. Yamamoto and K. Itoh, "Source coding theory for multiterminal communication systems with a remote source," *Trans. IECE Japan*, vol. E63, pp. 700–706, 1980.

[49] S. Draper, "Successive Structuring of Source Coding Algorithms for Data Fusion, Buffering, and Distribution in Networks," Ph.D. dissertation, MIT, Cambridge, MA, 2002.

[50] D. Rebollo-Monedero and B. Girod, "Generalization of the rate-distortion function of Wyner-Ziv coding of noisy sources in the quadratic-Gaussian case," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 2005, pp. 23–32.

[51] R. M. Gray, Conditional rate-distortion theory Stanford Univ., Stanford, CA, Tech. Rep. 6502-2, Oct. 1972.

[52] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder-II: General sources," *Inf. Contr.*, vol. 38, pp. 60–80, 1978.