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

Shao-Lun Huang, *Member, IEEE,* Xiangxiang Xu, *Student Member, IEEE,* and Lizhong Zheng, *Fellow, IEEE*

***Abstract*—In this paper, we propose an information-theoretic approach to design the functional representations to extract the hidden common structure shared by a set of random variables. The main idea is to measure the common information between the random variables by Watanabe's total correlation, and then find the hidden attributes of these random variables such that the common information is reduced the most given these attributes. We show that these attributes can be characterized by an exponential family specified by the eigen-decomposition of some pairwise joint distribution matrix. Then, we adopt the log-likelihood functions for estimating these attributes as the desired functional representations of the random variables, and show that such representations are informative to describe the common structure. Moreover, we design both the multivariate alternating conditional expectation (MACE) algorithm to compute the proposed functional representations for discrete data, and a novel neural network training approach for continuous or high-dimensional data. Furthermore, we show that our approach has deep connections to existing techniques, such as Hirschfeld-Gebelein-Rényi (HGR) maximal correlation, linear principal component analysis (PCA), and consistent functional map, which establishes insightful connections between information theory and machine learning. Finally, the performances of our algorithms are validated by numerical simulations.**

***Index Terms*—Alternating conditional expectations algorithm, information geometry, informative representation, total correlation, unsupervised learning**

I. Introduction

Given a set of d discrete random variables $X^d = (X_1, \ldots, X_d)$ with the (unknown) joint distribution P_{X^d}, and a sequence of observed sample vectors $\underline{x}^{(\ell)} = (x_1^{(\ell)}, \ldots, x_d^{(\ell)})$ i.i.d. generated from this joint distribution, for $\ell = 1, \ldots, n$, our goal in this paper is to efficiently and effectively extract the hidden common information structure (or simply called common structure) shared by these random variables from the observed sample vectors. This is a typical unsupervised learning problem, and such common structures can be useful in many machine learning scenarios. As a motivating example, in the MNIST digits recognition problem [1], we often divide the images into overlapping sub-images, such as in Fig. 1, and then train feature functions on the sub-images for learning the digits. In this problem, we can view each sub-image as a random variable X_i, and the training images as the observed data vectors. Since these sub-images are constructed from the written digits, the digit is the key common information shared by these sub-images. Therefore, effectively mining the information of the shared structure among these random variables can be helpful for recognizing the digits.

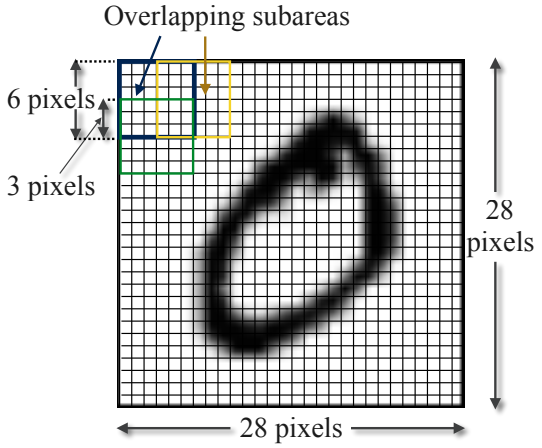

Fig. 1: The division of images into $8 \times 8 = 64$ overlapping subareas. Each subarea has 6×6 pixels, and nearby subareas overlap with 3 pixels.

In addition, the concept of extracting common structure shared by multiple random variables or objects has also appeared or implicitly posted in several disciplines. For instance, linear principal component analysis (PCA) [2], the most widely adopted unsupervised learning technique, can be viewed as resolving a principal direction that conveys the most common randomness among different dimensions of data vectors. In addition, the consistent functional map network [3]–[5], a recently proposed effective approach in computer vision, takes each X_i as a shape, and aims to find the shared components among different shapes. The main issue behind these problems is: how to design good low-dimensional functions of the random variables X^d, such that these functional representations are effective to reveal the common structure among these random variables. This can also be viewed as the unsupervised dimension reduction problem with the particular focus on extracting the common information of random variables. In this paper, our goal is to apply the ideas from information theory to design good algorithms for finding such useful functional representations.

Our approach can be delineated in the following steps. Firstly, we want to identify the targeted random variable U embedded in the random variables X^d with some joint

S.-L. Huang is with the Data Science and Information Technology Research Center, Tsinghua-Berkeley Shenzhen Institute, Shenzhen 518055, China (e-mail: shaolun.huang@sz.tsinghua.edu.cn).

X. Xu is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: xuxx14@mails.tsinghua.edu.cn).

L. Zheng is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA (e-mail: lizhong@mit.edu).

distribution P_{UX^d}, such that U contains much information about the common structure shared by X^d. For this purpose, we apply the Watanabe's total correlation (or simply called the total correlation [6]) to measure the amount of information shared by multiple random variables, and then find the optimal embedded U such that the reduction of the total correlation given the knowledge of U is maximized. To extract the effective low-dimensional features, we restrict the information volume of U about $X_1, \ldots, X_d$ to be small, so that we can concentrate on the most "learnable" part of information about the common structures from the data. We show that in this small information rate regime of U, the optimal embedded U can be characterized by an exponential family induced by the largest eigenvector of a pairwise joint distribution matrix. Then, we apply the log-likelihood function of estimating U from $X_1, \ldots, X_d$ in this exponential family as the functional representation for extracting the common structure. Since U is informative about the common structure and the log-likelihood function is the sufficient statistic of the observed data vectors about the target U, such a functional representation is effective to extract the common structure shared by these random variables. In addition, we extend this approach to searching for a sequence of mutually independent random variables $U^k = (U_1, \ldots, U_k)$, such that the reduction of the total correlation is maximized. It turns out that the log-likelihood functions for estimating U^k precisely correspond to the top k eigenvectors of the pairwise joint distribution matrix, which establishes a decomposition of the common information between multiple random variables to principal modes of the pairwise joint distribution matrix.

Moreover, we demonstrate that these functional representations can be directly computed from the observed data vectors by a multivariate alternating conditional expectation (MACE) algorithm, which generalizes the traditional alternating conditional expectation (ACE) algorithm [7] to more than two random variables. This offers an efficient and reliable way to compute useful functional representations from discrete data variables. Furthermore, for high-dimensional or continuous data variables such that the conditional expectations are hardly accurately estimated from the limited data samples, we show that the functional representations can be computed through neural networks by optimizing a pairwise correlation loss. This offers a novel neural network training architecture for jointly analyzing multi-modal data.

Finally, we show in Section IV that our approach shares deep connections and can be viewed as generalizations to several existing techniques, including the Hirschfeld-Gebelein-Rényi (HGR) maximal correlation [8], linear PCA, and consistent functional map network. This combines the knowledge from different domains, and offers a unified understanding for disciplines in information theory, statistics, and machine learning. We would also like to mention that the idea of studying the tradeoff between the total correlation and the common information rate was also employed in [9] [10] for Gaussian vectors in caching problems, while our works investigate this tradeoff for general discrete random variables. Moreover, the correlation explanation (CorEx) introduced by [11] also applied the total correlation as the information criterion to unsupervised learning. In particular, the authors in [11] solved an optimization problem by restricting the cardinality of U, and a rather complicated iterative algorithm for discrete data samples was derived. On the other hand, in this paper we restrict the information volume contained in U, which is a more natural constraint in information theory, and we obtain clean analytical solutions that can be computed on continuous or high-dimensional data by simple and efficient algorithms.

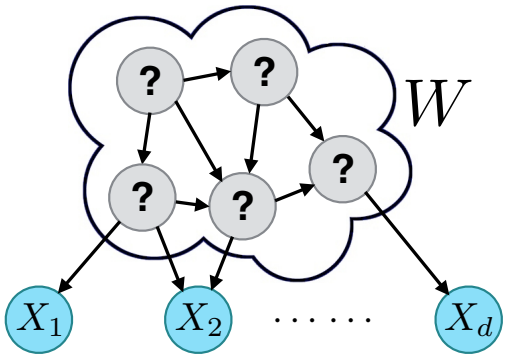

Fig. 2: The random variables $X_1, \ldots, X_d$ are conditional independently generated from some hidden structure W.

In the rest of this paper, we introduce the details of our information theoretic approach for extracting the common structure via functional representations, and present the resulted algorithm design, as well as their applications to practical problems.

II. The Information Theoretic Approach

Given discrete random variables $X^d = (X_1, \ldots, X_d)$ with the ranges $\mathcal{X}^d \triangleq \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ and joint distribution P_{X^d}, we model the common structure shared by these random variables as a high-dimensional latent variable W, in which the random variables $X_1, \ldots, X_d$ are conditionally independent given W, i.e., $P_{X^d|W} = \Pi_{i=1}^d P_{X_i|W}$, as depicted in Fig. 2. Our goal is to learn this common structure from i.i.d. sample vectors generated from P_{X^d}. Since the correlation between W and X_i's is generally complicated, it is difficult to directly identify and learn the structures of W without the labels and assumptions on the generating models of X_i's as in the unsupervised learning scenarios. Therefore, instead of identifying the high-dimensional latent variable W, we focus on learning the low-dimensional random variable U that contains much common information shared between X_i's, which can be viewed as the informative attribute for the common structure.

To identify such variable, we apply the total correlation[1] [6] to measure the amount of common information shared between multiple random variables. Then, the amount of information that an attribute U contains about the common structure shared by the random variables X_i's is measured by the reduction of the total correlation given the knowledge of U, defined as

$$\mathcal{L}(X^d|U)$$

[1]Specifically, for random variables $X_1, \ldots, X_d$, the total correlation is defined as the Kullback-Leibler (K-L) divergence $D(P_{X_1 \cdots X_d} \| P_{X_1} \cdots P_{X_d})$ between the joint distribution and the product of the marginal distributions.

$$\triangleq D(P_{X^d}\|P_{X_1}\cdots P_{X_d}) - D(P_{X^d}\|P_{X_1}\cdots P_{X_d}|U). \quad (1)$$

Our goal is to identify the targeted random variable U with the information rate constraint[2] $I(U;X^d) \le \delta$, for some given δ, such that the reduction of the total correlation is maximized. This can be formulated as the optimization problem:

$$\max_{P_{UX^d}} \quad \mathcal{L}(X^d|U) \quad (2a)$$

$$\text{subject to: } I(U;X^d) \le \delta. \quad (2b)$$

In particular, we would like to focus on the low-rate regime of U, which assumes δ to be small. This allows us to concentrate on the most representative low-dimensional attribute for describing the common structure. In addition, we make an extra constraint that

$$\min_u P_U(u) > \gamma, \quad (3)$$

for some finite $\gamma > 0$ irrelevant to δ, which is natural for many machine learning problems. While the optimization problem (2) in general has no analytical solution, in the regime of small δ, it can be solved by a local information geometric approach, in which the optimal solutions can be specified by the eigen-decomposition of some pairwise joint distribution matrix.

A. The Local Information Geometry

To delineate our approach and results, we define the matrix $\mathbf{B}$ from the pairwise joint distributions as

$$\mathbf{B} = \begin{bmatrix} \mathbf{I}_{(1)} & \mathbf{B}_{12} & \cdots & \mathbf{B}_{1d} \\ \mathbf{B}_{21} & \mathbf{I}_{(2)} & \cdots & \mathbf{B}_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_{d1} & \mathbf{B}_{d2} & \cdots & \mathbf{I}_{(d)} \end{bmatrix} \quad (4)$$

where $\mathbf{I}_{(i)}$ are $|\mathcal{X}_i| \times |\mathcal{X}_i|$ identity matrices, for all i, and $\mathbf{B}_{ij}$ are $(|\mathcal{X}_i| \times |\mathcal{X}_j|)$-dimensional matrices with the entry in the x_i-th row and x_j-th column defined as, for all $i \ne j$,

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

The eigen-decomposition of the matrix $\mathbf{B}$ has the following properties.

Lemma 1. *Let the eigenvalues and eigenvectors of the matrix $\mathbf{B}$ be $\lambda^{(0)} \ge \lambda^{(1)} \ge \cdots \ge \lambda^{(m-1)}$ and $\boldsymbol{\psi}^{(0)}, \boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(m-1)}$, respectively, where $m \triangleq \sum_{i=1}^d |\mathcal{X}_i|$ is the dimensionality of $\mathbf{B}$. In addition, let $\boldsymbol{v}_i$ be the $|\mathcal{X}_i|$-dimensional vector such that $v_i(x_i) = \sqrt{P_{X_i}(x_i)}$, then*

1) *$\mathbf{B}$ is a positive semidefinite matrix, i.e., $\lambda^{(m-1)} \ge 0$.*
2) *The largest eigenvalue $\lambda^{(0)} = d$ with the corresponding eigenvector $\boldsymbol{\psi}^{(0)} = \frac{1}{\sqrt{d}}\left[\boldsymbol{v}_1^{\mathrm{T}}, \ldots, \boldsymbol{v}_d^{\mathrm{T}}\right]^{\mathrm{T}}$.*
3) *The second largest eigenvalue $\lambda^{(1)} \ge 1$.*
4) *The last $d-1$ eigenvalues $\lambda^{(m-d+1)} = \cdots = \lambda^{(m-1)} = 0$, and the subspace of the corresponding $d-1$ eigenvectors is spanned by the vectors $\boldsymbol{\psi} = \left[\alpha_1\boldsymbol{v}_1^{\mathrm{T}}, \ldots, \alpha_d\boldsymbol{v}_d^{\mathrm{T}}\right]^{\mathrm{T}}$, such that the scalars α_i's satisfy $\sum_{i=1}^d \alpha_i = 0$.*
5) *For each $1 \le \ell \le m-d$, if we partition the corresponding eigenvector $\boldsymbol{\psi}^{(\ell)}$ into $|\mathcal{X}_i|$-dimensional subvectors $\boldsymbol{\psi}_i^{(\ell)}$, such that*

$$\boldsymbol{\psi}^{(\ell)} = \begin{bmatrix} \boldsymbol{\psi}_1^{(\ell)} \\ \vdots \\ \boldsymbol{\psi}_d^{(\ell)} \end{bmatrix}, \quad (5)$$

then $\boldsymbol{\psi}_i^{(\ell)}$ is orthogonal to $\boldsymbol{v}_i$, for all i.

Proof. See Appendix A. □

It will also be convenient to define the matrix $\tilde{\mathbf{B}}$:

$$\tilde{\mathbf{B}} \triangleq \mathbf{B} - d \cdot \boldsymbol{\psi}^{(0)}(\boldsymbol{\psi}^{(0)})^{\mathrm{T}}. \quad (6)$$

Then from Lemma 1, the eigenvalues of $\tilde{\mathbf{B}}$ are $\lambda^{(1)} \ge \cdots \ge \lambda^{(m-d)} \ge 0 = \lambda^{(m-d+1)} = \cdots = \lambda^{(m)}$, with the corresponding eigenvectors $\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(m-1)}, \boldsymbol{\psi}^{(0)}$. Moreover, we define a collection of functions $f_i^{(\ell)} : \mathcal{X}_i \mapsto \mathbb{R}$ as

$$f_i^{(\ell)}(x_i) = \frac{\psi_i^{(\ell)}(x_i)}{\sqrt{P_{X_i}(x_i)}}, \qquad \text{for all } i, \ell, \quad (7)$$

where $\boldsymbol{\psi}_i^{(\ell)}$ is the i-th subvector of $\boldsymbol{\psi}^{(\ell)}$ as defined in (5). Then, it follows from Lemma 1 and (7) that $f_i^{(\ell)}(X_i)$'s are zero-mean functions and $\sum_{i=1}^d \mathbb{E}[(f_i^{(\ell)}(X_i))^2] = 1$. In addition, these functions induce an exponential family of joint distributions for U, X^d.

Definition 1. Let $\mathcal{H}$ be the set of functions $h\colon \mathcal{U} \mapsto \mathbb{R}$ with zero mean and unit variance. Then, an exponential family $\mathcal{P}_{\exp}^{(\delta)}$ on U, X^d is defined as

$$\mathcal{P}_{\exp}^{(\delta)} = \Bigg\{ \frac{1}{Z} P_U(u) P_{X^d}(x^d) \cdot \exp\left(\sqrt{2\delta}\frac{h(u)}{\sqrt{\lambda^{(1)}}}\sum_{i=1}^d f_i^{(1)}(x_i)\right) : h \in \mathcal{H} \Bigg\},$$

where Z is the normalizing factor.

Note that this also defines a family of random variables U embedded in X^d corresponding to the collection of distributions in $\mathcal{P}_{\exp}^{(\delta)}$. It turns out that this exponential family characterizes the optimal solution of (2) in the regime of small δ, which is demonstrated as follows.

Theorem 1. *The optimal value of* (2a) *is*

$$\max_{P_{UX^d}} \mathcal{L}(X^d|U) = \delta\left(\lambda^{(1)} - 1\right) + o(\delta), \quad (8)$$

which is attainable by the distributions in $\mathcal{P}_{\exp}^{(\delta)}$. Moreover, for any distribution P_{UX^d} achieving (8)*, there exists a distribution $\hat{P}_{UX^d} \in \mathcal{P}_{\exp}^{(\delta)}$, such that for all $(u, x^d) \in \mathcal{U} \times \mathcal{X}^d$,*

$$\left|P_{UX^d}(u,x^d) - \hat{P}_{UX^d}(u,x^d)\right| = o\left(\sqrt{\delta}\right).$$

Proof. See Appendix B. □

[2]Note that $I(U;X^d)$ measures the amount of information of U about the whole X^d, while $\mathcal{L}(X^d|U)$ measures the amount of information only about the common structure. The constraint $I(U;X^d) \le \delta$ allows us to focus on low-dimensional attribute of W, in which we typically choose δ to be small.

From Theorem 1, the family of random variables U embedded in X^d defined by $\mathcal{P}^{(\delta)}_{\exp}$ is the set of attributes that contain the most amount of information about the common structure shared by X^d. To extract such information from data, we consider the log-likelihood function to estimate U from X^d:

$$\log \frac{P_{X^d|U=u}(x^d)}{P_{X^d}(x^d)} = \frac{\sqrt{2\delta}h(u)}{\sqrt{\lambda^{(1)}}} \sum_{i=1}^{d} f_i^{(1)}(x_i) + o\left(\sqrt{\delta}\right). \quad (9)$$

Although the log-likelihood functions for different $U = u$ in the exponential family $\mathcal{P}^{(\delta)}_{\exp}$ may have different magnitudes due to $h(u)$, all of them are proportional to the functional representation $\sum_{i=1}^{d} f_i^{(1)}(x_i)$ of the data vectors. This can be interpreted as the 1-dimensional subspace of the functional space of X^d that is the most informative about the shared structure. This is similar to what linear PCA [2] aims to achieve in the space of data, while we are searching for the optimal subspace of the general functional space. Later on we will show that our result is indeed a nonlinear generalization of linear PCA.

In addition, note that $\boldsymbol{\psi}^{(1)}$ is the second largest eigenvector of $\mathbf{B}$, which maximizes $\boldsymbol{\psi}^{\mathrm{T}}\mathbf{B}\boldsymbol{\psi}$ over all unit vectors $\boldsymbol{\psi}$ that are orthogonal to $\boldsymbol{\psi}^{(0)}$. This implies that the functions $f_i^{(1)}(X_i)$ defined from (7) form the optimal solution of the joint correlation optimization problem:

$$\begin{aligned} \max_{f_i\colon \mathcal{X}_i \mapsto \mathbb{R},\ i=1,\ldots,d} \ & \mathbb{E}\left[\sum_{i\neq j} f_i(X_i) f_j(X_j)\right] \\ \text{subject to:} \quad & \mathbb{E}\left[f_i(X_i)\right] = 0, \quad i = 1, \ldots, d \\ & \mathbb{E}\left[\sum_{i=1}^{d} f_i^2(X_i)\right] = 1, \quad i = 1, \ldots, d. \end{aligned}$$

Therefore, the functional representation $f_i^{(1)}(X_i)$ essentially searches for a 1-dimensional subspace for each functional space of X_i, such that the joint correlation between these subspaces is maximized. As a consequence, these subspaces and the corresponding functional representations convey much information about the common structure shared among these random variables.

B. The Informative k-dimensional Attributes

In addition to the largest eigenvector $\boldsymbol{\psi}^{(1)}$, the rest eigenvectors of $\tilde{\mathbf{B}}$ essentially lead to functional representations, which correspond to informative k-dimensional attributes for the common structure. To show that, we consider the optimization problem[3] for k-dimensional attribute $U^k = (U_1, \ldots, U_k)$:

$$\max_{P_{U^k X^d}} \mathcal{L}\left(X^d | U^k\right), \quad (10)$$

[3]For comparison, CorEx [11] selected the attributes $(U_1, \ldots, U_k)$ by considering the optimization problem for given cardinalities of U_i (cf. [11, Eq. (4)]):

$$\max_{X_{G_i}, P_{U_i|X_{G_i}},\ i=1,\ldots,k} \ \sum_{i=1}^{k} \mathcal{L}\left(X_{G_i} | U_i\right),$$

where $\{X_{G_1}, \ldots, X_{G_k}\}$ is required to form a partition of $\{X_1, \ldots, X_d\}$.

where $\mathcal{L}\left(X^d|U^k\right)$ is as defined in (1), and the maximization is over all joint distributions $P_{U^k X^d}$ such that the constituent variables U_i with ranges $\mathcal{U}_i$, for $i = 1, \ldots, k$, satisfy: 1) $\delta \geq I(U_1; X^d) \geq \cdots \geq I(U_k; X^d)$; 2) $\min_{u_i \in \mathcal{U}_i} P_{U_i}(u_i) > \gamma$ for all $i = 1, \ldots, d$ and some constant $\gamma > 0$ independent of δ; 3) $U_1, \ldots, U_k$ are mutually independent variables; 4) $U_1, \ldots, U_k$ are conditionally independent variables given X^d.

To solve the optimization problem (10), we define the following exponential family for k-dimensional attributes.

Definition 2. Let $\mathcal{H}_i$ be the set of functions $h_i\colon \mathcal{U}_i \mapsto \mathbb{R}$ with zero mean and unit variance, for $i = 1, \ldots, k$. Then, an exponential family $\mathcal{P}^{(\delta)}_{\exp,k}$ on U^k, X^d is defined as

$$\begin{aligned} \mathcal{P}^{(\delta)}_{\exp,k} = \Bigg\{ & \frac{1}{Z_k}\left[\prod_{j=1}^{k} P_{U_i}(u_i)\right] P_{X^d}(x^d) \\ & \cdot \exp\left(\sqrt{2\delta}\sum_{\ell=1}^{k_0} h_\ell(u_\ell) \sum_{j=1}^{k_0} \frac{q_{j\ell}}{\sqrt{\lambda^{(j)}}} \sum_{i=1}^{d} f_i^{(j)}(x_i)\right) \\ & : h_\ell \in \mathcal{H}_\ell, \mathbf{Q} = [q_{ij}]_{k_0 \times k_0}, \mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}_{k_0} \Bigg\}, \end{aligned}$$

where $f_i^{(j)}(x_i)$ is as defined in (7), Z_k is the normalizing factor, $k_0 = \min\{k, k^*\}$ and

$$k^* \triangleq \max\left\{i\colon \lambda^{(i)} > 1\right\}.$$

Then, the exponential family $\mathcal{P}^{(\delta)}_{\exp,k}$ characterizes the optimal solutions of (10).

Theorem 2. *The optimal value of* (10) *is*

$$\max_{P_{U^k X^d}} \mathcal{L}\left(X^d|U^k\right) = \delta\left(\sum_{\ell=1}^{k_0} \lambda^{(\ell)} - k_0\right) + o(\delta), \quad (11)$$

which is attainable by the distributions in $\mathcal{P}^{(\delta)}_{\exp,k}$. Moreover, for any distribution $P_{U^k X^d}$ achieving (11)*, there exists a distribution $\hat{P}_{U^k X^d} \in \mathcal{P}^{(\delta)}_{\exp,k}$, such that for all $(u^k, x^d) \in \mathcal{U}_1 \times \cdots \times \mathcal{U}_k \times \mathcal{X}^d$,*

$$\left|P_{U^k X^d}(u^k, x^d) - \hat{P}_{U^k X^d}(u^k, x^d)\right| = o\left(\sqrt{\delta}\right).$$

Proof. See Appendix C. □

Note that from Definition 2 and Theorem 2, when $k > k^*$, the optimal solution is to design $P_{U^{k^*} X^d}$ to follow the distributions in $\mathcal{P}^{(\delta)}_{\exp,k^*}$, and let the last $k - k^*$ attributes $U_{k^*+1}, \ldots, U_k$ be independent of X^d. This implies that only the top k^* attributes can effectively reduce the total conditional correlation, which leads to an intrinsic criterion for designing the dimensionality k of the attributes.

Moreover from Definition 2, the log-likelihood functions for the optimal attributes correspond to the functional representations $\sum_{i=1}^{d} f_i^{(\ell)}(x_i)$, for $\ell = 1, \ldots, k$. This generalizes (9) for providing the informative k-dimensional representations about the common structure shared by $X_1, \ldots, X_d$. Furthermore, it is shown in Appendix D that the functions $f_i^{(\ell)}$ as defined

in (7) form the optimal solution of the following optimization problem:

$$\max_{\underline{f}_i:\ \mathcal{X}_i\mapsto\mathbb{R}^k,\ i=1,\ldots,d} \mathbb{E}\left[\sum_{i\neq j}\underline{f}_i^{\mathrm{T}}(X_i)\underline{f}_j(X_j)\right] \tag{12a}$$

$$\text{subject to:}\quad \mathbb{E}\left[\underline{f}_i(X_i)\right]=\underline{0},\ \text{for all } i \tag{12b}$$

$$\mathbb{E}\left[\sum_{i=1}^{d}\underline{f}_i(X_i)\underline{f}_i^{\mathrm{T}}(X_i)\right]=\mathbf{I}_k, \tag{12c}$$

where $\mathbf{I}_k$ is the k-dimensional identity matrix. Therefore for $i=1,\ldots,d$, the functional representations $f_i^{(\ell)}(X_i), \ell=1,\ldots,k$, form the k-dimensional functional subspace of X_i, such that the joint correlation between these subspaces for different X_i's is maximized.

Example 1 (Common Bits Patterns Extraction). Suppose that $b_1,\ldots,b_r\in\{1,-1\}$ are mutually independent $\mathsf{Bern}(\frac{1}{2})$ bits, and each random variable $X_i=b_{\mathcal{I}_i}\triangleq(b_j)_{j\in\mathcal{I}_i}$ is a subset of these random bits, where $\mathcal{I}_i\subseteq\{1,\ldots,r\}$ denotes the index set. Then, our information theoretic approach essentially extracts the bit patterns that appear the most among the random variables X^d. To show that, we define $w(\mathcal{I})$ as the number of sets $\mathcal{I}_i$ $(i=1,\ldots,d)$ that include $\mathcal{I}$, i.e.[4],

$$w(\mathcal{I})\triangleq\sum_{i=1}^{d}\mathbb{1}_{\{\mathcal{I}\subset\mathcal{I}_i\}}. \tag{13}$$

In addition, we denote $\varnothing=\mathcal{J}_0,\ldots,\mathcal{J}_{2^r-1}$ as the 2^r subsets of $\{1,\ldots,r\}$ with the decreasing order $d=w(\mathcal{J}_0)\geq w(\mathcal{J}_1)\geq\cdots\geq w(\mathcal{J}_{2^r-1})$. Then, it is shown in Appendix E that the eigenvalues for the corresponding matrix $\mathbf{B}$ are

$$\lambda^{(\ell)}=w(\mathcal{J}_\ell),\quad \ell=0,\ldots,m-1. \tag{14}$$

where $m=\sum_{i=1}^{d}2^{|\mathcal{I}_i|}$ is the dimensionality of $\mathbf{B}$. Therefore, the eigenvalue $\lambda^{(\ell)}$ of the matrix $\mathbf{B}$ essentially counts the number of times the corresponding bits pattern $b_{\mathcal{J}_\ell}$ appears among the random variables X^d, and the largest eigenvalue indicates the most appeared bits pattern. Moreover for $\lambda^{(\ell)}>0$, the corresponding functions $f_i^{(\ell)}(X_i)$ $(i=1,\ldots,d)$ as defined in (7) are

$$f_i^{(\ell)}(X_i)=\begin{cases}\dfrac{1}{\sqrt{w(\mathcal{J}_\ell)}}\displaystyle\prod_{j\in\mathcal{J}_\ell}b_j & \text{if } \mathcal{J}_\ell\subset\mathcal{I}_i\\ 0 & \text{otherwise.}\end{cases} \tag{15}$$

Thus, the ℓ-th optimal functional representation of X^d (cf. Section II-B) is

$$\sum_{i=1}^{d}f_i^{(\ell)}(X_i)=\sqrt{w(\mathcal{J}_\ell)}\prod_{j\in\mathcal{J}_\ell}b_j,$$

which depends only on the bits indexed by $\mathcal{J}_\ell$. For instance, if $r=d=3$, and $X_1=\{b_1,b_2\}, X_2=\{b_2,b_3\}$, and $X_3=\{b_1,b_3\}$, then for all subsets of $\{1,2,3\}$, the values for the function $w(\cdot)$ as defined in (13) are

$$w(\varnothing)=3,\quad w(\{1\})=w(\{2\})=w(\{3\})=2,$$
$$w(\{1,2\})=w(\{2,3\})=w(\{3,1\})=1,\quad w(\{1,2,3\})=0.$$

Therefore, the corresponding eigenvalues of $\mathbf{B}$ are

$$\lambda^{(0)}=3,\quad \lambda^{(1)}=\lambda^{(2)}=\lambda^{(3)}=2,$$
$$\lambda^{(4)}=\lambda^{(5)}=\lambda^{(6)}=1,\quad \lambda^{(7)}=0.$$

Moreover, the corresponding $f_i^{(\ell)}(X_i)$'s satisfy

$$\sum_{i=1}^{3}f_i^{(\ell)}(X_i)=\sqrt{2}b_\ell,\quad \ell=1,2,3,$$

and

$$\sum_{i=1}^{3}f_i^{(\ell)}(X_i)=\begin{cases}b_1b_2 & \ell=4\\ b_2b_3 & \ell=5\\ b_3b_1 & \ell=6.\end{cases}$$

III. The Algorithm to Compute the Functional Representation from Data

While our information theoretic approach provides a guidance for searching informative functional representations, it remains to derive the algorithm to compute these functions from observed data vectors. Intuitively, one can first estimate the empirical distribution between $X_1,\ldots,X_d$ from the data samples, and then construct the matrix $\mathbf{B}$ 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 $\mathbf{B}$ may be extremely high especially for big data applications, so that the singular value decomposition (SVD) can not be computed directly.

A. The Multivariate Alternating Conditional Expectation (MACE) Algorithm

Alternatively, it is well-known that eigenvectors of a matrix can be efficiently computed by the power method [12]. The power method iteratively multiplies the matrix to an initial vector, and if all the eigenvalues are nonnegative, it converges to the eigenvector with respect to the largest eigenvalue with an exponential convergence rate. To apply the power method for computing the second largest eigenvector of $\mathbf{B}$, we choose an initial vector $\boldsymbol{\psi}=\begin{bmatrix}\boldsymbol{\psi}_1^{\mathrm{T}} & \cdots & \boldsymbol{\psi}_d^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}$, such that $\boldsymbol{\psi}_i$ is orthogonal to $\boldsymbol{v}_i$, for all i. This forces $\boldsymbol{\psi}$ to be orthogonal to $\boldsymbol{\psi}^{(0)}$, and since $\mathbf{B}$ is positive semidefinite, the power iteration will converge to the second largest eigenvector if $\boldsymbol{\psi}$ is not orthogonal to $\boldsymbol{\psi}^{(1)}$. Then, the algorithm iteratively computes the matrix multiplication $\boldsymbol{\psi}\leftarrow\mathbf{B}\boldsymbol{\psi}$, or equivalently

$$\boldsymbol{\psi}_i\leftarrow\boldsymbol{\psi}_i+\sum_{j\neq i}\mathbf{B}_{ij}\boldsymbol{\psi}_j, \tag{16}$$

[4]We adopt the Kronecker notation

$$\mathbb{1}_{\mathcal{A}}=\begin{cases}1 & \mathcal{A}\text{ is true}\\ 0 & \text{otherwise.}\end{cases}$$

Algorithm 1 The Multivariate ACE (MACE) Algorithm

Require : The data samples $\underline{x}^{(\ell)} = (x_1^{(\ell)}, \ldots, x_d^{(\ell)})$, $\ell = 1, \ldots, n$ of variables $X_1, \ldots, X_d$

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

repeat :

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

until : The $\mathbb{E}\left[\sum_{i\neq j} f_i(X_i) f_j(X_j)\right]$ stops increasing.

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 [13], the step (16) is equivalent to a conditional expectation operation on functions:

$$f_i(X_i) \leftarrow f_i(X_i) + \mathbb{E}\left[\sum_{j\neq i} f_j(X_j)\middle| X_i\right], \tag{17}$$

Therefore, the power method can be transferred to an algorithm based on the alternating conditional expectation (ACE) [7] algorithm as shown in Algorithm 1, which computes the optimal functional representation derived in Section II. Note that the choice of $\boldsymbol{\psi}_i$ to be orthogonal to $\boldsymbol{v}_i$ is transferred to the zero-mean choice of functions in the initialization step of the algorithm.

B. Finding k Functional Representations from Eigen-decomposition

The Algorithm 1 can be further extended to compute the top k eigenvectors $\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(k)}$, and the corresponding functional representations. To design the algorithm for computing these functions, we denote the ℓ-th functional representation as $\vec{f}^{(\ell)} = (f_1^{(\ell)}, \ldots, f_d^{(\ell)})$, where $f_i^{(\ell)}$ is as defined in (7). Then, since $\boldsymbol{\psi}^{(k)}$ is orthogonal to $\boldsymbol{\psi}^{(\ell)}$, for $\ell \leq k-1$, the k-th functional representation $\vec{f}^{(k)}$ can be computed by the power method similar to the first functional representation $\vec{f}^{(1)}$, but with extra orthogonality constraints

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

for $\ell \leq k-1$ to maintain the orthogonality to the first $k-1$ functional representations. Therefore, $\vec{f}^{(k)}$ can be computed by the power method as in Algorithm 1 with the extra step of Gram-Schmidt procedure to guarantee the orthogonality, which is illustrated in Algorithm 2. Note that the computation complexities of Algorithm 1 and Algorithm 2 are both linear to the size of the dataset, which is often much more efficient than the singular value decomposition of the matrix $\mathbf{B}$.

C. Generating Informative Functional Representations for High-Dimensional Data

While the Algorithm 2 generally requires less training samples than estimating the joint distribution and the matrix

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

Require : The data samples $\underline{x}^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})$, $i = 1, \ldots, n$ of variables $X_1, \ldots, X_d$, and the 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 for $\vec{f}^{(k)}$.
3. The Gram-Schmidt procedure: $\vec{f}^{(k)} \leftarrow \vec{f}^{(k)} - \sum_{\ell=1}^{k-1} \langle \vec{f}^{(\ell)}, \vec{f}^{(k)} \rangle \cdot \vec{f}^{(\ell)}$

until : The $\mathbb{E}\left[\sum_{i\neq j} f_i^{(k)}(X_i) f_j^{(k)}(X_j)\right]$ stops increasing.

$\mathbf{B}$ (or $\tilde{\mathbf{B}}$), in order to obtain an acceptable estimation for the conditional expectation step (17), it is still necessary to acquire training samples in the size comparable to the cardinality of the random variable X_i. This is often difficult for high-dimensional or continuous random variables in practice. In such cases, we propose a neural network based approach to generate the informative functional representations by deep neural networks. The key idea is to note that by Eckart-Young-Mirsky theorem [14], the top $k \leq m$ eigenvectors of $\tilde{\mathbf{B}}$ can be computed from the low-rank approximation problem:

$$\boldsymbol{\Psi}^* = \min_{\boldsymbol{\Psi}\in\mathbb{R}^{m\times k}} \left\|\tilde{\mathbf{B}} - \boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{T}}\right\|_{\mathrm{F}}^2 \tag{18}$$

where the columns of $\boldsymbol{\Psi}^*$ are the top k eigenvectors of $\tilde{\mathbf{B}}$. The unconstrained optimization problem (18) leads to a training loss for generating informative functions by neural networks.

Proposition 1. *Let $\boldsymbol{\Psi}_i$ be $|\mathcal{X}_i| \times k$ matrices, for $i = 1, \ldots, d$, such that $\boldsymbol{\Psi} = \left[\boldsymbol{\Psi}_1^{\mathrm{T}} \ \cdots \ \boldsymbol{\Psi}_d^{\mathrm{T}}\right]^{\mathrm{T}}$, and define k-dimensional functions $\underline{f}_i \colon \mathcal{X}_i \mapsto \mathbb{R}^k$, $i = 1, \ldots, d$, as $\underline{f}_i(x_i) = \boldsymbol{\Psi}_i^{\mathrm{T}}(x_i)/\sqrt{P_{X_i}(x_i)}$, where $\boldsymbol{\Psi}_i(x_i)$ denotes the x_i-th row of the matrix $\boldsymbol{\Psi}_i$. Then, it follows that*

$$\left\|\tilde{\mathbf{B}} - \boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{T}}\right\|_{\mathrm{F}}^2 = \left\|\tilde{\mathbf{B}}\right\|_{\mathrm{F}}^2 - 2H\left(\underline{f}_1(X_1), \ldots, \underline{f}_d(X_d)\right), \tag{19}$$

where

$$H\left(\underline{f}_1(X_1), \ldots, \underline{f}_d(X_d)\right) \triangleq \sum_{i=1}^d \sum_{j=1}^d H\left(\underline{f}_i(X_i), \underline{f}_j(X_j)\right),$$

and $H\left(\underline{f}_i(X_i), \underline{f}_j(X_j)\right)$ is defined as, for all i, j,

$$\begin{aligned} &H\left(\underline{f}_i(X_i), \underline{f}_j(X_j)\right) \\ &\triangleq \mathbb{E}\left[\underline{f}_i^{\mathrm{T}}(X_i)\underline{f}_j(X_j)\right] - \left(\mathbb{E}\left[\underline{f}_i(X_i)\right]\right)^{\mathrm{T}} \mathbb{E}\left[\underline{f}_j(X_j)\right] \\ &\quad - \frac{1}{2}\operatorname{tr}\left\{\mathbb{E}\left[\underline{f}_i(X_i)\underline{f}_i^{\mathrm{T}}(X_i)\right]\mathbb{E}\left[\underline{f}_j(X_j)\underline{f}_j^{\mathrm{T}}(X_j)\right]\right\}, \end{aligned}$$

where $\operatorname{tr}\{\cdot\}$ denotes the trace of its matrix argument.

Proof. See Appendix F. □

Note that $H\left(\underline{f}_i(X_i), \underline{f}_j(X_j)\right)$ coincides with the *H-score* [15] when the means of the functions are zero, hence

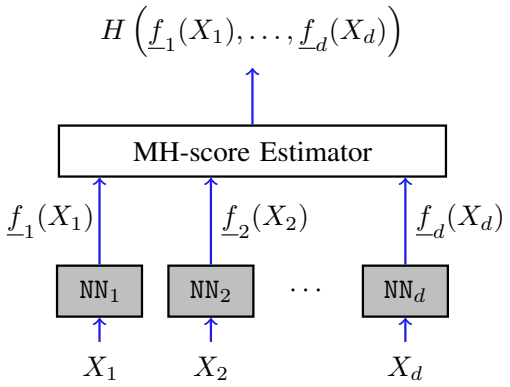

Fig. 3: The network architecture to estimate optimal functional representations, where each $\mathtt{NN}_i$ is a neural network to extract feature $\underline{f}_i(X_i)$ from the i-th input X_i.

we term $H\left(\underline{f}_1(X_1),\ldots,\underline{f}_d(X_d)\right)$ the *multivariate H-score* (MH-score). Then from (19), the optimization problem (18) is equivalent to the functional optimization problem

$$\max_{f_i:\,\mathcal{X}_i\mapsto\mathbb{R}^k,\ i=1,\ldots,d} H\left(\underline{f}_1(X_1),\ldots,\underline{f}_d(X_d)\right), \tag{20}$$

for solving the informative functional representations for the common structure. Moreover, since the H-score can be efficiently estimated from data samples (via, e.g., [16, Algorithm 1]), the MH-score can also be effectively computed using samples. Then, the optimization problem (20) leads to a neural network training strategy. Specifically, given the training samples of $X_1,\ldots,X_d$, we design d neural networks, where the i-th neural network $\mathtt{NN}_i$ takes X_i as the input and generates the representations $\underline{f}_i(X_i)$. Then, the weights of these neural networks are trained to minimize the negative MH-score as the loss function. Finally, the informative functional representations are generated by the trained d neural networks that attempt to optimize (20), as illustrated in Fig. 3.

IV. Connections to Existing Techniques

In this section, we demonstrate the relationship between our functional representations and the Hirschfeld-Gebelein-Rényi (HGR) maximal correlation [8], linear PCA [2], and the consistent functional map [3]. This demonstrates the deep connections between our approach and existing techniques, while offering novel information theoretic interpretations to machine learning algorithms.

A. The HGR 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 [8].

Definition 3 (Maximal Correlation)**.** For jointly distributed random variables X and Y, with discete ranges $\mathcal{X}$ and $\mathcal{Y}$ respectively, the maximal correlation between X and Y is defined as:

$$\rho(X;Y)\triangleq\max_{f:\,\mathcal{X}\mapsto\mathbb{R},\ g:\,\mathcal{Y}\mapsto\mathbb{R}}\mathbb{E}\left[f(X)g(Y)\right]$$

where the maximum is taken over zero-mean and unit-variance functions $f(X)$ and $g(Y)$.

The HGR maximal correlation has been shown useful not only as a statistical measurement, but also in designing machine learning algorithms for regression problems [17] [13] [18]. To draw the connection, note that in the bivariate case $d=2$, the functions derived in Section II are precisely the maximal correlation functions for two random variables. In addition, our functional representation for general cases essentially defines a generalized version of the maximal correlation [cf. (12)].

Definition 4. The generalized maximal correlation for jointly distributed random variables $X_1,\ldots,X_d$ with discrete ranges $\mathcal{X}_i$, for $i=1,\ldots,d$, is defined as

$$\rho^*(X_1,\cdots,X_d)\triangleq\max\frac{1}{d-1}\mathbb{E}\left[\sum_{i\neq j}f_i(X_i)f_j(X_j)\right] \tag{21}$$

for the functions $f_i:\,\mathcal{X}_i\mapsto\mathbb{R}$, with the constraints $\mathbb{E}\left[f_i(X_i)\right]=0,\ \mathbb{E}\left[\sum_{i=1}^d f_i^2(X_i)\right]=1$, for all i.

It is easy to verify that $0\leq\rho^*(X_1,\cdots,X_d)\leq 1$, and $\rho^*(X_1,\cdots,X_d)=0$ if and only if $X_1,\ldots,X_d$ are pairwise independent.

Note that there are some other generalizations to maximal correlations to multiple random variables. For example, the network maximal correlation (NMC) proposed in [19] defined a correlation measurement in the same way as (21) but with a slightly different constraint: $\mathbb{E}\left[f_i(X_i)\right]=0,\ \mathbb{E}\left[f_i^2(X_i)\right]=1,$ for all i. In addition, [20] proposed a maximally correlated principal component analysis (MCPCA) and demonstrated that (cf. [20, Theorem 5]) the solution of MCPCA coincides with the top singular vector of $\tilde{\mathbf{B}}$ in some special case. However, our approach and results essentially offer the information theoretic justification of generalizing the maximal correlation as extracting common structures shared among random variables, and also provide the guidance to algorithm designs.

B. Linear PCA

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

$$\frac{1}{n}\sum_{\ell=1}^n\sum_{i\neq j}\left(w_ix_i^{(\ell)}\right)\left(w_jx_j^{(\ell)}\right)=\mathbb{E}\left[\sum_{i\neq j}(w_iX_i)\cdot(w_jX_j)\right] \tag{22}$$

subject to the constraint

$$1=\sum_{i=1}^d w_i^2=\sum_{i=1}^d\mathbb{E}\left[(w_iX_i)^2\right], \tag{23}$$

where the expectations in (22) and (23) are taking over the empirical distributions $P_{X_iX_j}$ and P_{X_i} from the data vectors. Comparing to the Definition 4, we can see that our functional representation generalizes the linear PCA to nonlinear functional spaces of data. We would like to emphasize that [13] 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.

C. Consistent Functional Map

In computer vision, a typical question is to find the shared components among a collection of shapes, for example, the legs of chairs, when some noisy maps between these shapes are given. An effective approach to extract such shared structure between shape collections, called consistent functional map network, is recently proposed in [3]–[5]. The main idea of the consistent functional map network is to formulate the shared components as low-dimensional subspaces of the functional spaces of these shapes, and the given noisy maps between these shapes are formulated as the transition maps between these functional spaces. Then, the goal of the consistent functional map network is to find a low-dimensional subspace of the functional space of each shape, such that under a cycle of transition maps between shapes, this low-dimensional subspace remains the same.

Note that this idea is similar to the functional representation we derived in this paper, except that the consistent functional map network considers the transition maps between shapes that are deterministic maps, while we consider stochastic maps between random variables. In fact, it is shown in [3] that if we write the noisy maps between shapes i and j as $\mathbf{M}_{ij}$ (cf. the $X_{ij}^{\star}$ of [3, Eq. (8)]), then such subspaces can be solved by the eigen-decomposition of a matrix by replacing the stochastic transition map $\mathbf{B}_{ij}$ of $\mathbf{B}$ in (4) into the noisy (deterministic) maps $\mathbf{M}_{ij}$ between shapes. Therefore, the functional representation presented in this paper can be viewed as an extension of the consistent functional map network to general stochastic object, which can essentially be applied to a wider range of problems.

V. The Numerical Simulations

The functional representations of the data can be viewed as low-dimensional feature functions selected from the hidden common structure. In this section, we show that such selected low-dimensional feature functions can practically be useful by verifying the performance in the MNIST handwritten digit database [1] for digits recognition. In the MNIST database, there are $n = 60\,000$ 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 both Algorithm 2 and the low-rank approximation method described in Section III-C 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 handwritten digit recognition.

Algorithm 3 Quantizing Alphabets to Reduce the Cardinality

Require: training samples $\left\{x_i^{(\ell)} : \ell = 1, \ldots, n\right\}$
  Initialize: set the alphabet $\mathcal{X}_i \leftarrow \varnothing$.
  For $\ell = 1 : n$
    If $\exists\, x \in \mathcal{X}_i$, such that $d_H\left(x, x_i^{(\ell)}\right) \le 3$.
      Then set $x_i^{(\ell)} \leftarrow x$.
    Else $\mathcal{X}_i \leftarrow \mathcal{X}_i \cup \left\{x_i^{(\ell)}\right\}$
  End

To begin, we need to identify the random variables X_i in the MNIST problem. For this purpose, 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. Fig. 1 illustrates this division of images.

The purpose of dividing entire image into subareas is to reduce the complexity of training joint feature functions among image pixels, while capturing the correlations between nearby pixels. Then, each sub-image i of the 64 subareas can be viewed as a random variable X_i, for $i = 1, \ldots, 64$. Therefore, if we denote $x_i^{(\ell)}$ as the value of the sub-image i of the ℓ-th image of the MNIST database, then each random variable X_i has n training samples $x_i^{(1)}, \ldots, x_i^{(n)}$.

A. Apply the MACE Algorithm to MNIST

To apply the MACE Algorithm 2, we further quantize each image pixel into binary signals "0" and "1" with the quantization threshold 40. Note that each $x_i^{(\ell)}$ is essentially a 36-dimensional binary vector, thus the cardinality of the alphabet $|\mathcal{X}_i| = 2^{36}$. To reduce the cardinality, for each subarea i, we go through n training images to find all possible binary vectors in $\{0,1\}^{36}$, and then map these binary vectors into a smaller alphabet set, such that two binary vectors with Hamming distance no greater than three are mapped into the same alphabet. This quantization procedure is illustrated in Algorithm 3.

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

$$\vec{s}_\ell = \left(\vec{f}_1(x_1^{(\ell)}), \ldots, \vec{f}_{64}(x_{64}^{(\ell)})\right),$$

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) [21] to classify the vectors $\vec{s}_\ell$, for $\ell = 1, \ldots, n$ into ten groups with respect to the labels z_ℓ. This results in a linear classifier that associates a label $\hat{z}_\ell \in \{0, \ldots, 9\}$ to each score vector $\vec{s}_\ell$, and the label represents the recognized digit of the image corresponding to the score vector.

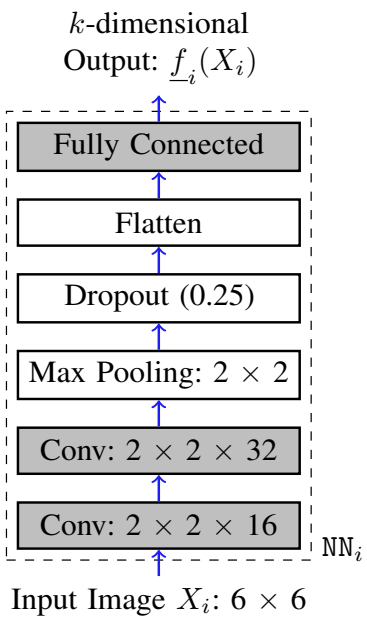

Fig. 4: The architecture of the i-th neural network $\mathtt{NN}_i$ that extracts feature $\underline{f}_i(X_i)$ from the input X_i.

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 feature functions $\vec{f}_i$. Then, the linear classifier is applied to recognize the digits in the test images, and the error probabilities of recognizing the digits via the score vectors with different values of k are demonstrated in the following table.

k	4	8	12	16	20	24
Error rate (%)	4.74	2.44	2.36	2.21	2.15	2.08

Note that our approach can be viewed as mapping the image pixels to the feature space by *one layer* of informative score functions and then apply the linear classification. It turns out that the error rate of our approach is comparable to the neural networks with *two layers* of feature mapping by the sigmoid functions (the error rate is 2.95% for a 3-layer fully-connected neural network with 500 and 150 units in two hidden layers [1], [22]). Moreover, the neural networks select the features with the aid of labels, while the feature functions in our approach are selected without the knowledge of label but from the shared structure between subareas. This essentially shows how the information from shared structures can be applied to practical problems by our algorithms.

B. Finding Functional Representations by Neural Networks

As illustrated in Section III-C, we first use 64 neural networks $\mathtt{NN}_1, \ldots, \mathtt{NN}_{64}$ to generate representations $\underline{f}_1(X_1), \ldots, \underline{f}_{64}(X_{64})$ from images, where each neural network $\mathtt{NN}_i$ consists of two convolutional layers, as shown in Fig. 4. Using the negative MH-score $-H\left(\underline{f}_1(X_1), \ldots, \underline{f}_{64}(X_{64})\right)$ as the loss function, we then train these 64 neural networks to obtain the optimal functional representations.

With the functional representations trained from the training set, we again adopt the linear SVM for the classification task and use this linear classifier to recognize the test images. The following table shows the classification error with different values of k.

k	4	8	12	16	20	24
Error rate (%)	3.46	1.73	1.43	1.17	1.15	1.11

Compared with the results from the MACE algorithm (Algorithm 2), the neural network approximation method has better performance. This performance gain is mainly from directly processing the subareas of images by the CNNs without the information loss in the quantization step. In addition, the features extracted in this unsupervised approach can achieve the performance comparable to the CNN-based supervised learning algorithms, such as LeNet-4, which has an error rate of 1.1% [22].

ACKNOWLEDGMENT

The work of Shao-Lun Huang was supported in part by the Natural Science Foundation of China under Grant 61807021, in part by the Shenzhen Science and Technology Research and Development Funds under Grant JCYJ20170818094022586, and in part by the Innovation and Entrepreneurship Project for Overseas High-Level Talents of Shenzhen under Grant KQJSCX20180327144037831.

REFERENCES

[1] Y. LeCun, C. Cortes, and C. J. Burges, "Mnist handwritten digit database," *AT&T Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, vol. 2, 2010.

[2] I. T. Jolliffe, "Principal component analysis. 1986," *Spring-verlag, New York*, 1986.

[3] Q. Huang, F. Wang, and L. Guibas, "Functional map networks for analyzing and exploring large shape collections," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 36:1–36:11, Jul. 2014.

[4] F. Wang, Q. Huang, and L. J. Guibas, "Image co-segmentation via consistent functional maps," in *Proceedings of the 2013 IEEE International Conference on Computer Vision*, Washington, DC, USA, 2013, pp. 849–856.

[5] F. Wang, Q. Huang, M. Ovsjanikov, and L. J. Guibas, "Unsupervised multi-class joint image segmentation," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2014, pp. 3142–3149.

[6] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of research and development*, vol. 4, no. 1, pp. 66–82, 1960.

[7] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *Journal of the American statistical Association*, vol. 80, no. 391, pp. 580–598, 1985.

[8] A. Rényi, "On measures of dependence," *Acta mathematica hungarica*, vol. 10, no. 3-4, pp. 441–451, 1959.

[9] G. J. Op't Veld and M. C. Gastpar, "Caching gaussians: Minimizing total correlation on the Gray-Wyner network," in *2016 Annual Conference on Information Science and Systems (CISS)*. IEEE, 2016, pp. 478–483.

[10] ——, "Total correlation of Gaussian vector sources on the Gray-Wyner network," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2016, pp. 385–392.

[11] G. Ver Steeg and A. Galstyan, "Discovering structure in high-dimensional data through correlation explanation," in *Advances in Neural Information Processing Systems*, 2014, pp. 577–585.

[12] G. H. Golub and C. Van Loan, "Matrix computations 4th edition," 2013.

[13] S.-L. Huang, A. Makur, L. Zheng, and G. W. Wornell, "On universal features for high-dimensional learning and inference," *submitted to IEEE Trans. Inform. Theory*, 2019. Preprint.

[14] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[15] S.-L. Huang, X. Xu, L. Zheng, and G. W. Wornell, "An information theoretic interpretation to deep neural networks," *arXiv preprint arXiv:1905.06600*, 2019.
[16] L. Wang, J. Wu, S.-L. Huang, L. Zheng, X. Xu, L. Zhang, and J. Huang, "An efficient approach to informative feature extraction from multimodal data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5281–5288.
[17] A. Makur, F. Kozynski, S.-L. Huang, and L. Zheng, "An efficient algorithm for information decomposition and extraction," in *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*. IEEE, 2015, pp. 972–979.
[18] M. Razaviyayn, F. Farnia, and D. Tse, "Discrete Rényi classifiers," in *Advances in Neural Information Processing Systems*, 2015, pp. 3276–3284.
[19] S. Feizi, A. Makhdoumi, K. Duffy, M. Medard, and M. Kellis, "Network maximal correlation," *arXiv preprint arXiv:1606.04789*, 2016.
[20] S. Feizi and D. Tse, "Maximally correlated principal component analysis," *arXiv preprint arXiv:1702.05471*, 2017.
[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
[22] Y. B. Y. LeCun, L. Bottou and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278 – 2324, Nov. 1998.
[23] S.-L. Huang and L. Zheng, "Linear information coupling problems," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 1029–1033.
[24] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

Shao-Lun Huang received the B.S. degree (Hons.) from the Department of Electronic Engineering, National Taiwan University, Taiwan, in 2008, and the M.S. and Ph.D. degrees from the Department of Electronic Engineering and Computer Sciences, Massachusetts Institute of Technology, in 2010 and 2013, respectively. From 2013 to 2016, he was working as a Postdoctoral Researcher jointly at the Department of Electrical Engineering, National Taiwan University, and the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology. Since 2016, he has been with the Tsinghua-Berkeley Shenzhen Institute, where he is currently an Assistant Professor. His research interests include information theory, communication theory, machine learning, with particular applications in deep neural networks.

Xiangxiang Xu (S'16) received the B.Eng. degree in electronic engineering from Tsinghua University, Beijing, China, in 2014, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. His research interests include information theory and machine learning, with applications in data analytics.

Lizhong Zheng (S'00–M'02–F'16) received the B.S. and M.S. degrees from the Department of Electronic Engineering at Tsinghua University, Beijing, China, in 1994 and 1997, respectively, and a Ph.D. degree from the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley (UC Berkeley), USA, in 2002. Since 2002, he has been working in the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology, Cambridge, USA, where he is currently a Professor of Electrical Engineering. His research interests include information theory, statistical inference, and wireless communications and networks. He was a recipient of the Eli Jury Award from UC Berkeley in 2002, the IEEE Information Theory Society Paper Award in 2003, the NSF CAREER Award in 2004, and the AFOSR Young Investigator Award in 2007. He became an IEEE Fellow in 2016.