An Efficient Algorithm for Information Decomposition and Extraction

Anuran Makur, Fabián Kozynski, Shao-Lun Huang, and Lizhong Zheng
EECS Department, Massachusetts Institute of Technology
Email: {a_makur, kozynski, shaolun, lizhong}@mit.edu

***Abstract*—The Hirschfeld-Gebelein-Rényi maximal correlation is a well-known measure of statistical dependence between two (possibly categorical) random variables. In inference problems, the maximal correlation functions can be viewed as so called features of observed data that carry the largest amount of information about some latent variables. These features are in general non-linear functions, and are particularly useful in processing high-dimensional observed data. The alternating conditional expectations (ACE) algorithm is an efficient way to compute these maximal correlation functions. In this paper, we use an information theoretic approach to interpret the ACE algorithm as computing the singular value decomposition of a linear map between spaces of probability distributions. With this approach, we demonstrate the information theoretic optimality of the ACE algorithm, analyze its convergence rate and sample complexity, and finally, generalize it to compute multiple pairs of correlation functions from samples.**

I. Introduction

The Hirschfeld-Gebelein-Rényi 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 [1]. We commence by formally defining this dependence measure as much of our ensuing discussion will be motivated by it.

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:

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

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. It turns out that the variational formulation of maximal correlation in Definition 1 shares deep ties with a class of statistical inference problems. We consider inference problems with the general structure of a Markov chain $U \to X \to Y$. Here, U represents some feature of the data that we wish to make decisions on. This feature is embedded in some data X, and we only get to observe a noisy version of this data Y. We refer to X as the latent variable, Y as the noisy observation, and the conditional distributions $P_{Y|X}$ as the observation model. A natural way to solve this inference problem is to learn the statistical model $P_{Y|U}$, i.e. a combination of the embedding of the feature U and the observation model, so that we can directly extract the information about U from the observations.

Unfortunately, this approach is difficult to use in many applications. For example, in the "Netflix problem," if we let X be the user ID and Y the movie ID, it is challenging to identify what feature U of a user is relevant to his or her choice of movies. A different approach to such problems is to focus only on the observation model. We try to find features of the observation Y that carries as much information about X as possible, and yet are simple enough so that further processing, such as clustering and kernel methods, can be applied to make the final decision. Most dimension reduction algorithms follow this approach. For example, one way to establish the optimality of principal component analysis (PCA) is to assume that X is Gaussian distributed, and passes through an observation model that adds white Gaussian noise. In this case, the principle components of the observed Y can be shown to carry the maximum amount of information about X.

We can interpret maximal correlation as a general formulation for this approach. The optimization problem in Definition 1 tries to find a feature $g(Y)$ that is highly correlated with some feature $f(X)$, or equivalently, has high predictive power towards some aspects of X. The advantage of finding such a feature is that $g(Y)$ can be a general real-valued function. In particular, it need not be a linear function of the data, and the data itself need not be real-valued (categorical data). Thus, this maximal correlation formulation provides a general basis to select features from high-dimensional data. Our goal is to extend this framework and develop practical algorithms. In the ensuing discussion, we will present efficient algorithms that solve the optimization problem in Definition 1 for real-world data, show that both the formulation of maximal correlation and the associated algorithms can be generalized to produce an arbitrary number of features, and demonstrate that the resulting approach is indeed different, superior, and more general compared to existing methods such as PCA.

II. The Geometry of Maximal Correlation

We now develop a geometric structure that offers an alternative view of the maximal correlation problem. Let $\mathcal{P}_\mathcal{X}$ and $\mathcal{P}_\mathcal{Y}$ denote spaces of distributions on $\mathcal{X}$ and $\mathcal{Y}$ respectively, where $|\mathcal{X}|, |\mathcal{Y}| < \infty$. Consider the observation model, $P_{Y|X} : \mathcal{P}_\mathcal{X} \to \mathcal{P}_\mathcal{Y}$, as a map that takes $P_X \in \mathcal{P}_\mathcal{X}$ to

$P_Y \in \mathcal{P}_\mathcal{Y}$: $\forall y \in \mathcal{Y},\ P_Y(y) = \sum_{x\in\mathcal{X}} P_{Y|X}(y|x)P_X(x)$, which we write in vector notation as $P_Y = P_{Y|X}\cdot P_X$. It turns out to be inconvenient to consider probability distributions P_X and P_Y as vectors. Instead, we select a reference distribution $P_{0,X} \in \text{relint}(\mathcal{P}_\mathcal{X})$, and for every $P_X \in \mathcal{P}_\mathcal{X}$, we write:

$$\forall x \in \mathcal{X},\ \ P_X(x) = P_{0,X}(x) + \sqrt{P_{0,X}(x)}\,\phi(x) \tag{1}$$

where ϕ is a *spherical perturbation vector* in $\mathbb{R}^{|\mathcal{X}|}$ satisfying:

$$\sum_{x\in\mathcal{X}} \sqrt{P_{0,X}(x)}\,\phi(x) = 0. \tag{2}$$

This defines a one-to-one correspondence, $P_X \leftrightarrow \phi$, between probability distributions in the neighborhood of $P_{0,X}$ and associated spherical perturbation vectors. Using this correspondence, we can think of the neighborhood of distributions around $P_{0,X}$ as a vector space $\Omega_X \triangleq \{\phi : \text{satisfying (2)}\}$. To condense notation, we also write (1) in vector form as:

$$P_X = P_{0,X} + \left[\sqrt{P_{0,X}}\right]\cdot\phi \tag{3}$$

where $\left[\sqrt{P_{0,X}}\right]$ is a diagonal matrix with entries $P_{0,X}(x)$, $x \in \mathcal{X}$, and P_X, $P_{0,X}$, and ϕ are all treated as column vectors. The reason for the seemingly unnatural choice of vector space Ω_X is the following result.

Lemma 1. *If $P_1^{(\epsilon)}$ and $P_2^{(\epsilon)}$ are two distributions in the neighborhood of $P_{0,X}$, with $P_1^{(\epsilon)} \leftrightarrow \epsilon\phi_1$ and $P_2^{(\epsilon)} \leftrightarrow \epsilon\phi_2$, then:*

$$D(P_1^{(\epsilon)}||P_2^{(\epsilon)}) = \frac{1}{2}\epsilon^2\|\phi_1 - \phi_2\|_2^2 + o\left(\epsilon^2\right).$$

Lemma 1 follows from the second order Taylor approximation of Kullback-Leibler (KL) divergence. The result does not change with the choice of $P_{0,X}$ in the neighborhood, or if we switch the order of $P_1^{(\epsilon)}$ and $P_2^{(\epsilon)}$. It portrays that the squared Euclidean ℓ_2-norm on Ω_X is a good approximation of the KL divergence when we focus on a small neighborhood of distributions. In the same spirit, we will also use the Euclidean inner product on Ω_X to describe projections and orthogonality in the variations of probability distributions.

We note that Ω_X can also be viewed as functional space over $\mathcal{X}$. The functions we often use in inference problems are the log-likelihood functions. Again considering $P_1^{(\epsilon)}$ and $P_2^{(\epsilon)}$ as defined above, we can write $\forall x \in \mathcal{X}$:

$$\begin{aligned} L_i(x) = \log\left(\frac{P_i^{(\epsilon)}(x)}{P_{0,X}(x)}\right) &= \log\left(\frac{P_{0,X}(x) + \epsilon\sqrt{P_{0,X}(x)}\phi_i(x)}{P_{0,X}(x)}\right) \\ &= \epsilon\,\frac{1}{\sqrt{P_{0,X}(x)}}\,\phi_i(x) + O\left(\epsilon^2\right) \end{aligned}$$

for $i = 1, 2$, where $\log(\cdot)$ denotes the natural logarithm. It can be easily verified that the log-likelihood ratio, $\log(P_1^{(\epsilon)}/P_2^{(\epsilon)})$, is associated in a similar manner to the vector $\phi_1 - \phi_2$. This establishes a three-way association between a distribution $P_X \in \mathcal{P}_\mathcal{X}$ that lies in a neighborhood of $P_{0,X}$, a vector $\phi \in \Omega_X$, and a function that we can evaluate over samples, $L : \mathcal{X} \to \mathbb{R}$, $L(\cdot) \triangleq \phi(\cdot)/\sqrt{P_{0,X}(\cdot)}$, which satisfies:

$$\mathbb{E}_{P_{0,X}}\left[L(X)\right] = \sum_{x\in\mathcal{X}} P_{0,X}(x)L(x) = 0. \tag{4}$$

We refer to such functions as *score functions*. They are used to extract features from data. In this sense, the restriction (4) is reasonable as adding a constant to a score function does not help extract any useful information. For the same reason, we also restrict score functions to be normalized:

$$\mathbb{E}_{P_{0,X}}\left[L^2(X)\right] = \sum_{x\in\mathcal{X}} P_{0,X}(x)L^2(x) = 1 \tag{5}$$

which is equivalent to restricting the corresponding vector ϕ to have unit norm.

As an example of how to exploit this structure, suppose we observe a sequence of samples $x_1, \ldots, x_n$, and wish to evaluate the log-likelihood ratio between $P_1 \leftrightarrow \phi_1$ and $P_2 \leftrightarrow \phi_2$:

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{P_1(x_i)}{P_2(x_i)}\right) = \widehat{\mathbb{E}}_n\left[L_1(X) - L_2(X)\right]$$

where $\widehat{\mathbb{E}}_n$ denotes expectation with respect to the empirical distribution $\widehat{P}_{x_1^n}$ of the samples. For large n, $\widehat{P}_{x_1^n}$ is typically restricted to a small neighborhood around some nominal distribution $P_{0,X}$, which represents our prior knowledge of the samples. Thus, we can associate $\widehat{P}_{x_1^n}$ with a corresponding perturbation vector ψ, and write:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{P_1(x_i)}{P_2(x_i)}\right) &= \sum_{x\in\mathcal{X}} \widehat{P}_{x_1^n}(x)\left(L_1(x) - L_2(x)\right) \\ &= \langle\psi, \phi_1 - \phi_2\rangle \end{aligned} \tag{6}$$

where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product, and the second equality follows from letting $\widehat{P}_{x_1^n} = P_{0,X} + \left[\sqrt{P_{0,X}}\right]\psi$, representing L_1 and L_2 by ϕ_1 and ϕ_2 respectively, and using (4). This result has an elegant geometric interpretation: evaluating the empirical average of the log-likelihood function is equivalent to projecting the displacement of the empirical distribution from the prior knowledge $P_{0,X}$ to the direction $\phi_1 - \phi_2$. In other words, we monitor the variation in the empirical distribution only along a specific direction: the one that is relevant to making decisions between P_1 and P_2.

For inference problems with memoryless models, the order of the samples is irrelevant in the decision making. So, the information of the data is carried by its empirical distribution. Evaluating the empirical average of various score functions on the data can therefore be viewed as monitoring the variations of the empirical distribution along different directions, or equivalently, extracting different partial information. When we know how the desired feature is statistically "encoded" in the data, we know which part of the information is "useful." For example, if $P_{X|U=1} = P_1$ and $P_{X|U=2} = P_2$, then (6) is a sufficient statistic for U, and orthogonal components of ψ can be discarded. Without this knowledge, we cannot deem any part of the information as irrelevant. However, processing, storage, or communication constraints often compel us to discard some partial information, as is typical in Big Data problems. Intelligently doing this without severely degrading performance requires new performance criteria and analytic structures to decompose information into parts that can potentially be dissipated. The geometric structure we introduced addresses such lossy information processing problems.

With our geometric structure, decomposing information reduces to decomposing the vector $\psi \in \Omega_X$. We fix an orthonormal basis $\{u_1, \ldots, u_{|\mathcal{X}|-1}\}$ of Ω_X, and compute the inner products $\langle \psi, u_i \rangle$ for $1 \le i \le |\mathcal{X}| - 1$. Each u_i corresponds to a score function: $\forall x \in \mathcal{X}$, $f_i(x) = u_i(x)/\sqrt{P_{0,X}(x)}$, and the orthogonality of $\{u_1, \ldots, u_{|\mathcal{X}|-1}\}$ implies that different score functions are uncorrelated with respect to $P_{0,X}$:

$$\delta_{ij} = \langle u_i, u_j \rangle = \sum_{x \in \mathcal{X}} P_{0,X}(x) f_i(x) f_j(x)$$

for every i, j, where δ_{ij} is the Kronecker delta. We refer to the inner products $\langle \psi, u_i \rangle$ as *scores* (not to be confused with score functions f_i). By the completeness of $\{u_1, \ldots, u_{|\mathcal{X}|-1}\}$, we can recover ψ if we collect all the scores. Hence, the set of real-valued scores, $\langle \psi, u_i \rangle = \widehat{\mathbb{E}}_n [f_i(X)]$, $1 \le i \le |\mathcal{X}| - 1$, is a different decomposition of the empirical distribution $\widehat{P}_{x_1^n}$.

If the data samples $X_1, \ldots, X_n$ are actually i.i.d. from $P_{0,X}$, and n is large enough so that the central limit theorem becomes a good approximation, then the scores, $\langle \psi, u_i \rangle$ for $1 \le i \le |\mathcal{X}| - 1$, are i.i.d. Gaussian distributed with variance proportional to $1/n$ (in an exponential followed by local approximation sense) [2]. At this point, there is no reason to believe any score is more valuable or informative than any other score. This remains true irrespective of our choice of basis. The story is different, however, if we observe the data through a memoryless observation model $P_{Y|X} : \mathcal{P}_\mathcal{X} \to \mathcal{P}_\mathcal{Y}$. We construct the vector spaces Ω_X and Ω_Y by choosing reference distributions $P_{0,X} \in \text{relint}(\mathcal{P}_\mathcal{X})$ and $P_{0,Y} \in \text{relint}(\mathcal{P}_\mathcal{Y})$ respectively, such that $P_{0,Y} = P_{Y|X} \cdot P_{0,X}$. Using the notation in (3), we can write $P_{Y|X}$ as a map from Ω_X to Ω_Y:

$$\begin{aligned} & P_Y = P_{Y|X} \cdot P_X \\ \Leftrightarrow \quad & P_{0,Y} + \left[\sqrt{P_{0,Y}}\right] \cdot \phi_Y = P_{Y|X} \cdot \left(P_{0,X} + \left[\sqrt{P_{0,X}}\right] \cdot \phi_X \right) \\ \Leftrightarrow \quad & \phi_Y = \left[\sqrt{P_{0,Y}}\right]^{-1} \cdot P_{Y|X} \cdot \left[\sqrt{P_{0,X}}\right] \cdot \phi_X \end{aligned}$$

where $P_X \leftrightarrow \phi_X$ and $P_Y \leftrightarrow \phi_Y$, and we define:

$$B \triangleq \left[\sqrt{P_{0,Y}}\right]^{-1} \cdot P_{Y|X} \cdot \left[\sqrt{P_{0,X}}\right] \tag{7}$$

as the map from Ω_X to Ω_Y. If we assume that the local approximation condition of Lemma 1 holds, then the variation ϕ_X between P_X and $P_{0,X}$ has a KL divergence proportional to $\|\phi_X\|_2^2$, and the induced variation ϕ_Y between P_Y and $P_{0,Y}$ has a KL divergence proportional to $\|\phi_Y\|_2^2$. We observe that for any variation $\phi_X \in \Omega_X$ with fixed KL divergence $\|\phi_X\|_2^2 = \delta$, the KL divergence of the induced variation, $\|\phi_Y\|_2^2$, depends on the direction of ϕ_X and the singular value decomposition (SVD) of B. Equivalently, depending on the SVD of B, some input features are corrupted severely by the noisy observation model, while others are more observable from the output end. Hence, we refer to B as the *divergence transition matrix* (DTM) of the observation model.

In the definition of the DTM (7), we did not take into account that both ϕ_X and ϕ_Y must satisfy the constraint (2). However, we can easily verify that $u_0 \triangleq [\sqrt{P_{0,X}(x)}, \forall x \in \mathcal{X}]^T$ and $v_0 \triangleq [\sqrt{P_{0,Y}(y)}, \forall y \in \mathcal{Y}]^T$ are the right and left singular vectors respectively, corresponding to a singular value of $\sigma_0 = 1$ for any DTM B. Translating this to the language of functional spaces, u_0 and v_0 correspond to constant functions over $\mathcal{X}$ and $\mathcal{Y}$ respectively. We could have incorporated constraint (2) into the definition of B using projections to the orthogonal complement subspaces of u_0 and v_0, but it is convenient to define B as we have. So, we will simply keep in mind that perturbation vectors in Ω_X and Ω_Y are orthogonal to u_0 and v_0, respectively, which means that they are spanned by the remaining right and left singular vectors of B, respectively. The next result presents some properties of the singular values and vectors of the DTM.

Theorem 2. *For any observation model $P_{Y|X}$, input reference distribution $P_{0,X} \in$ relint$(\mathcal{P}_\mathcal{X})$, output reference distribution $P_{0,Y} \in$ relint$(\mathcal{P}_\mathcal{Y})$, and corresponding DTM defined in (7), the following are true:*

1) *u_0 and v_0 are always a pair of right and left singular vectors respectively, corresponding to $\sigma_0 = 1$.*
2) *All other singular values satisfy: $1 = \sigma_0 \ge \sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_{K-1} \ge 0$, where $K = \min\{|\mathcal{X}|, |\mathcal{Y}|\}$.*
3) *$\sigma_1 = \rho(X;Y)$ is the maximal correlation, and the functions corresponding to its right and left singular vectors, u_1 and v_1, respectively:*

$$\forall x \in \mathcal{X}, \quad f^*(x) = \frac{u_1(x)}{\sqrt{P_{0,X}(x)}}$$

$$\forall y \in \mathcal{Y}, \quad g^*(y) = \frac{v_1(y)}{\sqrt{P_{0,Y}(y)}}$$

are the maximal correlation functions solving the extremal problem in Definition 1 of $\rho(X;Y)$.

4) *$\sigma_1 = \rho(X;Y)$ is also characterized by:*

$$\rho^2(X;Y) = \lim_{\epsilon \to 0} \sup_{\substack{Q_X \in \mathcal{P}_\mathcal{X}: \\ D(Q_X||P_{0,X}) = \frac{1}{2}\epsilon^2}} \frac{D(Q_Y||P_{0,Y})}{D(Q_X||P_{0,X})} \tag{8}$$

where $Q_Y \in \mathcal{P}_\mathcal{Y}$ satisfies $Q_Y = P_{Y|X} \cdot Q_X$ for any $Q_X \in \mathcal{P}_\mathcal{X}$. The supremum in (8) is achieved by choosing Q_X as a perturbation of $P_{0,X}$ along the direction of u_1, i.e. $Q_X = P_{0,X} + \epsilon \left[\sqrt{P_{0,X}}\right] \cdot u_1$.

We refer readers to [3] and the references therein for proofs of these properties. The fourth property (8) in Theorem 2 provides an information theoretic interpretation of maximal correlation and the geometric structure we described. It states that two distributions $P_{0,X}$ and Q_X become less distinguishable when passing through a noisy observation model, $P_{Y|X}$, in the sense that the KL divergence is reduced: $D(Q_Y||P_{0,Y}) \le D(Q_X||P_{0,X})$. This is simply the data processing inequality. More precisely, (8) illustrates that as long as $P_{0,X}$ and Q_X are in a small neighborhood, the reduction in the KL divergence is minimized when the difference between $P_{0,X}$ and Q_X is along the first right singular vector u_1. This is a tighter variant of the data processing inequality. So, depending on the SVD structure of B, different features of X are corrupted differently by the noisy observation. The feature that is least corrupted is the one corresponding to u_1 and v_1, i.e. the one that can be extracted by using the score functions f^* and g^*. Hence, returning to the

dimension reduction problem, when we do not know how the desired feature U is encoded in X but know the observation model $P_{Y|X}$, if we can only compute the empirical average of a single score function to capture some information about U, then the maximal correlation function g^* corresponding to the left singular vector v_1 is a sensible choice.

We remark that given a finite number of samples $y_1, \ldots, y_n$, the finite length realization of the observation model also adds extra noise to the empirical distribution of the observed samples $\widehat{P}_{y_1^n}$ [2]. It is proven in [4] that the distribution of this noise also depends on the SVD of B. Nonetheless, it is still true that features of X along the singular vectors with larger singular values are less corrupted by the observation channel.

The result (8) depends critically on the local assumption that P_X and Q_X are close. For example, in [5], a slightly different formulation with a ratio of mutual information terms was studied, and a rather different result was derived as the local assumption fails in this case. This nuance is elucidated in [3]. In practice, we often apply the score functions derived from the SVD structure to data without considering how well the local assumption holds. This is a heuristic engineering choice which is taken because it leads to structured and efficient algorithms.

Theorem 2 can be generalized to the following result, which addresses the problem of extracting $k \geq 1$ features from data.

Proposition 3. *For every $1 \leq k \leq K-1$, we have:*

$$\sigma_k = \max_{\substack{\{f_i : i=1,\ldots,k\} \\ \{g_i : i=1,\ldots,k\}}} \min_{1 \leq i \leq k} \mathbb{E}[f_i(X) g_i(Y)]$$

where the maximization is over all possible pairs of score functions with the following constraints:

$$\forall i \in \{1,\ldots,k\}\,,\ f_i : \mathcal{X} \to \mathbb{R},\ \ g_i : \mathcal{Y} \to \mathbb{R}$$
$$\forall i \in \{1,\ldots,k\}\,,\ \mathbb{E}[f_i(X)] = \mathbb{E}[g_i(Y)] = 0$$
$$\forall i,j \in \{1,\ldots,k\}\,,\ \mathbb{E}[f_i(X) f_j(X)] = \mathbb{E}[g_i(Y) g_j(Y)] = \delta_{ij}$$

and the expectations are taken over $P_{0,X}$ and $P_{0,Y}$. The optimizing score functions satisfy: $\forall x \in \mathcal{X}$, $f_i^*(x) = u_i(x)/\sqrt{P_{0,X}(x)}$, and $\forall y \in \mathcal{Y}$, $g_i^*(y) = v_i(y)/\sqrt{P_{0,Y}(y)}$, where u_i and v_i are the k pairs of right and left singular vectors of B corresponding to $\sigma_1, \ldots, \sigma_k$, respectively.

Proposition 3 is straightforward to prove using the SVD. We omit a statement regarding the information theoretic optimality of f_i^* and g_i^* akin to (8). Regardless, Proposition 3 suggests that upon observing a sequence of samples $y_1, \ldots, y_n$, if we are allowed to compute k scores, we should compute the empirical averages of $g_1, \ldots, g_k$, which correspond to singular vectors of B. These scores are the most informative in the sense of (8), and relate one-to-one to a collection of k features of the X sequence in the sense of Proposition 3.

III. Efficient Algorithms to Find Score Functions

In the approach we have presented so far, finding "optimal" score functions is equivalent to computing the first few singular vectors of B. As we are interested in problems where $\mathcal{X}$ and $\mathcal{Y}$ are very large, both computing the SVD of B and estimating B itself are often formidable tasks. Fortunately, with the help of the geometric structure we described, both these issues can be resolved with computationally efficient procedures. In this section, we will focus on computing a single score function g^* corresponding to σ_1. The next section will address the problem of finding multiple score functions.

Computing g^* is equivalent to finding the maximal correlation. This is a well-known problem in the literature, and a standard solution is the Alternating Conditional Expectations (ACE) algorithm [6]. We now delineate this algorithm using the SVD notation we introduced. For a $K \times K$ real matrix A (taken to be square without loss of generality) with ordered singular values $\sigma_0 \geq \sigma_1 \geq \ldots \geq \sigma_{K-1}$ and corresponding normalized right singular vectors $u_0, u_1, \ldots, u_{K-1} \in \mathbb{R}^K$, we can find u_0 using the *power method* from numerical linear algebra [7]. We start with an arbitrary vector $\phi \in \mathbb{R}^K$, and repeatedly multiply $A^T A$ to it. Since $A^T A = \sum_{i=0}^{K-1} \sigma_i^2 u_i u_i^T$ by the spectral theorem, and $\phi = \sum_{i=0}^{K-1} \alpha_i u_i$ for some $\alpha_i \in \mathbb{R}$ as $\{u_0, \ldots, u_{K-1}\}$ is an orthonormal basis, we can write: $\left(A^T A\right)^m \cdot \phi = \sum_{i=0}^{K-1} \sigma_i^{2m} \alpha_i u_i$. Assuming $\alpha_0 \neq 0$, as m becomes large, the component corresponding to σ_0 will dominate the sum, and the resulting vector is aligned with u_0. In practice, we scale the intermediate vectors to have unit norm once every few iterations for numerical stability. The power method converges geometrically (exponentially) with ratio σ_1^2/σ_0^2. We will ignore the $\sigma_0 = \sigma_1$ case, but even in this case, the power method outputs some linear combination of u_0 and u_1. Moreover, after computing u_0, we can compute u_1 by selecting an initial guess ϕ that is orthogonal to u_0.

We use this approach on the DTM B. Let our initial guess be $\phi \in \mathbb{R}^{|\mathcal{X}|}$ with corresponding score function $\forall x \in \mathcal{X}$, $f(x) = \phi(x)/\sqrt{P_{0,X}(x)}$, and let $\psi = B \cdot \phi \in \mathbb{R}^{|\mathcal{Y}|}$ with corresponding score function $\forall y \in \mathcal{Y}$, $g(y) = \psi(y)/\sqrt{P_{0,Y}(y)}$. Then, using (7), we have for every $y \in \mathcal{Y}$:

$$\begin{aligned} g(y) &= \frac{\psi(y)}{\sqrt{P_{0,Y}(y)}} = \frac{1}{\sqrt{P_{0,Y}(y)}} \sum_{x \in \mathcal{X}} B(x,y)\, \phi(x) \\ &= \frac{1}{\sqrt{P_{0,Y}(y)}} \sum_{x \in \mathcal{X}} \frac{P_{Y|X}(y|x)\sqrt{P_{0,X}(x)}}{\sqrt{P_{0,Y}(y)}} \sqrt{P_{0,X}(x)} f(x) \\ &= \mathbb{E}\left[f(X) | Y = y\right]. \end{aligned}$$

So, multiplying ϕ by B is equivalent to taking the conditional expectation of f: $\mathbb{E}[f(X)|Y=y]$. Likewise, multiplying ψ by B^T is equivalent to taking the other conditional expectation: $\mathbb{E}[g(Y)|X=x]$. Thus, the following ACE algorithm precisely solves for the singular vectors of B corresponding to σ_1.

Algorithm 1 ACE Algorithm

Require: knowledge of $P_{X,Y}$

1. Initialize: randomly pick $g(y)$, $y \in \mathcal{Y}$
   Center: $g(y) \leftarrow g(y) - \mathbb{E}[g(Y)]$

repeat

2a. $f(x) \leftarrow \mathbb{E}[g(Y)|X=x]$, $\forall x \in \mathcal{X}$

2b. $g(y) \leftarrow \mathbb{E}[f(X)|Y=y]$, $\forall y \in \mathcal{Y}$

2c. Regularize: $g(y) \leftarrow g(y)/\sqrt{\mathbb{E}\left[g^2(Y)\right]}$, $\forall y \in \mathcal{Y}$

until $\mathbb{E}[f(X)g(Y)]$ stops to increase

In Algorithm 1, the initial choice of $g(Y)$ is constrained to have zero mean. This is equivalent to setting ψ to be orthogonal to v_0, which corresponds to the constant function on $\mathcal{Y}$. This centering needs to be done only once at the initialization step, since we have $\mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0$ in all of the following steps. Moreover, the regularization step 2c does not have to be performed in every iteration; it is needed only once in a while to avoid arithmetic underflow.

In practice, the main obstacle of applying this algorithm is the requirement of the knowledge of $P_{X,Y}$. Especially for cases with large alphabets $\mathcal{X}$ and $\mathcal{Y}$, estimating $P_{X,Y}$ can require a lot of samples. A natural alternative is to replace the conditional expectations in steps 2a and 2b by empirical conditional averages. This gives the following algorithm.

Algorithm 2 ACE Algorithm with Finite Samples

Require: training samples $\{(x_i, y_i) : i = 1, \ldots, N\}$

1. Initialize: randomly pick $g(y)$, $y \in \mathcal{Y}$

repeat: pick a subset of n samples

2a. $f(x) \leftarrow \widehat{\mathbb{E}}_n[g(Y)|X = x]$, $\forall x \in \mathcal{X}$

2b. $g(y) \leftarrow \widehat{\mathbb{E}}_n[f(X)|Y = y]$, $\forall y \in \mathcal{Y}$

2c. Regularize: $g(y) \leftarrow g(y) - \widehat{\mathbb{E}}_n[g(Y)]$, $\forall y \in \mathcal{Y}$

$g(y) \leftarrow g(y)/\sqrt{\widehat{\mathbb{E}}_n[g^2(Y)]}$, $\forall y \in \mathcal{Y}$

until $\widehat{\mathbb{E}}_n[f(X)g(Y)]$ stops to increase

There are two main differences between Algorithm 2 and Algorithm 1. Firstly, the centering step that forces $g(y)$ to be zero mean is removed from the initialization step and performed repeatedly. This is because the empirical distribution $\widehat{P}_{x_1^n, y_1^n}$ might be different from $P_{X,Y}$. Thus, a non-zero mean might be introduced in the empirical averaging steps. As $\sigma_0 = 1$ is the largest singular value, we need to periodically prune the component along v_0; otherwise, it would dominate the resulting function. Secondly, to simplify the analysis, we assume that in each iteration we use a subset of n samples, and that these subsets are non-overlapping from one iteration to another. In practice, it is easy to use bootstrapping methods to reuse some of the samples. The key to the success of this algorithm is to select n large enough so that the empirical averages are close to the true conditional expectations. In our experiments, the algorithm converges exponentially fast. So, only a few iterations are typically required for convergence, and the difference between N and n is insignificant.

A. Sample Complexity of Learning the Maximal Correlation Functions

Since we are particularly interested in problems where $|\mathcal{X}|$ and $|\mathcal{Y}|$ are large and comparable, we assume that $|\mathcal{X}| = |\mathcal{Y}| = K$ for convenience. We further assume that the number of samples N is much larger than K, which means estimating the marginal distributions P_X and P_Y is manageable, but not large enough to provide an accurate estimate of the K^2-dimensional joint distribution. This allows us to make the simplifying assumption that P_X and P_Y are given.

It is notoriously difficult to estimate large dimensional distributions like $P_{X,Y}$. This is because a few of the entries, $P_{X,Y}(x, y)$ for some $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, are inevitably very small. Such entries appear very infrequently in the samples, and thus, many samples are needed to see the frequency of these values. We can perceive the ACE algorithm on samples as an effort to circumvent this situation. It tries only to estimate a part of the joint distribution, namely, the component corresponding to the second largest singular value and its corresponding singular vectors. We can then use this partial knowledge of $P_{X,Y}$ for the purposes of inference. Intuitively, we should expect the ACE algorithm to require fewer training samples than algorithms that attempt a full estimation of $P_{X,Y}$.

To rigorize this intuition, we consider the following estimation problem. Suppose (x_i, y_i), $i = 1, \ldots, n$, are drawn i.i.d. from an unknown joint distribution $P_{X,Y}$, and we are interested in estimating $\mathbb{E}[f(X)g(Y)]$ from these samples for a given pair of functions $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ satisfying:

$$\mathbb{E}\left[f^2(X)\right] = \mathbb{E}\left[g^2(Y)\right] = 1. \tag{9}$$

Note that we do not restrict the functions to be zero-mean. So, the functions can have components along u_0 and v_0, respectively. We seek to compute the rate at which the empirical average $\widehat{\mathbb{E}}_n[f(X)g(Y)]$ converges, and quantify how this rate varies for different choices of f and g. Specifically, we study the maximal correlation functions f^* and g^*, and the functions $\forall x \in \mathcal{X}$, $\check{f}(x) \triangleq \mathcal{I}_{x=x_0}(x)/\sqrt{P_X(x_0)}$ and $\forall y \in \mathcal{Y}$, $\check{g}(y) \triangleq \mathcal{I}_{y=y_0}(y)/\sqrt{P_Y(y_0)}$ for an arbitrary choice of $x_0 \in \mathcal{X}$ and $y_0 \in \mathcal{Y}$, where $\mathcal{I}$ denotes the indicator function. It can be checked that $\check{f}$ and $\check{g}$ satisfy (9), and that their true correlation is:

$$\mathbb{E}\left[\check{f}(X)\check{g}(Y)\right] = \frac{P_{X,Y}(x_0, y_0)}{\sqrt{P_X(x_0)P_Y(y_0)}} = B(x_0, y_0).$$

As we assume that both P_X and P_Y are precisely given, we treat the estimation of this correlation as the same as the estimation of the entry $P_{X,Y}(x_0, y_0)$ of the joint distribution. The correlation between f^* and g^* is generally high due to Definition 1. In contrast, most $\check{f}$ and $\check{g}$ of interest have smaller correlation than f^* and g^*. For problems with large alphabets, this gap is particularly significant. We will argue that because of this gap, the estimation of the maximal correlation, $\rho(X;Y)$, indeed requires a significantly smaller number of samples than the estimation of a particular entry $P_{X,Y}(x_0, y_0)$. In the ensuing discussion, we will retain the SVD notation we have defined for B with P_X replacing $P_{0,X}$ and P_Y replacing $P_{0,Y}$.

To quantify the sample complexity of an estimation, we consider the following criterion. For any given pair of functions f, g, we compute the number of samples, n, required such that the probability:

$$\mathbb{P}\left(\left|\frac{\widehat{\mathbb{E}}_n\left[f(X)g(Y)\right]}{\mathbb{E}\left[f(X)g(Y)\right]} - 1\right| \geq \Delta\right) \leq \gamma \tag{10}$$

for some small values $\Delta > 0$ and $\gamma > 0$. So, we allow the n-sample empirical average to differ by a factor of $(1 \pm \Delta)$ from the true value. We use this criterion to study the convergence of Algorithm 2. Consider a particular iteration where we have a guess $f : \mathcal{X} \to \mathbb{R}$ of the maximal correlation function.

Although f is not necessarily equal to f^*, we can decompose it with respect to the basis defined by the singular vectors of the DTM. To this end, we let $\phi \in \Omega_X$ be the perturbation vector corresponding to f: $\forall x \in \mathcal{X}$, $\phi(x) = f(x)\sqrt{P_X(x)}$, and expand ϕ in the form $\phi = \sum_{i=1}^{K-1} \alpha_i u_i$, or equivalently:

$$\forall x \in \mathcal{X}, \;\; f(x) = \sum_{i=1}^{K-1} \alpha_i f_i^*(x)$$

where each f_i^*, defined by $\forall x \in \mathcal{X}$, $f_i^*(x) = u_i(x)/\sqrt{P_X(x)}$, corresponds to the ith right singular vector of B, and the coefficients $\{\alpha_i : i = 1, \ldots, K-1\}$ can be computed using the inner products: $\alpha_i = \langle \phi, u_i \rangle = \mathbb{E}[f(X) f_i^*(X)]$.

After a round of empirical conditional expectation calculations, we get $g : \mathcal{Y} \to \mathbb{R}$ given by $\forall y \in \mathcal{Y}$, $g(y) = \widehat{\mathbb{E}}_n[f(X)|Y = y]$. As before, we decompose g with respect to the left singular vectors of B, which gives:

$$\forall y \in \mathcal{Y}, \;\; g(y) = \sum_{j=1}^{K-1} \beta_j g_j^*(y)$$

where each g_j^* is given by: $\forall y \in \mathcal{Y}$, $g_j^*(y) = v_j(y)/\sqrt{P_Y(y)}$. Moreover, we can write out each coefficient as:

$$\begin{aligned}
\beta_j &= \mathbb{E}\left[g(Y) g_j^*(Y)\right] \\
&= \sum_{y \in \mathcal{Y}} P_Y(y)\, g_j^*(y)\, \widehat{\mathbb{E}}_n \left[\sum_{i=1}^{K-1} \alpha_i f_i^*(X) \middle| Y = y \right] \\
&= \sum_{i=1}^{K-1} \alpha_i \widehat{\mathbb{E}}_n \left[f_i^*(X) g_j^*(Y) \right]
\end{aligned}$$

where we assume that the difference between the empirical marginal distribution $\widehat{P}_{y_1^n}$ and the true P_Y is negligible.

Thus, if n is large enough such that $\forall i, j \in \{1, \ldots, K-1\}$, $\widehat{\mathbb{E}}_n \left[f_i^*(X) g_j^*(Y) \right]$ is close to $\mathbb{E}\left[f_i^*(X) g_j^*(Y) \right] = \sigma_i \delta_{ij}$, then the orthogonal expansion coefficients corresponding to f_1^* and g_1^* dominate over iterations of the algorithm, and the algorithm converges to the maximal correlation functions exponentially fast. The sample complexity of Algorithm 2 is therefore determined by the number of samples in a block, n, which must be large enough such that (10) is satisfied for sufficiently small Δ and γ. The following result presents a tight characterization of the n needed for this purpose.

Theorem 4. *For any random variables X and Y with joint distribution $P_{X,Y}$, if X and Y are not independent, then for any $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ satisfying (9), we have:*

$$\begin{aligned}
&- \lim_{\Delta \to 0^+} \frac{1}{\Delta^2} \lim_{n \to \infty} \frac{1}{n} \log \left(\mathbb{P}\left(\left| \frac{\widehat{\mathbb{E}}_n \left[f(X) g(Y) \right]}{\mathbb{E}\left[f(X) g(Y) \right]} - 1 \right| \geq \Delta \right) \right) \\
&\qquad = \frac{1}{2} \frac{\mathbb{E}\left[f(X) g(Y) \right]^2}{\mathbb{VAR}\left(f(X) g(Y) \right)}.
\end{aligned}$$

Theorem 4 illustrates that the large deviations rate of decay of the probability in (10) is inversely proportional to the squared *coefficient of variation* of $f(X)g(Y)$ as $\Delta \to 0^+$. We now compare the estimation of the correlations between f_1^* and g_1^*, and $\check{f}$ and $\check{g}$. Using Theorem 4, the ratio of the number of samples required to achieve the same precision level Δ and confidence $1 - \gamma$ (between $\check{f}$, $\check{g}$ and f_1^*, g_1^*) is:

$$G(B) \triangleq \frac{\mathbb{E}\left[\check{f}(X) \check{g}(Y) \right]^2}{\mathbb{E}\left[f_1^*(X) g_1^*(Y) \right]^2} \frac{\mathbb{VAR}\left(f_1^*(X) g_1^*(Y) \right)}{\mathbb{VAR}\left(\check{f}(X) \check{g}(Y) \right)}. \tag{11}$$

This ratio depends on the true distribution $P_{X,Y}$, and is thus written as a function of the DTM B. We observe that in the first term of (11), $\mathbb{E}[f_1^*(X) g_1^*(Y)]$ is larger than $\mathbb{E}[\check{f}(X) \check{g}(Y)]$ in most cases of interest. On the other hand, when we square f_1^* and g_1^* element-wise, intuitively, the property of high correlation is lost. Hence, we expect the ratio between the variances in the second term of (11) to be insignificant, which means $G(B) < 1$ in most cases. This portrays that the maximal correlation functions are not only good information carriers, but the correlation between them is also easier to estimate compared to other pairs of functions. Consequently, the ACE algorithm converges fast; it requires fewer samples than that needed to estimate an entry of the joint distribution.

Unfortunately, we cannot prove this for every DTM. In principle, it might be possible to construct an example where the ratio between the variances is also significant. However, we cannot determine whether we are given such an "unfortunate" DTM when we face a problem. It is our belief that such cases are rare when K is large. We demonstrate this using a numerical experiment, where we randomly generate $P_{X,Y}$ using i.i.d. exponential entries followed by normalization. For each $P_{X,Y}$, we compute the ratio $G(B)$ corresponding to $\check{f}(X)$ and $\check{g}(Y)$ defined by some randomly chosen element $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$. We then plot the average values of $G(B)$ with respect to K. Figure 1 indicates that the average $G(B)$ decreases as K increases, and the relationship between $\log(K)$ and $\log(G)$ is linear. Intuitively, for a random $K \times K$ matrix $P_{X,Y}$, the second largest singular value of B, $\rho(X;Y)$, scales as $1/\sqrt{K}$ and $\mathbb{E}[\check{f}(X)\check{g}(Y)] = B(x_0, y_0)$ scales as $1/K$. Thus, we expect a saving in the sample complexity by a factor of $1/K$. Indeed, the relationship found in our simulation is $G(B) \approx C\, K^{\alpha}$, where C is a constant and $\alpha \approx -1.05$.

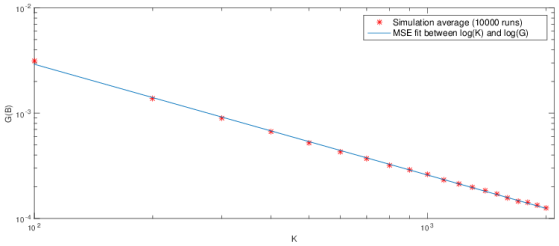

Fig. 1. Numerical experiment plot of average $G(B)$ as a function of K. In logarithmic axes, the relationship is almost linear, with an MSE fit of $\log(G) = -0.9921 - 1.0524 \log(K)$, and Pearson correlation of -0.9996.

B. Comparison to Principal Component Analysis

It is instructive to compare the ACE algorithm to PCA, because the two approaches have a clear resemblance in that the SVD is used in both. In PCA, the observed data are real-valued vectors $y_1, y_2, \ldots, y_n \in \mathbb{R}^m$. We stack these vectors together to form a matrix $Y = [y_1 \;\; y_2 \;\; \cdots \;\; y_n]$, and compute the

SVD of Y. Then, we can project each observed m-dimensional vector to $k < m$ leading left singular vectors of Y to form a reduced k-dimensional representation. In the ACE algorithm, we observe the samples $\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, \ldots, N\}$ where $|\mathcal{X}|, |\mathcal{Y}| < \infty$. We then compute a scaled version of the empirical conditional distribution for each $x \in \mathcal{X}$:

$$\forall y \in \mathcal{Y}, \;\; b_x(y) = \frac{\sqrt{P_X(x)}}{\sqrt{P_Y(y)}} \widehat{P}_{Y|X}(y|x) = \frac{\widehat{P}_{x_1^N, y_1^N}(x, y)}{\sqrt{P_X(x) P_Y(y)}}$$

where we assume the marginal distributions P_X and P_Y are given or precisely estimated from the data. These vectors are stacked together to form $B = [b_x, \, x \in \mathcal{X}]$, and we compute the SVD of B. The scores computed by the ACE algorithm are projections of $\widehat{P}_{y_1^N}$ (properly scaled) onto the leading left singular vectors of B. Hence, the two approaches are almost identical. The key difference is that in the ACE algorithm, we operate in the space of distributions rather than data. Consequently, "a strong advantage of the ACE procedure is the ability to incorporate variables of quite different type in terms of the set of values they can assume" [6].

Another way to relate PCA and the ACE algorithm stems from viewing the DTM as a map from the functional space over $\mathcal{X}$ to that over $\mathcal{Y}$. In the optimization of Definition 1, we look for general functions f and g such that $f(X)$ and $g(Y)$ are highly correlated. We can further restrict these functions to lie in linear subspaces of the functional spaces since we can still define the DTM as a linear map from a subspace of functions over $\mathcal{X}$ to a subspace of functions over $\mathcal{Y}$. The entire discussion regarding the SVD structure and iterative algorithms to compute optimal functions holds in this scenario.

A particular case of interest is when we have zero mean random vectors $\underline{X}, \underline{Y} \in \mathbb{R}^m$ that have covariance matrices K_X and K_Y, respectively, and cross-covariance matrix $K_{X,Y}$, and we constrain the correlation functions to be linear functions. With a little abuse of notation, we consider the linear functions $\forall \underline{x} \in \mathbb{R}^m$, $f(\underline{x}) = \underline{f}^T \underline{x}$ and $\forall \underline{y} \in \mathbb{R}^m$, $g(\underline{y}) = \underline{g}^T \underline{y}$ for some $\underline{f}, \underline{g} \in \mathbb{R}^m$. Then, we can specialize Definition 1 into:

$$\max_{\substack{\underline{f} \in \mathbb{R}^m, \, \underline{g} \in \mathbb{R}^m: \\ \mathbb{E}\left[(\underline{f}^T \underline{X})^2\right] = \mathbb{E}\left[(\underline{g}^T \underline{Y})^2\right] = 1}} \mathbb{E}\left[(\underline{f}^T \underline{X})(\underline{g}^T \underline{Y})\right]. \tag{12}$$

This is the setup of *canonical correlation analysis* (CCA), and the optimizing arguments are $\underline{f}^* = K_X^{-1/2} \underline{u}$ and $\underline{g}^* = K_Y^{-1/2} \underline{v}$, where $\underline{u}$ and $\underline{v}$ the right and left singular vectors of the matrix $C \triangleq K_X^{-1/2} K_{X,Y} K_Y^{-1/2}$, respectively [8]. This matrix resembles the definition of the DTM B in (7). If we wish to avoid directly solving the SVD, we can use a modified version of Algorithm 1. Since we may get a non-linear function after step 2a, $f(x) \leftarrow \mathbb{E}[g(Y)|X = x]$, $\forall x \in \mathcal{X}$, we project f onto the subspace of all linear functions using a pertinent inner product after step 2a. This is equivalent to minimizing the mean squared error: $\min_{\underline{f} \in \mathbb{R}^m} \mathbb{E}[(f(\underline{X}) - \underline{f}^T \underline{X})^2]$. Adding a similar projection step after step 2b, it is easy to verify that Algorithm 1 solves the CCA problem (12).

Such results with linear correlation functions rely only on the second moments of $\underline{X}$ and $\underline{Y}$. So, we can treat $\underline{X}$ and $\underline{Y}$ as though they are jointly Gaussian distributed (as commonly done in linear least squares estimation). Hence, the extremal problem in Definition 1 is a generalization of the Gaussian case. A further special case of CCA is when the noisy observation model actually adds white noise, i.e. $K_Y = K_X + \sigma^2 I$ and $K_{X,Y} = K_X$. This simplifies the CCA problem as the covariance matrices defining C are jointly diagonalizable. Consequently, $\underline{g}^* = K_Y^{-1/2} \underline{v}$ where $\underline{v}$ is given by the first eigenvector of K_Y; this is consistent with PCA.

IV. Finding Multiple Score Functions in Parallel

As we have discussed, when we use the ACE algorithm as a dimension reduction tool, our goal is to identify the feature of Y that is most correlated with some feature of X. This is meaningful in applications where we do not know how the desired feature U is embedded in the data X. So, we reduce the observed data Y to a manageable number of scores that capture sufficient information about U. It is therefore critical that we can compute any $k \geq 1$ of the most informative scores. On this front, the existing literature mostly follows the original formulation [6], where a generalization of the ACE algorithm finds a feature $f(X)$ and multiple score functions $g_j(Y_j)$, $j = 1, \ldots, p$, each operating on a part of the observed data Y_j, $j = 1, \ldots, p$, such that $\mathbb{E}[(f(X) - \sum_{j=1}^p g_j(Y_j))^2]$ is minimized.

With the geometric structure developed in this paper, a more natural approach is to find $k \geq 1$ pairs of score functions $f_i : \mathcal{X} \to \mathbb{R}$, $g_i : \mathcal{Y} \to \mathbb{R}$, $i = 1, \ldots, k$, such that each $g_i(Y)$ carries information about the corresponding $f_i(X)$. To avoid redundancy between score functions, we impose the additional constraints: $\forall i, j \in \{1, \ldots, k\}$, $\mathbb{E}[f_i(X) f_j(X)] = \mathbb{E}[g_i(Y) g_j(Y)] = \delta_{ij}$. Then, the optimal score functions are given by the following extremal problem:

$$\max_{\substack{\{f_i : i = 1, \ldots, k\} \\ \{g_i : i = 1, \ldots, k\}}} \min_{1 \leq i \leq k} \mathbb{E}\left[f_i(X) g_i(Y)\right]$$

which is solved in Proposition 3. These optimal correlation functions correspond to the left and right singular vectors of B associated to its k largest singular values excluding $\sigma_0 = 1$. The scores associated with $g_1(Y), \ldots, g_k(Y)$ reflect the variation of the distribution of Y in a k-dimensional subspace. As k increases, more information is captured in these scores. We can show that these score functions can be computed from training data with the lowest possible sample complexity in the sense of Theorem 4, and are information theoretically optimal in the sense of (8), i.e. they achieve the optimal tradeoff between the computational complexity, k, and the information loss. We next present an efficient algorithm generalizing Algorithm 2 to find these score functions.

Observe that we can sequentially compute the leading k singular vectors of a $K \times K$ real matrix A by repeatedly running the power iteration method with initial vectors that are orthogonal to all previously computed singular vectors. However, a more desirable approach is to compute the k singular vectors in parallel. We begin by randomly choosing $\phi_0, \phi_1, \ldots, \phi_{k-1} \in \mathbb{R}^K$ which form a matrix $\Phi = [\phi_0 \;\; \phi_1 \cdots \phi_{k-1}]$, and update these vectors by computing: $\Phi \leftarrow A^T A \cdot \Phi$. We must include

a regularization step that ensures that the vectors in Φ are mutually orthogonal. A common technique is to use the Gram-Schmidt process which forces each ϕ_i to be orthogonal to $\phi_0, \ldots, \phi_{i-1}$. It is straightforward to verify that this procedure converges to the k leading singular vectors of A.

Applying this parallelized algorithm to the DTM, we initialize by choosing k score functions: $g_1, g_2, \ldots, g_k : \mathcal{Y} \to \mathbb{R}$, which we stack into a vector: $\underline{g} : \mathcal{Y} \to \mathbb{R}^k$. This can be interpreted as associating each value $y \in \mathcal{Y}$ with a k-dimensional signature $\underline{g}(y)$. Given a collection of n samples $\{(x_i, y_i) : i = 1, \ldots, n\}$, we compute the empirical conditional expectation analogous to step 2a in Algorithm 2:

$$\underline{f}(x) \leftarrow \widehat{\mathbb{E}}_n \left[\underline{g}(Y)|X = x\right], \ \ \forall x \in \mathcal{X}.$$

The resulting $\underline{f} : \mathcal{X} \to \mathbb{R}^k$ can be construed as assigning a k-dimensional signature to each value $x \in \mathcal{X}$. To compute this empirical conditional expectation, it is convenient to start with an arbitrarily chosen initial guess of $\underline{f}$. Upon observing a sample (x_i, y_i), we can simply update the signature $\underline{f}(x_i)$ by moving it towards $\underline{g}(y_i)$ with a step size of $\Delta / P_X(x_i)$, where $P_X(x_i)$ is the frequency that x_i is observed and $\Delta > 0$ is essentially a learning rate. Moreover, the aforementioned Gram-Schmidt regularization step is computationally expensive as the process operates on K-dimensional vectors. To instead operate on the k-dimensional signatures, we construct the $K \times k$ matrix $\Psi = \left[\sqrt{P_Y}\right] \cdot G$, where $G = [g_1 \cdots g_k]$ and each g_i is a column vector with entries $g_i(y)$, $y \in \mathcal{Y}$. We then compute the spectral decomposition of $\Psi^T \Psi = U \Lambda U^T$, and whiten Ψ by updating: $\Psi \leftarrow \Psi \cdot (U \Lambda^{-\frac{1}{2}})$, or equivalently updating: $G \leftarrow G \cdot (U \Lambda^{-\frac{1}{2}})$, which corresponds to updating:

$$\underline{g}(y) \leftarrow \left(\Lambda^{-\frac{1}{2}} U^T\right) \cdot \underline{g}(y), \ \ \forall y \in \mathcal{Y}.$$

This simple procedure ensures that Ψ has rank k and a condition number of 1. When the algorithm converges, the column vectors of Ψ form a basis for the k-dimensional subspace spanned by the leading k singular vectors of B excluding the one corresponding to $\sigma_0 = 1$. Although we do not recover these singular vectors, we can still perform dimension reduction to the desired k-dimensional subspace. We collect these ideas together as Algorithm 3.

In Algorithm 3, we assume as before that the marginal distributions P_X and P_Y are easy to estimate. In fact, although we include steps 2a and 3a which estimate the marginals, we assume that all operations involving them in the algorithm are precise. Furthermore, we only use two different sets of independent samples in steps 2 and 3 in order to simplify the analysis. In reality, we can use partially or fully overlapping sets of samples. Finally, we note that the number of samples used in each iteration, n, is chosen according to Theorem 4.

Algorithm 3 Parallel ACE Algorithm with Finite Samples

Require: training samples $\{(x_i, y_i) : i = 1, \ldots, N\}$

1. Initialize: randomly pick $\underline{g}(y)$, $y \in \mathcal{Y}$ and $\underline{f}(x)$, $x \in \mathcal{X}$

repeat:

2. Pick a subset of n samples, and for each sample (x_i, y_i):
   - 2a. Update counter to estimate $P_X(x_i)$
   - 2b. $\underline{f}(x_i) \leftarrow \underline{f}(x_i) + \frac{\Delta}{P_X(x_i)} \ \underline{g}(y_i)$
3. Pick another subset of n samples, and for each sample:
   - 3a. Update counter to estimate $P_Y(y_i)$
   - 3b. $\underline{g}(y_i) \leftarrow \underline{g}(y_i) + \frac{\Delta}{P_Y(y_i)} \ \underline{f}(x_i)$
4. Regularize:
   - 4a. $\underline{g}(y) \leftarrow \underline{g}(y) - \widehat{\mathbb{E}}_n \left[\underline{g}(Y)\right], \ \forall y \in \mathcal{Y}$
   - 4b. $\Phi \leftarrow \sum_{y \in \mathcal{Y}} P_Y(y) \, \underline{g}(y) \, \underline{g}(y)^T$
   - 4c. $[U \ \ \Lambda] = \mathsf{eig}(\Phi)$
   - 4d. $\underline{g}(y) \leftarrow \left(\Lambda^{-\frac{1}{2}} U^T\right) \cdot \underline{g}(y), \ \forall y \in \mathcal{Y}$

until $\widehat{\mathbb{E}}_n \left[\underline{f}(X)^T \underline{g}(Y)\right]$ stops to increase

V. Conclusion

In this paper, we developed a trinity of isomorphic vector spaces including distributions in a neighborhood, spherical perturbations, and score functions. We then illustrated the elegant geometric formulation of maximal correlation as a singular value of a linear map between such spaces. This formulation engendered an iterative procedure, known as the ACE algorithm, to compute maximal correlation functions (informative features) from data. We characterized the information theoretic optimality of these functions in a local data processing sense. Furthermore, we provided a tight characterization of the sample complexity of the ACE algorithm to argue that as the cardinalities $|\mathcal{X}|$ and $|\mathcal{Y}|$ become large, estimating maximal correlation functions requires far fewer samples than estimating arbitrary elements of the DTM. Finally, we generalized the ACE algorithm to compute several informative and mutually "orthogonal" score functions in parallel. This general algorithm serves as a dimension reduction tool which can be construed as performing PCA in the space of distributions rather than data. As a final remark, we note that our algorithms can also be used to fuse information from multiple sources of data by taking unions over the corresponding alphabet sets. Justifying this rigorously is a key future research endeavor.

References

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

[2] S.-L. Huang, A. Makur, F. Kozynski, and L. Zheng, "Efficient statistics: Extracting information from iid observations," in *Proceedings of the 52nd Annual Allerton Conference on Communication, Control, and Computing*, Allerton House, UIUC, Illinois, USA, October 1-3 2014.

[3] A. Makur and L. Zheng, "Bounds between contraction coefficients," in *Proceedings of the 53rd Annual Allerton Conference on Communication, Control and Computing*, Allerton House, UIUC, Illinois, USA, September 29-October 2 2015.

[4] A. Makur, "A study of local approximations in information theory," Master's thesis, Massachusetts Institute of Technology, June 2015.

[5] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover," April 2013, arXiv:1304.6133 [cs.IT].

[6] 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, September 1985.

[7] J. W. Demmel, *Applied Numerical Linear Algebra*, 1st ed. Philadelphia: Society for Industrial and Applied Mathematics (SIAM), 1997.

[8] W. K. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*, 4th ed. New York: Springer, 2015.