An Information-theoretic Approach to Unsupervised Feature Selection for High-Dimensional Data

Shao-Lun Huang, Lin Zhang
DSIT Research Center
Tsinghua-Berkeley Shenzhen Institute
Shenzhen, China 518055
Email: {shaolun.huang, linzhang}@sz.tsinghua.edu.cn

Lizhong Zheng
Dep. of Electrical & Computer Eng.
Massachusetts Institute of Technology
Cambridge, MA 02139-4307
Email: lizhong@mit.edu

***Abstract*—In this paper, we model the unsupervised learning of a sequence of observed data vector as a problem of extracting joint patterns among random variables. In particular, we formulate an information-theoretic problem to extract common features of random variables by measuring the loss of total correlation given the feature. This problem can be solved by a local geometric approach, where the solutions can be represented as singular vectors of some matrices related to the pairwise distributions of the data. In addition, we illustrate how these solutions can be transferred to feature functions in machine learning, which can be computed by efficient algorithms from data vectors. Moreover, we present a generalization of the HGR maximal correlation based on these feature functions, which can be viewed as a nonlinear generalization to linear PCA. Finally, the simulation result shows that our extracted feature functions have great performance in real-world problems.**

I. Introduction

In unsupervised learning, it is assumed that a sequence of d-dimensional data vectors $\underline{x}^{(m)} = (x_1^{(m)}, \ldots, x_d^{(m)})$, for $m = 1, \ldots, n$, is observed. These data vectors are often statistically modelled as i.i.d. generated from an unknown joint distribution $P_{X_1 \cdots X_d}$ for some jointly distributed random variables $X_1 \cdots X_d$, and we want to learn features or patterns directly from these data vectors. In this paper, we would like to assume that there exists an unknown structure, modelled as a random variable W, such that the random variables $X_1, \ldots, X_d$ are conditionally independent outputs from W, i.e., $P_{X_1 \cdots X_d|W} = \Pi_{i=1}^d P_{X_i|W}$. Figure 1 illustrates this model. Then, our goal is to learn features of the hidden structure W directly from the observed data vectors $\underline{x}^{(m)}$ without prior knowledge of W. Note that Figure 1 can model a wide range of unsupervised learning problems. For example, in unsupervised image clustering, the hidden structure can be the collection of hidden features, such as the gender or age, and X_i can be subareas of the images.

It turns out that the main difficulty here is that there is no prior knowledge about W, and we need to design a good information criterion to select features that are more likely to describe W very well. For that, observe that W can be viewed as a sort of "common information" shared by random variables $X_1, \ldots, X_d$, and learning the features of W can be formulated as extracting common features among these random variables. Motivated by this observation, in this paper we measure the among of common information between random variables by the total correlation [1], and formulate an information-theoretic problem to select the feature U from the data variables to maximize the loss of total correlation given the feature:

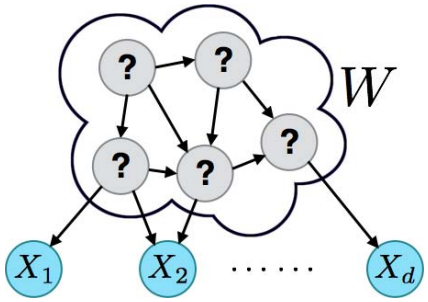

Fig. 1. The random variables $X_1, \ldots, X_d$ are conditional independently generated from some hidden structure W.

$$\max_{\substack{P_{UX_1\cdots X_d}:\\ I(U;X_1,\ldots,X_d)\le \frac{1}{2}\epsilon^2}} \quad C(X_1,\ldots,X_d) - C(X_1,\ldots,X_d|U) \quad (1)$$

where $C(X_1, \ldots, X_d) \triangleq D(P_{X_1 \cdots X_d} \| P_{X_1} \cdots P_{X_d})$ is the total correlation. In this paper, we are particularly interested in solving (1) in the small ϵ region, which allows us to focus on the most significant low rate feature that are often useful in practice. Moreover, this formulation reveals several important advantages.

Firstly, the problem (1) in the small ϵ region can be solved by a local geometric approach [2]- [4] with a clean analytical form. In addition, the solution can be represented by singular vectors of a matrix B, whose entries are weighted paiwise joint distributions between $X_1, \ldots, X_d$. In addition, the sufficient statistic for estimating the optimal U of (1) from $X_1, \ldots, X_d$ can be adopted as a feature function, which extracts all the relevant information of an observed data vector contained about U. Moreover, this feature function can be computed by an efficient algorithm from data vectors directly. Finally, the feature functions from our approach can be interpreted as a generalization of the well-known Hirschfeld-Gebelein-Rényi (HGR) maximal correlation [5] to multiple random variables. We also show that this can be viewed as a nonlinear generalization to linear principle component analysis (PCA). This offers critical insights between information theory, statistics, and machine learning, and the local geometric approach is precisely the key technique to draw these connections. In the rest of this paper, we demonstrate these results in detail, and show the application of the feature functions to a real-world problem.

978-1-5090-3097-2/17/$31.00 ©2017 IEEE

Related works : The local geometric approach employed in this paper was first introduced in [2] for solving communication problem. In addition, the authors in [3] [4] extended this approach to study the learning problem for a pair of random variables, where our work in this paper can be viewed as the generalization of this framework to multiple random variables. Moreover, the idea of applying the reduction of total correlation as a learning criterion was also observed in [6], where the authors solved an optimization problem by restricting the cardinality of U, and an rather complicated iterative algorithm was derived. On the other hand, in this paper we restrict the information volume contained in U, which is a much more natural constrain in information theory, and obtain clean analytical solutions that can be computed by a simple and efficient algorithm.

Notations : Throughout this paper, we use X, $\mathcal{X}$, $|\mathcal{X}|$ and x to denote the random variable, range, cardinality, and its value. For a matrix B, $B(i;j)$ denotes the entry of B in the i-th row and j-th column. Finally, we use $\sqrt{P_X}$ to denote an $|\mathcal{X}|$-dimensional vector with entries $\sqrt{P_X(x)}$, for all $x \in \mathcal{X}$.

II. The Local Geometric Approach

We commence by applying a local geometric technique [2] to solve (1). For this purpose, we make a subtle assumption that $\max_u P_U(u)/\min_u P_U(u) < \gamma$, for some finite $\gamma > 0$ irrelevant to ϵ. This assumption is natural to many practical problems. With this assumption, the constraint $I(U;X_1,\ldots,X_d) \leq \frac{1}{2}\epsilon^2$ for small ϵ implies that the conditional distribution $P_{X_1\cdots X_d|U}$ can be written as a perturbation to the marginal distribution[1]:

$$P_{X_1\cdots X_d|U=u}(x_1,\ldots,x_d) = \tag{2}$$
$$P_{X_1\cdots X_d}(x_1,\ldots,x_d) + \epsilon\sqrt{P_{X_1\cdots X_d}(x_1,\ldots,x_d)}\phi_u(x_1,\ldots,x_d)$$

where ϕ_u can be viewed as an $|\mathcal{X}_1|\cdot|\mathcal{X}_2|\cdots|\mathcal{X}_d|$ dimensional vector. Then, the mutual information can be expressed as

$$I(U;X_1,\ldots,X_d) = \frac{1}{2}\epsilon^2\mathbb{E}_U[\|\phi_U\|^2] + o(\epsilon^2),$$

where $\|\cdot\|$ denotes the l_2-norm. Thus, by ignoring the higher order term of ϵ as we are interested in the small ϵ region, the constraint $I(U;X_1,\ldots,X_d) \leq \frac{1}{2}\epsilon^2$ can be reduced to

$$\mathbb{E}_U[\|\phi_U\|^2] \leq 1.$$

In addition, the objective function of (1) can also be expressed in terms of mutual informations:

$$\begin{aligned} &C(X_1,\ldots,X_d) - C(X_1,\ldots,X_d|U) \\ &= \sum_{i=1}^{d} I(U;X_i) - I(U;X_1,\ldots,X_d), \end{aligned} \tag{3}$$

and for each i, the mutual information can be again approximated as the l_2-norm square

$$I(U;X_i) = \frac{1}{2}\epsilon^2\mathbb{E}_U[\|\psi_{i,U}\|^2] + o(\epsilon^2),$$

where $\psi_{i,U}$ is the $|\mathcal{X}_i|$-dimensional perturbation vector defined as

$$\psi_{i,u}(x_i) = \frac{P_{X_i|U=u}(x_i) - P_{X_i}(x_i)}{\epsilon\sqrt{P_{X_i}(x_i)}} \tag{4}$$

Then, the optimization problem (1), by ignoring the higher order terms of ϵ, can be transferred to a linear algebraic problem

$$\max_{\mathbb{E}_U[\|\phi_U\|^2]\leq 1} \sum_{i=1}^{d} \mathbb{E}_U[\|\psi_{i,U}\|^2]. \tag{5}$$

To solve (5), observe that P_{X_i} and $P_{X_i|U}$ are marginal distributions of $P_{X_1\cdots X_d}$ and $P_{X_1\cdots X_d|U}$, thus there is a correlation between ϕ_U and $\psi_{i,U}$:

$$\psi_{i,u}(x_i) = \sum_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d} \frac{\sqrt{P_{X_1\cdots X_d}(x_1,\ldots,x_d)}}{\sqrt{P_{X_i}(x_i)}}\phi_u(x_1,\ldots,x_d)$$

which can be represented in matrix form as $\psi_{i,u} = B_i\cdot\phi_u$, where B_i is an $|\mathcal{X}_i| \times (|\mathcal{X}_1|\cdot|\mathcal{X}_2|\cdots|\mathcal{X}_d|)$ matrix with entries

$$B_i(\hat{x}_i;(x_1,\ldots,x_n)) = \begin{cases} \frac{\sqrt{P_{X_1\cdots X_n}(x_1,\ldots,x_n)}}{\sqrt{P_{X_i}(\hat{x}_i)}} & \text{if } \hat{x}_i = x_i, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, if we define an $(|\mathcal{X}_1|+\cdots+|\mathcal{X}_m|)\times(|\mathcal{X}_1|\cdots|\mathcal{X}_m|)$-dimensional matrix $B_0 \triangleq [B_1^T\cdots B_d^T]^T$, then (5) can be written as

$$\max_{\mathbb{E}_U[\|\phi_U\|^2]\leq 1}\sum_{i=1}^{d}\mathbb{E}_U[\|B_i\cdot\phi_U\|^2] = \max_{\mathbb{E}_U[\|\phi_U\|^2]\leq 1}\mathbb{E}_U[\|B_0\cdot\phi_U\|^2] \tag{6}$$

Moreover, since ϕ_U is a perturbation vector of probability distributions, by summing over all $x_1,\ldots,x_d$ for both sides of (2), it has to satisfy an extra constraint

$$\sum_{x_1,\ldots,x_d}\sqrt{P_{X_1\cdots X_d}(x_1,\ldots,x_d)}\phi_u(x_1,\ldots,x_d) = 0,$$

which implies that ϕ_U is orthogonal to the vector $u_0 = \sqrt{P_{X_1\cdots X_d}}$. In particular, it is shown in [2] that u_0 is the right singular vector of B_0 with the largest singular value $\sigma_0 = \sqrt{n}$, and the corresponding left singular vector is $v_0 = \frac{1}{\sqrt{n}}\left[\sqrt{P_{X_1}}^T\cdots\sqrt{P_{X_d}}^T\right]^T$. Thus, the optimal solution of (6) is to align the vectors $\phi_{U=u}$, for all u, along the second largest right singular vector of B_0. It turns out that it is easier to compute the second largest left singular vector of B instead of the right one, since the left singular vector has much smaller dimensionality. This is equivalent to compute the second largest eigenvector of the matrix $B \triangleq B_0B_0^T$, which by definition can be written as

$$B = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1d} \\ B_{21} & B_{22} & \cdots & B_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ B_{d1} & B_{d2} & \cdots & B_{dd} \end{bmatrix} \tag{7}$$

where $B_{ij} \triangleq B_iB_j^T$ are $|\mathcal{X}_i|\times|\mathcal{X}_j|$-dimensional matrices with

[1]Note that the constraint $I(U;X_1,\ldots,X_d) \leq \epsilon^2/2$ itself does not imply that the conditional distribution has a perturbation form. See [7] for the details.

entries

$$B_{ij}(x_i;x_j)=\frac{P_{X_iX_j}(x_i,x_j)}{\sqrt{P_{X_i}(x_i)}\sqrt{P_{X_j}(x_j)}}$$

for $i\neq j$, and for each i, B_{ii} is an identity matrix. The above discussions are summarized as the following Theorem.

Theorem 1. *Let $\lambda^{(1)}$ and $\psi^{(1)}$ be the second largest eigenvalue and eigenvector of B, then*

$$\lambda^{(1)}=\lim_{\epsilon\to 0}\frac{2}{\epsilon^2}\max_{I(U;X_1,\ldots,X_d)\leq\frac{1}{2}\epsilon^2}\sum_{i=1}^{d}I(U;X_i),$$

which from (3) *implies that the optimum of* (1) *is $\lambda^{(1)}-1$. Moreover, let the optimal solution of* (1) *be $P^{(\epsilon)}_{X_1\cdots X_d|U}$, then*

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\frac{P^{(\epsilon)}_{X_1\cdots X_d|U}-P_{X_1\cdots X_d}}{\sqrt{P_{X_1\cdots X_d}}}\propto B_0^T\psi^{(1)},$$

where "$\propto$" denotes two vectors are aligned.

We will show in section III-A that the second largest eigenvector of BB^T can be computed efficiently by a modified ACE algorithm. Finally, we highlight a property of $\psi^{(1)}$ that will be used later.

Lemma 1. *Let $\psi^{(1)}=[\psi_1^T\cdots\psi_d^T]^T$, where each ψ_i is an $|\mathcal{X}_i|$-dimensional vector, then ψ_i is orthogonal to $\underline{P_{X_i}}$.*

Proof: Let $\phi^{(1)}$ be the second largest right singular vector of B_0, then $\psi_i=B_0\phi^{(1)}$. Since $\phi^{(1)}$ is a perturbation vector defined in (2), ψ_i is also a perturbation vector defined in (4), which implies its orthogonality to $\underline{P_{X_i}}$. ∎

III. Feature Extraction From Data Vectors

We shall now transfer the information-theoretic results obtained in section II to learning problems. Remember that our goal is to extract features from a sequence of observed data vectors $\underline{x}^{(m)}$ about the hidden common structure W, and the features are often represented by functions of the data $\underline{x}^{(m)}$ in machine learning. While the results in section II tells us the statistical relationship between the targeted common feature U and the data variables, it remains to transfer this knowledge to a functional representation of data vectors. This can typically be carried out by considering the sufficient statistic of inferring U via random variables $X_1,\ldots,X_d$, which is the log-likelihood

$$\log\frac{P_{X_1\cdots X_d|U}}{P_{X_1\cdots X_d}}\simeq\frac{P_{X_1\cdots X_d|U}-P_{X_1\cdots X_d}}{P_{X_1\cdots X_d}}\propto\frac{\phi^{(1)}}{\sqrt{P_{X_1\cdots X_d}}}$$

where $\phi^{(1)}$ is the second largest right singular vector of B. Here, we approximate the log-likelihood to the first order term of ϵ, and note that all the vectors ϕ_U should be aligned to $\phi^{(1)}$. This motivates us to define the feature function $f:\mathcal{X}_1\times\cdots\times\mathcal{X}_d\mapsto\mathbb{R}$ as

$$f(x_1,\ldots,x_d)\triangleq\frac{\phi^{(1)}(x_1,\ldots,x_d)}{\sqrt{P_{X_1\cdots X_d}(x_1,\ldots,x_d)}},\tag{8}$$

which is the functional representation of the data vectors that extracts all the relevant information about the target U. It is easy to verify that f is a zero-mean and unit-variance function.

A. The Algorithm to Compute Feature Functions

The feature function (8) is a high-dimensional function exponential to the number of random variables d; however, this function can be computed efficiently via the analyses in section II. The key step is to note that $\sqrt{\lambda^{(1)}}\phi^{(1)}=B_0^T\psi^{(1)}=\sum_{i=1}^{d}B_i^T\psi_i$, where $\psi^{(1)}=[\psi_1^T\cdots\psi_d^T]^T$. Thus, if we define functions $f_i:\mathcal{X}_i\mapsto\mathbb{R}$ as $f_i(x_i)=\psi_i(x_i)/\sqrt{P_{X_i}(x_i)}$, then it is easy to verify that

$$\sqrt{\lambda^{(1)}}\phi^{(1)}(x_1,\ldots,x_d)=\sqrt{P_{X_1\cdots X_d}(x_1,\ldots,x_d)}\sum_{i=1}^{d}f_i(x_i),$$

which implies

$$\sqrt{\lambda^{(1)}}f(x_1,\ldots,x_d)=\sum_{i=1}^{d}f_i(x_i).\tag{9}$$

Since $\sqrt{\lambda^{(1)}}$ is simply a normalization factor, we simply need to compute the functions $\vec{f}=(f_1,\ldots,f_d)$, which is equivalent to compute the second largest eigenvector $\psi^{(1)}$ of B.

Remark 1. From (9), we know that the feature function to infer the most informative target U about common information of $X_1,\ldots,X_d$ has an additive structure, i.e., it can be written as the sum of individual functions of each X_i. Note that the result of additive structure comes from applying the local geometric approach to solve the information theoretic problem (1) over general high-dimensional perturbation vectors ψ. This is particular attractive since we do not need to make any assumption on how the target U is embedded in $X_1,\ldots,X_d$, and the structure of the perturbation vector ψ in (2), but can still obtain the additive structure of feature functions. This demonstrates the critical role of the local geometric approach in applying information theory to machine learning.

To derive an algorithm to compute the eigenvector $\psi^{(1)}$ from observed data vectors $\underline{x}^{(m)}$, an intuitive way is to estimate the empirical distribution between $X_1,\ldots,X_d$ from data samples, and construct the matrix B to solve the eigen-decomposition. However, this is often not feasible in practice due to: (1) there may not be enough number of samples to estimate the joint distribution accurately, (2) the dimensionality of B may be extremely high especially for big data applications, so that the SVD can not be conducted directly. Alternatively, it is well-known that eigenvectors of a matrix can be efficiently computed by the well-known power method, which iterative multiplies the matrix to an initial vector, and converges to the largest eigenvector exponentially fast. To apply the power method for computing the second largest singular vector of B, from Lemma 1, we choose the initial vector $\psi=[\psi_1\cdots\psi_d]^T$, such that ψ_i is orthogonal to $\underline{P_{X_i}}$. This also forces ψ to be orthogonal to $\underline{\sqrt{P_{X_1\cdots X_d}}}$, which guarantees the convergence to the second largest eigenvalue. Then, the algorithm iteratively compute the matrix multiplication $\psi\leftarrow B\psi$, or equivalently

$$\psi_i\leftarrow\psi_i+\sum_{j\neq i}B_{ij}\psi_j,\tag{10}$$

for all i. Note that if we write $f_i(x_i)=\psi_i(x_i)/\sqrt{P_{X_i}(x_i)}$, then as shown in [3], the step (10) is mathematically equivalent to a conditional expectation operation on functions:

$$f_i(X_i)\leftarrow f_i(X_i)+\mathbb{E}\left[\sum_{j\neq i}f_j(X_j)\middle|X_i\right],$$

Therefore, the power method can be transferred to an algorithm

based on the alternative conditional expectation (ACE) [8] as shown in Algorithm 1, which computes the optimal feature functions for (9).

Algorithm 1 The modified ACE Algorithm

Require : The data samples of variables $X_1, \ldots, X_d$

1. Initialization: randomly pick zero-mean functions $\vec{f} = (f_1, \ldots, f_d)$.

repeat :

2. $f_i(X_i) \leftarrow f_i(X_i) + \mathbb{E}\left[\sum_{j\neq i} f_j(X_j)\middle| X_i\right]$.
3. $f_i(X_i) \leftarrow f_i(X_i)/\sqrt{\mathbb{E}\left[\sum_{i=1}^{d} f_i^2(X_i)\right]}$.

until $\vec{f}$ converges.

B. Computing Multiple Feature Functions

While the above discussions focuses on the second largest singular vector of B and the corresponding feature function, it is clear that one can also compute the rest eigenvectors and feature functions. It turns out that these feature functions demonstrates an optimal tradeoff between the number of selected feature functions, and the amount of information extracted about a targeted hidden variable [3]. Like the eigenvectors, the computation of multiple feature functions can be implemented in a successive manner. After the first $k-1$ sets of feature functions $\vec{f}^{(i)} = (f_1^{(i)}, \ldots, f_d^{(i)})$, for $i = 1, \ldots, k-1$, is computed, the k-th set of feature functions $\vec{f}^{(k)}$ has to be orthogonal to previous feature functions:

$$\langle \vec{f}^{(m)}, \vec{f}^{(k)} \rangle \triangleq \sum_{i=1}^{d} \mathbb{E}\left[f_i^{(m)}(X_i) f_i^{(k)}(X_i)\right] = 0, \text{ for } m \leq k-1$$

Therefore, $\vec{f}^{(k)}$ can be computed the same as Algorithm 1 but with an extra step of Gram-Schmidt procedure to guarantee the orthogonality, which is illustrated in Algorithm 2.

Algorithm 2 The Computation of $\vec{f}^{(k)}$

Require : The data samples of variables $X_1, \ldots, X_n$, and previously computed functions $\vec{f}^{(1)}, \ldots, \vec{f}^{(k-1)}$.

1. Initialization: randomly pick zero-mean functions $\vec{f}^{(k)} = (f_1^{(k)}, \ldots, f_d^{(k)})$.

repeat :

2. Run step 2 and 3 of Algorithm 1.
3. $\vec{f}^{(k)} \leftarrow \vec{f}^{(k)} - \sum_{m=1}^{k-1} \langle \vec{f}^{(m)}, \vec{f}^{(k)} \rangle \cdot \vec{f}^{(m)}$

until $\vec{f}^{(k)}$ converges.

IV. The Generalized Maximal Correlation

The HGR maximal correlation is a variational generalization of the well-known Pearson correlation coefficient, and was originally introduced as a normalized measure of the dependence between two random variables [5].

Definition 1 (Maximal Correlation)**.** For jointly distributed random variables X and Y, with ranges $\mathcal{X}$ and $\mathcal{Y}$ respectively, the maximal correlation between X and Y is defined as:

$$\rho(X;Y) \triangleq \sup_{\substack{f:\mathcal{X}\to\mathbb{R},\ g:\mathcal{Y}\to\mathbb{R}\ :\\ \mathbb{E}[f(X)]=\mathbb{E}[g(Y)]=0\\ \mathbb{E}[f^2(X)]=\mathbb{E}[g^2(Y)]=1}} \mathbb{E}\left[f(X)g(Y)\right]$$

where the supremum is taken over all Borel measurable functions. Furthermore, if X or Y is a constant almost surely, there exist no functions f and g which satisfy the constraints, and we define $\rho(X;Y) = 0$.

It is easily verified that $0 \leq \rho(X;Y) \leq 1$, and $\rho(X;Y) = 0$ if and only if X is independent of Y. In this section, we propose a generalization of HGR maximal correlation to multiple random variables, based on the feature functions in section III.

Definition 2. The generalized maximal correlation (GMC) for jointly distributed random variables $X_1, \ldots, X_d$ with ranges $\mathcal{X}_i$, for $i = 1, \ldots, d$, is defined as

$$\rho^*(X_1, \cdots, X_d) \triangleq \max \frac{1}{d-1} \mathbb{E}\left[\sum_{i\neq j} f_i(X_i) f_j(X_j)\right] \quad (11)$$

for the functions $f_i : \mathcal{X}_i \to \mathbb{R}$, with the constraints

$$\mathbb{E}\left[f_i(X_i)\right] = 0,\ \mathbb{E}\left[\sum_{i=1}^{d} f_i^2(X_i)\right] = 1, \text{ for all } i. \quad (12)$$

Proposition 1. *The optimal functions of* (11) *can be computed by Algorithm 1.*

Proof: Let f_i^* be the functions optimizing (11), and $\psi_i(x_i) = \sqrt{P_{X_i}(x_i)} f_i^*(x_i)$, then it is easy to verify that the vector $\psi = [\psi_1^T \cdots \psi_d^T]^T$ is the second largest eigenvector of B. ∎

As the HGR maximal correlation, the GMC satisfies some fundamental properties for correlation measurements, where the proofs of these properties are straightforward by definition.

Property 1. *For jointly distributed random variables $X_1, \ldots, X_d$, the GMC satisfies $\rho^*(X_1, \ldots, X_d) \leq 1$, and for $d \geq 3$, the equality holds if and only if there exists functions $f_i(X_i)$, such that for all i, j, $f_i(X_i) = f_j(X_j)$ with probability* 1.

Property 2. *For random variables $X_1, \ldots, X_d$, the GMC $\rho^*(X_1, \ldots, X_d) = 0$ if and only if the random variables are pairwise independent.*

Property 3. *For $d = 2$, the GMC reduces to the maximal correlation, i.e., $\rho^*(X;Y) = \rho(X,Y)$.*

It turns out that GMC is a nonlinear generalization of the linear PCA [9]. To see that, consider a sequence of data vectors $\underline{x}^{(m)} = (x_1^{(m)}, \ldots, x_d^{(m)}) \in \mathbb{R}^d$, for $m = 1, \ldots, n$, where the sample mean and variance for each dimension are zero and one, i.e., $\sum_{m=1}^{n} x_i^{(m)} = 0$, and $\frac{1}{n}\sum_{m=1}^{n} (x_i^{(m)})^2 = 1$, for all i. Then, the PCA aims to find the principle vector $\underline{w} = (w_1, \ldots, w_d)$ with unit norm such that $\sum_{m=1}^{n} \langle \underline{w}, \underline{x}^{(m)} \rangle^2$ is maximized; or equivalently, to maximize

$$\sum_{m=1}^{n} \sum_{i\neq j} \left(w_i x_i^{(m)}\right)\left(w_j x_j^{(m)}\right) = \mathbb{E}\left[\sum_{i\neq j} (w_i X_i) \cdot (w_j X_j)\right] \quad (13)$$

subject to the constraint

$$1 = \sum_{i=1}^{d} w_i^2 = \sum_{i=1}^{d} \mathbb{E}\left[(w_i X_i)^2\right], \quad (14)$$

where the expectations in (13) and (14) are over the empirical distributions $P_{X_i X_j}$ and P_{X_i} from the data vectors. Comparing to the definition 2, we can see that GMC generalizes the linear

PCA to nonlinear functional spaces of data. We would like to emphasize that [3] also provides a nonlinear generalization to PCA for the Gaussian distributed data vectors by the local geometric approach. Our approach presented in this paper essentially offers another generalization for general discrete data vectors.

Remark 2. There are some other generalizations to maximal correlations to multiple random variables. For example, the network maximal correlation (NMC) proposed in [10] defines a correlation measurement the same as (11) but with a slightly different constraint:

$$\mathbb{E}\left[f_i(X_i)\right] = 0,\ \mathbb{E}\left[f_i^2(X_i)\right] = 1,\ \text{for all } i.$$

In addition, [11] proposes a maximally correlated principal component analysis, which considers the largest singular value of a matrix closely related to (7). Our results essentially offer the information theoretic justification of generalizing the maximal correlation as extracting common features among random variables, and it turns out that the local geometric approach is the key technique to obtain these insights.

V. Simulation Results

In this section, we verify the performance of the selected features from our framework to the MNIST Handwritten Digit Database [12] for digits recognition. In the MNIST database, there are $N = 60000$ images contained in the training sets, and each image has a label that represents the digits "0" to "9". The images in this database are consisted of 28×28 pixels, where each image pixel takes the value ranging from 0 to 255. While this is a supervised learning problem, we will show that Algorithm 2 can be applied to select features from images directly without the knowledge of labels, and these features, although selected in an unsupervised way, have good performance in digital recognition. To apply Algorithm 2, we need to identify the random variables X_i's in the MNIST problem as follows:

1. Each image pixel is quantized into binary signals "0" and "1" with the quantization threshold 40.
2. We divide each image into $8 \times 8 = 64$ overlapping subareas, where each sub-image has 6×6 pixels, and two nearby subareas are overlapped with 3 pixels. Figure 2 illustrates this division of images.

Moreover, we quantize each subarea by Hamming distance 3, and represent each subarea as a random variable X_i, for $i = 1, \dots 64$.

After this pre-processing, 64 random variables X_i are specified, and each image n can be viewed as a 64-dimensional data vector $(x_1^{(n)}, \dots, x_{64}^{(n)})$, for $n = 1, \dots, N$. Then, we apply Algorithm 2 to compute k feature functions $\vec{f}_i = (f_i^{(1)}, \dots, f_i^{(k)})$ for each random variable X_i. These feature functions maps the pre-processed training image n into a $64k$-dimensional score vector

$$\vec{s}_i = \left(\vec{f}_1(x_1^{(n)}), \dots, \vec{f}_{64}(x_{64}^{(n)})\right).$$

which extracts non-linear features of the image. Note that in this step, we select the feature functions only from the image pixels but without the knowledge of the labels.

With the score vectors computed, at the second step we apply the linear support vector machine (SVM) [13] to classify the vectors $\vec{s}_i$, for $i = 1, \dots, N$ into ten groups with respect to the labels z_i. This results in a linear classifier that associates a label $\hat{z}_i \in \{0, \dots, 9\}$ to each score vector $\vec{s}_i$, and the label represents the recognized digit of the image corresponding to the score vector. To test the performance of this linear classifier in the set of test images, we first conduct the same pre-processing to the test images, and map the pre-processed test images into $64k$-dimensional score vectors by the score functions $\vec{f}_i$. Then, the linear classifier is applied to recognize the digits in the test images. The error probability of recognizing the digits via the score vectors with $k = 24$ is 2.1%, which outperformances the convolutional neural network with 2 layers.

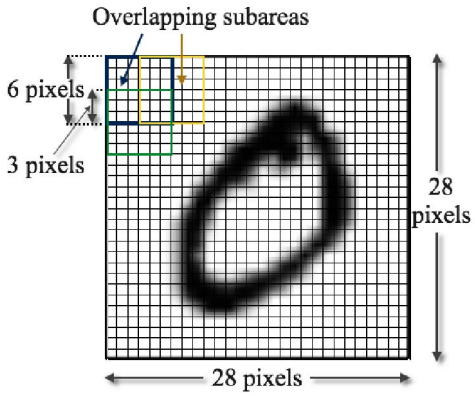

Fig. 2. The division of images into $8 \times 8 = 64$ overlapping subareas. Each subarea has 6×6 pixels, and nearby subareas overlap with 3 pixels.

References

[1] S. Watanabe, "Information Theoretical Analysis of Multivariate Correlation," *IBM Journal of Research and Development*, vol. 4, pp. 66–82, Jan. 1960.

[2] S.-L. Huang and L. Zheng, "Linear Information Coupling Problems," in *Proc. Int. Symp. Inform. Theory*, July 2012.

[3] S.-L. Huang, A. Makur, L. Zheng, and G. W. Wornell, "Universal feature for universal feature selection in high-dimensional inference: An information-theoretic framework," preprint, 2017.

[4] A. Makur, F. Kozynski, S.-L. Huang, and L. Zheng, "An efficient algorithm for information decomposition and extraction," in *Proc. Allerton Conf. Commun., Contr., Computing*, (Monticello, IL), Oct. 2015.

[5] A. Rényi, "On measures of dependence," Acta Mathematica Academiae Scientiarum Hungarica, vol. 10, no. 3-4, pp. 441–451, 1959.

[6] G. V. Steeg, A. Galstyan, "Discovering Structure in High-Dimensional Data Through Correlation Explanation," *Proceedings of the 27th International Conference on Neural Information Processing Systems,* pp. 577–585, Dec. 2014

[7] V. Anantharam, A. Gohari, S. Kamath, C. Nair, "On Maximal Correlation, Hypercontractivity, and the Data Processing Inequality studied by Erkip and Cover", arXiv: http://arxiv.org/abs/1304.6133

[8] L. Breiman, J. H. Friedman, "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association,* vol. 80, issue 391, 1985

[9] I. T. Jolliffe, *Principal Component Analysis*, Springer Verlag, 1986.

[10] S. Feizi, A. Makhdoumi, K. Duffy, M. Kellis, M. Médard, "Network Maximal Correlation," *Computer Science and Artificial Intelligence Laboratory Technical Report*, MIT, Cambridge.

[11] S. Feizi, D. Tse, "Maximally Correlated Principal Component Analysis", arXiv: https://arxiv.org/abs/1702.05471

[12] MNIST Handwritten Digit Database, from http://yann.lecun.com/exdb/mnist/

[13] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, pp.273 – 297, Sept. 1995.