ANALYZING DATA INTO QUANTIZED COMPONENTS

Konstantinos Diamantaras[a], Theophilos Papadimitriou[b], Konstantinos Goulianas[a]

[a]Technological Education Institute of Thessaloniki
Department of Information Technology
Sindos 57400, Greece
{kdiamant,gouliana}@it.teithe.gr

[b]Democritus University of Thrace
Department of Economics
Komotini 69100, Greece
papadimi@ierd.duth.gr

ABSTRACT

Signals in various applications are often generated by linear combinations of quantized components. The analysis of data into such components is treated here as a matrix analysis problem. We first show that the component alphabet can always be normalized to the levels 0, ..., $M-1$, without loss of generality. Then we study certain conditions under which the decomposition is possible. In particular, we present an analytical algorithm based on the differences of the observed points and the recursive estimation of the quantized components when the number of unique observed points is sufficiently large.

Index Terms— matrix factorization, data analysis, quantized components, blind source separation

1. INTRODUCTION

In various applications the data that are observed or collected can be described as linear combinations of components whose samples take values from a finite alphabet. For example, in digital communications, various popular modulation schemes such as PAM, BPSK, or QAM, generate quantized sources which are commonly observed at the receiver mixed with other sources or mixed with delayed copies of themselves due to multipath [1]. Also in social or economic analysis, observed quantities or statistical measurements depend on discrete variables. Oftentimes these variables are binary (e.g. gender, ownership of a car, etc) but multi-level variables are also common (number of children, education level, number of accident per month, e.a.) [2]. In such cases, the analysis of the observed data should be constrained to produce quantized (discrete) components. Traditional component methods such as PCA or ICA are not immediately applicable for this task.

1.1. Relation to prior work

The analysis of signals into components is a very old idea dating back to the pioneering work of Pearson [3] who studied the optimal fit of linear hyperplanes to the data. His approach turned out to be equivalent to the analysis of signals into uncorrelated components, later known as Principal Component Analysis (PCA) [4], a name coined by Hotelling in 1933 [5]. PCA is a cornerstone method for statistical analysis which has found numerous applications in a large number of fields such as data compression, image processing, face recognition, economic data analysis, etc. Recent decades saw the development of many alternative approaches to data analysis based on different assumptions. In Independent Component Analysis (ICA) [6, 7] the data are represented by linear or non-linear combinations of components that are assumed to be mutually statistically independent. Non-negative matrix factorization [8] is another popular data decomposition method where the factors of the data matrix are constrained to be non-negative. In sparse signal decomposition (also known as *sparse representation* or *compressive sensing*) [9, 10] the data are represented by a limited number of non-zero components by minimizing their l_0 or l_1 norm.

More recently, a number of approaches have proposed the statistical analysis of data into discrete components with applications mainly on document analysis and socioeconomical sciences. Such approaches include the Latent Dirichlet Allocation [11], Probabilistic Latent Semantic Indexing [12], Gamma-Poisson models [13], multinomial PCA [14]. These methods have been unified under the statistical perspective of the so called Discrete Principal Component Analysis [15]. Also, in [16, 17] Gutch e.a. have approached the problem of independent component analysis over finite fields using a probabilistic framework. The Blind Separation of multi-level sources [18] is another statistical approach for obtaining signal components that take values from a discrete alphabet. This last work viewed the problem from a statistical perspective and can be seen as the precursor to the present work.

In this paper we approach the problem of data analysis into quantized components from a matrix factorization point of view. We make no particular assumptions regarding the probability distribution of the input nor of the distribution of the basis (mixing) vectors. In that sense, our approach is more similar to the Non-negative matrix factorization problem.

2. PROBLEM FORMULATION

Consider a real, discrete time signal $\mathbf{x}(k) \in \mathbb{R}^m$ which is generated by a shifted linear combination of n real, quantized components $\bar{s}_i(k) \in \bar{\mathcal{A}}_M$, $i = 1, \cdots, n$:

$$\mathbf{x}(k) = \sum_{i=1}^{n} \mathbf{c}_i \bar{s}_i(k) + \bar{\mathbf{b}} \tag{1}$$

for some $\mathbf{c}_1$, ..., $\mathbf{c}_n$, $\bar{\mathbf{b}}$, where $\bar{\mathcal{A}}_M = \{\bar{\alpha}_0, \cdots, \bar{\alpha}_{M-1}\}$ is the alphabet consisting of M discrete symbols (also called *levels*). Without loss of generality we assume that the levels are arranged in increasing order and the distance between consecutive levels is equal to 1, thus: $\bar{\alpha}_p = \bar{\alpha}_0 + p, \quad p = 0, \cdots, M-1$. Subtracting the fixed offset

This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALES. Investing in knowledge society through the European Social Fund.

$\bar{\alpha}_0$ from the input symbols we obtain the new discrete components

$$s_i(k) = \bar{s}_i(k) - \bar{\alpha}_0$$

which take values from the "normal" alphabet

$$\mathcal{A}_M = \{0, \cdots, M-1\}$$

and (1) becomes

$$\mathbf{x}(k) = \sum_{i=1}^{n} \mathbf{c}_i s_i(k) + \mathbf{b} \tag{2}$$

with $\mathbf{b} = \bar{\mathbf{b}} + \bar{\alpha}_0 \sum_{i=1}^{n} \mathbf{c}_i$. Given a specific set of data for $k = 1, \cdots, K$, we can rewrite (1) in matrix form as

$$\mathbf{X} = \begin{bmatrix} \mathbf{C} \; \mathbf{b} \end{bmatrix} \begin{bmatrix} \mathbf{S} \\ \mathbf{u}_K^T \end{bmatrix} \tag{3}$$

where $\mathbf{u}_K$ is a K-dim vector of all 1's. Therefore, the signal decomposition (1) from any generic alphabet $\bar{\mathcal{A}}_M$ is equivalent to the signal decomposition (2) from the normalized alphabet $\mathcal{A}_M$.

Our goal is to solve the following problem:

Quantized Matrix Factorization: Given a matrix $\mathbf{X} \in \mathbb{R}^{m \times K}$ generated by Eq. (3) find $\mathbf{C} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$ and $\mathbf{S} \in \mathbb{R}^{n \times K}$ with $s_{ij} \in \mathcal{A}_M$, assuming that we know the number of symbols M and the number of components n.

Clearly for any permutation matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ we have $\mathbf{C}'\mathbf{S}' = \mathbf{C}\mathbf{S}$ with $\mathbf{C}' = \mathbf{C}\mathbf{P}^T$, $\mathbf{S}' = \mathbf{P}\mathbf{S}$. Therefore, the components can be recovered only upto an arbitrary permutation. The familiar scaling ambiguity present in ICA is absent here because we assume knowledge of $\mathcal{A}_M$, ie. we know the range of the inputs.

The most general problem formulation includes the presence of measurement error $\mathbf{e}$, so that Eq. (1) becomes

$$\mathbf{x}(k) = \sum_{i=1}^{n} \mathbf{c}_i \bar{s}_i(k) + \bar{\mathbf{b}} + \mathbf{e}(k). \tag{4}$$

However, we shall not treat this problem here, postponing its discussion for a future work. In the sequel we shall assume that the error component is zero.

3. ANALYSIS

Definition 1 *Define $col(\mathbf{Z})$ to be the set of columns of the matrix $\mathbf{Z}$.*

Definition 2 (Difference matrix) *For any matrix $\mathbf{Z} \in \mathbb{R}^{h \times w}$ we define the* Difference Matrix $\mathbf{D}_Z \in \mathbb{R}^{h \times w(w-1)}$ *as the matrix consisting of the differences $(\mathbf{z}_p - \mathbf{z}_q)$ for all pairs $\mathbf{z}_p$, $\mathbf{z}_q \in col(\mathbf{Z})$ with $p \neq q$.*

Let $\mathbf{A} \in \mathbb{R}^{n \times M^n}$ denote the matrix whose columns are all the elements of $\mathcal{A}_M^n$ appearing exactly once (i.e. contains all n-tuples of symbols taken from the alphabet $\mathcal{A}_M$):

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots & M-1 \\ 0 & 1 & \cdots & M-1 & 0 & \cdots & M_1 & \cdots & M-1 \\ \vdots & \vdots & & & & & & & \vdots \\ 0 & 0 & \cdots & M-1 & 0 & \cdots & M-1 & \cdots & M-1 \end{bmatrix} \tag{5}$$

Clearly $col(\mathbf{S}) \subseteq col(\mathbf{A})$. The difference matrix $\mathbf{D}_A$ has $M^n(M^n - 1)$ columns with many repetitions. We can show the following:

Theorem 1 *Each one of the vectors $\mathbf{e}_1 = [1, 0, \cdots, 0]^T$, $-\mathbf{e}_1$, $\mathbf{e}_2 = [0, 1, \cdots, 0]^T$, $-\mathbf{e}_2$, ..., $\mathbf{e}_n = [0, 0, \cdots, 1]^T$, $-\mathbf{e}_n$, appear exactly $(M-1)M^{n-1}$ times in the matrix $\mathbf{D}_A$. Moreover, these are the most frequently repeated columns in $\mathbf{D}_A$.*

PROOF. For the first part of the theorem, we shall focus, without loss of generality, on $\mathbf{e}_1$ as the proof for the other vectors is entirely analogous. Let $\mathbf{d} = \mathbf{a}_p - \mathbf{a}_q = \mathbf{e}_1$ for some $\mathbf{a}_p, \mathbf{a}_q \in col(\mathbf{A})$ and $N(d)$ denote the number of combinations by which we can achieve $a - b = d$ with $a, b \in \mathcal{A}_M$. We can achieve $d_1 = 1$ in $N(1) = (M-1)$ different ways, namely $d_1 = 1 - 0$ or $2 - 1$, ..., or $(M-1) - (M-2)$. We also have $N(0) = M$ so the rest of the vector $\mathbf{d}_{2:n} = \mathbf{a}_{i,2:n} - \mathbf{a}_{j,2:n}$ can be zero in $N(0)^{n-1} = M^{n-1}$ ways. Let $N(\mathbf{e}_1)$ denote the times the vector $[1, 0, \cdots, 0]^T$ appears in $\mathbf{D}_A$. Then $N(\mathbf{e}_1) = N(1)N(0)^{n-1} = (M-1)M^{n-1}$. Similarly, $N(\mathbf{e}_i) = N(1)N(0)^{n-1} = (M-1)M^{n-1}$, for all $i = 1, \cdots, n$.

For the second part of the theorem, consider any vector $\mathbf{d} = \mathbf{a}_p - \mathbf{a}_q$, $p \neq q$, such that $\mathbf{d} \notin \{\pm\mathbf{e}_1, \cdots, \pm\mathbf{e}_n\}$. It is not difficult to see that there always exists a vector $\mathbf{e}_p$, for some $p \in \{1, \cdots, n\}$, such that $|d_i| \geq |e_{p,i}|$ for all $i = 1, \cdots, n$, and furthermore, $|d_j| > |e_{p,j}|$ for at least one $j \in \{1, \cdots, n\}$. From straightforward computations we find that $N(d) < N(e)$ if $|d| > |e|$ and $N(d) = N(-d) = N(|d|)$. So

$$N(\mathbf{d}) = \prod_{i=1}^{n} N(|d_i|) < \prod_{i=1}^{n} N(|e_{p,i}|) = N(\mathbf{e}_p).$$

■

Similarly to $\mathbf{A}$ we may define $\mathbf{H} = \mathbf{C}\mathbf{A} \in \mathbb{R}^{m \times M^n}$, so $\mathbf{x} \in col(\mathbf{X}) \Leftrightarrow \mathbf{x} = \mathbf{h} + \mathbf{b}$, $\mathbf{h} \in col(\mathbf{H})$. Considering the difference matrix $\mathbf{D}_H$ we note that $\mathbf{D}_H = \mathbf{C}\mathbf{D}_A$.

If for any pair $\mathbf{d}_p, \mathbf{d}_q \in col(\mathbf{D}_A)$ we have

$$\mathbf{d}_p \neq \mathbf{d}_q \Rightarrow \mathbf{C}\mathbf{d}_p \neq \mathbf{C}\mathbf{d}_q \tag{6}$$

then the next result follows directly from Theorem 1:

Corollary 1 *If condition* (6) *holds, then each one of the vectors $\mathbf{c}_1$, $-\mathbf{c}_1$, $\mathbf{c}_2$, $-\mathbf{c}_2$, ..., $\mathbf{c}_n$, $-\mathbf{c}_n$ appear exactly $(M-1)M^{n-1}$ times in the matrix $\mathbf{D}_H$. Moreover, these are the most frequently repeated vectors in $\mathbf{D}_H$.*

PROOF. Condition (6) implies that there is a one-to-one correspondence between the columns of $\mathbf{D}_H$ and the columns of $\mathbf{D}_A$. Therefore, each vector $\pm\mathbf{c}_i = \pm\mathbf{C}\mathbf{e}_i$ appears $(M-1)M^{n-1}$ times in $\mathbf{D}_H$ while all other vectors appear less frequently. ■

Therefore, the most frequent columns of $\mathbf{D}_H$ are the columns of the "basis matrix" $\mathbf{C}$ together with their opposites in arbitrary order.

Note that we made no specific assumptions with respect to the dimension, m, of the observed signal $\mathbf{x}$ or the number of components, n. In general, m can be less than, equal to, or greater than n.

Let $\mathbf{U}$ be the $m \times L$ matrix formed by the unique columns of $\mathbf{X}$, so $\mathbf{X} = \mathbf{U}\mathbf{T}$ with $\mathbf{T} \in \mathbb{R}^{L \times K}$ being the appropriate expansion matrix ($t_{ij} \in \{0, 1\}$ and $\sum_{i=1}^{L} t_{ij} = 1$).

For given M and n, two extreme cases can be easily identified regarding the existence and uniqueness of the decomposition

- if $L > M^n$ then the matrix $\mathbf{X}$ admits no quantized factorization since the unique columns are more than the size of the input alphabet $|\mathcal{A}_M^n|$;

- if $L \leq n$ then $\mathbf{X}$ admits many decompositions. For instance, we have $\mathbf{X} = \mathbf{CS}$ for any matrix $\mathbf{S} \in \mathbb{R}^{n\times K}$ with $s_{ij} \in \mathcal{A}_M$ and $\mathrm{rank}(\mathbf{S}) = n$, and $\mathbf{C} = \mathbf{XS}^+$, where $\mathbf{S}^+$ is the pseudo-inverse of $\mathbf{S}$.

The interesting case is for $n < L \leq M^n$. For $L = M^n$ we have the following result:

Theorem 2 (Case $L = M^n$) *Consider the matrices $\mathbf{X}$ and $\mathbf{U}$ defined above. For a given number of levels, M, and number of components, n, if the following statements are true:*

- $L = M^n$
- *There exist n distinct, non-zero vectors $\mathbf{c}_1, \cdots, \mathbf{c}_n \in \mathbf{D}_U$ such that each one of them and each one of their opposites $-\mathbf{c}_1 \cdots, -\mathbf{c}_n$ are repeated $(M-1)M^{n-1}$ times in $\mathbf{D}_U$. Define*
$$\mathbf{C} = [\mathbf{c}_1, \cdots, \mathbf{c}_n] \tag{7}$$
- $\mathbf{C}$ *satisfies condition* (6)
- *The sets $S_0^i = \{\mathbf{x} \in col(\mathbf{U}) \mid \mathbf{x} - \mathbf{c}_i \notin col(\mathbf{U})\}$, $S_j^i = \{\mathbf{x} \in col(\mathbf{U}) \mid \mathbf{x} - \mathbf{c}_i \in S_{j-1}^i\}$, $j = 1, \cdots, M-1$, have cardinality M^n for all $i = 1, \cdots, n$*

then there exists a decomposition in the form of Eq. (3) *with $\mathbf{C}$ defined above and with $\mathbf{S} \in \mathbb{R}^{n\times K}$, $s_{ij} \in \mathcal{A}_M$.*

PROOF. For any given $i \in \{1, \cdots, n\}$, the sets S_j^i, $j = 0, \cdots, M-1$, are disjoint. Since there are M such sets with cardinality M^{n-1} while the total number of columns of $\mathbf{U}$ is $L = M^n$, it follows that each vector $\mathbf{x} \in col(\mathbf{U})$ must be a member of exactly one of the sets S_0^i, ..., S_{M-1}^i, for each i. If $\mathbf{x} \in S_{j_1}^1, \cdots, S_{j_p}^p, \cdots, S_{j_n}^n$ then the vector $[j_1, \cdots, j_p, \cdots, j_n]^T$ will be called the *signature* of $\mathbf{x}$. For any $\mathbf{x}' \in col(\mathbf{U})$ with signature $[j_1, \cdots, j_p \pm 1, \cdots, j_n]^T$ we have $\mathbf{x}' = \mathbf{x} \pm \mathbf{c}_p$. It easy to show that for any $\mathbf{x}$, $\mathbf{x}'$ with signature vectors $\mathbf{j}$, $\mathbf{j}'$, we have $\mathbf{x}' = \mathbf{x} + \sum_{i=1}^n (j_i' - j_i)\mathbf{c}_i$, or

$$\mathbf{x}' = \mathbf{x} + \mathbf{C} \cdot (\mathbf{j}' - \mathbf{j}) \tag{8}$$

Also there are not two distinct $\mathbf{x}, \mathbf{x}' \in col(\mathbf{U})$ with the same signature vector $\mathbf{j}$. Otherwise, for each $\boldsymbol{\ell} = [\ell_1, \cdots, \ell_n]^T \in \{0, \cdots, M-1\}^n$ we would have $\mathbf{y} = \mathbf{x} + \mathbf{C}(\boldsymbol{\ell} - \mathbf{j}) \neq \mathbf{x}' + \mathbf{C}(\boldsymbol{\ell} - \mathbf{j}) = \mathbf{y}'$. Then there would exist $2 \cdot M^n$ many distinct columns in $\mathbf{U}$, in contradiction to our assumption.

Let $\mathbf{x}_0$ the column with signature $[0, \cdots, 0]^T$. Then for the k-th column $\mathbf{x}(k) \in col(\mathbf{U})$ with signature $\mathbf{j}(k)$ we have $\mathbf{x}(k) = \mathbf{C}\mathbf{j}(k) + \mathbf{x}_0$. Defining the matrix $\mathbf{J} = [\,\mathbf{j}(1), \cdots, \mathbf{j}(M^n)\,]$ we have

$$\mathbf{U} = \begin{bmatrix}\mathbf{C}\ \mathbf{x}_0\end{bmatrix} \begin{bmatrix}\mathbf{J} \\ \mathbf{u}_L^T\end{bmatrix}$$

with $\mathbf{u}_L = [1, \cdots, 1]^T \in \mathbb{R}^L$. Right-multiplying by $\mathbf{T}$ we obtain

$$\mathbf{X} = \begin{bmatrix}\mathbf{C}\ \mathbf{x}_0\end{bmatrix} \begin{bmatrix}\mathbf{S} \\ \mathbf{u}_K^T\end{bmatrix}, \tag{9}$$

$$\mathbf{S} = \mathbf{JT}. \tag{10}$$

■

According the proof of theorem 2 the signature of a vector $\mathbf{x}$ yields the component vector $\mathbf{s}$ corresponding to $\mathbf{x}$. Based on the assumptions of Theorem 2 we shall develop a practical and efficient algorithm for quantization component analysis, taking advantage of the recursive relationship described by Eq. (8). Starting from an arbitrary point $\mathbf{x}$ with signature $\mathbf{j}$, any point $\mathbf{x}'$ for which $\mathbf{x}' = \mathbf{x} + k\mathbf{c}_i$, will have signature equal to $\mathbf{j}' = \mathbf{j} + k\mathbf{e}_i$, where $\mathbf{e}_i$ is the i-th column of the $n \times n$ identity matrix.

Algorithm 1 (Quantized Component Extraction)

1. *Identify the unique columns of $\mathbf{X}$ to form a matrix $\mathbf{U}$, and save the mapping $\mathbf{T}$, from the columns of $\mathbf{U}$ to the columns of $\mathbf{X}$.*
2. *Construct the difference matrix $\mathbf{D}_U$ by taking pairwise differences of all columns of $\mathbf{U}$. Take $\pm\mathbf{c}_1$, ..., $\pm\mathbf{c}_n$, to be the vectors which appear $(M-1)M^{n-1}$ times in $\mathbf{D}_U$. Select the signs arbitrarily.*
3. *Mark all columns of $\mathbf{U}$ as* "not visited" *and* "not done".
4. *Select an arbitrary vector $\mathbf{x} \in col(\mathbf{U})$. Set its signature to $\mathbf{j} = \mathbf{0}$ and mark it as* "visited" *and* "done".
5. *For all $i = 1, \cdots, n$, find the vectors $\mathbf{x}' \in col(\mathbf{U})$ for which $\mathbf{x}' = \mathbf{x} + k\mathbf{c}_i$, for some $k \in \{-(M-1), \cdots, M-1\}$ and set their signatures equal to $\mathbf{j}' = \mathbf{j} + k\mathbf{e}_i$. Mark all such vectors $\mathbf{x}'$ as* "done".
6. *While there are columns of $\mathbf{U}$ which are* "not done"*, select an arbitrary $\mathbf{x} \in col(\mathbf{U})$ with signature $\mathbf{j}$ so that it is* "done" *but* "not visited". *Go to Step 5.*
7. *Let $\mu_i = \min_{\mathbf{j}}\{j_i\}$, $i = 1, \cdots, n$. For all signatures $\mathbf{j}$ set $j_i \leftarrow j_i - \mu_i$; Call $\mathbf{x}_0$ the vector with signature $\mathbf{j} = \mathbf{0}$.*
8. *Construct $\mathbf{C}$ and $\mathbf{S}$ according to* (7), (10).

Note that the the arbitrary starting point in step 4 is assigned the signature $\mathbf{j} = \mathbf{0}$ although its true signature might actually be $\mathbf{j}^0 = [j_1^0, \cdots, j_n^0]^T \neq \mathbf{0}$. For this reason, the signatures obtained after Step 6 are shifted by $\mathbf{j}^0$. Thus Step 7 is required to shift the signatures back to the original range $0, \cdots, M-1$.

4. SIMULATION EXAMPLE

We have tested the Algorithm 1 using various test data. Here, for the sake of visualization, we present an example where the observed signal dimensionality is $m = 2$, the number of components is $n = 3$, and the input alphabet is $\mathcal{A}_3 = \{0, 1, 2\}$, for $M = 3$. The basis matrix is

$$\mathbf{C} = \begin{bmatrix} -0.1924 & -0.7648 & -1.4224 \\ 0.8886 & -1.4023 & 0.4882 \end{bmatrix}$$

and the offset $\mathbf{b} = [0\ 0]^T$. Figure 1 shows the subsets S_j^i corresponding to the basis vectors $\mathbf{c}_i$ and levels $j = 0, 1, 2$. The basis vectors $\mathbf{c}_1$, $\mathbf{c}_2$, $\mathbf{c}_3$, are estimated by clustering the pairwise differences $\mathbf{x} - \mathbf{x}'$ for all $\mathbf{x}, \mathbf{x}' \in \mathbf{U}$. Figure 2 shows how the computation of the signature spreads from a starting point $\mathbf{x}_0$ according to rule (8). Then a random point among $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ is chosen to continue the recursion, and so on. The recursion terminated in 27 iterations. The final estimated basis matrix was a perfect estimation of $\mathbf{C}$ up to a permutation of the columns:

$$\hat{\mathbf{C}} = \begin{bmatrix} -0.7648 & 0.1924 & -1.4224 \\ -1.4023 & -0.8886 & 0.4882 \end{bmatrix}.$$

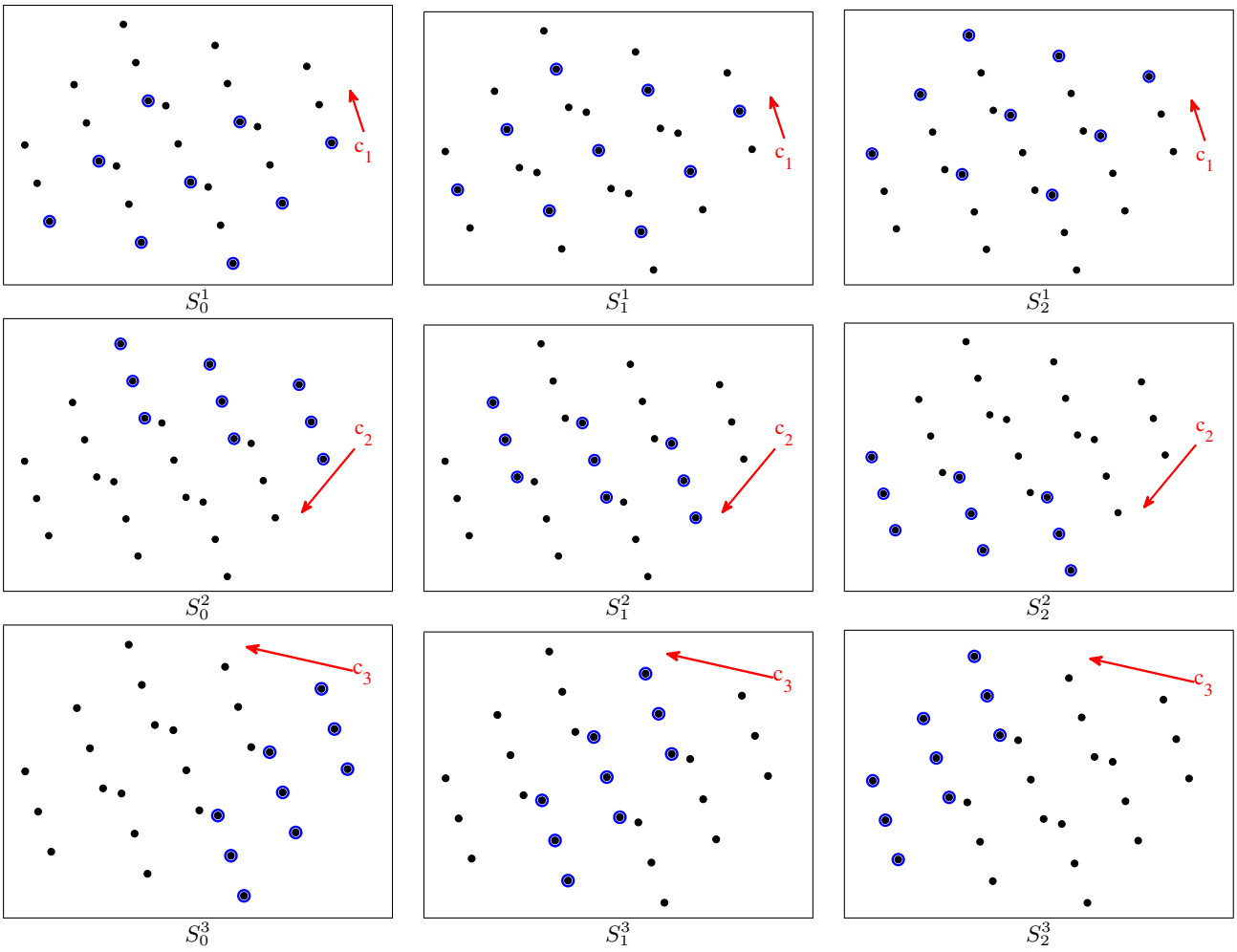

Fig. 1. The subsets S_j^i for a case with $m = 2$, $n = 3$, $M = 3$ and $\mathcal{A}_3 = \{0, 1, 2\}$. Each subplot shows the constellation of the vectors $\mathbf{x} \in col(\mathbf{X})$ and the points belonging to the corresponding subset S_j^i are marked with circles. Each subset S_j^i is a shifted version of the subset S_{j-1}^i by $\mathbf{c}_i$.

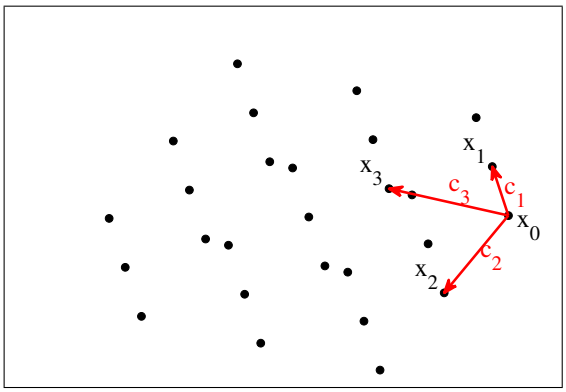

Fig. 2. Given the signature $[j_1, j_2, j_3]$ of vector $\mathbf{x}_0$ the signatures of $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ are $[j_1+1, j_2, j_3]$, $[j_1, j_2+1, j_3]$, and $[j_1, j_2, j_3+1]$, respectively.

5. CONCLUSIONS

We have treated the decomposition of signals into n quantized components as a matrix factorization problem. First it is shown that if the input alphabet consists of M equally spaced levels then the problem can be transformed, without loss of generality, into an equivalent one with alphabet normalized between 0 and $M-1$. Next, we show that under certain conditions and provided that enough observation samples are collected, the problem admits a solution unique up to permutation. For this case, we propose a deterministic iterative algorithm which estimates the basis (mixing) vectors from the pairwise differences of the unique observations and then proceeds to perfectly reconstruct the components in a recursive fashion.

6. REFERENCES

[1] J. Proakis, *Digital Communications*, 2000.

[2] S. Kolenikov and G. Angeles, "Socioeconomic status measurement with discrete proxy variables: Is principal component

analysis a reliable answer?," *Review of Income and Wealth*, vol. 55, no. 1, pp. 128–165, 2009.

[3] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[4] I. Jolliffe, *Principal Component Analysis*, John Wiley, 2nd edition, 2002.

[5] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[6] P. Comon, "Independent Component Analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, April 1994.

[7] A. Hyvärinen, E. Oja, and J. Karhunen, *Independent Component Analysis*, John Wiley, 2001.

[8] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2001, number 13, pp. 556–562.

[9] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[10] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, pp. 21–30, March 2008.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[12] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 1999, pp. 50–57, ACM.

[13] J. Canny, "Gap: a factor model for discrete data," in *in Proc. SIGIR 2004*, 2004, pp. 122–129.

[14] W. Buntine, "Variational extensions to em and multinomial pca," in *ECML, LNAI*. 2002, pp. 23–34, Springer.

[15] W. Buntine and A. Jakulin, "Discrete principal component analysis," in *Proceedings of the Subspace, Latent Structure and Feature Selection Techniques: Statistical and Optimisation perspectives*, Slovenia, Feb. 2005.

[16] H. W. Gutch, P. Gruber, and F. J. Theis, "ICA over Finite Fields," in *Latent Variable Analysis and Signal Separation*, V. Vigneron, V. Zarzoso, E. Moreau, R. Gribonval, and E. Vincent, Eds., LNCS 6365, pp. 645–652. Springer, 2010.

[17] H. W. Gutch, P. Gruber, A. Yeredor, and F. J. Theis, "ICA over finite fields - Separability and algorithms," *Signal Processing*, vol. 92, pp. 1796–1808, August 2012.

[18] K. I. Diamantaras and Th. Papadimitriou, "Histogram based blind identification and source separation from linear instantaneous mixtures," *Lecture Notes on Computer Sience (LNCS)*, vol. 5441, pp. 227–234, 2009.