Decorrelation of Neutral Vector Variables: Theory and Applications

Zhanyu Ma, *Member, IEEE*, Jing-Hao Xue, Arne Leijon, Zheng-Hua Tan, *Senior Member, IEEE*, Zhen Yang, *Member, IEEE*, and Jun Guo

***Abstract*—In this paper, we propose novel strategies for neutral vector variable decorrelation. Two fundamental invertible transformations, namely serial nonlinear transformation and parallel nonlinear transformation, are proposed to carry out the decorrelation. For a neutral vector variable, which is not multivariate Gaussian distributed, the conventional principal component analysis (PCA) cannot yield mutually independent scalar variables. With the two proposed transformations, a highly negatively correlated neutral vector can be transformed to a set of mutually independent scalar variables with the same degrees of freedom. We also evaluate the decorrelation performances for the vectors generated from a single Dirichlet distribution and a mixture of Dirichlet distributions. The mutual independence is verified with the distance correlation measurement. The advantages of the proposed decorrelation strategies are intensively studied and demonstrated with synthesized data and practical application evaluations.**

***Index Terms*—Neutral vector, neutrality, non-Gaussian, decorrelation, Dirichlet variable**

I. Introduction

In many pattern recognition and machine learning areas, Gaussian distributions, among other probability distributions, have been ubiquitously applied to describe data distribution, with the assumption that these data are Gaussian distributed [1]. However, in many applications the distribution of data is asymmetric or constrained [2]. For example, the pixel values in a color or grey image [3], [4], the ratings assigned to an item in collaborative filtering [5]–[7], and the epigenetic mark values in epigenome-wide-association studies [8], [9] have strictly bounded support (*e.g.*, $x \in [0, c]$). In speech enhancement, the spectrum coefficients [10], [11] are semi-bounded (*i.e.*, $x \in (0, +\infty)$). The l_2 norms of the spatial fading correlation [12] and the yeast gene expressions [13] are equal to 1 and such data convey directional property (*i.e.*, $\|\mathbf{x}\|_2 = 1$). A common property of the aforementioned data is that, these data have *not only* a specific support range, *but also* a non-bell distribution shape. Apparently, these properties do not match the natural properties of a Gaussian distribution (*i.e.*, the definition domain is unbounded and the distribution shape is symmetric). Therefore, such data are non-Gaussian distributed [14]. It has been demonstrated in many recent studies that explicitly utilizing the non-Gaussian characteristics can significantly improve the performance in practice [3], [4], [8]–[20].

One typical type of non-Gaussian distributed data, among others, is the one that represents proportions. In the frequently used mixture modeling technique [3], [21], [22], the weighting factors denote the proportions of each mixture component in the whole mixture model. In the text mining area, the Dirichlet distribution is used to model topic relations, *i.e.*, the proportions with which a specific topic appears in the total set of documents [23]–[25]. For analyzing color images, the normalized RGB space, which is often used as pure color space by discarding the illuminance [26]–[29], represents the proportions of RGB channels in the whole color space. In time series signal processing [30], [31], the difference between two adjacent line spectral frequencies (LSFs) conveys the proportion of frequency distance (in angle) to half of the unit circle's circumference. The LSFs are less sensitive to quantization noise than other representations and are widely used in speech coding [30], [32], [33]. Also, the parameters in the multinomial distribution [34], [35] represent probabilities for each particular event to happen in the trial sequence. In [36], a novel online kernel learning algorithm, called QKRLS, was developed, which is computationally efficient and can be used for online regression and classification.

Data representing proportions can be denoted by a $K+1$ dimensional vector $\mathbf{x} = [x_1, \ldots, x_K, x_{K+1}]^{\mathrm{T}}$ with K degrees of freedom. Each element x_k is nonnegative and the sum of all the elements in $\mathbf{x}$ is a constant (usually can be normalized to 1). Connor et al. [37] introduced the concept "neutrality" to investigate a particular type of independence for the elements in $\mathbf{x}$. Even though the resulting neutral vector represents a particular type of independence after a substraction-normalization operation [38], the elements in the neutral vector are mutually highly correlated, or rather, negatively correlated. Intuitively, if one proportion increases, then the remaining proportions would decrease correspondingly, since the summation of all the proportions is a constant.

For correlated random vector variables, principal component analysis (PCA) is a popular technique used for applications such as data decorrelation, dimension reduction, lossy data compression, and feature extraction [21], [39], [40]. It is also known as Karhunen-Loève transform (KLT) in transform coding [41], [42]. It can be considered as an orthogonal transformation of the correlated variables into a set of un-

Z. Ma and J. Guo are with Pattern Recognition and Intelligent System Lab., Beijing University of Posts and Telecommunications, Beijing, China.

J.-H. Xue is with the Department of Statistical Science, University College London, London, United Kingdom.

A. Leijon is with the School of Electrical Engineering, KTH - Royal Institute of Technology, Stockholm, Sweden.

Z.-H. Tan is with the Department of Electronic Systems, Aalborg University, Aalborg, Denmark.

Z. Yang is with the College of Computer Science, Beijing University of Technology, Beijing, China.

The corresponding author is Z. Ma. Email: mazhanyu@bupt.edu.cn

correlated scalar variables, which are named as principal components. This transformation is linear and invertible. PCA is the optimal decorrelation strategy for multivariate-Gaussian distributed data [21]. For data from a multivariate-Gaussian distribution, the resulting transformed scalar variables are not only mutually uncorrelated but also mutually independent. For data from other sources, PCA can only guarantee that the scalar variables are mutually uncorrelated.

Independent component analysis (ICA) is a computational method applied to separate a multivariate vector variable into a set of additive and mutually independent scalar variables (sources) [43], [44]. With the assumption that the source signals are independent of each other and the source signals are non-Gaussian distributed, ICA attempts to decorrelate a multivariate vector variable into mutually independent non-Gaussian scalar variables. ICA can be applied to several fields such as face recognition [45], blind source separation [46], and wireless communications [47].

A neutral vector has a bounded support (in $[0, c]$) and is negatively correlated (the off-diagonal elements in the covariance matrix is negative). Thus, it cannot follow a multivariate-Gaussian distribution. In this case, applying PCA to neutral vector can only yield mutually uncorrelated but not mutually independent scalar variables. Although ICA can yield mutually independent non-Gaussian scalar variables, it cannot preserve the bounded support property. By considering the neutrality, the highly correlated variables in a neutral vector can be decorrelated into a set of independent variables with nonlinear transformation. Moreover, such procedure does not depend on the eigenvalue decomposition of the covariance matrix.

In this paper, we propose two fundamental transformation strategies, namely the serial nonlinear transformation (SNT) and the parallel nonlinear transformation (PNT), to decorrelate neutral vectors. These invertible nonlinear transformations take the advantages of the completely neutrality. We prove that the above mentioned nonlinear transformations can decorrelate the neutral vector variable into a set of mutually independent variables. Particularly, if the neutral vector variable is Dirichlet distributed, each of the transformed variables follows the beta distribution, which is actually a special case of the Dirichlet distribution with two parameters.

Although nonlinear kernel functions can be introduced to carry out kernel PCA [48], [49] or kernel ICA [50], [51] such that the vector variable decorrelation can be implemented in a nonlinear manner, the proposed nonlinear transformation strategies are different from these ones. In kernel PCA, input vectors are firstly mapped into a feature space via a kernel function, and then the standard PCA is applied to conduct the decorrelation [21, Ch. 12.3]. Similar approaches are applied to kernel ICA. Therefore, kernel PCA and kernel ICA each contain two stages, which are nonlinear kernel mapping and linear decorrelation (in the feature space). In contrast to this, the proposed nonlinear transformation strategies (*i.e.*, SNT and PNT) do not require kernel mapping. It is a one-stage nonlinear operation in the decorrelation implementation.

For a neutral random vector, the decorrelation strategies are based on each observed vector *only* and does not require any statistical information (*e.g.*, the covariance matrix) of the whole observation set. In other words, the decorrelation strategies are model independent. Therefore, the proposed decorrelation strategies reduce the computational complexity, compared with PCA which requires eigenvalue decomposition of the covariance matrix. ICA has even higher computational costs than PCA. The decorrelation of a vector variable is important and very helpful in many applications (*e.g.*, source coding, dimension reduction, and feature selection [52], [53]). Hence, the proposed decorrelation strategies are novel and useful for the data with neutrality.

Algorithm 1 Serial Nonlinear Transformation

Input: Neutral vector $\mathbf{x} = [x_1, \ldots, x_K, x_{K+1}]^{\mathrm{T}}$
Set $\mathbf{x}_1 = \mathbf{x}$, $i = 1$
repeat
  Assign the value of the 1st element of $\mathbf{x}_i$ to u_i;
  $i = i + 1$, $\mathbf{x}_i = \mathbf{x}_{i-1}$, with the first element in $\mathbf{x}_{i-1}$ removed;
  Normalize the remaining elements in $\mathbf{x}_i$ as $\mathbf{x}_i = \mathbf{x}_i / \|\mathbf{x}_i\|_1$
until $i == K$
Output: Transformed vector $\mathbf{u} = [u_1, \ldots, u_K]^{\mathrm{T}}$.

The rest of this paper is organized as follows: we review the concept of neutrality in Sec. II. The proposed transformation strategies are introduced in Sec. III where the proof of mutually independence is also provided. In Sec. IV, we take the Dirichlet distribution as an example for neutral vectors. Comprehensive evaluations of the proposed strategies with synthesized and real data are presented in Sec. V. We draw some conclusions in Sec. VI.

II. Neutral Vector Variable

Assuming we have a random vector variable $\mathbf{x} = [x_1, x_2, \ldots, x_K, x_{K+1}]^{\mathrm{T}}$, where $x_k > 0$ and $\sum_{k=1}^{K+1} x_k = 1$. Let $\mathbf{x}_{k1} = [x_1, \ldots, x_k]^{\mathrm{T}}$ and $\mathbf{x}_{k2} = [x_{k+1}, \ldots, x_{K+1}]^{\mathrm{T}}$. The vector $\mathbf{x}_{k1}$ is neutral if $\mathbf{x}_{k1}$ is independent of $\mathbf{w}_k = \frac{1}{1-s_k}\mathbf{x}_{k2}$, for $1 \leq k \leq K$ [37], [54], where $s_k = \sum_{i=1}^{k} x_i$ and $s_0 = 0$. If for all k, $\mathbf{x}_{k1}$ are neutral, then $\mathbf{x}$ is defined as a *completely neutral* vector [37], [55]. A neutral vector with $(K + 1)$ elements has K degrees of freedom.

The idea of neutrality was introduced by Connor et al. [37] for describing constrained variables with the property mentioned above. It was originally developed for biological applications. According to the above definition, the neutral vector conveys a particular type of independence among its elements, even though the element variables themselves are mutually negatively correlated. A complete neutral vector variable has a set of properties, we list those will be used here:

***Property** 2.1 (Mutually Independence):* For completely neutral vector $\mathbf{x}$, define $z_k = \frac{x_k}{1-s_{k-1}}$ and $z_1 = x_1$, we have $z_1, z_2, \ldots, z_K$ are mutually independent.

***Property** 2.2 (Aggregation Property):* Mutually Independence For a completely neutral vector $\mathbf{x}$, when adding any adjacent elements x_r and x_{r+1} together, the resulting K-dimensional vector $\mathbf{x}^{r \uplus r+1} = [x_1, \ldots, x_r + x_{r+1}, \ldots, x_{K+1}]$ is a completely neutral vector again.

The proofs of the above properties can be found in Appendix A and B.

Usually, the dimensions in a completely neutral vector should be equally treated. In other words, the positions of the dimensions do not affect the properties of the vector. In

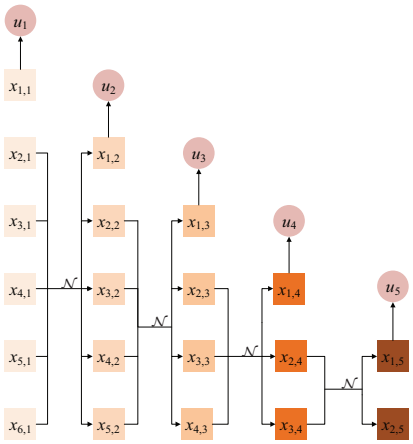


Fig. 1. An example of SNT with $K = 5$. The transformed variables are $u_1 = x_{1,1}$, $u_2 = x_{1,2}$, $u_3 = x_{1,3}$, $u_4 = x_{1,4}$ and $u_5 = x_{1,5}$. $x_{i,j}$ denotes the i^{th} element in $\mathbf{x}_j$. $\mathcal{N}$ is the l_1-norm normalization.

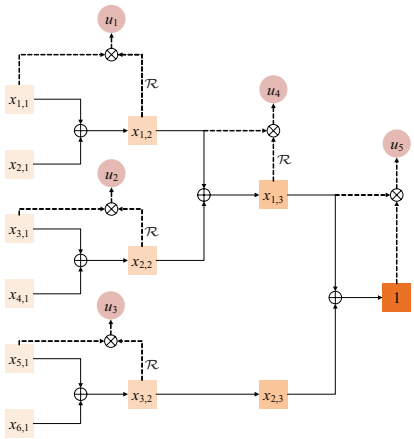


Fig. 2. An example of PNT with $K = 5$. The transformed coefficients are $u_1 = x_{1,1}/x_{1,2}$, $u_2 = x_{3,1}/x_{2,2}$, $u_3 = x_{5,1}/x_{3,2}$, $u_4 = x_{1,2}/x_{1,3}$ and $u_5 = x_{1,3}$. $\mathcal{R}$ represents the reciprocal operation.

order to explicitly convey this fact, we make the following definition:

Definition For a *completely neutral* vector $\mathbf{x}$, if *arbitrarily* permuted version of $\mathbf{x}$ is still *completely neutral*, then this vector is *exchangeably completely neutral*.

In the field of statistical analysis, a typical variable which has the above mentioned properties is the Dirichlet variable. For the Bayesian analysis of mixture models [21], [22], the weighting factors of the mixture components are usually modelled by a Dirichlet distribution. Recently, the Dirichlet process (*e.g.*, [56], [57]) was applied for nonparametric Bayesian analysis. If we represent the Dirichlet process with the so-called stick-breaking process [56], the independence among different generating steps can be expressed explicitly as a neutral vector with infinite dimensionality. The Dirichlet process is the cornerstone of non-parametric Bayesian analysis and applied to a variety of practical signal and feature analysis problems. Thus, the concept of neutral vectors is very useful in many signal processing, pattern recognition, and other practical applications.

III. Transformations for Neutral Vectors

In most signal processing applications, the transformations we use are linear or linear according to some nonlinear kernel functions. Even though we could apply PCA directly to the neutral random vector variable, this linear transformation could only decorrelate the data, but cannot guarantee the independence if the data are not Gaussian. Furthermore, PCA does not exploit the neutrality [14]. In this case, PCA is not optimal for decorrelating neutral vectors. By considering the exchangeably complete neutrality, we propose two nonlinear invertible transformations, namely the serial nonlinear transformation (SNT) and the parallel nonlinear transformation (PNT). Each of the proposed nonlinear transformations can decorrelate the vector variable into a set of mutually independent variables. In contrast to PCA, the transformations do not require any statistical information (*e.g.*, the covariance matrix) of the observed vector set. Thus, it avoids the eigenvalue analysis for PCA and, therefore, the computational complexity is reduced.

A. Serial Nonlinear Transformation

For an exchangeable completely neutral vector variable with K degrees of freedom, if we process the vector variable with the SNT strategy described in Algorithm 1, the $K+1$ dimensional vector variable $\mathbf{x}$ is transformed to a vector $\mathbf{u}$ with K variables. These K scalar variables are mutually independent.

The proof of mutually independence of $\mathbf{u}$ is straightforward. In the transformed vector $\mathbf{u}$, the first element $u_1 = x_1 = z_1$. The second element of $\mathbf{u}$ is

$$u_2 = \frac{x_2}{1 - s_1} = z_2. \tag{1}$$

Similarly, we have

$$u_3 = \frac{\frac{x_3}{1-s_1}}{1 - \frac{x_2}{1-s_1}} = \frac{x_3}{1 - s_2} = z_3. \tag{2}$$

More generally, we can obtain that

$$u_k = \frac{\frac{x_k}{1-s_{k-2}}}{1 - \frac{x_{k-1}}{1-s_{k-2}}} = \frac{x_k}{1 - s_{k-1}} = z_k, \quad 3 \le k \le K. \tag{3}$$

According to Property 2.1, Eq. 3 shows that the K variables in $\mathbf{u}$ are mutually independent. Since the SNT is invertible and $\mathbf{x}$ has K degrees of freedom, $\mathbf{u}$ contains the same amount of information as $\mathbf{x}$. An example of the SNT with $K = 5$ is illustrated in Fig. 1.

B. Parallel Nonlinear Transformation

The SNT algorithm needs K rounds of iterations to finalize the transformation. In order to facilitate the operation, we can also carry out the nonlinear transformation in a parallel way, which is referred as the parallel nonlinear transformation (PNT). The PNT scheme is introduced in Alg. 2. At each iteration, the dimension of the processed vector is reduced by half. Finally, we still get a vector variable $\mathbf{u}$ with K mutually independent element variables. An example of PNT with $K = 5$ is shown in Fig. 2. The proofs of independence can be found in Appendix C.

C. Fast Parallel Nonlinear Transformation

According to Alg. 2, the implementation of PNT needs to check if the length of $\mathbf{x}_i$ in each iteration is even or odd.

Algorithm 2 Parallel Nonlinear Transformation [32]

Step 1. Initialization
Set $\mathbf{x}_1 = \mathbf{x}$, $i = 2$
Step 2. Aggregation
$L = \text{length}(\mathbf{x}_{i-1}) - 1$
if L is even **then**
  for $l = 1, l \leq L/2, l++$ **do**
    $x_{l,i} = x_{2l-1,i-1} + x_{2l,i-1}$
    $u_{l,i-1} = \frac{x_{2l-1,i-1}}{x_{l,i}}$
  end for
  $\mathbf{x}_i = [x_{1,i}, \ldots, x_{l,i}, x_{L+1,i-1}]^T$
  $\mathbf{u}_{i-1} = [u_{1,i-1}, \ldots, u_{l,i-1}]^T$
else
  for $l = 1, l < (L+1)/2, l++$ **do**
    $x_{l,i} = x_{2l-1,i-1} + x_{2l,i-1}$
    $u_{l,i-1} = \frac{x_{2l-1,i-1}}{x_{l,i}}$
  end for
  $\mathbf{x}_i = [x_{1,i}, \ldots, x_{l,i}]^T$
  $\mathbf{u}_{i-1} = [u_{1,i-1}, \ldots, u_{l,i-1}]^T$
end if
Step 3. Stop criterion
if $\text{length}(\mathbf{x}_i) == 2$ **then**
  $\mathbf{u}_i = x_{1,i}$, go to step 4
else
  $i = i + 1$, go to step 2.
end if
Step 4. Return the transformed coefficients $\mathbf{u} = [\mathbf{u}_1^T, \ldots, \mathbf{u}_i^T]^T$.

This is due to the fact that the number of elements in $\mathbf{x}$ is not always equal to power of 2. Inspired by the fast Fourier transform [58], we design a fast PNT (FPNT) algorithm to facilitate the practical computation with zero-padding. Zero-padding is a technique usually employed to make the length of a vector equal to a power of 2, by adding zeros to the end of the vector so that the total number of elements equals the next higher power of 2. The vector $\mathbf{x}$ is expanded with zero-padding to the next higher power of 2. During each iteration in the transformation, the vector length reduces to half, until the length of the vector reduces to two. This algorithm skips the check of parity, and, therefore, the practical computational time is reduced. It is convenient to implement in practice. It is worthy to note that this FPNT algorithm has similar computational complexity to the PNT flow chart shown in Alg. 2. The FPNT algorithm is introduced in Algorithm 3.

IV. Dirichlet Variable: An Example

In the above nonlinear transformations, we did not assign any explicit distribution to the neutral vector variable. Indeed, the transformation itself does not require us to know the specific distribution of the vector variable, with the assumption that the vector variable is exchangeably completely neutral. In this section, we will take the Dirichlet variable as an intuitive example. It has been showed in [54] that the Dirichlet distribution is characterized by neutrality and a vector drawn from a Dirichlet distribution is *completely* neutral. Moreover, any permutation of such vector (which is generated from a Dirichelt distribution) is also a *completely* neutral vector (*i.e.*, exchangeably completely neutral). Note that, a *completely* neutral vector *may not* have such permutation property [37].

The Dirichlet density function is defined as

$$\mathbf{Dir}(\mathbf{x};\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K+1}\alpha_k)}{\prod_{k=1}^{K+1}\Gamma(\alpha_k)} \prod_{k=1}^{K+1} x_k^{\alpha_k - 1}, x_k \geq 0, \sum_{k=1}^{K+1} x_k = 1, \alpha_k > 0. \quad (4)$$

If we take any element x_k from $\mathbf{x}$ and denote the remaining normalized elements as $\mathbf{x}_{\backslash k} = \frac{1}{1-x_k}[x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_{K+1}]^{\mathrm{T}}$, it can be shown that [59]

$$f(x_k, \mathbf{x}_{\backslash k}) = \mathbf{Beta}(x_k; \alpha_k, \sum_{i=1, i\neq k}^{K+1} \alpha_i) \times \mathbf{Dir}(\mathbf{x}_{\backslash k}; \boldsymbol{\alpha}_{\backslash k}), \quad (5)$$

where $\boldsymbol{\alpha}_{\backslash k} = [\alpha_1, \ldots, \alpha_{k-1}, \alpha_{k+1}, \ldots, \alpha_{K+1}]^{\mathrm{T}}$ and

$$\mathbf{Beta}(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \quad (6)$$

is the beta density function, which is exactly a Dirichlet density function with two parameters a and b. Thus a Dirichlet variable $\mathbf{x} = [x_1, \ldots, x_K, x_{K+1}]^{\mathrm{T}}$ is a neutral vector. Furthermore, the Dirichlet variable has the aggregation property as [59]

$$\mathbf{x}_{i+j} \sim \mathbf{Dir}(\mathbf{x}_{i+j}; \boldsymbol{\alpha}_{i+j}), \quad (7)$$

where $\mathbf{x}_{i+j} = [x_1, \ldots, x_i + x_j, \ldots, x_{K+1}]^{\mathrm{T}}$ and $\boldsymbol{\alpha}_{i+j} = [\alpha_1, \ldots, \alpha_i + \alpha_j, \ldots, \alpha_{K+1}]^{\mathrm{T}}$. These properties can be easily shown by the principles of variable substitution.

For the SNT strategy, the transformed variable u_k is beta distributed as

$$u_k \sim \mathbf{Beta}(u_k; \alpha_k, \sum_{i=k+1}^{K+1} \alpha_i), \quad (8)$$

which can be proved by the neutrality and the aggregation properties. For each loop in the PNT algorithm (Algorithm 2), we define a new parameter vector $\boldsymbol{\alpha}_i$ for the ith loop ($i \geq 2$). The update rule for $\boldsymbol{\alpha}_i$ is the same as $\mathbf{x}_i$ and $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}$. In the ith loop, we can obtain a Dirichlet distribution by aggregating the elements $x_{3,i-1}, \ldots, x_{L+1,i-1}$ together as

$$\begin{aligned}&[x_{1,i-1}, x_{2,i-1}, \sum_{l=3}^{L+1} x_{l,i-1}]^{\mathrm{T}} \\ \sim &\mathbf{Dir}(x_{1,i-1}, x_{2,i-1}, \sum_{l=3}^{L+1} x_{l,i-1}; \alpha_{1,i-1}, \alpha_{2,i-1}, \sum_{l=3}^{L+1} \alpha_{l,i-1}).\end{aligned} \quad (9)$$

By considering that $\sum_{l=3}^{L+1} x_{l,i-1}$ is a neutral variable, the normalized version of the remaining two variables $x_{1,i-1}, x_{2,i-1}$ are again Dirichlet distributed with two parameters. This is equivalent to a beta distribution. Thus the obtained coefficient $u_{1,i-1} = x_{1,i-1}/(x_{1,i-1} + x_{2,i-1})$ follows a beta distribution as

$$u_{1,i-1} \sim \mathbf{Beta}(u_{1,i-1}; \alpha_{1,i-1}, \alpha_{2,i-1}). \quad (10)$$

Based on the same reasoning, we can show that $u_{l,i-1}$ is also beta distributed. Thus, with SNT or PNT, the Dirichlet variable can be decorrelated into a vector with the same degrees of freedom. Due to the complete neutrality, the element variables in the transformed vector are mutually independent, and each element variable is beta distributed.

Algorithm 3 Fast Parallel Nonlinear Transformation

Input: Neutral vector $\mathbf{x} = [x_1, \ldots, x_K, x_{K+1}]^{\mathrm{T}}$
Set $T = \lceil \log_2 (K+1) \rceil$ and $P = 2^T - (K+1)$
Set $\mathbf{x}_{\text{zp}} = [\mathbf{x}^{\mathrm{T}}, \mathbf{0}_P^{\mathrm{T}}]^{\mathrm{T}}$ (zero-padding) †
Set $\mathbf{x}_1 = \mathbf{x}_{\text{zp}}$
for $t = 1, t \leq T, t++$ **do**
  $\mathbf{u}_t^{\text{temp}} = \mathbf{x}_t^{\text{odd}} ./ (\mathbf{x}_t^{\text{odd}} + \mathbf{x}_t^{\text{even}})$ ‡
  Set $\mathbf{u}_t$ to be a vector containing only the elements that are not equal to one in $\mathbf{u}_t^{\text{temp}}$
  $\mathbf{x}_{t+1} = \mathbf{x}_t^{\text{odd}} + \mathbf{x}_t^{\text{even}}$
end for
Output: Transformed vector $\mathbf{u} = [\mathbf{u}_1^{\mathrm{T}}, \ldots, \mathbf{u}_T^{\mathrm{T}}]^{\mathrm{T}}$.

† $\mathbf{0}_P$ is a $P \times 1$ vector contains only 0.
‡ $\mathbf{x}_t^{\text{odd}}$ and $\mathbf{x}_t^{\text{even}}$ represent the odd and even elements in $\mathbf{x}_t$, respectively. The operator $./$ denotes element-wise division. Moreover, we define $\frac{0}{0} = 1$.

V. Experimental Results and Discussions

The importance of independence arises in many applications. The proposed nonlinear transformation methods can decorrelate a neutral vector variable into a set of mutually independent scalar variables. In order to illustrate the decorrelation performance, the distance correlation (DC) [60], [61], which measures statistical dependence between two random variables, is applied to evaluate the mutual independence of the scalar variables after transformation. Unlike the commonly used Pearson correlation coefficient [62], [63], the DC is zero if and only if the random variables are statistically mutually independent [64]. Given a set of paired samples $(X_n, Y_n),\ n = 1, \ldots, N$, all pairwise Euclidean distances a_{ij} and b_{ij} are calculated as

$$a_{ij} = \|X_i - X_j\|, \quad b_{ij} = \|Y_i - Y_j\|, \quad i, j = 1, \ldots, N. \tag{11}$$

Taking the doubly centered distances, we have

$$A_{ij} = a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot},\ B_{ij} = b_{ij} - \bar{b}_{i\cdot} - \bar{b}_{\cdot j} + \bar{b}_{\cdot\cdot}, \tag{12}$$

where $\bar{a}_{i\cdot}$ denotes the mean of the ith row, $\bar{a}_{\cdot j}$ is the mean of the jth column, and $\bar{a}_{\cdot\cdot}$ stands for the grand mean of the matrix. The same definitions apply to $\bar{b}_{i\cdot}$, $\bar{b}_{\cdot j}$, and $\bar{b}_{\cdot\cdot}$.

In order to evaluate the statistical significance of the DC, a permutation test is employed. The p-value for the permutation test is calculated as follows:

1) For the original data (X_n, Y_n), create a new data set (X_n, Y_{n^*}), where n^* denotes a permutation of the set $\{1, \ldots, N\}$. The permutation set is selected randomly as drawing without replacement;
2) Calculate a DC for the randomized data
3) Repeat the above two steps a large number of times, the p-value for this permutation test is the proportion of the DC values in step 2 that are larger than the DC from the original data.

The null hypothesis in this case is that the two variables involved are independent of each other (the DC is 0). When the corresponding p-value is smaller than 0.05, the null-hypothesis is rejected so that these two variables are *not* independent (but could still be uncorrelated). Hence, p-value greater than 0.05 indicates independence. We choose the significance level as 0.05 in the remaining parts of this paper.

In this section, we firstly compare PNT/SNT with PCA and ICA, with evaluation of decortication performance. Next, we demonstrate the decorrelation performance of PNT (in terms of mutual independence) with both synthesized and real data. Afterwards, we apply the proposed strategy to real-life applications to improve corresponding practical performance.

A. Comparisons of SNT, PNT, PCA, and ICA

1) Computational Complexity: In practical applications, the computational complexity of decorrelation is usually a concern. We now analyze the computational complexities of SNT and PNT, respectively, and compare them with that of the conventionally used PCA and ICA strategies.

- SNT and PNT
  As described in Algorithm 1, each iteration yields one element in the target vector $\mathbf{u}$. Hence, when decorrelating a $(K+1)$ neutral vector variable (with K degrees of freedom) into a set of K independent scalar variables, K iterations are required. During each iteration, one summation and L division should be operated for the purpose of normalization, where L is the number of elements in the intermediate vector $\mathbf{x}_i$. Therefore, if we treat the summation as one floating-point operation and the division as eight times of that[1], the computational complexity for SNT is $\mathcal{O}(NK^2)$.
  When applying Algorithm 3 to decorrelate the neutral vector in a parallel manner, at most $\lceil \log_2 (K+1) \rceil$ iterations are required. Within each iteration, about $L/2$ summations and $L/2$ divisions with an even L or $(L+1)/2$ summations and $(L+1)/2$ divisions with an odd L are needed. Therefore, with the same consideration of the floating-point operation above, the computational complexity for PNT is $\mathcal{O}(NK \log K)$, since $L = K$ at the first iteration and L will reduce to (approximately) half in each of the consequent iteration.
  With the above analysis, we can conclude that the PNT algorithm is more efficient than the SNT algorithm and preferable in practice, although both algorithms can nonlinearly transform the neutral vector into a set of mutually independent scalars.
- PCA
  The operation of PCA includes two parts: 1) eigenvalue analysis of the covariance matrix and 2) decorrelation of the vector. Many approaches exist for an eigenvalue analysis. To our best knowledge, the fastest method so-far is the method proposed by Luk et al. [66]. The computational cost is about $\mathcal{O}(K^2 \log K)$ for a $K \times K$ covariance matrix. For the decorrelation, multiplying the source vector with the eigenvector matrix will have computational cost around $\mathcal{O}(K^2)$. Therefore, the computational cost for PCA is, on average, $\mathcal{O}(NK^2 \log K)$.
  Hence, the proposed SNT- and PNT-based decorrelation methods are more efficient than the PCA-based method.
- ICA
  Although robust source separation performance can be achieved by ICA, the drawback of algorithms for carrying out ICA is the high computational complexity [67]. Typical algorithms for ICA requires centering, whitening, and dimension reduction as preprocessing steps to facilitate the calculation. Unlike PNT/SNT or PCA which converges fast, the convergence of ICA also depends on the number of iterations. Hence, analytically tractable solution does not exist. As introduced in [68], the computational cost for ICA, with M iterations, is $\mathcal{O}(MNK^2)$

2) Decorrelation Performance: We generated different amounts of samples from a single Dirichlet distribution, where the parameters are chosen to be $\boldsymbol{\alpha} = [2, 5, 6, 3, 7]^{\mathrm{T}}$. The proposed PNT method, which was shown more efficient than the SNT method, was applied to decorrelate the generated samples. With different amounts of data, the DCs between possible pairs of all the transformed variables were evaluated

[1]According to T. Minka's Lightspeed Matlab toolbox [65] http://research.microsoft.com/en-us/um/people/minka/software/lightspeed/.

TABLE I

EVALUATION OF THE DECORRELATION PERFORMANCE ON THE DATA GENERATED FROM A DIRICHLET DISTRIBUTION WITH $\boldsymbol{\alpha} = [2, 5, 6, 3, 7]^{\mathrm{T}}$. THE NULL HYPOTHESIS IS THAT THE RELATED TWO DIMENSIONS ARE INDEPENDENT FROM EACH OTHER (*i.e.*, THE DC IS 0). THE FIRST ROW: p-VALUES FOR THE GENERATED DATA. THE SECOND ROW: p-VALUES FOR THE DECORRELATED DATA VIA PNT. THE THIRD ROW: p-VALUES FOR THE DECORRELATED DATA VIA PCA. THE FOURTH ROW: p-VALUES FOR THE DECORRELATED DATA VIA ICA. THE p-VALUES THAT ARE SMALLER THAN 0.05 ARE MARKED WITH UNDERLINE, INDICATING THAT THE CORRESPONDING TWO RANDOM VARIABLES ARE NOT INDEPENDENT.

(a) $N = 100$, original.

	x_1	x_2	x_3	x_4
x_1	0	0.198	0.127	0.376
x_2		0	0.007	0.140
x_3			0	0.067
x_4				0

(b) $N = 200$, original.

	x_1	x_2	x_3	x_4
x_1	0	0.054	0.063	0.189
x_2		0	0.001	0.024
x_3			0	0.047
x_4				0

(c) $N = 400$, original.

	x_1	x_2	x_3	x_4
x_1	0	0.010	0.004	0.069
x_2		0	0.000	0.001
x_3			0	0.002
x_4				0

(d) $N = 800$, original.

	x_1	x_2	x_3	x_4
x_1	0	0.000	0.000	0.007
x_2		0	0.000	0.000
x_3			0	0.000
x_4				0

(e) $N = 100$, with PNT.

	u_1	u_2	u_3	u_4
u_1	0	0.455	0.426	0.546
u_2		0	0.481	0.405
u_3			0	0.495
u_4				0

(f) $N = 200$, with PNT.

	u_1	u_2	u_3	u_4
u_1	0	0.464	0.527	0.455
u_2		0	0.621	0.625
u_3			0	0.508
u_4				0

(g) $N = 400$, with PNT.

	u_1	u_2	u_3	u_4
u_1	0	0.583	0.484	0.668
u_2		0	0.538	0.402
u_3			0	0.582
u_4				0

(h) $N = 800$, with PNT.

	u_1	u_2	u_3	u_4
u_1	0	0.519	0.360	0.367
u_2		0	0.561	0.496
u_3			0	0.564
u_4				0

(i) $N = 100$, with PCA.

	u_1	u_2	u_3	u_4
u_1	0	0.307	0.565	0.606
u_2		0	0.211	0.330
u_3			0	0.207
u_4				0

(j) $N = 200$, with PCA.

	u_1	u_2	u_3	u_4
u_1	0	0.142	0.511	0.625
u_2		0	0.075	0.152
u_3			0	0.019
u_4				0

(k) $N = 400$, with PCA.

	u_1	u_2	u_3	u_4
u_1	0	0.048	0.395	0.472
u_2		0	0.003	0.084
u_3			0	0.000
u_4				0

(l) $N = 800$, with PCA.

	u_1	u_2	u_3	u_4
u_1	0	0.001	0.258	0.197
u_2		0	0.000	0.008
u_3			0	0.000
u_4				0

(m) $N = 100$, with ICA.

	u_1	u_2	u_3	u_4
u_1	0	0.080	0.098	0.104
u_2		0	0.095	0.092
u_3			0	0.086
u_4				0

(n) $N = 200$, with ICA.

	u_1	u_2	u_3	u_4
u_1	0	0.124	0.126	0.136
u_2		0	0.142	0.145
u_3			0	0.108
u_4				0

(o) $N = 400$, with ICA.

	u_1	u_2	u_3	u_4
u_1	0	0.073	0.222	0.324
u_2		0	0.123	0.134
u_3			0	0.155
u_4				0

(p) $N = 800$, with ICA.

	u_1	u_2	u_3	u_4
u_1	0	0.091	0.241	0.174
u_2		0	0.329	0.353
u_3			0	0.114
u_4				0

and the corresponding p-values are listed in Tab. I(e), I(f), I(g), and I(h), respectively. To make extensive comparison, we also applied the PCA-based decorrelation method and the ICA-based decorrelation method, respectively, to the generated data and summarized the decorrelation performance in Tab. I(i)-Tab. I(p).

When the amount of samples is small (*e.g.*, $N = 100$), the generated data cannot reveal neutrality completely (*e.g.*, in Tab. I(a), the p-value for the DC between x_1 and x_2 is larger than 0.05. This indicates that these two variables are independent of each other, which is in conflict with the definition of neutrality.), PNT, PCA, and ICA methods can decorrelate the "semi"-neutral vector variable into a set of mutually independent scalar variables. As the amount of sample increases, the neutrality of the data becomes clear (*i.e.*, all the p-values are smaller than 0.05 in Tab. I(b), I(c), and I(d)). It can be observed that both the PNT and the ICA algorithms can yield mutually independent variables for all the cases (p-value is larger than 0.05). In contrast, the PCA algorithm can only lead to partially mutual independence.

In summary, the proposed strategy can nonlinearly transform the highly negatively correlated neutral vector variable into a set of mutually independent scalar variables. Compared with PCA, PNT and ICA show better decorrelation performance for the data with neutral property, with a wide range of amounts of samples. In order to remove the effect of randomness, we ran 50 rounds of simulations and the mean values are reported in Tab. I. Each round of simulation includes data generation, PNT decorrelation, PCA decorrelation, ICA decorrelation, and DC calculation.

3) Discussions: We compared the computational complexities of SNT, PNT, PCA, and ICA in Sec. V-A1. The proposed SNT and PNT methods have less computational complexity compared to PCA and ICA. In all of these methods, PNT has the least computational complexity. ICA has the largest computational complexity (usually, M is a number larger than $\log K$). At the meantime, it does not have analytically tractable solution and needs many iterations to converge.

When evaluating these methods with decorrelation performance, we only used PNT to represent the proposed nonlinear transformation strategies. It can be observed that both PNT and ICA have good decorrelatoin performance (in terms of mutual independence measured by DC) for neutral vector variables, with a wide range of data amounts. PCA does not perform well for neutral vector variables when N increases.

In summary, for neutral vector variable, PNT performs better than PCA and ICA, in terms of both decorrelation and computational complexity. Comparing with PNT and PCA, ICA does not have an analytically tractable solution. Therefore, ICA algorithms typically resort to iterative procedures with either difficulties or high computational load. Hence, we compare only PNT and PCA in the following experiments.

B. Synthesized Data Evaluation

1) Mixture of Dirichlet Distributions: In real applications, the data we obtained are usually multimodally distributed. The neutral vector variable is, however, uni-modally distributed by definition. Hence, it is of sufficient interest to study the decorrelation performance of the proposed method on the data sampled from a mixture of Dirichlet distributions. In this section, we generated a set of data from a mixture of Dirichlet distributions to evaluate the decorrelation performance. The chosen model contains two mixture components, which has mixture coefficients as $\pi_1 = 0.3,\ \pi_2 = 0.7$, and component parameters as $\boldsymbol{\alpha}_1 = [2, 5, 6, 3, 7]^{\mathrm{T}},\ \boldsymbol{\alpha}_2 = [10, 2, 8, 2, 18]^{\mathrm{T}}$. Table II shows the decorrelation performance on the whole data set. The upper row illustrates the decorrelation performance for the data set with $N = 50$ samples. As mentioned in the previous section, small amount of data from a single component cannot completely reveal the neutrality. Hence, the data

TABLE II

EVALUATION OF THE DECORRELATION PERFORMANCE ON THE DATA GENERATED FROM A MIXTURE OF DIRICHLET DISTRIBUTIONS WITH $\pi_1 = 0.3,\ \pi_2 = 0.7$, AND $\boldsymbol{\alpha}_1 = [2,5,6,3,7]^{\mathrm{T}},\ \boldsymbol{\alpha}_2 = [10,2,8,2,18]^{\mathrm{T}}$. THE UPPER ROW: p-VALUES FOR THE DATA SET WITH $N = 50$ SAMPLES. THE BOTTOM ROW: p-VALUES FOR THE DATA SET WITH $N = 800$ SAMPLES. THE p-VALUES THAT ARE SMALLER THAN 0.05 ARE MARKED WITH UNDERLINE, INDICATING THAT THE CORRESPONDING TWO RANDOM VARIABLES ARE NOT INDEPENDENT.

(a) Whole data set, original.

	x_1	x_2	x_3	x_4
x_1	0	0.107	$\underline{0.021}$	$\underline{0.001}$
x_2		0	0.246	$\underline{0.019}$
x_3			0	0.359
x_4				0

(b) Whole data set, with PNT.

	u_1	u_2	u_3	u_4
u_1	0	$\underline{0.031}$	$\underline{0.029}$	$\underline{0.000}$
u_2		0	0.321	0.109
u_3			0	0.147
u_4				0

(c) Cluster 1, with PNT.

	u_1	u_2	u_3	u_4
u_1	0	0.471	0.610	0.480
u_2		0	0.463	0.513
u_3			0	0.422
u_4				0

(d) Cluster 2, with PNT.

	u_1	u_2	u_3	u_4
u_1	0	0.468	0.410	0.502
u_2		0	0.614	0.559
u_3			0	0.534
u_4				0

(e) Whole data set, original.

	x_1	x_2	x_3	x_4
x_1	0	$\underline{0.000}$	$\underline{0.000}$	$\underline{0.000}$
x_2		0	$\underline{0.001}$	$\underline{0.000}$
x_3			0	$\underline{0.023}$
x_4				0

(f) Whole data set, with PNT.

	u_1	u_2	u_3	u_4
u_1	0	$\underline{0.000}$	$\underline{0.000}$	$\underline{0.000}$
u_2		0	$\underline{0.000}$	$\underline{0.000}$
u_3			0	$\underline{0.000}$
u_4				0

(g) Cluster 1, with PNT.

	u_1	u_2	u_3	u_4
u_1	0	0.529	0.484	0.429
u_2		0	0.511	0.630
u_3			0	0.469
u_4				0

(h) Cluster 2, with PNT.

	u_1	u_2	u_3	u_4
u_1	0	0.483	0.459	0.414
u_2		0	0.531	0.474
u_3			0	0.517
u_4				0

generated from a mixture of Dirichlet distributions may still have mutual independence between some pairs of dimensions (*e.g.*, in Tab. II(a), the p-value for the DC between x_2 and x_3 is larger than 0.05, which indicates mutual independence.) In such case, when applying the PNT algorithm to the whole data set, it yields only *partially* mutual independence (see Tab. II(b)). For each data cluster, the PNT algorithm works well, as expected (see Tab. II(c) and II(d)). With large amount of data ($N = 800$), the data generated from each mixture component have strong neutral property so that the whole data set are highly correlated but *not* neutral (see Tab. II(e)). In this case, the PNT algorithm does not work (see Tab. II(f)). This is because the proposed decorrelation strategy is based on the assumption of neutrality and it may not work for the data that are not neutral. However, if we partition the data into clusters where each cluster contains data vectors that are neutral, the PNT algorithm can perfectly leads to mutual independence between any possible pairs of decorrelated dimensions (see Tab. II(g) and II(h)).

2) Coding Gain/Removal of Memory Advantage: One advantage of the proposed nonlinear transformation strategy occurs in high rate quantization of vectors. In the application of source coding, the source vectors are usually highly correlated. Hence, it is natural to decorrelate the vector into a set of mutually independent scalars so that the vector quantization (VQ) can be replaced by a set of scalar quantization (SQ) without losing the memory advantage [69]. This can be quantified by the so-called coding gain measurement [69], [70]. For different quantization methods, the coding gain can be measured as (or proportional to) the ratio of quantization distortions, with a given number of bits for quantization.

As shown in [71], with the high rate assumption, the distortion incurred by quantizing a vector approaches a simple quadratically weighted error as

$$d(\mathbf{x}, \widehat{\mathbf{x}}) = (\mathbf{u} - \widehat{\mathbf{u}})^{\mathrm{T}}\, \mathcal{J}_{\mathcal{T}}^{\mathrm{T}}(\mathbf{u}) \mathcal{J}_{\mathcal{T}}(\mathbf{u})\, (\mathbf{u} - \widehat{\mathbf{u}})\,, \tag{13}$$

where $\mathcal{J}_{\mathcal{T}}$ is the Jacobian matrix of the *inverse* PNT algorithm $\mathbf{x} = \mathcal{T}(\mathbf{u})$. The distortion in the $\mathbf{x}$ domain, incurred by quantizing $\mathbf{u}$, can be approximated as [32]

$$D_{\mathbf{x}}(\mathbf{u}) \cong \sum_{k=1}^{K} \mathbf{E}\left[\mathcal{J}_{\mathcal{T}}^{\mathrm{T}}(\mathbf{u}) \mathcal{J}_{\mathcal{T}}(\mathbf{u})\right]_{k,k} \times D(u_k), \tag{14}$$

where K is the dimensionality of $\mathbf{u}$ and $\mathbf{E}[\cdot]$ denotes expectation operation. In the above equation, we denote $D(u_k)$ as the distortion incurred by quantization of u_k in the $\mathbf{u}$ domain. By assuming that $\mathbf{x}$ is Dirichlet distributed with known parameters, we can apply the PNT algorithm to transform $\mathbf{x}$ to $\mathbf{u}$, and u_k is beta distributed (see (10)) [32]. With the high rate theory and entropy constrained quantization [69], we can derive that, with R bits and probability density function (PDF)-optimized bit allocation strategy [72], the distortion in the $\mathbf{x}$ domain incurred by quantizing $\mathbf{u}$ is [32]

$$D_{\mathbf{x}}(\mathbf{u}) = \frac{K}{12} \times 2^{-\frac{2}{K} \times \left[R - \sum_{k=1}^{K} h(u_k)\right]} \times \sqrt[K]{\prod_{k=1}^{K} \mathbf{E}\left[\mathcal{J}_{\mathcal{T}}^{\mathrm{T}}(\mathbf{u}) \mathcal{J}_{\mathcal{T}}(\mathbf{u})\right]_{k,k}}\,,$$

where $h(u_k)$ is the differential entropy of u_k.

On the other hand, if we quantize each element in $\mathbf{x}$ according to its marginal distribution (this means we replace a vector quantizer by a set of scalar quantizer without decorrelation), the distortion is

$$D_{\mathbf{x}}(\mathbf{x}) = \frac{K}{12} \times 2^{-\frac{2}{K} \times \left[R - \sum_{k=1}^{K} h(x_k)\right]}. \tag{15}$$

For a $(K+1)$-dimensional Dirichlet distribution with parameter $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_{K+1}]^T$, the marginal distribution for the kth dimension is

$$x_k \sim \mathbf{Beta}(x_k; \alpha_k, \sum_{i=1, i\neq k}^{K+1} \alpha_i). \tag{16}$$

Thus we can measure the coding gain as the ratio of two distortions

$$G = \frac{D_{\mathbf{x}}(\mathbf{x})}{D_{\mathbf{x}}(\mathbf{u})} = \frac{2^{\frac{2}{K} \sum_{k=1}^{K} [h(x_k) - h(u_k)]}}{\sqrt[K]{\prod_{k=1}^{K} \mathbf{E}\left[\mathcal{J}_{\mathcal{T}}^{\mathrm{T}}(\mathbf{u}) \mathcal{J}_{\mathcal{T}}(\mathbf{u})\right]_{k,k}}}. \tag{17}$$

In the above equation, the ratio $G > 1$ indicates less distortion can be achieved by the proposed nonlinear transformation. The larger this ratio is, the more benefit we obtain from the transformation. In order to evaluate the coding gain G extensively, we evaluated the coding gain with different $\boldsymbol{\alpha}$ and different dimensionalities. To give an example, the inverse nonlinear transformation and the elements in $\mathcal{J}_{\mathcal{T}}(\mathbf{u})$ with $K = 4$ are listed in Tab. III. The expectation term in the denominator of (17) can be calculated in a closed-form expression with the fact that u_i is beta distributed and the parameters can be calculated from the original Dirichlet parameters (see (10) for more details).

The coding gains with $K = 4, 5, 6$ are plotted in Fig. 3. For each K, we randomly generated the elements in $\boldsymbol{\alpha}$ from $[10, 50]$. In total 100 rounds of simulations were conducted for each K. It can be observed that the proposed nonlinear transformation yield a coding gain greater than 1 for different

TABLE III
THE INVERSE PNT ALGORITHM $\mathbf{x} = \mathcal{T}(\mathbf{u})$ AND THE JACOBIAN MATRIX $\mathcal{J}_{\mathcal{T}}(\mathbf{u})$ FOR $(K = 4)$.

$$\mathbf{x} = \mathcal{T}(\mathbf{u}): \quad \begin{array}{l} x_1 = u_1u_3u_4 \\ x_2 = (1-u_1)u_3u_4 \\ x_3 = u_2(1-u_3)u_4 \\ x_4 = (1-u_2)(1-u_3)u_4 \end{array} \qquad \mathcal{J}_{\mathcal{T}}(\mathbf{u}) = \begin{bmatrix} u_3u_4 & 0 & u_1u_4 & u_1u_3 \\ -u_3u_4 & 0 & (1-u_1)u_4 & (1-u_1)u_3 \\ 0 & (1-u_3)u_4 & -u_2u_4 & u_2(1-u_3) \\ 0 & -(1-u_3)u_4 & -(1-u_2)u_4 & (1-u_2)(1-u_3) \end{bmatrix}$$

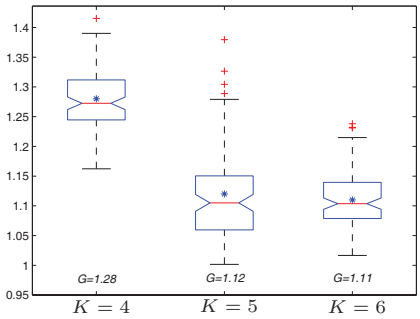


Fig. 3. Coding gains for different K shown as box plot. The central red mark is the median, the blue star mark is the mean, the edges of the box are the 25^{th} and 75^{th} percentiles. The outliers are marked with red crosses. The mean values are listed at bottom.

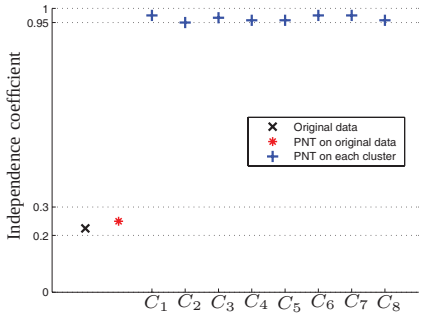


Fig. 4. Independence coefficients of different data set. C_i denotes the i^{th} cluster obtained by the EM algorithm. The amount of samples is $N = 800$ and the number of mixture components (clusters) is 8.

dimensions. This is because the memory advantage of VQ over SQ has been removed.

3) Discussion: The synthesized data experiments above demonstrated the superior performance of the proposed nonlinear transformation strategy for neutral data. The data generated from a mixture of Dirichlet distributions are multimodally distributed so that they are not neutral. In this case, we can partition the data into different clusters. By assuming that the data assigned to each cluster were generated from a single Dirichlet distribution, the proposed method can be applied to these data and results in promising decorrelation performance. Decorrelation of highly negatively correlated vector plays an important role in many applications. In the next section, we will apply this idea to real data applications.

C. Real Data Evaluation

Decorrelation of a highly correlated vector variable into a set of mutually independent variables leads to many advantages in real applications [21], [32], [39], [40], [69]. In this section, we evaluate the decorrelation performance of the proposed strategy for real life data that fit the definition of neutral vector (nonnegative and l_1 norm equals one). To this end, we assume such "neutral-like"[2] data have neutral property and apply the PNT algorithm to nonlinearly transform them. The performance improvement in practical applications is also presented.

*1) **Vector Quantization of Line Spectral Frequency Parameters**:* Quantization of the LSF parameters of the linear predictive coding (LPC) model is an essential part of speech transmission [32], [73], [74]. The LSF parameters are usually 10-dimensional for narrow band speech and 16-dimensional for wide band speech. Hence, vector quantization (VQ) is required. Generally speaking, VQ has memory, shape, and space-filling advantages over scalar quantization (SQ) [69], [73]. However, it is *impractical* to design a full vector quantizer because 1) the size of codebook increases exponentially with the dimension of data, which leads to high storage complexity; 2) the effort of training a codebook and searching for an index in the codebook is also exponentially increased with the data's dimension, which is computationally costly. Especially, when the dimension is high, *e.g.*, > 10, the above VQ is not feasible. In practical VQ implementation, the frequently used method is to decorrelate the LSF parameters into a set of mutually independent scalars so that the memory advantage of VQ over SQ can be removed [69], [70], [73]. Then, a set of SQs will be employed to replace the VQ.

In the design of PDF-optimized VQ, the Gaussian distribution and the corresponding Gaussian mixture model (GMM) have been intensively applied to model the distribution of the LSF parameters [30], [75], [76]. However, since the LSF parameters are in the interval $(0, \pi)$ and are strictly ordered, it is *not* Gaussian distributed. For the purpose of more efficient modeling, the LSF parameters can be converted to the so-called ΔLSF parameters [32], [72]. The ΔLSF parameters are nonnegative and the summation equals 1[3]. As the ΔLSF parameters fit the the definition, we suppose that they follow Dirichlet distributions and apply a Dirichlet mixture model (DMM) to describe the underlying distribution of the data. As data generated from a Dirichlet distribution have neutral property, the proposed nonlinear strategy is applied to decorrelate the ΔLSF parameters. A practical VQ is carried out based on the neutrality.

- *Evaluation of Independence*
  The ΔLSF parameters are 16-dimensional[4] for wide band speech data. It is space consuming to list a 16×16 mutual independence p-value table. Thus, we calculated *independence coefficient* (IC), which is defined as the proportion of the number of mutually independent pairs to the number of all the possible pairs[5] to measure the decorrelation performance. The higher this proportion is,

[2]Hereby, we name the vector 1) contains nonnegative elements and 2) has unit/constant l_1-norm as "neutral-like" data.

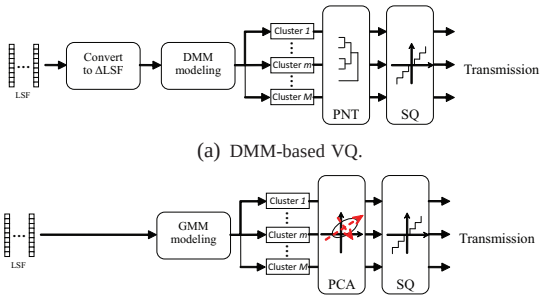
[3]Strictly speaking, the summation of the ΔLSF parameters equals π, which can be scaled so that the summation equals 1. The scaled ΔLSF parameters represent the proportions of the ΔLSF on the unit circle [32].

[4]We show only the results for wide band data here. Similar performance can also be obtained for narrow band data.

[5]For a $K \times K$ matrix, the number of all the possible pair is $\frac{K(K-1)}{2}$, without consideration of self pairs.

the better the decorrelation performance is[6].
As described in Sec. V-B3, we firstly applied the PNT algorithm to the ΔLSF parameters. As shown in Fig. 4, the IC of PNT for the original data is small, which means the decorrelation performance of PNT is not significant. This is due to the fact that the ΔLSF parameters are multimodally distributed. We applied the EM algorithm [32] to partition the ΔLSF parameters into different clusters. With the assumption that the data in each cluster are Dirichlet distributed (hence, they are neutral vectors), we applied the PNT algorithm to the data in each cluster, respectively. The ICs of PNT for each cluster are also plotted in Fig. 4. It is clearly shown that most of the pairs (more than 95%) are mutually independent. Hence, the mutual correlation for each cluster has been significantly removed by PNT.

- *Improvement in VQ*
Motivated by the coding gain advantage in Sec. V-B2, we designed and implemented a DMM-based VQ based on the neutral properties. The LSF parameters were partitioned into I[7] clusters with a DMM which contains I mixture components [72]. With the above introduced procedure, the PNT algorithm is applied to realize the decorrelation for each cluster and a set of mutually independent scalar elements are obtained. As the memory advantage of VQ over SQ is removed by explicitly using the neutrality, we carried out a PDF-optimized VQ for the LSF parameters. The benefits are two fold:
    1) *Saving of the storage, training and searching costs.* With average bit rate (in per vector sense) R, there are $\log_2 M$ bits spent on indexing the mixture component and $R_q = R - \log_2 M$ bits spent on VQ. Hence, by assuming all the components are identical to each other, a codebook with 2^{R_q} codewords is required for each mixture component. In the SQ case, the bit for each cluster (*i.e.*, mixture component) will be further placed on each dimension based on its differential entropy. On average, $\frac{R_q}{16}$ is assigned to each dimension and only $16 \times 2^{\frac{R_q}{16}}$ is needed for each component. Usually, R is a number about $40 \sim 50$. Hence, the required number of codewords is significantly reduced and the storage cost is saved. The well-known Lloyd algorithm [77], [78] and the Linde-Buzo-Gray (LGB) algorithm [79], [80] are usually utilized for obtaining the codebook. In the case of VQ, the training is carried out in a 16-dimensional space. Meanwhile, the training is executed in one-dimensional space for SQ. Obviously, training a codebook in 16-dimensional space is more computationally costly than that in one-dimensional space, and, therefore, the training cost is saved. For the same reasoning, the searching cost is also significantly reduced when replacing VQ by SQ.
    2) *Saving of Bit rates.* The ultimate goal of PDF-optimized VQ is to spend as less bits as possible while satisfying the quantization distortion requirement. A practical VQ for the LSF parameters, which is based on the DMM modeling and the proposed nonlinear transformation strategy, was introduced in [32]. With the transparent coding[8] criterion, we evaluated the log spectral distortion (LSD) obtained from the DMM-based VQ and compared it with the state-of-the-art GMM-based VQ [81]. The GMM-based VQ partitioned the LSF parameters into I clusters with the EM algorithm for GMM. Next, the LSF parameters are decorrelated with PCA. Finally, a PDF-optimized GMM-based VQ is carried as well. Fig. 5 shows the designs for the DMM-based VQ and the GMM-based VQ. The VQ performance comparisons are summarized in Tab. IV. It is clearly demonstrated that the DMM-based VQ improves the performance by about 3 bits/vector. This is due to the fact that the proposed nonlinear transformation strategy removes the memory advantage and makes the implementation of practical VQ feasible. More details can be found in [32].



(a) DMM-based VQ.

(b) GMM-based VQ.

Fig. 5. Flow chart of DMM-based VQ and GMM-based VQ.

*2) **EEG Signal Classification**:* For persons who suffer from neuromuscular diseases, brain-computer interface (BCI) connects them with computers by recording and analyzing the brain signals. As non-invasively acquired signal, the Electroencephalogram (EEG) signal is the most studied and applied one in the design of a BCI system [82], [83]. For the EEG signal obtained from one channel, various types of features have been extracted from the signal for the purpose of classification. The marginal discrete wavelet transform (mDWT) vector, among others, is a typical feature that is widely adopted [84]–[86]. The elements in a DWT vector reveal features related to the transient nature of the signal. The marginalization operation, which yields the mDWT vector, makes the DWT vector insensitive to time alignment [84]. The data set used in this paper is from the BCI competition III [87]. During one EEG signal trial recording, a subject had to perform imagined movements of either the left small finger or the tongue. The data set contains 278 trials for training and 100 trials for test. The trials in the training and test sets are evenly distributed and labeled, respectively. For each trial, 64

[6]The largest ratio is 1, which means all the possible pairs are mutually independent.

[7]Usually, I equals a power of 2.

[8]Transparent coding criterion: 1) 1 dB LSD on average, 2) less than 2% outliers in $2 \sim 4$ dB range, and 3) no outlier larger than 4 dB.

TABLE IV
COMPARISONS OF VQ PERFORMANCE. THE NUMBER OF MIXTURE COMPONENTS IS $M = 256$. $706k$ LSF VECTORS WERE USED FOR TRAINING AND $258k$ WERE USED FOR EVALUATION. THE SPEECH DATA ARE FROM THE TIMIT DATABASE [91].

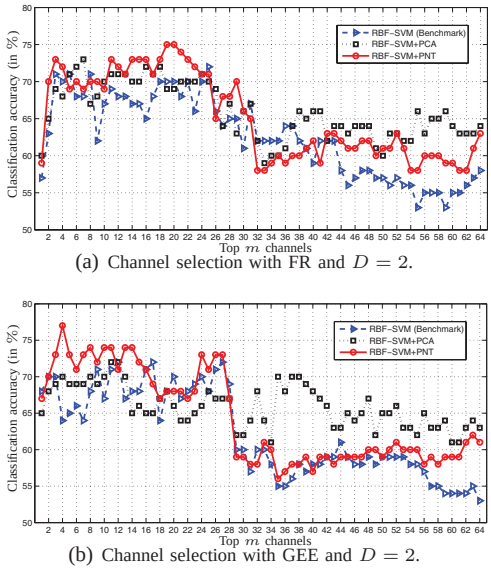
VQ Type	bits/vec.	LSD (dB)	LSD outliers (in %)	
			$2-4$ dB	>4 dB
DMM-based VQ	44	1.039	1.200	0.000
	45	0.997	0.830	0.000
GMM-based VQ	47	1.029	0.776	0.005
	48	0.971	0.920	0.003

channel data of length 3000 samples were provided. The mDWT vector contains nonnegative elements and has unit l_1-norm. Hence, we applied the nonlinear transformation method to decorrelate the mDWT vector for the purposed of classification accuracy improvement.

In our previous work [88], we have successfully applied the proposed PNT method in EEG signal classification. The so-called multivariate Beta distribution (mvBeta)-based classifier was introduced based on the feature selection strategy in the transformed feature domain and has been applied to classify the EEG signals. In this paper, we will make thorough study to show that the obtained gain in classification accuracy is indeed from the application of the PNT method to the mDWT vectors.

- *Channel Selection*
  Not all the channels are closely relevant to the classification task. Before conducting the classification task, it is of importance to select more relevant channels so that the classification accuracy can be improved. The Fisher ratio (FR) and the generalization error estimation (GEE) [88], [89] were applied to select channels. The channels are ranked according to their FRs and GEEs, respectively. In the classification stage, we exploit the mDWT vectors from the top m channels.
- *Feature Selection*
  Feature selection is an important problem in EEG signal classification [84], [88], [90]. For each selected channel, the dimension of the mDWT vector is 5 (the degrees of freedom is 4). We applied the PNT algorithm to decorrelate the mDWT vectors from the training set. A set of 4-dimensional vectors, each of which contains mutually independent elements were obtained. We sorted the 4 dimensions according to their variances in descending order. The mDWT vectors from the test set were also decorrelated via PNT. The dimension reordering was carried out based on the variance order from the training set. According to the reordered dimensions, we selected the relevant D dimensions for classification.

- *Performance Improvement*
  For binary classification task, the support vector machine (SVM) is a classic and the widely applied classifier [21], [92]–[94]. We evaluated the above introduced feature selection strategy by comparing the classification accuracies. For each channel selection method, an SVM with radial basis function (RBF) kernel was trained as the benchmark, respectively. With LIBSVM toolbox [95], we adjusted the parameters in the RBF-SVM so that the cross validation of training accuracy is the highest. *All* mDWT vectors from the training set were used for the parameter adjustment. To make fair comparisons, we also applied PCA to decorrelate the mDWT vectors. The mDWT vectors in the test set were transformed with the eigenvectors obtained from the training set. The relevant dimensions were also selected according to the variances. The classification results were obtained with the top m channels (ranked via FR or GEE). For each channel, the most relevant D features (ranked via variance) were selected. In total, we obtained $(m \times D)$-dimensional feature vector to train the RBF-SVM. It can be observed that the RBF-SVM+PNT yields the highest recognition accuracies, for FR case and GEE case, respectively. Figure 6 illustrates the classification results with top m channels when $D = 2$. The highest classification rates are both obtained with $D = 2$, which indicates that feature selection via variance indeed benefits the classification. The RBF-SVM+PNT yields the highest recognition accuracy for FR case (75% with with $D = 2$ and $m = 19, 20$) and GEE case (77% with $D = 2$ and $m = 4$), respectively.



(a) Channel selection with FR and $D = 2$.

(b) Channel selection with GEE and $D = 2$.

Fig. 6. Classification accuracy comparisons of RBF-SVM (benchmark, no transformation), RBF-SVM+PCA, and RBF-SVM+PNT with $D = 2$. The results have been presented in [88].

3) Discussion: The LSF parameters in the LPC model and the mDWT parameters in the EEG signal contain nonnegative elements and have unit/constant l_1-norm, respectively. Although it is difficult (or even not feasible) to prove the neutrality for such neutral-like data, we can still exploit the neutrality to apply the PNT-based nonlinear transformation strategy for the purpose of decorrelation and improve practical performance. Compared with the PCA-based linear transformation strategy, the PNT-based nonlinear transformation showed advantages in both applications.

VI. CONCLUSIONS

Nonlinear transformations for neutral vector variable were proposed and studied in this paper. By explicitly utilizing the neutrality of neutral vector variables, we introduced the serial

nonlinear transformation and parallel nonlinear transformation methods to decorrelate a neutral vector variable into a set of mutually independent element variables. The mutual independence was theoretically proved. The computational costs of the proposed decorrelation methods were analyzed and compared with the PCA-based and ICA-based approaches. It has been shown that the computational costs of the proposed methods are the smallest.

As a typical case, the vector variable following the Dirichlet distribution is a completely neutral vector. The transformed element variables are all beta distributed. With the distance correlation metric, the decorrelation performance of the proposed nonlinear transformation was demonstrated to be superior to those of PCA and ICA with both synthesized and real life data. Moreover, we applied the proposed nonlinear transformation in two applications, *i.e.*, quantization of line spectral frequency parameters in the speech linear predictive model and EEG signal classification. Extensive experimental results showed that, when carrying out decorrelation and feature selection for neutral-like data, the proposed parallel nonlinear transformation (PNT)-based nonlinear transformation can achieve better practical performance and is preferable to the conventionally applied PCA-based linear transformation.

APPENDIX A
PROOF OF PROPERTY 2.1

This property can be readily proved by iteratively using the neutral property of $\mathbf{x}_{k1}$. The reader is referred to [37, pp.196] for more details.

APPENDIX B
PROOF OF PROPERTY 2.2

Due to the completely neutral property, we have $\mathbf{x}_{k1} \perp \mathbf{w}_k$, $1 \leq k \leq K$, where $\perp$ denotes independence. For the K-dimensional vector $\mathbf{x}^{r \uplus r+1}$[9],

1) When $1 \leq k < r$, it can be recognized that the elements in $\mathbf{x}_{k1}^{r \uplus r+1}$ are identical to those in $\mathbf{x}_{k1}$. The only difference between $\mathbf{w}_k^{r \uplus r+1}$ and $\mathbf{w}_k$ is that $\mathbf{w}_k^{r \uplus r+1}$ contains element $\frac{x_r + x_{r+1}}{1-s_k}$ while $\mathbf{w}_k$ contains $[\frac{x_r}{1-s_k}, \frac{x_{r+1}}{1-s_k}]$. Based on these facts, we can immediately show that $\mathbf{x}_{k1}^{r \uplus r+1}$ is independent of all the elements in $\mathbf{w}_k^{r \uplus r+1}$ except for $\frac{x_r + x_{r+1}}{1-s_k}$. On the other hand, we also have

$$\mathbf{x}_{k1}^{r \uplus r+1} \perp \mathbf{w}_k \Rightarrow \quad \mathbf{x}_{k1}^{r \uplus r+1} \perp [\tfrac{x_r}{1-s_k}, \tfrac{x_{r+1}}{1-s_k}] \tag{18}$$

$$\Rightarrow \quad \mathbf{x}_{k1}^{r \uplus r+1} \perp \tfrac{x_r + x_{r+1}}{1-s_k}, \tag{19}$$

Hence, it can be proved that $\mathbf{x}_{k1}^{r \uplus r+1}$ is independent of $\frac{x_r + x_{r+1}}{1-s_k}$ and, therefore, $\mathbf{x}_{k1}^{r \uplus r+1}$ is neutral for $1 \leq k < r$.

2) When $r < k < K$, $\mathbf{w}_k^{r \uplus r+1} = \mathbf{w}_k$ and the distinct elements in $\mathbf{x}_{k1}^{r \uplus r+1}$ and $\mathbf{x}_{k1}$ are $x_r + x_{r+1}$ and $[x_r, x_{r+1}]$, respectively. For the same reasoning, we can also prove that $\mathbf{x}_{k1}^{r \uplus r+1}$ is neutral for $r < k \leq K$.

Based on these, we conclude that $\mathbf{x}_{k1}^{r \uplus r+1}$ is neutral for $1 \leq k \leq K$ and $\mathbf{x}^{r \uplus r+1}$ is completely neutral.

[9]We use similar notation as defined at the beginning of Sec. II.

APPENDIX C
PROOF OF INDEPENDENCE AFTER PNT

1) Independence within Subvector $\mathbf{u}_i$: According to the PNT scheme in Alg. 2, at the i^{th} iteration, we obtain a new vector $\mathbf{x}_i = [x_{1,i-1} + x_{2,i-1}, x_{3,i-1} + x_{4,i-1}, \ldots]^{\mathrm{T}}$, where we denote the l^{th} element in the $\mathbf{x}_i$ as $x_{l,i}$ and define $\mathbf{x}_1 = \mathbf{x}$. With Property 2.2 (the aggregation property), it can be readily shown that $\mathbf{x}_i$ is completely neutral for any i.

In the i^{th} iteration, the elements in $\mathbf{u}_i$ are $u_{l,i} = \frac{x_{2l-1,i}}{x_{2l-1,i} + x_{2l,i}}$. For any two elements $u_{m,i}$ and $u_{n,i}$ (we assume $m < n$ here), we have the following relation

$$[\ldots, x_{2m-1,i}, x_{2m,i}]^{\mathrm{T}} \perp [\ldots, w_{2n-1,i}, w_{2n,i}, \ldots]^{\mathrm{T}}, \tag{20}$$

which is due to the completely neutrality of $\mathbf{x}_i$. Here, $w_{2n-1,i} = \frac{x_{2n-1,i}}{1-s_{2m}}$. By recognizing $u_{m,i} = \frac{x_{2m-1,i}}{x_{2m-1,i} + x_{2m,i}}$ and $u_{n,i} = \frac{x_{2n-1,i}}{x_{2n-1,i} + x_{2n,i}} = \frac{w_{2n-1,i}}{w_{2n-1,i} + w_{2n,i}}$ and denoting $\bar{u}_{m,i} = 1 - u_{m,i}$ and $\bar{u}_{n,i} = 1 - u_{n,i}$, the relation between $[u_{m,i}, \bar{u}_{m,i}, u_{n,i}, \bar{u}_{n,i}]^{\mathrm{T}}$ and $[x_{2m-1,i}, x_{2m,i}, w_{2n-1,i}, w_{2n,i}]^{\mathrm{T}}$ can be presented as

$$[u_{m,i}, \bar{u}_{m,i}, u_{n,i}, \bar{u}_{n,i}]^{\mathrm{T}} = \mathcal{H}\left([x_{2m-1,i}, x_{2m,i}, w_{2n-1,i}, w_{2n,i}]^{\mathrm{T}}\right). \tag{21}$$

The Jacobian matrix of the above transformation is

$$\mathcal{J}_{\mathcal{H}} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}, \tag{22}$$

where

$$\mathbf{A} = \begin{bmatrix} \frac{\partial u_{m,i}}{\partial x_{2m-1,i}} & \frac{\partial u_{m,i}}{\partial x_{2m,i}} \\ \frac{\partial \bar{u}_{m,i}}{\partial x_{2m-1,i}} & \frac{\partial \bar{u}_{m,i}}{\partial x_{2m,i}} \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} \frac{\partial u_{n,i}}{\partial w_{2n-1,i}} & \frac{\partial u_{n,i}}{\partial w_{2n,i}} \\ \frac{\partial \bar{u}_{n,i}}{\partial w_{2n-1,i}} & \frac{\partial \bar{u}_{n,i}}{\partial w_{2n,i}} \end{bmatrix}. \tag{23}$$

By the principles of variable substitution, we have

$$\begin{aligned} & f(x_{2m-1,i}, x_{2m,i}, w_{2n-1,i}, w_{2n,i}) \\ &= |\det(\mathcal{J}_{\mathcal{H}})| f(u_{m,i}, \bar{u}_{m,i}, u_{n,i}, \bar{u}_{n,i}) \\ &= |\det(\mathbf{A})||\det(\mathbf{B})| f(u_{m,i}, \bar{u}_{m,i}, u_{n,i}, \bar{u}_{n,i}). \end{aligned} \tag{24}$$

Similarly, the following relations also hold

$$\begin{aligned} f(x_{2m-1,i}, x_{2m,i}) &= |\det(\mathbf{A})| f(u_{m,i}, \bar{u}_{m,i}) \\ f(w_{2n-1,i}, w_{2n,i}) &= |\det(\mathbf{B})| f(u_{n,i}, \bar{u}_{n,i}). \end{aligned} \tag{25}$$

Combining (20), (24), and (25), we can obtain

$$f(u_{m,i}, \bar{u}_{m,i}, u_{n,i}, \bar{u}_{n,i}) = f(u_{m,i}, \bar{u}_{m,i}) f(u_{n,i}, \bar{u}_{n,i}) \tag{26}$$

and infer that $u_{m,i} \perp u_{n,i}$. Hence, the elements within the group $\mathbf{u}_i$ are mutually independent. Note that this proof is different from that shown in [32], as no permutation property of $\mathbf{x}$ is used.

2) Independence between Subvectors $\mathbf{u}_i$ and $\mathbf{u}_j$: In Algorithm 2, each iteration yields one subvector $\mathbf{u}_i$ based on $\mathbf{x}_i$. Taking two arbitrary subvectors $\mathbf{u}_i$ and $\mathbf{u}_j$ (we suppose $i < j$) and selecting arbitrary elements $u_{p,i}$ and $u_{q,j}$ from each subvector, respectively, we have the following transformation

$$[u_{p,i}, u_{q,j}, \bar{u}_{q,j}]^{\mathrm{T}} = \mathcal{G}\left([u_{p,i}, x_{2q-1,j}, x_{2q,j}]^{\mathrm{T}}\right), \tag{27}$$

where $u_{q,j} = \frac{x_{2q-1,j}}{x_{2q-1,j} + x_{2q,j}}$ and $\bar{u}_{q,j} = 1 - u_{q,j}$. Similar as the proof procedure in Sec. C-1, we get the Jacobian matrix of the transformation $\mathcal{G}$ as

$$\mathcal{J}_{\mathcal{G}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & & \\ 0 & & \mathbf{C} \end{bmatrix}, \ \mathbf{C} = \begin{bmatrix} \frac{\partial u_{q,j}}{\partial x_{2q-1,j}} & \frac{\partial u_{q,j}}{\partial x_{2q,j}} \\ \frac{\partial \bar{u}_{q,j}}{\partial x_{2q-1,j}} & \frac{\partial \bar{u}_{q,j}}{\partial x_{2q,j}} \end{bmatrix}. \tag{28}$$

With the fact[10] that $u_{p,i} \perp [x_{2q-1,j}, x_{2q,j}]^{\mathrm{T}}$, we have

$$\begin{aligned} f(u_{p,i}, u_{q,j}, \bar{u}_{q,j}) =& \frac{1}{\mid \det(\mathbf{C}) \mid} f(u_{p,i}, x_{2q-1,j}, x_{2q,j}) \\ =& \frac{1}{\mid \det(\mathbf{C}) \mid} f(u_{p,i}) f(x_{2q-1,j}, x_{2q,j}). \end{aligned} \tag{29}$$

In addition to this, we also have

$$f(x_{2q-1,j}, x_{2q,j}) = \mid \det(\mathbf{C}) \mid f(u_{q,j}, \bar{u}_{q,j}). \tag{30}$$

Thus, substituting (30) into (29), we finally get

$$f(u_{p,i}, u_{q,j}, \bar{u}_{q,j}) = f(u_{p,i}) f(u_{q,j}, \bar{u}_{q,j}), \tag{31}$$

which indicates $u_{p,i} \perp u_{q,j}$. Then it can be concluded that any two subvectors are mutually independent.

Combining the conclusion of independence within and among the subvectors, the mutual independence of the element variables in $\mathbf{u}$ is proved.

ACKNOWLEDGEMENT

This work is partly supported by the National Nature Science Foundation of China (NSFC) grant No. 61402047, 61511130081, 6167103, and 61273217, Beijing Natural Science Foundation (BNSF) grant No. 4162044, the Beijing Excellent Talent Development Foundation, the Importation and Development of High- Caliber Talents Project of Beijing Municipal Institutions (No. CIT&TCD201404052).

REFERENCES

[1] S. Park, E. Serpedin, and K. Qaraqe, "Gaussian assumption: The least favorable but the most useful," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 183–186, May 2013.

[2] T. M. Nguyen and Q. M. J. Wu, "A nonsymmetric mixture model for unsupervised image segmentation," *IEEE Transactions on Cybernetics*, vol. 43, no. 2, pp. 751–765, April 2013.

[3] Z. Ma and A. Leijon, "Bayesian estimation of beta mixture models with variational inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, vol. 33, no. 11, pp. 2160–2173, 2011.

[4] N. Bouguila, D. Ziou, and E. Monga, "Practical Bayesian estimation of a finite beta mixture through gibbs sampling and its applications," *Statistics and Computing*, vol. 16, pp. 215 –225, 2006.

[5] Z. Ma, A. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, and J. Guo, "Variational Bayesian matrix factorization for bounded support data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 4, pp. 876–889, 2015.

[6] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using Markov chain Monte carlo," in *Proceedings of International Conference on Machine Learning*, 2008, pp. 880–887.

[7] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorizationo," in *Proceedings of Advanced Neural Information Processing Systems*, 2008, pp. 1257–1264.

[8] Y. Ji, C. Wu, P. Liu, J. Wang, and K. R. Coombes, "Application of beta mixture models in bioinformatics," *Bioinformatics applications note*, vol. 21, pp. 2118–2122, 2005.

[9] Z. Ma and A. E. Teschendorff, "A variational bayes beta mixture model for feature selection in DNA methylation studies," *Journal of Bioinformatics and Computational Biology*, vol. 11, no. 4, 2013.

[10] N. Mohammadiha, R. Martin, and A. Leijon, "Spectral domain speech enhancement using HMM state-dependent super-Gaussian priors," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 253–256, March 2013.

[11] N. Mohammadiha and A. Leijon, "Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 998–1011, May 2013.

[12] K. Mammasis, R. W. Stewart, and J. S. Thompson, "Spatial fading correlation model using mixtures of von Mises Fisher distributions," *IEEE Transactions on Wireless Communications*, vol. 8, no. 4, pp. 2046–2055, April 2009.

[10]This can be directly observed from the exchangeably complete neutrality.

[13] J. Taghia, Z. Ma, and A. Leijon, "Bayesian estimation of the von-Mises Fisher mixture model with variational inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 9, pp. 1701–1715, Sept 2014.

[14] Z. Ma, "Non-Gaussian statistical models and their applications," Ph.D. dissertation, KTH - Royal Institute of Technology, 2011.

[15] J. Jung, S. R. Lee, H. Park, S. Lee, and I. Lee, "Capacity and error probability analysis of diversity reception schemes over generalized-K fading channels using a mixture Gamma distribution," *IEEE Trans. on Wirel. Commun.*, vol. 13, no. 9, pp. 4721–4730, Sept 2014.

[16] L. Zão and R. Coelho, "Generation of coloured acoustic noise samples with non-Gaussian distributions," *IET Signal Processing*, vol. 6, no. 7, pp. 684–688, September 2012.

[17] D. Xu, C. Shen, and F. Shen, "A robust particle filtering algorithm with non-Gaussian measurement noise using student-t distribution," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 30–34, 2014.

[18] Z. Si, R. Thobaben, and M. Skoglund, "Rate-compatible LDPC convolutional codes achieving the capacity of the BEC," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 4021–4029, June 2012.

[19] Z. Si, R. Thobaben, and M. Skoglund, "Bilayer LDPC convolutional codes for decode-and-forward relaying," *IEEE Transactions on Communications*, vol. 61, no. 8, pp. 3086–3099, August 2013.

[20] B. Chen, L. Xing, H. Zhao, N. Zheng, and J. C. Príncipe, "Generalized correntropy for robust adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3376–3387, July 2016.

[21] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[22] G. J. McLachlan and D. Peel, *Finite Mixture Models*. Wiley, 2000.

[23] D. M. Blei, "Probabilstic model of text and images," Ph.D. dissertation, University of California, Berkeley, 2004.

[24] D. M. Blei and J. D. Lafferty, "Correlated topic models," in *Advances in Neural Information Processing Systems*, 2006.

[25] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[26] G. Wyszecki and W. S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*. New York: Wiley, 2000.

[27] N. Bouguila and D. Ziou, "High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on minimum message length," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1716 –1731, Oct. 2007.

[28] S. Boutemedjet, N. Bouguila, and D. Ziou, "A hybrid feature extraction selection approach for high-dimensional non-Gaussian data clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1429–1443, Aug 2009.

[29] P. K. Rana, J. Taghia, Z. Ma, and M. Flierl, "Probabilistic multiview depth image enhancement using variational inference," *IEEE Journal of Selected Topics in Signal Processing*, 2015.

[30] A. D. Subramaniam and B. D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 130–142, Mar 2003.

[31] S. S. Yedlapalli and K. V. S. Hari, "The line spectral frequency model of a finite-length sequence," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 3, pp. 646 –658, Jun. 2010.

[32] Z. Ma, A. Leijon, and W. B. Kleijn, "Vector quantization of LSF parameters with a mixture of Dirichlet distributions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1777 – 1790, Sep. 2013.

[33] Z. Ma, S. Chatterjee, W. B. Kleijn, and J. Guo, "Dirichlet mixture modeling to estimate an empirical lower bound for LSF quantization signal processing," *Signal Processing*, vol. 104, pp. 291–295, Nov. 2014.

[34] C. Forbes, M. Evans, N. Hastings, and B. Peacock, *Statistical distributions*. Wiley, 2011.

[35] C. Chen, W. Buntine, N. Ding, L. Xie, and L. Du, "Differential topic models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 2, pp. 230–242, Feb 2015.

[36] B. Chen, S. Zhao, P. Zhu, and J. C. Príncipe, "Quantized kernel recursive least squares algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 9, pp. 1484–1491, Sept 2013.

[37] R. J. Connor and J. E. Mosimann, "Concepts of independence for proportions with a generalization of the Dirichlet distribution," *Journal of American Statistial Association*, vol. 64, no. 325, pp. 194–206, 1969.

[38] I. R. James, "Products of independent beta variables with applications to Connor and Mosimann's generalized Dirichlet distribution," *Journal of American Statistical Association*, vol. 67, no. 340, pp. 910–912, 1972.

[39] F. De la Torre, "A least-squares framework for component analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1041–1055, June 2012.

[40] F. Han and H. Liu, "High dimensional semiparametric scale-invariant principal component analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 2016–2032, Oct 2014.

[41] H. Chen and B. Zeng, "New transforms tightly bounded by DCT and KLT," *IEEE Sig. Proc. Lett.*, vol. 19, no. 6, pp. 344–347, 2012.

[42] M. U. Torun and A. N. Akansu, "An efficient method to derive explicit KLT kernel for first-order autoregressive discrete process," *IEEE Transactions on Signal Processing*, vol. 61, no. 15, pp. 3944–3953, 2013.

[43] J. V. Stone, *Independent component analysis: a tutorial introduction*. MIT Press, 2004.

[44] H. Nguyen and R. Zheng, "Binary independent component analysis with or mixtures," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3168–3181, July 2011.

[45] K.-C. Kwak and W. Pedrycz, "Face recognition using an enhanced independent component analysis approach," *IEEE Transactions on Neural Networks*, vol. 18, no. 2, pp. 530–541, March 2007.

[46] I. Santamaria, "Handbook of blind source separation: Independent component analysis and applications (common, p. and jutten, ; 2010 [book review]," *IEEE Signal Processing Magazine*, vol. 30, no. 2, pp. 133–134, March 2013.

[47] L. R. Arnaut and C. S. Obiekezie, "Source separation for wideband energy emissions using complex independent component analysis," *IEEE Transactions on Electromagnetic Compatibility*, vol. 56, no. 3, pp. 559–570, June 2014.

[48] B. Schölkopf, A. Smola, and K.-R. Müller, *Kernel principal component analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 583–588. [Online]. Available: http://dx.doi.org/10.1007/BFb0020217

[49] C. Varon, C. Alzate, and J. A. K. Suykens, "Noise level estimation for model selection in kernel PCA denoising," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 11, pp. 2650–2663, Nov 2015.

[50] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2002.

[51] Y. Xiao, Z. Zhu, Y. Zhao, Y. Wei, and S. Wei, "Kernel reconstruction ICA for sparse representation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 6, pp. 1222–1232, June 2015.

[52] Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1279–1289, June 2016.

[53] Y. Yuan, J. Wan, and Q. Wang, "Congested scene classification via efficient unsupervised feature learning and density estimation," *Pattern Recognition*, vol. 56, pp. 159–169, 2016.

[54] I. R. James and J. E. Mosimann, "A new characterization of the Dirichlet distribution through neutrality," *The Annals of Statistics*, vol. 8, no. 1, pp. 183–189, 1980.

[55] R. K. S. Hankin, "A generalization of the Dirichlet distribution," *Journal of Statistical Software*, vol. 33, no. 11, pp. 1–18, 2010.

[56] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, vol. 1, pp. 121–144, 2005.

[57] P. Orbanz, "Construction of nonparametric Bayesian models from parametric Bayes equations," in *Advances in Neural Information Processing Systems*, 2010.

[58] E. O. Brigham, *The Fast Fourier Transfor*. Prentice-Hall, 2002.

[59] B. A. Frigyik, A. Kapila, and M. R. Gupta, "Introduction to the Dirichlet distribution and related processes," Department of Electrical Engineering, University of Washington, Tech. Rep., 2010.

[60] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing independence by correlation of distances," *Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[61] G. J. Székely and M. L. Rizzo, "Brownian distance covariance," *Annals of Applied Statistics*, vol. 3, no. 4, pp. 1233–1303, 2009.

[62] K. Pearson, "Notes on regression and inheritance in the case of two parents," in *Proceedings of the Royal Society of London*, 1895, pp. 240–242.

[63] R. R. Wilcox, *Introduction to robust estimation and hypothesis testing*. Academic Press, 2005.

[64] G. J. Székely and M. L. Rizzo, "On the uniqueness of distance covariance," *Statistics & Probability Letters*, vol. 82, no. 12, pp. 2278–2282, 2012.

[65] T. Minka, "The lightspeed matlab toolboxes." [Online]. Available: http://research.microsoft.com/en-us/um/people/minka/software/lightspeed/

[66] F. T. Luk and S. Qiao, *A fast singular value algorithm for Hankel matrices*. Boston, MA, USA: American Mathematical Society, 2003, pp. 169–177.

[67] S. Shwartz, M. Zibulevsky, and Y. Y. Schechner, *ICA Using Kernel Entropy Estimation with NlogN Complexity*. Springer Berlin Heidelberg, 2004, ch. Independent Component Analysis and Blind Signal Separation, pp. 422–429.

[68] V. Laparra, G. Camps-Valls, and J. Malo, "Iterative Gaussianization: From ICA to random rotations," *IEEE Transactions on Neural Networks*, vol. 22, no. 4, pp. 537–549, April 2011.

[69] W. B. Kleijn, *A basis for source coding*, 2010, KTH lecture notes.

[70] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1991.

[71] W. R. Gardner and B. D. Rao, "Theoretical analysis of the high-rate vector quantization of LPC parameters," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 367 –381, sep 1995.

[72] Z. Ma and A. Leijon, "Modeling speech line spectral frequencies with Dirichlet mixture models," in *Proceedings of Interspeech*, 2010.

[73] Y. Lee, W. Jung, and M. Y. Kim, "GMM-based KLT-domain switched-split vector quantization for LSF coding," *IEEE Signal Processing Letters*, vol. 18, no. 7, pp. 415–418, July 2011.

[74] L. Wang, Z. Chen, and F. Yin, "A novel hierarchical decomposition vector quantization method for high-order LPC parameters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 212–221, Jan 2015.

[75] S. Chatterjee and T. V. Sreenivas, "Predicting VQ performance bound for LSF coding," *IEEE Signal Processing Letters*, vol. 15, pp. 166 –169, 2008.

[76] M. A. Ramirez, "Intra-predictive switched split vector quantization of speech spectra," *IEEE Signal Processing Letters*, vol. 20, no. 8, pp. 791–794, Aug 2013.

[77] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[78] Y. H. Kim and A. Ortega, "Quantizer design for energy-based source localization in sensor networks," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5577–5588, Nov 2011.

[79] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, Jan. 1980.

[80] Y. Koren, I. Yavneh, and A. Spira, "A multigrid approach to the scalar quantization problem," *IEEE Transactions on Information Theory*, vol. 51, no. 8, pp. 2993–2998, Aug 2005.

[81] S. Chatterjee and T. V. Sreenivas, "Low complexity wideband LSF quantization using GMM of uncorrelated Gaussian mixtures," in *16th European Signal Processing Conference (EUSIPCO)*, 2008.

[82] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 4, no. 2, p. R1, 2007.

[83] H. Cecotti and A. Graser, "Convolutional neural networks for P300 detection with application to brain-computer interfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 433–445, March 2011.

[84] A. Subasi, "EEG signal classification using wavelet feature extraction and a mixture of expert model," *Expert Systems with Applications*, vol. 32, no. 4, pp. 1084 – 1093, 2007.

[85] Z. Ma, Z.-H. Tan, and S. Prasad, "EEG signal classification with super-dirichlet mixture model," in *Proceedings of IEEE Statistical Signal Processing Workshop*, Aug. 2012, pp. 440 – 443.

[86] G. Xu, J. Han, Y. Zou, and X. Zeng, "A 1.5-D multi-channel EEG compression algorithm based on NLSPIHT," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1118–1122, Aug 2015.

[87] "BCI competition III," http://www.bbci.de/competition/iii.

[88] Z. Ma, Z.-H. Tan, and J. Guo, "Feature selection for neutral vector in EEG signal classification," *NEUROCOMPUTING*, vol. 174, pp. 937–945, Jan. 2016.

[89] T. N. Lal, M. Schroder, T. Hinterberger, J.Weston, M. Bogdan, N. Birbaumer, and B. Scholkopf, "Support vector channel selection in BCI," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1003–1010, Jun. 2004.

[90] H.-I. Suk and S.-W. Lee, "A novel Bayesian framework for discriminative feature extraction in brain-computer interfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 286–299, Feb 2013.

[91] "DARPA-TIMIT," *Acoustic-phonetic continuous speech corpus, NIST Speech Disc 1.1-1*, 1990.

[92] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, 1995.

[93] X. Huang, L. Shi, and J. A. K. Suykens, "Support vector machine classifier with pinball loss," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 984–997, May 2014.

[94] S. Nan, L. Sun, B. Chen, Z. Lin, and K. A. Toh, "Density-dependent quantized least squares support vector machine for large data sets," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–13, 2015, accepted and to appear.

[95] C.-C. Chang and C.-J. Lin., "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 17, pp. 1–27, 2011.

Zhanyu Ma has been an Associate Professor at Beijing University of Posts and Telecommunications, Beijing, China, since 2014. He is also an adjunct Associate Professor at Aalborg University, Aalborg, Denmark, since 2015. He received his Ph.D. degree in Electrical Engineering from KTH (Royal Institute of Technology), Sweden, in 2011. From 2012 to 2013, he has been a Postdoctoral research fellow in the School of Electrical Engineering, KTH, Sweden. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in multimedia signal processing, data mining, biomedical signal processing, and bioinformatics.

Jing-Hao Xue received the Dr. Eng. degree in signal and information processing from Tsinghua University in 1998 and the Ph.D. degree in statistics from the University of Glasgow in 2008. Since 2008 he has worked in the Department of Statistical Science at University College London as a Lecturer and Senior Lecturer. His current research interests include statistical classification, high-dimensional data analysis, computer vision, and pattern recognition.

Arne Leijon is a Professor in Hearing Technology at the KTH (Royal Inst of Technology) Sound and Image Processing Lab, Stockholm, Sweden, since 1994. His main research interest concerns applied signal processing in aids for people with hearing impairment, and methods for individual fitting of these aids, based on psychoacoustic modelling of sensory information transmission and subjective sound quality. He received the M. S. degree in Engineering Physics in 1971, and a Ph.D. degree in Information Theory in 1989, both from Chalmers University of Technology, Gothenburg, Sweden.

Zheng-Hua Tan is an Associate Professor in the Department of Electronic Systems at Aalborg University, Aalborg, Denmark,since May 2001. His research interests include speech and speaker recognition, noise-robust speech processing, multimedia signal and information processing, human-robot interaction, and machine learning. He has served as an Editorial Board Member/Associate Editor for Elsevier Computer Speech and Language, Elsevier Digital Signal Processing and Elsevier Computers and Electrical Engineering. He was a Lead Guest Editor for the IEEE Journal of Selected Topics in Signal Processing.

Zhen Yang received the PhD degree in signal processing from the Beijing University of Posts and Telecommunications. He is a associate professor of computer science and engineering at Beijing University of Technology. His research interests include data mining, machine learning, trusted computing, and content security. He is also a senior Member of the Chinese Institute of Electronics and a member of the IEEE.

Jun Guo received B.E. and M.E. degrees from Beijing University of Posts and Telecommunications (BUPT), China in 1982 and 1985, respectively, Ph.D. degree from the Tohuku-Gakuin University, Japan in 1993. At present he is a professor and a vice president of BUPT. His research interests include pattern recognition theory and application, information retrieval, content based information security, and bioinformatics. He has published over 200 papers on the journals and conferences including SCIENCE, Nature Scientific Reports, IEEE Trans. on PAMI, Pattern Recognition, AAAI, CVPR, ICCV, SIGIR, etc.