Source-Optimized Clustering for Distributed Source Coding

Gerhard Maierbacher João Barros

Abstract— **Motivated by the design of low-complexity distributed quantizers and iterative decoding algorithms that leverage the correlation in the data picked up by a large-scale sensor network, we address the problem of finding correlation preserving clusters. To construct a factor graph describing the statistical dependencies between sensor measurements, we develop a hierarchical clustering algorithm that minimizes the Kullback Leibler Distance between known and approximated source statistics. Finally, we show how the clustering result can be exploited in the design of index assignments for distributed quantization and source-channel decoders of manageable complexity.**

Index Terms— **sensor networks, distributed source coding, clustering, trees, quantization, Kullback-Leibler distance**

I. Introduction

In distributed sensing scenarios, where correlated data has to be gathered by a large number of low-complexity, power-restricted sensors, efficient source coding and data gathering techniques are key towards reducing the required number of transmissions and enabling extended network life-time.

Inspired by the seminal work of Slepian and Wolf [17], characterizing the fundamental limits of separate encoding of correlated sources, several authors have contributed with coding solutions (see e.g. [18] and references therein). Focusing on scalar quantization, [4] proposes to eliminate the redundancy in the quantized values of a pair of sources by reusing sets of quantization indices that minimize the overall end-to-end distortion. The constructions in [14] and [15] use the syndromes of channel codes with appropriate distance properties. On the other hand, the contributions in [10] and [2] address the optimization of multiterminal quantizers for combined use with Slepian Wolf codes, e.g. based on LDPC or punctured turbo codes [18]. In [5], we presented a low-complexity solution for the special case of the symmetric Gaussian CEO Problem [6]. In addition to distributed source coding, clustering algorithms for correlated data, largely inspired by contributions in statistical data analysis [9], have proved to be useful in the context of sensor networks, particularly in the design of energy-efficient and decentralized data gathering protocols, see e.g. [7] and [12].

The main goal of the present paper is to ensure the *scalability* of distributed quantization, i.e. its applicability to scenarios with a very large number of encoders (100 or more), for which *source*-optimized clustering provides a natural solution. Previous work [8] along this line presented a solution for the scalability problem on the decoding side by running the sum-product algorithm on a carefully chosen factor graph approximation of the source correlation. Focusing on the encoding side, we provide a scalable solution for distributed quantization based on source-optimized hierarchical clustering. Our main idea is to reduce the number of quantization bits in a systematic way by exploiting correlation preserving clusters that minimize the Kullback-Leibler Distance (KLD) between the given source statistics and a factor graph approximation. Simulation results underline the efficiency and scalability of the proposed approach, which can be applied to a variety of index assignments for low complexity distributed quantization.

The rest of the paper is organized as follows. In Section II we give a precise formulation of the problem setup. The clustering procedure itself is addressed in Section III and numerical results are discussed in Section IV. Section V provides some concluding remarks.

II. Problem Setup

We start by introducing our notation. Random variables are always denoted by capital letters U, where its realization is denoted by the corresponding lowercase letter u. Vectors of random variables are denoted by bold capital letters and always assumed to be column vectors $\mathbf{U} = (U_1, U_2, \cdots, U_M)^T$, whereas its realization is denoted by the corresponding bold lowercase letter $\mathbf{u} = (u_1, u_2, \cdots, u_M)^T$. Index sets are denoted by capital calligraphic letters $\mathcal{M}$ unless otherwise noted, and $|\mathcal{M}|$ denotes the set's cardinality. We follow the convention, that variables indexed by a set denote a set of variables, e.g. if $\mathcal{M} = \{1, 2, 3\}$ then $u_{\mathcal{M}} = \{u_1, u_2, u_3\}$, and use the same concept to define variable vectors, such that $\mathbf{u}_{\mathcal{M}} = (u_1, u_2, u_3)^T$. The covariance is defined by $Cov\{\mathbf{a}, \mathbf{b}\} \triangleq E\{\mathbf{a}\mathbf{b}^T\} - E\{\mathbf{a}\}E\{\mathbf{b}\}^T$, where $E\{\cdot\}$ is the expectation operator. An M-dimensional random variable with realizations $\mathbf{u} = (u_1\, u_2, \cdots, u_M)^T$, $u_m \in \mathbb{R}$, is Gaussian distributed with mean $\boldsymbol{\mu} = E\{\mathbf{u}\}$ and covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}\{\mathbf{u}, \mathbf{u}\}$, when its PDF $p_{\mathbf{U}}(\mathbf{u})$ is given by

$$p_{\mathbf{U}}(\mathbf{u}) = \exp(-\frac{1}{2}(\mathbf{u} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{u} - \boldsymbol{\mu}))/(2\pi|\boldsymbol{\Sigma}|)^{1/2}. \quad (1)$$

Such a PDF is simply denoted as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The expression $\mathbf{0}_M$ is the length-M all-zero column vector, $\mathbf{I}_M$ is the $M \times M$ identity matrix, and $|\mathbf{A}|$ is the determinant of $\mathbf{A}$.

A. System Model

Each sensor indexed by m, with $m \in \mathcal{M} = \{1, 2, \cdots, M\}$, observes continuous real-valued source samples $u_m(t)$ at time t. For simplicity, we consider only the spatial correlation of measurements and not their temporal dependence such that the

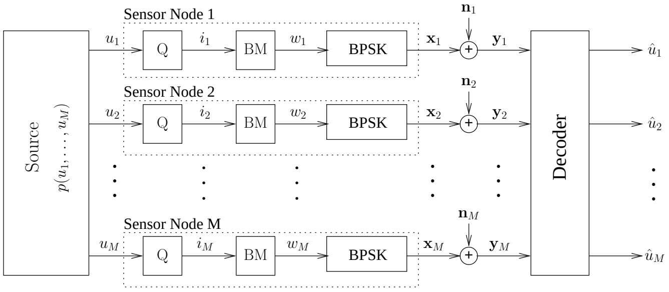


Fig. 1. System model - M correlated samples are separately encoded by M sensor nodes, consisting of a scalar quantizer (Q) a index assignment stage (BM) and a simple modulator (BPSK). The data is then transmitted over an array of independent AWGN channels and decoded jointly at the receiver.

The authors are with the Department of Computer Science and the Laboratory of Artificial Intelligence and Computer Science, Universidade do Porto, Portugal. This work was partly supported by the Fundação para a Ciêcia e Tecnologia (Portuguese Foundation for Science and Technology) under grant POSC/EIA/62199/2004.

time index t is dropped and only one time step is considered. The sample vector $\mathbf{u} = \mathbf{u}_{\mathcal{M}} = (u_1, u_2, \cdots, u_M)^T$ at any given time t is assumed to be one realization of a M-dimensional Gaussian random variable, whose PDF $p_{\mathbf{U}}(\mathbf{u})$ is given by $\mathcal{N}(\mathbf{0}_M, \mathbf{R})$ with

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,M} \\ \rho_{2,1} & 1 & \cdots & \rho_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{M,1} & \rho_{M,2} & \cdots & 1 \end{bmatrix}.$$

It follows that the samples u_m have zero mean $E\{u_m\}$ and unit variance $Cov\{u_m, u_m\}$. Gaussian models for capturing the spatial correlation between sensors at different locations are discussed in [16], whereas examples of reasonable models for the correlation coefficients $\rho_{m,m'} = E\{u_m u_{m'}\}$ of physical processes unfolding in a field can be found in [3].

We assume that the sensors are low-complexity devices consisting of a scalar quantizer (Q), an index assignment (BM) and a modulator (BPSK), see Figure 1. Each sensor m quantizes its observation u_m onto the quantization index $i_m \in \mathcal{I}_m = \{1, 2, \ldots, |\mathcal{I}_m|\}$, which is then mapped onto the output codeword $w_m \in \mathcal{W}_m = \{1, 2, \ldots, |\mathcal{W}_m|\}$ as defined by the index assignment stage.

The number of bits used for the codewords is denoted as Q_m such that $Q_m = \lceil \log_2(|\mathcal{W}_m|) \rceil$. In the following, the vector of quantization indices is denoted by $\mathbf{i} = \mathbf{i}_{\mathcal{M}} = (i_1, i_2, \cdots, i_M)^T$ and the vector of codewords is denoted by $\mathbf{w} = \mathbf{w}_{\mathcal{M}} = (w_1, w_2, \cdots, w_M)^T$.

Prior to transmission the modulator maps the binary representation of the codewords w_m, which is determined by standart Gray mapping, to a tuple $\mathbf{x}_m$ of channel symbols, which are transmitted to the remote receiver. In our examples we use binary phase shift keying (BPSK), such that in a discrete-time baseband description of our transmission scheme w_m is mapped to Q_m symbols $\mathbf{x}_m = (x_{m,1} \ldots x_{m,Q_m})$, $x_{m,q} \in \{+1, -1\}$.

We assume a multiple access channel with orthogonal accessing, such that $p(\mathbf{y}_1, \ldots, \mathbf{y}_M | \mathbf{x}_1, \ldots, \mathbf{x}_M)$ factors into $\prod_{m=1}^{M} p(\mathbf{y}_m | \mathbf{x}_m)$. The reachback channel then consists of an array of additive white Gaussian noise channel with noise variance η^2. Assuming coherent demodulation, we write the channel outputs as $\mathbf{y}_m = \mathbf{x}_m + \mathbf{n}_m$, $m = 1, 2, \ldots, M$ where $\mathbf{n}_m$ is distributed according to $\mathcal{N}(\mathbf{0}_Q, \sigma^2 \mathbf{I}_Q)$.

The decoder uses the channel output vector $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_M)^T$ and the available knowledge of the source correlation $\mathbf{R}$ to produce the estimates $\hat{u}_m$ of the measurements u_m. Assuming, that the mean square error (MSE) $E\{(\hat{u}_m - \tilde{u}_m)^2\}$, between the estimate $\hat{u}_m$ and the source representation $\tilde{u}_m$, corresponding to the quantization index i_m, is the fidelity criterion to be minimized by the decoder, the conditional mean estimator (CME) [13] should be applied:

$$\hat{u}_m = E\{\tilde{u}_m(i_m) | \mathbf{y}\} = \sum_{\forall i \in \mathcal{I}_m} \tilde{u}_m(i) \cdot p(i_m = i | \mathbf{y}). \quad (2)$$

Notice, that for PDF-optimized quantizers this estimator also minimizes the MSE $E\{(\hat{u}_m - u_m)^2\}$ between the estimate $\hat{u}_m$ and the originally observed value u_m [8]. The required posterior probabilities $p(i_m | \mathbf{y})$ are given by

$$p(i_m = i | \mathbf{y}) = \gamma \cdot \sum_{\forall \mathbf{i}: i_m = i} p(\mathbf{y} | \mathbf{i}) p(\mathbf{i}), \quad (3)$$

where $\mathbf{i} = \mathbf{i}_{\mathcal{M}} = (i_1 \, i_2 \ldots i_M)^T$ and $\gamma = 1/p(\mathbf{y})$ is a constant normalizing the sum over the product of probabilities to one. Since the channels are independent $p(\mathbf{y}|\mathbf{i})$ factors into $\prod_{m=1}^{M} p(\mathbf{y}_m | i_m)$.

Notice, that the index assignments used at the encoders can be described statistically by transition probabilities $p(w_m | i_m)$, $\forall i_m \in \mathcal{I}_m$ and $w_m \in \mathcal{W}_m$, such that $p(w_m | i_m) = 1$, if i_m is mapped onto w_m, and $p(w_m | i_m) = 0$, otherwise. Using this statistical representation, we are able to substitute the probabilities $p(\mathbf{y}_m | i_m)$ in (3) by $p(\mathbf{y}_m | i_m) = \sum_{\forall w \in \mathcal{W}_m} p(\mathbf{y}_m | w_m = w) \cdot p(w_m = w | i_m)$, allowing us to fully take into account the index assignments for decoding.

B. Main Goals

Under the system model described above, distributed source coding can be achieved by (jointly) designing index assignments with $|\mathcal{W}_m| < |\mathcal{I}_m|$ for several encoders[1], such that the spatial redundancy within the data is exploited to reduce the data rate to $Q_m = \lceil \log_2(|\mathcal{W}_m|) \rceil$ [bits/sample] while using as optimization criteria the resulting end-to-end distortion after joint decoding. One possible way to design such index assignments was proposed by Flynn and Gray in [4], who suggested a heuristic search algorithm reusing the indices of a high resolution quantizer. This conceptual approach can be adapted for our purposes as outlined in Section IV. The central problem in the design of distributed source codes, which do not rely on random-binning techniques [1, p. 410], is that systematic code design is in most cases only feasible for a small number of encoders. Seeking a scalable solution for large-scale sensor networks, we focus on finding correlation preserving clusters of limited size, giving rise to a computational feasible design process for the encoders within each of those clusters.

In previous work [8] it was shown that the complexity of optimal decoding using (2) grows exponentially with the number of sensors M. Therefore, a scalable solution for decoding based on factor graphs and the sum-product (SP) algorithm was proposed, using an approximated PDF $\hat{p}_{\mathbf{U}}(\mathbf{u})$ instead of $p_{\mathbf{U}}(\mathbf{u})$ as basis for decoding. To reduce decoding complexity, $p_{\mathbf{U}}(\mathbf{u})$ was modeled by an underlying factorization, which itself can be represented by a factor graph and used for efficient decoding. Specifically, factorizations forming a constrained chain rule expansion (CCRE) minimizing the Kullback-Leibler distance (KLD) between $p_{\mathbf{U}}(\mathbf{u})$ and $\hat{p}_{\mathbf{U}}(\mathbf{u})$, to be introduced in Section III-A, were deemed to be a suitable choice for multiterminal sensing scenarios under a minimum mean squared error (MMSE) criterion. Capitalizing on these insights, our goal is to construct KLD optimized CCRE corresponding to correlation preserving clusters that enable efficient index assignments for large-scale distributed quantization and sum-product decoding.

III. Source optimized Clustering

A. Preliminaries

A PDF $p_{\mathbf{U}}(\mathbf{u})$ can be approximated by assuming that $p(\mathbf{u})$ can be factorized by $\hat{p}(\mathbf{u}) = \prod_{n=1}^{N} f_n(\mathbf{u}_{\mathcal{S}_n})$, where $\mathcal{S}_n \subseteq \mathcal{M}$, with $n = 1, 2, \cdots, N$, are subsets of source indices, such that

[1] Such index assignments generally increase the distortion of the system, because information is lost during the mapping process, i.e. more than one quantization index i_m might lead to one and the same codeword w_m. However, since the rate can be reduced considerably, this method offers a way to achieve a wider range of rate/distortion trade-offs

$\bigcup_{n=1}^{N} \mathcal{S}_n = \mathcal{M}$. Since generally $p(\mathbf{u}) \neq \hat{p}(\mathbf{u})$, the resulting PDF $\hat{p}_{\mathbf{U}}(\mathbf{u})$ is an approximation of $p_{\mathbf{U}}(\mathbf{u})$.

A constrained chain rule expansion (CCRE) of $p(\mathbf{u})$ can be obtained from the regular chain rule expansion by removing some of the conditioning variables. More formally, a factorization

$$\hat{p}(\mathbf{u}) = \prod_{n=1}^{N} f_n(\mathbf{u}_{\mathcal{S}_n}) = \prod_{n=1}^{N} p(\mathbf{u}_{\mathcal{A}_n}|\mathbf{u}_{\mathcal{B}_n}), \quad (4)$$

where $\mathcal{A}_n$, $\mathcal{B}_n$ and $\mathcal{S}_n = \mathcal{A}_n \cup \mathcal{B}_n$ are subsets of the elements in $\mathcal{M}$, is a CCRE of $p(\mathbf{u})$, if the following constraints are met:

$$\mathcal{A}_n \cap \mathcal{B}_n = \emptyset, \ \mathcal{B}_n \subseteq \bigcup_{l=1}^{n-1} \mathcal{A}_l, \ \bigcup_{n=1}^{N} \mathcal{A}_n = \mathcal{M}. \quad (5)$$

Thus, the set $\mathcal{B}_1$ is always empty and for the usual chain rule expansion holds $\mathcal{B}_n = \bigcup_{l=1}^{n-1} \mathcal{A}_l$. We call a CCRE symmetric, if any $\mathcal{B}_n$ with $n = 2, 3, \cdots, N$ is a subset of $\mathcal{S}_l$ for some $l < n$.

The Kullback-Leibler distance (KLD) between a PDF $p_{\mathbf{U}}(\mathbf{u})$ and its approximation $\hat{p}_{\mathbf{U}}(\mathbf{u})$

$$D(p_{\mathbf{U}}(\mathbf{u})||\hat{p}_{\mathbf{U}}(\mathbf{u})) = \int \cdots \int p(\mathbf{u}) \log_2 \frac{p(\mathbf{u})}{\hat{p}(\mathbf{u})} \, d\mathbf{u} \quad (6)$$

is measured in bit [1] and can be used as optimization criteria when constructing source factorizations. In [8] it was shown, that the KLD can be calculated explicitly for a CCRE's of Gaussian PDF's $\mathcal{N}(\mathbf{0}_M, \mathbf{R})$ using

$$D(p_{\mathbf{U}}(\mathbf{u})||\hat{p}_{\mathbf{U}}(\mathbf{u})) = -\frac{1}{2}\log_2|\mathbf{R}| + \frac{1}{2}\sum_{n=1}^{N}\log_2\frac{|\mathbf{R}_{\mathcal{S}_n}|}{|\mathbf{R}_{\mathcal{B}_n}|}, \quad (7)$$

where $\mathbf{R}_{\mathcal{S}_n}$ and $\mathbf{R}_{\mathcal{B}_n}$ are the covariance matrices of the Gaussian PDF's $p_{\mathbf{U}_{\mathcal{S}_n}}(\mathbf{u}_{\mathcal{S}_n})$ and $p_{\mathbf{U}_{\mathcal{B}_n}}(\mathbf{b}_{\mathcal{B}_n})$, respectively.

B. Hierarchical Clustering Algorithm

The algorithm described in the following is based on the principles of hierarchical clustering [9] and can be seen as a variant of the Ward algorithm [11]. The goal is to cluster the set of sources $\mathcal{M}$, into subsets $\Lambda_c \subseteq \mathcal{M}$, such that $\bigcup_{c=1}^{C} \Lambda_c = \mathcal{M}$ and $\Lambda_i \cap \Lambda_j = \emptyset$, for all $i \neq j$ with $\{i, j\} \in \Gamma$, $\Gamma = \{1, \cdots, C\}$ and $C = |\mathcal{M}|$. The maximum cluster size is denoted by S, such that $|\Lambda_c| \leq S$, for all $c \in \Gamma$.

The clusters are constructed by a successive merging process, starting with a set of single-element clusters $\Lambda'_s = \{s\}$ with indices $s \in \Gamma' = \{1, \cdots, M\}$, which are merged step-by-step together, taking into account the KLD between the approximated PDF $\check{p}_{\mathbf{U}}(\mathbf{u})$, based on the factorization $\check{p}(\mathbf{u}) = \prod_{\forall s \in \Gamma'} p(\mathbf{u}_{\Lambda'_s})$, and the original PDF $p_{\mathbf{U}}(\mathbf{u})$ as an *objective* function. The first step of the procedure is to select two of those clusters which, when united, reduce the number of clusters by one. The algorithm determines for each possible pair of clusters (Λ'_k, Λ'_l), with $\{k, l\} \in \Gamma'$ and $k \neq l$, the current value of the objective function to find the pair (Λ'_k, Λ'_l) leading to the smallest KLD between original and approximated PDF. As soon as all possible mergings are evaluated the indices of the selected clusters $\{k, l\}$ are removed from Γ' and the index of the newly created cluster r is added to Γ'. This procedure is repeated until only a single cluster remains and a history of all mergings performed during the different stages of the optimization procedure is obtained. In Figure 2 (a) the merging process is illustrated for an exemplary scenario. A graphical representation of the mergings performed during the optimization process (the so called dendrogram [9]) is shown in Figure 2 (b).

Using (7), the objective function is given by

$$D(p_{\mathbf{U}}(\mathbf{u})||\check{p}_{\mathbf{U}}(\mathbf{u})) = -\frac{1}{2}\log_2|\mathbf{R}| + \frac{1}{2}\sum_{\forall s \in \Gamma'}\log_2|\mathbf{R}_{\Lambda'_s}|, \quad (8)$$

where $\mathbf{R}_{\Lambda'_s}$ denotes the covariance matrix for the source variables collected in the vector $\mathbf{u}_{\Lambda'_s}$ which can be easily obtained from $\mathbf{R}$ using the techniques described in [8]. Since the objective function has to be evaluated many times during the optimization process, we will express $D(\cdot)$ based on intermediate results to reduce computational complexity. The KLD improvement imposed by the statistical dependencies within an arbitrary cluster Λ'_s, with $s \in \Gamma'$, can be expressed as

$$\Delta D_s = \frac{1}{2} \cdot \log_2|\mathbf{R}_{\Lambda'_s}|. \quad (9)$$

The differential KLD improvement associated with the merging of an arbitrary pair of clusters (Λ'_k, Λ'_l), with $\{k, l\} \in \Gamma'$ with $k \neq l$, into a new cluster $\Lambda'_{k\cup l} = \Lambda'_k \cup \Lambda'_l$, can be calculated by

$$\Delta D'_{k\cup l} = (\Delta D_{k\cup l} - \Delta D_k - \Delta D_l), \quad (10)$$

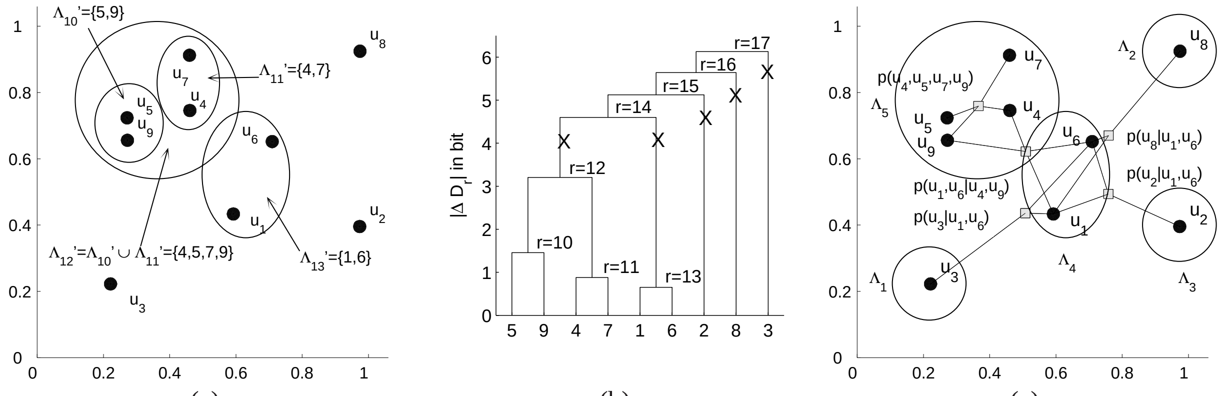
where $\Delta D_{k\cup l}$ is the KLD improvement imposed by cluster $\Lambda'_{k\cup l}$ calculated according to (9). Knowing $\Delta D'_{k\cup l}$, the impact of currently considered merging onto the objective function according to (8) can be evaluated and only $\Delta D'_{k\cup l}$ as well as $\Delta D_{k\cup l}$ need to be calculated for each merging performed. ΔD_k and ΔD_l can be obtained from previous results and do not need to be re-calculated each time.

The source clusters Λ_c with a maximum targeted cluster size S can then be constructed using the dendrogram derived before, see *Fig. 2*(b). We start at the root of the dendrogram —which basically is a tree— and descend along its branches to lower hierarchical levels. While moving from one level to the next lower one, the dendrogram branches into two subtrees. The number of leafs, i.e. the number of sources m connected to each subtree are counted and if the number of leafs of one (or both) subtree(s) is smaller/ equal than S, we cut the corresponding subtree out of the dendrogram. This pruning process is repeated until all leafs are removed from the dendrogram. When the pruning is finished, the subtrees are labeled by an successively increasing index $c = 1, \cdots, C$. We define the set of cluster indices as $\Gamma = \{1, \cdots, C\}$. The source clusters Λ_c can be then simply be determined from the subtrees by assigning the variables m, associated with each subtree's leafs, to the corresponding cluster, see example in Figure 2 (b).

Because of the hierarchical merging concept based on local decisions, the proposed clustering algorithm is in general sub-optimal. However, the hierarchical approach has the advantage, that the resulting dendrogram can be used elegantly to construct clusters with a bounded number of source variables S (with e.g. partitional clustering methods [9] this would not be an easy task).

C. PDF Optimized Source Factorization

In the last section we have shown, how KLD optimized clusters fitting our purposes can be constructed. The second step towards our goal of obtaining a source factorization is to incorporate derived clusters into a symmetric CCRE of form $\hat{p}(\mathbf{u}) = \prod_{n=1}^{N} f_n(\mathbf{u}_{\mathcal{S}_n})$ (conditions in (5) have to be met), which is achieved by *linking* the unconnected clusters Λ_c, with $c \in \Gamma$, successively together.



(a) (b) (c)

Fig. 2. Example - PDF optimized clustering procedure with $M = 9$ uniformly distributed sensors picking-up observations $u_1, ..., u_9$: (a) Mergings performed during the hierarchical clustering procedure for the first four iteration steps leading to resulting clusters Λ'_r with indices $r = 10, \cdots, 13$. (b) Tree representation of the mergings performed during different the stages of the optimization process (dendrogram). The KLD improvement $|\Delta D_r|$ in [bit] of the created clusters Λ'_r is provided for all iteration steps. For a maximum cluster size of $S = 4$, the branches of the tree, which are cut during the pruning process, are marked with a cross and the source clusters $\Lambda_1 = \{3\}$, $\Lambda_2 = \{8\}$, $\Lambda_3 = \{2\}$, $\Lambda_4 = \{1, 6\}$ and $\Lambda_5 = \{4, 5, 7, 9\}$ can be defined. (c) Source factorization obtained when using $A = 2$ source variables in the originating and $B = 2$ source variables in the destination cluster. The factor graph represents the symmetric CCRE $\hat{p}(\mathbf{u}) = p(u_4, u_5, u_7, u_9) \cdot p(u_1, u_6|u_4, u_9) \cdot p(u_2|u_1, u_6) \cdot p(u_8|u_1, u_6) \cdot p(u_3|u_1, u_6)$.

Assuming, that we want to link (already connected) cluster Λ_k with cluster Λ_l, we have to choose a set of variables $\mathcal{G}_k \subseteq \Lambda_k$ and a set of variables $\mathcal{H}_l \subseteq \Lambda_l$ to establish a link between both clusters. Since complexity of the scalable decoding highly depends on the number of variables within the single factors (refer [8] for details), we introduce the design parameters A and B, such that $|\mathcal{G}_k| \leq A$ and $|\mathcal{H}_l| \leq B$ for all $\{k, l\} \in \Gamma$. While establishing the link between the clusters the factors $p(\mathbf{u}_{\mathcal{H}_l}|\mathbf{u}_{\mathcal{G}_k})$ and $p(\mathbf{u}_{\mathcal{G}_l}|\mathbf{u}_{\mathcal{H}_l})$ are added to the source factorization and we define the KLD improvement associated with this link as

$$\delta D_{k\rightarrow l} = \frac{1}{2}\log_2 \frac{|\mathbf{R}_{\mathcal{G}_k \cup \mathcal{H}_l}|}{|\mathbf{R}_{\mathcal{G}_k}|}. \qquad (11)$$

After selecting an arbitrary cluster as starting point for the linking process, we take one of the unconnected clusters (i.e. a cluster, which is not yet considered in the source factorization) and link it with the already connected clusters (i.e. incorporate it into the source factorization). Assuming, that we choose the cluster $r \in \Gamma$ as starting point of optimization, we define the source factorization as the factor $p(\mathbf{u}_{\Lambda_r})$, define a set of linked clusters $\Gamma' = \{r\}$ and define a set of unconnected clusters $\overline{\Gamma'} = \Gamma\backslash\{r\}$. During the optimization process, we successively select a pair of clusters (Λ_k, Λ_l), with $k \in \Gamma'$, $l \in \overline{\Gamma'}$, choose the subsets $\mathcal{G}_k \in \Lambda_k$ and $\mathcal{H}_l \in \Lambda_l$, with $|\mathcal{G}_k| \leq A$ and $|\mathcal{H}_l| \leq B$, such that the profit $|\delta D_{k\rightarrow l}|$ associated with establishing this link is maximized and add the factors $p(\mathbf{u}_{\mathcal{H}_l}|\mathbf{u}_{\mathcal{G}_k})$ and $p(\mathbf{u}_{\mathcal{G}_l}|\mathbf{u}_{\mathcal{H}_l})$ to the source factorization. For each step of the procedure, the index l is added to the set of linked clusters, i.e. $\Gamma' = \Gamma' \cup \{l\}$, and removed from the set unconnected clusters, i.e. $\overline{\Gamma'} = \overline{\Gamma'}\backslash\{l\}$. This is done until all clusters are linked, i.e. $|\Gamma'| = |\Gamma|$. Figure 2 (c) shows the resulting source factorization for our previous example.

The linking process described before is generally not optimal. It can be shown that the optimal way to link a given set of clusters together leads to the (more complex) directed spanning tree problem, which we will show and address in future work. For the case where $A = B = 1$ the directed spanning tree problem however degrades to the undirected case, which is directly reflected in presented linking procedure and therefore leads to an optimal result for this special case.

IV. Results and Discussion

To underline the effectiveness and efficiency of our low-complexity coding strategies we present numerical performance results for an exemplary scenario of randomly placed sensors. We consider a square unit area element with $M = 100$ uniformly distributed sensors. The sensor measurements u_m are Gaussian distributed $\mathcal{N}(0, 1)$. As outlined in Section II-A, we assume, that sensor measurements $\mathbf{u} = \mathbf{u}_{\mathcal{M}} = (u_1, \cdots, u_M)^T$ are distributed according to a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}_M, \mathbf{R})$, where the correlation between a pair of sensors u_k and u_l decreases exponentially with the distance $d_{k,l}$ between them, such that $\rho_{k,l} = \exp(-\beta \cdot d_{k,l})$. Since the performance of our techniques depend on the correlations between the sensors, we consider two different source models, one with $\beta = 0.5$ (strongly correlated sensor measurements) and one with $\beta = 2$ (weakly correlated measurements). All scalar quantizers at the encoders are Lloyd-Max optimized to minimize the mean squared error (MSE) in the sensor readings $u_1, \cdots, u_M$ using identical resolution for quantization and identical rates for data transmission, i.e. $|\mathcal{I}_m| = L$ and $Q_m = Q$, for $m = 1, ...M$, where $L \leq 16$ was chosen. The clusters Λ_c are derived as described in Section III-B with a maximum cluster size of $S = 4$ and the index assignments are constructed consecutively for all $c \in \Gamma$, with $\Gamma = \{1, \cdots, C\}$, see Figure 3. The index assignments of all clusters Λ_c with $|\Lambda_c| > 1$, are constructed using a heuristic search algorithm inspired by the techniques

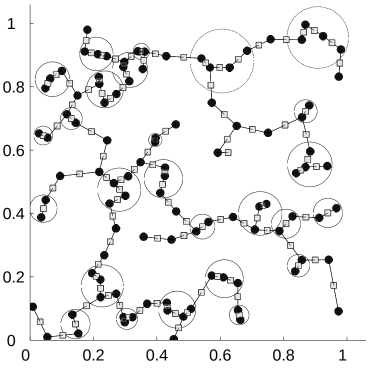


Fig. 3. Simulation scenario - Graphical representation of the KLD optimized source factorization for $M = 100$ uniformly distributed sensors with correlation factor $\beta = 0.5$. The clusters with a maximum size of $S = 4$ (indicated by circles) were created using the hierarchical clustering method and linked together by choosing $A = B = 1$.

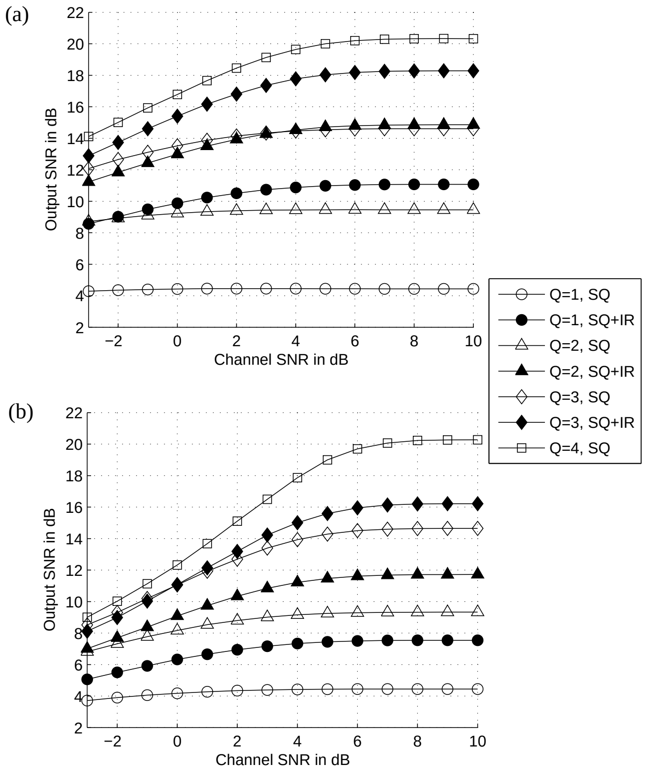


Fig. 4. Simulation results - Performance of the system with $M = 100$ sensors for correlation factor $\beta = \{0.5,\ 2\}$ when simple scalar quantization (SQ) alone and scalar quantization with a subsequent index-reuse (SQ+IR) is used at the encoder. All encoders use identical rates of Q [bits/sample] for the data transmission.

described by Flynn and Gray in [4], which reuses the indices of a high-resolution quantizer to successively reduce the data-rate needed for transmission using the overall end-to-end distortion $d_{\Lambda_c} = E\{||\mathbf{u}_{\Lambda_c} - \hat{\mathbf{u}}_{\Lambda_c}||^2\}$ of all sources $m \in \Lambda_c$ as optimization criteria. Since it is not possible to construct index assignments for clusters with $|\Lambda_c| = 1$, we chose in this case a scalar quantizer (Lloyd-Max optimized as before) with decreased resolution and no index assignments, such that $Q_m = Q$ is still guaranteed for all encoders $m \in \mathcal{M}$. The source factorization used for decoding is based on the clusters Λ_c, $c \in \Gamma$, which are linked together as described in Section III-C. To reduce the decoder complexity even further, we approximate the factors $f_n(\mathbf{u}_{\mathcal{S}_n})$, with $n = 1, \cdots, N$ and $|\mathcal{S}_n| > 2$, by a factorization itself (using the same techniques as described in Section III-C) to create an overall source factorization, whose factors contain two variables at maximum (see Figure 3). The decoder is based on the sum-product algorithm as described in [8], where the required PMF's were obtained by Monte Carlo simulation, using Lloyd-Max optimized quantizers with resolution $L_m = L$ for $m = 1, 2, \cdots, M$. To evaluate the performance of the coding strategies we measure the output signal-to-noise ratio (SNR) given by

$$\text{Output SNR} = 10 \cdot \log_{10}\left(\frac{||\mathbf{u}||^2}{||\mathbf{u} - \hat{\mathbf{u}}||^2}\right) \text{ in dB} \qquad (12)$$

versus the channel SNR $= 10 \cdot (E_s/N_0)$ in dB averaged over a $M \times 10000$ source samples. The simulation results of our system are depicted in Figure 4 (a) for strongly and in Figure 4 (b) for weakly correlated sources. In both scenarios, we consider the performance achieved when using scalar quantization alone at the encoder, i.e. where the performance is mainly governed by the properties of the decoder, and the performance achieved when scalar quantization with a subsequent index-reuse is used for encoding. Notice, that only the index assignments yielding best possible performance were chosen for the experiments (e.g. a rate of $Q_m = 1$ [bits/sample] may be obtained from quantizers of $|\mathcal{I}_m| = 4$, $|\mathcal{I}_m| = 8$ or $|\mathcal{I}_m| = 16$). Our simulation results reveal, that simple index assignment techniques applied to local clusters can achieve significant performance gains using our coding approach, especially for low data-rates.

V. Conclusions

Concerned with the scalability of distributed quantization in large-scale sensor networks, we proposed a practical solution for code design based on a very low-complexity encoding stage (scalar quantization followed by simple index assignments) and a specific hierarchical clustering algorithm.

Our simulation results reveal, that despite the simplicity of the proposed encoding techniques, significant performance gains can be achieved by the proposed joint encoding/ decoding approach. Possible extensions include other classes of index assignments, e.g. based on channel codes with appropriate distance properties [14], and cluster-based distributed inference.

References

[1] T. M. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., 1991.

[2] Bernd Girod David Rebollo-Monedero, Rui Zhang. Design of Optimal Quantizers for Distributed Source Coding. In *Proceedings of the Data Compression Conference (DCC03)*, 2003.

[3] C. R. Dietrich and G. N. Newsam. Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, 18(4):1088–1107, 1997.

[4] T. J. Flynn and R. M. Gray. Encoding of Correlated Observations. *IEEE Trans. Inform. Theory*, IT-33(6):773–787, 1987.

[5] Gerhard Maierbacher, João Barros. Low-Complexity Coding for the CEO Problem with many Encoders. In *Twenty-sixt Syposium on Information Theory in the Benelux*, Brussels, Belgium, 2005.

[6] Harish Viswanathan, Toby Berger. The quadratic Gaussian CEO problem. *IEEE Trans. Inform. Theory*, 43:1549–1559, 1997.

[7] Wendi Rabiner Heinzelman, Anantha Chandrakasan, and Hari Balakrishnan. Energy-efficient communication protocol for wireless microsensor networks. In *HICSS*, 2000.

[8] J. Barros and M. Tuechler. Scalable Decoding on Factor Trees: A Practical Solution for Sensor Networks. *IEEE Transactions on Communications*, 54:284–294, 02 2006.

[9] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[10] Gilles Van Assche Jean Cardinal. Joint Entropy-Constrained Multiterminal Quantization. In *Proceedings of the International Symposium on Information Theory*, Lausanne, Switzerland, 2002.

[11] J. Ward Jr. Hierachical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.

[12] SangHak Lee, JuneJae Yoo, and TaeChoong Chung. Distance-based energy efficient clustering for wireless sensor networks. In *Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks (LCN04)*, pages 567–568, 2004.

[13] H. V. Poor. *An Introduction to Signal Detection and Estimation*. Springer-Verlag, 1994.

[14] S. S. Pradhan and K. Ramchandran. Distributed Source Coding Using Syndromes (DISCUS): Design and Construction. In *Proc. IEEE Data Compression Conf. (DCC)*, Snowbird, UT, 1999.

[15] S. Sandeep Pradhan, Kannan Ramchandran. Generalized Coset Codes for Distributed Binning. *IEEE Trans. Inform. Theory*, 51:3457–3474, 2005.

[16] A. Scaglione and S. D. Servetto. On the Interdependence of Routing and Data Compression in Multi-Hop Sensor Networks. In *Proc. ACM MobiCom*, Atlanta, GA, 2002.

[17] D. Slepian and J. K. Wolf. Noiseless Coding of Correlated Information Sources. *IEEE Trans. Inform. Theory*, IT-19(4):471–480, 1973.

[18] S. Cheng Z. Xiong, A. D. Liveris. Distributed source coding for sensor networks. *IEEE Signal Processing Magazine*, 09 2004.