On Estimation of Modal Decompositions

Anuran Makur, Gregory W. Wornell, and Lizhong Zheng
EECS Department, Massachusetts Institute of Technology
Email: {a_makur, gww, lizhong}@mit.edu

***Abstract*—A modal decomposition is a useful tool that deconstructs the statistical dependence between two random variables by decomposing their joint distribution into orthogonal modes. Historically, modal decompositions have played important roles in statistics and information theory, e.g., in the study of maximal correlation. They are defined using the singular value decompositions of divergence transition matrices (DTMs) and conditional expectation operators corresponding to joint distributions. In this paper, we first characterize the set of all DTMs, and illustrate how the associated conditional expectation operators are the only weak contractions among a class of natural candidates. While modal decompositions have several modern machine learning applications, such as feature extraction from categorical data, the sample complexity of estimating them in such scenarios has not been analyzed. Hence, we also establish some non-asymptotic sample complexity results for the problem of estimating dominant modes of an unknown joint distribution from training data.**

I. Introduction

Modal decompositions of bivariate distributions were originally discovered in statistics by Hirschfeld [1], and have since been rediscovered in various related contexts. The purpose of such decompositions was to extend dimensionality reduction techniques such as *principal component analysis* [2], [3] and *canonical correlation analysis* (CCA) [4] to categorical data. In fact, CCA can be perceived as both a specialization and a generalization of such decompositions [5, Section 4.5.2]. Such decompositions were further analyzed to understand the notion of maximal correlation by Gebelein [6], Rényi [7], and later by Witsenhausen [8], as well as by Sarmanov [9] who elucidated the equivalence between squared maximal correlation and the *strong data processing inequality* (SDPI) for χ^2-divergence. Yet another independent development followed in the work of Lancaster [10], [11], whose line of analysis evolved into a study of when bivariate distributions could be decomposed using orthogonal polynomials; we refer readers to [12], which generalizes *Mehler's decomposition* [13] for jointly Gaussian distributions (cf. [14]), and the references therein for further details. Modal decompositions were then rediscovered again and further developed by the French school of data analysis [15], which exploited them as a data visualization tool known as *correspondence analysis* [16], [17]. Finally, in the context of non-parametric regression, Breiman and Friedman formulated the *alternating conditional expectations* (ACE) algorithm to (yet again) numerically compute modal decompositions [18]. (We also refer readers to the unified exposition of several of these ideas in [19].) While this rich history serves as a testament to the importance of modal decompositions in statistics and machine learning, there has not been any rigorous and non-asymptotic sample complexity analysis for estimating modal decompositions from training data for the purposes of modern data science applications.

In this paper, we make two main contributions. Firstly, after formally introducing modal decompositions of conditional expectation operators and *divergence transition matrices* (DTMs) in section I-A, we completely characterize DTMs and elucidate the special property of our choice of conditional expectation operators in section II. Secondly, we illustrate several non-asymptotic sample complexity results for estimating modal decompositions from training data in section III, which at least partially remedy the paucity of such results in the literature. Furthermore, in section I-B, we motivate our analysis of modal decompositions by elucidating their utility as approximate isometric embeddings of categorical data into Euclidean spaces.

A. Modal Decompositions of Bivariate Distributions

Consider the random variables X and Y taking values in the finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, with joint distribution $P_{X,Y}$. For simplicity, assume that the marginals satisfy $P_X(x) > 0$ and $P_Y(y) > 0$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Define the Hilbert space $\mathcal{L}^2(\mathcal{X}, P_X)$ of all real-valued functions on $\mathcal{X}$ with correlation as inner product:

$$\forall f_1, f_2 \in \mathcal{L}^2(\mathcal{X}, P_X),\ \langle f_1, f_2 \rangle_{P_X} \triangleq \sum_{x \in \mathcal{X}} P_X(x) f_1(x) f_2(x) \quad (1)$$

and induced $\mathcal{L}^2$-norm $\|\cdot\|_{P_X}$, and similarly, define $\mathcal{L}^2(\mathcal{Y}, P_Y)$.

We will analyze the singular value decomposition (SVD) structure of two equivalent representations of $P_{X,Y}$. The first representation is the *conditional expectation operator* $\mathbf{P}_{X|Y} : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ that maps any $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ to:

$$\forall y \in \mathcal{Y},\ \ \left(\mathbf{P}_{X|Y} f\right)(y) \triangleq \mathbb{E}[f(X)|Y = y]\,. \quad (2)$$

The second representation is the (aforementioned) DTM $\mathbf{B} \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|}$, whose (y,x)th entry is defined as [20]:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \ B(x,y) \triangleq \frac{P_{X,Y}(x,y)}{\sqrt{P_X(x) P_Y(y)}}\,. \quad (3)$$

(It is worth mentioning that the DTM $\mathbf{B}$ parallels the spectral graph theoretic concept of a symmetric normalized *Laplacian matrix*—see, e.g., [21, Section II-D], [22, Section 2.2].)

Let $K \triangleq \min\{|\mathcal{X}|, |\mathcal{Y}|\}$, and denote the orthonormal sets of right and left singular vectors of $\mathbf{P}_{X|Y}$ as $f_0^*, \ldots, f_{K-1}^* \in \mathcal{L}^2(\mathcal{X}, P_X)$ and $g_0^*, \ldots, g_{K-1}^* \in \mathcal{L}^2(\mathcal{Y}, P_Y)$, respectively, with corresponding singular values $\sigma_0 \geq \sigma_1 \geq \cdots \geq \sigma_{K-1} \geq 0$:

$$\forall i \in \{0, \ldots, K-1\},\ \ \mathbf{P}_{X|Y} f_i^* = \sigma_i\, g_i^*\,. \quad (4)$$

As shown in [7, Theorem 1] (cf. [23, Prop 2, Appendix A]), $\sigma_0 = 1$, $f_0^*(x) = 1$ for all $x \in \mathcal{X}$, and $g_0^*(y) = 1$ for all $y \in \mathcal{Y}$. Moreover, for any $i \in \{1, \ldots, K-1\}$, σ_i denotes the *maximal correlation*, cf. [1], [6], [7], [24, Section III]:

$$\sigma_i = \max_{\substack{f \in \mathcal{L}^2(\mathcal{X}, P_X),\, g \in \mathcal{L}^2(\mathcal{Y}, P_Y): \\ \forall j < i,\ \mathbb{E}[f(X) f_j^*(X)] = \mathbb{E}[g(Y) g_j^*(Y)] = 0 \\ \mathbb{E}\left[f(X)^2\right] = \mathbb{E}\left[g(Y)^2\right] = 1}} \mathbb{E}[f(X) g(Y)]\,, \quad (5)$$

where the optimal functions are f_i^* and g_i^*. Equivalently, we have the following SVD for $\mathbf{B}$ [24, Prop 2]:

$$\mathbf{B} = \sum_{i=0}^{K-1} \sigma_i \, \boldsymbol{\psi}_i^Y \left(\boldsymbol{\psi}_i^X\right)^{\mathrm{T}}, \quad (6)$$

where the sets of singular vectors $\boldsymbol{\psi}_0^X, \ldots, \boldsymbol{\psi}_{K-1}^X \in \mathbb{R}^{|\mathcal{X}|}$ and $\boldsymbol{\psi}_0^Y, \ldots, \boldsymbol{\psi}_{K-1}^Y \in \mathbb{R}^{|\mathcal{Y}|}$ (which are orthonormal with respect to the standard Euclidean inner product) satisfy the relations:

$$\forall i \in \{0, \ldots, K-1\}, \forall x \in \mathcal{X},\ \psi_i^X(x) = f_i^*(x) \sqrt{P_X(x)}\,, \quad (7)$$

$$\forall i \in \{0, \ldots, K-1\}, \forall y \in \mathcal{Y},\ \psi_i^Y(y) = g_i^*(y) \sqrt{P_Y(y)}\,, \quad (8)$$

where $\psi_i^X(x)$ and $\psi_i^Y(y)$ are the xth and yth entries of $\boldsymbol{\psi}_i^X$ and $\boldsymbol{\psi}_i^Y$, respectively, and we have $\psi_0^X(x) = \sqrt{P_X(x)}$ for all $x \in \mathcal{X}$ and $\psi_0^Y(y) = \sqrt{P_Y(y)}$ for all $y \in \mathcal{Y}$.

As shown in the complete manuscript [24, Prop 2], (4) and (6) can be recast as a *modal decomposition* of $P_{X,Y}$:

$$P_{X,Y}(x,y) = P_X(x)\, P_Y(y) \left(1 + \sum_{i=1}^{K-1} \sigma_i \, f_i^*(x) \, g_i^*(y)\right) \quad (9)$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, where $\mathbb{E}[f_i^*(X)] = \mathbb{E}[g_i^*(Y)] = 0$ and $\mathbb{E}\left[f_i^*(X) f_j^*(X)\right] = \mathbb{E}\left[g_i^*(Y) g_j^*(Y)\right] = \mathbb{1}_{i=j}$ for every $i, j \in \{1, \ldots, K-1\}$ (and $\mathbb{1}_{\mathcal{A}}$ denotes the indicator function that equals 1 if the proposition $\mathcal{A}$ is true and 0 otherwise). This elegantly decomposes the statistical dependence between X and Y into orthogonal modes, and elucidates the relative importance of these modes via the singular values. The decomposition (9) and maximal correlations (5) have been either the subject of or crucial in many recent studies on, e.g., hypercontractivity [25], SDPIs and functional inequalities [23], [26], [27], estimation theory, security, and privacy [28], feature extraction and dimensionality reduction [29], [30], and neural networks [31]. In the next subsection, we elaborate on one motivation for our analysis of modal decompositions; we refer readers to [24, Sections IV and V] for various other motivations and characterizations of modal decompositions.

B. Motivation: Embeddings of Categorical Data

Consider the problem of clustering the elements of $\mathcal{X}$ or $\mathcal{Y}$ in a manner that captures the salient dependencies between X and Y. For example, in the context of the "Netflix problem" [32], where $\mathcal{X}$ is the set of subscriber indices and $\mathcal{Y}$ is the set of movie indices, clustering the subscribers according to what movies they watch can help to build effective recommendation systems. However, since X and Y are categorical, in order to utilize simple clustering algorithms such as κ-means clustering [33]–[35], we must embed elements of $\mathcal{X}$ or $\mathcal{Y}$ into points in $\mathbb{R}^k$ with $k \in \{1, \ldots, K-1\}$. So, our objective is to extract real-valued *features* of X that carry as much information about Y as possible. We can interpret (5) as a natural formulation that achieves this goal. Indeed, the dominant pairs of singular vectors $(f_1^*(X), g_1^*(Y)), \ldots, (f_k^*(X), g_k^*(Y))$ yield features of X and Y that are maximally correlated with each other and carry orthogonal modes of information to avoid redundancy. These features produce the following *embeddings* of $\mathcal{X}$ and $\mathcal{Y}$ into the Euclidean space $\mathbb{R}^k$:

$$\mathcal{X} \ni x \ \mapsto\ \left[\sigma_1 f_1^*(x) \cdots \sigma_k f_k^*(x)\right]^{\mathrm{T}} \in \mathbb{R}^k\,, \quad (10)$$

$$\mathcal{Y} \ni y \ \mapsto\ \left[\sigma_1 g_1^*(y) \cdots \sigma_k g_k^*(y)\right]^{\mathrm{T}} \in \mathbb{R}^k\,, \quad (11)$$

which permit us to cluster elements of $\mathcal{X}$ or $\mathcal{Y}$ by clustering the corresponding embedded points in $\mathbb{R}^k$. These embeddings are particularly useful in high-dimensional settings where K is large, and we use $k \ll K$ to perform *dimensionality reduction*.

As expounded in [5, Section 4.5.3], our proposed embeddings (10) and (11) are very closely related to *diffusion maps* [36], which were introduced as a general conceptual framework for understanding "kernel eigenmap methods" such as *Laplacian eigenmaps* [37], and have been exploited in several machine learning tasks such as *manifold learning* and *spectral clustering* (see, e.g., [38, Section 2]). Indeed, much like diffusion maps, first consider the embedding $\mathcal{X} \ni x \mapsto P_{Y|X}(\cdot|x)$ of $\mathcal{X}$ into $\mathbb{R}^{|\mathcal{Y}|-1}$ using the conditional distributions. Observe using (9) that this association can be represented by the *isometric embedding* $\mathcal{X} \ni x \mapsto \left[\sigma_1 f_1^*(x) \cdots \sigma_{K-1} f_{K-1}^*(x)\right]^{\mathrm{T}} \in \mathbb{R}^{K-1}$. It is straightforward to verify that the ℓ^2-distance between the isometric embeddings of any two elements $x, x' \in \mathcal{X}$ precisely captures the (χ^2-divergence like) *diffusion distance* between the corresponding conditional distributions:

$$D_{\mathsf{diff}}^2(x, x') \triangleq \sum_{y \in \mathcal{Y}} \frac{\left(P_{Y|X}(y|x) - P_{Y|X}(y|x')\right)^2}{P_Y(y)} \quad (12)$$

$$= \sum_{i=1}^{K-1} \sigma_i^2 \left(f_i^*(x) - f_i^*(x')\right)^2. \quad (13)$$

Thus, clustering isometric embeddings of $\mathcal{X}$ using ℓ^2-distance corresponds to clustering conditional distribution embeddings of $\mathcal{X}$ using diffusion distance. To reduce the dimensionality of this isometric embedding when K is large and $k \ll K$, we can truncate it to produce the approximate isometric embedding in (10). When σ_{k+1} is very small, it is easy to see via (13) that the ℓ^2-distance between two embeddings of the form (10) approximately captures the diffusion distance between the corresponding conditional distributions.

In practical settings, we usually have access to n samples of training data $\{(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} : i \in \{1, \ldots, n\}\}$ that are drawn i.i.d. from the *unknown* true distribution $P_{X,Y}$. For instance, each sample (X_i, Y_i) might convey that subscriber X_i has streamed movie Y_i. In order to learn the embeddings (10) and (11), we need to estimate the k dominant pairs of singular vectors $(f_1^*, g_1^*), \ldots, (f_k^*, g_k^*)$ using the training data. The natural approach to do this is to efficiently compute the dominant singular vectors of the DTM corresponding to the

empirical distribution of the data using numerical linear algebra techniques such as the *orthogonal iteration* method, the *QR iteration* algorithm and its numerically enhanced variants, and *Krylov subspace* based methods (e.g., *Lanczos algorithm*) [39], [40]. Specifically, applying the orthogonal iteration method to DTMs is equivalent to applying the renowned ACE algorithm [18] to conditional expectation operators; see [24, Section VI], [5, Section 4.4] for more details and convergence properties.

II. Characterization of Operators

In this section, we present two main results. The first result in section II-A characterizes the set of all DTMs and develops basic properties of the map that takes bivariate distributions to their DTMs. The second result in section II-B portrays how our unique choice of Hilbert spaces ensures that our conditional expectation operators are weak contractions.

A. Characterization of DTMs

We begin by introducing some useful notation. Let $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ denote the set of all bivariate distributions over $\mathcal{X}\times\mathcal{Y}$ with entry-wise strictly positive marginals, and $\mathcal{P}_\circ^{\mathcal{X}\times\mathcal{Y}}$ denote the relative interior of $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ (i.e., all entry-wise strictly positive bivariate distributions). Moreover, define the *DTM function* $\mathbf{B}:\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}\to\mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ according to (3), so that $\mathbf{B}=\mathbf{B}(P_{X,Y})$ with some abuse of notation. Let $\mathcal{B}^{\mathcal{X}\times\mathcal{Y}}\triangleq\mathbf{B}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$ be the set of all DTMs, i.e., the range of the DTM function, and let $\mathcal{B}_\circ^{\mathcal{X}\times\mathcal{Y}}\triangleq\mathbf{B}(\mathcal{P}_\circ^{\mathcal{X}\times\mathcal{Y}})$ denote the set of all DTMs corresponding to distributions in $\mathcal{P}_\circ^{\mathcal{X}\times\mathcal{Y}}$. Then, the ensuing theorem characterizes $\mathcal{B}^{\mathcal{X}\times\mathcal{Y}}$ and $\mathcal{B}_\circ^{\mathcal{X}\times\mathcal{Y}}$, and establishes that $\mathbf{B}$ is an equivalent representation of $P_{X,Y}$.

Theorem 1 (Characterization of DTMs)**.** *The following hold:*

1) *A matrix $\mathbf{M}$ is a DTM corresponding to a distribution in $\mathcal{P}_\circ^{\mathcal{X}\times\mathcal{Y}}$ if and only if $\mathbf{M}>\mathbf{0}$ (entry-wise) and $\|\mathbf{M}\|_{\mathrm{s}}=1$:*
$$\mathcal{B}_\circ^{\mathcal{X}\times\mathcal{Y}}=\left\{\mathbf{M}\in\mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}\colon \mathbf{M}>\mathbf{0} \text{ and } \|\mathbf{M}\|_{\mathrm{s}}=1\right\},$$
where $\|\cdot\|_{\mathrm{s}}$ denotes the spectral norm, and $\mathbf{0}$ is a matrix with all entries equal to 0 of appropriate dimension.
2) *A matrix $\mathbf{M}$ is a DTM corresponding to a distribution in $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ if and only if $\mathbf{M}\geq\mathbf{0}$ (entry-wise), $\|\mathbf{M}\|_{\mathrm{s}}=1$, and both $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ and $\mathbf{M}\mathbf{M}^{\mathrm{T}}$ have entry-wise strictly positive eigenvectors corresponding to their unit eigenvalue:*
$$\begin{aligned}\mathcal{B}^{\mathcal{X}\times\mathcal{Y}}=\big\{&\mathbf{M}\in\mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}\colon \mathbf{M}\geq\mathbf{0},\ \|\mathbf{M}\|_{\mathrm{s}}=1,\\ &\exists\,\boldsymbol{\psi}^X>\mathbf{0},\ \mathbf{M}^{\mathrm{T}}\mathbf{M}\,\boldsymbol{\psi}^X=\boldsymbol{\psi}^X,\ \text{and}\\ &\exists\,\boldsymbol{\psi}^Y>\mathbf{0},\ \mathbf{M}\mathbf{M}^{\mathrm{T}}\boldsymbol{\psi}^Y=\boldsymbol{\psi}^Y\big\}.\end{aligned}$$
3) *$\mathbf{B}\colon\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}\to\mathcal{B}^{\mathcal{X}\times\mathcal{Y}}$ is bijective and continuous.*

Theorem 1 is established using the *Perron-Frobenius theorem* (cf. [41, Theorems 8.2.2 and 8.3.1]) in [24, Appendix II-B]. In view of part 2 of Theorem 1, it is worth noting that a non-negative square matrix $\mathbf{A}\geq\mathbf{0}$, such as the Gramian $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ or dual Gramian $\mathbf{M}\mathbf{M}^{\mathrm{T}}$, has entry-wise strictly positive left and right eigenvectors corresponding to its spectral radius $\rho(\mathbf{A})$ if and only if the triangular block form of $\mathbf{A}$ is a direct sum of irreducible non-negative square matrices whose spectral radii are also $\rho(\mathbf{A})$, cf. [42, Chapter 2, Section 3].

B. Representation of Conditional Expectation Operators

It is reasonable to ask why we focus on SVDs of DTMs, i.e., conditional expectation operators with the specific choices of Hilbert spaces in section I-A, as opposed to other commonly used representations of $P_{X,Y}$, e.g., information density [43], [44]. To partially address this question, we demonstrate that our choices of Hilbert spaces uniquely produce conditional expectation operators that are weak contractions over a reasonable class of candidates.

Given $P_{X,Y}\in\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$, the map $\mathbf{P}_{X|Y}$ is completely characterized by the conditional distribution $P_{X|Y}$ via (2). However, to make $\mathbf{P}_{X|Y}$ a well-defined linear operator with an SVD, we must endow its input and output vector spaces of functions with inner products. Fix the output Hilbert space of $\mathbf{P}_{X|Y}$ to be $\mathcal{L}^2(\mathcal{Y},P_Y)$. While this produces a canonical choice of input Hilbert space $\mathcal{L}^2(\mathcal{X},P_X)$, let us instead select the Hilbert space $\mathcal{L}^2(\mathcal{X},Q_X)$ for any (entry-wise) strictly positive distribution $Q_X>\mathbf{0}$ over $\mathcal{X}$. We then define the corresponding induced *operator norm* of $\mathbf{P}_{X|Y}:\mathcal{L}^2(\mathcal{X},Q_X)\to\mathcal{L}^2(\mathcal{Y},P_Y)$ as:
$$\left\|\mathbf{P}_{X|Y}\right\|_{Q_X\to P_Y}\triangleq\max_{f\in\mathcal{L}^2(\mathcal{X},Q_X)\setminus\{\mathbf{0}\}}\frac{\left\|\mathbf{P}_{X|Y}f\right\|_{P_Y}}{\left\|f\right\|_{Q_X}},\tag{14}$$
where we also use $\mathbf{0}$ to represent the everywhere zero function. The next theorem conveys that the only choice of input Hilbert space that makes $\mathbf{P}_{X|Y}$ a *weak contraction* is $\mathcal{L}^2(\mathcal{X},P_X)$.

Theorem 2 (Weak Contraction)**.** *The minimum operator norm of $\mathbf{P}_{X|Y}$ over all choices of $Q_X>\mathbf{0}$ is:*
$$\min_{Q_X>\mathbf{0}}\left\|\mathbf{P}_{X|Y}\right\|_{Q_X\to P_Y}=\left\|\mathbf{P}_{X|Y}\right\|_{P_X\to P_Y}=1,$$
where $Q_X^*=P_X$ is the unique minimizer. Furthermore, for any $Q_X>\mathbf{0}$, we have the following χ^2-divergence bound:
$$\left\|\mathbf{P}_{X|Y}\right\|_{Q_X\to P_Y}^2\geq 1+\chi^2(P_X\|Q_X)\triangleq\sum_{x\in\mathcal{X}}\frac{P_X(x)^2}{Q_X(x)}.$$

Proof. Note that for all $Q_X>\mathbf{0}$, we have $f_0^*\in\mathcal{L}^2(\mathcal{X},Q_X)$ with $\|f_0^*\|_{Q_X}=1$ and $\|\mathbf{P}_{X|Y}f_0^*\|_{P_Y}=\|g_0^*\|_{P_Y}=1$, where f_0^* and g_0^* are everywhere unity functions. As a result, we get $\|\mathbf{P}_{X|Y}\|_{Q_X\to P_Y}\geq 1$. Moreover, since $\sigma_0=1$, we know that $Q_X=P_X$ achieves this lower bound; indeed, for every $f\in\mathcal{L}^2(\mathcal{X},P_X)$, the conditional Jensen's inequality yields:
$$\mathbb{E}\Big[\mathbb{E}[f(X)|Y]^2\Big]\leq\mathbb{E}\big[\mathbb{E}\big[f(X)^2\big|Y\big]\big]=\mathbb{E}\big[f(X)^2\big].$$
This proves that the minimum operator norm of $\mathbf{P}_{X|Y}$ is 1.

To prove that $Q_X^*=P_X$ is the unique minimizer, it suffices to establish the χ^2-divergence bound, because χ^2-divergence is zero if and only if its input distributions are equal. For any $Q_X>\mathbf{0}$, a direct calculation shows that the *adjoint* operator $\mathbf{P}_{X|Y}^*:\mathcal{L}^2(\mathcal{Y},P_Y)\to\mathcal{L}^2(\mathcal{X},Q_X)$ of $\mathbf{P}_{X|Y}:\mathcal{L}^2(\mathcal{X},Q_X)\to\mathcal{L}^2(\mathcal{Y},P_Y)$ is given by (see [24, Appendix II-C] for details):
$$\forall x\in\mathcal{X},\quad\big(\mathbf{P}_{X|Y}^*g\big)(x)=\frac{P_X(x)}{Q_X(x)}\,\mathbb{E}[g(Y)|X=x]$$
for every $g\in\mathcal{L}^2(\mathcal{Y},P_Y)$, where the conditional expectation is with respect to $P_{Y|X}$ associated with $P_{X,Y}$. Now observe

that $(\mathbf{P}^*_{X|Y} g_0^*)(x) = P_X(x)/Q_X(x)$ for all $x \in \mathcal{X}$. Hence, $\|g_0^*\|_{P_Y} = 1$ and we have:

$$\left\|\mathbf{P}^*_{X|Y} g_0^*\right\|^2_{Q_X} = \sum_{x\in\mathcal{X}} Q_X(x) \frac{P_X(x)^2}{Q_X(x)^2} = 1 + \chi^2(P_X||Q_X)\,.$$

Since $\|\mathbf{P}_{X|Y}\|^2_{Q_X\to P_Y} = \|\mathbf{P}^*_{X|Y}\|^2_{P_Y\to Q_X} \geq \|\mathbf{P}^*_{X|Y} g_0^*\|^2_{Q_X}$, we obtain the desired bound. This completes the proof. ∎

Theorem 2 illustrates that given $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$, the only inner products that make $\mathbf{P}_{X|Y} = \mathbb{E}[\cdot|Y]$ and $\mathbb{E}[\cdot|X]$ adjoints and weak contractions are those with respect to P_X and P_Y.

III. Sample Complexity Analysis

In this section, we present two main sample complexity results for estimation of the dominant $k \in \{1,\ldots,K-1\}$ modes in (9). As mentioned at the end of section I-B, we will assume that the true joint distribution $P_{X,Y}$ is unknown, but we have n samples of labeled training data $\{(X_i,Y_i) \in \mathcal{X}\times\mathcal{Y} : i \in \{1,\ldots,n\}\}$ drawn i.i.d. from $P_{X,Y}$. Therefore, we will use the empirical distribution $\hat{P}^n_{X,Y}$ on $\mathcal{X}\times\mathcal{Y}$:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\;\; \hat{P}^n_{X,Y}(x,y) \triangleq \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i=x}\,\mathbb{1}_{Y_i=y} \tag{15}$$

as a proxy for $P_{X,Y}$. Note that when K is large, the training data is often enough to accurately estimate P_X and P_Y, but not $P_{X,Y}$. Furthermore, in many scenarios, we may have additional (inexpensive) unlabeled training data available, which also improves the estimation accuracy of P_X and P_Y. In light of these and other analytical tractability considerations, we will assume that the true marginals P_X and P_Y are known, and for some $p_0 > 0$, they satisfy the (universal) lower bound:

$$\min_{x\in\mathcal{X}} P_X(x) \geq p_0 \quad \text{and} \quad \min_{y\in\mathcal{Y}} P_Y(y) \geq p_0\,. \tag{16}$$

To estimate the dominant modes in (9), define the singular vector estimates $\hat{f}_i^* \in \mathcal{L}^2(\mathcal{X}, P_X)$ and $\hat{g}_i^* \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ for $i \in \{1,\ldots,K\}$, and the singular value estimates $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_K \geq 0$, via the "empirical" modal decomposition:

$$\hat{P}^n_{X,Y}(x,y) = P_X(x)\,P_Y(y)\left(1 + \sum_{i=1}^K \hat{\sigma}_i\,\hat{f}_i^*(x)\,\hat{g}_i^*(y)\right), \tag{17}$$

where $\mathbb{E}\big[\hat{f}_i^*(X)\hat{f}_j^*(X)\big] = \mathbb{E}\big[\hat{g}_i^*(Y)\hat{g}_j^*(Y)\big] = \mathbb{1}_{i=j}$ for $i,j \in \{1,\ldots,K\}$. The decomposition (17) corresponds to the SVD of a quasi-DTM $\hat{\mathbf{B}} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ with (y,x)th entry given by:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\; \hat{B}(x,y) \triangleq \frac{\hat{P}^n_{X,Y}(x,y) - P_X(x)P_Y(y)}{\sqrt{P_X(x)P_Y(y)}}. \tag{18}$$

Indeed, the matrix $\hat{\mathbf{B}}$ has the SVD:

$$\hat{\mathbf{B}} = \sum_{i=1}^K \hat{\sigma}_i\,\hat{\boldsymbol{\psi}}_i^Y \big(\hat{\boldsymbol{\psi}}_i^X\big)^{\mathrm{T}}, \tag{19}$$

where the sets of orthonormal singular vectors $\hat{\boldsymbol{\psi}}_1^X,\ldots,\hat{\boldsymbol{\psi}}_K^X \in \mathbb{R}^{|\mathcal{X}|}$ and $\hat{\boldsymbol{\psi}}_1^Y,\ldots,\hat{\boldsymbol{\psi}}_K^Y \in \mathbb{R}^{|\mathcal{Y}|}$ have elements:

$$\forall i \in \{1,\ldots,K\}, \forall x \in \mathcal{X},\;\; \hat{\psi}_i^X(x) = \hat{f}_i^*(x)\sqrt{P_X(x)}\,, \tag{20}$$

$$\forall i \in \{1,\ldots,K\}, \forall y \in \mathcal{Y},\;\; \hat{\psi}_i^Y(y) = \hat{g}_i^*(y)\sqrt{P_Y(y)}\,. \tag{21}$$

Finally, for convenience, we define (and will analyze) the following zero-mean singular vector estimates for every $i \in \{1,\ldots,K\}$: $\check{f}_i^*(x) \triangleq \hat{f}_i^*(x) - \mathbb{E}\big[\hat{f}_i^*(X)\big]$ for all $x \in \mathcal{X}$, and $\check{g}_i^*(y) \triangleq \hat{g}_i^*(y) - \mathbb{E}\big[\hat{g}_i^*(Y)\big]$ for all $y \in \mathcal{Y}$.

A. Estimation of Dominant Maximal Correlations

We first determine the number of samples required to obtain accurate estimates $\hat{\sigma}_1,\ldots,\hat{\sigma}_k$ of $\sigma_1,\ldots,\sigma_k$ in terms of the (squared) ℓ^1-loss. The next theorem conveys an exponential concentration of measure inequality for the ℓ^1-loss.

Theorem 3 (Estimation Tail Bound I). *For any* $0 \leq \delta \leq \frac{\sqrt{k}}{p_0\sqrt{2}}$:

$$\mathbb{P}\left(\sum_{i=1}^k \big|\hat{\sigma}_i - \sigma_i\big| \geq \delta\right) \leq \exp\left(\frac{1}{4} - \frac{n\,p_0^2\,\delta^2}{8k}\right),$$

where $\exp(\cdot)$ *denotes the natural exponential.*

Proof Sketch. We outline the proof here, and refer readers to [24, Appendix VI-A] for details. First, for each $i \in \{1,\ldots,n\}$, let $\mathbf{Z}_i \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ be a random matrix with (y,x)th element:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\;\; Z_i(x,y) \triangleq \frac{\mathbb{1}_{X_i=x}\mathbb{1}_{Y_i=y} - P_X(x)P_Y(y)}{\sqrt{P_X(x)P_Y(y)}}\,.$$

Accordingly, $\mathbf{Z}_1,\ldots,\mathbf{Z}_n$ are i.i.d., and $\hat{\mathbf{B}} = \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i$ due to (15). For every $i \in \{1,\ldots,n\}$, define the corresponding zero-mean random matrices $\tilde{\mathbf{Z}}_i \triangleq \mathbf{Z}_i - \mathbb{E}\big[\mathbf{Z}_i\big]$. Each $\tilde{\mathbf{Z}}_i$ is almost surely bounded in Frobenius norm $\|\cdot\|_{\mathrm{F}}$:

$$\left\|\tilde{\mathbf{Z}}_i\right\|^2_{\mathrm{F}} = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \frac{\big(\mathbb{1}_{X_i=x}\mathbb{1}_{Y_i=y} - P_{X,Y}(x,y)\big)^2}{P_X(x)P_Y(y)} \tag{22}$$

$$\leq \frac{1}{p_0^2}\left(\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathbb{1}_{X_i=x}\mathbb{1}_{Y_i=y} + P_{X,Y}(x,y)^2\right) \tag{23}$$

$$\leq \frac{2}{p_0^2}\,, \tag{24}$$

where (23) follows from (16) and the fact that for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $-2\,\mathbb{1}_{X_i=x}\mathbb{1}_{Y_i=y}P_{X,Y}(x,y) \leq 0$, and (24) holds because the first summation in (23) is equal to unity while the second summation is upper bounded by unity. Furthermore, the "total variance" of each $\tilde{\mathbf{Z}}_i$ is also bounded:

$$\mathbb{E}\left[\left\|\tilde{\mathbf{Z}}_i\right\|^2_{\mathrm{F}}\right] \leq \frac{1}{p_0^2}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathrm{var}[\mathbb{1}_{X_i=x}\mathbb{1}_{Y_i=y}] \leq \frac{1}{p_0^2}\,, \tag{25}$$

where $\mathrm{var}[\cdot]$ denotes the variance operator, the first inequality follows from (22) and (16), and the second inequality holds because $\mathrm{var}[\mathbb{1}_{X_i=x}\mathbb{1}_{Y_i=y}] \leq P_{X,Y}(x,y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The boundedness conditions (24) and (25) ensure that we can apply a vector version of Bernstein's inequality to the i.i.d. sequence $\tilde{\mathbf{Z}}_1,\ldots,\tilde{\mathbf{Z}}_n$ (as we will do shortly).

Next, note that $\hat{\sigma}_1,\ldots,\hat{\sigma}_k$ are the k largest singular values of $\hat{\mathbf{B}}$. Moreover, due to (3), (6), and the structure of $\boldsymbol{\psi}_0^X$ and $\boldsymbol{\psi}_0^Y$ delineated in section I-A, we get:

$$\mathbb{E}\big[\mathbf{Z}_1\big] = \mathbf{B} - \boldsymbol{\psi}_0^Y\big(\boldsymbol{\psi}_0^X\big)^{\mathrm{T}} = \sum_{j=1}^{K-1} \sigma_j\,\boldsymbol{\psi}_j^Y\big(\boldsymbol{\psi}_j^X\big)^{\mathrm{T}},$$

which implies that $\sigma_1, \ldots, \sigma_k$ are the k largest singular values of $\mathbb{E}[\mathbf{Z}_1]$. To complete the proof, we employ the following singular value perturbation bound.

Lemma 4 (Stability of Singular Values)**.** *Given two matrices $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{k_1 \times k_2}$, for every $k \in \{1, \ldots, \min\{k_1, k_2\}\}$:*

$$\sum_{i=1}^{k} \left|\sigma_i(\mathbf{A}_1) - \sigma_i(\mathbf{A}_2)\right| \leq \sqrt{k}\, \left\|\mathbf{A}_1 - \mathbf{A}_2\right\|_{\mathrm{F}},$$

where $\sigma_i(\cdot)$ denotes the ith largest singular value of its matrix argument with $i \in \{1, \ldots, \min\{k_1, k_2\}\}$.

Lemma 4 is derived in [24, Lemma 104] via a weak majorization result for singular values known as the *Lidskii inequality*, cf. [45, Theorem 3.4.5]. Applying it to our problem yields:

$$\sum_{i=1}^{k} \left|\hat{\sigma}_i - \sigma_i\right| \leq \sqrt{k}\, \left\|\hat{\mathbf{B}} - \mathbb{E}[\mathbf{Z}_1]\right\|_{\mathrm{F}} = \sqrt{k}\, \left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathbf{Z}}_i \right\|_{\mathrm{F}}. \quad (26)$$

Finally, for any $0 \leq \delta \leq \frac{\sqrt{k}}{p_0 \sqrt{2}}$, we have:

$$\mathbb{P}\left(\sum_{i=1}^{k} \left|\hat{\sigma}_i - \sigma_i\right| \geq \delta \right) \leq \mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathbf{Z}}_i \right\|_{\mathrm{F}} \geq \frac{\delta}{\sqrt{k}} \right) \quad (27)$$

$$\leq \exp\left(\frac{1}{4} - \frac{n\, p_0^2\, \delta^2}{8k} \right), \quad (28)$$

where (27) follows from (26), and to obtain (28) we have used the bounds (24) and (25) along with the vector generalization of *Bernstein's inequality* in [46, Theorem 2.4] (also see [24, Lemma 103]). This completes the proof. ■

Theorem 3 shows that estimating $\sigma_1, \ldots, \sigma_k$ via $\hat{\sigma}_1, \ldots, \hat{\sigma}_k$ to within a fixed ℓ^1-norm error and a fixed confidence level requires the number of samples n to grow linearly with k. A key consequence of Theorem 3 is the following corollary, which presents a corresponding squared ℓ^1-risk bound.

Corollary 5 (Squared ℓ^1-Risk Bound I)**.** *For every sufficiently large n such that $n \geq 16 \log(4kn)$:*

$$\mathbb{E}\left[\left(\sum_{i=1}^{k} \left|\hat{\sigma}_i - \sigma_i\right| \right)^2 \right] \leq \frac{6k + 8k \log(nk)}{p_0^2 n},$$

where $\log(\cdot)$ denotes the natural logarithm.

Corollary 5 is proved in [24, Appendix VI-B]. We remark that several other consequences of Theorem 3 and Corollary 5, such as estimation bounds for Ky Fan k-norms, can be found in [24, Section VI-B].

B. Estimation of Dominant Features

In this final subsection, we determine the number of samples required to obtain accurate estimates $\check{f}_*^k = (\check{f}_1^*, \ldots, \check{f}_k^*)$ of the singular vectors $f_*^k = (f_1^*, \ldots, f_k^*)$. (By symmetry, analogous results can be obtained for the estimation of $g_*^k = (g_1^*, \ldots, g_k^*)$ using $\check{g}_*^k = (\check{g}_1^*, \ldots, \check{g}_k^*)$.) Despite the existence of invariant subspace stability results, e.g., the *Davis-Kahan theorems* and *Wedin's theorems* (see [47, Chapter V, Sections 3 and 4]), the individual singular vectors of an operator often vary greatly under perturbations. So, instead of directly analyzing the convergence of $\check{f}_*^k$ to f_*^k, our development focuses on measuring the accuracy of these estimates with the loss function:

$$\sum_{i=1}^{k} \left\|\mathbf{P}_{X|Y} f_i^*\right\|_{P_Y}^2 - \sum_{i=1}^{k} \left\|\mathbf{P}_{X|Y} \check{f}_i^*\right\|_{P_Y}^2 \geq 0, \quad (29)$$

where the first term is equal to $\sigma_1^2 + \cdots + \sigma_k^2$, the second term can be construed as an estimator of the first term, and the non-negativity above is argued in [24, Section VI-B, Lemma 3, Equation (546)]. To facilitate further interpretation of this loss function, we refer readers to the detailed exposition in [24] of how this loss function captures the extent to which the estimates $\check{f}_*^k$ preserve as much of a "k-rank approximation" of the mutual information between X and Y (see [24, Equation (73)]) as possible under local approximations.

As before, the next theorem portrays an exponential concentration of measure inequality for the loss function in (29).

Theorem 6 (Estimation Tail Bound II)**.** *For any $0 \leq \delta \leq 4k$:*

$$\mathbb{P}\left(\sum_{i=1}^{k} \left\|\mathbf{P}_{X|Y} f_i^*\right\|_{P_Y}^2 - \left\|\mathbf{P}_{X|Y} \check{f}_i^*\right\|_{P_Y}^2 \geq \delta \right) \leq \left(|\mathcal{X}| + |\mathcal{Y}|\right) \exp\left(-\frac{n\, p_0\, \delta^2}{64\, k^2} \right),$$

where we clarify that the probability measure $\mathbb{P}$ is given by the law of $\check{f}_*^k$.

Theorem 6 is derived in [24, Appendix VI-D] using a matrix generalization of Bernstein's inequality, cf. [48, Theorem 1.6], and the *Weyl inequality* from matrix perturbation theory, cf. [41, Corollary 7.3.5(a)]. It illustrates that estimating f_*^k via $\check{f}_*^k$ to within a fixed error and confidence level requires n to grow quadratically with k. As before, a key consequence of Theorem 6 is the following corollary, which presents a bound on the *mean squared error* (MSE) between $\sum_{i=1}^{k} \left\|\mathbf{P}_{X|Y} \check{f}_i^*\right\|_{P_Y}^2$ and $\sum_{i=1}^{k} \left\|\mathbf{P}_{X|Y} f_i^*\right\|_{P_Y}^2$.

Corollary 7 (MSE Risk Bound II)**.** *For every sufficiently large n such that $\frac{p_0 n}{64} \geq \frac{1}{|\mathcal{X}|+|\mathcal{Y}|}$ and $\frac{p_0 n}{4} \geq \log\left(\frac{p_0 n}{64}\left(|\mathcal{X}| + |\mathcal{Y}|\right)\right)$:*

$$\mathbb{E}\left[\left(\sum_{i=1}^{k} \left\|\mathbf{P}_{X|Y} f_i^*\right\|_{P_Y}^2 - \left\|\mathbf{P}_{X|Y} \check{f}_i^*\right\|_{P_Y}^2 \right)^2 \right] \leq \frac{64k^2 \left(\log\left(p_0 n \left(|\mathcal{X}| + |\mathcal{Y}|\right)\right) - 3 \right)}{p_0 n}.$$

Corollary 7 is established in [24, Appendix VI-E].

The non-asymptotic sample complexity results in this section provide an initial theoretical foundation for the important problem of estimating modal decompositions. We believe that these results can be sharpened and generalized (by removing some of our assumptions) using tools from the rich matrix estimation literature (see [49] and the references therein). In closing, we again refer readers to the manuscript [24, Section VI] for several complementary sample complexity bounds as well as some pertinent large deviations theoretic analysis.

REFERENCES

[1] H. O. Hirschfeld, "A connection between correlation and contingency," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 31, no. 4, pp. 520–524, October 1935.

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

[3] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441 and 498–520, September 1933.

[4] ——, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, December 1936.

[5] A. Makur, "Information contraction and decomposition," Sc.D. Thesis in Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA, May 2019.

[6] H. Gebelein, "Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung," *Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 21, no. 6, pp. 364–379, December 1941, in German.

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

[8] H. S. Witsenhausen, "On sequences of pairs of dependent random variables," *SIAM Journal on Applied Mathematics*, vol. 28, no. 1, pp. 100–113, January 1975.

[9] O. V. Sarmanov, "Maximal correlation coefficient (non-symmetric case)," *Doklady Akademii Nauk SSSR*, vol. 121, no. 1, pp. 52–55, 1958, in Russian.

[10] H. O. Lancaster, "The structure of bivariate distributions," *The Annals of Mathematical Statistics*, vol. 29, no. 3, pp. 719–736, September 1958.

[11] ——, *The Chi-Squared Distribution*. New York, NY, USA: John Wiley & Sons Inc., 1969.

[12] A. Makur and L. Zheng, "Polynomial singular value decompositions of a family of source-channel models," *IEEE Transactions on Information Theory*, vol. 63, no. 12, pp. 7716–7728, December 2017.

[13] F. G. Mehler, "Ueber die entwicklung einer function von beliebig vielen variablen nach laplaceschen functionen höherer ordnung," *Journal für die reine und angewandte Mathematik*, vol. 66, pp. 161–176, 1866, in German.

[14] E. Abbe and L. Zheng, "A coordinate system for Gaussian networks," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 721–733, February 2012.

[15] J.-P. Benzécri, *L'Analyse des Données, Tôme 2: L'Analyse des Correspondances*. Paris, France: Dunod, 1973, in French.

[16] M. J. Greenacre, *Theory and Applications of Correspondence Analysis*. San Diego, CA, USA: Academic Press, March 1984.

[17] M. Greenacre and T. Hastie, "The geometric interpretation of correspondence analysis," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 437–447, June 1987.

[18] 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.

[19] A. Buja, "Theory of bivariate ACE," Department of Statistics, University of Washington, Seattle, WA, USA, Tech. Rep. 74, December 1985.

[20] S. Borade and L. Zheng, "Euclidean information theory," in *Proceedings of the IEEE International Zurich Seminar on Communications*, Zurich, Switzerland, March 12-14 2008, pp. 14–17.

[21] D. Qiu, A. Makur, and L. Zheng, "Probabilistic clustering using maximal matrix norm couplings," in *Proceedings of the 56th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, October 2-5 2018, pp. 1020–1027.

[22] G. Schiebinger, M. J. Wainwright, and B. Yu, "The geometry of kernelized spectral clustering," *The Annals of Statistics*, vol. 43, no. 2, pp. 819–846, April 2015.

[23] A. Makur and L. Zheng, "Linear bounds between contraction coefficients for f-divergences," July 2018, arXiv:1510.01844v4 [cs.IT].

[24] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng, "On universal features for high-dimensional learning and inference," November 2019, arXiv:1911.09105 [cs.LG].

[25] 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].

[26] M. Raginsky, "Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3355–3389, June 2016.

[27] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and Bayesian networks," in *Convexity and Concentration*, ser. The IMA Volumes in Mathematics and its Applications, E. Carlen, M. Madiman, and E. M. Werner, Eds., vol. 161. New York, NY, USA: Springer, 2017, pp. 211–249.

[28] F. du Pin Calmon, A. Makhdoumi, M. Médard, M. Varia, M. Christiansen, and K. R. Duffy, "Principal inertia components and applications," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 5011–5038, August 2017.

[29] A. Makur, F. Kozynski, S.-L. Huang, and L. Zheng, "An efficient algorithm for information decomposition and extraction," in *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, September 29-October 2 2015, pp. 972–979.

[30] S.-L. Huang, A. Makur, L. Zheng, and G. W. Wornell, "An information-theoretic approach to universal feature selection in high-dimensional inference," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Aachen, Germany, June 25-30 2017, pp. 1336–1340.

[31] H. Hsu, S. Salamatian, and F. P. Calmon, "Correspondence analysis using neural networks," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, Naha, Japan, April 16-18 2019, pp. 2671–2680.

[32] J. Bennett and S. Lanning, "The Netflix prize," in *Proceedings of KDD Cup and Workshop*, San Jose, CA, USA, August 12 2007, pp. 3–6.

[33] H. Steinhaus, "Sur la division des corps matériels en parties," *Bulletin de l'Académie Polonaise des Sciences, Classe III*, vol. 4, no. 12, pp. 801–804, 1956, in French.

[34] J. Max, "Quantizing for minimum distortion," *IRE Transactions on Information Theory*, vol. 6, no. 1, pp. 7–12, March 1960.

[35] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. IT-28, no. 2, pp. 129–137, March 1982.

[36] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis, Elsevier*, vol. 21, no. 1, pp. 5–30, July 2006.

[37] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proceedings of the Advances in Neural Information Processing Systems 14 (NIPS)*, Vancouver, BC, Canada, December 3-8 2001, pp. 585–591.

[38] A. S. Bandeira, "Ten lectures and forty-two open problems in the mathematics of data science," October 2016, Department of Mathematics, MIT, Cambridge, MA, USA, Lecture Notes 18.S096 (Fall 2015).

[39] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed., ser. Johns Hopkins Studies in the Mathematical Sciences. Baltimore, MD, USA: The Johns Hopkins University Press, 1996.

[40] J. W. Demmel, *Applied Numerical Linear Algebra*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics (SIAM), 1997.

[41] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. New York, NY, USA: Cambridge University Press, 2013.

[42] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, ser. Classics in Applied Mathematics. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics (SIAM), 1994, vol. 9.

[43] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22–29, March 1990.

[44] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 752–772, May 1993.

[45] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. New York, NY, USA: Cambridge University Press, 1991.

[46] E. J. Candès and Y. Plan, "A probabilistic and RIPless theory of compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 11, pp. 7235–7254, November 2011.

[47] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, ser. Computer Science and Scientific Computing. New York, NY, USA: Academic Press, 1990.

[48] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, August 2012.

[49] S. Chatterjee, "Matrix estimation by universal singular value thresholding," *The Annals of Statistics*, vol. 43, no. 1, pp. 177–214, February 2015.