Foundations and Trends® in Communications and Information Theory

Universal Features for High-Dimensional Learning and Inference

Information Theoretic and Geometric Perspectives

Suggested Citation: Shao-Lun Huang, Anuran Makur, Gregory W. Wornell and Lizhong Zheng (2024), “Universal Features for High-Dimensional Learning and Inference”, Foundations and Trends® in Communications and Information Theory: Vol. 21, No. 1-2, pp 1–299. DOI: 10.1561/0100000107.

Shao-Lun Huang
Tsinghua-Berkeley Shenzhen Institute
shaolun.huang@sz.tsinghua.edu.cn

Anuran Makur
Purdue University
amakur@purdue.edu

Gregory W. Wornell
Massachusetts Institute of Technology
gww@mit.edu

Lizhong Zheng
Massachusetts Institute of Technology
lizhong@mit.edu

now
the essence of knowledge
Boston — Delft

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Contents

The version of record is available at: http://dx.doi.org/10.1561/0100000107

The version of record is available at: http://dx.doi.org/10.1561/0100000107

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Universal Features for High-Dimensional Learning and Inference

Shao-Lun Huang[1], Anuran Makur[2], Gregory W. Wornell[3] and Lizhong Zheng[4]

[1] *Tsinghua-Berkeley Shenzhen Institute, China; shaolun.huang@sz.tsinghua.edu.cn*
[2] *Purdue University, USA; amakur@purdue.edu*
[3] *Massachusetts Institute of Technology, USA; gww@mit.edu*
[4] *Massachusetts Institute of Technology, USA; lizhong@mit.edu*

ABSTRACT

This monograph develops unifying perspectives on the problem of identifying universal low-dimensional features from high-dimensional data for inference tasks in settings involving learning. For such problems, natural notions of universality are introduced, and a local equivalence among them is established. The analysis is naturally expressed via information geometry, which provides both conceptual and computational insights. The development reveals the complementary roles of the singular value decomposition, Hirschfeld-Gebelein-Rényi maximal correlation, the canonical correlation and principle component analyses of Hotelling and Pearson, Tishby's information bottleneck, Wyner's and Gács-Körner common information, Ky Fan k-norms, and Breiman and Friedman's alternating conditional expectations algorithm. Among other uses, the framework facilitates

Shao-Lun Huang, Anuran Makur, Gregory W. Wornell and Lizhong Zheng (2024), "Universal Features for High-Dimensional Learning and Inference", Foundations and Trends® in Communications and Information Theory: Vol. 21, No. 1-2, pp 1–299.
DOI: 10.1561/0100000107.

understanding and optimizing aspects of learning systems, including multinomial logistic (softmax) regression and neural network architecture, matrix factorization methods for collaborative filtering and other applications, rank-constrained multivariate linear regression, and forms of semi-supervised learning.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

1

Introduction

In many contemporary and emerging applications of machine learning and statistical inference, the phenomena of interest are characterized by variables defined over large alphabets. Familiar examples, among many others, include the relationship between individual consumers and products that may be of interest to them, and the relationship between images and text in a visual search setting. In such scenarios, not only are the data high-dimensional, but the collection of possible inference tasks is also large. At the same time, training data available to learn the underlying relationships is often quite limited relative to its dimensionality.

From this perspective, for a given level of training data, there is a need to understand which inference tasks can be most effectively carried out, and, in turn, what features of the data are most relevant to them. A natural framework for addressing such questions rather broadly can be traced back to the pioneering work of Hirschfeld [112], building on that of Pearson [223], [224].

In this monograph, we develop an interpretation of the fundamental problem as one of extracting "universally good" features, and establish that diverse notions of such universality lead to precisely the same fea-

The version of record is available at: http://dx.doi.org/10.1561/0100000107

tures. The development emphasizes an information theoretic treatment of the associated questions, and in particular we adopt a convenient "local" information geometric analysis that provides useful insight. In turn, as we describe, the interpretation of such features in terms of a suitable singular value decomposition (SVD) facilitates their computation in a host of applications.

While a variety of the included results exist in one form or another, the treatment aims to be as self-contained as possible. It emphasizes a development from first-principles together with common, unifying terminology and notation, and pointers to the rich embodying literature, both historical and contemporary. Results with no direct or indirect attribution appear here for the first time, to the best of our knowledge. Additionally, to make the treatment as accessible as possible, proofs are largely deferred to appendices, allowing the main text to focus on the statements of key results, their interpretation, and their application.

For the same reason, the development emphasizes distributions over finite alphabets, with the continuous alphabet case largely (but not entirely) focused on Gaussian distributions. These allow the key insights to be revealed while avoiding a variety of technical issues and conditions that would otherwise arise.

At its core, the methodology envisioned by Hirschfeld is based on a particular decomposition of the joint distribution for a pair of variables (X, Y) whose relationship is of interest. Accordingly, we begin by developing and characterizing this decomposition.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

2

The Modal Decomposition of Joint Distributions

As a foundation, in this section we describe the modal decomposition of bivariate distributions over finite alphabets into constituent features that arises out of Hirschfeld's analysis. We develop this decomposition in terms of the SVD of a convenient matrix characterization of the distribution and the associated conditional expectation operator. Several illustrative examples are provided in Section 2.2.

To start, let X and Y denote random variables over finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, with joint distribution[1] $P_{X,Y}$. Without loss of generality we assume throughout that the marginals satisfy $P_X(x) > 0$ and $P_Y(y) > 0$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, since otherwise the associated symbols may be removed from their respective alphabets. Accordingly, we let $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ denote the set of all such distributions.

For an arbitrary *feature*[2] $f\colon \mathcal{X} \to \mathbb{R}$, let $g\colon \mathcal{Y} \to \mathbb{R}$ be the feature induced by f through conditional expectation with respect to $P_{X|Y}(\cdot|y)$, i.e.,

$$g(y) = \mathbb{E}\big[f(X)\big|Y = y\big], \qquad y \in \mathcal{Y}. \tag{2.1}$$

[1] We use (upper case) P notation for the probability mass functions of discrete-valued random variables.

[2] The literature sometimes refers to these as embeddings, referring to functions of embeddings as features. However, our treatment does not require this distinction.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Then we can express (2.1) in the form

$$\begin{aligned} g(y) &= \frac{1}{P_Y(y)} \sum_{x\in\mathcal{X}} P_{X,Y}(x,y)\, f(x) \\ &= \frac{1}{\sqrt{P_Y(y)}} \sum_{x\in\mathcal{X}} \frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)}\,\sqrt{P_Y(y)}} \sqrt{P_X(x)}\, f(x), \end{aligned}$$

i.e.,

$$\xi^Y(y) = \sum_{x\in\mathcal{X}} B(x,y)\, \xi^X(x), \tag{2.2}$$

where we have defined

$$B(x,y) \triangleq \frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)}\,\sqrt{P_Y(y)}}, \quad x\in\mathcal{X},\ y\in\mathcal{Y}, \tag{2.3}$$

and

$$\xi^X(x) \triangleq \sqrt{P_X(x)}\, f(x) \tag{2.4a}$$

$$\xi^Y(y) \triangleq \sqrt{P_Y(y)}\, g(y). \tag{2.4b}$$

Clearly ξ^X and ξ^Y in (2.4) are equivalent representations for f and g respectively. But B in (2.3) is also an equivalent representation for $P_{X,Y}$, as we will verify shortly. Moreover, (2.2) expresses that B has an interpretation as a conditional expectation operator, and thus is equivalent to $P_{X|Y}$.

Next consider an arbitrary feature $\tilde{g}\colon \mathcal{Y} \to \mathbb{R}$, and let $\tilde{f}\colon \mathcal{X} \to \mathbb{R}$ be the feature induced by $\tilde{g}$ through conditional expectation with respect to $P_{Y|X}(\cdot|x)$, i.e.,

$$\tilde{f}(x) = \mathbb{E}\big[\tilde{g}(Y)\big|X = x\big]. \tag{2.5}$$

Then using the notation (2.3) and that analogous to (2.4), i.e.,

$$\tilde{\xi}^X(x) = \sqrt{P_X(x)}\, \tilde{f}(x) \tag{2.6a}$$

$$\tilde{\xi}^Y(y) = \sqrt{P_Y(y)}\, \tilde{g}(y), \tag{2.6b}$$

we can express (2.5) in the form

$$\tilde{\xi}^X(x) = \sum_{y\in\mathcal{Y}} \underbrace{\bar{B}(y,x)}_{\triangleq B(x,y)}\, \tilde{\xi}^Y(y), \tag{2.7}$$

where $\bar{B}$ is the adjoint of B. Likewise $\bar{B}$ is an equivalent representation for $P_{X,Y}$ and, in turn, $P_{Y|X}$.

It is convenient to represent B as a matrix. Specifically, we let $\mathbf{B}$ denote the $|\mathcal{Y}| \times |\mathcal{X}|$ matrix whose (y,x)th entry is $B(x,y)$, i.e.,

$$\mathbf{B} = \left[\sqrt{\mathbf{P}_Y}\right]^{-1} \mathbf{P}_{Y,X} \left[\sqrt{\mathbf{P}_X}\right]^{-1}, \tag{2.8}$$

where $\sqrt{\mathbf{P}_X}$ denotes a $|\mathcal{X}| \times |\mathcal{X}|$ diagonal matrix whose xth diagonal entry is $\sqrt{P_X(x)}$, where $\sqrt{\mathbf{P}_Y}$ denotes a $|\mathcal{Y}| \times |\mathcal{Y}|$ diagonal matrix whose yth diagonal entry is $\sqrt{P_Y(y)}$, and where $\mathbf{P}_{Y,X}$ denotes the $|\mathcal{Y}| \times |\mathcal{X}|$ matrix whose (y,x)th entry is $P_{Y,X}(y,x)$. In [132], $\mathbf{B}$ is referred to as the *divergence transfer matrix (DTM)* associated with $P_{X,Y}$.[3]

Although we generally restrict our attention to the case in which the marginals P_X and P_Y are positive, note that extending the DTM definition to arbitrary nonnegative marginals is straightforward. In particular, it suffices make the x'th column of $\mathbf{B}$ all zeros if $P_X(x') = 0$ for some $x' \in \mathcal{X}$, and, similarly, the y'th row of $\mathbf{B}$ all zeros if $P_Y(y') = 0$ for some $y' \in \mathcal{Y}$, i.e., (2.3) is extended via

$$B(x,y) \triangleq 0, \quad \begin{aligned} &\text{all } x \in \mathcal{X},\ y \in \mathcal{Y} \text{ such that} \\ &P_X(x) = 0 \text{ or } P_Y(y) = 0. \end{aligned} \tag{2.9}$$

Useful alternate forms of B and $\bar{B}$ are [cf. (2.3)]

$$\begin{aligned} B(x,y) &= \frac{P_{Y|X}(y|x)}{\sqrt{P_Y(y)}} \sqrt{P_X(x)} \\ \bar{B}(x,y) &= \frac{P_{X|Y}(x|y)}{\sqrt{P_X(x)}} \sqrt{P_Y(y)}, \end{aligned}$$

from which we obtain the alternate matrix representations

$$\mathbf{B} = \left[\sqrt{\mathbf{P}_Y}\right]^{-1} \mathbf{P}_{Y|X} \left[\sqrt{\mathbf{P}_X}\right] \tag{2.10}$$

$$\mathbf{B}^{\mathrm{T}} = \left[\sqrt{\mathbf{P}_X}\right]^{-1} \mathbf{P}_{X|Y} \left[\sqrt{\mathbf{P}_Y}\right], \tag{2.11}$$

[3] The work of [132], building on [36], focuses on a communication network setting. Subsequently, [122]–[124], [192] develop connections to learning that motivate aspects of, e.g., the present monograph.

where $\mathbf{P}_{Y|X}$ denotes the $|\mathcal{Y}| \times |\mathcal{X}|$ left (column) stochastic transition probability matrix whose (y,x)th entry is $P_{Y|X}(y|x)$, and where, similarly, $\mathbf{P}_{X|Y}$ denotes the $|\mathcal{X}| \times |\mathcal{Y}|$ left (column) stochastic transition probability matrix whose (x,y)th entry is $P_{X|Y}(x|y)$.

The SVD of $\mathbf{B}$ takes the form

$$\begin{aligned} \mathbf{B} &= \sum_{i=0}^{K-1} \sigma_i \, \boldsymbol{\psi}_i^Y \, (\boldsymbol{\psi}_i^X)^\mathrm{T} \\ \text{i.e.,} \quad B(x,y) &= \sum_{i=0}^{K-1} \sigma_i \, \psi_i^X(x) \, \psi_i^Y(y), \end{aligned} \tag{2.12a}$$

with

$$K \triangleq \min\{|\mathcal{X}|, |\mathcal{Y}|\}, \tag{2.12b}$$

where σ_i denotes the ith singular value, where $\boldsymbol{\psi}_i^Y$ and $\boldsymbol{\psi}_i^X$ are the corresponding left and right singular vectors, and where by convention we order the singular values according to

$$\sigma_0 \geq \sigma_1 \geq \cdots \geq \sigma_{K-1}. \tag{2.12c}$$

The following proposition establishes that B (and thus $\bar{B}$) is a contractive operator, a proof of which is provided in Appendix A.1.

Proposition 2.1. For $\mathbf{B}$ defined via (2.8) we have

$$\|\mathbf{B}\|_\mathrm{s} = 1, \tag{2.13}$$

where $\|\cdot\|_\mathrm{s}$ denotes the spectral (i.e., operator) norm of its matrix argument.[4] Moreover, in (2.12), the left and right singular vectors $\boldsymbol{\psi}_0^X$ and $\boldsymbol{\psi}_0^Y$ associated with singular value

$$\sigma_0 = 1 \tag{2.14a}$$

have elements

$$\psi_0^X(x) \triangleq \sqrt{P_X(x)} \qquad \text{and} \qquad \psi_0^Y(y) \triangleq \sqrt{P_Y(y)}. \tag{2.14b}$$

[4] The spectral norm of an arbitrary matrix $\mathbf{A}$ is

$$\|\mathbf{A}\|_\mathrm{s} = \max_i \sigma_i(\mathbf{A}),$$

where $\sigma_i(\mathbf{A})$ denotes the ith singular value of $\mathbf{A}$.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

It follows immediately from the second part of Proposition 2.1 that $\mathbf{B}$ is an equivalent representation for $P_{X,Y}$. Indeed, given $\mathbf{B}$, we can compute the singular vectors ψ_0^X and ψ_0^Y, from which we obtain P_X and P_Y via (2.14b). In turn, using these marginals together with $\mathbf{B}$, whose (y, x)th entry is (2.3), yields $P_{X,Y}(x, y) = B(x, y)\sqrt{P_X(x)}\sqrt{P_Y(y)}$. We provide a more complete characterization of the class of DTMs, i.e., $\mathbf{B}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$ in Appendix A.2. In so doing, we extend the equivalence result above, establishing the continuity of bijective mapping between $P_{X,Y}$ and $\mathbf{B}$.

The SVD (2.12) provides a key expansion of the joint distribution $P_{X,Y}(x, y)$. In particular, we have the following result.

Proposition 2.2. Let $\mathcal{X}$ and $\mathcal{Y}$ denote finite alphabets. Then for any $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$, there exist features $f_i^*\colon \mathcal{X} \to \mathbb{R}$ and $g_i^*\colon \mathcal{Y} \to \mathbb{R}$, for $i = 1, \dots, K-1$, such that

$$P_{X,Y}(x, y) = P_X(x)\, P_Y(y) \left[1 + \sum_{i=1}^{K-1} \sigma_i\, f_i^*(x)\, g_i^*(y)\right], \tag{2.15}$$

where $\sigma_1, \dots, \sigma_{K-1}$ are as defined in (2.12), and where[5]

$$\mathbb{E}\left[f_i^*(X)\right] = 0, \quad i \in \{1, \dots, K-1\} \tag{2.16a}$$

$$\mathbb{E}\left[g_i^*(Y)\right] = 0, \quad i \in \{1, \dots, K-1\} \tag{2.16b}$$

$$\mathbb{E}\left[f_i^*(X)\, f_j^*(X)\right] = \mathbb{1}_{i=j}, \quad i, j \in \{1, \dots, K-1\} \tag{2.16c}$$

$$\mathbb{E}\left[g_i^*(Y)\, g_j^*(Y)\right] = \mathbb{1}_{i=j}, \quad i, j \in \{1, \dots, K-1\}. \tag{2.16d}$$

Moreover, f_i^* and g_i^* are related to the singular vectors in (2.12) according to

$$f_i^*(x) \triangleq \frac{\psi_i^X(x)}{\sqrt{P_X(x)}}, \quad i = 1, \dots, K-1 \tag{2.17a}$$

$$g_i^*(y) \triangleq \frac{\psi_i^Y(y)}{\sqrt{P_Y(y)}}, \quad i = 1, \dots, K-1, \tag{2.17b}$$

[5] We use the Kronecker notation

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

The version of record is available at: http://dx.doi.org/10.1561/0100000107

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.

Proof. It suffices to note that

$$\begin{aligned}
B(x,y) &= \frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)}\sqrt{P_Y(y)}} && (2.18)\\
&= \sqrt{P_X(x)}\,\sqrt{P_Y(y)} + \sum_{i=1}^{K-1} \sigma_i\,\psi_i^X(x)\,\psi_i^Y(y) && (2.19)\\
&= \sqrt{P_X(x)}\,\sqrt{P_Y(y)}\\
&\qquad + \sum_{i=1}^{K-1} \sigma_i \sqrt{P_X(x)}\,f_i^*(x)\,\sqrt{P_Y(y)}\,g_i^*(y) && (2.20)\\
&= \sqrt{P_X(x)}\,\sqrt{P_Y(y)}\left[1 + \sum_{i=1}^{K-1} \sigma_i\,f_i^*(x)\,g_i^*(y)\right], && (2.21)
\end{aligned}$$

where to obtain (2.18) we have used (2.3), to obtain (2.19) we have used (2.12a) with (2.14), and where to obtain (2.20) we have made the choices (2.17), which we note satisfy the constraints (2.16). In particular, (2.16a) follows from the fact that ψ_0^X and ψ_i^X are orthogonal, for $i = 1, \ldots, K-1$, and, likewise, (2.16b) follows from the fact that ψ_0^Y and ψ_i^Y are orthogonal, for $i = 1, \ldots, K-1$. Finally, (2.16c) and (2.16d) follow from the remaining orthogonality relations among the ψ_i^X and ψ_i^Y, respectively. ■

The expansion (2.15) in Proposition 2.2 effectively forms the basis of the methodology introduced by Hirschfeld [112], who sought to extend the applicability of the methods of Pearson [224], [225]. An independent development began with the work of Gebelein [91], upon which the work of Rényi [232] was based.[6] Building on the insights of Hirschfeld, but independent of the work of Gebelein and Rényi, related aspects of such analysis were also explored by Lancaster [158], [160]. The analysis was reinvented again and further developed in [30], [31] using the terminology "correspondence analysis," and further interpreted in [102],

[6] The associated analysis was expressed in terms of eigenvalue decompositions of $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ instead of the SVD of $\mathbf{B}$, since the latter was not widely-used at the time.

[165]. A particular focus of the correspondence analysis literature is on visualization tools and techniques. Subsequent developments appear in [94], [204], and more recent expositions and expansions include [164], [212], and the practical guide [103]. More recently still, [47]–[49] studies aspects of what correspondence analysis terms "principal inertia components" (which are the squared singular values $\sigma_1^2, \dots, \sigma_{K-1}^2$) in the context of information and estimation theory. In particular, generalizing the first principal inertia component (i.e. σ_1^2), the work introduces the term "k-correlation" to refer to $\sigma_1^2 + \cdots + \sigma_k^2$, establishes some properties of k-correlation such as convexity and a data-processing inequality (DPI) [47, Section II], and demonstrates some applications in the context of estimation.

The features (2.17) in (2.15) can be interpreted as suitably normalized sufficient statistics for inferences involving X and Y. Indeed, since

$$P_{Y|X}(y|x) = P_Y(y) \left[1 + \sum_{i=1}^{K-1} \sigma_i \, f_i^*(x) \, g_i^*(y) \right] \tag{2.22a}$$

$$P_{X|Y}(x|y) = P_X(x) \left[1 + \sum_{i=1}^{K-1} \sigma_i \, f_i^*(x) \, g_i^*(y) \right], \tag{2.22b}$$

it follows that[7]

$$f_*^{K-1}(x) \triangleq (f_1^*(x), \dots, f_{K-1}^*(x))$$

is a sufficient statistic for inferences about y based on x, i.e., we have the Markov structure

$$Y \leftrightarrow f_*^{K-1}(X) \leftrightarrow X.$$

Analogously,

$$g_*^{K-1}(y) \triangleq (g_1^*(y), \dots, g_{K-1}^*(y))$$

is a sufficient statistic for inferences about x based on y, i.e., we have the Markov structure

$$X \leftrightarrow g_*^{K-1}(Y) \leftrightarrow Y.$$

[7] Throughout, we use the convenient sequence notation $a^l \triangleq (a_1, \dots, a_l)$.

Moreover, we have[8]

$$X \leftrightarrow f_*^{K-1}(X) \leftrightarrow g_*^{K-1}(Y) \leftrightarrow Y. \tag{2.23}$$

In turn, we have the mutual information (data-processing) relation[9]

$$I(X;Y) = I\big(f_*^{K-1}(X); g_*^{K-1}(Y)\big). \tag{2.24}$$

Additionally, note that Proposition 2.2 has further consequences that are direct result of its connection to the SVD of $\mathbf{B}$. In particular, since the left and right singular vectors are related according to

$$\sigma_i\, \boldsymbol{\psi}_i^Y = \mathbf{B}\, \boldsymbol{\psi}_i^X \tag{2.25a}$$

$$\sigma_i\, \boldsymbol{\psi}_i^X = \mathbf{B}^{\mathrm{T}} \boldsymbol{\psi}_i^Y, \tag{2.25b}$$

it follows from (2.17) that the f_i^* and g_i^* are related according to

$$\sigma_i\, f_i^*(x) = \mathbb{E}\big[g_i^*(Y)\big|X = x\big] \tag{2.26a}$$

$$\sigma_i\, g_i^*(y) = \mathbb{E}\big[f_i^*(X)\big|Y = y\big], \tag{2.26b}$$

for $i = 1, \ldots, K-1$. Moreover, in turn, we obtain, for $i, j \in \{1, \ldots, K-1\}$,

$$\begin{aligned} \mathbb{E}\big[f_i^*(X)\, g_j^*(Y)\big] &= \mathbb{E}\big[\mathbb{E}\big[f_i^*(X)\big|Y = y\big] g_j^*(Y)\big] \\ &= \mathbb{E}\big[\sigma_i\, g_i^*(Y)\, g_j^*(Y)\big] \\ &= \sigma_i\, \mathbb{1}_{i=j}. \end{aligned} \tag{2.27}$$

[8] Indeed, with $F \triangleq f_*^{K-1}(X)$ and $G \triangleq g_*^{K-1}(Y)$, we have

$$P_{G|F,X}(g, f, x) = P_{G|F}(g, f)$$

since $G \leftrightarrow Y \leftrightarrow F \leftrightarrow X$ so $G \leftrightarrow F \leftrightarrow X$, and

$$P_{Y|G,F,X}(y|g, f, x) = P_{Y|G}(y|g).$$

since $Y \leftrightarrow G \leftrightarrow X \leftrightarrow F$ so $Y \leftrightarrow G \leftrightarrow (X, F)$.

[9] Indeed, since

$$X \leftrightarrow F \triangleq f_*^{K-1}(X) \leftrightarrow G \triangleq g_*^{K-1}(Y) \leftrightarrow Y,$$

we have, with $I(\cdot;\cdot)$ denoting mutual information,

$$I(X;Y) = I(F;Y) + \underbrace{I(X;Y|F)}_{=0} = I(F;Y,G) = I(F;G) + \underbrace{I(F;Y|G)}_{=0} = I(F;G).$$

2.1 The Canonical Dependence Matrix

In our development, it is convenient for the analysis to remove the zeroth mode from $\mathbf{B}$. We do this by defining the matrix $\tilde{\mathbf{B}}$ whose (y, x)th entry is

$$\tilde{B}(y,x) \triangleq \frac{P_{X,Y}(x,y) - P_X(x)\,P_Y(y)}{\sqrt{P_X(x)}\,\sqrt{P_Y(y)}} = \sum_{i=1}^{K-1} \sigma_i\,\psi_i^X(x)\,\psi_i^Y(y), \quad (2.28)$$

where in the last equality we have expressed its SVD in terms of that for $\mathbf{B}$, and from which we see that $\tilde{\mathbf{B}}$ has singular values

$$1 \geq \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{K-1} \geq \sigma_K = 0,$$

where we have defined the zero singular value σ_K as a notational convenience. Note that we can interpret $\tilde{B}$ as the conditional expectation operator $\mathbb{E}[\cdot|Y = y]$ restricted to the (sub)space of zero-mean features $f(X)$, which produces a corresponding zero-mean features $g(Y)$. We refer to $\tilde{\mathbf{B}}$, which we can equivalently write in the form[10]

$$\begin{aligned}\tilde{\mathbf{B}} &= \left[\sqrt{\mathbf{P}_Y}\right]^{-1} \left[\mathbf{P}_{Y|X} - \mathbf{P}_Y\,\mathbf{1}\,\mathbf{1}^{\mathrm{T}}\right] \left[\sqrt{\mathbf{P}_X}\right] && (2.29)\\ &= \sum_{i=1}^{K-1} \sigma_i\,\boldsymbol{\psi}_i^Y\,(\boldsymbol{\psi}_i^X)^{\mathrm{T}}, && (2.30)\end{aligned}$$

as the *canonical dependence matrix (CDM)*. Some additional perspectives on this representation of the conditional expectation operator—and thus the particular choice of SVD—are provided in Appendix A.3.

It is worth emphasizing that restricting attention to features of X and Y that are zero-mean is without loss of generality, as there is an invertible mapping between any set of features and their zero-mean counterparts. As a result, we will generally impose this constraint.

2.2 Examples

We conclude with some simple, illustrative examples.

[10] As first used in Appendix A.1, we use $\mathbf{1}$ to denote a vector of all ones (with dimension implied by context).

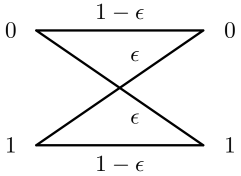

Figure 2.1: A binary symmetric channel with parameter $\epsilon \in (0, 1/2]$.

Example 2.3. Suppose X is uniformly distributed over $\mathcal{X} = \{0, 1\}$, and suppose that Y is the output of a binary symmetric channel whose input is X; specifically, with $\mathcal{Y} = \{0, 1\}$ and for some parameter $\epsilon \in [0, 1/2]$, we have

$$P_{Y|X}(y|x) = \begin{cases} 1-\epsilon & y = x \\ \epsilon & y \neq x, \end{cases} \tag{2.31}$$

as depicted in Figure 2.1. Then Y is also uniformly distributed, and

$$\tilde{B}(x, y) = \begin{cases} 1/2-\epsilon & x = y \\ \epsilon - 1/2 & x \neq y, \end{cases}$$

i.e.,

$$\tilde{\mathbf{B}} = \left(\frac{1}{2} - \epsilon\right) \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

It follows that

$$\sigma_1 = 1 - 2\epsilon \qquad \text{and} \qquad \boldsymbol{\psi}_1^X = \boldsymbol{\psi}_1^Y = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

whence

$$f_1^*(x) = (-1)^x$$

and $g_1^* = f_1^*$. By comparison, the log-likelihood ratio for this model is (when $\epsilon > 0$)

$$\ell(x) \triangleq \log \frac{P_{X|Y}(x|0)}{P_{X|Y}(x|1)} = (-1)^x \log\left(\frac{1-\epsilon}{\epsilon}\right) \propto f_1^*(x).$$

Example 2.4. Suppose X is uniformly distributed over $\mathcal{X} = \{0, 1\}$, and suppose that Y is the output of a binary erasure channel whose input is

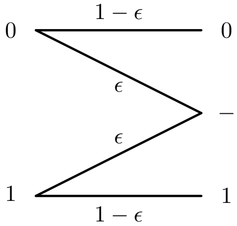

Figure 2.2: A binary erasure channel with parameter $\epsilon \in (0, 1/2]$.

X; specifically, with $\mathcal{Y} = \{0, -, 1\}$ and for some parameter $\epsilon \in [0, 1/2]$, we have

$$P_{Y|X}(y|x) = \begin{cases} 1-\epsilon & y = x \\ \epsilon & y = -, \end{cases} \tag{2.32}$$

as depicted in Figure 2.2. Then

$$P_Y(y) = \begin{cases} (1-\epsilon)/2 & y \in \{0, 1\} \\ \epsilon & y = - \end{cases}$$

and

$$\tilde{B}(x, y) = \begin{cases} \sqrt{1-\epsilon}/2 & (x, y) \in \{(0, 0), (1, 1)\} \\ -\sqrt{1-\epsilon}/2 & (x, y) \in \{(0, 1), (1, 0)\} \\ 0 & y = -, \end{cases}$$

i.e.,

$$\tilde{\mathbf{B}} = \frac{\sqrt{1-\epsilon}}{2} \begin{bmatrix} 1 & -1 \\ 0 & 0 \\ -1 & 1 \end{bmatrix}.$$

It follows that

$$\sigma_1 = \sqrt{1-\epsilon}, \qquad \boldsymbol{\psi}_1^X = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad \boldsymbol{\psi}_1^Y = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix},$$

whence

$$f_1^*(x) = (-1)^x \qquad \text{and} \qquad g_1^*(y) = \begin{cases} (-1)^y/\sqrt{1-\epsilon} & y \in \{0, 1\} \\ 0 & y = -. \end{cases}$$

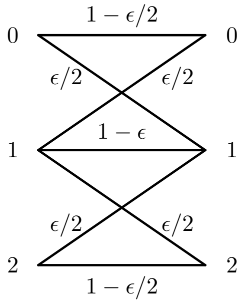

Figure 2.3: A ternary channel with parameter $\epsilon \in (0, 1/2]$.

By comparison, the likelihood ratios for this model are, e.g.,

$$\mathscr{L}_1(x) = \frac{P_{X|Y}(x|0)}{P_{X|Y}(x|-)} = (-1)^x + 1 = f_1^*(x) + 1$$

$$\mathscr{L}_2(x) = \frac{P_{X|Y}(x|1)}{P_{X|Y}(x|-)} = (-1)^{1-x} + 1 = f_1^*(1-x) + 1.$$

Example 2.5. Suppose X is uniformly distributed over $\mathcal{X} = \{0, 1, 2\}$, and suppose Y is the output of a ternary channel whose input is X. In particular, with $\mathcal{Y} = \{0, 1, 2\}$ and for some parameter $\epsilon \in (0, 1/2]$ we have

$$P_{Y|X}(y|x) = \begin{cases} 1-\epsilon/2 & (x,y) \in \{(0,0),(2,2)\} \\ 1-\epsilon & (x,y) = (1,1) \\ \epsilon/2 & (x,y) \in \{(0,1),(2,1),(1,0),(1,2)\} \\ 0 & (x,y) \in \{(0,2),(2,0)\}, \end{cases}$$

as depicted in Figure 2.3. Then Y is also uniformly distributed, and

$$\tilde{\mathbf{B}} = \begin{bmatrix} 1-\epsilon/2 & \epsilon/2 & 0 \\ \epsilon/2 & 1-\epsilon & \epsilon/2 \\ 0 & \epsilon/2 & 1-\epsilon/2 \end{bmatrix} - \frac{1}{3} \cdot \mathbf{1} \cdot \mathbf{1}^{\mathrm{T}}.$$

It follows that

$$\sigma_1 = 1 - \frac{\epsilon}{2}, \qquad \sigma_2 = 1 - \frac{3}{2}\epsilon,$$

and

$$\boldsymbol{\psi}_1^X = \boldsymbol{\psi}_1^Y = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \qquad \boldsymbol{\psi}_2^X = \boldsymbol{\psi}_2^Y = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix},$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

whence

$$f_1^*(x) = \begin{cases} \sqrt{3/2} & x = 0 \\ 0 & x = 1 \\ -\sqrt{3/2} & x = 2, \end{cases} \qquad f_2^*(x) = \begin{cases} 1/\sqrt{2} & x = 0 \\ -\sqrt{2} & x = 1 \\ 1/\sqrt{2} & x = 2, \end{cases}$$

and $g_i^* = f_i^*$ for $i \in \{1, 2\}$. By comparison, likelihood ratios for this model are, e.g.,

$$\mathscr{L}_1(x) = \frac{P_{X|Y}(x|0)}{P_{X|Y}(x|1)} = \begin{cases} a & x = 0 \\ b & x = 1 \\ 0 & x = 2 \end{cases} \qquad a \triangleq \frac{1 - \epsilon/2}{\epsilon/2}$$

$$\mathscr{L}_2(x) = \frac{P_{X|Y}(x|2)}{P_{X|Y}(x|1)} = \begin{cases} 0 & x = 0 \\ b & x = 1 \\ a & x = 2 \end{cases} \qquad b \triangleq \frac{\epsilon/2}{1 - \epsilon}.$$

It is straightforward to verify that there exists an invertible transformation between (f_1^*, f_2^*) and $(\mathscr{L}_1, \mathscr{L}_2)$; in particular, one can readily construct a one-to-one function $\vartheta\colon f_1^*(\mathcal{X}) \times f_2^*(\mathcal{X}) \to \mathscr{L}_1(\mathcal{X}) \times \mathscr{L}_2(\mathcal{X})$ such that $\vartheta(f_1^*(x), f_2^*(x)) = (\mathscr{L}_1(x), \mathscr{L}_2(x))$ for all $x \in \mathcal{X}$.

Example 2.6. Suppose X has a (positive) distribution P_X over $\mathcal{X} = \{1, \dots, K\}$ for some $K > 0$, and $Y = X$. Then $P_{X,Y}(x, y) = P_X(x)\, \mathbb{1}_{y=x}$ and $P_Y = P_X$, and $\tilde{\mathbf{B}}$ has entries

$$\tilde{B}(x, y) = \begin{cases} 1 - P_X(x) & x = y \\ -\sqrt{P_X(x)\, P_Y(y)} & x \neq y, \end{cases}$$

which we note is symmetric and idempotent (i.e., an orthogonal projection matrix): $\tilde{\mathbf{B}}^2 = \tilde{\mathbf{B}} = \tilde{\mathbf{B}}^{\mathrm{T}}$. Hence, $\sigma_1, \dots, \sigma_{K-1} \in \{0, 1\}$. Moreover,[11]

$$\tilde{\mathbf{B}} - \mathbf{I} = \boldsymbol{\psi}_0^X (\boldsymbol{\psi}_0^X)^{\mathrm{T}},$$

with $\boldsymbol{\psi}_0^X$ as defined via (2.14b). It follows that $\sigma_1 = \dots = \sigma_{K-1} = 1$, and the corresponding (left and right) singular vectors span the orthogonal complement of the vector $\boldsymbol{\psi}_0^X$.

[11] We use $\mathbf{I}$ to denote the identity matrix of appropriate dimension.

That all the singular values (except σ_K) are unity reflects that Y is perfectly predictable from X (and vice-versa). Note that we obtain the same modal decomposition structure when, more generally, $Y = \vartheta(X)$ for some one-to-one map $\vartheta \colon \mathcal{X} \to \vartheta(\mathcal{X})$, i.e, $\mathcal{Y} = \vartheta(\mathcal{X})$ is a relabeling of the symbols in $\mathcal{X}$.

Example 2.7. For $\delta \in (0,1)$, suppose $\mathcal{X} = \mathcal{Y} = \{0,1,2\}$, and

$$P_{X,Y}(x,y) = \begin{cases} \delta & (x,y) = (0,0) \\ (1-\delta)/4 & (x,y) \in \{1,2\} \times \{1,2\} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$P_X(x) = \begin{cases} \delta & x = 0 \\ (1-\delta)/2 & x \in \{1,2\}, \end{cases}$$

and $P_Y = P_X$, and thus

$$\tilde{\mathbf{B}} = \begin{bmatrix} 1-\delta & -\sqrt{\delta(1-\delta)/2} & -\sqrt{\delta(1-\delta)/2} \\ -\sqrt{\delta(1-\delta)/2} & \delta/2 & \delta/2 \\ -\sqrt{\delta(1-\delta)/2} & \delta/2 & \delta/2 \end{bmatrix}.$$

It follows that

$$\sigma_i = \begin{cases} 1 & i = 0 \\ 0 & i \in \{1,2\} \end{cases} \qquad \text{and} \qquad \boldsymbol{\psi}_1^X = \boldsymbol{\psi}_2^Y = \begin{bmatrix} -\sqrt{1-\delta} \\ \sqrt{\delta/2} \\ \sqrt{\delta/2} \end{bmatrix},$$

and hence

$$f_1^*(x) = \begin{cases} -\sqrt{(1-\delta)/\delta} & x = 0 \\ \sqrt{\delta/(1-\delta)} & x \in 1,2 \end{cases}$$

with $g_1^* = f_1^*$. That there is a unit singular value reflects that part of Y is perfectly predictable from X (and vice-versa).

Example 2.8 ([81]). For $n > 0$, suppose $C, A_1, \dots, A_{n-1}, B_1, \dots, B_{n-1}$ are independent and uniformly distributed on $\{0,1\}$, and let

$$X = C + \sum_{i=1}^{n-1} A_i 2^i \qquad \text{and} \qquad Y = C + \sum_{i=1}^{n-1} B_i 2^i.$$

Then X and Y are uniformly distributed on $\mathcal{X} = \mathcal{Y} = \{0, 1, \dots, 2^n - 1\}$ and

$$P_{X,Y}(x, y) = \begin{cases} 1/2^{n+1} & x - y \text{ is even} \\ 0 & x - y \text{ is odd,} \end{cases}$$

so

$$\tilde{B}(x, y) = (-1)^{x-y}/2^n, \qquad x \in \mathcal{X},\ y \in \mathcal{Y}.$$

For example, for $n = 2$, we have

$$\tilde{\mathbf{B}} = \frac{1}{4} \begin{bmatrix} 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{bmatrix}.$$

It follows that $\sigma_i = \mathbb{1}_{i=1}$ and $\psi_1^Y = \psi_1^X$ with $\psi_1^X(x) = (-1)^x/2^{n/2}$, whence $f_1^*(x) = (-1)^x$ and $g_1^* = f_1^*$. That there is a unit singular value again reflects that part of Y is perfectly predictable from X (and vice-versa).

Example 2.9 ([75]). Suppose $U \in \{-1, 1\}$, $V \in \{-1, 1\}$, and $W \in \{0, 1\}$ are independent random variables with $\mathbb{P}(U = 1) = \mathbb{P}(V = 1) = 1/2$ and $\mathbb{P}(W = 1) = \epsilon \in (0, 1)$. In turn, let $X = UW$ and $Y = VW$, so $\mathcal{X} = \mathcal{Y} = \{-1, 0, 1\}$. Then since $|X| = |Y|$, it follows that $\sigma_1 = 1$. A more detailed analysis yields

$$\tilde{\mathbf{B}} = \begin{bmatrix} (1-\epsilon)/2 & -\sqrt{\epsilon(1-\epsilon)/2} & (1-\epsilon)/2 \\ -\sqrt{\epsilon(1-\epsilon)/2} & \epsilon & -\sqrt{\epsilon(1-\epsilon)/2} \\ (1-\epsilon)/2 & -\sqrt{\epsilon(1-\epsilon)/2} & (1-\epsilon)/2 \end{bmatrix},$$

from which we obtain

$$\sigma_1 = 1, \quad \sigma_2 = 0 \qquad \text{and} \qquad f_1^*(x) = (-1)^{|x|} \left(\frac{1-\epsilon}{\epsilon} \right)^{|x|-1/2},$$

and $g_1^* = f_1^*$. More generally, for an arbitrary numeric random variable W we obtain $\sigma_1 = 1$, as discussed in [75].

Example 2.10. Suppose $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$ and

$$P_{X,Y}(x, y) = \begin{cases} 1/2 & (x, y) = (0, 0) \\ 1/6 & (x, y) \in \{(1, 1), (2, 2)\} \\ 1/12 & (x, y) \in \{(1, 2), (2, 1)\} \\ 0 & (x, y) \in \{(0, 1), (1, 0), (0, 2), (2, 0)\}, \end{cases}$$

so

$$P_X(x) = \begin{cases} 1/2 & x = 0 \\ 1/4 & x \in \{1,2\} \end{cases}$$

and $P_Y = P_X$. Then

$$\tilde{\mathbf{B}} = \begin{bmatrix} 1/2 & -1/\sqrt{8} & -1/\sqrt{8} \\ -1/\sqrt{8} & 5/12 & 1/12 \\ -1/\sqrt{8} & 1/12 & 5/12 \end{bmatrix}.$$

It follows that

$$\sigma_1 = 1 \qquad \text{and} \qquad \sigma_2 = 1/3,$$

and

$$\boldsymbol{\psi}_1^X = \boldsymbol{\psi}_1^Y = \begin{bmatrix} -1/\sqrt{2} \\ 1/2 \\ 1/2 \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\psi}_2^X = \boldsymbol{\psi}_2^Y = \begin{bmatrix} 0 \\ 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix},$$

whence

$$f_1^*(x) = \begin{cases} -1 & x = 0 \\ 1 & x \in \{1,2\} \end{cases} \qquad \text{and} \qquad f_2^*(x) = \begin{cases} 0 & x = 0 \\ 1 & x = 1 \\ -1 & x = 2, \end{cases}$$

and $(g_1^*, g_2^*) = (f_1^*, f_2^*)$. We note that, consistent with (2.27), we have $\mathbb{P}(f_1^*(X) = g_1^*(Y)) = 1$.

3

Variational Characterization of the Modal Decomposition

There is a natural and insightful variational characterization of the modal decomposition of Section 2. In this section, we develop this alternative view in terms of standard SVD analysis, following an approach to which both Gebelein [91] and Rényi [232] made foundational contributions. Accordingly, the result is often referred to as Hirschfeld-Gebelein-Rényi (HGR) maximal correlation analysis. Using this analysis, we interpret HGR maximal correlation as the Ky Fan k-norm of the CDM, and obtain the features defining the modal decomposition via an optimization.

We develop the desired variational characterization of the feature functions (2.17), viz., (f_i^*, g_i^*), $i = 1, 2, \dots K - 1$, in Section 3.2, after first summarizing the requisite linear algebra in the following section.

3.1 Variational Characterizations of the SVD

Some classical variational results on the SVD that will be useful in our analysis. First, we have the following lemma (see, e.g., [114, Corollary 4.3.39, p. 248]).

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Lemma 3.1. Given an arbitrary $k_1 \times k_2$ matrix $\mathbf{A}$ and any $k \in \{1, \dots, \min\{k_1, k_2\}\}$, we have

$$\max_{\left\{\mathbf{M} \in \mathbb{R}^{k_2 \times k} :\, \mathbf{M}^{\mathrm{T}}\mathbf{M} = \mathbf{I}\right\}} \|\mathbf{A}\mathbf{M}\|_{\mathrm{F}}^2 = \sum_{i=1}^{k} \sigma_i(\mathbf{A})^2, \tag{3.1}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of its matrix argument,[1] and where $\sigma_1(\mathbf{A}) \geq \cdots \geq \sigma_{\min\{k_1,k_2\}}(\mathbf{A})$ denote the (ordered) singular values of $\mathbf{A}$. Moreover, the maximum in (3.1) is achieved by

$$\mathbf{M} = \begin{bmatrix} \boldsymbol{\psi}_1(\mathbf{A}) & \cdots & \boldsymbol{\psi}_k(\mathbf{A}) \end{bmatrix}, \tag{3.2}$$

with $\boldsymbol{\psi}_i(\mathbf{A})$ denoting the right singular vector of $\mathbf{A}$ corresponding to $\sigma_i(\mathbf{A})$, for $i = 1, \dots, \min\{k_1, k_2\}$.

Second, the following lemma, essentially due to von Neumann (see, e.g., [174] [114, Theorem 7.4.1.1]), will also be useful in our analysis, and can be obtained using Lemma 3.1 in conjunction with the Cauchy-Schwarz inequality.

Lemma 3.2. Given an arbitrary $k_1 \times k_2$ matrix $\mathbf{A}$, we have

$$\max_{\substack{\left\{\mathbf{M}_1 \in \mathbb{R}^{k_1 \times k},\, \mathbf{M}_2 \in \mathbb{R}^{k_2 \times k} :\right. \\ \left.\mathbf{M}_1^{\mathrm{T}}\mathbf{M}_1 = \mathbf{M}_2^{\mathrm{T}}\mathbf{M}_2 = \mathbf{I}\right\}}} \operatorname{tr}\left(\mathbf{M}_1^{\mathrm{T}}\mathbf{A}\mathbf{M}_2\right) = \sum_{i=1}^{k} \sigma_i(\mathbf{A}), \tag{3.3}$$

with $\sigma_1(\mathbf{A}) \geq \cdots \geq \sigma_{\min\{k_1,k_2\}}(\mathbf{A})$ denoting the (ordered) singular values of $\mathbf{A}$. Moreover, the maximum in (3.3) is achieved by

$$\mathbf{M}_j = \begin{bmatrix} \boldsymbol{\psi}_1^{(j)}(\mathbf{A}) & \cdots & \boldsymbol{\psi}_k^{(j)}(\mathbf{A}) \end{bmatrix}, \quad j = 1, 2, \tag{3.4}$$

with $\boldsymbol{\psi}_i^{(1)}(\mathbf{A})$ and $\boldsymbol{\psi}_i^{(2)}(\mathbf{A})$ denoting the left and right singular vectors, respectively, of $\mathbf{A}$ corresponding to $\sigma_i(\mathbf{A})$, for $i = 1, \dots, \min\{k_1, k_2\}$.

[1] Specifically, the Frobenius norm of an arbitrary matrix $\mathbf{A}$ is

$$\|\mathbf{A}\|_{\mathrm{F}} \triangleq \operatorname{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right) = \sum_i \sigma_i(\mathbf{A})^2,$$

where $\sigma_i(\mathbf{A})$ denotes the ith singular value of $\mathbf{A}$, and were $\operatorname{tr}(\cdot)$ denotes the trace of its matrix argument.

3.2 Maximal Correlation Features

We now have the following result, which relates the modal decomposition and correlation maximization, and reveals the role of Ky Fan k-norms (as defined in, e.g., [114, Section 7.4.8]) in the analysis.

Proposition 3.3. For any $k \in \{1, \dots, K-1\}$, the dominant k features (2.17) in Proposition 2.2, i.e.,

$$f_*^k \triangleq (f_1^*, \dots, f_k^*) \quad \text{and} \quad g_*^k \triangleq (g_1^*, \dots, g_k^*), \tag{3.5}$$

are obtained via[2]

$$(f_*^k, g_*^k) = \operatorname*{arg\,min}_{(f^k, g^k) \in \mathcal{F}_k \times \mathcal{G}_k} \mathbb{E}\left[\left\|f^k(X) - g^k(Y)\right\|^2\right] = \operatorname*{arg\,max}_{(f^k, g^k) \in \mathcal{F}_k \times \mathcal{G}_k} \sigma(f^k, g^k), \tag{3.6a}$$

where

$$\sigma(f^k, g^k) \triangleq \mathbb{E}\left[\left(f^k(X)\right)^{\mathrm{T}} g^k(Y)\right] \tag{3.6b}$$

and

$$\mathcal{F}_k \triangleq \left\{f^k : \mathbb{E}\left[f^k(X)\right] = \mathbf{0}, \quad \mathbb{E}\left[f^k(X)\, f^k(X)^{\mathrm{T}}\right] = \mathbf{I}\right\} \tag{3.6c}$$

$$\mathcal{G}_k \triangleq \left\{g^k : \mathbb{E}\left[g^k(Y)\right] = \mathbf{0}, \quad \mathbb{E}\left[g^k(Y)\, g^k(Y)^{\mathrm{T}}\right] = \mathbf{I}\right\}. \tag{3.6d}$$

Moreover, the resulting maximal correlation is

$$\sigma(f_*^k, g_*^k) = \mathbb{E}\left[\left(f_*^k(X)\right)^{\mathrm{T}} g_*^k(Y)\right] = \sum_{i=1}^{k} \sigma_i, \tag{3.7}$$

which we note is the Ky Fan k-norm of $\tilde{\mathbf{B}}$.[3]

[2] We use $\|\cdot\|$ to denote the Euclidean norm, i.e., $\left\|a^k\right\| = \sqrt{\sum_{i=1}^k a_i^2}$ for any k and a^k.

[3] We use $\|\cdot\|_{(k)}$ to denote the Ky Fan k-norm of its argument, i.e., for $\mathbf{A} \in \mathbb{R}^{k_1 \times, k_2}$,

$$\|\mathbf{A}\|_{(k)} \triangleq \sum_{i=1}^{k} \sigma_i(\mathbf{A}), \tag{3.8}$$

with $\sigma_1(\mathbf{A}) \geq \cdots \geq \sigma_k(\mathbf{A})$ denoting the singular values of $\mathbf{A}$, for $k \in \{1, \min\{k_1, k_2\}\}$.

Proof. First, note that the constraints (3.6c) and (3.6d) express (2.16) in Proposition 2.2. Next, to facilitate our development, we define [cf. (2.4)]

$$\xi_i^X(x) \triangleq \sqrt{P_X(x)}\, f_i(x), \quad x \in \mathcal{X} \tag{3.9a}$$

$$\xi_i^Y(y) \triangleq \sqrt{P_Y(y)}\, g_i(y), \quad y \in \mathcal{Y}. \tag{3.9b}$$

for $i = 1, \dots, K$ We refer to ξ_i^X and ξ_i^Y as the *feature vectors* associated with the feature functions f_i and g_i, respectively, and we further use $\boldsymbol{\xi}_i^X$ and $\boldsymbol{\xi}_i^Y$ to denote column vectors whose xth and yth entries are $\xi_i^X(x)$ and $\xi_i^Y(y)$, respectively. Then

$$\sigma(f^k, g^k) = \sum_{i=1}^{k} \sigma_i(f_i, g_i) \tag{3.10a}$$

with

$$\sigma_i(f_i, g_i) = \mathbb{E}[f_i(X)\, g_i(Y)] = (\boldsymbol{\xi}_i^Y)^{\mathrm{T}} \mathbf{B}\, \boldsymbol{\xi}_i^X = (\boldsymbol{\xi}_i^Y)^{\mathrm{T}} \tilde{\mathbf{B}}\, \boldsymbol{\xi}_i^X, \tag{3.10b}$$

where the last equality in (3.10b) follows from the mean constraints in (3.6c) and (3.6d), which imply, for $i = 1, \dots, K$,

$$\sum_{x \in \mathcal{X}} \sqrt{P_X(x)}\, \xi_i^X(x) = \sum_{y \in \mathcal{Y}} \sqrt{P_Y(y)}\, \xi_i^Y(y) = 0.$$

In turn, from (3.10) we have

$$\sigma(f^k, g^k) = \sum_{i=1}^{k} (\boldsymbol{\xi}_i^Y)^{\mathrm{T}} \tilde{\mathbf{B}}\, \boldsymbol{\xi}_i^X = \operatorname{tr}\Big((\boldsymbol{\Xi}^Y)^{\mathrm{T}} \tilde{\mathbf{B}}\, \boldsymbol{\Xi}^X \Big), \tag{3.11}$$

where

$$\boldsymbol{\Xi}^X \triangleq \begin{bmatrix} \boldsymbol{\xi}_1^X & \cdots & \boldsymbol{\xi}_k^X \end{bmatrix} \tag{3.12a}$$

$$\boldsymbol{\Xi}^Y \triangleq \begin{bmatrix} \boldsymbol{\xi}_1^Y & \cdots & \boldsymbol{\xi}_k^Y \end{bmatrix}. \tag{3.12b}$$

Moreover, from the covariance constraints in (3.6c) and (3.6d) we have

$$(\boldsymbol{\Xi}^X)^{\mathrm{T}} \boldsymbol{\Xi}^X = (\boldsymbol{\Xi}^Y)^{\mathrm{T}} \boldsymbol{\Xi}^Y = \mathbf{I}. \tag{3.13}$$

Hence, applying Lemma 3.2 we immediately obtain that (3.11) is maximized subject to (3.13) by the feature vectors

$$\boldsymbol{\Xi}^X = \boldsymbol{\Psi}^X_{(k)} \tag{3.14a}$$

$$\boldsymbol{\Xi}^Y = \boldsymbol{\Psi}^Y_{(k)}, \tag{3.14b}$$

with

$$\boldsymbol{\Psi}^X_{(k)} \triangleq \begin{bmatrix} \boldsymbol{\psi}^X_1 & \cdots & \boldsymbol{\psi}^X_k \end{bmatrix} \tag{3.15a}$$

$$\boldsymbol{\Psi}^Y_{(k)} \triangleq \begin{bmatrix} \boldsymbol{\psi}^Y_1 & \cdots & \boldsymbol{\psi}^Y_k \end{bmatrix}, \tag{3.15b}$$

whence f_i^* and g_i^* as given by (2.17), for $i = 1, \dots, k$. The final statement of the proposition follows immediately from the properties of the SVD; specifically,

$$(\boldsymbol{\psi}^Y_i)^{\mathrm{T}} \tilde{\mathbf{B}} \, \boldsymbol{\psi}^X_i = \sigma_i, \quad i = 1, \dots, k,$$

i.e.,

$$\left(\boldsymbol{\Psi}^Y_{(k)}\right)^{\mathrm{T}} \tilde{\mathbf{B}} \, \boldsymbol{\Psi}^X_{(k)} = \boldsymbol{\Sigma}_{(k)}, \tag{3.16}$$

with $\boldsymbol{\Sigma}_{(k)}$ denoting a $(k \times k)$ diagonal matrix with diagonal entries $\sigma_1, \dots \sigma_k$. ■

The quantity (3.7) is often referred to as the Hirschfeld-Gebelein-Rényi (HGR) maximal correlation associated with the distribution $P_{X,Y}$, particularly in the special case $k = 1$. Note that when $k = 1$, the Ky Fan k-norm specializes to the spectral norm. In practice, larger values of k are generally more useful for measuring the degree of dependence. Indeed, as follows from the discussion in [81], [106], [149], $\|\tilde{\mathbf{B}}\|_{\mathrm{s}}$ can achieve its maximum of unity even when X and Y are nearly independent, as Example 2.8 (reproduced from [81]) illustrates when n is large. By contrast, in this example, $\|\tilde{\mathbf{B}}\|_{(k)} = 1$ when $k = 2^n - 1$, which is much smaller than its maximum possible value of $2^n - 1$ in this regime, reflecting the weak dependence.

4

Local Information Geometry

The introduction to modal decompositions in the preceding sections emphasizes a Euclidean geometry. In contrast, the standard information geometry based on Kullback-Leibler (KL) divergence [8], [9], [63], [71], which underlies the information-theoretic analysis of inference and learning, is nonEuclidean. However, information geometry is *locally* Euclidean, and thus valuable information-theoretic perspectives can be obtained through a local geometric analysis on the (relative interior of the) probability simplex. This analysis corresponds to the use of χ^2-divergence. The roots of such analysis date back at least to the work of Pearson, whose *mean-square contingency* measure is based on this divergence [223], and, later, Hirschfeld [112].

In this section, we develop the required foundations of this analysis. In the resulting Euclidean information space, distributions are represented as information vectors, and features as feature vectors, and we develop an equivalence between them via log-likelihoods. Via this geometry, we develop a suitable notion of weakly dependent variables for which we obtain a decomposition of mutual information and through which we interpret truncated modal decompositions as "information efficient." Additionally, we characterize the error exponents in local decision-making in terms of (mismatched) feature projections.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Further interpretation of the features f_*^{K-1} and g_*^{K-1} arising out of the modal decomposition of Section 2 benefits from developing the underlying inner product space. More specifically, a local analysis of information geometry leads to key information-theoretic interpretations of (2.17) as *universal* features. Accordingly, we begin with some basic definitions.

4.1 Basic Concepts, Terminology, and Notation

Let $\mathcal{P}^{\mathcal{Z}}$ denote the space of distributions on some finite alphabet $\mathcal{Z}$, where $|\mathcal{Z}| < \infty$, and let $\operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$ denote the relative interior of $\mathcal{P}^{\mathcal{Z}}$, i.e., the subset of positive distributions.

Definition 4.1 (ϵ-Neighborhood)**.** For a given $\epsilon > 0$, the ϵ-neighborhood of a reference distribution $P_0 \in \operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$ is the set of distributions in a (Neyman) χ^2-divergence [211] ball of radius ϵ^2 about P_0, i.e.,

$$\mathcal{N}_{\epsilon}^{\mathcal{Z}}(P_0) \triangleq \left\{ P' \in \mathcal{P}^{\mathcal{Z}} \colon \chi^2(P' \| P_0) \leq \epsilon^2 \right\}, \tag{4.1a}$$

where for $P \in \mathcal{P}^{\mathcal{Z}}$ and $Q \in \operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$,

$$\chi^2(P \| Q) \triangleq \sum_{z \in \mathcal{Z}} \frac{\left(Q(z) - P(z)\right)^2}{Q(z)}. \tag{4.1b}$$

The use of χ^2-divergence is both convenient and natural, as it is a second-order approximation to KL divergence [71]; specifically,[1]

[1] More generally, [71, Theorem 4.1] establishes that χ^2-divergence is a second-order approximation to any f-divergence

$$D_f(P \| Q) = \sum_{z \in \mathcal{Z}} Q(z)\, f\!\left(\frac{P(z)}{Q(z)}\right)$$

for which $f''(1)$ exists and is positive, where $f \colon (0, \infty) \to \mathbb{R}$ is convex and satisfies $f(1) = 0$.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

$$\begin{aligned}
D(P\|Q) &= \sum_{z\in\mathcal{Z}} P(z)\,\log\frac{P(z)}{Q(z)} \\
&= \sum_{z\in\mathcal{Z}} [Q(z) + (P(z) - Q(z))]\,\log\left[1 + \left(\frac{P(z)-Q(z)}{Q(z)}\right)\right] \\
&= \sum_{z\in\mathcal{Z}} [Q(z) + (P(z) - Q(z))] \sum_{l=1}^{\infty} \frac{(-1)^{l-1}}{l} \left(\frac{P(z)-Q(z)}{Q(z)}\right)^{l} \\
&= \frac{1}{2}\chi^2(P\|Q) + o(\chi^2(P\|Q)), \qquad \chi^2(P\|Q) \to 0,
\end{aligned}$$

where we have used the Taylor series expansion

$$\log(1+\omega) = \sum_{l=1}^{\infty} \frac{(-1)^{l-1}}{l}\,\omega^l.$$

In the sequel, we assume that all the distributions of interest, including all empirical distributions that may be observed, lie in such an ϵ-neighborhood of the prescribed P_0. While we don't restrict ϵ to be small, most of the information-theoretic insights arise from the asymptotics corresponding to $\epsilon \to 0$.

An equivalent representation for a distribution $P \in \mathcal{N}_\epsilon^{\mathcal{Z}}(P_0)$ is in terms of its *information vector*

$$\phi(z) \triangleq \frac{P(z) - P_0(z)}{\epsilon\sqrt{P_0(z)}}, \tag{4.2}$$

which we note satisfies

$$\|\phi\| \le 1, \tag{4.3}$$

with $\|\cdot\|$ denoting the usual Euclidean norm.[2] We will sometimes find it convenient to express $\phi = \phi(\cdot)$ as a $|\mathcal{Z}|$-dimensional column vector $\boldsymbol{\phi}$, according to some arbitrarily chosen but fixed ordering of the elements of $\mathcal{Z}$.

[2] Specifically, for ϕ defined on $\mathcal{Z}$,

$$\|\phi\|^2 \triangleq \sum_{z\in\mathcal{Z}} \phi(z)^2.$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Hence, we can equivalently interpret the $(|\mathcal{Z}|-1)$-dimensional neighborhood $\mathcal{N}_\epsilon^{\mathcal{Z}}(P_0)$ as the set of distributions whose corresponding information vectors lie in the unit Euclidean ball about the origin. Note that since

$$\sum_{z\in\mathcal{Z}} \sqrt{P_0(z)}\,\phi(z) = 0, \tag{4.4}$$

the $(|\mathcal{Z}|-1)$-dimensional vector space subset

$$\mathcal{I}^{\mathcal{Z}}(P_0) = \Big\{\phi\colon \langle\sqrt{P_0},\phi\rangle = 0 \text{ and } \|\phi\| \le 1\Big\}, \tag{4.5}$$

with $\langle\cdot,\cdot\rangle$ denoting the usual Euclidean inner product,[3] characterizes all the possible information vectors: $\phi \in \mathcal{I}^{\mathcal{Z}}(P_0)$ if and only if $P \in \mathcal{N}_\epsilon^{\mathcal{Z}}(P_0)$, for all ϵ sufficiently small. It is convenient to refer to $\mathcal{I}^{\mathcal{Z}}(P_0)$ as *information space*. When the relevant reference distribution P_0 is clear from context we will generally omit it from our notation, and simply use $\mathcal{I}^{\mathcal{Z}}$ to refer to this space.

For a feature function $h\colon \mathcal{Z}\to\mathbb{R}$, we refer to

$$\xi(z) \triangleq \sqrt{P_0(z)}\,h(z) \tag{4.6}$$

as its associated *feature vector*.[4] As with information vectors, we will sometimes find it convenient to express $\xi = \xi(\cdot)$ as a $|\mathcal{Z}|$-dimensional column vector $\boldsymbol{\xi}$, according to the chosen ordering of the elements of $\mathcal{Z}$. Moreover, there is an effective equivalence of feature vectors and information vectors, which the following proposition establishes. A proof is provided in Appendix B.1.

Proposition 4.2. Let $P_0 \in \operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$ be an arbitrary reference distribution, and ϵ a positive constant. Then for any distribution $P \in \mathcal{P}^{\mathcal{Z}}$,

$$h(z) = \frac{1}{\epsilon}\left(\frac{P(z)}{P_0(z)} - 1\right) \tag{4.7}$$

[3] Specifically, for ϕ_1 and ϕ_2 defined on $\mathcal{Z}$,

$$\langle\phi_1,\phi_2\rangle \triangleq \sum_{z\in\mathcal{Z}} \phi_1(z)\,\phi_2(z).$$

[4] Note that is a simple generalization of the terminology introduced after (3.9).

is a feature function satisfying

$$\mathbb{E}_{P_0}\big[h(Z)\big] = 0, \tag{4.8}$$

and has as its feature vector the information vector of $P(z)$, i.e.,

$$\xi(z) = \phi(z) = \frac{P(z) - P_0(z)}{\epsilon\sqrt{P_0(z)}}. \tag{4.9}$$

Conversely, for any feature function $h\colon \mathcal{Z} \to \mathbb{R}$ such that (4.8) holds,

$$P(z) = P_0(z)\big(1 + \epsilon h(z)\big) \tag{4.10}$$

is a valid distribution for all ϵ sufficiently small, and has as its information vector the feature vector of h, i.e.,

$$\phi(z) = \xi(z) = \sqrt{P_0(z)}\, h(z). \tag{4.11}$$

The following corollary of Proposition 4.2 specific to the case of (relative) *log-likelihood* feature functions is further useful in our analysis. A proof is provided in Appendix B.2.

Corollary 4.3. Let $P_0 \in \operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$ be an arbitrary reference distribution and ϵ a positive constant. Then for any distribution $P \in \mathcal{N}_\epsilon^{\mathcal{Z}}(P_0)$ with associated information vector ϕ, the feature vector ξ_{LL} associated with the relative log-likelihood feature function[5]

$$h_{\mathrm{LL}}(z) \triangleq \frac{1}{\epsilon}\left(\log\frac{P(z)}{P_0(z)} - \mathbb{E}_{P_0}\left[\log\frac{P(Z)}{P_0(Z)}\right]\right), \quad z \in \mathcal{Z} \tag{4.12}$$

satisfies[6]

$$\xi_{\mathrm{LL}}(z) = \phi(z) + o(1), \quad \epsilon \to 0, \quad z \in \mathcal{Z}. \tag{4.13}$$

Conversely, every feature function $h\colon \mathcal{Z} \to \mathbb{R}$ satisfying $\mathbb{E}_{P_0}\big[h(Z)\big] = 0$ can be interpreted to first order as a (relative) log-likelihood, i.e., can be expressed in the form

$$h(z) = \frac{1}{\epsilon}\left(\log\frac{P(z)}{P_0(z)} - \mathbb{E}_{P_0}\left[\log\frac{P(Z)}{P_0(Z)}\right]\right) + o(1), \quad \epsilon \to 0, \quad z \in \mathcal{Z}, \tag{4.14}$$

for some $P \in \mathcal{P}^{\mathcal{Z}}$

[5] Throughout, all logarithms are base e, i.e., natural.

[6] Note that the $o(1)$ term has zero mean with respect to P_0, consistent with $\xi_{\mathrm{LL}} \in \mathcal{I}^{\mathcal{Z}}(P_0)$.

A consequence of Proposition 4.2 is that we do not need to distinguish between feature vectors and information vectors in the underlying inner product space. Indeed, note that when without loss of generality we normalize a feature h so that both (4.8) and

$$\mathbb{E}_{P_0}\left[h(Z)^2\right] = 1,$$

are satisfied, then we have $\xi \in \mathcal{I}^{\mathcal{Z}}(P_0)$, where ξ is the feature vector associated with h, as defined in (4.6).

The following lemma, verified in Appendix B.3, interprets inner products between feature vectors and information vectors.

Lemma 4.4. For any $P_0 \in \operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$, let h be a feature function satisfying (4.8) with associated feature vector $\xi \in \mathcal{I}^{\mathcal{Z}}(P_0)$. Then for any $\epsilon > 0$ and $P \in \mathcal{N}_{\epsilon}^{\mathcal{Z}}(P_0)$ with associated information vector $\phi \in \mathcal{I}^{\mathcal{Z}}(P_0)$,

$$\mathbb{E}_P[h(Z)] = \epsilon \langle \phi, \xi \rangle.$$

The squared-norm of a feature vector is its variance; specifically, for a feature function h satisfying (4.8) so $\xi \in \mathcal{I}^{\mathcal{Z}}(P_0)$,

$$\mathbb{E}_{P_0}\left[h(Z)^2\right] = \|\xi\|^2. \tag{4.15}$$

However, it is natural to interpret the squared-norm of an information vector in terms of KL divergence[7] with respect to P_0, which follows as a special case of the following more general lemma. A proof is provided in Appendix B.4.

Lemma 4.5. For a given $P_0 \in \operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$ and $\epsilon > 0$, let $P_1, P_2 \in \mathcal{N}_{\epsilon}^{\mathcal{Z}}(P_0)$ be arbitrary, and let ϕ_1 and ϕ_2 denote the corresponding information vectors, respectively. Then

$$D(P_1 \| P_2) \triangleq \sum_{z \in \mathcal{Z}} P_1(z) \log \frac{P_1(z)}{P_2(z)} = \frac{\epsilon^2}{2} \|\phi_1 - \phi_2\|^2 + o(\epsilon^2), \quad \epsilon \to 0. \tag{4.16}$$

[7] For $P, Q \in \mathcal{P}^{\mathcal{Z}}$, we use the usual

$$D(P\|Q) = \sum_{z \in \mathcal{Z}} P(z) \log \frac{P(z)}{Q(z)}$$

to denote KL divergence of Q from P.

Moreover, for $P \in \mathcal{N}_\epsilon^{\mathcal{Z}}(P_0)$ and with ϕ denoting its information vector,[8] we have as a special case

$$D(P \| P_0) = \frac{\epsilon^2}{2} \|\phi\|^2 + o(\epsilon^2), \quad \epsilon \to 0, \tag{4.17}$$

since $\phi_0 \equiv 0$ is the information vector associated with P_0.

Note that as (4.16) reflects, divergence is locally symmetric in P_1 and P_2—specifically, to first order in ϵ^2.

Additionally, in (4.16) we recognize $\phi_1 - \phi_2$ as, to first order, the information vector associated with the log-likelihood ratio feature function

$$h_{\mathrm{LLR}}(z) \triangleq \frac{1}{\epsilon}\left(\log \frac{P_1(z)}{P_2(z)} - \mathbb{E}_{P_0}\left[\log \frac{P_1(z)}{P_2(z)}\right]\right). \tag{4.18a}$$

In particular, since

$$\log \frac{P_1(z)}{P_2(z)} = \log \frac{P_1(z)}{P_0(z)} - \log \frac{P_2(z)}{P_0(z)},$$

it follows from the first part of Corollary 4.3 that (4.18a) has feature vector

$$\xi_{\mathrm{LLR}}(z) = \phi_1(z) - \phi_2(z) + o(1), \quad \epsilon \to 0, \quad z \in \mathcal{Z}. \tag{4.18b}$$

It is also important to appreciate that (4.16) is invariant to the choice of reference distribution within the neighborhood, which is an immediate consequence of the following result, verified in Appendix B.5.

Lemma 4.6. For a given $P_0 \in \operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$ and $\epsilon > 0$ sufficiently small that $\mathcal{N}_\epsilon^{\mathcal{Z}}(P_0) \subset \operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$, let $P_1, P_2 \in \mathcal{N}_\epsilon^{\mathcal{Z}}(P_0)$ be arbitrary, and let ϕ_1 and ϕ_2 be the corresponding information vectors. Then for any $\tilde{P}_0 \in \mathcal{N}_\epsilon^{\mathcal{Z}}(P_0)$, the information vectors

$$\tilde{\phi}_1(z) \triangleq \frac{P_1(z) - \tilde{P}_0(z)}{\epsilon\sqrt{\tilde{P}_0(z)}} \quad \text{and} \quad \tilde{\phi}_2(z) \triangleq \frac{P_2(z) - \tilde{P}_0(z)}{\epsilon\sqrt{\tilde{P}_0(z)}}$$

satisfy, for each $z \in \mathcal{Z}$,

$$\tilde{\phi}_1(z) - \tilde{\phi}_2(z) = \big(\phi_1(z) - \phi_2(z)\big)\big(1 + o(1)\big), \quad \epsilon \to 0. \tag{4.19}$$

[8] Note, for comparison, that $\chi^2(P\|P_0) = \epsilon^2 \|\phi\|^2$.

4.2 Weakly Dependent Variables

An instance of local analysis corresponds to weak dependence between variables, a concept we formally define as follows.

Definition 4.7 (ϵ-Dependence)**.** Let Z and W be defined over alphabets $\mathcal{Z}$ and $\mathcal{W}$, respectively, and distributed according to $P_{Z,W} \in \mathcal{P}^{\mathcal{Z}\times\mathcal{W}}$, where $\mathcal{P}^{\mathcal{Z}\times\mathcal{W}}$ is the (usual) restriction of the simplex to distributions with positive marginals. Then Z and W are ϵ-dependent if there exists an $\epsilon > 0$ such that[9]

$$P_{Z,W} \in \mathcal{N}_{\epsilon}^{\mathcal{Z}\times\mathcal{W}}(P_Z P_W), \tag{4.21}$$

where P_Z and P_W are the marginal distributions associated with $P_{Z,W}$.

As related notions of ϵ-dependence, we can replace (4.21) with one of

$$P_{W|Z}(\cdot|z) \in \mathcal{N}_{\epsilon}^{\mathcal{W}}(P_W), \quad \text{all } z \in \mathcal{Z} \tag{4.22}$$

$$P_{Z|W}(\cdot|w) \in \mathcal{N}_{\epsilon}^{\mathcal{Z}}(P_Z), \quad \text{all } w \in \mathcal{W}. \tag{4.23}$$

These notions are all locally equivalent, which the following lemma establishes; a proof is provided in Appendix B.6.

Lemma 4.8. Let Z and W be defined over alphabets $\mathcal{Z}$ and $\mathcal{W}$, respectively, and distributed according to $P_{Z,W} \in \mathcal{P}^{\mathcal{Z}\times\mathcal{W}}$, where $\mathcal{P}^{\mathcal{Z}\times\mathcal{W}}$ is the (usual) restriction of the simplex to distributions with positive marginals. When

$$\liminf_{\epsilon\to 0} P_Z(z) > 0, \quad \text{all } z \in \mathcal{Z} \tag{4.24a}$$

$$\liminf_{\epsilon\to 0} P_W(w) > 0, \quad \text{all } w \in \mathcal{W}, \tag{4.24b}$$

[9] Note that the condition (4.21) is equivalent to

$$\chi^2(P_{Z,W} \| P_Z P_W) \leq \epsilon^2, \tag{4.20}$$

the left-hand side of which defines mutual information with respect to χ^2-divergence. This mutual information was historically referred to as "mean-square contingency," a concept introduced by Pearson [112], [223]. Note that $\chi^2(P_{X,Y} \| P_X P_Y) = \sigma_1^2 + \cdots + \sigma_{K-1}^2$. As such, the k-correlation quantity used in [47, Section II] represents an approximation of classical mean-square contingency by truncation.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

the following statements are equivalent as $\epsilon \to 0$:

$$P_{Z,W} \in \mathcal{N}^{\mathcal{Z}\times\mathcal{W}}_{\mathcal{O}(\epsilon)}(P_Z P_W) \tag{4.25a}$$

$$P_{W|Z}(\cdot|z) \in \mathcal{N}^{\mathcal{W}}_{\mathcal{O}(\epsilon)}(P_W), \quad \text{all } z \in \mathcal{Z} \tag{4.25b}$$

$$P_{Z|W}(\cdot|w) \in \mathcal{N}^{\mathcal{Z}}_{\mathcal{O}(\epsilon)}(P_Z), \quad \text{all } w \in \mathcal{W}. \tag{4.25c}$$

Accordingly, any of (4.25) can be used to characterize $\mathcal{O}(\epsilon)$-dependence. In the sequel, except where the distinction is needed, with some abuse of terminology we will use ϵ-dependence and $\mathcal{O}(\epsilon)$-dependence interchangeably.

A further asymptotic equivalence between the notion of ϵ-dependence based on χ^2-divergence and one based on KL divergence is established by the following lemma, whose proof is provided in Appendix B.7.

Lemma 4.9. Under the hypotheses of Lemma 4.8,

$$I(Z;W) = \mathcal{O}(\epsilon^2) \quad \text{if and only if} \quad P_{Z,W} \in \mathcal{N}^{\mathcal{Z}\times\mathcal{W}}_{\mathcal{O}(\epsilon)}(P_Z P_W). \tag{4.26}$$

Finally, for completeness, we have the following asymptotic equivalences among notions of ϵ-dependence based on KL divergence, analogous to Lemma 4.8. A proof is provided in Appendix B.8.

Lemma 4.10. Under the hypotheses of Lemma 4.8, the following statements are equivalent as $\epsilon \to 0$:

$$I(Z;W) = \mathcal{O}(\epsilon^2) \tag{4.27a}$$

$$D(P_{W|Z}(\cdot|z) \| P_W) = \mathcal{O}(\epsilon^2), \quad \text{all } z \in \mathcal{Z} \tag{4.27b}$$

$$D(P_{Z|W}(\cdot|w) \| P_Z) = \mathcal{O}(\epsilon^2), \quad \text{all } w \in \mathcal{W}. \tag{4.27c}$$

We will exploit the various equivalences (4.25)–(4.27) in our analysis.

4.3 The Modal Decomposition of Mutual Information

The modal decomposition (2.15) of $P_{X,Y}$ leads directly to a corresponding decomposition of mutual information when X and Y are weakly dependent. In particular, we have the following result.

Lemma 4.11. Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ with $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ be ϵ-dependent random variables, and let $\tilde{\mathbf{B}}$ denote the associated CDM. Then

$$I(X;Y) = \frac{1}{2}\|\tilde{\mathbf{B}}\|_{\mathrm{F}}^2 + o(\epsilon^2) = \frac{1}{2}\sum_{i=1}^{K-1} \sigma_i^2 + o(\epsilon^2), \tag{4.28}$$

where the summation is $\mathcal{O}(\epsilon^2)$, as $\epsilon \to 0$.

Proof. It suffices to make the choices $P = P_{X,Y}$ and $P_0 = P_X P_Y$ in (4.17) of Lemma 4.5, and recognize that the corresponding information vector—which is convenient to express as a matrix in this case—has elements

$$\phi(x,y) = \frac{P_{X,Y}(x,y) - P_X(x)\,P_Y(y)}{\epsilon\sqrt{P_X(x)}\,\sqrt{P_Y(y)}} = \frac{1}{\epsilon}\tilde{B}(x,y). \tag{4.29}$$

Then, since the Frobenius norm of an information vector in matrix form coincides with its Euclidean norm, and since for any matrix $\mathbf{A}$ whose singular values are $\sigma_1(\mathbf{A}),\dots,\sigma_l(\mathbf{A})$ for some l, we have $\|\mathbf{A}\|_{\mathrm{F}}^2 = \sum_{i=1}^{l}\sigma_i(\mathbf{A})^2$, (4.28) follows. Finally, that the first term on the right-hand side of (4.28) is at most $\mathcal{O}(\epsilon^2)$ follows from applying the constraint (4.3) to the information vector defined via (4.29). ■

A key interpretation of the decomposition (4.28) is as follows. For each $1 \le k \le K-1$, the bivariate function

$$P_{X,Y}^{(k)}(x,y) \triangleq P_X(x)\,P_Y(y)\left(1 + \sum_{i=1}^{k}\sigma_i\, f_i^*(x)\, g_i^*(y)\right) \tag{4.30a}$$

obtained by truncating (2.15) sums to unity and, for all ϵ sufficiently small, is nonnegative for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$, so has the interpretation as a joint distribution for new variables $(X^{(k)}, Y^{(k)})$, i.e., $P_{X,Y}^{(k)} = P_{X^{(k)},Y^{(k)}}$, having the same (original) marginals P_X and P_Y for all such k. Moreover, these new variables have mutual information

$$I(X^{(k)};Y^{(k)}) = \frac{1}{2}\sum_{i=1}^{k}\sigma_i^2 + o(\epsilon^2), \quad \epsilon \to 0. \tag{4.30b}$$

Hence, the kth term in the expansion contributes an increment of $\sigma_k^2/2 + o(\epsilon^2)$ to the mutual information. From this perspective, in the weak-dependence regime the chosen ordering captures the largest proportion of mutual information from the fewest number of terms. Valuable complementary perspectives on these order-k distributions will become apparent later in the development.

4.4 The Local Geometry of Decision Making

In our development, it will be useful to exploit a geometric interpretation of traditional binary hypothesis testing, which we now describe. In particular, suppose we observe m samples $z_1^m = (z_1, \dots, z_m)$ drawn in an independent, identically distributed (i.i.d.) manner from either distribution P_1 or distribution P_2, where $P_1, P_2 \in \mathcal{N}_\epsilon^{\mathcal{Z}}(P_0)$. As in Section 4.1, let ϕ_1 and ϕ_2 denote the associated information vectors.

For this problem, for some $1 \leq k \leq K-1$, consider a sequence of k-dimensional statistics

$$\ell^k = (\ell_1, \dots, \ell_k) \tag{4.31a}$$

with

$$\ell_l = \frac{1}{m} \sum_{j=1}^{m} h_l(z_j), \quad l \in \{1, \dots, k\}, \tag{4.31b}$$

for some feature functions $h^k = (h_1, \dots, h_k)$ with associated feature vectors $\xi^k = (\xi_1, \dots, \xi_k)$.

Without loss of generality we restrict our attention to normalized feature functions such that the statistics

$$h^k(Z) = (h_1(Z), \dots, h_k(Z))$$

are zero mean, unit-variance, and uncorrelated with respect to P_0, i.e.,

$$\mathbb{E}_{P_0}[h_i(Z)] = 0, \quad i \in \{1, \dots, k\} \tag{4.32a}$$

$$\mathbb{E}_{P_0}[h_i(Z)\, h_j(Z)] = \mathbb{1}_{i=j}, \quad i, j \in \{1, \dots, k\}. \tag{4.32b}$$

Indeed, if $h^k(Z)$ had any other mean and (nonsingular) covariance structure, then we could apply an invertible transformation to ℓ^k to generate an equivalent statistic $\tilde{\ell}^k$ with the desired structure.[10] Note

[10] In particular, with $\boldsymbol{\ell}$ denoting the vector representation of ℓ^k, if $\boldsymbol{\ell}$ has mean vector $\boldsymbol{\mu}_{\boldsymbol{\ell}}$ and covariance matrix is $\boldsymbol{\Lambda}_{\boldsymbol{\ell}}$, then $\tilde{\boldsymbol{\ell}} \triangleq \boldsymbol{\Lambda}_{\boldsymbol{\ell}}^{-1/2}(\boldsymbol{\ell} - \boldsymbol{\mu}_{\boldsymbol{\ell}})$, with $\boldsymbol{\Lambda}_{\boldsymbol{\ell}}^{1/2}$ denoting any square root matrix of $\boldsymbol{\Lambda}_{\boldsymbol{\ell}}$, has mean $\boldsymbol{\mu}_{\tilde{\boldsymbol{\ell}}} = \mathbf{0}$ and covariance matrix $\boldsymbol{\Lambda}_{\tilde{\boldsymbol{\ell}}} = \mathbf{I}$ as desired.

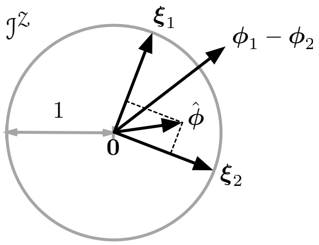
$\mathcal{I}^{\mathcal{Z}}$
$\boldsymbol{\xi}_1$
$\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2$
1
$\hat{\boldsymbol{\phi}}$
$\mathbf{0}$
$\boldsymbol{\xi}_2$

Figure 4.1: The local geometry of decision making for distinguishing i.i.d. samples z^m over alphabet $\mathcal{Z}$ from one of $P_1, P_2 \in \mathcal{P}^{\mathcal{Z}}$ based on a statistic $\ell = (1/m)\sum_{i=1}^{m} h^k(z_i)$ involving feature functions h^k. In information space $\mathcal{I}^{\mathcal{Z}}(P_0)$, where the reference distribution P_0 maps to the origin, ℓ corresponds to the projection of the information vector $\hat{\boldsymbol{\phi}}$ for $\hat{P}$ onto subspace spanned by the feature vectors $\boldsymbol{\xi}^k$ for h^k. The optimum decision rule projects $\hat{\boldsymbol{\phi}}$ directly onto $\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2$, the feature vector associated with the log-likelihood ratio h_{LLR}.

that for feature functions normalized according to (4.32a), the feature vectors lie in information space, i.e.,

$$\xi_i \in \mathcal{I}^{\mathcal{Z}}(P_0), \qquad i \in \{1, \dots, k\},$$

and when the feature functions are further normalized according to (4.32b), the associated feature vectors are orthonormal, i.e.,

$$\langle \xi_i, \xi_j \rangle = \mathbb{1}_{i=j}, \qquad i, j \in \{1, \dots, k\}. \tag{4.33}$$

With $\hat{P}$ denoting the empirical distribution of the data z_1^m, we can express (4.31b) in the form

$$\ell_l = \sum_{z \in \mathcal{Z}} \hat{P}(z)\, h_l(z) = \epsilon \langle \hat{\phi}, \xi_l \rangle$$

where we have used Lemma 4.4 with $P = \hat{P}$, and where

$$\hat{\phi}(z) = \frac{1}{\epsilon} \frac{\hat{P}(z) - P_0(z)}{\sqrt{P_0(z)}}. \tag{4.34}$$

is the *observed* information vector [cf. (4.2)]. The associated geometry is depicted in Figure 4.1.

Our main result is as follows, a proof of which is provided in Appendix B.9.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Lemma 4.12. Given a reference distribution $P_0 \in \operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$ a constant $\epsilon > 0$ and integers m and k, let $z_1, \dots, z_m$ denote i.i.d. samples from one of P_1 or P_2, where $P_1, P_2 \in \mathcal{N}_\epsilon^{\mathcal{Z}}(P_0)$ and both positive. Then the error probability p_e for deciding whether P_1 or P_2 is the generating distribution, based on a statistic ℓ^k of the form (4.31) with normalized feature functions h^k, decays exponentially in m as $m \to \infty$, with (Chernoff) exponent

$$\lim_{m\to\infty} \frac{-\log p_\mathrm{e}}{m} \triangleq E_{h^k} = \sum_{l=1}^{k} E_{h_l}, \tag{4.35a}$$

where

$$E_{h_l} \triangleq \frac{\epsilon^2}{8} \langle \phi_1 - \phi_2, \xi_l \rangle^2 + o(\epsilon^2), \quad \epsilon \to 0. \tag{4.35b}$$

The k-fold local efficiency $\nu(h^k)$ of the rule defined by h^k quantifies the goodness of the exponent (4.35) in Lemma 4.12 relative to the ideal exponent

$$E \triangleq \frac{\epsilon^2}{8} \|\phi_1 - \phi_2\|.$$

Specifically,

$$\nu(h^k) \triangleq \lim_{\epsilon\to 0} \frac{E_{h^k}}{E} = \frac{\sum_{l=1}^{k} \langle \phi_1 - \phi_2, \xi_l \rangle^2}{\|\phi_1 - \phi_2\|^2}. \tag{4.36}$$

It follows from Bessel's inequality that $0 \le \nu(h^k) \le 1$, and from (4.18) that the upper bound is achieved by the choices

$$h_1 = h_\mathrm{LLR}, \quad \text{and} \quad h_i \equiv 0,\ i \in \{2, \dots, k\}, \tag{4.37}$$

i.e., the log-likelihood ratio is an optimum statistic, as expected. In the sequel, we focus on inference scenarios in which such a statistic cannot be used directly.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

5

Universal Feature Characterizations

In this section, we introduce several different notions of feature universality. In turn, using the local analysis of Section 4, we show that these diverse characterizations of universality all yield precisely the same features—those that arise in the modal decomposition of the joint distribution $P_{X,Y}$ as developed in Sections 2 and 3.

The section is structured as follows. We begin by noting that the modal decomposition features characterize a locally exponential family for the conditional distributions. For the remaining characterizations, we introduce latent attribute variables. In Section 5.4 we obtain the modal decomposition features as the solution to a game between system designer and nature, where the system designer must choose features to detect attributes that nature chooses at random after these features are fixed. In Section 5.5, we obtain the same features as the solution to a cooperative game in which the system designer and nature seek the most detectable attributes and locally sufficient statistics for their detection. In Section 5.6, we obtain the same features as the solution to a local symmetric version of Tishby's information bottleneck problem that seeks mutual information-maximizing attributes and the associated locally sufficient statistics. And in Section 5.7, we show that superposi-

The version of record is available at: http://dx.doi.org/10.1561/0100000107

tions of these same features arise as sufficient statistics in the solution to a local version of Wyner's common information, which using variational analysis we show specializes to the nuclear (trace) norm of the CDM. In turn, Section 5.8 develops the Markov structure relating the resulting common information variable to the attributes optimizing the information bottleneck. Finally, for comparison, Section 5.9 shows how these features appear in the characterization of Gács-Körner common information.

5.1 A Preliminary Exponential Family Perspective

As an initial viewpoint, when X and Y are weakly dependent, their conditional distributions are exponential families in which the features in the modal decomposition (2.15) are natural statistics. Specifically, when X and Y are ϵ-dependent according to Definition 4.7, we have, starting from (2.22a),

$$P_{Y|X}(y|x) = P_Y(y) \exp\left\{ \sum_{i=1}^{K-1} \sigma_i \, f_i^*(x) \, g_i^*(y) + o(\epsilon) \right\}, \quad \epsilon \to 0, \tag{5.1}$$

where we have used the Taylor series approximation $\mathrm{e}^{\omega} = 1 + \omega + o(\omega)$ and that the exponent in the first term in (5.1) is $\mathcal{O}(\epsilon)$. We recognize the posterior (5.1) as an exponential family with natural parameters $g_*^{K-1}(y)$ and natural statistics $f_*^{K-1}(x)$. Moreover, by symmetry we have, or equivalently via (2.22b),

$$P_{X|Y}(x|y) = P_X(x) \exp\left\{ \sum_{i=1}^{K-1} \sigma_i \, f_i^*(x) \, g_i^*(y) + o(\epsilon) \right\}, \quad \epsilon \to 0, \tag{5.2}$$

from which we see that for inferences about X from Y, the roles of the features are reversed in the associated posterior: $f_*^{K-1}(x)$ are the natural parameters and $g_*^{K-1}(y)$ are the natural statistics.

Since exponential families with such structure are widely used in discriminative models for learning, we can interpret the (5.1) and (5.2) as indicating universal feature choices. Moreover, dimensionally-reduced families of the form

$$P^{(k)}_{Y|X}(y|x) = P_Y(y) \exp\left\{ \sum_{i=1}^{k} \sigma_i \, f_i^*(x) \, g_i^*(y) + \mathcal{o}(\epsilon) \right\}$$

$$P^{(k)}_{X|Y}(x|y) = P_X(x) \exp\left\{ \sum_{i=1}^{k} \sigma_i \, f_i^*(x) \, g_i^*(y) + \mathcal{o}(\epsilon) \right\}$$

for some $1 \leq k \leq K-1$ represent approximations that maximize the retained mutual information, as per the discussion in Section 4.3 surrounding (4.30). We develop and interpret these posterior distributions further in Section 8. However, there are other senses in which the features in (2.15) are universal, which we develop first, and which require a data model we now introduce.

5.2 Latent Attribute and Statistic Model

In the sequel, we develop universal features from the modal decomposition (2.15) via the introduction of latent (auxiliary) variables. Latent variable models have a long history in facilitating both the interpretation and exploitation of relationships in data. While the original focus was on linear relationships, corresponding to factor analysis as introduced by Spearman [254], the modern view is considerably broader; see, e.g., [33] for a discussion.

As we now describe, our treatment models scenarios in which the inference task involving X and Y is not known in advance through the introduction of latent *attribute* variables whose values we seek to determine. We emphasize at the outset that in this model, we treat $P_{X,Y}$ as known or, equivalently, to have been sufficiently reliably estimated from training samples, a process for which we will later discuss.

We begin by formalizing the notion of an attribute.[1]

Definition 5.1 (ϵ-Attribute)**.** Given $\epsilon > 0$ and $P_Z \in \operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$ for some $\mathcal{Z}$, then W on some alphabet $\mathcal{W}$ with $2 \leq |\mathcal{W}| \leq |\mathcal{Z}|$ and having distribution $P_W \in \operatorname{relint}(\mathcal{P}^{\mathcal{W}})$ is an ϵ-attribute of Z if W is ϵ-dependent on Z, i.e.,

$$P_{Z|W}(\cdot|w) \in \mathcal{N}^{\mathcal{Z}}_{\epsilon}(P_Z), \quad w \in \mathcal{W},$$

[1] More generally, we use *attribute* to refer to an ϵ-attribute in which there is no restriction on ϵ, i.e., it can be arbitrarily large.

$P_{Z|W}(\cdot|w) \notin \mathcal{N}_0^{\mathcal{Z}}(P_Z)$ for all $w \in \mathcal{W}$, and W is conditionally independent of all other variables in the model given Z.

Such attributes are specified by a collection of parameters. In particular, we have the following.

Definition 5.2 (ϵ-Attribute Configuration). Given $\epsilon > 0$ and $P_Z \in \operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$ for some $\mathcal{Z}$, then ϵ-attribute W of Z is characterized by its *configuration*

$$\begin{aligned}\mathcal{C}_\epsilon^{\mathcal{Z}}(P_Z) \triangleq \Big\{ \mathcal{W},\, \{P_W(w),\, w \in \mathcal{W}\},\, &\{P_{Z|W}(\cdot|w),\, w \in \mathcal{W}\} \colon \\ &P_{Z|W}(\cdot|w) \in \mathcal{N}_\epsilon^{\mathcal{Z}}(P_Z),\ w \in \mathcal{W}, \\ &\sum_{w \in \mathcal{W}} P_W(w)\, P_{Z|W}(z|w) = P_Z(z),\ z \in \mathcal{Z} \Big\}, \end{aligned} \tag{5.3}$$

which can be equivalently expressed in the form

$$\begin{aligned}\mathcal{C}_\epsilon^{\mathcal{Z}}(P_Z) = \Big\{ \mathcal{W},\, \{P_W(w),\, w \in \mathcal{W}\},\, &\{\boldsymbol{\phi}_w^{Z|W},\, w \in \mathcal{W}\} \colon \\ &\boldsymbol{\phi}_w^{Z|W} \in \mathcal{I}^{\mathcal{Z}}(P_Z),\ w \in \mathcal{W}, \\ &\sum_{w \in \mathcal{W}} P_W(w)\, \phi_w^{Z|W}(z) = 0,\ z \in \mathcal{Z} \Big\}, \end{aligned} \tag{5.4}$$

where

$$\phi_w^{Z|W}(z) \triangleq \frac{P_{Z|W}(z|w) - P_Z(z)}{\epsilon \sqrt{P_Z(z)}}, \quad z \in \mathcal{Z},\ w \in \mathcal{W} \tag{5.5}$$

define the information vectors associated with the ϵ-attribute W.

In Definition 5.2, we note that the equivalent form (5.4) is a consequence of the fact the constraint

$$\sum_{w \in \mathcal{W}} P_W(w)\, P_{Z|W}(z|w) = P_Z(z),$$

implies the information vectors must satisfy

$$\sum_{w \in \mathcal{W}} P_W(w)\, \phi_w^{Z|W}(z) = 0. \tag{5.6}$$

In the context of a given model $P_{X,Y}$, the attribute variables U and V for X and Y, respectively, are characterized by the Markov structure

$$U \leftrightarrow X \leftrightarrow Y \leftrightarrow V. \tag{5.7}$$

More generally, in the case of m samples drawn from $P_{X,Y}$, our model has the Markov structure

$$U \leftrightarrow X^m \leftrightarrow Y^m \leftrightarrow V \tag{5.8a}$$

with the conditional independence and memoryless structure

$$P_{X^m|U}(x^m|u) = \prod_{i=1}^{m} P_{X|U}(x_i|u) \tag{5.8b}$$

$$P_{Y^m|V}(y^m|v) = \prod_{i=1}^{m} P_{Y|V}(y_i|v) \tag{5.8c}$$

$$P_{X^m,Y^m}(x^m, y^m) = \prod_{i=1}^{m} P_{X,Y}(x_i, y_i). \tag{5.8d}$$

The attributes U and V can be interpreted as instances of *class* variables, whose values correspond to different aspects of X and Y, respectively.

Our development focuses on the case where U and V depend only weakly on X and Y. Specifically, we consider the ϵ-dependence

$$\begin{aligned} P_{X|U}(\cdot|u) &\in \mathcal{N}_\epsilon^{\mathcal{X}}(P_X), \quad \text{for all } u \in \mathcal{U} \\ P_{Y|V}(\cdot|v) &\in \mathcal{N}_\epsilon^{\mathcal{Y}}(P_Y), \quad \text{for all } v \in \mathcal{V}. \end{aligned}$$

Via Lemma 4.5, ϵ-dependence can be equivalently expressed as the condition

$$D(P_{X|U}(\cdot|u)\|P_X) \leq \frac{\epsilon^2}{2}(1 + o(1)), \quad \epsilon \to 0, \quad \text{for all } u \in \mathcal{U} \tag{5.9a}$$

$$D(P_{Y|V}(\cdot|v)\|P_Y) \leq \frac{\epsilon^2}{2}(1 + o(1)), \quad \epsilon \to 0, \quad \text{for all } v \in \mathcal{V}, \tag{5.9b}$$

which, in turn, of course implies

$$I(X;U) = \sum_{u\in\mathcal{U}} P_U(u)\, D(P_{X|U}(\cdot|u)\|P_X) \leq \frac{\epsilon^2}{2}(1 + o(1)), \quad \epsilon \to 0 \tag{5.10a}$$

$$I(Y;V) = \sum_{v\in\mathcal{V}} P_V(v)\, D(P_{Y|V}(\cdot|v)\|P_Y) \leq \frac{\epsilon^2}{2}(1+o(1)), \quad \epsilon \to 0 \tag{5.10b}$$

For inferences about attributes U and V, we will generally consider statistics of the form

$$S^k \triangleq \frac{1}{m}\sum_{i=1}^{m} f^k(X_i) \quad \text{and} \quad T^k \triangleq \frac{1}{m}\sum_{i=1}^{m} g^k(Y_i), \tag{5.11}$$

for some $k \in \{1,\dots,K-1\}$ and feature choices $f^k\colon \mathcal{X} \to \mathbb{R}^k$ and $g^k\colon \mathcal{Y} \to \mathbb{R}^k$. Moreover, in accordance with our earlier discussion, without loss of generality we restrict our attention to normalized features, i.e. $(f^k, g^k) \in \mathcal{F}_k \times \mathcal{G}_k$ with $\mathcal{F}_k$ and $\mathcal{G}_k$ as defined in (3.6c) and (3.6d), respectively. As we will develop, the particular choices

$$S_*^k \triangleq \frac{1}{m}\sum_{i=1}^{m} f_*^k(X_i) \quad \text{and} \quad T_*^k \triangleq \frac{1}{m}\sum_{i=1}^{m} g_*^k(Y_i), \tag{5.12}$$

with f_*^k and g_*^k as defined in (3.5) play a special role.

Finally, we will sometimes extend the model (5.8) to the case of multidimensional U and V with special structure, which we term *multi-attributes.*[2]

Definition 5.3 (ϵ-Multi-Attribute)**.** Given $\epsilon > 0$, l, and Z over some alphabet $\mathcal{Z}$, then an attribute W of Z with configuration $\mathcal{C}_{l\epsilon}^{\mathcal{Z}}(P_Z)$ over alphabet $\mathcal{W}$ is an l-dimensional ϵ-multi-attribute W^l over alphabet $\mathcal{W} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_l$ if the variables W^l are:
1) such that

$$|\mathcal{W}_i| \geq 2 \quad \text{and} \quad P_{W_i} \in \operatorname{relint}(\mathcal{P}^{\mathcal{W}_i}), \quad i \in \{1,\dots,l\};$$

2) ϵ-dependent on Z, i.e.,

$$\left.\begin{aligned} &P_{Z|W_i}(\cdot|w_i) \in \mathcal{N}_\epsilon^{\mathcal{Z}}(P_Z), \\ &P_{Z|W_i}(\cdot|w_i) \notin \mathcal{N}_0^{\mathcal{Z}}(P_Z), \end{aligned}\right. \quad \text{all } w_i \in \mathcal{W}_i \text{ and } i \in \{1,\dots,l\};$$

[2] As in the case of attributes, we more generally use *multi-attribute* to refer to an ϵ-multi-attribute in which there is no restriction on ϵ, i.e., it can be arbitrarily large. Key to a multi-attribute is the simultaneous marginal and conditional independence structure. As an example, if the automobile in a digital image Z has color W_1 and motor type W_2, then one might reasonably model (W_1, W_2) as a multi-attribute.

3) conditionally independent given Z, i.e.,

$$P_{W^l|Z}(w^l|z) = \prod_{i=1}^{l} P_{W_i|Z}(w_i|z), \quad \text{all } w^l \in \mathcal{W},\, z \in \mathcal{Z};$$

and
4) (marginally) independent, i.e.,

$$P_{W^l}(w^l) = \prod_{i=1}^{l} P_{W_i}(w_i), \quad \text{all } w^l \in \mathcal{W}.$$

We use $\mathcal{C}_\epsilon^{\mathcal{Z},l}(P_Z)$ to denote the configuration of such a ϵ-multi-attribute variable.

Multi-attribute variables have the following key orthogonality property. A proof is provided in Appendix C.1.

Lemma 5.4. For some $\epsilon > 0$ and integer $l \geq 1$, let W^l be an ϵ-multi-attribute of $Z \in \mathcal{Z}$ over alphabet $\mathcal{W} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_l$. Then with the information vector notation

$$\phi_{w_i}^{Z|W_i}(z) \triangleq \frac{P_{Z|W_i}(z|w_i) - P_Z(z)}{\epsilon\sqrt{P_Z(z)}}, \quad i = 1, \dots, l, \tag{5.13}$$

we have, for $i, j \in \{1, \dots, l\}$,

$$\langle \phi_{w_i}^{Z|W_i}, \phi_{w_j}^{Z|W_j} \rangle = 0, \ \text{for all } i \neq j,\, w_i \in \mathcal{W}_i \text{ and } w_j \in \mathcal{W}_j.$$

In addition, multi-attributes admit the following information vector decomposition.[3]

Lemma 5.5. For some $\epsilon > 0$ and integer $l \geq 1$, let W^l be an ϵ-multi-attribute of $Z \in \mathcal{Z}$ over alphabet $\mathcal{W} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_l$. Then with the information vector notation

$$\phi_{w^l}^{Z|W^l}(z) \triangleq \frac{P_{Z|W^l}(z|w^l) - P_Z(z)}{\epsilon\sqrt{P_Z(z)}}, \tag{5.14}$$

and $\phi_{w_i}^{Z|W_i}$ as defined in (5.13), we have

$$\boldsymbol{\phi}_{w^l}^{Z|W^l} = \sum_{i=1}^{l} \boldsymbol{\phi}_{w_i}^{Z|W_i} + o(1), \quad \epsilon \to 0. \tag{5.15}$$

[3] Note that this decomposition implies that an ϵ-multi-attribute is an $l\epsilon$-attribute.

A proof is provided in Appendix C.2, and exploits the following simple approximation.

Fact 5.6. For any integer $l \geq 1$ and constants ϵ and $a_1, \dots, a_l$, then

$$\prod_{i=1}^{l} (1 + \epsilon a_i) = 1 + \epsilon \sum_{i=1}^{l} a_i + o(\epsilon), \quad \epsilon \to 0. \tag{5.16}$$

For multi-attributes U^k and V^k of X and Y, respectively, we use

$$\phi_{u_i}^{X|U_i}(x) \triangleq \frac{P_{X|U_i}(x|u_i) - P_X(x)}{\epsilon\sqrt{P_X(x)}} \tag{5.17a}$$

$$\phi_{v_i}^{Y|V_i}(y) \triangleq \frac{P_{Y|V_i}(y|v_i) - P_Y(y)}{\epsilon\sqrt{P_Y(y)}} \tag{5.17b}$$

to denote the information vectors corresponding to $P_{X|U_i}(\cdot|u_i)$ and $P_{Y|V_i}(\cdot|v_i)$, respectively.

Note that for the extended Markov model (5.8), orthogonality for multi-attribute U^k of X^m further implies that the U^k are conditionally independent given X_j, each $j \in \{1, \dots, m\}$, and X^m are conditionally independent given U_i, each $i \in \{1, \dots, k\}$, i.e.,

$$P_{U^k|X_j}(u^k|x_j) = \prod_{i=1}^{k} P_{U_i|X_j}(u_i|x_j) \tag{5.18a}$$

$$P_{X^m|U_i}(x^m|u_i) = \prod_{j=1}^{m} P_{X_j|U_i}(x_j|u_i), \tag{5.18b}$$

and the orthogonality for multi-attribute V^k of Y^m implies V^k are conditionally independent given Y_j, each $j \in \{1, \dots, m\}$, and Y^m are conditionally independent given V_i, each $i \in \{1, \dots, k\}$, i.e.,

$$P_{V^k|Y_j}(v^k|y_j) = \prod_{i=1}^{k} P_{V_i|Y_j}(v_i|y_j) \tag{5.19a}$$

$$P_{Y^m|V_i}(y^m|v_i) = \prod_{j=1}^{m} P_{Y_j|V_i}(y_j|v_i). \tag{5.19b}$$

5.3 Induced Local Geometries of Attribute Variables

We now express the relationships between U, V and X, Y geometrically. In particular, we show that the local geometry of $P_{X|U}(\cdot|u)$ in the simplex $\mathcal{P}^{\mathcal{X}}$ induces a corresponding local geometry for $P_{Y|U}(\cdot|u)$ in the simplex $\mathcal{P}^{\mathcal{Y}}$ via the operator B, and, likewise, the local geometry of $P_{Y|V}(\cdot|v)$ in the simplex $\mathcal{P}^{\mathcal{Y}}$ induces a corresponding local geometry for $P_{X|V}(\cdot|v)$ in the simplex $\mathcal{P}^{\mathcal{X}}$ via the adjoint.

Indeed, the Markov relation $U \leftrightarrow X \leftrightarrow Y$ implies

$$P_Y(y) = \sum_{x\in\mathcal{X}} P_{Y|X}(y|x)\, P_X(x)$$
$$P_{Y|U}(y|u) = \sum_{x\in\mathcal{X}} P_{Y|X}(y|x)\, P_{X|U}(x|u),$$

from which we conclude that a neighborhood of P_X in the simplex $\mathcal{P}^{\mathcal{X}}$ maps to a neighborhood of P_Y in the simplex $\mathcal{P}^{\mathcal{Y}}$. In particular, with P_X and P_Y as the reference distributions in $\mathcal{P}^{\mathcal{X}}$ and $\mathcal{P}^{\mathcal{Y}}$, respectively, the information vectors

$$\phi_u^{X|U}(x) = \frac{P_{X|U}(x|u) - P_X(x)}{\epsilon\sqrt{P_X(x)}} \tag{5.20a}$$

$$\phi_u^{Y|U}(y) = \frac{P_{Y|U}(y|u) - P_Y(y)}{\epsilon\sqrt{P_Y(y)}} \tag{5.20b}$$

associated with the distributions $P_{X|U}(\cdot|u)$ and $P_{Y|U}(\cdot|u)$, respectively, satisfy

$$\phi_u^{Y|U}(y) = \frac{1}{\sqrt{P_Y(y)}} \sum_{x\in\mathcal{X}} P_{Y|X}(y|x)\sqrt{P_X(x)}\,\phi_u^{X|U}(x). \tag{5.21}$$

With $\boldsymbol{\phi}_u^{X|U}$ and $\boldsymbol{\phi}_u^{Y|U}$ denoting the associated column vectors, using (2.10) we can equivalently express (5.21) in the matrix form

$$\boldsymbol{\phi}_u^{Y|U} = \mathbf{B}\,\boldsymbol{\phi}_u^{X|U}. \tag{5.22}$$

Evidently, $\mathbf{B}$ maps a local divergence sphere in $\mathcal{P}^{\mathcal{X}}$ to a local divergence ellipsoid in $\mathcal{P}^{\mathcal{Y}}$ whose principal axes correspond to the left singular vectors of $\mathbf{B}$, as Figure 5.1 depicts.

Analogously, the Markov relation $X \leftrightarrow Y \leftrightarrow V$ implies

$$P_X(x) = \sum_{y\in\mathcal{Y}} P_{X|Y}(x|y)\, P_Y(y)$$
$$P_{X|V}(x|v) = \sum_{y\in\mathcal{Y}} P_{X|Y}(x|y)\, P_{Y|V}(y|v),$$

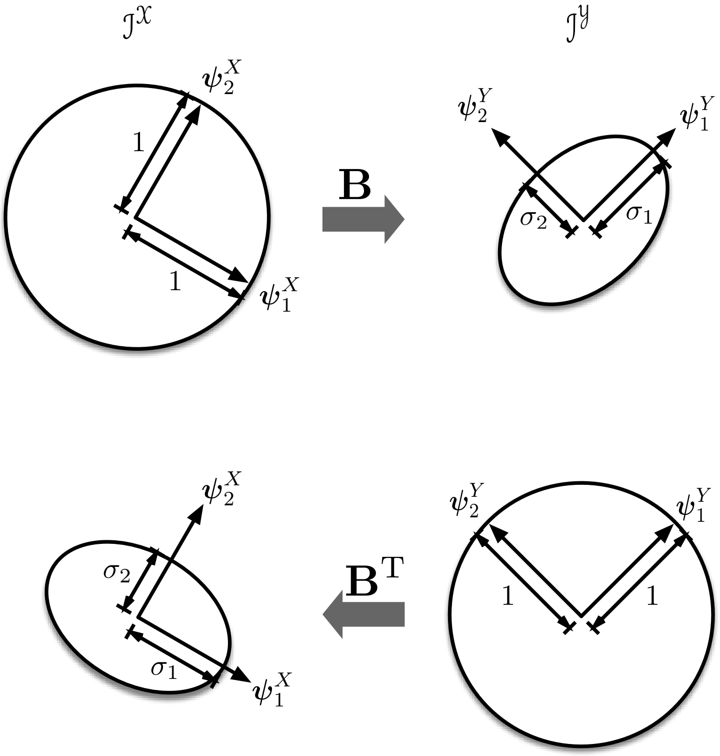

Figure 5.1: The information geometry associated with the DTM $\mathbf{B}$. For $i = 1, \dots, K-1$, the unit information vector ψ_i^X in $\mathcal{I}^{\mathcal{X}}$ maps via $\mathbf{B}$ to the shorter information vector $\sigma_i \psi_i^Y$ in $\mathcal{I}^{\mathcal{Y}}$, and the unit information vector ψ_i^Y in $\mathcal{I}^{\mathcal{Y}}$ maps via $\mathbf{B}^{\mathrm{T}}$ to the shorter information vector $\sigma_i \psi_i^X$ in $\mathcal{I}^{\mathcal{X}}$.

from which we conclude that a neighborhood of P_Y in the simplex $\mathcal{P}^{\mathcal{Y}}$ maps to a neighborhood of P_X in the simplex $\mathcal{P}^{\mathcal{X}}$. In particular, with the same reference distributions, and using (2.11), we obtain that the information vectors

$$\phi_v^{Y|V}(y) = \frac{P_{Y|V}(y|v) - P_Y(y)}{\epsilon\sqrt{P_Y(y)}} \tag{5.23a}$$

$$\phi_v^{X|V}(x) = \frac{P_{X|V}(x|v) - P_X(x)}{\epsilon\sqrt{P_X(x)}} \tag{5.23b}$$

associated with the distributions $P_{Y|V}(\cdot|v)$ and $P_{X|V}(\cdot|v)$ are related according to

$$\boldsymbol{\phi}_v^{X|V} = \mathbf{B}^{\mathrm{T}}\boldsymbol{\phi}_v^{Y|V}, \tag{5.24}$$

where $\boldsymbol{\phi}_v^{Y|V}$ and $\boldsymbol{\phi}_v^{X|V}$ denote the corresponding column vectors. In this case, as Figure 5.1 also depicts, $\mathbf{B}^{\mathrm{T}}$ maps a local divergence sphere in $\mathcal{P}^{\mathcal{Y}}$ to a local divergence ellipsoid in $\mathcal{P}^{\mathcal{X}}$ whose principal axes correspond to the right singular vectors of $\mathbf{B}$.

Finally, we note that there are constraints on the distributions governing the attributes U and V. In particular, we have [cf. (5.6)]

$$\sum_{u\in\mathcal{U}} P_U(u)\,\phi_u^{X|U}(x) = 0, \quad x \in \mathcal{X} \tag{5.25a}$$

$$\sum_{v\in\mathcal{V}} P_V(v)\,\phi_v^{Y|V}(y) = 0, \quad y \in \mathcal{Y}. \tag{5.25b}$$

5.4 Minimum Error Probability Universal Features

In this section, we model the universal feature selection problem as the following game between a system designer and nature. First, nature chooses the distribution for latent attribute variables (U, V) in the Markov chain (5.7) at random. Next, before nature reveals its chosen distributions, the system designer chooses feature functions f^k and g^k knowing $P_{X,Y}$ and the probability law according to which nature chooses its distribution. Finally, after revealing its chosen distributions, the system designer implements a test for determining (U, V) with minimum error probability from statistics formed from its chosen features applied to samples of (X, Y). The details are as follows.

Let $\mathcal{C}^{\mathcal{X}}_{\epsilon}(P_X)$ and $\mathcal{C}^{\mathcal{Y}}_{\epsilon}(P_Y)$ denote configurations for attributes U and V, respectively, in the sense of Definition 5.2, i.e.,

$$\begin{aligned}\mathcal{C}^{\mathcal{X}}_{\epsilon}(P_X) = \Bigl\{&\mathcal{U},\, \{P_U(u),\, u \in \mathcal{U}\},\, \{P_{X|U}(\cdot|u),\, u \in \mathcal{U}\}\colon\\ &\qquad P_{X|U}(\cdot|u) \in \mathcal{N}^{\mathcal{X}}_{\epsilon}(P_X),\;\; u \in \mathcal{U},\\ &\qquad \sum_{u\in\mathcal{U}} P_U(u)\, P_{X|U}(x|u) = P_X(x),\;\; x \in \mathcal{X}\Bigr\} \qquad (5.26a)\\ = \Bigl\{&\mathcal{U},\, \{P_U(u),\, u \in \mathcal{U}\},\, \{\boldsymbol{\phi}^{X|U}_u,\, u \in \mathcal{U}\}\colon\\ &\qquad \boldsymbol{\phi}^{X|U}_u \in \mathcal{J}^{\mathcal{X}},\;\; u \in \mathcal{U},\\ &\qquad\qquad \sum_{u\in\mathcal{U}} P_U(u)\, \phi^{X|U}_u(x) = 0,\;\; x \in \mathcal{X}\Bigr\}\end{aligned}$$

and

$$\begin{aligned}\mathcal{C}^{\mathcal{Y}}_{\epsilon}(P_Y) = \Bigl\{&\mathcal{V},\, \{P_V(v),\, v \in \mathcal{V}\},\, \{P_{Y|V}(\cdot|v),\, v \in \mathcal{V}\}\colon\\ &\qquad P_{Y|V}(\cdot|v) \in \mathcal{N}^{\mathcal{Y}}_{\epsilon}(P_Y),\;\; v \in \mathcal{V},\\ &\qquad \sum_{v\in\mathcal{V}} P_V(v)\, P_{Y|V}(y|v) = P_Y(y),\;\; y \in \mathcal{Y}\Bigr\}, \qquad (5.26b)\\ = \Bigl\{&\mathcal{V},\, \{P_V(v),\, v \in \mathcal{V}\},\, \{\boldsymbol{\phi}^{Y|V}_v,\, v \in \mathcal{V}\}\colon\\ &\qquad \boldsymbol{\phi}^{Y|V}_v \in \mathcal{J}^{\mathcal{Y}},\;\; v \in \mathcal{V},\\ &\qquad\qquad \sum_{v\in\mathcal{V}} P_V(v)\, \phi^{Y|V}_v(y) = 0,\;\; y \in \mathcal{Y}\Bigr\}.\end{aligned}$$

In choosing a configuration pair, nature uses a probability law in which the ensemble for each attribute is characterized by rotational invariance, our definition for which relies on the following concept of spherical symmetry [58], [74].

Definition 5.7 (Spherical Symmetry). A $k_1 \times k_2$ random matrix $\mathbf{Z}$ is spherically symmetric if for any orthogonal $k_1 \times k_1$ and $k_2 \times k_2$ matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$, respectively, we have

$$\mathbf{Z} \overset{\mathrm{d}}{=} \mathbf{Q}_1^{\mathrm{T}} \mathbf{Z}\, \mathbf{Q}_2, \tag{5.27}$$

where $\overset{\mathrm{d}}{=}$ denotes equality in distribution.

Moreover, the following consequence of spherical symmetry is useful in our analysis; a proof is provided in Appendix C.3.[4]

Lemma 5.8. Let $\mathbf{Z}$ be a $k_1 \times k_2$ spherically symmetric random matrix. Then if $\mathbf{A}_1$ and $\mathbf{A}_2$ are any fixed matrices of compatible dimensions, then

$$\mathbb{E}\Big[\big\|\mathbf{A}_1^{\mathrm{T}}\mathbf{Z}\mathbf{A}_2\big\|_{\mathrm{F}}^2\Big] = \frac{1}{k_1 k_2}\,\|\mathbf{A}_1\|_{\mathrm{F}}^2\,\|\mathbf{A}_2\|_{\mathrm{F}}^2\,\mathbb{E}\Big[\|\mathbf{Z}\|_{\mathrm{F}}^2\Big]. \tag{5.28}$$

Our ensemble of interest is defined in terms of spherical symmetry as follows.

Definition 5.9 (Rotation Invariant Ensemble). Given $\epsilon > 0$, a rotationally invariant ensemble (RIE) for an ϵ-attribute W of a variable Z is a collection of attribute configurations of the form (5.4), together with a probability measure over the collection such that $\mathbf{\Phi}^{Z|W}$, the $|\mathcal{Z}| \times |\mathcal{W}|$ matrix whose columns are $\boldsymbol{\phi}_w^{Z|W}$, $w \in \mathcal{W}$,[5] is spherically symmetric.

In what follows, we denote the error probability in decisions for U and V based on S^k, respectively, via

$$p_{\mathrm{e}}^{U|S}(\mathcal{C}_\epsilon^{\mathcal{X}}(P_X), f^k) \quad \text{and} \quad p_{\mathrm{e}}^{V|S}(\mathcal{C}_\epsilon^{\mathcal{Y}}(P_Y), f^k), \tag{5.30a}$$

and those for the decisions based on T^k via, respectively,

$$p_{\mathrm{e}}^{U|T}(\mathcal{C}_\epsilon^{\mathcal{X}}(P_X), g^k) \quad \text{and} \quad p_{\mathrm{e}}^{V|T}(\mathcal{C}_\epsilon^{\mathcal{Y}}(P_Y), g^k), \tag{5.30b}$$

[4] The proof makes use of the notation $\mathbf{e}_i$ for the (elementary) vector whose ith element is 1, and all other elements are 0, which we will more generally find convenient in our analysis.

[5] With some abuse of terminology and notation, we will say the wth column of $\mathbf{\Phi}^{Z|W}$ is $\mathbf{\Phi}^{Z|W}\mathbf{e}_w = \boldsymbol{\phi}_w^{Z|W}$, to avoid cumbersome exposition. More precisely, given an (arbitrary) bijective function $\mathbb{I}_{\mathcal{W}} \colon \mathcal{W} \to \{1, \dots, |\mathcal{W}|\}$ with inverse $\mathbb{I}_{\mathcal{W}}^{-1}$,

$$\mathbf{\Phi}^{Z|W} \triangleq \begin{bmatrix} \boldsymbol{\phi}_{\mathbb{I}_{\mathcal{W}}^{-1}(1)}^{Z|W} & \cdots & \boldsymbol{\phi}_{\mathbb{I}_{\mathcal{W}}^{-1}(|\mathcal{W}|)}^{Z|W} \end{bmatrix}, \tag{5.29}$$

i.e., the $\mathbb{I}_{\mathcal{W}}(w)$th column of $\mathbf{\Phi}^{Z|W}$ is $\mathbf{\Phi}^{Z|W}\mathbf{e}_{\mathbb{I}_{\mathcal{W}}(w)} = \boldsymbol{\phi}_w^{Z|W}$.

where S^k and T^k are as defined in (5.11) for feature choices $f^k\colon \mathcal{X} \to \mathbb{R}^k$ and $g^k\colon \mathcal{Y} \to \mathbb{R}^k$. In turn, we define the error exponents

$$\bar{E}^{U|S}(f^k) \triangleq \lim_{m\to\infty} -\frac{\mathbb{E}_{\mathrm{RIE}}\left[\log p_{\mathrm{e}}^{U|S}(\mathcal{C}_\epsilon^{\mathcal{X}}(P_X), f^k)\right]}{m} \tag{5.31a}$$

$$\bar{E}^{V|S}(f^k) \triangleq \lim_{m\to\infty} -\frac{\mathbb{E}_{\mathrm{RIE}}\left[\log p_{\mathrm{e}}^{V|S}(\mathcal{C}_\epsilon^{\mathcal{Y}}(P_Y), f^k)\right]}{m} \tag{5.31b}$$

$$\bar{E}^{U|T}(g^k) \triangleq \lim_{m\to\infty} -\frac{\mathbb{E}_{\mathrm{RIE}}\left[\log p_{\mathrm{e}}^{U|T}(\mathcal{C}_\epsilon^{\mathcal{X}}(P_X), g^k)\right]}{m} \tag{5.31c}$$

$$\bar{E}^{V|T}(g^k) \triangleq \lim_{m\to\infty} -\frac{\mathbb{E}_{\mathrm{RIE}}\left[\log p_{\mathrm{e}}^{V|T}(\mathcal{C}_\epsilon^{\mathcal{Y}}(P_Y), g^k)\right]}{m}, \tag{5.31d}$$

where $\mathbb{E}_{\mathrm{RIE}}[\cdot]$ denotes expectation with respect to the RIEs for $\mathcal{C}_\epsilon^{\mathcal{X}}(P_X)$ and $\mathcal{C}_\epsilon^{\mathcal{Y}}(P_Y)$.

Our main result is the following proposition, which identifies the features the system designer should choose, and the exponent of the resulting error probability. A proof is provided in Appendix C.4.

Proposition 5.10. Given $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ and attributes U and V of X and Y, respectively, each drawn from a RIE for some $\epsilon > 0$, then for any dimension $k \in \{1, \dots, K-1\}$,[6]

$$\begin{multline}\left(\bar{E}^{U|S}(f^k), \bar{E}^{V|S}(f^k), \bar{E}^{U|T}(g^k), \bar{E}^{V|T}(g^k)\right) \\ \leq \left(\bar{E}_0^{X|U}\,\epsilon^2\,k,\ \bar{E}_0^{Y|V}\,\epsilon^2\sum_{i=1}^{k}\sigma_i^2, \bar{E}_0^{X|U}\,\epsilon^2\sum_{i=1}^{k}\sigma_i^2,\ \bar{E}_0^{Y|V}\,\epsilon^2\,k\right) + o(\epsilon^2)\end{multline} \tag{5.32}$$

as $\epsilon \to 0$, where $\bar{E}_0^{X|U}$ and $\bar{E}_0^{Y|V}$ are positive constants that do not depend on ϵ, k, or $P_{X,Y}$. Moreover, all the inequalities in (5.32) simultaneously hold with equality for the choices (f_*^k, g_*^k) as defined in (3.5), i.e., the associated multi-objective maximization has a unique[7] Pareto-optimal solution.

[6] For arbitrary sequences a^l and b^l of arbitrary length l, we use $a^l \leq b^l$ to denote that $a_i \leq b_i$ for $i \in \{1, \dots, l\}$.

[7] Note that while the optimized multi-objective function is unique, the features that achieve them need not be, as is the case when there are repeated singular values.

We emphasize that the result does not depend on any details of the probability law governing nature's choice other than the RIE property. In particular, f_*^k, g_*^k are optimum no matter what priors we might place over various parameters of the configurations $\mathcal{C}_\epsilon^{\mathcal{X}}(P_X), \mathcal{C}_\epsilon^{\mathcal{Y}}(P_Y)$ generating the RIE. In this sense, their optimality is fairly strong.

5.5 Universal Features via a Cooperative Game

In this section, we show how the same universal features arise as the solution to a *cooperative game*, which further reveals the latent variable configurations for which these features are effectively sufficient statistics. In this game, for a given $k \in \{1, \dots, K-1\}$, nature chooses configurations $\mathcal{C}_\epsilon^{\mathcal{X},k}(P_X)$ and $\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y)$ of multi-attribute variable collections $U = U^k$ and $V = V^k$ in (5.8), and the system designer chooses the features f^k and g^k. Their shared goal is to identify variables (U^k, V^k) that are, in an appropriate sense, most detectable from the statistics (S^k, T^k) as defined in (5.11) in terms of these features.

The specific shared goal of nature and the system designer is to maximize the probability that the least detectable of $U_1, \dots, U_i$ and the least detectable of $V_1, \dots, V_i$, for $i = 1, \dots, k$, are correctly detected, as $m \to \infty$.

For the analysis of this game, the following min-max characterization of singular values is useful. [114, Theorem 4.2.6].

Lemma 5.11 (Courant-Fischer)**.** Let $\mathbf{A}$ be a $k_1 \times k_2$ matrix with singular values $\sigma_1(\mathbf{A}) \geq \cdots \geq \sigma_k(\mathbf{A})$ where $k = \min\{k_1, k_2\}$. Then for every $i \in \{1, \dots, k\}$,

$$\sigma_i(\mathbf{A}) = \max_{\{\mathcal{S} \subset \mathbb{R}^k :\ \dim(\mathcal{S}) = i\}} \ \min_{\{\boldsymbol{\phi} \in \mathcal{S} :\ \|\boldsymbol{\phi}\| = 1\}} \|\mathbf{A}\boldsymbol{\phi}\|, \tag{5.33}$$

where $\mathcal{S}$ denotes a subspace, and the maximum is achieved by $\boldsymbol{\phi} = \boldsymbol{\psi}_i^{\mathrm{R}}$, a right singular vector of $\mathbf{A}$ corresponding to $\sigma_i(\mathbf{A})$.

In addition, in our development the following well-known inequality, which follows from the fact that the spectral norm $\|\cdot\|_{\mathrm{s}}$ (defined on p. 8) is the matrix norm induced by the (Euclidean) vector norm $\|\cdot\|$, is convenient.

Fact 5.12. For any compatible matrices $\mathbf{A}_1$ and $\mathbf{A}_2$, we have

$$\|\mathbf{A}_1\mathbf{A}_2\|_{\mathrm{F}} \leq \|\mathbf{A}_1\|_{\mathrm{s}} \, \|\mathbf{A}_2\|_{\mathrm{F}}.$$

In the sequel, we denote the error probabilities in decisions based on S^k about each of constituent elements of U^k and V^k, respectively, via

$$p_{\mathrm{e}}^{U_i|S}\big(\mathcal{C}_\epsilon^{\mathcal{X},k}(P_X), f^k\big) \quad \text{and} \quad p_{\mathrm{e}}^{V_i|S}\big(\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y), f^k\big), \tag{5.34a}$$

and those based on T^k via, respectively,

$$p_{\mathrm{e}}^{U_i|T}\big(\mathcal{C}_\epsilon^{\mathcal{X},k}(P_X), g^k\big) \quad \text{and} \quad p_{\mathrm{e}}^{V_i|T}\big(\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y), g^k\big), \tag{5.34b}$$

for $i \in \{1, \dots, k\}$, where S^k and T^k are as defined in (5.11) for feature choices $f^k\colon \mathcal{X} \to \mathbb{R}^k$ and $g^k\colon \mathcal{Y} \to \mathbb{R}^k$. In turn, we define the error exponents

$$E^{U_i|S}\big(\mathcal{C}_\epsilon^{\mathcal{X},k}(P_X), f^k\big) \triangleq \lim_{m\to\infty} \frac{-\log p_{\mathrm{e}}^{U_i|S}\big(\mathcal{C}_\epsilon^{\mathcal{X},k}(P_X), f^k\big)}{m} \tag{5.35a}$$

$$E^{V_i|S}\big(\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y), f^k\big) \triangleq \lim_{m\to\infty} \frac{-\log p_{\mathrm{e}}^{V_i|S}\big(\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y), f^k\big)}{m} \tag{5.35b}$$

$$E^{U_i|T}\big(\mathcal{C}_\epsilon^{\mathcal{X},k}(P_X), g^k\big) \triangleq \lim_{m\to\infty} \frac{-\log p_{\mathrm{e}}^{U_i|T}\big(\mathcal{C}_\epsilon^{\mathcal{X},k}(P_X), g^k\big)}{m} \tag{5.35c}$$

$$E^{V_i|T}\big(\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y), g^k\big) \triangleq \lim_{m\to\infty} \frac{-\log p_{\mathrm{e}}^{V_i|T}\big(\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y), g^k\big)}{m}. \tag{5.35d}$$

Our main result is as follows. A proof is provided in Appendix C.5.

Proposition 5.13. Given $k \in \{1, \dots, K-1\}$ and $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$, let $\mathcal{C}_\epsilon^{\mathcal{X},k}(P_X)$ and $\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y)$ denote configurations of ϵ-multi-attribute variables U^k and V^k of X and Y, respectively, for some $\epsilon > 0$. Then

$$\begin{aligned}\Big(&\min_{j\leq i} E^{U_j|S}\big(\mathcal{C}_\epsilon^{\mathcal{X},k}(P_X), f^k\big),\ i \in \{1,\dots,k\},\\ &\min_{j\leq i} E^{V_j|S}\big(\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y), f^k\big),\ i \in \{1,\dots,k\},\\ &\min_{j\leq i} E^{U_j|T}\big(\mathcal{C}_\epsilon^{\mathcal{X},k}(P_X), g^k\big),\ i \in \{1,\dots,k\},\\ &\min_{j\leq i} E^{V_j|T}\big(\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y), g^k\big),\ i \in \{1,\dots,k\}\Big)\end{aligned}$$

$$\leq \Bigg(\frac{\epsilon^2}{2},\ i \in \{1,\dots,k\},$$
$$\frac{\epsilon^2}{2}\sigma_i^2,\ i \in \{1,\dots,k\},$$
$$\frac{\epsilon^2}{2}\sigma_i^2,\ i \in \{1,\dots,k\},$$
$$\frac{\epsilon^2}{2},\ i \in \{1,\dots,k\} \Bigg) + o(\epsilon^2), \quad \epsilon \to 0. \tag{5.36}$$

Moreover, the inequalities in (5.36) all hold with equality when (f^k, g^k) are chosen to be (f_*^k, g_*^k) as defined in (3.5), and $\mathcal{C}_\epsilon^{\mathcal{X},k}(P_X)$ and $\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y)$ as chosen to be, respectively,

$$\begin{aligned}\mathcal{C}_{\epsilon,*}^{\mathcal{X},k}(P_X) \triangleq \Big\{ &\mathcal{U}_i = \{+1,-1\},\ \{P_{U_i}(u_i) = 1/2,\ u_i \in \mathcal{U}_i\},\\ &\{P_{X|U_i}(x|u_i) = P_X(x)(1+\epsilon u_i\, f_i^*(x)),\\ &u_i \in \mathcal{U}_i,\ x \in \mathcal{X}\},\ i = 1,\dots,k \Big\}\end{aligned} \tag{5.37a}$$

and

$$\begin{aligned}\mathcal{C}_{\epsilon,*}^{\mathcal{Y},k}(P_Y) \triangleq \Big\{ &\mathcal{V}_i = \{+1,-1\}, \quad \{P_{V_i}(v_i) = 1/2,\ v_i \in \mathcal{V}_i\},\\ &\{P_{Y|V_i}(y|v_i) = P_Y(y)(1+\epsilon v_i\, g_i^*(y)),\\ &v_i \in \mathcal{V}_i,\ y \in \mathcal{Y}\},\ i = 1,\dots,k \Big\},\end{aligned} \tag{5.37b}$$

i.e., the associated multi-objective maximization has a unique Pareto-optimal solution.

In addition, from the Markov structure (5.7) and the modal structure (2.26), we immediately obtain the following corollary.

Corollary 5.14. The optimizing multi-attribute variables U^k and V^k in Proposition 5.13 have the property that

$$P_{X|V_i}(x|v_i) = P_X(x)(1+\epsilon v_i\, \sigma_i\, f_i^*(x)) \tag{5.38a}$$
$$P_{Y|U_i}(y|u_i) = P_Y(y)(1+\epsilon u_i\, \sigma_i\, g_i^*(y)), \tag{5.38b}$$

for $i = 1,\dots,k$.

Given data (x^m, y^m) from the extended Markov model (5.8), it further follows that (S_*^k, T_*^k) defined via (5.12) is, as $\epsilon \to 0$, a sufficient statistic for inferences about the optimizing multi-attributes (U^k, V^k), i.e., in this limit, we have the Markov chains

$$(U^k, V^k) \leftrightarrow (S_*^k, T_*^k) \leftrightarrow (X^m, Y^m) \tag{5.39}$$

and

$$U^k \leftrightarrow S_*^k \leftrightarrow T_*^k \leftrightarrow V^k. \tag{5.40}$$

In particular, we have the following result, a proof of which is provided in Appendix C.6.

Corollary 5.15. In the solution to the optimization in Proposition 5.13 for the extended model (5.8),

$$P_{U^k,V^k|X^m,Y^m}(u^k, v^k|x^m, y^m) = \frac{1}{4^k}\left(1 + \epsilon m \sum_{i=1}^{k}(u_i\, s_i^* + v_i\, t_i^*)\right) + o(\epsilon), \tag{5.41}$$

as $\epsilon \to 0$, with, consistent with (5.12),

$$s_i^* = \frac{1}{m}\sum_{j=1}^{m} f_i^*(x_j) \quad \text{and} \quad t_i^* = \frac{1}{m}\sum_{j=1}^{m} g_i^*(y_j). \tag{5.42}$$

Moreover,

$$P_{U^k|S_*^k,T_*^k,V^k}(u^k|s_*^k, t_*^k, v^k) = \frac{1}{2^k}\left(1 + \epsilon m \sum_{i=1}^{k} u_i\, s_i^*\right) + o(\epsilon) \tag{5.43a}$$

$$P_{V^k|S_*^k,T_*^k,U^k}(v^k|s_*^k, t_*^k, u^k) = \frac{1}{2^k}\left(1 + \epsilon m \sum_{i=1}^{k} u_i\, t_i^*\right) + o(\epsilon), \tag{5.43b}$$

as $\epsilon \to 0$.

In essence, Corollary 5.15 shows that in making inferences about the (optimizing) attributes U^k and V^k from high-cardinality data (X^m, Y^m), it is (asymptotically) sufficient to extract a low-dimensional real-valued sufficient statistic (S^k, T^k). Moreover, it is sufficient to extract a statistic of dimension k corresponding to the number of "significant" singular values of $\tilde{\mathbf{B}}$. As significantly, the sufficient statistic pair (S_*^k, T_*^k) is obtained by *separate* processing of X^m and Y^m; joint processing of (X^m, Y^m) is not required.

In addition, as suggested by Corollary 5.15 and revealed in its proof,

$$P_{U^k,V^k|X^m,Y^m}(u^k, v^k|x^m, y^m) = \prod_{i=1}^{k} P_{U_i,V_i|X^m,Y^m}(u_i, v_i|x^m, y^m) \tag{5.44a}$$

with

$$P_{U_i,V_i|X^m,Y^m}(u_i, v_i|x^m, y^m) = \frac{1}{4}\left(1 + \epsilon m(u_i\, s_i^* + v_i\, t_i^*)\right) + o(\epsilon), \tag{5.44b}$$

as $\epsilon \to 0$, from which we see that to achieve the optimum exponents given by (5.36), it is sufficient for decisions about the attribute pair (U_i, V_i) to be made based on the statistic (S_i^*, T_i^*) alone. Moreover, (5.44a) reveals that although not imposed as a constraint, the optimizing configuration is such that the U^k are conditionally independent of Y^m (and the V^k are conditionally independent of X^m).

5.6 Universal Features via an Information Bottleneck

In this section, we show that the same configurations $\mathcal{C}_{\epsilon,*}^{\mathcal{X},k}(P_X)$ and $\mathcal{C}_{\epsilon,*}^{\mathcal{Y},k}(P_Y)$ that are optimum in the cooperative game of Section 5.5 are the solution to a natural mutual information maximization problem, which provides a third viewpoint from which to interpret f_*^k and g_*^k as universal features.

Our main result is as follows. A proof is provided in Appendix C.7.

Proposition 5.16. Given $\epsilon > 0$, $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$, and ϵ-multi-attribute variables $U = U^k$ and $V = V^k$ for some k in the Markov chain (5.7), then

$$I(U^k; V^k) \leq \frac{\epsilon^4}{2} \sum_{i=1}^{k} \sigma_i^2 + o(\epsilon^4), \quad \epsilon \to 0. \tag{5.45}$$

Moreover, the inequality in (5.45) holds with equality when the configurations $\mathcal{C}_{\epsilon,*}^{\mathcal{X},k}(P_X)$ and $\mathcal{C}_{\epsilon,*}^{\mathcal{Y},k}(P_Y)$ of U^k and V^k, respectively, are given by (5.37), in which case

$$P_{U^k,V^k}(u^k,v^k) = \frac{1}{4^k}\left(1+\epsilon^2\sum_{i=1}^{k}\sigma_i\,u_i\,v_i\right) + o(\epsilon^2), \quad u^k,v^k \in \{+1,-1\}^k. \tag{5.46}$$

We interpret the optimizing multi-attribute pair (U^k,V^k) in Proposition 5.16, with joint distribution (5.46), as expressing the dominant components of the dependency in the relationship between X and Y as determined by $P_{X,Y}$. This is reflected in the fact that

$$P_{U^k,V^k}(u^k,v^k) = \prod_{i=1}^{k} P_{U_i,V_i}(u_i,v_i)$$

with, for $i,j = 1,\dots,k$,

$$P_{U_i,V_j}(u_i,v_j) = \frac{1}{4}(1+\epsilon^2\,\sigma_i\,u_i\,v_j\,\mathbb{1}_{i=j}) + o(\epsilon^2), \tag{5.47}$$

for which [cf. (4.28)]

$$I(U_i;V_j) = \frac{\epsilon^4}{2}\,\sigma_i^2\,\mathbb{1}_{i=j} + o(\epsilon^4). \tag{5.48}$$

An immediate consequence of Proposition 5.16 is that for observations X^m,Y^m from the model (5.8), we have that (S_*^k,T_*^k) is a (locally) sufficient statistic for inferences about U^k,V^k, i.e., Corollary 5.15 applies. This local sufficiency can be equivalently expressed in the form

$$\lim_{\epsilon\to 0}\frac{I(U^k,V^k;X^m,Y^m)}{I(U^k,V^k;S_*^k,T_*^k)} = 1. \tag{5.49}$$

Note too that (5.46) provides a higher-order characterization of P_{U^k,V^k} than that derived from (5.41). Indeed, from the latter (setting $m=1$ for convenience) we obtain only

$$\begin{aligned}
&P_{U^k,V^k}(u^k,v^k)\\
&\quad= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\,P_{U^k,V^k|X,Y}(u^k,v^k,x,y)\\
&\quad= \frac{1}{4^k}\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\left(1+\epsilon m\sum_{i=1}^{k}(u_i\,f_i^*(x)+v_i\,g_i^*(y))\right) + o(\epsilon)\\
&\quad= \frac{1}{4^k} + o(\epsilon), \quad \epsilon\to 0.
\end{aligned}$$

Additionally, it follows from the discussion in Section 4.2 that when $\mathcal{U}_i = \mathcal{V}_i = \{+1, -1\}$ and $P_{U_i} = P_{V_i} \equiv 1/2$, then the ϵ-multi-attribute constraints $P_{X|U_i}(\cdot|u_i) \in \mathcal{N}_\epsilon^{\mathcal{X}}(P_X)$ and $P_{Y|V_i}(\cdot|u_i) \in \mathcal{N}_\epsilon^{\mathcal{Y}}(P_Y)$ are equivalent to $I(X, U_i) \leq \epsilon^2/2$ and $I(Y; V_i) \leq \epsilon^2/2$ as $\epsilon \to 0$, for $i = 1, \dots, k$. As a result, Proposition 5.16 can be equivalently expressed in the form of a solution to a information bottleneck problem [258] in the weak dependence regime.[8] In particular, we have the following immediate corollary.[9]

Corollary 5.17. Given $\epsilon > 0$, $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ and variables $U = U^k$ and $V = V^k$ in the Markov chain (5.7), then

$$\max_{U^k, V^k} I(U^k; V^k) = \frac{\epsilon^4}{2} \sum_{i=1}^{k} \sigma_i^2 + o(\epsilon^4), \quad \epsilon \to 0,$$

where the maximization is over all configurations of (U^k, V^k) such that: constituent variables (U_i, V_i), $i = 1, \dots, k$ satisfy:

1) $\max\{I(U_i; X), I(V_i; Y)\} \leq \epsilon^2/2$;

2) they are binary and equiprobable, i.e., $U_i, V_i \in \{+1, -1\}$ and $P_{U_i} = P_{V_i} \equiv 1/2$;

3) U^k and V^k are each collections of independent variables; and

4) the U^k and V^k are each collections of conditionally independent variables given X and Y, respectively.

Moreover, the maximum is achieved by the configurations (5.37).

As an aside, different but related one-sided information bottleneck problems can also be analyzed within the same framework of analysis.

[8] For an early application of the use of information bottleneck techniques in learning, see [253], [259].

[9] As the proof reveals, sufficiently weak pairwise dependence will suffice for condition 3, but we impose mutual independence for convenience. Moreover, while condition 4 alone implies a degree of weak marginal dependence, it is insufficient. Finally, conditions 3 and 4 together can be viewed, in some sense, as "entropy maximizing" conditions.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

For example, the following result is proved in Appendix C.8.[10]

Proposition 5.18. Given $\epsilon > 0$, $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ and variables $U = U^k$ and $V = V^k$ in the Markov chain (5.7), then

$$\max_{V^k} I(V^k; X) = \max_{U^k} I(U^k; Y) = \frac{\epsilon^2}{2} \sum_{i=1}^{k} \sigma_i^2 + o(\epsilon^2), \quad \epsilon \to 0,$$

where the maximization is over all configurations of (U^k, V^k) such that: constituent variables U_i, V_i, $i = 1, \dots, k$ satisfy:

1) $\max\{I(U_i; X), I(V_i; Y)\} \leq \epsilon^2/2$;
2) they are binary and equiprobable, i.e., $U_i, V_i \in \{+1, -1\}$ and $P_{U_i} = P_{V_i} \equiv 1/2$;
3) U^k and V^k are each collections of independent variables; and
4) the U^k and V^k are collections of conditionally independent variables given X and Y, respectively.

Moreover, the maximum is achieved by the configurations (5.37).

As a further comment, work on aspects of the more general information bottleneck problem and the associated role of hypercontractivity analysis in its treatment includes [3], [10]–[12], [35], [142], [155], [196], [226], [228]. From this perspective, the development of this section reveals that a number of the subtleties that complicate such analysis and lead to anomalous behavior are avoided by the restriction to local variables.

5.7 Universal Features via Wyner Common Information

As we develop in this section, there is a key relationship between the optimizing multi-attributes (U^k, V^k) in Section 5.6 (and Section 5.5),

[10] Other variations of this result correspond to avoiding the binary, equiprobable and mutual information constraints and instead using $D(P_{X|U}(\cdot|u) \| P_X) \leq \epsilon^2/2$ for all $u \in \mathcal{U}$. Alternatively, by the equidistant property of capacity-achieving output distributions, we can equivalently express this divergence constraint as $\max_{P_U} I(X; U) \leq \epsilon^2/2$.

and the Wyner common information $C(X,Y)$ in the pair (X,Y) characterized by a given joint distribution $P_{X,Y}$. Recall that Wyner common information can be expressed in terms of an auxiliary variable W according to [279]

$$C(X,Y) \triangleq \min_{\substack{P_{W|X,Y}: \\ X \leftrightarrow W \leftrightarrow Y}} I(W;X,Y), \tag{5.50}$$

and satisfies $C(X,Y) \geq I(X;Y)$.

The results we obtain are for the case where X and Y are extra-weakly dependent; specifically, for some $\epsilon > 0$,

$$\|\tilde{\mathbf{B}}\|_* \leq \epsilon, \tag{5.51}$$

where $\|\cdot\|_*$ denotes the nuclear norm of its argument.[11] We refer to X and Y as sub-ϵ dependent in this case, since by standard norm inequalities [97]

$$\|\tilde{\mathbf{B}}\|_{\mathrm{F}} \leq \|\tilde{\mathbf{B}}\|_*, \tag{5.52}$$

whence $P_{X,Y} \in \mathcal{N}_\epsilon^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y)$, i.e., sub-$\epsilon$ dependence implies ϵ-dependence.[12] We further use $\bar{\mathcal{N}}_\epsilon^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y)$ to denote the joint distributions in $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ with sub-ϵ dependence given marginals P_X and P_Y, via which (5.52) expresses

$$\bar{\mathcal{N}}_\epsilon^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y) \subset \mathcal{N}_\epsilon^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y). \tag{5.53}$$

Under sub-ϵ dependence, we define the following restricted common information.

Definition 5.19 (ϵ-Common Information)**.** Given $P_{X,Y} \in \bar{\mathcal{N}}_\epsilon^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y)$ for $\epsilon > 0$, the ϵ-common information is

$$C_\epsilon(X,Y) = \min_{P_{W|X,Y} \in \mathcal{P}_\epsilon} I(W;X,Y), \tag{5.54}$$

[11] Specifically, the nuclear norm of an arbitrary matrix $\mathbf{A}$ is

$$\|\mathbf{A}\|_* \triangleq \operatorname{tr}\left(\sqrt{\mathbf{A}^{\mathrm{T}}\mathbf{A}}\right) = \sum_i \sigma_i(\mathbf{A}),$$

where $\sigma_i(\mathbf{A})$ denotes the ith singular value of $\mathbf{A}$. Note that the nuclear norm is the Ky Fan k-norm with $k = \operatorname{rank}(\mathbf{A})$, i.e., $\|\mathbf{A}\|_* = \|\mathbf{A}\|_{(\operatorname{rank}(\mathbf{A}))}$.

[12] Of course, since $\|\mathbf{A}\|_* \leq \sqrt{\operatorname{rank}(\mathbf{A})}\,\|\mathbf{A}\|_{\mathrm{F}}$ for any $\mathbf{A}$, we also have that ϵ-dependence implies sub-$(K\epsilon)$ dependence. In turn, $\mathcal{O}(\epsilon)$-dependence and sub-$\mathcal{O}(\epsilon)$ dependence are equivalent.

where

$$\begin{aligned}\mathcal{P}_\epsilon \triangleq \Big\{ & P_{W|X,Y} \in \mathcal{P}^{\mathcal{W}},\ \text{some } \mathcal{W}\colon X \leftrightarrow W \leftrightarrow Y \text{ and} \\ & P_{X|W}(\cdot|w) \in \mathcal{N}^{\mathcal{X}}_{\sqrt{\delta(\epsilon)}}(P_X),\ P_{Y|W}(\cdot|w) \in \mathcal{N}^{\mathcal{Y}}_{\sqrt{\delta(\epsilon)}}(P_Y), \\ & \text{for all } w \in \mathcal{W} \text{ and } \delta(\cdot) > 0 \text{ such that } \lim_{\epsilon\to 0} \delta(\epsilon) \to 0. \Big\}. \end{aligned} \tag{5.55}$$

In Definition 5.19, a configuration of W such that $P_{W|X,Y} \in \mathcal{P}_\epsilon$ takes the form

$$\begin{aligned}\mathcal{C}^{\mathcal{X},\mathcal{Y}}_\epsilon(P_{X,Y}) \triangleq \Big\{ & \mathcal{W},\ \{P_W(w),\ w \in \mathcal{W}\}, \\ & \{P_{X|W}(\cdot|w),\ w \in \mathcal{W}\}, \\ & \{P_{Y|W}(\cdot|w),\ w \in \mathcal{W}\}\Big\} \end{aligned} \tag{5.56}$$

subject to the constraints

$$P_{X|W}(\cdot|w) \in \mathcal{N}^{\mathcal{X}}_{\sqrt{\delta(\epsilon)}}(P_X),\ w \in \mathcal{W}, \tag{5.57a}$$

$$P_{Y|W}(\cdot|w) \in \mathcal{N}^{\mathcal{Y}}_{\sqrt{\delta(\epsilon)}}(P_Y),\ w \in \mathcal{W}, \tag{5.57b}$$

for some δ such that $\delta(\epsilon) \to 0$ as $\epsilon \to 0$, and

$$P_{X|W}(x|w)\, P_{Y|W}(y|w) = P_{X,Y|W}(x,y|w). \tag{5.58}$$

In turn, (5.58) implies the constraint

$$\sum_{w\in\mathcal{W}} P_W(w)\, P_{X|W}(x|w)\, P_{Y|W}(y|w) = P_{X,Y}(x,y), \tag{5.59}$$

which further implies the constraints

$$\sum_{w\in\mathcal{W}} P_W(w)\, P_{X|W}(x|w) = P_X(x) \tag{5.60a}$$

$$\sum_{w\in\mathcal{W}} P_W(w)\, P_{Y|W}(y|w) = P_Y(y). \tag{5.60b}$$

Defining the information vectors

$$\phi^{X|W}_w(x) \triangleq \frac{P_{X|W}(x|w) - P_X(x)}{\sqrt{\delta(\epsilon)}\sqrt{P_X(x)}} \tag{5.61a}$$

$$\phi^{Y|W}_w(y) \triangleq \frac{P_{Y|W}(y|w) - P_Y(y)}{\sqrt{\delta(\epsilon)}\sqrt{P_Y(y)}}, \tag{5.61b}$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

and

$$\tilde{\phi}_w^{X,Y|W}(x,y) \triangleq \frac{P_{X,Y|W}(x,y|w) - P_X(x)\,P_Y(y)}{\sqrt{2\delta(\epsilon)}\sqrt{P_X(x)\,P_Y(y)}}, \tag{5.61c}$$

we can equivalently express (5.56) in the form

$$\begin{aligned} \mathcal{C}_\epsilon^{\mathcal{X},\mathcal{Y}}(P_{X,Y}) \triangleq \Big\{ \mathcal{W},\ \{P_W(w),\ w \in \mathcal{W}\},& \\ \{\boldsymbol{\phi}_w^{X|W},\ w \in \mathcal{W}\}& \\ \{\boldsymbol{\phi}_w^{Y|W},\ w \in \mathcal{W}\}\Big\}& \end{aligned} \tag{5.62}$$

subject to the constraints

$$\boldsymbol{\phi}_w^{X|W} \in \mathcal{I}^{\mathcal{X}}, \quad w \in \mathcal{W}, \tag{5.63a}$$

$$\boldsymbol{\phi}_w^{Y|W} \in \mathcal{I}^{\mathcal{Y}}, \quad w \in \mathcal{W}, \tag{5.63b}$$

which correspond to (5.57), and

$$\tilde{\phi}_w^{X,Y|W}(x,y) = \check{\phi}_w^{X,Y|W}(x,y) + \sqrt{\frac{\delta(\epsilon)}{2}}\,\phi_w^{X|W}(x)\,\phi_w^{Y|W}(y) \tag{5.64a}$$

with

$$\check{\phi}_w^{X,Y|W}(x,y) \triangleq \frac{1}{\sqrt{2}}\Big(\sqrt{P_Y(y)}\,\phi_w^{X|W}(x) + \sqrt{P_X(x)}\,\phi_w^{Y|W}(y)\Big), \tag{5.64b}$$

as well as

$$\delta(\epsilon)\sum_{w\in\mathcal{W}} P_W(w)\,\phi_w^{X|W}(x)\,\phi_w^{Y|W}(y) = \tilde{B}(y,x), \tag{5.65}$$

for $x \in \mathcal{X},\ y \in \mathcal{Y}$, and

$$\sum_{w\in\mathcal{W}} P_W(w)\,\phi_w^{X|W}(x) = 0, \quad x \in \mathcal{X}, \tag{5.66a}$$

$$\sum_{w\in\mathcal{W}} P_W(w)\,\phi_w^{Y|W}(y) = 0, \quad y \in \mathcal{Y}, \tag{5.66b}$$

which correspond to (5.58)–(5.60), respectively. In particular, we obtain

(5.66) from (5.60) using (5.61a)–(5.61b), and we obtain (5.64) from

$$
\begin{aligned}
&P_{X,Y|W}(x,y|w) && (5.67)\\
&\quad= P_{X|W}(x|w)\,P_{Y|W}(y|w) && (5.68)\\
&\quad= \Big(P_X(x)+\sqrt{\delta(\epsilon)}\,\sqrt{P_X(x)}\,\phi_w^{X|W}(x)\Big)\\
&\qquad\qquad\cdot\Big(P_Y(y)+\sqrt{\delta(\epsilon)}\,\sqrt{P_Y(y)}\,\phi_w^{Y|W}(y)\Big)\\
&\quad= P_X(x)\,P_Y(y)+\sqrt{\delta(\epsilon)}\sqrt{P_X(x)\,P_Y(y)}\\
&\qquad\cdot\Big[\sqrt{P_Y(y)}\,\phi_w^{X|W}(x)+\sqrt{P_X(x)}\,\phi_w^{Y|W}(y)\\
&\qquad\qquad+\sqrt{\delta(\epsilon)}\,\phi_w^{X|W}(x)\,\phi_w^{Y|W}(y)\Big], && (5.69)
\end{aligned}
$$

where we have used (5.58) and (5.61a)–(5.61b), and where we recognize the term in brackets as $\sqrt{2}\,\tilde{\phi}_w^{X,Y|W}(x)$ according to (5.61c). Finally, we obtain (5.65) from the expectation of (5.69) with respect to P_W, yielding

$$
\begin{aligned}
&P_{X,Y}(x,y)\\
&\quad=\sum_{w\in\mathcal{W}}P_W(w)\,P_{X|W}(x|w)\,P_{Y|W}(y|w)\\
&\quad=P_X(x)\,P_Y(y)\\
&\qquad+\sqrt{P_X(x)\,P_Y(y)}\,\delta(\epsilon)\sum_{w\in\mathcal{W}}P_W(w)\,\phi_w^{X|W}(x)\,\phi_w^{Y|W}(y), \quad (5.70)
\end{aligned}
$$

where we have used (5.66), and where we recognize $\tilde{B}(y,x)$ as defined in (2.28) as the final factor in (5.70).

The following variational characterization of the nuclear (i.e., trace) norm (see, e.g., [16], [230]) is useful in our development.

Lemma 5.20. Given an arbitrary $k_1\times k_2$ matrix $\mathbf{A}$, we have

$$
\min_{\substack{\{k,\ \mathbf{M}_1\in\mathbb{R}^{k_1\times k},\ \mathbf{M}_2\in\mathbb{R}^{k\times k_2}:\\ \mathbf{M}_1\mathbf{M}_2=\mathbf{A}\}}}\left(\frac{1}{2}\|\mathbf{M}_1\|_{\mathrm{F}}^2+\frac{1}{2}\|\mathbf{M}_2\|_{\mathrm{F}}^2\right)=\|\mathbf{A}\|_*. \quad (5.71)
$$

In particular, we obtain that the ϵ-common information is given by the nuclear norm of $\tilde{\mathbf{B}}$. A proof is provided in Appendix C.9.

Proposition 5.21. Given $P_{X,Y} \in \bar{\mathcal{N}}_{\epsilon}^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y)$ for $\epsilon > 0$, we have[13]

$$C(X,Y) \le C_{\epsilon}(X,Y) = \|\tilde{\mathbf{B}}\|_{*} + o(\epsilon), \quad \epsilon \to 0, \tag{5.72a}$$

where

$$\|\tilde{\mathbf{B}}\|_{*} = \sum_{i=1}^{K-1} \sigma_i, \tag{5.72b}$$

which is achieved by the configuration

$$\begin{aligned} &C_{*}^{\mathcal{X},\mathcal{Y}}(P_{X,Y}) \\ &\quad = \Bigg\{ \mathcal{W} = \{\pm 1, \dots, \pm(K-1)\}, \\ &\qquad P_W(w) = \frac{\sigma_{|w|}}{2\|\tilde{\mathbf{B}}\|_{*}}, \\ &\qquad P_{X|W}(x|w) = P_X(x)\Big(1 + \operatorname{sgn}(w)\, \|\tilde{\mathbf{B}}\|_{*}^{1/2}\, f_{|w|}^{*}(x)\Big), \\ &\qquad P_{Y|W}(y|w) = P_Y(y)\Big(1 + \operatorname{sgn}(w)\, \|\tilde{\mathbf{B}}\|_{*}^{1/2}\, g_{|w|}^{*}(y)\Big) \Bigg\}. \end{aligned} \tag{5.73}$$

and $\delta(\epsilon) = \epsilon$ in (5.55).

We note that while in general the cardinality of W in the characterization of Wyner common information is known only to satisfy the upper bound $|\mathcal{W}| \le |\mathcal{X}| \times |\mathcal{Y}|$, we obtain that cardinality $|\mathcal{W}| = 2(K-1)$, which is much smaller when $\mathcal{X}$ and/or $\mathcal{Y}$ is large, suffices to achieve ϵ-common information as $\epsilon \to 0$.

Given data (x^m, y^m) from the extended model

$$X^m \leftrightarrow W \leftrightarrow Y^m, \tag{5.74a}$$

[13] Since $I(W;X,Y) \ge \max\{I(W;X), I(W;Y)\}$ by the chain rule, it follows that our result does not change if we further include in (5.55) of Definition 5.19 all distributions $P_{X|W}(\cdot|w)$ and $P_{Y|W}(\cdot|w)$ that for all $w \in \mathcal{W}$ do not depend on ϵ, since they will give rise to nonvanishing $I(W;X)$ and $I(W;Y)$. In essence, the configurations our definition omits are those for which the P_W is increasingly severely imbalanced as $\epsilon \to 0$.

with

$$P_{X^m|W}(x^m|w) = \prod_{i=1}^{m} P_{X|W}(x_i|w) \tag{5.74b}$$

$$P_{Y^m|W}(y^m|w) = \prod_{i=1}^{m} P_{Y|W}(y_i|w) \tag{5.74c}$$

$$P_{X^m,Y^m}(x^m, y^m) = \prod_{i=1}^{m} P_{X,Y}(x_i, y_i), \tag{5.74d}$$

it further follows from Proposition 5.21 that

$$R_*^{K-1} \triangleq S_*^{K-1} + T_*^{K-1} \tag{5.75}$$

with S_*^{K-1} and T_*^{K-1} as defined via (5.12), is, as $\epsilon \to 0$, a sufficient statistic for inferences about the ϵ-common information variable W, i.e., we have the Markov chain

$$W \leftrightarrow R_*^{K-1} \leftrightarrow (S_*^{K-1}, T_*^{K-1}) \leftrightarrow (X^m, Y^m), \quad \epsilon \to 0. \tag{5.76}$$

In particular, we have the following result, a proof of which is provided in Appendix C.10.

Corollary 5.22. In the solution to the optimization in Proposition 5.21 for the extended model (5.74),

$$P_{W|X^m,Y^m}(w|x^m, y^m) = \frac{\sigma_{|w|}}{2\|\tilde{\mathbf{B}}\|_*}\Big(1 + m\,\mathrm{sgn}(w)\,\|\tilde{\mathbf{B}}\|_*^{1/2}\, r_{|w|}^*\Big) + o(\sqrt{\epsilon}), \tag{5.77}$$

as $\epsilon \to 0$, where, consistent with (5.75),

$$r_i^* = s_i^* + t_i^*, \tag{5.78}$$

and s_i^* and t_i^* are as defined in (5.42).

The sufficiency relation of Corollary 5.22 can be equivalently expressed in the form

$$\lim_{\epsilon \to 0} \frac{I(W; X^m, Y^m)}{I(W; R_*^{K-1})} = 1.$$

In essence, Corollary 5.22 shows that in making inferences about the ϵ-common information variable W from high-cardinality data (X^m, Y^m), it is sufficient to extract a low-dimensional real-valued sufficient statistic R^{K-1}. And we emphasize that a consequence of the way common information is defined is that the sufficient statistic R_*^{K-1} involves separate processing of X^m and Y^m.

5.8 Relating Common Information to Dominant Structure

The ϵ-common information variable W of Proposition 5.21 can be related to the dominant structure sequence pair (U^{K-1}, V^{K-1}) of Proposition 5.16 (and Proposition 5.13). To develop this, let us equivalently express W as

$$W \triangleq W^{K-1} = (W_1, \dots, W_{K-1}), \tag{5.79a}$$

where each W_i is a variable defined over alphabet

$$\mathcal{W}_\circ \triangleq \{-1, 0, +1\} \tag{5.79b}$$

according to

$$W_i \triangleq \begin{cases} +1 & W = i \\ -1 & W = -i \\ 0 & \text{otherwise.} \end{cases} \tag{5.79c}$$

We can interpret W_i as effectively capturing the ϵ-common information in (U_i, V_i), which is defined, consistent with Definition 5.19, as

$$C_\epsilon(U_i, V_i) = \min_{P_{\tilde{W}_i|U_i,V_i} \in \tilde{\mathcal{P}}_\epsilon} I(\tilde{W}_i; U_i, V_i), \tag{5.80}$$

where

$$\begin{aligned} \tilde{\mathcal{P}}_\epsilon \triangleq \Big\{ & P_{\tilde{W}_i|U_i,V_i} \in \mathcal{P}^{\tilde{\mathcal{W}}_i},\ \text{some } \tilde{\mathcal{W}}_i \colon U_i \leftrightarrow \tilde{W}_i \leftrightarrow V_i \text{ and} \\ & P_{U_i|\tilde{W}_i}(\cdot|\tilde{w}_i) \in \mathcal{N}^{\mathcal{U}_i}_{\sqrt{\delta(\epsilon)}}(P_{U_i}),\ P_{V_i|\tilde{W}_i}(\cdot|\tilde{w}_i) \in \mathcal{N}^{\mathcal{V}_i}_{\sqrt{\delta(\epsilon)}}(P_{V_i}), \\ & \text{for all } \tilde{w}_i \in \tilde{\mathcal{W}}_i \text{ and } \delta(\cdot) > 0 \text{ such that } \lim_{\epsilon \to 0} \delta(\epsilon) \to 0. \quad \Big\}. \end{aligned} \tag{5.81}$$

In particular, we have the following result, a proof of which is provided in Appendix C.11.

Corollary 5.23. Given $P_{X,Y} \in \bar{\mathcal{N}}_\epsilon^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y)$ for $\epsilon > 0$, and let W^{K-1} be the representation (5.79) of the optimizing ϵ-common information variable W in Proposition 5.21. Then

$$C_\epsilon(X,Y) = I(W;X,Y) = \sum_{i=1}^{K-1} I(W_i;X,Y) + o(\epsilon), \quad \epsilon \to 0, \tag{5.82}$$

where

$$I(W_i;X,Y) = \sigma_i + o(\epsilon), \quad \epsilon \to 0, \quad i \in \{1,\dots,K-1\}. \tag{5.83}$$

Moreover, if (U^{K-1}, V^{K-1}) are the optimizing $\tilde{\epsilon}$-multi-attributes in Proposition 5.16 for some $\tilde{\epsilon} > 0$, then

$$C_\epsilon(U_i,V_i) = \tilde{\epsilon}^2 \, I(W_i;X,Y) + o(\tilde{\epsilon}^2\epsilon), \quad \tilde{\epsilon},\epsilon \to 0, \quad i \in \{1,\dots,K-1\}. \tag{5.84}$$

The associated data processing implications follow from the extended Markov structure (5.74). In particular, an (asymptotically) sufficient statistic for making decisions about W_i from (X^m, Y^m) is

$$R_i^* = S_i^* + T_i^*, \tag{5.85}$$

i.e., we have the Markov chain

$$W_i \leftrightarrow R_i^* \leftrightarrow (S_i^*, T_i^*) \leftrightarrow (X^m, Y^m), \quad \epsilon \to 0.$$

In particular, we have the following result, a proof of which is provided in Appendix C.12.

Corollary 5.24. For W_i as defined in (5.79),

$$\begin{aligned} & P_{W_i|X^m,Y^m}(w_i|x^m,y^m) \\ & \quad = \begin{cases} \dfrac{\sigma_i}{2\,\|\tilde{\mathbf{B}}\|_*}\left(1 + m w_i \,\|\tilde{\mathbf{B}}\|_*^{1/2}\, r_i^*\right) + o(\sqrt{\epsilon}) & w_i = \pm 1 \\ \left(1 - \dfrac{\sigma_i}{\|\tilde{\mathbf{B}}\|_*}\right) + o(\sqrt{\epsilon}) & w_i = 0, \end{cases} \end{aligned} \tag{5.86}$$

as $\epsilon \to 0$, whose dominant term depends on x^m, y^m only through r_i^*.

By comparison, $\tilde{W}_i$ satisfies the asymptotic Markov structure

$$\tilde{W}_i \leftrightarrow Z_i \leftrightarrow R_i^* \leftrightarrow (X^m, Y^m),$$

in the limit $\tilde{\epsilon}, \epsilon \to 0$, where

$$Z_i \triangleq U_i + V_i. \tag{5.87}$$

In particular, we have the following result, a proof of which is provided in Appendix C.13.

Corollary 5.25. For $\tilde{W}_i$ as defined in (5.80),

$$\begin{aligned} &P_{\tilde{W}_i|Z_i,X^m,Y^m}(\tilde{w}_i|z_i,x^m,y^m) \\ &\qquad = o(\tilde{\epsilon}\sqrt{\epsilon}) + \begin{cases} (1+\operatorname{sgn}(\tilde{w}_i\, z_i)\,\sqrt{\sigma_i})/2 & z_i = \pm 2 \\ 0 & z_i = 0, \end{cases} \end{aligned} \tag{5.88a}$$

as $\tilde{\epsilon}, \epsilon \to 0$, whose dominant term depends on x^m, y^m (and thus r_i^*) only through z_i, and

$$\begin{aligned} &P_{Z_i|X^m,Y^m}(z_i|x^m,y^m) \\ &\qquad = o(\tilde{\epsilon}) + \begin{cases} (1+\tilde{\epsilon}m\ \operatorname{sgn}(z_i)\, r_i^*)/4 & z_i = \pm 2 \\ 1/2 & z_i = 0, \end{cases} \end{aligned} \tag{5.88b}$$

as $\tilde{\epsilon} \to 0$, whose dominant term depends on x^m, y^m only through r_i^*.

5.9 Universal Features and Gács-Körner Common Information

In contrast to the formulation of Wyner, a notion of common information better suited to some applications is that introduced by Gács and Körner [89], and further developed by Witsenhausen [277]. Recall that with functions $f\colon \mathcal{X} \to \mathcal{W}$ and $g\colon \mathcal{Y} \to \mathcal{W}$, the Gács-Körner common information can be expressed in the form[14]

$$\bar{C}(X,Y) \triangleq \max_{\substack{f,g:\\ \mathbb{P}(f(X)=g(Y))=1}} H(f(X)), \tag{5.89}$$

[14] We use $H(Z)$ to denote the entropy of a random variable Z.

and satisfies $\bar{C}(X, Y) \leq I(X; Y)$, in contrast to Wyner common information. Moreover, as shown in [4], [5] (and, additionally, [141]), an equivalent characterization in the form of a dual to (5.50) is

$$\bar{C}(X, Y) = \max_{\substack{P_{W|X,Y}: \\ W \leftrightarrow X \leftrightarrow Y, \\ X \leftrightarrow Y \leftrightarrow W}} I(W; X, Y). \tag{5.90}$$

As established in [277], Gács-Körner common information can be expressed in terms of universal features rather simply, in contrast to Wyner common information. To see this, the following lemma is useful; a proof is provided in Appendix C.14.

Lemma 5.26. For given $k \in \{1, \ldots, K-1\}$, P_X, P_Y, f_*^k, and g_*^k, consider the family of $P_{X,Y}$ of the form (2.15) satisfying (2.16) and having $\sigma_i = 0$ for $k < i \leq K - 1$. Then[15]

$$\max_{(\sigma_1, \ldots, \sigma_k)} I(X; Y) = H(f_*^k(X)) = \sum_{x \in \mathcal{X}} P_X(x) \log\left(1 + \sum_{i=1}^{k} f_i^*(x)^2\right), \tag{5.91}$$

and the maximum is achieved when $\sigma_1 = \cdots = \sigma_k = 1$.

In particular, use of this lemma yields the following result, whose proof is provided in Appendix C.15.

Proposition 5.27. In terms of the decomposition (2.15), the Gács-Körner common information takes the form

$$\bar{C}(X, Y) = \begin{cases} \sum_{x \in \mathcal{X}} P_X(x) \log\left(1 + \sum_{i=1}^{k} f_i^*(x)^2\right) = H(f_*^k(X)) & \sigma_1 = 1 \\ 0 & \text{otherwise}, \end{cases} \tag{5.92}$$

where $k \in \{1, \ldots, K - 1\}$ is defined via the property that $\sigma_i = 1$ for $i \leq k$ and $\sigma_i < 1$ for $i > k$.

[15] By symmetry, it further follows that

$$\max_{(\sigma_1, \ldots, \sigma_k)} I(X; Y) = H\big(g_*^k(Y)\big) = \sum_{y \in \mathcal{Y}} P_Y(y) \log\left(1 + \sum_{i=1}^{k} g_i^*(y)^2\right).$$

As (5.92) reflects, maximizing f and g in (5.89) are the maximal correlation features when the maximal correlation is unity, in which case

$$\mathbb{P}(f_*^k(X) = g_*^k(Y)) = 1,$$

and the corresponding optimizing alphabet is

$$\mathcal{W} = f_*^k(\mathcal{X}) = g_*^k(\mathcal{Y}) \subset \mathbb{R}^k.$$

Distributions for which $\bar{C}(X, Y) \neq 0$ include those of Examples 2.7, 2.8, 2.9, and 2.10, for instance. Evidently, $\bar{C}(X, Y) \neq 0$ only for distributions $P_{X,Y}$ with very special structure, and in particular $\bar{C}(X, Y) = 0$ when X and Y are weakly dependent.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

6

Learning Modal Decompositions

Having now characterized and interpreted universal features in a variety of complementary ways, in this section we now turn our attention to the problem of their estimation from training data. Indeed, since the universal features developed in Section 5 are naturally expressed in terms of the modal decomposition (2.15) of the joint distribution $P_{X,Y}$, efficient learning of this decomposition from data is key to the practical applicability of these features. We discuss aspects of such issues in this section.

A suitable development arises out of orthogonal iteration method of computing an SVD, which itself arises out of the variational characterization of the SVD. In particular, Section 6.1 shows that the statistical interpretation of orthogonal iteration takes the form of what is referred to as the alternating conditional expectations (ACE) algorithm introduced by Breiman and Friedman [38]. In turn, Section 6.2 develops aspects of the sample complexity of feature recovery, in partial support of the empirical observation that in practice the dominant modes can typically be recovered with comparatively less training data.

We begin by more fully describing the scenario of interest. To this point in our development, for pedagogical reasons our model has been

The version of record is available at: http://dx.doi.org/10.1561/0100000107

that $P_{X,Y}$ is known, by which we mean practically that it has been reliably estimated from some training phase. When this is the case, the problem of computing the modal decomposition is simply one of computing the SVD of the associated CDM $\tilde{\mathbf{B}}$. However, in practice, of course, learning $P_{X,Y}$ from samples is an important aspect of the overall feature selection process in the inference pipeline.

From this perspective, there is a need to understand both the computational and sample complexity of universal feature recovery. Accordingly, it will be convenient in our exposition to first consider the modal computation from known $P_{X,Y}$, from which some of the computational complexity issues can be appreciated, then use the resulting foundation to address the problem of learning the modal decomposition from samples, through which sample complexity behavior can be examined.

6.1 Computing the Modal Decomposition

Given $P_{X,Y}$, computation of the SVD of $\tilde{\mathbf{B}}$ is a straightforward exercise in numerical linear algebra. In particular, from $P_{X,Y}$ we compute marginals P_X and P_Y, then construct $\tilde{\mathbf{B}}$ via (2.29), then apply any of many well-established numerical methods for computing the SVD of a matrix—see, e.g., [97], [237], [264]. However, in this section we further develop an interpretation of the resulting computation in the context of probabilistic analysis that will be insightful in the sequel. We emphasize at the outset that the results of this subsection are largely not new, but rather establish the viewpoints and interpretations we need for our subsequent development.

6.1.1 Orthogonal Iteration

Among the oldest and simplest approaches to the computation of the principal singular value and vector of a matrix is referred to as *power iteration* or the *power method* [97], [237], [264]. Moreover, power iteration can be used in a sequential manner to recover any number of the largest singular values and their corresponding singular vectors. However, when the first $1 < k \leq K - 1$ dominant modes are desired, it is more efficient

The version of record is available at: http://dx.doi.org/10.1561/0100000107

to compute them in parallel via a generalization of power iteration. The most direct generalization is referred to as *orthogonal iteration* [97, Section 7.3.2].[1] This algorithm has a corresponding relation to the generalized variational characterizations of the SVD we have used throughout our analysis.

To implement orthogonal iteration, we start with some $|\mathcal{X}| \times k$ matrix $\bar{\mathbf{\Psi}}_{(k)}^{X,0}$, which is typically chosen at random, and then execute the following iterative procedure:

1. Set $l = 0$.

2. Orthogonalize $\bar{\mathbf{\Psi}}_{(k)}^{X,l}$ to obtain $\hat{\mathbf{\Psi}}_{(k)}^{X,l}$ via the (thin or reduced) QR decomposition [97]
$$\bar{\mathbf{\Psi}}_{(k)}^{X,l} = \hat{\mathbf{\Psi}}_{(k)}^{X,l}\,\mathbf{\Theta}_{(k)}^{X,l}, \tag{6.1}$$
in which $\mathbf{\Theta}_{(k)}^{X,l}$ is a $k \times k$ upper triangular matrix.

3. Compute
$$\bar{\mathbf{\Psi}}_{(k)}^{Y,l} = \tilde{\mathbf{B}}\,\hat{\mathbf{\Psi}}_{(k)}^{X,l}, \tag{6.2}$$
then orthogonalize to obtain $\hat{\mathbf{\Psi}}_{(k)}^{Y,l}$ via the (thin or reduced) QR decomposition
$$\bar{\mathbf{\Psi}}_{(k)}^{Y,l} = \hat{\mathbf{\Psi}}_{(k)}^{Y,l}\,\mathbf{\Theta}_{(k)}^{Y,l}, \tag{6.3}$$
in which $\mathbf{\Theta}_{(k)}^{Y,l}$ is a $k \times k$ upper triangular matrix.

4. Produce the update
$$\bar{\mathbf{\Psi}}_{(k)}^{X,l+1} = \tilde{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{\Psi}}_{(k)}^{Y,l}. \tag{6.4}$$

5. Increment l and return to Step 2.

The QR decompositions employed in the orthogonalizations in this algorithm can be directly implemented using, e.g., the Gram-Schmidt procedure. However, numerical stability can be improved in practice through the use of Householder transformations [97].

[1] A refinement of orthogonal iteration referred to as *QR iteration* [97, Section 7.3.3] forms the basis of most practical implementations, and can be used in conjunction with various acceleration techniques.

The convergence of this procedure depends on the properties of

$$\mathbf{A} = \big(\boldsymbol{\Psi}^{X}_{(k)}\big)^{\mathrm{T}} \bar{\boldsymbol{\Psi}}^{X,0}_{(k)}, \tag{6.5}$$

where $\boldsymbol{\Psi}^{X}_{(k)}$ is the matrix of dominant singular vectors defined in (3.15). In particular, using $a_{i,j}$ to denote the (i,j)th entry of $\mathbf{A}$, when $\sigma_1, \dots, \sigma_{k+1}$ are unique and there exist distinct $j_1, \dots, j_k$ such that $a_{i,j_i} \neq 0$ for each i, then we obtain, as $l \to \infty$, [97, Theorem 7.3.1]

$$\begin{aligned}
\hat{\boldsymbol{\Psi}}^{X,l}_{(k)} &\to \boldsymbol{\Psi}^{X}_{(k)} \\
\hat{\boldsymbol{\Psi}}^{Y,l}_{(k)} &\to \boldsymbol{\Psi}^{Y}_{(k)} \\
\big(\hat{\boldsymbol{\Psi}}^{Y,l}_{(k)}\big)^{\mathrm{T}} \tilde{\mathbf{B}}\, \hat{\boldsymbol{\Psi}}^{X,l}_{(k)} &\to \boldsymbol{\Sigma}_{(k)},
\end{aligned}$$

with $\boldsymbol{\Sigma}_{(k)}$ as defined following (3.16). Moreover, convergence is exponentially fast.[2] When $\sigma_1, \dots, \sigma_k$ are not distinct (but still $\sigma_k > \sigma_{k+1}$), natural generalizations of these results are obtained [97, Theorem 7.3.1].

6.1.2 Statistical Interpretation as the ACE Algorithm

The use of orthogonal iteration to compute the dominant modes in (2.15) has a direct statistical interpretation that corresponds to what is referred to as the alternating conditional expectations (ACE) algorithm [38], [270].

In particular, choosing the correspondences

$$\begin{aligned}
\hat{f}_i(x) &\triangleq \frac{\hat{\psi}_i^{X,l}(x)}{\sqrt{P_X(x)}}, & \bar{f}_i(x) &\triangleq \frac{\bar{\psi}_i^{X,l}(x)}{\sqrt{P_X(x)}}, \\
\hat{g}_i(y) &\triangleq \frac{\hat{\psi}_i^{Y,l}(y)}{\sqrt{P_Y(y)}}, & \bar{g}_i(y) &\triangleq \frac{\bar{\psi}_i^{Y,l}(y)}{\sqrt{P_Y(y)}},
\end{aligned}$$

for $i = 1, \dots, k$, with

$$\begin{aligned}
\hat{\boldsymbol{\Psi}}^{X,l}_{(k)} &= \begin{bmatrix} \hat{\boldsymbol{\psi}}^{X,l}_1 & \cdots & \hat{\boldsymbol{\psi}}^{X,l}_k \end{bmatrix}, & \bar{\boldsymbol{\Psi}}^{X,l}_{(k)} &= \begin{bmatrix} \bar{\boldsymbol{\psi}}^{X,l}_1 & \cdots & \bar{\boldsymbol{\psi}}^{X,l}_k \end{bmatrix}, \\
\hat{\boldsymbol{\Psi}}^{Y,l}_{(k)} &= \begin{bmatrix} \hat{\boldsymbol{\psi}}^{Y,l}_1 & \cdots & \hat{\boldsymbol{\psi}}^{Y,l}_k \end{bmatrix}, & \bar{\boldsymbol{\Psi}}^{Y,l}_{(k)} &= \begin{bmatrix} \bar{\boldsymbol{\psi}}^{Y,l}_1 & \cdots & \bar{\boldsymbol{\psi}}^{Y,l}_k \end{bmatrix},
\end{aligned}$$

[2] However, it is worth emphasizing that the closer a pair of dominant singular values are to each other, the poorer the convergence rate.

we can rewrite the procedure of Section 6.1.1 in the form of Algorithm 1, which iteratively computes both the dominant k features (f_*^k, g_*^k) and

$$\sigma^{(k)} = \sum_{i=1}^{k} \sigma_i,$$

the Ky Fan k-norm of $\tilde{\mathbf{B}}$. As such, the convergence behavior follows immediately from the corresponding analysis of orthogonal iteration.[3]

To obtain Algorithm 1, we use that (6.1) and (6.3) can be equivalently expressed in the form[4]

$$\begin{aligned} \bar{f}^k(x) &= \big(\mathbf{\Theta}^X_{(k)}\big)^{\mathrm{T}} \hat{f}^k(x), \quad x \in \mathcal{X} \\ \bar{g}^k(y) &= \big(\mathbf{\Theta}^Y_{(k)}\big)^{\mathrm{T}} \hat{g}^k(y), \quad y \in \mathcal{Y}, \end{aligned}$$

via which we also obtain

$$\begin{aligned} \sum_{x\in\mathcal{X}} P_X(x)\, \bar{f}^k(x)\, \bar{f}^k(x)^{\mathrm{T}} &= (\bar{\mathbf{\Psi}}^X_{(k)})^{\mathrm{T}} \bar{\mathbf{\Psi}}^X_{(k)} = (\mathbf{\Theta}^X_{(k)})^{\mathrm{T}} \mathbf{\Theta}^X_{(k)} \\ \sum_{y\in\mathcal{Y}} P_Y(y)\, \bar{g}^k(y)\, \bar{g}^k(y)^{\mathrm{T}} &= (\bar{\mathbf{\Psi}}^Y_{(k)})^{\mathrm{T}} \bar{\mathbf{\Psi}}^Y_{(k)} = (\mathbf{\Theta}^Y_{(k)})^{\mathrm{T}} \mathbf{\Theta}^Y_{(k)}, \end{aligned}$$

since

$$(\hat{\mathbf{\Psi}}^X_{(k)})^{\mathrm{T}} \hat{\mathbf{\Psi}}^X_{(k)} = (\hat{\mathbf{\Psi}}^Y_{(k)})^{\mathrm{T}} \hat{\mathbf{\Psi}}^Y_{(k)} = \mathbf{I}.$$

Additionally, we use that, via (2.28), (6.2) and (6.4) can be equivalently expressed in the form

$$\begin{aligned} \bar{g}^k(y) &= \frac{1}{\sqrt{P_Y(y)}} \sum_{x\in\mathcal{X}} \tilde{B}(y,x) \sqrt{P_X(x)}\, \hat{f}^k(x) \\ &= \mathbb{E}\big[\hat{f}^k(X)\big|Y=y\big] - \mathbb{E}\big[\hat{f}^k(X)\big] \\ \bar{f}^k(x) &= \frac{1}{\sqrt{P_X(x)}} \sum_{x\in\mathcal{X}} \tilde{B}(y,x) \sqrt{P_Y(y)}\, \hat{g}^k(y) \\ &= \mathbb{E}\big[\hat{g}^k(Y)\big|X=x\big] - \mathbb{E}\big[\hat{g}^k(Y)\big]. \end{aligned}$$

[3] We emphasize that the Cholesky decompositions in Algorithm 1 are unique when the associated covariance matrices are full rank, which is the case when: 1) the singular values are distinct; and 2) the covariance matrix of the initialization is positive definite, with components correlated with each of the dominant features in the modal decomposition.

[4] For convenience, in this derivation, we drop the dependence on iteration (superscript l) from our the notation, consistent with the notation in Algorithm 1.

Algorithm 1 ACE Algorithm for Multiple Mode Computation

```
Require: Joint distribution $P_{X,Y}$, number of modes $k$
    1. Initialization: randomly choose $\bar{f}^k(x),\ \forall x \in \mathcal{X}$
    repeat
      2a. Center: $\bar{f}^k(x) \leftarrow \bar{f}^k(x) - \mathbb{E}[\bar{f}^k(X)], \quad \forall x \in \mathcal{X}$
      2b. Cholesky factor:
            $\mathbb{E}[\bar{f}^k(X)\,\bar{f}^k(X)^{\mathrm{T}}] = (\mathbf{\Theta}^X_{(k)})^{\mathrm{T}}\mathbf{\Theta}^X_{(k)}$
      2c. Whiten:
            $\hat{f}^k(x) \leftarrow (\mathbf{\Theta}^X_{(k)})^{-\mathrm{T}}\,\bar{f}^k(x),\ \forall x \in \mathcal{X}$
      2d. $\bar{g}^k(y) \leftarrow \mathbb{E}[\hat{f}^k(X)|Y = y],\ \forall y \in \mathcal{Y}$
      2e. Center: $\bar{g}^k(y) \leftarrow \bar{g}^k(y) - \mathbb{E}[\bar{g}^k(Y)], \quad \forall y \in \mathcal{Y}$
      2f. Cholesky factor:
            $\mathbb{E}[\bar{g}^k(Y)\,\bar{g}^k(Y)^{\mathrm{T}}] = (\mathbf{\Theta}^Y_{(k)})^{\mathrm{T}}\mathbf{\Theta}^Y_{(k)}$
      2g. Whiten:
            $\hat{g}^k(y) \leftarrow (\mathbf{\Theta}^Y_{(k)})^{-\mathrm{T}}\,\bar{g}^k(y),\ \forall y \in \mathcal{Y}$
      2h. $\bar{f}^k(x) \leftarrow \mathbb{E}[\hat{g}^k(Y)|X = x],\ \forall x \in \mathcal{X}$
      2i. $\hat{\sigma}^{(k)} \leftarrow \mathbb{E}[\bar{f}^k(X)^{\mathrm{T}}\hat{g}^k(Y)]$
    until $\hat{\sigma}^{(k)}$ stops increasing.
```

Note, too, that via standard Bayesian estimation theory, the conditional expectations in Algorithm 1 can be interpreted as minimum mean-square error (MMSE) estimates, and thus we can equivalently write the corresponding steps 2h and 2d as optimizations; specifically, we have, respectively, the variational characterizations

$$\bar{f}^k(\cdot) \leftarrow \operatorname*{arg\,min}_{f^k(\cdot)} \mathbb{E}\Big[\big\|f^k(X) - \hat{g}^k(Y)\big\|^2\Big] \tag{6.6a}$$

$$\bar{g}^k(\cdot) \leftarrow \operatorname*{arg\,min}_{g^k(\cdot)} \mathbb{E}\Big[\big\|\hat{f}^k(X) - g^k(Y)\big\|^2\Big], \tag{6.6b}$$

where we note that these optimizations can, themselves, be carried out by an iterative procedure. Such implementations can be attractive when there are constraints on $f_*^k(\cdot)$ and $g_*^k(\cdot)$ based on, e.g., domain knowledge and/or other considerations.

6.2 Estimating the Modal Decomposition from Data

When the joint distribution $P_{X,Y}$ is unknown, as is common in applications, but we have (labeled) training data

$$\mathcal{T} \triangleq \{(x_1, y_1), \dots, (x_n, y_n)\}, \tag{6.7}$$

drawn i.i.d. from $P_{X,Y}$, we can replace $P_{X,Y}$ in Algorithm 1 with the empirical distribution

$$\hat{P}_{X,Y}(x, y) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x=x_i}\, \mathbb{1}_{y=y_i} \tag{6.8}$$

to generate an estimate of the modal decomposition. In this case, the expectations in Algorithm 1 are all with respect to the corresponding empirical distributions. In particular, those in steps 2a-b and 2e-f are with respect to, respectively,

$$\hat{P}_X(x) \triangleq \sum_y \hat{P}_{X,Y}(x, y) \quad \text{and} \quad \hat{P}_Y(y) \triangleq \sum_x \hat{P}_{X,Y}(x, y),$$

while those in steps 2d and 2h are with respect to, respectively,

$$\hat{P}_{X|Y}(x|y) \triangleq \frac{\hat{P}_{X,Y}(x,y)}{\hat{P}_Y(y)} \quad \text{and} \quad \hat{P}_{Y|X}(y|x) \triangleq \frac{\hat{P}_{X,Y}(x,y)}{\hat{P}_X(x)},$$

and that in step 2i is with respect to $\hat{P}_{X,Y}(x,y)$.

Evidently, in this version of Algorithm 1, the computational complexity scales with the number of training samples n. There are a variety of ways to reduce this complexity in practice. For example, among other possibilities, in each basic iteration we can choose to operate on only a (comparatively small) randomly chosen subset of the training data and exploit bootstrapping techniques.

It is also worth emphasizing that in some scenarios we may have both *labeled* and *unlabeled* training data available, the latter of which correspond to samples $x_1, \dots, x_{n'}$ and $y_1, \dots, y_{n'}$, drawn i.i.d. from P_X and P_Y, respectively. While labeled data is typically expensive to obtain, since the labeling process often involves a significant amount of manual labor, unlabeled data is comparatively inexpensive to obtain, since no correspondences need be identified. As such, it is often possible to accurately estimate P_X and P_Y. In such scenarios, the corresponding version of Algorithm 1 replaces $P_{X,Y}$ with an estimate based the labeled training data subject the marginal constraints P_X and P_Y, which can be constructed in a variety of ways.[5]

In the sequel, we quantify the accuracy of the modal decomposition when estimated from data. In light of the preceding discussion, for this analysis, $\hat{f}_i^*$, $\hat{g}_i^*$, and $\hat{\sigma}_i$ for $i = 1, \dots, K$ are defined via the modal decomposition

$$\hat{P}_{X,Y}(x,y) = P_X(x)\, P_Y(y) \left[1 + \sum_{i=1}^{K} \hat{\sigma}_i\, \hat{f}_i^*(x)\, \hat{g}_i^*(y) \right], \tag{6.9}$$

where $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_K \geq 0$ and $\mathbb{E}[\hat{f}_i^*(X)\, \hat{f}_j^*(X)] = \mathbb{E}[\hat{g}_i^*(Y)\, \hat{g}_j^*(Y)] = \mathbb{1}_{i=j}$ for $i, j \in \{1, \dots, K\}$. We emphasize that in (6.9) we only have $\hat{\sigma}_K = 0$ and $\mathbb{E}[\hat{f}_i(X)] = \mathbb{E}[\hat{g}_i(Y)] = 0$ for $i \in \{1, \dots, K\}$, when $\hat{P}_X =$

[5] Such estimation procedures for $P_{X,Y}$ are often referred to as *raking* and have a long history. A common approach is to use *iterative scaling* (*iterative proportional fitting*) [71], [134]. See also, e.g., [95], [104], [275].

P_X and $\hat{P}_Y = P_Y$.[6] Accordingly, in analysis, we will frequently find it convenient to use the equivalent zero-mean features

$$\check{f}_i^*(x) \triangleq \hat{f}_i^*(x) - \mathbb{E}\big[\hat{f}_i^*(X)\big] \tag{6.10a}$$

$$\check{g}_i^*(y) \triangleq \hat{g}_i^*(y) - \mathbb{E}\big[\hat{g}_i^*(Y)\big], \tag{6.10b}$$

for $i = 1, \dots, K$.

The decomposition (6.9) corresponds to the singular value decomposition of the quasi-CDM $\hat{\mathbf{B}}$ whose (y, x)th entry is

$$\hat{B}(x, y) \triangleq \frac{\hat{P}_{X,Y}(x, y) - P_X(x)\, P_Y(y)}{\sqrt{P_X(x)}\sqrt{P_Y(y)}}, \tag{6.11}$$

i.e.,

$$\hat{\mathbf{B}} \triangleq \Big[\sqrt{\mathbf{P}_Y}\Big]^{-1} \Big[\hat{\mathbf{P}}_{Y,X} - \mathbf{P}_Y\, \mathbf{P}_X\Big] \Big[\sqrt{\mathbf{P}_X}\Big]^{-1} \tag{6.12}$$

$$= \sum_{i=1}^{K} \hat{\sigma}_i \hat{\boldsymbol{\psi}}_i^Y \big(\hat{\boldsymbol{\psi}}_i^X\big)^{\mathrm{T}}, \tag{6.13}$$

where the singular vectors in (6.13) have elements

$$\hat{\psi}_i^X(x) \triangleq \sqrt{P_X(x)}\, \hat{f}_i^*(x), \qquad x \in \mathcal{X}, \tag{6.14a}$$

$$\hat{\psi}_i^Y(y) \triangleq \sqrt{P_Y(y)}\, \hat{g}_i^*(y), \qquad y \in \mathcal{Y}, \tag{6.14b}$$

for $i = 1, \dots, K$.

There are several aspects of the modal decomposition estimation whose sample complexity is of interest, which we now address.

6.2.1 Sample Complexity of Maximal Correlation Estimates

In this section, we determine the number of samples required to obtain accurate estimates $\hat{\sigma}_1, \dots, \hat{\sigma}_k$ of $\sigma_1, \dots, \sigma_k$, for $k \in \{1, \dots, K\}$.[7] Specifically, we focus on the measure

$$\mu_1^k(P_{X,Y}, \hat{P}_{X,Y}) \triangleq \sum_{i=1}^{k} |\hat{\sigma}_i - \sigma_i|, \tag{6.15}$$

and related quantities.

We begin with the following tail probability bound, a proof of which is provided in Appendix D.1.

[6] As such these properties effectively hold when $n \gg \max\{|\mathcal{X}|, |\mathcal{Y}|\}$.

[7] Although the case $k = K$ is typically less of interest (since $\sigma_K = 0$), we include it for completeness. For this case, $f_K^*(X)$ and $g_K^*(Y)$ can be chosen freely subject to the constraint that they are uncorrelated with $f_*^{K-1}(X)$ and $g_*^{K-1}(Y)$, respectively.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Proposition 6.1. For $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$, let $p_0 > 0$ be such that

$$P_X(x) \geq p_0 \quad \text{and} \quad P_Y(y) \geq p_0, \qquad \text{all } x \in \mathcal{X},\, y \in \mathcal{Y}. \tag{6.16}$$

Then for $k \in \{1, \dots, K\}$ and $0 \leq \delta \leq \sqrt{k/2}/p_0$,

$$\mathbb{P}\left(\sum_{i=1}^{k} |\hat{\sigma}_i - \sigma_i| \geq \delta\right) \leq \exp\left\{\frac{1}{4} - \frac{p_0^2\,\delta^2 n}{8k}\right\}, \tag{6.17}$$

where $\hat{\sigma}_i$ for $i = 1, \dots, K$ are defined via (6.9) with $\hat{P}_{X,Y}$ denoting the empirical distribution based on n training samples.

A key consequence of Proposition 6.1 is the following corollary.

Corollary 6.2. Suppose $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ is such that (6.16) is satisfied for some $p_0 > 0$. Then for $k \in \{1, \dots, K\}$ and n sufficiently large that $n \geq 16 \log(4kn)$,

$$\mathbb{E}\left[\left(\sum_{i=1}^{k} |\hat{\sigma}_i - \sigma_i|\right)^2\right] \leq \frac{6k + 8k\log(nk)}{p_0^2 n}, \tag{6.18}$$

where $\hat{\sigma}_i$ for $i = 1, \dots, K$ are defined via (6.9) with $\hat{P}_{X,Y}$ denoting the empirical distribution based on n training samples.

The proof of Corollary 6.2, provided in Appendix D.2, makes use of the following simple lemma, which is a straightforward exercise in calculus.

Lemma 6.3. Given $a, b > 0$, the convex function $\varphi_{a,b} \colon \mathbb{R} \to \mathbb{R}$ defined via

$$\varphi_{a,b}(\omega) \triangleq \omega + a\mathrm{e}^{-b\omega}, \tag{6.19}$$

has its minimum at

$$\omega_* \triangleq \operatorname*{arg\,min}_{\omega} \varphi_{a,b}(\omega) = \frac{1}{b}\log(ab), \tag{6.20}$$

where it takes value

$$\min_{\omega} \varphi_{a,b}(\omega) = \varphi_{a,b}(\omega_*) = \frac{1 + \log(ab)}{b}. \tag{6.21}$$

Additional consequences of Proposition 6.1 and Corollary 6.2 are that

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

and

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

respectively, which follow from the triangle inequality; specifically,

$$\Big|\|\mathbf{A}_1\|_{(k)} - \|\mathbf{A}_2\|_{(k)}\Big| = \left|\sum_{i=1}^{k}(\sigma_i(\mathbf{A}_1) - \sigma_i(\mathbf{A}_2))\right| \leq \sum_{i=1}^{k}|\sigma_i(\mathbf{A}_1) - \sigma_i(\mathbf{A}_2)|, \tag{6.22}$$

for any $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{k_1 \times k_2}$ and $k \in \{1, \min\{k_1, k_2\}\}$, with $\sigma_1(\cdot) \geq \cdots \geq \sigma_{\min\{k_1,k_2\}}(\cdot)$ denoting the ordered singular values of its (matrix) argument.[8]

And still further consequences of Proposition 6.1 and Corollary 6.2 are that

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

and

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

respectively, which follow from the standard norm inequality

$$\|a^k\| \leq \|a^k\|_1 \triangleq \sum_{i=1} |a_i|, \qquad \text{any } k \text{ and } a^k. \tag{6.23}$$

Finally, for ϵ-dependent X and Y, variables $X^{(k)}, Y^{(k)}$ defined via (4.30a) have mutual information $I(X^{(k)}; Y^{(k)})$ given by (4.30). Accordingly, a natural estimate of this mutual information is

$$\hat{I}(X^{(k)}; Y^{(k)}) \triangleq \frac{1}{2}\sum_{i=1}^{k}\hat{\sigma}_i^2,$$

[8] Note that (6.22), in turn, means that, more generally, Lemma D.2 in Appendix D.1 also quantifies the stability of Ky Fan k-norms.

for which the error is

$$\hat{I}(X^{(k)};Y^{(k)}) - I(X^{(k)};Y^{(k)}) = \frac{1}{2}\sum_{i=1}^{k}(\hat{\sigma}_i^2 - \hat{\sigma}_i^2) + o(\epsilon^2), \quad \epsilon \to 0. \tag{6.24}$$

The results of Proposition 6.1 and Corollary 6.2 can be used to bound the error (6.24); specifically, we have the following corollary, whose proof is provided in Appendix D.3.

Corollary 6.4. Suppose $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ is such that (6.16) is satisfied for some $p_0 > 0$, and let $\hat{\sigma}_i$ for $i = 1, \ldots, K$ be defined via (6.9) with $\hat{P}_{X,Y}$ denoting the empirical distribution based on n training samples. Then for any $k \in \{1, \ldots, K\}$ and $0 \leq \delta \leq \sqrt{k/2}/p_0^2$,

$$\mathbb{P}\left(\left|\frac{1}{2}\sum_{i=1}^{k}(\hat{\sigma}_i^2 - \sigma_i^2)\right| \geq \delta\right) \leq \exp\left\{\frac{1}{4} - \frac{p_0^4\,\delta^2 n}{8k}\right\}, \tag{6.25}$$

and, for n such that $n \geq 16\log(4kn)$,

$$\mathbb{E}\left[\left|\frac{1}{2}\sum_{i=1}^{k}(\hat{\sigma}_i^2 - \sigma_i^2)\right|^2\right] \leq \frac{6k + 8k\log(nk)}{p_0^4 n}. \tag{6.26}$$

More generally, it should be emphasized that, as the proofs of Proposition 6.1, Corollary 6.2, and Corollary 6.4 reveal, their results hold not only for the k dominant singular values $\hat{\sigma}_i$ and σ_i, but for arbitrary (corresponding) subsets of k singular values.

It is also worth noting that Proposition 6.1, Corollary 6.2, and Corollary 6.4 all suggest that when some value of x (or some value of y) occurs with low probability, then estimating the singular values can require a correspondingly large amount of data. An instance of this is Example 2.7 when δ is small.

6.2.2 Sample Complexity of Feature Estimates

In this section, we determine the number of samples required to obtain accurate estimates

$$\check{f}_*^k = (\check{f}_1^*, \ldots, \check{f}_k^*) \quad \text{and} \quad \check{g}_*^k = (\check{g}_1^*, \ldots, \check{g}_k^*)$$

of the features f_*^k and g_*^k, respectively, for $k \in \{1, \ldots, K-1\}$. Our development focuses on measuring the accuracy of these estimates by the extent to which they preserve as much of the mutual information between X and Y as possible, in the local sense, corresponding to $\sigma_1^2 + \cdots + \sigma_k^2$. Specifically, we measure this via[9]

$$\mu_2^k(P_{X,Y}, \hat{P}_{X,Y}) \triangleq \mathbb{E}_{P_Y}\Big[\big\|\mathbb{E}_{P_{X|Y}}[f_*^k(X)]\big\|^2 - \big\|\mathbb{E}_{P_{X|Y}}[\check{f}_*^k(X)]\big\|^2\Big]. \tag{6.27}$$

To facilitate interpretation of the measure (6.27), note that since

$$\|\tilde{\mathbf{B}}\,\boldsymbol{\Xi}^X\|_{\mathrm{F}}^2 = \mathbb{E}_{P_Y}\Big[\big\|\mathbb{E}_{P_{X|Y}}[f^k(X)]\big\|^2\Big], \tag{6.28}$$

with $\boldsymbol{\Xi}^X$ denoting the collection of feature vectors associated with f^k as defined in (3.12a), we have

$$\max_{f^k \in \mathcal{F}_k} \mathbb{E}_{P_Y}\Big[\big\|\mathbb{E}_{P_{X|Y}}[f^k(X)]\big\|^2\Big] = \mathbb{E}_{P_Y}\Big[\big\|\mathbb{E}_{P_{X|Y}}[f_*^k(X)]\big\|^2\Big] = \sum_{i=1}^{k} \sigma_i^2,$$

where $\mathcal{F}_k$ is as defined in (3.6c).

We begin with the following tail probability bound, a proof of which is provided in Appendix D.4.

Proposition 6.5. Let $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ be such that (6.16) is satisfied for some $p_0 > 0$. Then for $k \in \{1, \dots, K\}$ and $0 \leq \delta \leq 4k$,[10]

$$\mathbb{P}_{\check{f}_*^k}\Big(\mathbb{E}_{P_Y}\Big[\big\|\mathbb{E}_{P_{X|Y}}[f_*^k(X)]\big\|^2 - \big\|\mathbb{E}_{P_{X|Y}}[\check{f}_*^k(X)]\big\|^2\Big] \geq \delta\Big) \\ \leq (|\mathcal{X}| + |\mathcal{Y}|)\exp\left\{-\frac{p_0\,\delta^2 n}{64k^2}\right\}, \tag{6.29}$$

where $\check{f}_i^*$ for $i = 1, \dots, K-1$ are defined via (6.10a) and (6.9), with $\hat{P}_{X,Y}$ denoting the empirical distribution based on n training samples.

Note that by symmetry, it also follows immediately from Proposition 6.5 that

$$\mathbb{P}_{\check{g}_*^k}\Big(\mathbb{E}_{P_X}\Big[\big\|\mathbb{E}_{P_{Y|X}}[g_*^k(Y)]\big\|^2 - \big\|\mathbb{E}_{P_{Y|X}}[\check{g}_*^k(Y)]\big\|^2\Big] \geq \delta\Big) \\ \leq (|\mathcal{X}| + |\mathcal{Y}|)\exp\left\{-\frac{p_0\,\delta^2 n}{64k^2}\right\}, \tag{6.30}$$

[9] To avoid unnecessarily cumbersome expressions, we have left implicit the conditioning on $\check{f}_*^k(\cdot)$ in the expectation in (6.27).

[10] With slight abuse of notation, we use $\mathbb{P}_{\check{f}_*^k}(\cdot)$ to denote probability with respect to the distribution governing the random map $\check{f}_*^k(\cdot)$.

where, analogously, $\check{g}_i^*$ for $i = 1, \dots, K-1$ are defined via (6.10b) and (6.9).

A key consequence of Proposition 6.5 is the following corollary, whose proof, provided in Appendix D.5, makes use of Lemma 6.3.

Corollary 6.6. Let $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ be such that (6.16) is satisfied for some $p_0 > 0$. Then for $k \in \{1, \dots, K\}$ and n sufficiently large that

$$\frac{p_0 n}{64} \geq \frac{1}{(|\mathcal{X}| + |\mathcal{Y}|)} \tag{6.31a}$$

and

$$\frac{p_0 n}{4} \geq \log\biggl(\frac{p_0 n}{64}(|\mathcal{X}| + |\mathcal{Y}|)\biggr), \tag{6.31b}$$

we have

$$\mathbb{E}_{\check{f}_*^k}\Biggl[\biggl(\mathbb{E}_{P_Y}\Bigl[\bigl\|\mathbb{E}_{P_{X|Y}}\bigl[f_*^k(X)\bigr]\bigr\|^2 - \bigl\|\mathbb{E}_{P_{X|Y}}\bigl[\check{f}_*^k(X)\bigr]\bigr\|^2\Bigr]\biggr)^2\Biggr] \\ \leq \frac{64k^2\Bigl(\log\bigl[p_0 n(|\mathcal{X}| + |\mathcal{Y}|)\bigr] - 3\Bigr)}{p_0 n}, \tag{6.32}$$

where $\check{f}_i^*$ for $i = 1, \dots, K-1$ are defined via (6.10a) and (6.9), with $\hat{P}_{X,Y}$ denoting the empirical distribution based on n training samples, and where (with slight abuse of notation) we use $\mathbb{E}_{\check{f}_*^k}[\cdot]$ to denote expectation with respect to the distribution governing the random map $\check{f}_*^k$.

As with the case of maximal correlation estimates, we note that Proposition 6.5 and Corollary 6.6 suggest that a large amount of data is required to estimate the features when some value of x (or some value of y) occurs with low probability. An instance of this is, again, Example 2.7 when δ is small.

6.2.3 Complementary Sample Complexity Bounds

In Proposition 6.1, we bound the sample complexity of maximal correlation estimates via a Frobenius norm bound, while in Proposition 6.5, we bound the sample complexity of feature estimates via a spectral norm bound. However, we may interchange these analyses, using spectral norm bounds to analyze sample complexity of maximal correlation estimates, and Frobenius norm bounds to analyze the sample complexity of feature estimates.

In this way, we obtain complementary results. In particular, we have the following alternative bound on the sample complexity of maximal correlation estimates, a proof of which is provided in Appendix D.6.

Proposition 6.7. For $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$, let $p_0 > 0$ be such that (6.16) is satisfied. Then for $k \in \{1, \dots, K\}$ and $0 \leq \delta \leq k$,

$$\mathbb{P}\left(\sum_{i=1}^{k} |\hat{\sigma}_i - \sigma_i| \geq \delta\right) \leq (|\mathcal{X}| + |\mathcal{Y}|) \exp\left\{-\frac{p_0\,\delta^2 n}{4k^2}\right\}, \tag{6.33}$$

where $\hat{\sigma}_i$ for $i = 1, \dots, K$ are defined via (6.9) with $\hat{P}_{X,Y}$ denoting the empirical distribution based on n training samples.

Moreover, Proposition 6.7 can be used to obtain an alternative version of Corollary 6.2.

Likewise, we have the following alternative bound on the sample complexity of feature estimates, a proof of which is provided in Appendix D.7.

Proposition 6.8. Let $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ be such that (6.16) is satisfied for some $p_0 > 0$. Then for $k \in \{1, \dots, K\}$ and $0 \leq \delta \leq (4/p_0)/\sqrt{k/2}$,

$$\begin{aligned}\mathbb{P}_{\check{f}^k_*}\Big(\mathbb{E}_{P_Y}\Big[\big\|\mathbb{E}_{P_{X|Y}}[f^k_*(X)]\big\|^2 - \big\|\mathbb{E}_{P_{X|Y}}[\check{f}^k_*(X)]\big\|^2\Big] \geq \delta\Big)&\\ \leq \exp\left\{\frac{1}{4} - \frac{p_0^2\,\delta^2 n}{128k}\right\},&\end{aligned} \tag{6.34}$$

where $\check{f}^*_i$ for $i = 1, \dots, K-1$ are defined via (6.10a) and (6.9), with $\hat{P}_{X,Y}$ denoting the empirical distribution based on n training samples.

Proposition 6.8 can similarly be used to obtain an alternative version of Corollary 6.6.

Comparing (6.33) to (6.17) for maximum correlation estimates, and (6.34) to (6.29) for feature estimates, we see that the alternative bounds depend on the parameters in different ways, and apply in different regimes. As such, each may be better than the other in different regimes.

6.2.4 A Related Measure of Feature Quality

A natural measure of feature quality closely related to that defined in (6.27) is

$$\mu_{2'}^{k}(P_{X,Y},\hat{P}_{X,Y}) \triangleq \left\|\mathbb{E}\big[f_*^k(X)\,g_*^k(Y)^{\mathrm{T}}\big]-\mathbb{E}\big[\check{f}_*^k(X)\,\check{g}_*^k(Y)^{\mathrm{T}}\big]\right\|_{\mathrm{F}}. \tag{6.35}$$

See Appendix D.8 for an analysis of (6.35), which establishes that sample complexity bounds very similar to those for (6.27) apply.

6.2.5 Sample Complexity Error Exponent Analysis

Further sample complexity results can be obtained in the limit $n \to \infty$ via large deviations analysis, complementing the results of Sections 6.2.1–6.2.3.

In this analysis, for a given $\hat{P}_{X,Y}$ we focus on the empirical DTM $\hat{\mathbf{B}}$ whose (y,x)th entry is[11]

$$\hat{B}(x,y) \triangleq \frac{\hat{P}_{X,Y}(x,y)}{\sqrt{\hat{P}_X(x)}\sqrt{\hat{P}_Y(y)}}, \tag{6.36}$$

for which $\hat{f}_i^*$, $\hat{g}_i^*$, and $\hat{\sigma}_i$ for $i=1,\dots,K$ are defined via the modal decomposition

$$\hat{P}_{X,Y}(x,y)=\hat{P}_X(x)\,\hat{P}_Y(y)\left[1+\sum_{i=1}^{K-1}\hat{\sigma}_i\,\hat{f}_i^*(x)\,\hat{g}_i^*(y)\right], \tag{6.37}$$

where $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_{K-1} \geq 0$ and

$$\mathbb{E}_{\hat{P}_X}\big[\hat{f}_i^*(X)\,\hat{f}_j^*(X)\big]=\mathbb{E}_{\hat{P}_Y}\big[\hat{g}_i^*(Y)\,\hat{g}_j^*(Y)\big]=\mathbb{1}_{i=j},\quad i,j\in\{1,\dots,K-1\}.$$

Eq. (6.37) corresponds to the singular value decomposition

$$\hat{\mathbf{B}}=\sum_{i=0}^{K-1}\hat{\sigma}_i\hat{\boldsymbol{\psi}}_i^Y\,(\hat{\boldsymbol{\psi}}_i^X)^{\mathrm{T}}, \tag{6.38}$$

[11] By contrast, in the preceding sections the analysis focused on the quasi-CDM defined via (6.11), which uses true instead of empirical marginals and removes the zeroth mode, resulting in the decomposition (6.9). However, the re-use of notation is convenient. Also, more generally $\hat{B}(x,y) \triangleq 0$ for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$ such that $\hat{P}_X(x)=0$ or $\hat{P}_Y(y)=0$, consistent with (2.9).

where for $i = 1, \ldots, K-1$ the singular vectors in (6.38) are related to the feature vectors in (6.37) as in (6.14), and where [cf. (2.14)]

$$\hat{\sigma}_0 = 1, \qquad \hat{\boldsymbol{\psi}}_0^X(x) = \sqrt{\hat{P}_X(x)}, \qquad \hat{\boldsymbol{\psi}}_0^Y(y) = \sqrt{\hat{P}_Y(y)}.$$

Motivated by the analyses in the previous sections, we consider the following neighborhoods of $P_{X,Y}$:

$$\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y}) \triangleq \left\{\hat{P}_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}} \colon \|\hat{\mathbf{B}} - \mathbf{B}\|_{\mathrm{F}} \leq \delta\right\} \tag{6.39a}$$

$$\mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y}) \triangleq \left\{\hat{P}_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}} \colon \|\hat{\mathbf{B}} - \mathbf{B}\|_{\mathrm{s}} \leq \delta\right\} \tag{6.39b}$$

$$\mathcal{S}_\delta^{k}(P_{X,Y}) \triangleq \left\{\hat{P}_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}} \colon \left|\|\mathbf{B}\boldsymbol{\Psi}_{(k)}^X\|_{\mathrm{F}}^2 - \|\mathbf{B}\hat{\boldsymbol{\Psi}}_{(k)}^X\|_{\mathrm{F}}^2\right| \leq \delta\right\}, \quad k \in \{1, \ldots, K-1\}, \tag{6.39c}$$

for any $\delta > 0$, where $\hat{\mathbf{B}}$ is the DTM corresponding to $\hat{P}_{X,Y}$. We denote the k dominant singular vectors using $\hat{\boldsymbol{\psi}}_i^X$, and $\hat{\boldsymbol{\psi}}_i^Y$, and define

$$\hat{\boldsymbol{\Psi}}_{(k)}^X \triangleq \begin{bmatrix} \hat{\boldsymbol{\psi}}_1^X & \cdots & \hat{\boldsymbol{\psi}}_k^X \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Psi}}_{(k)}^Y \triangleq \begin{bmatrix} \hat{\boldsymbol{\psi}}_1^Y & \cdots & \hat{\boldsymbol{\psi}}_k^Y \end{bmatrix}. \tag{6.40}$$

Our main result relates the error exponents for (6.39c) to those of (6.39a) and (6.39b); a proof is provided in Appendix D.9.

Proposition 6.9. For any $P_{X,Y} \in \operatorname{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$, any $0 < \delta < B_{\min}(P_{X,Y})$ with

$$B_{\min}(P_{X,Y}) \triangleq \min_{x\in\mathcal{X}, y\in\mathcal{Y}} B(x,y) > 0, \tag{6.41}$$

and $\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y})$, $\mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y})$, and $\mathcal{S}_\delta^{k}(P_{X,Y})$ as defined in (6.39), we have

$$E\big(\mathcal{S}_{4\delta\sqrt{k}}^{k}(P_{X,Y})\big) \geq E_-\big(\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y})\big) = E_*\big(\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y})\big) \tag{6.42a}$$

$$E\big(\mathcal{S}_{4\delta k}^{k}(P_{X,Y})\big) \geq E_-\big(\mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y})\big) = E_*\big(\mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y})\big) \geq E_*\big(\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y})\big), \tag{6.42b}$$

for $k \in \{1, \ldots, K-1\}$, where for $\mathcal{S} \subseteq \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$,

$$E(\mathcal{S}) \triangleq -\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\hat{P}_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}} \backslash \mathcal{S}\Big) \tag{6.43}$$

$$E_*(\mathcal{S}) \triangleq -\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\hat{P}_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}} \backslash \mathcal{S}\Big) \tag{6.44}$$

$$E_-(\mathcal{S}) \triangleq \inf_{\hat{P}_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}} \backslash \mathcal{S}} D(\hat{P}_{X,Y} \| P_{X,Y}). \tag{6.45}$$

Proposition 6.9 quantifies, e.g., the relative difficulties of achieving small

$$\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \left(\hat{B}(x,y) - B(x,y)\right)^2,$$

which corresponds to (6.39a), small[12]

$$\max_{(f,g)\in\mathcal{F}_1\times\mathcal{G}_1} \Big(\mathbb{E}\left[f(X)\,g(Y)\right] - \mathbb{E}_{\hat{P}_{X,Y}}\left[f(X)\,g(Y)\right]\Big),$$

which corresponds to (6.39b) when $\hat{P}_X = P_X$ and $\hat{P}_Y = P_Y$ (and which is a good approximation for moderately large n), and small

$$\left|\mathbb{E}_{P_Y}\left[\left\|\mathbb{E}_{P_{X|Y}}\left[f_*^k(X)\right]\right\|^2 - \left\|\mathbb{E}_{P_{X|Y}}\left[\hat{f}_*^k(X)\right]\right\|^2\right]\right|,$$

which corresponds to (6.39c)—and which is closely related to achieving small

$$\left\|\mathbb{E}\left[f_*^k(X)\,g_*^k(Y)^{\mathrm{T}}\right] - \mathbb{E}\left[\hat{f}_*^k(X)\,\hat{g}_*^k(Y)^{\mathrm{T}}\right]\right\|_{\mathrm{F}},$$

according to the discussion of Section 6.2.4.

Finally, the following lemma provides a local characterization of related Chernoff exponents.[13] A proof is provided in Appendix D.10.

Lemma 6.10. For $P_Z \in \operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$ and every $h\colon \mathcal{Z} \to \mathbb{R}$ such that $\mathbb{E}[h(Z)] \neq 0$ and $\operatorname{var}[h(Z)] > 0$, and with $\hat{P}_Z$ denoting the empirical distribution formed from n i.i.d. samples of P_Z, we have

$$\lim_{\gamma\to 0^+}\lim_{n\to\infty} \frac{2}{\gamma^2 n}\log\mathbb{P}\left(\left|\frac{\mathbb{E}_{\hat{P}_Z}[h(Z)]}{\mathbb{E}[h(Z)]} - 1\right| \geq \gamma\right) - \frac{\left(\mathbb{E}[h(Z)]\right)^2}{\operatorname{var}[h(Z)]}. \tag{6.46}$$

We can apply Lemma 6.10 to $h(Z) = f(X)\,g(Y)$ with $Z = (X,Y)$ for different choices of f and g. In particular, it follows immediately that for any $P_{X,Y} \in \operatorname{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$, and any f and g such that

$$\begin{aligned} \mathbb{E}\left[f(X)^2\right] = \mathbb{E}\left[g(Y)^2\right] &= 1 \\ \mathbb{E}\left[f(X)\,g(Y)\right] &\neq 0 \\ \operatorname{var}\left[f(X)\,g(Y)\right] &> 0, \end{aligned}$$

[12] This is the special ($k = 1$) case of expressing $\|\mathbf{B} - \hat{\mathbf{B}}\|_{(k)}$ in the form

$$\max_{(f^k,g^k)\in\mathcal{F}_k\times\mathcal{G}_k} \Big[\mathbb{E}\left[f^k(X)^{\mathrm{T}}g^k(Y)\right] - \mathbb{E}_{\hat{P}_{X,Y}}\left[f^k(X)^{\mathrm{T}}g^k(Y)\right]\Big].$$

[13] For related and expanded local exponent analysis, see, e.g., [126], [127].

we have

$$-\lim_{\Delta\to 0^+}\frac{1}{\Delta^2}\lim_{n\to\infty}\frac{1}{n}\log\left[\mathbb{P}\left(\left|\frac{\mathbb{E}_{\hat{P}_{X,Y}}[f(X)\,g(Y)]}{\mathbb{E}[f(X)\,g(Y)]}-1\right|\geq\Delta\right)\right]$$
$$=\frac{1}{2}\,\frac{\mathbb{E}\big[f(X)\,g(Y)\big]^2}{\operatorname{var}\big[f(X)\,g(Y)\big]}. \tag{6.47}$$

As one instance of (6.47), we can choose $f = f_1^*$ and $g = g_1^*$ to quantify the sample complexity of estimating σ_1. As another, we can choose

$$f(x)=\frac{\mathbb{1}_{x=x_0}}{\sqrt{P_X(x)}}\quad\text{and}\quad g(y)=\frac{\mathbb{1}_{y=y_0}}{\sqrt{P_Y(y)}},$$

for some (x_0, y_0), to quantify the sample complexity of estimating the (x_0, y_0)th entry of $\mathbf{B}$, i.e.,

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

6.2.6 Additional Perspectives

While the preceding results represent useful insights into the sample complexity of modal decomposition estimation, ultimately more work is needed to fully characterize the behavior of such estimates. Moreover, one can view modal decompositions of the empirical joint distribution truncated to the dominant modes as a kind of rank-based "smoothing" of the empirical distribution to improve its quality when the number of samples is small relative to the alphabet sizes involved. As such, it can, in principle, be compared to (and potentially augmented with) a variety of ways that maximum likelihood estimates have traditionally been improved in such settings. These include parametric methods such as those based on exponential families [25], [65], [99], [199], [210], for example. They also include nonparametric smoothing, early developments of which appear in [45], [221], [234], [276]; see, e.g., [251] for an overview. And they include smoothing methods specifically tailored to settings with very large alphabets, such as additive-constant (i.e., generalized Laplace) smoothing [37], [137], [139], [163], [183], [220], Good-Turing estimation [98], [145], [198], [216], Jelinek-Mercer smoothing [138], and absolute discounting (i.e., Kneser-Ney smoothing) [152], [209].

7

Collaborative Filtering and Matrix Factorization

A variety of high-dimensional learning and inference problems can be addressed within the framework of analysis of the preceding sections. One example is the design of recommender systems [201] based on collaborative filtering [96]. These systems aim to predict the preferences of individual users for various items from (limited) knowledge of some of their and other users' item preferences, such as may be obtained from ratings data and/or records of prior choices. Among the most successful forms of collaborative filtering to date have been matrix factorization methods based on latent factor models and involving low-rank approximation techniques [153], [233], [255], a subset of which are formulated as matrix completion problems [50], [52], [53], [147] or variations thereof [90], [267].

In applying such methods, the system designer must (implicitly or otherwise) choose: 1) how to model the available data expressing user preferences; 2) what matrix representation to factor; and 3) a criterion for evaluating the quality of candidate factorizations. The large literature in this area reflects the many choices available.

In this section, we use the context of collaborative filtering to develop key matrix factorization perspectives associated with the modal

The version of record is available at: http://dx.doi.org/10.1561/0100000107

decomposition. In particular, we formulate the problem of collaborative filtering as one of Bayesian (multi-)attribute matching, and find that the optimum such filtering is achieved using a truncated modal decomposition. In addition, we demonstrate how the resulting filtering corresponds to the optimum low-rank approximation to the empirical CDM, and note how to differs in some significant respects from some other commonly proposed factorizations for such applications.

7.1 Bayesian Attribute Matching

As a convenient context popularized in [28], consider the content-provider problem of recommending movies to subscribers. Let $\mathcal{X}$ be the collection of subscribers, and let $\mathcal{Y}$ be the collection of available movies. In turn, $(X, Y) = (x, y) \in \mathcal{X} \times \mathcal{Y}$ denotes the event that the next instant a movie is watched, it will be subscriber x watching movie y, and $P_{X,Y}(x, y)$ denotes the probability of this event.

With this notation, the associated conditional $P_{Y|X}(y|x)$ denotes the probability that if x is the next subscriber to watch a movie, he/she will select movie y. From this perspective, the recommendation problem can be interpreted as identifying values of y for which this conditional probability is high for the given x, or more generally sampling from $P_{Y|X}(\cdot|x)$. Alternatively, if one seeks to avoid biasing the recommendation according to P_Y and replace it with a uniform distribution, we sample from the distribution proportional to $P_{X|Y}(x|\cdot)$ instead.

In practice, we must estimate $P_{X,Y}$ from data $(x_1, y_1), \dots, (x_n, y_n)$, where (x_j, y_j) is a record of subscriber $X = x_j$ having selected movie $Y = y_j$ to watch at some point in the past. In particular, we treat these n records as i.i.d. samples from $P_{X,Y}$. In the regime of interest, there are comparatively few training samples n relative to the joint alphabet size $\mathcal{X} \times \mathcal{Y}$, so to obtain meaningful results the procedure for estimating $P_{X,Y}$ must take this into account. In the sequel, this is accomplished by exploiting attribute variables, as we now develop.

In developing the key concepts, it is convenient to initially treat $P_{X,Y}$ as known, then return to the scenario of interest in which only the empirical distribution $\hat{P}_{X,Y}$ is available.

To begin, we view the multi-attribute variables U^k and V^k in the Markov chain $U^k \leftrightarrow X \leftrightarrow Y \leftrightarrow V^k$ obtained in Section 5.5 (and Section 5.6) as the dominant attributes of subscribers and movies, respectively. In turn the corresponding S_*^k and T_*^k, as defined in (5.12), represent sufficient statistics for the detection of these attributes.

Conceptually, for each movie y, there is an associated movie multi-attribute $V^k(y)$ generated randomly from y according to $P_{V^k|Y}(\cdot|y)$, as defined via (5.37b), that expresses its dominant characteristic. Likewise, for the target subscriber x, there is an associated movie multi-attribute $V_\circ^k(x)$ generated randomly from x according to $P_{V^k|X}(\cdot|x)$, as defined via (5.38a), that expresses his/her preferred movie characteristic. Having defined these multi-attributes, we can express the recommendation problem as one of Bayesian decision-making among multiple (not mutually exclusive) hypotheses. Specifically, $\mathcal{E}_y(x)$ denotes the event that there is an attribute match with movie y for subscriber x, i.e.,

$$\mathcal{E}_y(x) \triangleq \{V^k(y) = V_\circ^k(x)\}. \tag{7.1}$$

The following result characterizes the movie recommendation rule that results from maximizing the expected number of matches. A proof is provided in Appendix E.1.

Proposition 7.1. Given $k \in \{1, \dots, K-1\}$, $l \in \{1, \dots, |\mathcal{Y}|\}$, $P_{X,Y} \in \operatorname{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$, and a collection $\hat{\mathcal{Y}}(x)$ of l (distinct) movies for subscriber $x \in \mathcal{X}$, let the number that are a match be

$$M \triangleq \sum_{y\in\hat{\mathcal{Y}}(x)} \mathbb{1}_{\mathcal{E}_y(x)}, \tag{7.2}$$

where $\mathcal{E}_y(x)$ as defined in (7.1), with the movie multi-attributes $V^k(y)$ (for movie $y \in \mathcal{Y}$) and $V_\circ^k(x)$ (for user $x \in \mathcal{X}$) being independent and distributed according to, respectively, $P_{V^k|Y}(\cdot|y)$ and $P_{V^k|X}(\cdot|x)$ as defined via Proposition 5.13 and Corollary 5.14, i.e., according to[1]

$$P_{V_\circ^k,V^k,X,Y}(v_\circ^k, v^k, x, y) = P_{V^K|X}(v_\circ^k|x)\, P_{V^K|Y}(v^k|y)\, P_{X,Y}(x,y). \tag{7.3}$$

[1] Note that (7.3) implies, e.g., 1) $V_\circ^k \leftrightarrow (X,Y) \leftrightarrow V^k$; 2) $X \leftrightarrow Y \leftrightarrow V^k$; and 3) $V_\circ^k \leftrightarrow X \leftrightarrow Y$. These, in turn, imply, e.g., $V_\circ^k \leftrightarrow X \leftrightarrow V^k$ and $V_\circ^k \leftrightarrow X \leftrightarrow V^k$.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Then

$$\mathbb{E}[M] \le \frac{1}{2^k} \sum_{y\in\hat{\mathcal{Y}}^*(x)} \left(1+\epsilon^2\sum_{i=1}^{k} \sigma_i\, f_i^*(x)\, g_i^*(y)\right) + o(\epsilon^2), \tag{7.4}$$

as $\epsilon \to 0$, where

$$\hat{\mathcal{Y}}^*(x) \triangleq \{y_1^*(x), \ldots, y_l^*(x)\} \tag{7.5}$$

is constructed sequentially according to

$$y_1^*(x) \triangleq \operatorname*{arg\,max}_{y\in\mathcal{Y}} \sum_{i=1}^{k} \sigma_i\, f_i^*(x)\, g_i^*(y) \tag{7.6a}$$

$$y_j^*(x) \triangleq \operatorname*{arg\,max}_{y\in\mathcal{Y}\setminus\{y_1^*(x),\ldots,y_{j-1}^*(x)\}} \sum_{i=1}^{k} \sigma_i f_i^*(x)\, g_i^*(y), \quad j \in \{2,\ldots,l\}. \tag{7.6b}$$

Moreover, the inequality in (7.4) holds with equality when we choose $\hat{\mathcal{Y}}(x) = \hat{\mathcal{Y}}^*(x)$.

Note that in the case $l = 1$, the criterion in Proposition 7.1 specializes to the probability of a decision error, which our result establishes is minimized by the use of the *maximum a posteriori* (MAP) decision rule, generalizing the familiar result for Bayesian hypothesis testing. More generally, Proposition 7.1 establishes that we maximize the expected number of matches in our list by an MAP *list* decision rule: we recommend to subscriber x the l movies having the l highest probabilities of an attribute match. Note, too, that using (2.26b) we can write the k-dimensional (weighted) inner product that is the core computation in (7.6) in the form

$$\sum_{i=1}^{k} \sigma_i\, f_i^*(x)\, g_i^*(y) = \sum_{i=1}^{k} g_i^*(y)\, \mathbb{E}[g_i^*(Y)|X=x],$$

which provides additional interpretation of the maximum inner product decision rule.

7.2 Interpretation as Matrix Factorization

To interpret Bayesian attribute matching as a form of matrix factorization, we have the following result establishing the decision rule as a maximum likelihood one based on a rank-reduced approximation to $P_{X,Y}$.

Corollary 7.2. Given $k \in \{1, \dots, K-1\}$, $l \in \{1, \dots, |\mathcal{Y}|\}$, and $P_{X,Y} \in \operatorname{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$, the optimum recommendation list in Proposition 7.1 can be expressed in the form

$$\begin{aligned} y_1^*(x) &= \operatorname*{arg\,max}_{y\in\mathcal{Y}} P_{X|Y}^{(k)}(x|y) \\ y_j^*(x) &= \operatorname*{arg\,max}_{y\in\mathcal{Y}\setminus\{y_1^*(x),\dots,y_{j-1}^*(x)\}} P_{X|Y}^{(k)}(x|y), \quad j \in \{2, \dots, l\}, \end{aligned}$$

where

$$P_{X|Y}^{(k)}(x|y) \triangleq \frac{P_{X,Y}^{(k)}(x,y)}{P_Y(y)},$$

with $P_{X,Y}^{(k)}$ as defined in (4.30a).

This corollary is an immediate consequence of the fact that the objective function in (7.6) can be equivalently written in terms of $P_{X,Y}^{(k)}$; specifically,

$$\sum_{i=1}^{k} \sigma_i\, f_i^*(x)\, g_i^*(y) = \frac{P_{X,Y}^{(k)}(x,y)}{P_X(x)\, P_Y(y)} - 1.$$

Moreover, $P_{X,Y}^{(k)}$ is the distribution corresponding to $\tilde{\mathbf{B}}^{(k)}$, which is the rank-k approximation to $\tilde{\mathbf{B}}$ obtained by retaining the dominant k modes in the SVD of $\tilde{\mathbf{B}}$. Equivalently, $\tilde{\mathbf{B}}^{(k)}$ is the rank-constrained approximation to $\tilde{\mathbf{B}}$ obtained by minimizing $\|\tilde{\mathbf{B}} - \tilde{\mathbf{B}}^{(k)}\|_{\mathrm{F}}$, which follows from the well-known matrix approximation theorem [84] [114, Corollary 7.4.1.3(a) and Section 7.4.2] [97, Theorem 2.4.8]:[2]

Lemma 7.3 (Eckart-Young)**.** If $\mathbf{A}$ and $\tilde{\mathbf{A}}$ are $k_1 \times k_2$ matrices such that $\mathbf{A}$ has singular values $\sigma_1(\mathbf{A}) \geq \cdots \geq \sigma_{\min\{k_1,k_2\}}(\mathbf{A})$ and $\operatorname{rank}(\tilde{\mathbf{A}}) \leq k$, then

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_{\mathrm{F}}^2 \geq \sum_{i=k+1}^{\min\{k_1,k_2\}} \sigma_i^2. \tag{7.7}$$

[2] As discussed in [256], the original version of this approximation theorem was actually due to Schmidt [244].

Algorithm 2 Collaborative Filtering by Attribute-Matching

```
Require: Subscriber list $\mathcal{X}$, movie list $\mathcal{Y}$, selections history $\mathcal{T}$, i.e., $\hat{P}_{X,Y}$,
    dimension $k$, recommendation list size $l$, target subscriber $x$
    1. Estimate $k$ modes of $P_{X,Y}$ from $\hat{P}_{X,Y}$ via ACE:
           $\hat{\sigma}_i,\ \hat{f}_i(\cdot),\ \hat{g}_i(\cdot), \quad \text{for } i = 1, \ldots, k$
    2. Initialize recommendation list: $\hat{\mathcal{Y}}^* = \varnothing$
    3. Initialize candidates list: $\bar{\mathcal{Y}} = \mathcal{Y}$
    for $j = 1, \ldots, l$ do
        4a. $y^* \leftarrow \arg\max_{y \in \bar{\mathcal{Y}}} \sum_{i=1}^{k} \hat{\sigma}_i\, \hat{f}_i^*(x)\, \hat{g}_i^*(y)$
        4b. Update recommendation list: $\hat{\mathcal{Y}}^* \leftarrow \hat{\mathcal{Y}}^* \cup \{y^*\}$
        4c. Update candidates list: $\bar{\mathcal{Y}} \leftarrow \bar{\mathcal{Y}} \setminus \{y^*\}$
    end for
```

7.3 Collaborative Filtering Based on Attribute Matching

The preceding results lead directly to a straightforward collaborative filtering procedure. In particular, given a history (6.7) of n prior movie selections by users, modeled as drawn i.i.d. from $P_{X,Y}$, we form the empirical distribution $\hat{P}_{X,Y}$, and use this distribution in Proposition 7.1 and Corollary 7.2. Consistent with the discussion in Section 6.2, we focus on the case in which P_X and P_Y can be accurately estimated, but $P_{X,Y}$ cannot.

As such, we effectively obtain the dominant k modes from the modal decomposition (6.9) using Algorithm 1 with the empirical distribution, then use the resulting $\hat{\sigma}_i$, $\hat{f}_i^*$, and $\hat{g}_i^*$, for $i = 1, \ldots, k$, to form the score function

$$\sum_{i=1}^{k} \hat{\sigma}_i\, \hat{f}_i^*(x)\, \hat{g}_i^*(y),$$

whose maxima over y for a given subscriber x produce the movie recommendations. As the analysis in Section 6.2 reflects, we can expect the estimated modes to be accurate provided k is sufficiently small relative to n. The complete procedure takes the form of Algorithm 2.

It is worth emphasizing that the attribute matching approach to collaborative filtering dictates factoring $\hat{\mathbf{B}}$ as defined in (6.12), which differs from other approaches used in the literature. For example, popular alternatives include factoring the matrix representation $\hat{\mathbf{P}}_{Y,X}$ of $\hat{P}_{X,Y}$ itself [153], and factoring the matrix representation for pointwise mutual information [60] (information density [107]), i.e., the matrix whose (y, x)th entry is

$$\log \frac{\hat{P}_{X,Y}(x, y)}{P_X(x)\, P_Y(y)}, \tag{7.8}$$

as arises in natural language processing [173], [205].

7.4 Extensions

Finally, a natural alternative to the procedure of Corollary 7.2 are the recommendations

$$y_1^*(x) = \operatorname*{arg\,max}_{y \in \mathcal{Y}} P_{Y|X}^{(k)}(y|x) \tag{7.9a}$$

$$y_j^*(x) = \operatorname*{arg\,max}_{y \in \mathcal{Y} \setminus \{y_1^*(x), \dots, y_{j-1}^*(x)\}} P_{Y|X}^{(k)}(y|x), \quad j \in \{2, \dots, l\}, \tag{7.9b}$$

where

$$P_{Y|X}^{(k)}(y|x) \triangleq \frac{P_{X,Y}^{(k)}(x, y)}{P_X(x)}.$$

Unlike that of Corollary 7.2, this procedure includes the effect of P_Y in its recommendations. It is obtained by replacing (7.2) with

$$M \triangleq \sum_{y \in \hat{\mathcal{Y}}(x)} P_Y(y)\, \mathbb{1}_{\mathcal{E}_y(x)},$$

and analytically extending the local analysis to $\epsilon = 1$, i.e., relaxing the weak dependence constraint.

8

Softmax Regression

In this section, we develop a further characterization of the universal features arising out of the modal decomposition as the optimizing parameters in softmax regression (i.e., multinomial logistic regression) in the weak-dependence regime. Softmax regression [34], which originated with the introduction of logistic regression by Cox [64], has proven to be an extraordinarily useful classification architecture in a wide range of practical applications, and has well known approximation properties—see, e.g., [24], [73], [115]. As such, viewing our results from this perspective yields valuable additional interpretations and insights. More generally, this form of regression can be expressed as an elementary form of neural network, and thus its analysis is useful in relating the preceding results to aspects of the contemporary neural network architectures that are of particular interest for emerging applications.

The section is structured as follows. In Section 8.1, we analyze a local version of multinomial logistic regression. Specifically, under weak dependence we show that softmax weights correspond to (normalized) conditional expectations. In turn, in Section 8.2, we establish that the resulting discriminative model matches, to first order, that of a Gaussian mixture without any Gaussian assumptions in the analysis.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Most significantly, in Section 8.3 we further show that the optimizing features are, again, those of the modal decomposition, in which case the associated softmax weights are proportional to the "dual" features of the decomposition. Our analysis additionally quantifies the performance limits in this regime in terms of the associated singular values. As we discuss in Section 8.4, this analysis implies a relationship between the ACE algorithm and methods used to train at least some classes of neural networks.

8.1 A Local Analysis of Softmax Regression

In softmax regression, for a class index Y and k-dimensional real-valued data S we fit a posterior of the form

$$\tilde{P}^{g,\beta}_{Y|S}(y|s) = P_Y(y) \exp\Big\{s^{\mathrm{T}} g(y) + \beta(y) - \alpha(s)\Big\},$$

to some $P_{S,Y}$ by choosing parameters g and β. Note that g is defined by[1] $|\mathcal{Y}|$ parameter vectors $g(1), \dots, g(|\mathcal{Y}|)$, each of dimension k. Likewise, β is defined by $|\mathcal{Y}|$ scalar parameters $\beta(1), \dots, \beta(|\mathcal{Y}|)$.

We characterize the optimizing softmax parameters in the weak-dependence regime as follows; a proof is provided in Appendix F.1.

Proposition 8.1. Given $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ such that X, Y are ϵ-dependent for some $\epsilon > 0$, a dimension k, and $s = f(x)$ for some $f\colon \mathcal{X} \to \mathbb{R}^k$, let

$$P_{S,Y}(s, y) = \sum_{\{x\colon f(x)=s\}} P_{X,Y}(x, y)$$

be the induced distribution, and let P_S denote the associated induced marginal, with $\boldsymbol{\mu}_S$ its mean, and $\boldsymbol{\Lambda}_S$ its covariance, which we assume to be nonsingular.[2] Let

[1] In this section, we omit the superscript notation that we previously used to explicitly indicate the dimension of a multi-dimensional variable.

[2] It is comparatively straightforward to verify that when $\boldsymbol{\Lambda}_S$ is singular, the proposition holds provided we replace the inverse $\boldsymbol{\Lambda}_S^{-1}$ wherever it appears with the Moore-Penrose pseudoinverse $\boldsymbol{\Lambda}_S^{\dagger}$ and add to $g_{*,S}(y)$ any $g_{\varnothing}(y) \in \mathrm{null}(\boldsymbol{\Lambda}_S)$ such that $\mathbb{E}\big[g_{\varnothing}(Y)\big] = \mathbf{0}$.

$$\begin{aligned}\tilde{\mathcal{P}}_s^{\mathcal{Y}}(P_Y) \triangleq \Big\{P \in \mathcal{P}^{\mathcal{Y}}\colon &\\ & P = \tilde{P}_{Y|S}^{g,\beta}(y|s) \triangleq P_Y(y)\exp\Big\{s^{\mathrm{T}} g(y) + \beta(y) - \alpha(s)\Big\}\\ & \text{for some } \beta\colon \mathcal{Y} \to \mathbb{R} \text{ and } g\colon \mathcal{Y} \to \mathbb{R}^k\Big\}\end{aligned} \tag{8.1}$$

denote the exponential family with natural statistic $g(y)$ and natural parameter $s \in \mathcal{S}$, where $\mathcal{S} \triangleq f(\mathcal{X})$. Then

$$\begin{aligned}\min_{\tilde{P}_{Y|S}(\cdot|s) \in \tilde{\mathcal{P}}_s^{\mathcal{Y}}(P_Y)} & \sum_{s\in\mathcal{S}} P_S(s)\, D(P_{Y|S}(\cdot|s) \,\|\, \tilde{P}_{Y|S}(\cdot|s))\\ &= I(Y;S) - \frac{1}{2}\,\mathbb{E}\Big[\big\|\mathbf{\Lambda}_S^{-1/2}(\boldsymbol{\mu}_{S|Y}(Y) - \boldsymbol{\mu}_S)\big\|^2\Big] + o(\epsilon^2)\end{aligned} \tag{8.2}$$

as $\epsilon \to 0$, with

$$\boldsymbol{\mu}_{S|Y}(y) \triangleq \mathbb{E}[S|Y = y], \tag{8.3}$$

and is achieved by the parameters

$$g(y) = g_{*,S}(y) \triangleq \mathbf{\Lambda}_S^{-1}\Big(\boldsymbol{\mu}_{S|Y}(y) - \boldsymbol{\mu}_S\Big) + o(\epsilon) \tag{8.4a}$$

$$\beta(y) = \beta_{*,S}(y) \triangleq -\boldsymbol{\mu}_S^{\mathrm{T}}\, g_{*,S}(y) + o(\epsilon), \tag{8.4b}$$

i.e.,

$$\tilde{P}_{Y|S}^{*}(y|s) \propto P_Y(y)\exp\Big\{(s-\boldsymbol{\mu}_S)^{\mathrm{T}}\mathbf{\Lambda}_S^{-1}(\boldsymbol{\mu}_{S|Y}(y)-\boldsymbol{\mu}_S)\Big\}(1 + o(1)), \tag{8.4c}$$

as $\epsilon \to 0$.

8.2 Relationships to Gaussian Mixture Analysis

It is useful to note that the optimum posterior (8.4c) in Proposition 8.1 matches that for a Gaussian mixture in which the components depend weakly on the class index, despite the fact that there are no Gaussian assumptions in the proposition. In particular, suppose that $P_{S|Y}(\cdot|y) = \mathtt{N}(\boldsymbol{\mu}_{S|Y}(y), \mathbf{\Lambda}_{S|Y})$, where $\mathbf{\Lambda}_{S|Y}$ is positive definite and, as the notation reflects, does not depend on y, and where $\boldsymbol{\mu}_{S|Y}(y) \triangleq \boldsymbol{\mu}_S + \epsilon\mathbf{e}(y)$ with $\mathbb{E}[\mathbf{e}(Y)] = \mathbf{0}$ and $\epsilon > 0$. Then

$$
\begin{aligned}
&P_{Y|S}(y|s)\\
&\quad\propto P_Y(y)\, P_{S|Y}(s|y)\\
&\quad\propto P_Y(y)\exp\Big\{-\frac{1}{2}(s-\boldsymbol{\mu}_{S|Y}(y))^{\mathrm{T}}\boldsymbol{\Lambda}_{S|Y}^{-1}(s-\boldsymbol{\mu}_{S|Y})\Big\}\\
&\quad= P_Y(y)\exp\Big\{-\frac{1}{2}\Big[(s-\boldsymbol{\mu}_S)^{\mathrm{T}}\boldsymbol{\Lambda}_{S|Y}^{-1}(s-\boldsymbol{\mu}_S)\\
&\qquad\qquad\qquad\qquad+2\,(s-\boldsymbol{\mu}_{S|Y}(y))^{\mathrm{T}}\boldsymbol{\Lambda}_{S|Y}^{-1}(\boldsymbol{\mu}_S-\boldsymbol{\mu}_{S|Y}(y))\\
&\qquad\qquad\qquad\qquad\quad+(\boldsymbol{\mu}_S-\boldsymbol{\mu}_{S|Y}(y))^{\mathrm{T}}\boldsymbol{\Lambda}_{S|Y}^{-1}(\boldsymbol{\mu}_S-\boldsymbol{\mu}_{S|Y}(y))\Big]\Big\} \qquad (8.5)\\
&\quad\propto P_Y(y)\exp\Big\{(s-\boldsymbol{\mu}_{S|Y}(y))^{\mathrm{T}}\boldsymbol{\Lambda}_{S|Y}^{-1}(\boldsymbol{\mu}_{S|Y}(y)-\boldsymbol{\mu}_S)\Big\}(1+o(1)) \qquad (8.6)\\
&\quad= P_Y(y)\exp\Big\{(s-\boldsymbol{\mu}_S)^{\mathrm{T}}\boldsymbol{\Lambda}_S^{-1}(\boldsymbol{\mu}_{S|Y}(y)-\boldsymbol{\mu}_S)\Big\}(1+o(1)), \qquad (8.7)
\end{aligned}
$$

as $\epsilon \to 0$, where to obtain (8.5) we have used the simple expansion

$$s-\boldsymbol{\mu}_{S|Y}(y)=(s-\boldsymbol{\mu}_S)+(\boldsymbol{\mu}_S-\boldsymbol{\mu}_{S|Y}(y)),$$

and to obtain (8.6) we have used that in the exponent in (8.5) the first term does not depend on y, the second term is $\mathcal{O}(\epsilon)$, and the third term is $o(\epsilon)$. To obtain (8.7) we have used that $\boldsymbol{\mu}_{S|Y}-\boldsymbol{\mu}_S$ and $\boldsymbol{\Lambda}_S^{-1}-\boldsymbol{\Lambda}_{S|Y}^{-1}$ are both $o(1)$. Hence, (8.7) and (8.4c) match, to first order.

8.3 Optimum Feature Design

Proposition 8.1 describes the optimizing softmax weights g and biases β for a given choice of f. When we further optimize over the choice of f, we obtain a direct relation to the modal decomposition of Proposition 2.2 and the universal posterior (5.1). In particular, we have the following result, a proof of which is provided in Appendix F.2.

Corollary 8.2. Given dimension $k \in \{1,\dots,K-1\}$, if $P_{X,Y}$ is such that f_*^k as defined in (3.5) is injective (i.e., a one-to-one function), then

$$\min_{\substack{\{\text{injective } f:\\ s=f(x)\}}}\sum_{s\in\mathcal{S}}P_S(s)\,D(P_{Y|S}(\cdot|s)\,\|\,\tilde{P}^*_{Y|S}(\cdot|s))=\frac{1}{2}\sum_{i=k+1}^{K}\sigma_i^2+o(\epsilon^2), \tag{8.8}$$

as $\epsilon \to 0$, and is achieved by

$$f_i(x) = f_i^*(x), \quad x \in \mathcal{X}, \quad i = 1, \dots, k, \tag{8.9}$$

with f_i^* as defined in Proposition 2.2 . Moreover, for this choice of f, the parameters

$$g_{*,S}(y) = (g_1^{*,S}(y), \dots, g_k^{*,S}(y)) \quad \text{and} \quad \beta_{*,S}(y)$$

in Proposition 8.1 take the form

$$g_i^{*,S}(y) = \sigma_i \, g_i^*(y), \quad y \in \mathcal{Y}, \quad i = 1, \dots, k, \tag{8.10a}$$

$$\beta_{*,S} = 0, \tag{8.10b}$$

where $g_i^*(y)$ and σ_i are as defined in Proposition 2.2, and thus

$$\tilde{P}^*_{Y|S_*}(y|f_*^k(x)) \propto P_Y(y) \exp\left\{ \sum_{i=1}^k \sigma_i f_i^*(x) \, g_i^*(y) \right\} (1 + o(1)), \quad \epsilon \to 0. \tag{8.11}$$

8.4 A Neural Network Perspective

Since softmax regression can be interpreted as a simple neural network classifier with a single hidden layer [100], [207], the preceding results can be equivalently expressed in terms of the optimization of such networks. Moreover, with the interpretation of universal features as a solution to a local information bottleneck as developed in Section 5.6, these results shed insight into recent analyses of deep learning based on such bottlenecks [2], [7], [250], [260] as well as related information-theoretic analyses [206].

The neural network architecture associated with softmax analysis is, in our notation, as depicted in Figure 8.1. The input layer uses a so-called "one-hot" representation of the input x, corresponding to weights $\mathbb{1}_{x=j}$. Next, in the hidden layer, features $s_i = f_i(x)$ of the input x are generated using weights $f_i(j)$. Finally, in the output layer, the (unnormalized) log-posterior $\tau(y)$ is constructed according to

$$\tau(y) = \sum_{i=1}^k s_i \, g_i(y) + \beta(y), \qquad y = 1, \dots, |\mathcal{Y}|$$

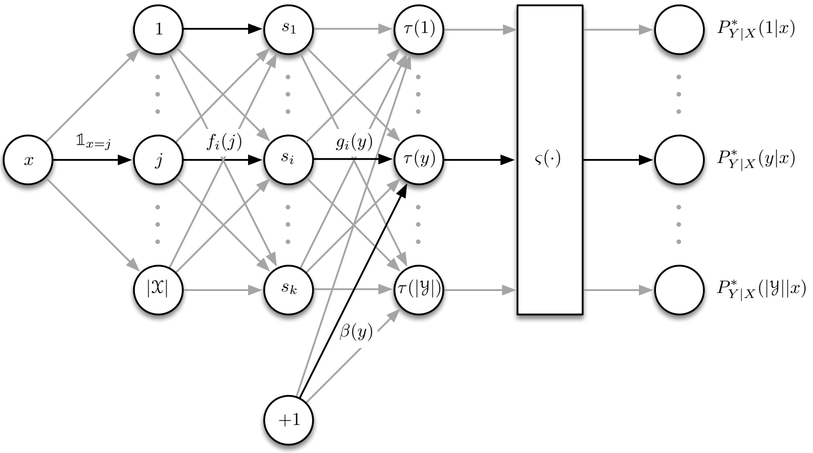

Figure 8.1: A neural network representation of the softmax regression framework. In this network, we use a one-hot representation of the input x, corresponding to the kronecker weights $\mathbb{1}_{x=j}$. The hidden layer is characterized by the feature weights $f_i(j)$ for $i \in \{1, \ldots, k\}$, $j \in \{1, \ldots, |\mathcal{X}|\}$, and the output layer is parameterized by the weights $g_i(y)$ and biases $\beta(y)$, for $i \in \{1, \ldots, k\}$, $y \in \{1, \ldots, |\mathcal{Y}|\}$. The softmax processing, as defined in (8.12), is represented by the sigmoid function $\varsigma(\cdot)$ and operates on the unnormalized log-posterior $\tau(y)$.

using output layer weights $g_i(y)$ and biases $\beta(y)$. The $\tau(y)$ are then combined and normalized to produce the posterior via the softmax processing

$$P^*_{Y|X}(y|x) = \frac{\mathrm{e}^{\tau(y)}}{\sum\limits_{y'=1}^{|\mathcal{Y}|} \mathrm{e}^{\tau(y')}} = \varsigma\left(-\log \sum_{y' \neq y} \mathrm{e}^{\tau(y')-\tau(y)}\right), \tag{8.12a}$$

where

$$\varsigma(\omega) \triangleq \frac{1}{1 + \mathrm{e}^{-\omega}} \tag{8.12b}$$

is the sigmoid function [100].

For such networks and their multi-layer generalizations, the optimization of the weights and biases is typically carried out using stochastic gradient descent (SGD) [100]. As such, Proposition 8.1 implies that SGD is effectively computing empirical conditional expectations, and as such corresponds to an approximation to one step of the ACE algorithm.

More generally, Corollary 8.2 establishes that jointly optimizing the softmax parameters and data embeddings (features) can be accomplished iteratively via the full ACE algorithm.

To see this, let $\hat{P}_{S,Y}$ denote the empirical distribution for induced training data

$$\mathcal{T}_f = \{(s_1, y_1), \dots, (s_n, y_n)\}$$

generated from $P_{S,Y}$, and note that

$$\sum_{s\in\mathcal{S}} \hat{P}_S(s)\, D\big(\hat{P}_{Y|S}(\cdot|s)\big\|\tilde{P}^{g,\beta}_{Y|S}(\cdot|s)\big) = \hat{H}(Y|S) - \underbrace{\mathbb{E}_{\hat{P}_{S,Y}}\big[\log \tilde{P}^{g,\beta}_{Y|S}(Y|S)\big]}_{\triangleq \ell(g,\beta)},$$

where $\ell(g, \beta)$ is the log-likelihood function, and $\hat{H}(Y|S)$ denotes the empirical conditional entropy. Then if the number of training samples is sufficiently large that, effectively,[3]

$$\hat{P}_Y(y) = \sum_{s\in\mathcal{S}} \hat{P}_{S,Y}(s, y) = P_Y(y), \quad y \in \mathcal{Y},$$

the maximum-likelihood parameters are, via Proposition 8.1

$$\hat{g}_{*,S}(y) = \hat{\boldsymbol{\Lambda}}_S^{-1}\Big(\hat{\boldsymbol{\mu}}_{S|Y}(y) - \hat{\boldsymbol{\mu}}_S\Big) + o(\epsilon) \tag{8.13a}$$

$$\hat{\beta}_{*,S}(y) = -\hat{\boldsymbol{\mu}}_S^{\mathrm{T}}\, \hat{g}_{*,S}(y) + o(\epsilon), \tag{8.13b}$$

where

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_S &= \mathbb{E}_{\hat{P}_S}[S] \\
\hat{\boldsymbol{\mu}}_{S|Y}(y) &= \mathbb{E}_{\hat{P}_{S|Y}(\cdot|y)}[S] \\
\hat{\boldsymbol{\Lambda}}_S &= \mathbb{E}_{\hat{P}_S}\Big[(S - \hat{\boldsymbol{\mu}}_S)(S - \hat{\boldsymbol{\mu}}_S)^{\mathrm{T}}\Big],
\end{aligned}$$

with

$$\begin{aligned}
\hat{P}_S(s) &= \sum_{y\in\mathcal{Y}} \hat{P}_{S,Y}(s, y), \quad s \in \mathcal{S} \\
\hat{P}_{S|Y}(s|y) &= \frac{\hat{P}_{S,Y}(s, y)}{P_Y(y)}, \quad s \in \mathcal{S},\ y \in \mathcal{Y}.
\end{aligned}$$

[3] There are only $|\mathcal{Y}| - 1$ degrees of freedom in P_Y, so that when $|\mathcal{S}|$ is large, as we assume, P_Y can be more accurately estimated from a given number of samples than $P_{S,Y}$, since the latter is described by $|\mathcal{S}||\mathcal{Y}| - 1$ degrees of freedom.

Likewise, further optimizing the likelihood for the training data (6.7) with respect to f such that $s = f(x)$ under the condition that $\hat{P}_S(s) = P_S(s)$ for $s \in \mathcal{S}$ yields that

$$\hat{f}_i(x) = \hat{f}_i^*(x) \text{ and } \hat{g}_i(y) = \hat{\sigma}_i\, \hat{g}_i^*(y), \quad i \in \{1, \ldots, k\},$$

where $\hat{f}_i$, $\hat{g}_i$, and $\hat{\sigma}_i$ are as defined in the analysis of Section 6.2 and can (effectively) be computed via Algorithm 1; specifically, they characterize the modal decomposition of the empirical distribution as expressed by (6.9).

Such analysis suggests the potential for alternatives to SGD that more directly approximate empirical conditional expectation, and for interpretations of the iterative matrix factorizations inherent in, e.g., [83], [268]. Moreover, the analysis provides an upper bound (8.8) on performance against which the performance of various weight optimization strategies can be measured.

Finally—and perhaps more importantly—we can view existing neural network implementations as a tool for efficiently computing conditional expectations. Indeed, direct computation of empirical conditional expectations can be prohibitively expensive in practice for typical alphabet sizes, which the use of SGD can circumvent.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

9

Gaussian Distributions and Linear Features

While the preceding sections have focused on distributions over finite alphabets, in this section we turn our attention to the case of continuous-valued variables, emphasizing the case in which X, Y are Gaussian. Our treatment closely parallels the preceding one for finite alphabets, and has its roots in the pioneering work of Pearson [224] and Hotelling [116], [117]. Indeed, as we will develop, the resulting features in this case are linear, and strongly connected to both canonical correlation analysis (CCA) [109], [110], [117] and principal component analysis (PCA) [116], [140], [224]. More generally, the associated framework provides an analysis for the case of arbitrary distributions of continuous-valued variables subject to linear processing constraints.

The section is organized as follows. We first establish some notation in Section 9.1, then proceed in Section 9.2 to construct the modal decomposition of covariance via the SVD of the canonical correlation matrix (CCM), and in Section 9.3 to obtain the familiar formulation of Hotelling's canonical correlation analysis (CCA) via the corresponding variational characterization. We further define a local Gaussian geometry in Section 9.4, the associated notion of weakly correlated variables in Section 9.5, and then, in Section 9.6, construct a local modal decom-

The version of record is available at: http://dx.doi.org/10.1561/0100000107

position of joint distributions of such variables in terms of the CCA features, which are linear. Section 9.6 also includes a brief discussion of aspects of extensions to nonlinear features and nonGaussian distributions. In Section 9.7 we introduce Gaussian attribute models, and then show that the CCA features arise in the solution to universal feature problem formulations. In particular, Section 9.8 shows they arise in the solution of an attribute estimation game in which nature chooses the attribute at random after the system designer chooses the linear features from which it will be estimated using a minimum mean-square error (MMSE) criterion, and Section 9.9 shows they arise in the solution of the corresponding cooperative MMSE attribute estimation game; these analyses are global. Section 9.10, shows the CCA features arising in the solution to the local symmetric version of Tishby's Gaussian information bottleneck problem, and Section 9.11 describes how superpositions of CCA features arise in the solution to the (global) Gaussian version of Wyner's common information problem; locally this common information is given by the nuclear norm of the CCM. Section 9.12 describes the Markov relationships between the dominant attributes in the solution to the information bottleneck and the common information variable. Section 9.13 interprets the features arising out of Pearson's principal component analysis (PCA) as a special case of the preceding analyses in which the underlying variables are simultaneously diagonalizable, and Section 9.14 discusses the estimation of CCA features, interpreting the associated SVD computation as a version of the ACE algorithm in which the features are linearly constrained. Section 9.15 develops Gaussian attribute matching, and interprets the resulting procedure as one of optimum rank-constrained linear estimation, and Section 9.16 develops a form of rank-constrained linear regression as the counterpart to softmax regression, and distinguishes it from classical formulations.

9.1 Gaussian Variables

We begin with some convenient terminology, notation, and conventions. Our development focuses on Gaussian variables that take the form of (column) vectors. We use $\mathtt{N}(\mu_Z, \mathbf{\Lambda}_Z)$ to denote the corresponding

The version of record is available at: http://dx.doi.org/10.1561/0100000107

distribution of such a variable[1] $Z \in \mathbb{R}^{K_Z}$, where μ_Z and $\mathbf{\Lambda}_Z$ denote the associated mean vector and covariance matrix, respectively, parameterizing the distribution, i.e.,[2]

$$P_Z(z) = \frac{|\mathbf{\Lambda}_Z|^{-1/2}}{(2\pi)^{K_Z/2}} \exp\left\{-\frac{1}{2}\,(z-\mu_Z)^{\mathrm{T}}\mathbf{\Lambda}_Z^{-1}\,(z-\mu_Z)\right\}, \tag{9.1}$$

with $|\cdot|$ denoting the determinant of its argument. Without loss of generality, we restrict our attention to variables Z such that $\mathbf{\Lambda}_Z$ is (strictly) positive definite since otherwise we may eliminate the associated redundancy by reducing the dimensionality of Z until the covariance matrix is positive definite. Also, for simplicity of exposition we restrict our attention to zero-mean variables whenever possible, while recognizing that nonzero means are unavoidable when conditioning on other such variables. The extension to the more general case is straightforward.

It will frequently be convenient to work with the following equivalent representation of a random variable.

Definition 9.1 (Normalized Variable). For a variable $Z \in \mathbb{R}^{K_Z}$ with mean μ_Z and covariance $\mathbf{\Lambda}_Z$, the corresponding normalized variable is

$$\tilde{Z} \triangleq \mathbf{\Lambda}_Z^{-1/2}(Z-\mu_Z) \tag{9.2}$$

and has mean $\mathbf{0}$ and covariance $\mathbf{I}$.

In the sequel, we will generally use ˜ notation to indicate variables normalized according to Definition 9.1.

Next, consider an arbitrary pair of Gaussian variables, $Z \in \mathbb{R}^{K_Z}$ and $W \in \mathbb{R}^{K_W}$, which are jointly represented by

$$C = \begin{bmatrix} Z \\ W \end{bmatrix} \sim \mathtt{N}(\mathbf{0}, \mathbf{\Lambda}_C), \tag{9.3a}$$

[1] In the Gaussian development, to avoid certain notational conflicts we drop the use of boldface characters for random vectors, but retain them for nonrandom ones, and to further simplify notation, we also forgo the use of superscripts to indicate the dimension of a variable, as in Section 8.

[2] We use (upper case) P notation for the probability density functions of continuous-valued random variables.

where

$$\mathbf{\Lambda}_C = \mathbb{E}[CC^{\mathrm{T}}] = \begin{bmatrix} \mathbf{\Lambda}_Z & \mathbf{\Lambda}_{ZW} \\ \mathbf{\Lambda}_{WZ} & \mathbf{\Lambda}_W \end{bmatrix}, \tag{9.3b}$$

so $Z \sim \mathtt{N}(\mathbf{0}, \mathbf{\Lambda}_Z)$, $W \sim \mathtt{N}(\mathbf{0}, \mathbf{\Lambda}_W)$, $\mathbf{\Lambda}_{WZ} = \mathbb{E}[WZ^{\mathrm{T}}]$, and $\mathbf{\Lambda}_{ZW} = \mathbf{\Lambda}_{WZ}^{\mathrm{T}}$.

It will frequently be convenient to express the relationship between such variables in the familiar innovations form, the notation for which we summarize as follows.

Lemma 9.2 (Innovations Form and MMSE Estimation)**.** For any zero-mean jointly Gaussian variables Z, W characterized by $\mathbf{\Lambda}_Z$, $\mathbf{\Lambda}_W$, and $\mathbf{\Lambda}_{ZW}$, we have

$$Z = \mathbf{\Gamma}_{Z|W}\, W + \nu_{W\to Z}, \tag{9.4}$$

with gain matrix

$$\mathbf{\Gamma}_{Z|W} \triangleq \mathbf{\Lambda}_{ZW}\mathbf{\Lambda}_W^{-1}, \tag{9.5}$$

and where $\nu_{W\to Z} \sim \mathtt{N}(\mathbf{0}, \mathbf{\Lambda}_{Z|W})$ is independent of W and thus

$$\mathbf{\Lambda}_{Z|W} = \mathbb{E}\big[\nu_{W\to Z}\, \nu_{W\to Z}^{\mathrm{T}}\big] = \mathbf{\Lambda}_Z - \mathbf{\Lambda}_{ZW}\mathbf{\Lambda}_W^{-1}\mathbf{\Lambda}_{WZ}. \tag{9.6}$$

Moreover, the MMSE estimate of Z based on W follows immediately as

$$\hat{Z}(W) = \mathbb{E}[Z|W] = \mathbf{\Gamma}_{Z|W}\, W, \tag{9.7}$$

for which the mean-square error (MSE) in the resulting estimation error $\nu_{W\to Z}$ is, from (9.6),

$$\lambda_{\mathrm{e}}^{Z|W} \triangleq \mathbb{E}\Big[\big\|\nu_{W\to Z}\big\|^2\Big] = \mathrm{tr}\big(\mathbf{\Lambda}_{Z|W}\big). \tag{9.8}$$

9.2 The Modal Decomposition of Covariance

For the model of interest involving zero-mean jointly Gaussian $X \in \mathbb{R}^{K_X}$ and $Y \in \mathbb{R}^{K_Y}$ with covariances $\mathbf{\Lambda}_X$ and $\mathbf{\Lambda}_Y$, respectively, and cross-covariance $\mathbf{\Lambda}_{XY}$, it follows that the correlation structure among the equivalent normalized variables

$$\tilde{A} \triangleq \begin{bmatrix} \tilde{X} \\ \tilde{Y} \end{bmatrix} \tag{9.9}$$

is

$$\mathbf{\Lambda}_{\tilde{A}} = \begin{bmatrix} \mathbf{I} & \tilde{\mathbf{B}}^{\mathrm{T}} \\ \tilde{\mathbf{B}} & \mathbf{I} \end{bmatrix}, \tag{9.10}$$

where

$$\tilde{\mathbf{B}} \triangleq \mathbf{\Lambda}_Y^{-1/2}\, \mathbf{\Lambda}_{YX}\, \mathbf{\Lambda}_X^{-1/2} = \mathbf{\Lambda}_Y^{-1/2}\, \mathbf{\Gamma}_{Y|X}\, \mathbf{\Lambda}_X^{1/2} \tag{9.11}$$

plays the role in Gaussian analysis corresponding to $\tilde{\mathbf{B}}$ in the discrete case [124]. We recognize $\tilde{\mathbf{B}}$ as, of course, the central quantity in CCA, and refer to it as the *canonical correlation matrix (CCM).*

The SVD of $\tilde{\mathbf{B}}$ takes the form

$$\tilde{\mathbf{B}} = \mathbf{\Psi}^Y \mathbf{\Sigma} \left(\mathbf{\Psi}^X\right)^{\mathrm{T}} = \sum_{i=1}^{K} \sigma_i\, \boldsymbol{\psi}_i^Y \left(\boldsymbol{\psi}_i^X\right)^{\mathrm{T}}, \tag{9.12}$$

with

$$K \triangleq \min\{K_X, K_Y\}, \tag{9.13}$$

where $\mathbf{\Sigma}$ is an $K_Y \times K_X$ diagonal matrix whose K diagonal entries are $\sigma_1, \dots, \sigma_K$, where

$$\mathbf{\Psi}^X = \begin{bmatrix} \boldsymbol{\psi}_1^X & \cdots & \boldsymbol{\psi}_{K_X}^X \end{bmatrix} \tag{9.14a}$$

$$\mathbf{\Psi}^Y = \begin{bmatrix} \boldsymbol{\psi}_1^Y & \cdots & \boldsymbol{\psi}_{K_Y}^Y \end{bmatrix} \tag{9.14b}$$

are $K_X \times K_X$ and $K_Y \times K_Y$ orthogonal matrices, respectively, and where, as before, we order the singular values according to $\sigma_1 \geq \cdots \geq \sigma_K$.

Analogous to the case of finite alphabets, $\tilde{\mathbf{B}}$ is a contractive operator representing conditional expectation, i.e., $\sigma_i \leq 1$ for $i = 1, \dots, K$, as is the case for $\tilde{\mathbf{B}}$ in the finite-alphabet case. In particular, this follows from the following standard result, a derivation of which we provide for convenience in Appendix G.1.

Fact 9.3. Let $\mathbf{M}$ be a matrix such that

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{I} & \mathbf{M} \\ \mathbf{M}^{\mathrm{T}} & \mathbf{I} \end{bmatrix}$$

is symmetric, and let $\sigma_i(\mathbf{M})$ denote the ith singular value of $\mathbf{M}$. Then $\mathbf{\Lambda}$ is positive semidefinite if and only if $\sigma_i(\mathbf{M}) \leq 1$ for all i. More specifically, the ith pair of eigenvalues of $\mathbf{\Lambda}$ are $1 \pm \sigma_i(\mathbf{M})$ and the remaining eigenvalues are all unity.

In turn, the SVD (9.12) yields the following modal decomposition of the covariance $\mathbf{\Lambda}_{YX}$.

Proposition 9.4. Let $X \in \mathbb{R}^{K_X}$, $Y \in \mathbb{R}^{K_Y}$ be zero-mean jointly Gaussian variables characterized by $\mathbf{\Lambda}_X$, $\mathbf{\Lambda}_Y$, and $\mathbf{\Lambda}_{XY}$, and let (9.12) denote the SVD of its CCM (9.11). Then there exist invertible linear mappings (coordinate transformation)

$$S^* \triangleq \mathbf{f}^*(X) = \begin{bmatrix} f_1^*(X) & \cdots & f_{K_X}^*(X) \end{bmatrix}^{\mathrm{T}} \triangleq (\mathbf{F}^*)^{\mathrm{T}} X \tag{9.15a}$$

$$T^* \triangleq \mathbf{g}^*(Y) = \begin{bmatrix} g_1^*(Y) & \cdots & g_{K_Y}^*(Y) \end{bmatrix}^{\mathrm{T}} \triangleq (\mathbf{G}^*)^{\mathrm{T}} Y \tag{9.15b}$$

satisfying

$$\mathbb{E}\big[\mathbf{f}^*(X)\, \mathbf{f}^*(X)^{\mathrm{T}}\big] = (\mathbf{F}^*)^{\mathrm{T}} \mathbf{\Lambda}_X\, \mathbf{F}^* = \mathbf{I} \tag{9.16a}$$

$$\mathbb{E}\big[\mathbf{g}^*(Y)\, \mathbf{g}^*(Y)^{\mathrm{T}}\big] = (\mathbf{G}^*)^{\mathrm{T}} \mathbf{\Lambda}_Y\, \mathbf{G}^* = \mathbf{I}, \tag{9.16b}$$

such that

$$\mathbb{E}\big[\mathbf{g}^*(Y)\, \mathbf{f}^*(X)^{\mathrm{T}}\big] = \mathbf{\Sigma}, \tag{9.17}$$

i.e.,

$$\mathbf{\Lambda}_{YX} = (\mathbf{G}^*)^{-\mathrm{T}} \mathbf{\Sigma}\, (\mathbf{F}^*)^{-1} = \mathbf{\Lambda}_Y\, \mathbf{G}^*\, \mathbf{\Sigma}\, (\mathbf{F}^*)^{\mathrm{T}} \mathbf{\Lambda}_X. \tag{9.18}$$

Proof. Let

$$\mathbf{F}^* \triangleq \mathbf{\Lambda}_X^{-1/2}\, \mathbf{\Psi}^X \tag{9.19a}$$

$$\mathbf{G}^* \triangleq \mathbf{\Lambda}_Y^{-1/2}\, \mathbf{\Psi}^Y, \tag{9.19b}$$

which we note satisfy (9.16)

$$(\mathbf{F}^*)^{\mathrm{T}} \mathbf{\Lambda}_X\, \mathbf{F}^* = (\mathbf{\Psi}^X)^{\mathrm{T}} \mathbf{\Lambda}_X^{-1/2} \mathbf{\Lambda}_X \mathbf{\Lambda}_X^{-1/2} \mathbf{\Psi}^X = \mathbf{I}$$

$$(\mathbf{G}^*)^{\mathrm{T}} \mathbf{\Lambda}_Y\, \mathbf{G}^* = (\mathbf{\Psi}^Y)^{\mathrm{T}} \mathbf{\Lambda}_Y^{-1/2} \mathbf{\Lambda}_Y \mathbf{\Lambda}_Y^{-1/2} \mathbf{\Psi}^Y = \mathbf{I}.$$

Moreover, since

$$(\mathbf{F}^*)^{-1} = (\mathbf{\Psi}^X)^{\mathrm{T}} \mathbf{\Lambda}_X^{1/2}$$

$$(\mathbf{G}^*)^{-1} = (\mathbf{\Psi}^Y)^{\mathrm{T}} \mathbf{\Lambda}_Y^{1/2},$$

it follows that (9.18) is satisfied, i.e.,

$$(\mathbf{G}^*)^{-\mathrm{T}} \mathbf{\Sigma}\, (\mathbf{F}^*)^{-1} = \mathbf{\Lambda}_Y^{1/2}\, \mathbf{\Psi}^Y \mathbf{\Sigma}\, (\mathbf{\Psi}^X)^{\mathrm{T}} \mathbf{\Lambda}_X^{1/2} = \mathbf{\Lambda}_{YX},$$

where to obtain the last equality we have used (9.12). ■

One consequence of Proposition 9.4 are the following conditional expectation relations, which are derived in Appendix G.2

Corollary 9.5. The features $\mathbf{f}^*$ and $\mathbf{g}^*$ defined via (9.19) satisfy

$$\boldsymbol{\Sigma}\,\mathbf{f}^*(X) = \mathbb{E}\big[\mathbf{g}^*(Y)\big|X\big] \tag{9.20a}$$

$$\boldsymbol{\Sigma}\,\mathbf{g}^*(Y) = \mathbb{E}\big[\mathbf{f}^*(X)\big|Y\big]. \tag{9.20b}$$

Finally, note that $\tilde{\mathbf{B}}$ has the interpretation of the gain matrix in estimates of $\tilde{Y}$ based on $\tilde{X}$ (and vice-versa). In particular, from Lemma 9.2 it is readily verified that we have the innovations form

$$\tilde{Y} = \tilde{\mathbf{B}}\tilde{X} + \tilde{\nu}, \tag{9.21}$$

i.e., $\boldsymbol{\Gamma}_{\tilde{Y}|\tilde{X}} = \tilde{\mathbf{B}}$, with

$$\mathbb{E}\big[\tilde{\nu}\tilde{\nu}^{\mathrm{T}}\big] = \boldsymbol{\Lambda}_{\tilde{Y}|\tilde{X}} = \mathbf{I} - \tilde{\mathbf{B}}\tilde{\mathbf{B}}^{\mathrm{T}},$$

so the resulting MSE in the MMSE estimate

$$\hat{\tilde{Y}}(\tilde{X}) = \mathbb{E}\big[\tilde{Y}\big|\tilde{X}\big] = \tilde{\mathbf{B}}\tilde{X} \tag{9.22}$$

is

$$\tilde{\lambda}_{\mathrm{e}} \triangleq \mathbb{E}\big[\|\tilde{\nu}\|^2\big] = \mathrm{tr}\big(\mathbf{I} - \tilde{\mathbf{B}}\tilde{\mathbf{B}}^{\mathrm{T}}\big) = K_Y - \|\tilde{\mathbf{B}}\|_{\mathrm{F}}^2. \tag{9.23}$$

As such, the SVD (9.12) has the further interpretation as a modal decomposition of the MMSE estimator (9.22).

9.3 Variational Characterization of the Modal Decomposition

As in the case of finite alphabets, the linear features $(\mathbf{f}^*, \mathbf{g}^*)$ in Proposition 9.4 can be equivalently obtained from a variational characterization, via which we obtain the key connection to CCA [109], [110], [117].[3]

[3] In some of the literature, the term CCA is used to refer to a general maximal correlation framework, and the term *linear CCA* is used to designate the special case in which the features are restricted to be linear. We largely avoid this terminology, and use CCA to refer only to the latter, consistent with its original conception.

Proposition 9.6. For any $k \in \{1, \dots, K\}$, let $\mathbf{F}^*_{(k)}$ and $\mathbf{G}^*_{(k)}$ denote the first k columns of $\mathbf{F}^*$ and $\mathbf{G}^*$, respectively, in Proposition 9.4, i.e.,

$$S^*_{(k)} \triangleq \left(\mathbf{F}^*_{(k)}\right)^{\mathrm{T}} X \triangleq \left[f^*_1(X) \quad \cdots \quad f^*_k(X)\right]^{\mathrm{T}} \tag{9.24a}$$

$$T^*_{(k)} \triangleq \left(\mathbf{G}^*_{(k)}\right)^{\mathrm{T}} Y \triangleq \left[g^*_1(Y) \quad \cdots \quad g^*_k(Y)\right]^{\mathrm{T}}. \tag{9.24b}$$

Then

$$\begin{aligned}\left(\mathbf{F}^*_{(k)}, \mathbf{G}^*_{(k)}\right) &= \operatorname*{arg\,min}_{(\mathbf{F}_{(k)}, \mathbf{G}_{(k)}) \in \mathcal{L}} \mathbb{E}\left[\left\|\mathbf{F}^{\mathrm{T}}_{(k)} X - \mathbf{G}^{\mathrm{T}}_{(k)} Y\right\|^2\right] \\ &= \operatorname*{arg\,max}_{(\mathbf{F}_{(k)}, \mathbf{G}_{(k)}) \in \mathcal{L}} \sigma(\mathbf{F}_{(k)}, \mathbf{G}_{(k)}),\end{aligned} \tag{9.25}$$

where

$$\sigma(\mathbf{F}_{(k)}, \mathbf{G}_{(k)}) \triangleq \mathbb{E}\left[\left(\mathbf{F}^{\mathrm{T}}_{(k)} X\right)^{\mathrm{T}} \mathbf{G}^{\mathrm{T}}_{(k)} Y\right] = \mathrm{tr}\left(\mathbf{G}^{\mathrm{T}}_{(k)} \mathbf{\Lambda}_{YX} \mathbf{F}_{(k)}\right) \tag{9.26}$$

and

$$\mathcal{L} \triangleq \left\{\left(\mathbf{F}_{(k)}, \mathbf{G}_{(k)}\right) \colon \mathbf{F}^{\mathrm{T}}_{(k)} \mathbf{\Lambda}_X \, \mathbf{F}_{(k)} = \mathbf{G}^{\mathrm{T}}_{(k)} \mathbf{\Lambda}_Y \, \mathbf{G}_{(k)} = \mathbf{I}\right\}. \tag{9.27}$$

Moreover, the resulting maximal correlation (generalized Pearson correlation coefficient) is

$$\sigma(\mathbf{F}^*_{(k)}, \mathbf{G}^*_{(k)}) = \mathrm{tr}\left(\left(\mathbf{G}^*_{(k)}\right)^{\mathrm{T}} \mathbf{\Lambda}_{YX} \mathbf{F}^*_{(k)}\right) = \sum_{i=1}^{k} \sigma_i, \tag{9.28}$$

the Ky Fan k-norm of $\tilde{\mathbf{B}}$.

Proof. Without loss of generality, we reparameterize $\mathbf{F}_{(k)}$ and $\mathbf{G}_{(k)}$ in terms of new matrices[4] $\mathbf{\Xi}^X$ and $\mathbf{\Xi}^Y$ according to

$$\mathbf{F}_{(k)} = \mathbf{\Lambda}_X^{-1/2} \, \mathbf{\Xi}^X \tag{9.29a}$$

$$\mathbf{G}_{(k)} = \mathbf{\Lambda}_Y^{-1/2} \, \mathbf{\Xi}^Y, \tag{9.29b}$$

in which case

$$\sigma(\mathbf{F}_{(k)}, \mathbf{G}_{(k)}) = \mathrm{tr}\left(\mathbf{G}^{\mathrm{T}}_{(k)} \mathbf{\Lambda}_{YX} \mathbf{F}_{(k)}\right) = \mathrm{tr}\left(\left(\mathbf{\Xi}^Y\right)^{\mathrm{T}} \tilde{\mathbf{B}} \, \mathbf{\Xi}^X\right), \tag{9.30}$$

[4] We refer to $\mathbf{\Xi}^X$ and $\mathbf{\Xi}^Y$ as the *feature weights* associated with the linear features S and T, and note that they play the role in Gaussian analysis corresponding to that played by the feature vectors $\mathbf{\Xi}^X$ and $\mathbf{\Xi}^Y$ in the discrete case.

and (9.27) dictates that

$$(\boldsymbol{\Xi}^X)^{\mathrm{T}}\boldsymbol{\Xi}^X = (\boldsymbol{\Xi}^Y)^{\mathrm{T}}\boldsymbol{\Xi}^Y = \mathbf{I}. \tag{9.31}$$

From Lemma 3.2, it follows immediately that (9.30) is maximized subject to (9.31) when we choose

$$\boldsymbol{\Xi}^X = \boldsymbol{\Psi}^X_{(k)} \tag{9.32a}$$

$$\boldsymbol{\Xi}^Y = \boldsymbol{\Psi}^Y_{(k)}, \tag{9.32b}$$

where

$$\boldsymbol{\Psi}^X_{(k)} \triangleq \begin{bmatrix} \boldsymbol{\psi}^X_1 & \cdots & \boldsymbol{\psi}^X_k \end{bmatrix} \tag{9.33a}$$

$$\boldsymbol{\Psi}^Y_{(k)} \triangleq \begin{bmatrix} \boldsymbol{\psi}^Y_1 & \cdots & \boldsymbol{\psi}^Y_k \end{bmatrix}, \tag{9.33b}$$

and the resulting maximal correlation is (9.28), i.e.,

$$\mathbf{F}^*_{(k)} = \boldsymbol{\Lambda}_X^{-1/2}\boldsymbol{\Psi}^X_{(k)} \tag{9.34a}$$

$$\mathbf{G}^*_{(k)} = \boldsymbol{\Lambda}_Y^{-1/2}\boldsymbol{\Psi}^Y_{(k)}, \tag{9.34b}$$

as claimed. ■

9.4 Local Gaussian Information Geometry

It will sometimes be useful to define a local analysis for Gaussian variables. For such analysis, there is a natural counterpart of the χ^2-divergence (4.1b) used in the finite-alphabet case. In particular, we will make use of the following notion of neighborhood.

Definition 9.7 (Gaussian ϵ-Neighborhood). For a given $\epsilon > 0$, the ϵ-neighborhood of a K_0-dimensional Gaussian distribution $P_0 = \mathtt{N}(\mu_0, \boldsymbol{\Lambda}_0)$ with positive definite $\boldsymbol{\Lambda}_0$ is the set of Gaussian distributions in the following generalized divergence ball of radius ϵ^2 about P_0, i.e.,

$$\mathcal{N}^{K_0}_\epsilon(P_0) \triangleq \{P' = \mathtt{N}(\mu, \boldsymbol{\Lambda})\colon \bar{D}(P' \| P_0) \le \epsilon^2 K_0\}, \tag{9.35a}$$

where for $P = \mathtt{N}(\mu_P, \boldsymbol{\Lambda}_P)$ and $Q = \mathtt{N}(\mu_Q, \boldsymbol{\Lambda}_Q)$ with positive definite $\boldsymbol{\Lambda}_Q$,

$$\bar{D}(P\|Q) \triangleq (\mu_P - \mu_Q)^{\mathrm{T}}\boldsymbol{\Lambda}_Q^{-1}(\mu_P - \mu_Q) + \frac{1}{2}\|\boldsymbol{\Lambda}_Q^{-1/2}(\boldsymbol{\Lambda}_P - \boldsymbol{\Lambda}_Q)\boldsymbol{\Lambda}_Q^{-1/2}\|_{\mathrm{F}}^2. \tag{9.35b}$$

The divergence (9.35b) is a second-order approximation to KL divergence for Gaussian distributions; specifically,

$$\begin{aligned}
&2D\bigl(\mathtt{N}(\mu_P,\mathbf{\Lambda}_P) \,\|\, \mathtt{N}(\mu_Q,\mathbf{\Lambda}_Q)\bigr)\\
&\quad= \operatorname{tr}\bigl(\mathbf{\Lambda}_Q^{-1}\mathbf{\Lambda}_P\bigr) - K - \log\bigl|\mathbf{\Lambda}_P\mathbf{\Lambda}_Q^{-1}\bigr| + (\mu_P-\mu_Q)^{\mathrm{T}}\mathbf{\Lambda}_Q^{-1}(\mu_P-\mu_Q)\\
&\quad= \operatorname{tr}\bigl(\mathbf{\Lambda}_Q^{-1/2}\mathbf{\Lambda}_P\mathbf{\Lambda}_Q^{-1/2} - \mathbf{I}\bigr) - \log\bigl|\mathbf{\Lambda}_Q^{-1/2}\mathbf{\Lambda}_P\mathbf{\Lambda}_Q^{-1/2}\bigr|\\
&\qquad\qquad\qquad\qquad\qquad\qquad + (\mu_P-\mu_Q)^{\mathrm{T}}\mathbf{\Lambda}_Q^{-1}(\mu_P-\mu_Q)\\
&\quad= \frac{1}{2}\bigl\|\mathbf{\Lambda}_Q^{-1/2}\mathbf{\Lambda}_P\mathbf{\Lambda}_Q^{-1/2} - \mathbf{I}\bigr\|_{\mathrm{F}}^2 + (\mu_P-\mu_Q)^{\mathrm{T}}\mathbf{\Lambda}_Q^{-1}(\mu_P-\mu_Q)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + o\bigl(\|\mathbf{\Lambda}_P-\mathbf{\Lambda}_Q\|_{\mathrm{F}}^2\bigr)\\
&\quad= \frac{1}{2}\bigl\|\mathbf{\Lambda}_Q^{-1/2}(\mathbf{\Lambda}_P-\mathbf{\Lambda}_Q)\mathbf{\Lambda}_Q^{-1/2}\bigr\|_{\mathrm{F}}^2 + (\mu_P-\mu_Q)^{\mathrm{T}}\mathbf{\Lambda}_Q^{-1}(\mu_P-\mu_Q)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + o\bigl(\|\mathbf{\Lambda}_P-\mathbf{\Lambda}_Q\|_{\mathrm{F}}^2\bigr)\\
&\quad= \bar{D}\bigl(\mathtt{N}(\mu_P,\mathbf{\Lambda}_P) \,\|\, \mathtt{N}(\mu_Q,\mathbf{\Lambda}_Q)\bigr) + o\bigl(\|\mathbf{\Lambda}_P-\mathbf{\Lambda}_Q\|_{\mathrm{F}}^2\bigr),
\end{aligned}$$

as $\|\mathbf{\Lambda}_P - \mathbf{\Lambda}_Q\|_{\mathrm{F}} \to 0$, where we have used the second-order Taylor series approximation

$$\log|\mathbf{I}+\mathbf{A}| = \operatorname{tr}(\mathbf{A}) - \frac{1}{2}\|\mathbf{A}\|_{\mathrm{F}}^2 + o\bigl(\|\mathbf{A}\|_{\mathrm{F}}^2\bigr), \qquad \|\mathbf{A}\|_{\mathrm{F}} \to 0.$$

Just as $D(\cdot\|\cdot)$, is invariant to a change of coordinates, $\bar{D}(\cdot\|\cdot)$ is invariant to invertible linear transformation of variables, i.e., mappings of the form $Z' = \mathbf{A}\,Z + \mathbf{c}$ with nonsingular $\mathbf{A}$. In particular, we have the following result.

Lemma 9.8. Let $\mathtt{N}(\mu_P,\mathbf{\Lambda}_P)$ and $\mathtt{N}(\mu_Q,\mathbf{\Lambda}_Q)$ be K_0-dimensional Gaussian distributions with nonsingular $\mathbf{\Lambda}_Q$. Then for any nonsingular matrix $\mathbf{A}$ vector $\mathbf{c}$ of compatible dimensions,

$$\begin{aligned}
&\bar{D}\bigl(\mathtt{N}(\mu_P,\mathbf{\Lambda}_P) \,\|\, \mathtt{N}(\mu_Q,\mathbf{\Lambda}_Q)\bigr)\\
&\qquad= \bar{D}\bigl(\mathtt{N}(\mathbf{A}\mu_P+\mathbf{c},\mathbf{A}\mathbf{\Lambda}_P\mathbf{A}^{\mathrm{T}}) \,\|\, \mathtt{N}(\mathbf{A}\mu_Q+\mathbf{c},\mathbf{A}\mathbf{\Lambda}_Q\mathbf{A}^{\mathrm{T}})\bigr). \qquad (9.36)
\end{aligned}$$

A proof of this invariance is provided in Appendix G.3, and makes use of the following simple identity.

Lemma 9.9. For any symmetric matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ of equal dimension,

$$\bigl\|\mathbf{M}_1^{1/2}\mathbf{M}_2\mathbf{M}_1^{1/2}\bigr\|_{\mathrm{F}}^2 = \operatorname{tr}\bigl(\mathbf{M}_1\mathbf{M}_2\mathbf{M}_1\mathbf{M}_2\bigr). \qquad (9.37)$$

9.5 Weakly Correlated Variables

An instance of the local analysis of Section 9.4 corresponds to weak correlation between variables, a concept we formally define as follows.

Definition 9.10 (ϵ-Correlation)**.** Let Z and W be zero-mean jointly Gaussian with dimensions K_Z and K_W, respectively. Then Z and W are ϵ-correlated when

$$P_{Z,W} \in \mathcal{N}_{\epsilon}^{K_Z+K_W}(P_Z P_W), \tag{9.38}$$

where P_Z and P_W are the marginal distributions associated with $P_{Z,W}$.

The following lemma, a proof of which is provided in Appendix G.4, is useful in further characterizing ϵ-correlated variables.

Lemma 9.11. For any $\epsilon > 0$ and zero-mean, ϵ-correlated jointly Gaussian variables Z, W characterized by $\mathbf{\Lambda}_Z$, $\mathbf{\Lambda}_W$, and $\mathbf{\Lambda}_{ZW}$, we have

$$\bar{D}(P_{Z,W} \| P_Z P_W) = \epsilon^2 \|\mathbf{\Phi}^{Z|W}\|_{\mathrm{F}}^2, \tag{9.39}$$

where

$$\mathbf{\Phi}^{Z|W} \triangleq \frac{1}{\epsilon} \mathbf{\Lambda}_Z^{-1/2} \mathbf{\Lambda}_{ZW} \mathbf{\Lambda}_W^{-1/2}, \tag{9.40}$$

which we refer to as the *innovation matrix*.

In particular, it follows immediately from Lemma 9.11 that Z, W being ϵ-correlated is equivalent to the condition

$$\|\mathbf{\Phi}^{Z|W}\|_{\mathrm{F}}^2 \le K_Z + K_W. \tag{9.41}$$

It also follows that Z, W are ϵ-correlated when, on average, $P_{Z|W}(\cdot|w) \in \mathcal{N}_{\epsilon}^{K_Z+K_W}(P_Z)$. The following lemma is useful in establishing this result; a proof is provided in Appendix G.5.

Lemma 9.12. For $\epsilon > 0$ and zero-mean, ϵ-correlated jointly Gaussian variables Z, W characterized by $\mathbf{\Lambda}_Z$, $\mathbf{\Lambda}_W$, and $\mathbf{\Lambda}_{ZW}$,

$$\bar{D}(P_{Z|W}(\cdot|w) \| P_Z) = \epsilon^2 \|\mathbf{\Phi}^{Z|W} w\|^2 + o(\epsilon^2), \quad \epsilon \to 0, \tag{9.42}$$

where $\mathbf{\Phi}^{Z|W}$ is as defined in (9.40).

Our further equivalent condition for ϵ-correlation is then a consequence of the following result, a proof of which is provided in Appendix G.6.

Lemma 9.13. For $\epsilon > 0$ and zero-mean, ϵ-correlated jointly Gaussian variables Z, W characterized by $\mathbf{\Lambda}_Z$, $\mathbf{\Lambda}_W$, and $\mathbf{\Lambda}_{ZW}$, we have[5]

$$\mathbb{E}_{P_W}\left[\bar{D}(P_{Z|W}(\cdot|W)\|P_Z)\right] = \bar{D}(P_{Z,W}\|P_Z P_W)(1+\mathrm{o}(1)), \quad \epsilon \to 0. \tag{9.44}$$

Finally, yet another such equivalent notion of ϵ-correlation is

$$I(Z;W) = D(P_{Z,W}\|P_Z P_W) \le \epsilon^2 (K_Z + K_W) \tag{9.45}$$

where for Gaussian distributions, KL divergence takes the familiar form

$$\begin{aligned} &D\big(\mathtt{N}(\mu_P, \mathbf{\Lambda}_P)\|\mathtt{N}(\mu_Q, \mathbf{\Lambda}_Q)\big) \\ &\quad = \frac{1}{2}\Big[(\mu_P - \mu_Q)^{\mathrm{T}} \mathbf{\Lambda}_Q^{-1}(\mu_P - \mu_Q) + \mathrm{tr}\big(\mathbf{\Lambda}_Q^{-1}\mathbf{\Lambda}_P - \mathbf{I}\big) - \log\big|\mathbf{\Lambda}_P \mathbf{\Lambda}_Q^{-1}\big|\Big]. \end{aligned} \tag{9.46}$$

To establish (9.45), we make use of the following simple fact, whose proof is provided in Appendix G.7.

Fact 9.14. For $\delta > 0$ and any matrix $\mathbf{A}$,

$$\log\big|\mathbf{I} - \delta \mathbf{A}\,\mathbf{A}^{\mathrm{T}}\big| = -\delta \|\mathbf{A}\|_{\mathrm{F}}^2 + \mathrm{o}(\delta), \quad \delta \to 0.$$

As a first step, we have the following result, a proof of which is provided in Appendix G.8.

Lemma 9.15. For $\epsilon > 0$ and zero-mean, ϵ-correlated jointly Gaussian variables Z, W characterized by $\mathbf{\Lambda}_Z$, $\mathbf{\Lambda}_W$, and $\mathbf{\Lambda}_{ZW}$,

$$D\big(P_{Z|W}(\cdot|w)\;\|\;P_Z\big) = \frac{1}{2}\,\bar{D}\big(P_{Z|W}(\cdot|w)\;\|\;P_Z\big) + \mathrm{o}(\epsilon^2), \quad \epsilon \to 0. \tag{9.47}$$

[5] Note, too, that by symmetry we have

$$\mathbb{E}_{P_Z}\left[\bar{D}(P_{W|Z}(\cdot|Z)\|P_W)\right] = \mathbb{E}_{P_W}\left[\bar{D}(P_{Z|W}(\cdot|W)\|P_Z)\right]. \tag{9.43}$$

Indeed, $\mathbf{\Phi}^{W|Z} = \big(\mathbf{\Phi}^{Z|W}\big)^{\mathrm{T}}$.

The equivalence (9.45) is then an immediate consequence of the following corollary, whose proof is provided in Appendix G.9.

Corollary 9.16. For $\epsilon > 0$ and zero-mean, ϵ-correlated jointly Gaussian variables Z, W characterized by $\mathbf{\Lambda}_Z$, $\mathbf{\Lambda}_W$, and $\mathbf{\Lambda}_{ZW}$,

$$I(Z;W) = \frac{1}{2}\bar{D}(P_{Z,W}\|P_Z P_W)(1+o(1)), \quad \epsilon \to 0. \tag{9.48}$$

Finally, the following lemma is a useful generalization; a proof is provided in Appendix G.10.

Lemma 9.17. Let $P_{X,Y}$ and $Q_{X,Y}$ denote candidate jointly Gaussian distributions for ϵ-correlated variables X, Y with given covariances $\mathbf{\Lambda}_X$, $\mathbf{\Lambda}_Y$, and $\epsilon > 0$, where $\mathbf{\Lambda}^P_{XY}$ and $\mathbf{\Lambda}^Q_{XY}$ denote the respective cross-covariances. Then

$$D(P_{X,Y}\|Q_{X,Y}) = \frac{1}{2}\|\tilde{\mathbf{B}}_P - \tilde{\mathbf{B}}_Q\|_{\mathrm{F}}^2 + o(\epsilon^2), \quad \epsilon \to 0, \tag{9.49}$$

where [cf. (9.11)]

$$\tilde{\mathbf{B}}_P = \mathbf{\Lambda}_Y^{-1/2}\mathbf{\Lambda}^P_{YX}\mathbf{\Lambda}_X^{-1/2} \tag{9.50a}$$

$$\tilde{\mathbf{B}}_Q = \mathbf{\Lambda}_Y^{-1/2}\mathbf{\Lambda}^Q_{YX}\mathbf{\Lambda}_X^{-1/2}. \tag{9.50b}$$

9.6 Modal Decomposition of Jointly Gaussian Distributions

Section 9.2 describes how the SVD of $\tilde{\mathbf{B}}$ provides a modal decomposition of covariance for the jointly Gaussian X, Y model. As related analysis, this section describes how in the weak correlation regime, this SVD also provides a modal decomposition of mutual information and, more generally, the joint distribution $P_{X,Y}$.

First, since

$$\mathbf{\Phi}^{Y|X} = \frac{1}{\epsilon}\tilde{\mathbf{B}}, \tag{9.51}$$

specializing Lemma 9.11, we obtain that X, Y are ϵ-correlated when

$$\|\tilde{\mathbf{B}}\|_{\mathrm{F}}^2 = \sum_{i=1}^{K}\sigma_i^2 \leq \epsilon^2(K_X + K_Y). \tag{9.52}$$

In turn, when X, Y are ϵ-correlated we have, specializing Corollary 9.16,

$$I(X;Y) = \frac{1}{2}\sum_{i=1}^{K} \sigma_i^2 + o(\epsilon^2). \tag{9.53}$$

An interpretation of (9.53) is obtained in terms of the modal decomposition of $P_{X,Y}$, as we now describe. In this Gaussian scenario, in contrast to the finite alphabet case, the SVD is both a logarithmic-domain one and asymptotic. In particular, observe that with $\mathbf{\Lambda}_{\tilde{A}}$ as given by (9.10) for $\tilde{A}$ as defined in (9.9), we have

$$\begin{aligned}
\mathbf{\Lambda}_{\tilde{A}}^{-1} &= \begin{bmatrix} (\mathbf{I}-\tilde{\mathbf{B}}^{\mathrm{T}}\tilde{\mathbf{B}})^{-1} & -\tilde{\mathbf{B}}^{\mathrm{T}}(\mathbf{I}-\tilde{\mathbf{B}}\tilde{\mathbf{B}}^{\mathrm{T}})^{-1} \\ -\tilde{\mathbf{B}}(\mathbf{I}-\tilde{\mathbf{B}}^{\mathrm{T}}\tilde{\mathbf{B}})^{-1} & (\mathbf{I}-\tilde{\mathbf{B}}\tilde{\mathbf{B}}^{\mathrm{T}})^{-1} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I}+\tilde{\mathbf{B}}^{\mathrm{T}}\tilde{\mathbf{B}} & -\tilde{\mathbf{B}}^{\mathrm{T}} \\ -\tilde{\mathbf{B}} & \mathbf{I}+\tilde{\mathbf{B}}\tilde{\mathbf{B}}^{\mathrm{T}} \end{bmatrix} + o(\epsilon^2) \qquad (9.54) \\
&= \begin{bmatrix} \mathbf{I} & -\tilde{\mathbf{B}}^{\mathrm{T}} \\ -\tilde{\mathbf{B}} & \mathbf{I} \end{bmatrix} + o(\epsilon), \quad \epsilon \to 0, \qquad (9.55)
\end{aligned}$$

whence

$$\begin{aligned}
P_{\tilde{X},\tilde{Y}}(\tilde{x},\tilde{y}) &= \exp\left\{-\frac{K_A}{2}\log(2\pi) - \frac{1}{2}\tilde{x}^{\mathrm{T}}\tilde{x} - \frac{1}{2}\tilde{y}^{\mathrm{T}}\tilde{y} + \tilde{y}^{\mathrm{T}}\tilde{\mathbf{B}}\,\tilde{x} + o(\epsilon)\right\} \\
&= P_{\tilde{X}}(\tilde{x})\,P_{\tilde{Y}}(\tilde{y})\,\exp\left\{\sum_{i=1}^{K}\sigma_i\,\tilde{x}^{\mathrm{T}}\boldsymbol{\psi}_i^X(\boldsymbol{\psi}_i^Y)^{\mathrm{T}}\tilde{y} + o(\epsilon)\right\} \\
&= P_{\tilde{X}}(\tilde{x})\,P_{\tilde{Y}}(\tilde{y})\,\prod_{i=1}^{K}\exp\left\{\sigma_i\,\tilde{x}^{\mathrm{T}}\boldsymbol{\psi}_i^X(\boldsymbol{\psi}_i^Y)^{\mathrm{T}}\tilde{y}\right\}(1+o(1)),
\end{aligned}$$

as $\epsilon \to 0$, with $K_A = K_X + K_Y$.

As a result, we have

$$\begin{aligned}
P_{X,Y}(x,y) &= |\mathbf{\Lambda}_X|^{-1/2}|\mathbf{\Lambda}_Y|^{-1/2}\,P_{\tilde{X},\tilde{Y}}(\mathbf{\Lambda}_X^{-1/2}x, \mathbf{\Lambda}_Y^{-1/2}y) \\
&= P_X(x)\,P_Y(y)\left(\prod_{i=1}^{K}\mathrm{e}^{\sigma_i\,f_i^*(x)\,g_i^*(y)}\right)(1+o(1)), \qquad (9.56)
\end{aligned}$$

as $\epsilon \to 0$, where f_i^* and g_i^* are the linear functions (9.15) determined in Proposition 9.4.

Furthermore, meaningful approximations to this joint distribution arise by considering, for $k < K$,

$$\tilde{\mathbf{B}}_*^{(k)} \triangleq \mathbf{\Psi}_{(k)}^Y\,\mathbf{\Sigma}_{(k)}(\mathbf{\Psi}_{(k)}^X)^{\mathrm{T}}, \tag{9.57a}$$

where

$$\boldsymbol{\Sigma}_{(k)} \triangleq \left(\boldsymbol{\Psi}^{Y}_{(k)}\right)^{\mathrm{T}} \tilde{\mathbf{B}}\, \boldsymbol{\Psi}^{X}_{(k)} \tag{9.57b}$$

is a diagonal matrix whose diagonal elements are $\sigma_1, \ldots, \sigma_k$. In particular, if we let $X^{(k)}$ and $Y^{(k)}$ denote zero-mean jointly Gaussian variables with the same marginals as X and Y, respectively, but covariance[6] [cf. (9.18)]

$$\begin{aligned}\boldsymbol{\Lambda}^{(k)*}_{YX} \triangleq \boldsymbol{\Lambda}_{Y^{(k)}X^{(k)}} &= \boldsymbol{\Lambda}_{Y}^{1/2}\, \tilde{\mathbf{B}}^{(k)}_{*}\, \boldsymbol{\Lambda}_{X}^{1/2} \\ &= \left(\mathbf{G}^{*}_{(k)}\right)^{\dagger\mathrm{T}} \boldsymbol{\Sigma}_{(k)} \left(\mathbf{F}^{*}_{(k)}\right)^{\dagger} \\ &= \boldsymbol{\Lambda}_{Y}\, \mathbf{G}^{*}_{(k)}\, \boldsymbol{\Sigma}_{(k)} \left(\mathbf{F}^{*}_{(k)}\right)^{\mathrm{T}} \boldsymbol{\Lambda}_{X},\end{aligned} \tag{9.58}$$

where $\mathbf{F}^{*}_{(k)}$ and $\mathbf{G}^{*}_{(k)}$ are as defined in (9.34) and $\boldsymbol{\Sigma}_{(k)}$ is as defined in (9.57b), then it follows that the joint distribution of these new variables takes the form

$$\begin{aligned}P^{(k)*}_{X,Y}(x,y) &\triangleq P_{X^{(k)},Y^{(k)}}(x,y) \\ &= P_X(x)\, P_Y(y) \left(\prod_{i=1}^{k} \mathrm{e}^{\sigma_i\, f^{*}_{i}(x)\, g^{*}_{i}(y)}\right)(1 + o(1)),\end{aligned}$$

and

$$I(X^{(k)}; Y^{(k)}) = \frac{1}{2} \sum_{i=1}^{k} \sigma_i^2 + o(\epsilon^2).$$

9.7 Latent Gaussian Attributes and Statistical Model

In this section, we describe useful interpretations of the modal decomposition for Gaussian variables in terms of latent variable analysis, in a manner analogous to that of Section 5.2 for distributions over finite-alphabets. In this case, our development is more directly related to its roots in factor analysis [13], [26] as introduced by Spearman [254].

[6] Note that $\boldsymbol{\Lambda}^{(k)*}_{YX}$ so-defined is a valid cross covariance matrix, i.e.,

$$\begin{bmatrix} \boldsymbol{\Lambda}_{Y} & \boldsymbol{\Lambda}^{(k)*}_{YX} \\ \left(\boldsymbol{\Lambda}^{(k)*}_{YX}\right)^{\mathrm{T}} & \boldsymbol{\Lambda}_{X} \end{bmatrix}$$

is positive definite, as can be verified using Fact 9.3.

We begin with the introduction of latent Gaussian attribute variables. Although our definition includes a correlation constraint, in the Gaussian case analysis we do not limit our attention to the vanishing correlation regime.

Definition 9.18 (Gaussian ϵ-Attribute)**.** For [cf. (9.41)]

$$0 < \epsilon \leq \sqrt{\frac{K_W}{K_Z + K_W}}, \tag{9.59}$$

the variable $W \in \mathbb{R}^{K_W}$ is a Gaussian ϵ-attribute of $Z \in \mathbb{R}^{K_Z}$ if: 1) $K_W \leq K_Z$ and $\mathbf{\Lambda}_W$ is nonsingular; 2) W, Z are jointly Gaussian; 3) W, Z are ϵ-correlated but $\mathbf{\Lambda}_{WZ} \neq \mathbf{0}$; and 4) W conditionally independent of all other variables in the model given Z.

Definition 9.19 (Gaussian ϵ-Attribute Configuration)**.** Given a zero-mean Gaussian variable $Z \in \mathbb{R}^{K_Z}$ with covariance $\mathbf{\Lambda}_Z$, then for ϵ satisfying (9.59), an ϵ-attribute W of Z is characterized by its configuration [cf. (9.41)]

$$\mathcal{C}_\epsilon^{K_Z}(\mathbf{\Lambda}_Z) = \left\{ K_W,\, \mathbf{\Lambda}_W,\, \mathbf{\Phi}^{Z|W} \colon \left\| \mathbf{\Phi}^{Z|W} \right\|_{\mathrm{F}}^2 \leq K_Z + K_W \right\}, \tag{9.60}$$

with $\mathbf{\Phi}^{Z|W}$ as defined in (9.40).

As in the case of discrete variables, the notion of a multi-attribute is also useful in the Gaussian case.

Definition 9.20 (Gaussian ϵ-Multi-Attribute)**.** A Gaussian ϵ-multi-attribute is a Gaussian ϵ-attribute satisfying the additional property that

$$\left\| \mathbf{\Phi}^{Z|W} \right\|_{\mathrm{s}}^2 \leq \frac{K_Z + K_W}{K_W}, \tag{9.61}$$

with $\mathbf{\Phi}^{Z|W}$ as defined in (9.40).

Note that (9.61) is a stronger version of the ϵ correlation property, since $\left\| \mathbf{\Phi}^{Z|W} \right\|_{\mathrm{s}} \leq \left\| \mathbf{\Phi}^{Z|W} \right\|_{\mathrm{F}} \leq K_W \left\| \mathbf{\Phi}^{Z|W} \right\|_{\mathrm{s}}$.

Definition 9.21 (Gaussian ϵ-Multi-Attribute Configuration)**.** Given a zero-mean Gaussian variable $Z \in \mathbb{R}^{K_Z}$ with covariance $\mathbf{\Lambda}_Z$, then for ϵ

satisfying (9.59), an ϵ-multi-attribute W of Z is characterized by its configuration [cf. (9.61)]

$$\bar{\mathcal{C}}_{\epsilon}^{K_Z}(\mathbf{\Lambda}_Z) = \left\{K_W,\, \mathbf{\Lambda}_W,\, \mathbf{\Phi}^{Z|W} \colon \|\mathbf{\Phi}^{Z|W}\|_{\mathrm{s}}^2 \leq \frac{K_Z + K_W}{K_W}\right\}, \tag{9.62}$$

with $\mathbf{\Phi}^{Z|W}$ as defined in (9.40).

For inferences about an attribute W, we consider features of the form

$$h(Z) = \mathbf{H}^{\mathrm{T}} Z = (\mathbf{\Xi})^{\mathrm{T}} \tilde{Z},$$

where $\mathbf{\Xi}$ is the associated feature weight matrix. Without loss of generality, we restrict our attention to normalized (zero-mean) $h(Z)$, so that[7]

$$\mathbb{E}\big[h(Z)\, h(Z)^{\mathrm{T}}\big] = \mathbf{H}^{\mathrm{T}} \mathbf{\Lambda}_Z \mathbf{H} = (\mathbf{\Xi})^{\mathrm{T}} \mathbf{\Xi} = \mathbf{I}.$$

In the context of a given model $P_{X,Y}$, the Gaussian ϵ-attribute variables U and V for X and Y, respectively, are characterized by the (now Gauss-) Markov structure

$$U \leftrightarrow X \leftrightarrow Y \leftrightarrow V, \tag{9.63}$$

where $U \in \mathbb{R}^{K_U}$ and $V \in \mathbb{R}^{K_V}$ for some dimensions K_U and K_V.

The following familiar fact will be useful, whose short proof we provide for convenience in Appendix G.12.

Fact 9.22. Normalized zero-mean Gaussian variables $\tilde{Z}_1, \tilde{Z}_2, \tilde{Z}_3$ form a Markov chain $\tilde{Z}_1 \leftrightarrow \tilde{Z}_2 \leftrightarrow \tilde{Z}_3$ if and only if

$$\mathbf{\Lambda}_{\tilde{Z}_1 \tilde{Z}_3} = \mathbf{\Lambda}_{\tilde{Z}_1 \tilde{Z}_2} \mathbf{\Lambda}_{\tilde{Z}_2 \tilde{Z}_3}. \tag{9.64}$$

[7] Indeed, if they were not, so long as the columns of $\mathbf{\Xi}$ are linearly independent, so $(\mathbf{\Xi})^{\mathrm{T}}\mathbf{\Xi}$ is invertible, we can transform $\mathbf{H}$ into $\tilde{\mathbf{H}}$ via

$$\tilde{\mathbf{H}} = \mathbf{H}\Big((\mathbf{\Xi})^{\mathrm{T}}\mathbf{\Xi}\Big)^{-1/2} = \underbrace{\mathbf{\Xi}\Big((\mathbf{\Xi})^{\mathrm{T}}\mathbf{\Xi}\Big)^{-1/2}}_{\triangleq \tilde{\mathbf{\Xi}}},$$

where we note $(\tilde{\mathbf{\Xi}})^{\mathrm{T}}\tilde{\mathbf{\Xi}} = \mathbf{I}$.

As an application of Fact 9.22, we have, for example,

$$\boldsymbol{\Phi}^{Y|U} = \tilde{\mathbf{B}}\,\boldsymbol{\Phi}^{X|U} \tag{9.65a}$$

$$\boldsymbol{\Phi}^{X|V} = \tilde{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\Phi}^{Y|V}. \tag{9.65b}$$

For inferences about attributes U and V, we will generally consider statistics of the form

$$S_{(k)} \triangleq \left(\mathbf{F}_{(k)}\right)^{\mathrm{T}} X \triangleq \begin{bmatrix} f_1(X) & \cdots & f_k(X) \end{bmatrix}^{\mathrm{T}} \tag{9.66a}$$

$$T_{(k)} \triangleq \left(\mathbf{G}_{(k)}\right)^{\mathrm{T}} Y \triangleq \begin{bmatrix} g_1(Y) & \cdots & g_k(Y) \end{bmatrix}^{\mathrm{T}} \tag{9.66b}$$

for some dimension $k \in \{1, \dots, K\}$ and feature matrices $\mathbf{F}_{(k)} \in \mathbb{R}^{K_X \times k}$ and $\mathbf{G}_{(k)} \in \mathbb{R}^{K_Y \times k}$. Without loss of generality we restrict our attention to normalized features, i.e. $(\mathbf{F}_{(k)}, \mathbf{G}_{(k)}) \in \mathcal{L}$ with $\mathcal{L}$ as defined in (9.27). As we will develop, the particular choices $S^*_{(k)}, T^*_{(k)}$ defined in (9.24) play a special role.

For arbitrary jointly Gaussian W and Z, we use $\lambda_{\mathrm{e}}^{W|Z}$ to denote the MSE in the MMSE estimate of W based on Z, so with respect to our specific variables of interest, $\lambda_{\mathrm{e}}^{U|S}(\mathbf{F}_{(k)})$, $\lambda_{\mathrm{e}}^{V|S}(\mathbf{F}_{(k)})$, $\lambda_{\mathrm{e}}^{U|T}(\mathbf{G}_{(k)})$, and $\lambda_{\mathrm{e}}^{V|T}(\mathbf{G}_{(k)})$ denote the associated MSEs, with their dependencies on $\mathbf{F}_{(k)}$ and $\mathbf{G}_{(k)}$ made explicit.

9.8 MMSE Universal Features

In this formulation, we seek to determine optimum k-dimensional features for estimating a pair of unknown Gaussian attributes (U, V) for (X, Y) in the Gauss-Markov model (9.63), where $k \in \{1, \dots, K\}$.

As in the finite-alphabet setting, we view the configurations of attributes U and V as randomly drawn by nature from a RIE. In this case, this ensemble is also defined via the spherical symmetry of Definition 5.7.

Definition 9.23 (Gaussian Rotation Invariant Ensemble). Given ϵ satisfying (9.59), the Gaussian rotationally invariant ensemble (RIE) for an attribute W of a Gaussian variable Z is the collection of all jointly

Gaussian attribute configurations of the form (9.60) together with a probability measure over the collection such that $\mathbf{\Phi}^{Z|W}$ is spherically symmetric.

Let $\mathcal{C}_{\epsilon_X}^{K_X}(\mathbf{\Lambda}_X)$ and $\mathcal{C}_{\epsilon_Y}^{K_Y}(\mathbf{\Lambda}_Y)$ denote configurations for attributes U and V, respectively, in the sense of Definition 9.19, i.e.,

$$\mathcal{C}_{\epsilon_X}^{K_X}(\mathbf{\Lambda}_X) = \Big\{K_U,\, \mathbf{\Lambda}_U,\, \mathbf{\Phi}^{X|U} \colon \|\mathbf{\Phi}^{X|U}\|_{\mathrm{F}}^2 \le K_U + K_X\Big\} \tag{9.67a}$$

$$\mathcal{C}_{\epsilon_Y}^{K_Y}(\mathbf{\Lambda}_Y) = \Big\{K_V,\, \mathbf{\Lambda}_V,\, \mathbf{\Phi}^{Y|V} \colon \|\mathbf{\Phi}^{Y|V}\|_{\mathrm{F}}^2 \le K_V + K_Y\Big\}, \tag{9.67b}$$

where [cf. (9.59)]

$$0 < \epsilon_X^2 \le \frac{K_U}{K_U + K_X} \quad \text{and} \quad 0 < \epsilon_Y^2 \le \frac{K_V}{K_V + K_Y}. \tag{9.68}$$

In what follows, we denote the MSE in the MMSE estimates U and V based on $S_{(k)}$ as defined in (9.66a), respectively, via

$$\lambda_{\mathrm{e}}^{U|S}\big(\mathcal{C}_{\epsilon_X}^{K_X}(\mathbf{\Lambda}_X), \mathbf{F}_{(k)}\big) \quad \text{and} \quad \lambda_{\mathrm{e}}^{V|S}\big(\mathcal{C}_{\epsilon_Y}^{K_Y}(\mathbf{\Lambda}_Y), \mathbf{F}_{(k)}\big), \tag{9.69a}$$

and those for the MMSE estimates based on $T_{(k)}$ as defined in (9.66b) via, respectively,

$$\lambda_{\mathrm{e}}^{U|T}\big(\mathcal{C}_{\epsilon_X}^{K_X}(\mathbf{\Lambda}_X), \mathbf{G}_{(k)}\big) \quad \text{and} \quad \lambda_{\mathrm{e}}^{V|T}\big(\mathcal{C}_{\epsilon_Y}^{K_Y}(\mathbf{\Lambda}_Y), \mathbf{G}_{(k)}\big). \tag{9.69b}$$

In turn, we let

$$\bar{\lambda}_{\mathrm{e}}^{U|S}(\mathbf{F}_{(k)}) \triangleq \mathbb{E}_{\mathrm{RIE}}\big[\lambda_{\mathrm{e}}^{U|S}\big(\mathcal{C}_{\epsilon_X}^{K_X}(\mathbf{\Lambda}_X), \mathbf{F}_{(k)}\big)\big] \tag{9.70a}$$

$$\bar{\lambda}_{\mathrm{e}}^{V|S}(\mathbf{F}_{(k)}) \triangleq \mathbb{E}_{\mathrm{RIE}}\big[\lambda_{\mathrm{e}}^{V|S}\big(\mathcal{C}_{\epsilon_Y}^{K_Y}(\mathbf{\Lambda}_Y), \mathbf{F}_{(k)}\big)\big] \tag{9.70b}$$

$$\bar{\lambda}_{\mathrm{e}}^{U|T}(\mathbf{G}_{(k)}) \triangleq \mathbb{E}_{\mathrm{RIE}}\big[\lambda_{\mathrm{e}}^{U|T}\big(\mathcal{C}_{\epsilon_X}^{K_X}(\mathbf{\Lambda}_X), \mathbf{G}_{(k)}\big)\big] \tag{9.70c}$$

$$\bar{\lambda}_{\mathrm{e}}^{V|T}(\mathbf{G}_{(k)}) \triangleq \mathbb{E}_{\mathrm{RIE}}\big[\lambda_{\mathrm{e}}^{V|T}\big(\mathcal{C}_{\epsilon_Y}^{K_Y}(\mathbf{\Lambda}_Y), \mathbf{G}_{(k)}\big)\big], \tag{9.70d}$$

where $\mathbb{E}_{\mathrm{RIE}}[\cdot]$ denotes expectation with respect to the Gaussian RIEs for $\mathcal{C}_{\epsilon_X}^{K_X}(\mathbf{\Lambda}_X)$ and $\mathcal{C}_{\epsilon_Y}^{K_Y}(\mathbf{\Lambda}_Y)$.

For this scenario, we have following result, a proof of which is provided in Appendix G.13.

Proposition 9.24. Given zero-mean jointly Gaussian $X \in \mathbb{R}^{K_X}$, $Y \in \mathbb{R}^{K_Y}$ characterized by $\mathbf{\Lambda}_X$, $\mathbf{\Lambda}_Y$, and $\mathbf{\Lambda}_{XY}$, and attributes U and V of X

and Y, respectively, each drawn from a Gaussian RIE for some ϵ_X and ϵ_Y, respectively, satisfying (9.68), then for any dimension $k \in \{1, \dots, K\}$, the multi-objective minimization

$$\min_{(\mathbf{F}_{(k)}, \mathbf{G}_{(k)}) \in \mathcal{L}} \Big(\bar{\lambda}_{\mathrm{e}}^{U|S}(\mathbf{F}_{(k)}), \bar{\lambda}_{\mathrm{e}}^{V|S}(\mathbf{F}_{(k)}), \bar{\lambda}_{\mathrm{e}}^{U|T}(\mathbf{G}_{(k)}), \bar{\lambda}_{\mathrm{e}}^{V|T}(\mathbf{G}_{(k)}) \Big) \tag{9.71}$$

has a unique Pareto optimal solution, which is achieved by $\big(\mathbf{F}_{(k)}^*, \mathbf{G}_{(k)}^*\big)$ as defined in (9.19). Moreover,

$$\bar{\lambda}_{\mathrm{e}}^{U|S}(\mathbf{F}_{(k)}^*) = \mathrm{tr}(\mathbf{\Lambda}_U) \Big[1 - \epsilon_X^2 \bar{E}_0^{X|U}\, k \Big] \tag{9.72a}$$

$$\bar{\lambda}_{\mathrm{e}}^{V|S}(\mathbf{F}_{(k)}^*) = \mathrm{tr}(\mathbf{\Lambda}_V) \left[1 - \epsilon_Y^2 \bar{E}_0^{Y|V} \sum_{i=1}^{k} \sigma_i^2 \right] \tag{9.72b}$$

$$\bar{\lambda}_{\mathrm{e}}^{U|T}(\mathbf{G}_{(k)}^*) = \mathrm{tr}(\mathbf{\Lambda}_U) \left[1 - \epsilon_X^2 \bar{E}_0^{X|U} \sum_{i=1}^{k} \sigma_i^2 \right] \tag{9.72c}$$

$$\bar{\lambda}_{\mathrm{e}}^{V|T}(\mathbf{G}_{(k)}^*)) = \mathrm{tr}(\mathbf{\Lambda}_V) \Big[1 - \epsilon_Y^2 \bar{E}_0^{Y|V}\, k \Big], \tag{9.72d}$$

where $\bar{E}_0^{X|U}$ and $\bar{E}_0^{Y|V}$ are nonnegative constants that do not depend on ϵ_X, ϵ_Y, k, or $P_{X,Y}$.

We emphasize that Proposition 9.24 is not asymptotic: we do not require $\epsilon_X, \epsilon_Y \to 0$.

9.9 MMSE Cooperative Game

A characterization of the associated cooperative game for MSE minimization, in which nature chooses the attribute that can be most accurately estimated, is given by the following. A proof is provided in Appendix G.14.

Proposition 9.25. Given zero-mean jointly Gaussian $X \in \mathbb{R}^{K_X}$, $Y \in \mathbb{R}^{K_Y}$ characterized by $\mathbf{\Lambda}_X$, $\mathbf{\Lambda}_Y$, and $\mathbf{\Lambda}_{XY}$, parameters ϵ_X, ϵ_Y of multi-attributes U and V, respectively, satisfying (9.68), and a dimension $k \in \{1, \dots, K\}$, then the multi-objective minimization

$$\min_{\substack{(\mathcal{C}^{K_X}_{\epsilon_X}(\mathbf{\Lambda}_X),\mathcal{C}^{K_Y}_{\epsilon_Y}(\mathbf{\Lambda}_Y))\in\mathcal{C}_{(k)},\\ (\mathbf{F}_{(k)},\mathbf{G}_{(k)})\in\mathcal{L}}} \Big(\lambda_{\mathrm{e}}^{U|S}(\mathcal{C}^{K_X}_{\epsilon_X}(\mathbf{\Lambda}_X),\mathbf{F}_{(k)}), \lambda_{\mathrm{e}}^{V|S}(\mathcal{C}^{K_Y}_{\epsilon_Y}(\mathbf{\Lambda}_Y),\mathbf{F}_{(k)}), \lambda_{\mathrm{e}}^{U|T}(\mathcal{C}^{K_X}_{\epsilon_X}(\mathbf{\Lambda}_X),\mathbf{G}_{(k)}), \lambda_{\mathrm{e}}^{V|T}(\mathcal{C}^{K_Y}_{\epsilon_Y}(\mathbf{\Lambda}_Y),\mathbf{G}_{(k)})\Big), \tag{9.73}$$

where

$$\mathcal{C}_{(k)} \triangleq \Big\{(\mathcal{C}^{K_X}_{\epsilon_X}(\mathbf{\Lambda}_X),\mathcal{C}^{K_Y}_{\epsilon_Y}(\mathbf{\Lambda}_Y))\colon K_U = K_V = k, \left\|\mathbf{\Lambda}_U^{-1}\right\|_{\mathrm{s}} \le 1, \quad \left\|\mathbf{\Lambda}_V^{-1}\right\|_{\mathrm{s}} \le 1, \Big\}, \tag{9.74}$$

has a unique Pareto optimal solution, which is achieved by $(\mathbf{F}^*_{(k)},\mathbf{G}^*_{(k)})$ as defined in (9.19), and $(\bar{\mathcal{C}}^{K_X}_{\epsilon_X,*}(\mathbf{\Lambda}_X),\bar{\mathcal{C}}^{K_Y}_{\epsilon_Y,*}(\mathbf{\Lambda}_Y))$ characterized by

$$\mathbf{\Lambda}_U = \mathbf{\Lambda}_V = \mathbf{I} \tag{9.75a}$$

and

$$\mathbf{\Lambda}_{XU} = \epsilon_X \sqrt{\frac{K_X+k}{k}}\,\mathbf{\Lambda}_X\,\mathbf{F}^*_{(k)} \tag{9.75b}$$

$$\mathbf{\Lambda}_{YV} = \epsilon_Y \sqrt{\frac{K_Y+k}{k}}\,\mathbf{\Lambda}_Y\,\mathbf{G}^*_{(k)}. \tag{9.75c}$$

Moreover,

$$\lambda_{\mathrm{e}}^{U|S}(\bar{\mathcal{C}}^{K_X}_{\epsilon_X,*}(\mathbf{\Lambda}_X),\mathbf{F}^*_{(k)},\mathbf{G}^*_{(k)}) = k - \epsilon_X^2(K_X+k) \tag{9.76a}$$

$$\lambda_{\mathrm{e}}^{V|S}(\bar{\mathcal{C}}^{K_Y}_{\epsilon_Y,*}(\mathbf{\Lambda}_Y),\mathbf{F}^*_{(k)},\mathbf{G}^*_{(k)}) = k - \epsilon_Y^2\bigg(\frac{K_Y+k}{k}\bigg)\sum_{i=1}^{k}\sigma_i^2 \tag{9.76b}$$

$$\lambda_{\mathrm{e}}^{U|T}(\bar{\mathcal{C}}^{K_X}_{\epsilon_X,*}(\mathbf{\Lambda}_X),\mathbf{F}^*_{(k)},\mathbf{G}^*_{(k)}) = k - \epsilon_X^2\bigg(\frac{K_X+k}{k}\bigg)\sum_{i=1}^{k}\sigma_i^2 \tag{9.76c}$$

$$\lambda_{\mathrm{e}}^{V|T}(\bar{\mathcal{C}}^{K_Y}_{\epsilon_Y,*}(\mathbf{\Lambda}_Y),\mathbf{F}^*_{(k)},\mathbf{G}^*_{(k)}) = k - \epsilon_Y^2(K_Y+k). \tag{9.76d}$$

Note that Proposition 9.25 is also not asymptotic: it does not require $\epsilon_X,\epsilon_Y \to 0$. Note, too, that via Proposition 9.25 we obtain the multi-attributes U and V for which the features $(\mathbf{F}^*_{(k)},\mathbf{G}^*_{(k)})$ are sufficient statistics.

9.10 The Local Gaussian Information Bottleneck

The following result establishes the optimum attributes in the MMSE cooperative game of Section 9.9 coincide with those of a Gaussian version of the local information double bottleneck problem. A proof is provided in Appendix G.15.

Proposition 9.26. Let $X \in \mathbb{R}^{K_X}$, $Y \in \mathbb{R}^{K_Y}$ be zero-mean jointly Gaussian variables characterized by $\mathbf{\Lambda}_X$, $\mathbf{\Lambda}_Y$, and $\mathbf{\Lambda}_{XY}$, and given $\epsilon_X, \epsilon_Y > 0$, let U and V be Gaussian ϵ_X- and ϵ_Y-multi-attributes of X and Y, respectively, with $K_U = K_V = k$. Then

$$I(U;V) \le \frac{\epsilon_X^2 \epsilon_Y^2}{2} \left(\frac{K_X+k}{k}\right)\left(\frac{K_Y+k}{k}\right) \sum_{i=1}^{k} \sigma_i^2 + o(\epsilon_X^2 \epsilon_Y^2), \quad \epsilon_X, \epsilon_Y \to 0, \tag{9.77}$$

where the inequality holds with equality when the configurations of U and V are given by (9.75), in which case

$$\mathbf{\Lambda}_{UV} = \epsilon_X \epsilon_Y \sqrt{\frac{K_X+k}{k}} \sqrt{\frac{K_Y+k}{k}}\, \mathbf{\Sigma}_{(k)}, \tag{9.78}$$

where $\mathbf{\Sigma}_{(k)}$ is as defined in (9.57b).

Proposition 9.26 can be equivalently expressed in the form of a solution to a symmetric version of the Gaussian information bottleneck problem [55] in the weak dependence regime. In particular, we have the following corollary, a proof of which is provided in Appendix G.16.

Corollary 9.27. Let X, Y be zero-mean jointly Gaussian variables characterized by $\mathbf{\Lambda}_X$, $\mathbf{\Lambda}_Y$, and $\mathbf{\Lambda}_{XY}$, and let U and V be variables in the Gauss-Markov chain (9.63) such that we satisfy the independence relations $\mathbf{\Lambda}_U = \mathbf{\Lambda}_V = \mathbf{I}$, the conditional independence relations that[8] $\mathbf{\Lambda}_{XU}^{\mathrm{T}} \mathbf{\Lambda}_X^{-1} \mathbf{\Lambda}_{XU}$ and $\mathbf{\Lambda}_{YV}^{\mathrm{T}} \mathbf{\Lambda}_Y^{-1} \mathbf{\Lambda}_{YV}$ are diagonal, and the dependence constraints $\max\{I(U_i;X), I(V_i;Y)\} \le \epsilon^2/2$ for $i = 1, \ldots, k$. Then

$$\max_{U,V} I(U;V) = \frac{\epsilon^4}{2} \sum_{i=1}^{k} \sigma_i^2 + o(\epsilon^4), \quad \epsilon \to 0. \tag{9.79}$$

[8] The elements of U are conditionally independent given X when $\mathbf{\Lambda}_{U|X} = \mathbf{\Lambda}_U - \mathbf{\Lambda}_{XU}^{\mathrm{T}} \mathbf{\Lambda}_X^{-1} \mathbf{\Lambda}_{XU}$ is diagonal, and similarly for the V, Y relation.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Moreover, the maximum is achieved by the configurations [cf. (9.75b)–(9.75c)]

$$\mathbf{\Lambda}_{XU} = \epsilon \mathbf{\Lambda}_X \, \mathbf{F}^*_{(k)} \tag{9.80a}$$

$$\mathbf{\Lambda}_{YV} = \epsilon \mathbf{\Lambda}_Y \, \mathbf{G}^*_{(k)}, \tag{9.80b}$$

in which case

$$\mathbf{\Lambda}_{UV} = \epsilon^2 \, \mathbf{\Sigma}_{(k)}. \tag{9.81}$$

with $\mathbf{\Sigma}_{(k)}$ as defined in (9.57b).

It further follows that $(S^*_{(k)}, T^*_{(k)})$ is a sufficient statistic for inferences about the optimizing (U, V), i.e., for any dimension $k \in \{1, \dots, K\}$ we have the Markov chains

$$(U, V) \leftrightarrow (S^*_{(k)}, T^*_{(k)}) \leftrightarrow (X, Y) \tag{9.82}$$

and

$$U \leftrightarrow S^*_{(k)} \leftrightarrow T^*_{(k)} \leftrightarrow V. \tag{9.83}$$

In particular, we have the following result; a proof is provided in Appendix G.17.

Corollary 9.28. In the solution to the optimization in Proposition 9.26,

$$P_{U,V|X,Y}(u, v|x, y) = P_{U|X}(u|x) \, P_{V|Y}(v|y), \tag{9.84}$$

with

$$P_{U|X}(\cdot|x) = \mathtt{N}(\epsilon s^*_{(k)} \, , \, (1 - \epsilon^2) \, \mathbf{I}) \tag{9.85a}$$

$$P_{V|Y}(\cdot|y) = \mathtt{N}(\epsilon t^*_{(k)} \, , \, (1 - \epsilon^2) \, \mathbf{I}), \tag{9.85b}$$

where [cf. (9.24)] $s^*_{(k)} = \mathbf{F}^*_{(k)} \, x$ and $t^*_{(k)} = \mathbf{G}^*_{(k)} \, y$, and where we note that (9.85) depend on (x, y) only through $(s^*_{(k)}, t^*_{(k)})$. Moreover,

$$P_{U|S^*_{(k)}, T^*_{(k)}, V}(u|s^*_{(k)}, t^*_{(k)}, v) = P_{U|S^*_{(k)}}(u|s^*_{(k)}) \tag{9.86a}$$

$$P_{V|S^*_{(k)}, T^*_{(k)}, U}(v|s^*_{(k)}, t^*_{(k)}, u) = P_{V|T^*_{(k)}}(v|t^*_{(k)}), \tag{9.86b}$$

and

$$P_{V|X}(\cdot|x) = \mathtt{N}(\epsilon \mathbf{\Sigma}_{(k)} \, s^*_{(k)} \, , \, \mathbf{I} - \epsilon^2 \, \mathbf{\Sigma}^2_{(k)}) \tag{9.87a}$$

$$P_{U|Y}(\cdot|y) = \mathtt{N}(\epsilon \mathbf{\Sigma}_{(k)} \, t^*_{(k)} \, , \, \mathbf{I} - \epsilon^2 \, \mathbf{\Sigma}^2_{(k)}). \tag{9.87b}$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

We emphasize that the sufficient statistic pair $(S^*_{(k)}, T^*_{(k)})$ involves separate processing of X and Y. We also emphasize that Corollary 9.28 is not an asymptotic result—it holds for finite ϵ.

The more classical one-sided Gaussian information bottleneck problem [55] can also be analyzed in the weak-dependence regime. For example, we have the following result, a proof of which is provided in Appendix G.18.

Proposition 9.29. Let $X \in \mathbb{R}^{K_X}$, $Y \in \mathbb{R}^{K_Y}$ be jointly Gaussian variables characterized by $\mathbf{\Lambda}_X$, $\mathbf{\Lambda}_Y$, and $\mathbf{\Lambda}_{XY}$, and given $\epsilon > 0$, let U and V be variables in the Gauss-Markov chain (9.63) such that we satisfy the independence relations $\mathbf{\Lambda}_U = \mathbf{\Lambda}_V = \mathbf{I}$, the conditional independence relations that $\mathbf{\Lambda}^{\mathrm{T}}_{XU}\mathbf{\Lambda}_X^{-1}\mathbf{\Lambda}_{XU}$ and $\mathbf{\Lambda}^{\mathrm{T}}_{YV}\mathbf{\Lambda}_Y^{-1}\mathbf{\Lambda}_{YV}$ are diagonal, and the dependence constraints $\max\{I(U_i;X), I(V_i;Y)\} \leq \epsilon^2/2$ for $i = 1, \dots, k$. Then

$$\max_U I(U;Y) = \max_V I(V;X) = \frac{\epsilon^2}{2}\sum_{i=1}^{k}\sigma_i^2 + o(\epsilon^2), \quad \epsilon \to 0. \tag{9.88}$$

Moreover, the maximum is achieved by the configurations (9.80).

Note, finally, that (S, T) given by (9.24a) and (9.24b) are sufficient statistics for inferences about the resulting (U, V), which we emphasize are obtained by *separate* processing of X and Y.

9.11 Gaussian Common Information

We now develop the relationship between the optimizing Gaussian multi-attributes (U, V) in Section 9.10 (and Section 9.9), and the common information associated with the pair (X, Y) characterized by a given joint distribution $P_{X,Y}$.

In the Gaussian case, common information can be readily evaluated, without the local restriction of the finite-alphabet case, and takes the following form, as shown in [241, Corollary 1]. For convenience, the proof is provided in Appendix G.19.[9]

[9] As a possible indirect application, aspects of the results of this section may provide useful guidance on the design of *Bayesian CCA* methods [151], [269] that build on the latent variable perspectives of CCA in [18].

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Proposition 9.30. Let $X \in \mathbb{R}^{K_X}$, $Y \in \mathbb{R}^{K_Y}$ be zero-mean jointly Gaussian variables characterized by $\mathbf{\Lambda}_X$, $\mathbf{\Lambda}_Y$, and $\mathbf{\Lambda}_{XY}$. Then

$$C(X,Y) = \min_{\substack{P_{W|X,Y}:\\ X \leftrightarrow W \leftrightarrow Y}} I(W;X,Y) = \frac{1}{2}\sum_{i=1}^{K} \log\left(\frac{1+\sigma_i}{1-\sigma_i}\right). \tag{9.89}$$

Moreover, an optimizing $P_{W|X,Y}$ is Gaussian with

$$\mathbf{\Lambda}_{XW} = \mathbf{\Lambda}_X\,\mathbf{F}^*_{(K)}\,\mathbf{\Sigma}^{1/2}_{(K)} \tag{9.90a}$$

$$\mathbf{\Lambda}_{YW} = \mathbf{\Lambda}_Y\,\mathbf{G}^*_{(K)}\,\mathbf{\Sigma}^{1/2}_{(K)}. \tag{9.90b}$$

Note that since for $0 < \omega < 1$,

$$\frac{1}{2}\log\frac{1+\omega}{1-\omega} \geq \omega,$$

we have

$$C(X;Y) \geq \sum_{i=1}^{K} \sigma_i = \|\tilde{\mathbf{B}}\|_*,$$

where the bound is tight in the limit of weak correlation, i.e.,

$$\frac{C(X;Y)}{\|\tilde{\mathbf{B}}\|_*} \to 1 \qquad \text{as } \|\tilde{\mathbf{B}}\|_* \to 0.$$

We further have

$$W \leftrightarrow R^*_{(K)} \leftrightarrow (S^*_{(K)}, T^*_{(K)}) \leftrightarrow (X,Y). \tag{9.91}$$

where

$$R^*_{(K)} \triangleq S^*_{(K)} + T^*_{(K)}, \tag{9.92}$$

with $S^*_{(K)}$ and $T^*_{(K)}$ as defined in (9.24). In particular, we have the following result, a proof of which is provided in Appendix G.20.

Corollary 9.31. In the solution to the optimization in Proposition 9.30

$$\mathbb{E}[W|X,Y] = \mathbf{\Sigma}^{1/2}_{(K)}\bigl(\mathbf{I}+\mathbf{\Sigma}_{(K)}\bigr)^{-1} R^*_{(K)} \tag{9.93a}$$

$$\mathbf{\Lambda}_{W|X,Y} = \bigl(\mathbf{I}-\mathbf{\Sigma}_{(K)}\bigr)\bigl(\mathbf{I}+\mathbf{\Sigma}_{(K)}\bigr)^{-1}. \tag{9.93b}$$

9.12 Relating Common Information to Dominant Structure

The common information auxiliary variable W of Proposition 9.30 is naturally related to the multi-attributes (U, V) of Proposition 9.26 (and Proposition 9.25) when we choose the normalization

$$\epsilon_X = \sqrt{\frac{k}{K_X + k}} \quad \text{and} \quad \epsilon_Y = \sqrt{\frac{k}{K_Y + k}},$$

or, equivalently, $\epsilon = 1$ in Corollary 9.27. In particular, the following result, a proof of which is provided in Appendix G.21, establishes that common information can be equivalently characterized by

$$C(X,Y) = \min_{\substack{P_{W|X,Y}: \\ X \leftrightarrow W \leftrightarrow Y \\ W \leftrightarrow (U,V) \leftrightarrow (X,Y)}} I(W; X, Y). \tag{9.94}$$

so that the optimizing W satisfies

$$W \leftrightarrow (U,V) \leftrightarrow (S^*_{(K)}, T^*_{(K)}) \leftrightarrow (X,Y). \tag{9.95}$$

Corollary 9.32. Let X, Y be zero-mean jointly Gaussian variables characterized by $\mathbf{\Lambda}_X$, $\mathbf{\Lambda}_Y$, and $\mathbf{\Lambda}_{XY}$, and let (U, V) be the dominant K-dimensional multi-attributes of Corollary 9.27 with $\epsilon = 1$. If $\check{W}$ is chosen so that $\check{W} \leftrightarrow (U, V) \leftrightarrow (X, Y)$ is a Gauss-Markov chain with

$$\mathbf{\Lambda}_{\check{W}U} = \mathbf{\Lambda}_{\check{W}V} = \mathbf{\Sigma}^{1/2}_{(K)}, \tag{9.96}$$

and $\mathbf{\Lambda}_{\check{W}} = \mathbf{I}$, then

$$I(\check{W}; X, Y) = C(X,Y), \tag{9.97}$$

where $C(X, Y)$ is as given in Proposition 9.30.

When $\check{W}$ is constructed according to Corollary 9.32, we have the additional Markov structure

$$\check{W} \leftrightarrow (U+V) \leftrightarrow R^*_{(K)} \leftrightarrow (X,Y). \tag{9.98}$$

Specifically, we have the following readily verified result.

Corollary 9.33. With $\check{W}$ as constructed in Corollary 9.32, we have

$$\mathbb{E}[\check{W}|U,V] = \mathbf{\Sigma}^{1/2}_{(K)}\big(\mathbf{I} + \mathbf{\Sigma}_{(K)}\big)^{-1}(U+V) \tag{9.99}$$

$$\mathbf{\Lambda}_{\check{W}|U,V} = \big(\mathbf{I} - \mathbf{\Sigma}_{(K)}\big)\big(\mathbf{I} + \mathbf{\Sigma}_{(K)}\big)^{-1}. \tag{9.100}$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

9.13 An Interpretation of PCA

PCA [116], [140], [224] can be interpreted as a special case of the preceding results. Specifically, in some important instances, the form of dimensionality reduction realized by PCA corresponds to the optimum k-dimensional statistics $S_* = \mathbf{f}^*(Y)$ and $T_* = \mathbf{g}^*(X)$ as defined in (9.24a) and (9.24b), respectively, for the universal estimation of the unknown k-dimensional attributes U and V under any of our formulations.

Example 9.34. As an illustration, suppose we have the innovations form

$$Y = X + \nu_{X\to Y},$$

where X and Y are K-dimensional, and where $\mathbf{\Lambda}_\nu = \sigma_\nu^2\,\mathbf{I}$ but $\mathbf{\Lambda}_X$ is arbitrary. Moreover, let

$$\mathbf{\Lambda}_X = \mathbf{\Upsilon}\,\mathbf{\Lambda}\,\mathbf{\Upsilon}^{\mathrm{T}}$$

denote the diagonalization of $\mathbf{\Lambda}_X$, so the columns of

$$\mathbf{\Upsilon} = \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_K \end{bmatrix}, \tag{9.101}$$

are orthonormal, and $\mathbf{\Lambda}$ is diagonal with entries $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K$. Then it is immediate that $\mathbf{\Lambda}_Y$ has diagonalization

$$\mathbf{\Lambda}_Y = \mathbf{\Upsilon}\left(\mathbf{\Lambda} + \sigma_\nu^2\,\mathbf{I}\right)\mathbf{\Upsilon}^{\mathrm{T}}.$$

In this case, it follows immediately that $\tilde{\mathbf{B}}$ has SVD

$$\tilde{\mathbf{B}} = \mathbf{\Lambda}_Y^{-1/2}\mathbf{\Lambda}_X^{1/2} = \underbrace{\mathbf{\Upsilon}}_{=\mathbf{\Psi}^Y}\,\underbrace{\left(\mathbf{I} + \sigma_\nu^2\,\mathbf{\Lambda}^{-1}\right)^{-1/2}}_{=\mathbf{\Sigma}}\,\underbrace{\mathbf{\Upsilon}^{\mathrm{T}}}_{=\left(\mathbf{\Psi}^X\right)^{\mathrm{T}}}. \tag{9.102}$$

As a result, we have, for a given $1 \leq k \leq K$, that (9.24a) specializes to

$$\mathbf{F}^*_{(k)} = \left(\mathbf{\Upsilon}\,\mathbf{\Lambda}^{-1/2}\mathbf{\Upsilon}^{\mathrm{T}}\right)\mathbf{\Upsilon}_{(k)} = \mathbf{\Upsilon}_{(k)}\,\mathbf{\Lambda}_{(k)}^{-1/2}, \tag{9.103}$$

where $\mathbf{\Upsilon}_{(k)}$ denotes the $K \times k$ matrix consisting of the first k columns of $\mathbf{\Upsilon}$, i.e.,

$$\mathbf{\Upsilon}_{(k)} = \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_k \end{bmatrix},$$

and where $\mathbf{\Lambda}_{(k)}$ denotes the $k \times k$ upper left submatrix of $\mathbf{\Lambda}$, i.e., the matrix whose diagonal entries are $\lambda_1, \dots, \lambda_k$. Likewise, (9.24b) specializes to

$$\mathbf{G}^*_{(k)} = \left(\mathbf{\Upsilon} \left(\mathbf{\Lambda} + \sigma_\nu^2 \mathbf{I}\right)^{-1/2} \mathbf{\Upsilon}^{\mathrm{T}}\right) \mathbf{\Upsilon}_{(k)} = \mathbf{\Upsilon}_{(k)} \left(\mathbf{\Lambda}_{(k)} + \sigma_\nu^2 \mathbf{I}\right)^{-1/2}. \tag{9.104}$$

In turn, it follows from (9.103) and (9.104) that the k-dimensional PCA vector

$$S^{\mathrm{PCA}} = \mathbf{f}^{\mathrm{PCA}}(X) \triangleq \mathbf{\Upsilon}^{\mathrm{T}}_{(k)} X \tag{9.105a}$$

is a sufficient statistic for inferences about the unknown U and V based on X, and

$$T^{\mathrm{PCA}} = \mathbf{g}^{\mathrm{PCA}}(Y) \triangleq \mathbf{\Upsilon}^{\mathrm{T}}_{(k)} Y \tag{9.105b}$$

is a sufficient statistic for such inferences based on Y, i.e., we have the Markov structure

$$(U, V) \leftrightarrow S^{\mathrm{PCA}} \leftrightarrow X \tag{9.106a}$$

$$(U, V) \leftrightarrow T^{\mathrm{PCA}} \leftrightarrow Y. \tag{9.106b}$$

Beyond this illustrative example, for a general jointly Gaussian pair (X, Y), the statistics

$$S_{(k)} = \left(\mathbf{F}^*_{(k)}\right)^{\mathrm{T}} X \quad \text{and} \quad T_{(k)} = \left(\mathbf{G}^*_{(k)}\right)^{\mathrm{T}} Y$$

specialize to (invertible transformations of) the PCA statistics (9.105) whenever $K_X = K_Y = K$ and $\mathbf{\Lambda}_X$ and $\mathbf{\Lambda}_Y$ are simultaneously diagonalizable, i.e., when they share the same set of eigenvectors (9.101), which is equivalent to the condition that $\mathbf{\Lambda}_X$ and $\mathbf{\Lambda}_Y$ commute (see, e.g., [114, Theorem 1.3.12]). In fact, if $\mathbf{\Lambda}_X$ has distinct eigenvalues and commutes with $\mathbf{\Lambda}_Y$, then there is a polynomial $\pi(\cdot)$ of degree at most $K - 1$ such that $\mathbf{\Lambda}_Y = \pi(\mathbf{\Lambda}_X)$, which follows from the Cayley-Hamilton theorem (see, e.g., [114, Theorem 2.4.3.2 and Problem 1.3.P4]).

9.14 Learning Covariance Modal Decompositions

The linear features $\mathbf{f}(x)$ and $\mathbf{g}(y)$ in the modal decomposition of covariance $\mathbf{\Lambda}_{XY}$ are readily constructed via an iterative procedure. In particular, a natural approach corresponds to applying orthogonal iteration to $\tilde{\mathbf{B}}$ to generate the dominant modes of its SVD. The resulting procedure has a statistical interpretation as an ACE algorithm.

To obtain the Gaussian version of the ACE algorithm Algorithm 1, it suffices to note that the conditional expectations in this case are all linear—specifically,

$$\mathbb{E}[\mathbf{F}_{(k)}^{\mathrm{T}} X | Y = y] = \mathbf{F}_{(k)}^{\mathrm{T}} \mathbb{E}[X|Y = y] = \mathbf{F}_{(k)}^{\mathrm{T}} \boldsymbol{\Lambda}_{XY} \, \boldsymbol{\Lambda}_Y^{-1} \, y \tag{9.107a}$$

and

$$\mathbb{E}[\mathbf{G}_{(k)}^{\mathrm{T}} Y | X = x] = \mathbf{G}_{(k)}^{\mathrm{T}} \mathbb{E}[Y|X = x] = \mathbf{G}_{(k)}^{\mathrm{T}} \boldsymbol{\Lambda}_{XY}^{\mathrm{T}} \boldsymbol{\Lambda}_X^{-1} \, x, \tag{9.107b}$$

and that [cf. (9.17)]

$$\mathbb{E}\Big[\mathbf{F}_{(k)}^{\mathrm{T}} X \left(\mathbf{G}_{(k)}^{\mathrm{T}} Y\right)^{\mathrm{T}}\Big] = \mathbf{F}_{(k)}^{\mathrm{T}} \boldsymbol{\Lambda}_{XY} \, \mathbf{G}_{(k)}. \tag{9.107c}$$

The resulting procedure then takes the form of Algorithm 3. Computational complexity behavior is analogous to the corresponding algorithm for discrete data. As in our discussion of Section 6.1.2, steps 2f and 2c can be equivalently expressed in their respective variational forms [cf. (6.6)]

$$\bar{\mathbf{F}}_{(k)} \leftarrow \operatorname*{arg\,min}_{\mathbf{F}} \mathbb{E}\Big[\big\|\mathbf{F}_{(k)}^{\mathrm{T}} X - \hat{\mathbf{G}}_{(k)}^{\mathrm{T}} Y\big\|^2\Big] \tag{9.108a}$$

$$\bar{\mathbf{G}}_{(k)} \leftarrow \operatorname*{arg\,min}_{\mathbf{G}} \mathbb{E}\Big[\big\|\hat{\mathbf{F}}_{(k)}^{\mathrm{T}} X - \mathbf{G}_{(k)}^{\mathrm{T}} Y\big\|^2\Big], \tag{9.108b}$$

and evaluated iteratively or otherwise.

When the covariance structure $\boldsymbol{\Lambda}_X$, $\boldsymbol{\Lambda}_Y$, and $\boldsymbol{\Lambda}_{XY}$ is unknown, but we have training data

$$\mathcal{T} \triangleq \{(x_1, y_1), \ldots, (x_n, y_n)\}, \tag{9.109}$$

drawn i.i.d. from the associated Gaussian distribution, we can use sample covariance matrices in place of the true ones in Algorithm 3, viz.,

Algorithm 3 Gaussian ACE, Multiple Mode Computation

```
Require: Covariance matrices $\mathbf{\Lambda}_{XY}$, $\mathbf{\Lambda}_X$, and $\mathbf{\Lambda}_Y$; dimension $k$
    1. Initialization: randomly choose $\bar{\mathbf{F}}_{(k)}$
    repeat
        2a. Cholesky factor:
            $\bar{\mathbf{F}}_{(k)}^{\mathrm{T}}\mathbf{\Lambda}_X\bar{\mathbf{F}}_{(k)} = \big(\mathbf{\Theta}_{(k)}^X\big)^{\mathrm{T}}\mathbf{\Theta}_{(k)}^X$
        2b. Whiten:
            $\hat{\mathbf{F}}_{(k)} = \bar{\mathbf{F}}_{(k)}\,\big(\mathbf{\Theta}_{(k)}^X\big)^{-1}$
        2c. $\bar{\mathbf{G}}_{(k)} \leftarrow \mathbf{\Lambda}_Y^{-1}\,\mathbf{\Lambda}_{YX}\,\hat{\mathbf{F}}_{(k)}$
        2d. Cholesky factor:
            $\bar{\mathbf{G}}_{(k)}^{\mathrm{T}}\mathbf{\Lambda}_Y\bar{\mathbf{G}}_{(k)} = \big(\mathbf{\Theta}_{(k)}^Y\big)^{\mathrm{T}}\mathbf{\Theta}_{(k)}^Y$
        2e. Whiten:
            $\hat{\mathbf{G}}_{(k)} = \bar{\mathbf{G}}_{(k)}\,\big(\mathbf{\Theta}_{(k)}^Y\big)^{-1}$
        2f. $\bar{\mathbf{F}}_{(k)} \leftarrow \mathbf{\Lambda}_X^{-1}\,\mathbf{\Lambda}_{YX}^{\mathrm{T}}\hat{\mathbf{G}}_{(k)}$
        2g. $\hat{\sigma}^{(k)} \leftarrow \mathrm{tr}\big(\hat{\mathbf{G}}_{(k)}^{\mathrm{T}}\mathbf{\Lambda}_{YX}\,\bar{\mathbf{F}}_{(k)}\big)$
    until $\hat{\sigma}^{(k)}$ stops increasing.
```

$$
\begin{aligned}
\hat{\mathbf{\Lambda}}_X &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\boldsymbol{\mu}}_X)\,(x_i - \hat{\boldsymbol{\mu}}_X)^{\mathrm{T}}\\
\hat{\mathbf{\Lambda}}_Y &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\boldsymbol{\mu}}_Y)\,(y_i - \hat{\boldsymbol{\mu}}_Y)^{\mathrm{T}}\\
\hat{\mathbf{\Lambda}}_{XY} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\boldsymbol{\mu}}_X)\,(y_i - \hat{\boldsymbol{\mu}}_Y)^{\mathrm{T}},
\end{aligned}
$$

for example, where

$$
\hat{\boldsymbol{\mu}}_X = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \hat{\boldsymbol{\mu}}_Y = \frac{1}{n}\sum_{i=1}^{n} y_i.
$$

Sample complexity analysis of modal estimation in this Gaussian case can be carried out in a manner analogous to that described in Section 6.2.

9.15 Gaussian Attribute Matching

The preceding analysis can be used to develop the natural Gaussian counterpart of the Bayesian attribute matching formulation of collaborative filtering in Section 7. Our analysis can be interpreted as a formulation of a problem of high-dimensional linear estimation. As such, they can be viewed in a broader context that includes related results on shrinkage-based methods (see, e.g., [168] and the references therein) that generalize James-Stein estimators [136]. Recent work more directly analogous to the analyses of matrix factorization in, e.g., collaborative filtering as discussed in Section 7 include, e.g., [208], [230]. As such, this section provides an additional interpretation of such relationships.

For the purposes of illustration, consider a simple problem of low-level computer vision. Let Y denote a vector representing a (e.g., rasterized) K_Y-pixel target image of some scene of interest, and let X denote a vector representing a (linearly) distorted K_X-pixel source image of the scene. Such distortions could include, e.g., complex geometric transformations, nonuniform sampling, spatially-varying filtering, and noise. Then $P_{Y|X}(\cdot|x)$ denotes the probability density for the target image associated with a given source image x.

Given a choice for $k \in \{1, \dots, K\}$, the k-dimensional variables U and V in the Gaussian Markov chain (9.63) correspond to the dominant attributes of source and target images, respectively, and where $S^*_{(k)}$ and $T^*_{(k)}$ represent sufficient statistics for the estimation of these attributes.

Conceptually, for each target image y, there is an associated target attribute $V(y)$ generated randomly from y according to $P_{V|Y}(\cdot|y)$ that expresses the dominant attribute of the target image. Likewise, for the source image x, there is an associated target attribute $V_\circ(x)$ generated randomly from x according to $P_{V|X}(\cdot|x)$.

Next, let $\Delta_y(x)$ denote how close the target attribute of target image y is to the target attribute of the source image x, i.e.,

$$\Delta_y(x) \triangleq V(y) - V_\circ(x),$$

and define the set

$$\hat{\mathcal{Y}}(x) \triangleq \operatorname*{arg\,min}_{y \in \mathbb{R}^{K_Y}} \mathbb{E}\Big[\big\|\Delta_y(x)\big\|^2\Big]. \tag{9.110}$$

The following characterization of $\hat{\mathcal{Y}}(x)$—the collection of target images whose attributes match that of the source image most closely—is useful in our development. A proof is provided in Appendix G.22.

Lemma 9.35. Given $k \in \{1, \dots, K\}$ and zero-mean jointly Gaussian X, Y characterized by $\mathbf{\Lambda}_X$, $\mathbf{\Lambda}_Y$, and $\mathbf{\Lambda}_{XY}$, define k-dimensional Gaussian multi-attributes in the Gauss-Markov structure (9.63) according to Corollary 9.27 for some $\epsilon > 0$. Then for a given $x \in \mathbb{R}^{K_X}$ and $\hat{\mathcal{Y}}(x)$ as defined (9.110), it follows $y \in \hat{\mathcal{Y}}$ if and only if the associated (linear) features are related according to

$$g_i^*(y) = \sigma_i \, f_i^*(x), \qquad i = 1, \dots, k. \tag{9.111}$$

Among the target images y for which the attribute match with x is closest, we seek the most likely, which we denote using $y^*(x)$. We have the following characterization of $y^*(x)$. A proof is provided in Appendix G.23.

Proposition 9.36. Given $k \in \{1, \dots, K\}$ and zero-mean jointly Gaussian $X \in \mathbb{R}^{K_X}$, $Y \in \mathbb{R}^{K_Y}$ characterized by $\mathbf{\Lambda}_X$, $\mathbf{\Lambda}_Y$, and $\mathbf{\Lambda}_{XY}$, define k-dimensional Gaussian multi-attributes in the Gauss-Markov chain (9.63) according to Corollary 9.27 for some $\epsilon > 0$. Then for a given $x \in \mathbb{R}^{K_X}$ and $\hat{\mathcal{Y}}(x)$ as defined (9.110), we have that

$$y^*(x) \triangleq \operatorname*{arg\,max}_{y \in \hat{\mathcal{Y}}(x)} P_Y(y), \tag{9.112}$$

with $P_Y = \mathtt{N}(\mathbf{0}, \mathbf{\Lambda}_Y)$ denoting the marginal for Y, satisfies

$$y^*(x) = \left(\mathbf{G}^*_{(k)}\right)^{\dagger\mathrm{T}} \mathbf{\Sigma}_{(k)} \left(\mathbf{F}^*_{(k)}\right)^{\mathrm{T}} x, \tag{9.113a}$$

where the Moore-Penrose pseudoinverse of $\mathbf{G}_{(k)}$ takes the form

$$\left(\mathbf{G}^*_{(k)}\right)^{\dagger} = \left(\mathbf{G}^*_{(k)}\right)^{\mathrm{T}} \mathbf{\Lambda}_Y. \tag{9.113b}$$

The optimizing $y^*(x)$ in Proposition 9.36 has the interpretation as an MMSE estimate based not on $\mathbf{\Lambda}_{XY}$ but on the approximation $\mathbf{\Lambda}_{XY}^{(k)*}$ of rank k defined in (9.58). In particular, the following corollary is an immediate consequence of (9.58).

Corollary 9.37. An equivalent characterization of (9.113) in Proposition 9.36 is

$$y^*(x) = \mathbf{\Lambda}_{YX}^{(k)*} \, \mathbf{\Lambda}_X^{-1} \, x, \tag{9.114}$$

where $\mathbf{\Lambda}_{YX}^{(k)*}$ is as defined in (9.58).

Connections to PCA

As we now illustrate, PCA naturally arises in special cases of the preceding matching framework. In particular, returning to the scenario of the example in Section 9.13, we first interchange the roles of U and V, and X and Y, obtaining that the optimum target image $x^*(y)$ for a given source image y based on Gaussian attribute matching is

$$x^*(y) = \left(\mathbf{F}^*_{(k)}\right)^{\dagger\mathrm{T}} \mathbf{\Sigma}_{(k)} \left(\mathbf{G}^*_{(k)}\right)^{\mathrm{T}} y, \tag{9.115a}$$

where

$$\left(\mathbf{F}^*_{(k)}\right)^{\dagger} = \left(\mathbf{F}^*_{(k)}\right)^{\mathrm{T}} \mathbf{\Lambda}_X. \tag{9.115b}$$

Gaussian attribute matching in this scenario takes a familiar form. In particular, specializing (9.115) using (9.102), (9.103), and (9.104), we obtain

$$\begin{aligned} x^*(y) &= \mathbf{\Upsilon}_{(k)} \mathbf{\Lambda}_{(k)}^{1/2} \left(\mathbf{I} + \sigma_\nu^2 \mathbf{\Lambda}_{(k)}^{-1}\right)^{-1/2} \left(\mathbf{\Lambda}_{(k)} + \sigma_\nu^2 \mathbf{I})\right)^{-1/2} \mathbf{\Upsilon}_{(k)}^{\mathrm{T}} y \\ &= \mathbf{\Upsilon}_{(k)} \mathbf{\Lambda}_{(k)} \left(\mathbf{\Lambda}_{(k)} + \sigma_\nu^2 \mathbf{I})\right)^{-1} \mathbf{\Upsilon}_{(k)}^{\mathrm{T}} y. \end{aligned}$$

The result is, of course, a standard approach to simple (linear) denoising, whereby a signal of interest is expanded in the basis prescribed by PCA, only the dominant modes are retained, and the associated coefficients are appropriately attenuated. As such, our analysis provides an additional interpretation of such processing, further insights into which also arise in the next section.

9.16 Rank-Constrained Linear Regression

In this section, we develop the counterpart to softmax regression analysis of Section 8 for jointly Gaussian variables, which is a form of rank-constrained linear regression. Such regression problems have a long history. Indeed, Young [287] recognized the relationship between early factor analysis and low-rank approximation. Subsequent results on the topic appear in, e.g., [229], and, later, in [39, Theorem 10.2.1] [135]. Later still, interpretations of the special case of PCA in terms of neural networks appeared in, e.g., [21], [213], [214], and the general case of CCA in [146], in which an alternative to the ACE algorithm of Section 9.14 is involved in its implementation. The results of this section provide some complementary perspectives.

To begin, the counterpart of Proposition 8.1 is immediate in the Gaussian case. In particular, the following simple result expresses that the particular exponential family form is not restrictive.

Proposition 9.38. Let $X \in \mathbb{R}^{K_X}$, $Y \in \mathbb{R}^{K_Y}$ be zero-mean jointly Gaussian variables characterized by $\mathbf{\Lambda}_X$, $\mathbf{\Lambda}_Y$, and $\mathbf{\Lambda}_{XY}$. Furthermore, given a dimension $k \in \{1, \dots, K\}$, let $S = \mathbf{F}^{\mathrm{T}} X$ for some $K_X \times k$ matrix $\mathbf{F}$, so $\mathbf{\Lambda}_{YS} = \mathbf{\Lambda}_{YX}\, \mathbf{F}$ and $\mathbf{\Lambda}_S = \mathbf{F}^{\mathrm{T}} \mathbf{\Lambda}_X \mathbf{F}$ are the induced covariances. Then the joint probability density for S, Y takes the form

$$P_{S,Y}(s,y) = P_Y(y)\, P_{S|Y}(s|y) = \mathtt{N}(y; \mathbf{0}, \mathbf{\Lambda}_Y)\, \mathtt{N}(s; \mathbf{G}^{\mathrm{T}} y, \mathbf{\Lambda}_S - \mathbf{G}^{\mathrm{T}} \mathbf{\Lambda}_Y\, \mathbf{G}), \tag{9.116}$$

where

$$\mathbf{G}^{\mathrm{T}} Y \triangleq \mathbb{E}[S|Y]. \tag{9.117}$$

Proof. It suffices to exploit that since (9.117) is the MMSE estimate of S given Y, we have

$$S = \mathbf{G}^{\mathrm{T}} Y + \nu,$$

where the error ν is independent of Y. Moreover,

$$\mathbf{\Lambda}_{S|Y} = \mathbf{\Lambda}_S - \mathbf{\Lambda}_{SY} \mathbf{\Lambda}_Y^{-1} \mathbf{\Lambda}_{YS} = \mathbf{\Lambda}_S - \mathbf{G}^{\mathrm{T}} \mathbf{\Lambda}_Y\, \mathbf{G},$$

where to obtain the last equality we have used that

$$\mathbf{G}^{\mathrm{T}} = \mathbf{\Lambda}_{SY}\, \mathbf{\Lambda}_Y^{-1}.$$

since S, Y are jointly Gaussian. ■

In turn, the counterpart to Corollary 8.2 is the following result optimizing $\mathbf{F}$, whose proof is provided in Appendix G.24.

Proposition 9.39. Let $X \in \mathbb{R}^{K_X}$, $Y \in \mathbb{R}^{K_Y}$ be ϵ-correlated zero-mean jointly Gaussian variables whose joint density $P_{X,Y}$ is characterized by $\mathbf{\Lambda}_X$, $\mathbf{\Lambda}_Y$, and $\mathbf{\Lambda}_{XY}$. Furthermore, given a dimension $k \in \{1, \dots, K\}$, let

$$\tilde{\mathcal{P}}_k^{K_X,K_Y}(\mathbf{\Lambda}_X,\mathbf{\Lambda}_Y) \triangleq \left\{ P\colon P = \mathtt{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{\Lambda}_X & \tilde{\mathbf{\Lambda}}_{XY} \\ \tilde{\mathbf{\Lambda}}_{XY}^{\mathrm{T}} & \mathbf{\Lambda}_Y \end{bmatrix}\right), \right.$$

$$\left. \text{some } \tilde{\mathbf{\Lambda}}_{XY} \text{ with } \operatorname{rank}(\tilde{\mathbf{\Lambda}}_{XY}) \leq k \right\} \quad (9.118)$$

denote the collection of zero-mean jointly Gaussian distributions with rank-constrained cross-covariance. Then for $\tilde{P}_{X,Y} \in \tilde{\mathcal{P}}_k^{K_X,K_Y}(\mathbf{\Lambda}_X,\mathbf{\Lambda}_Y)$,

$$D(P_{X,Y}\|\tilde{P}_{X,Y}) \geq \sum_{i=k+1}^{K} \sigma_i^2 + o(\epsilon^2), \quad \epsilon \to 0,$$

where the inequality holds with equality when $\tilde{P}_{X,Y}$ has cross-covariance $\tilde{\mathbf{\Lambda}}_{YX} = \mathbf{\Lambda}_{YX}^{(k)*}$, with $\mathbf{\Lambda}_{YX}^{(k)*}$ as given by (9.58).

This result expresses that among all k-dimensional linear restrictions $S = \mathbf{F}^{\mathrm{T}} X$ of the data, that corresponding to $\mathbf{F} = \mathbf{F}_{(k)}^*$ is optimum.

In turn, given the choice $S = \left(\mathbf{F}_{(k)}^*\right)^{\mathrm{T}} X$, the matrix $\mathbf{G}_*^{(k)}$ defines the weights in the associated estimate of Y; specifically, [cf. (9.113)]

$$\hat{Y}^* = (\mathbf{\Lambda}_{XY}^{(k)*})^{\mathrm{T}} \mathbf{\Lambda}_X^{-1} X = \mathbf{\Lambda}_Y\, \mathbf{G}_{(k)}^*\, \mathbf{\Sigma}_{(k)}\, S. \quad (9.119)$$

We emphasize that the estimate (9.119), in which $\mathbf{\Lambda}_{XY}^{(k)*}$ can be equivalently expressed in the form

$$\mathbf{\Lambda}_{XY}^{(k)*} = (\mathbf{\Lambda}_X^{1/2}\, \mathbf{\Psi}_{(k)}^X)(\mathbf{\Lambda}_X^{1/2}\, \mathbf{\Psi}_{(k)}^X)^{\dagger}\, \mathbf{\Lambda}_{XY}, \quad (9.120)$$

is generally different from the MMSE estimator limited to rank k, which can be expressed in the following form [135], a derivation of which is provided in Appendix G.25.

Proposition 9.40. For zero-mean, jointly Gaussian $X \in \mathbb{R}^{K_X}$ and $Y \in \mathbb{R}^{K_Y}$ characterized by covariance $\mathbf{\Lambda}_X$, $\mathbf{\Lambda}_Y$, and $\mathbf{\Lambda}_{XY}$, then given $k \in \{1, \dots, K\}$,

$$\hat{Y}^{\circ} = \operatorname*{arg\,min}_{\substack{\{\hat{Y}\colon \hat{Y}=\tilde{\mathbf{\Gamma}}_{Y|X}X,\\ \operatorname{rank}(\tilde{\mathbf{\Gamma}}_{Y|X})\leq k\}}} \mathbb{E}\left[\|Y-\hat{Y}\|^2\right] = (\mathbf{\Lambda}_{XY}^{(k)\circ})^{\mathrm{T}} \mathbf{\Lambda}_X^{-1} X, \quad (9.121a)$$

where

$$\mathbf{\Lambda}_{XY}^{(k)\circ} = (\mathbf{\Lambda}_X^{1/2}\, \tilde{\mathbf{\Psi}}_{(k)}^X)(\mathbf{\Lambda}_X^{1/2}\, \tilde{\mathbf{\Psi}}_{(k)}^X)^{\dagger}\, \mathbf{\Lambda}_{XY}, \quad (9.121b)$$

with $\tilde{\mathbf{\Psi}}^X_{(k)}$ denoting the first (dominant) k columns of $\tilde{\mathbf{\Psi}}^X$ in the (alternative) SVD [cf. (9.12)]

$$\mathbf{\Lambda}_{YX}\,\mathbf{\Lambda}_X^{-1/2} = \mathbf{\Lambda}_Y^{1/2}\,\tilde{\mathbf{B}} = \tilde{\mathbf{\Psi}}^Y\,\tilde{\mathbf{\Sigma}}\,\big(\tilde{\mathbf{\Psi}}^X\big)^{\mathrm{T}} \tag{9.121c}$$

in which $\tilde{\mathbf{\Psi}}^X$ and $\tilde{\mathbf{\Psi}}^Y$ are orthogonal matrices and $\tilde{\mathbf{\Sigma}}$ is a diagonal matrix.

We note, in particular, that the generally different estimators (9.121) and (9.119) coincide when $\mathbf{\Lambda}_Y = \mathbf{I}$, since the SVDs (9.12) and (9.121c) are identical in this case.

Finally, the implied rank-constrained linear regression procedure is as follows. First, we assume that sufficient unlabeled training data is available that $\mathbf{\Lambda}_X$ and $\mathbf{\Lambda}_Y$ are accurately recovered. Second, from the labeled training data we obtain the empirical covariance $\hat{\mathbf{\Lambda}}_{XY}$. We then let $\hat{P}_{X,Y}$ denote the distribution of zero-mean jointly Gaussian variables characterized by $\mathbf{\Lambda}_X$, $\mathbf{\Lambda}_Y$, and $\hat{\mathbf{\Lambda}}_{XY}$ and apply Proposition 9.39 with $P_{X,Y} = \hat{P}_{X,Y}$ to obtain that the (locally) divergence-minimizing (cross-entropy maximizing) regression parameters are given by

$$\hat{\mathbf{\Lambda}}^{(k)*}_{YX} \triangleq \mathbf{\Lambda}_Y\,\hat{\mathbf{G}}^*_{(k)}\,\hat{\mathbf{\Sigma}}_{(k)}\,\big(\hat{\mathbf{F}}^*_{(k)}\big)^{\mathrm{T}}\mathbf{\Lambda}_X, \tag{9.122}$$

where $\hat{\mathbf{F}}_{(k)}$, $\hat{\mathbf{G}}_{(k)}$, and $\hat{\mathbf{\Sigma}}_{(k)}$ correspond to the k dominant modes in the modal decomposition of the empirical cross-covariance, viz., [cf. (9.18)]

$$\hat{\mathbf{\Lambda}}_{YX} = \mathbf{\Lambda}_Y\,\hat{\mathbf{G}}^*\,\hat{\mathbf{\Sigma}}\,\big(\hat{\mathbf{F}}^*\big)^{\mathrm{T}}\mathbf{\Lambda}_X. \tag{9.123}$$

In turn, the quality of the model fit is given by

$$D(\hat{P}_{X,Y}\,\|\,\hat{P}^{(k)*}_{X,Y}) = \sum_{i=k+1}^{K}\hat{\sigma}_i^2 + o(\epsilon^2), \quad \epsilon \to 0, \tag{9.124}$$

where $\hat{P}^{(k)*}_{X,Y}$ denotes the (optimized) distribution of zero-mean jointly Gaussian variables characterized by $\mathbf{\Lambda}_X$, $\mathbf{\Lambda}_Y$, and $\hat{\mathbf{\Lambda}}^{(k)*}_{XY}$, and $\hat{\sigma}_1, \dots, \hat{\sigma}_K$ are the diagonal entries of $\hat{\mathbf{\Sigma}}$.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

10

Nonlinear Features and nonGaussian Distributions

While a comprehensive treatment is beyond the scope of this monograph, in this section we briefly discuss selected aspects of modal decompositions involving nonlinear features and nonGaussian variables, for completeness. More generally, this is an active area of research, and additional emerging directions are summarized in Section 13.

10.1 Nonlinear Features for Gaussian Distributions

If we seek a modal decomposition of the form (2.15) for the Gaussian case, an infinite number of terms must be involved: the modal decomposition takes the form

$$P_{X,Y}(x,y) = P_X(x)\,P_Y(y)\left(1+\sum_{i=1}^{\infty}\tilde{\sigma}_i\,\tilde{f}_i^*(x)\,\tilde{g}_i^*(y)\right), \tag{10.1a}$$

with

$$\begin{aligned} \mathbb{E}\big[\tilde{f}_i^*(\tilde{X})\big] = \mathbb{E}\big[\tilde{g}_i^*(\tilde{Y})\big] &= 0,\ \ i=1,2,\dots \\ \mathbb{E}\big[\tilde{f}_i^*(\tilde{X})\,\tilde{f}_j^*(\tilde{X})\big] = \mathbb{E}\big[\tilde{g}_i^*(\tilde{Y})\,\tilde{g}_j^*(\tilde{Y})\big] &= \mathbb{1}_{i=j},\ \ i,j=1,2,\dots. \end{aligned} \tag{10.1b}$$

And in such an expansion, it is important to emphasize that the terms involving linear features need not dominate. In the sequel, we briefly develop this insight.

In a modal decomposition of the form (10.1), only some of the features $\tilde{f}_1^*, \tilde{g}_1^*, \tilde{f}_2^*, \tilde{g}_2^*, \dots$ are linear, which is implied by expanding the exponentiation operator in (9.56) using a Taylor Series. For instance, when $K_X = K_Y = 1$, one obtains Mehler's decomposition [200]

$$\begin{aligned} P_{\tilde{X},\tilde{Y}}(\tilde{x},\tilde{y}) &= P_{\tilde{X}}(\tilde{x})\, P_{\tilde{Y}}(\tilde{y}) \left[1 + \sum_{i=1}^{\infty} \rho^{i}\, \pi_i(\tilde{x})\, \pi_i(\tilde{y}) \right] \\ &= P_{\tilde{X}}(\tilde{x})\, P_{\tilde{Y}}(\tilde{y}) \left[1 + \sum_{i=1}^{\infty} |\rho|^{i} \operatorname{sgn}(\rho)^{i}\, \pi_i(\tilde{x})\, \pi_i(\tilde{y}) \right], \qquad (10.2a) \end{aligned}$$

where $\rho = \lambda_{\tilde{X}\tilde{Y}} = \mathbb{E}\left[\tilde{X}\tilde{Y}\right]$, and where π_i is the (scaled) ith-order Hermite polynomial

$$\pi_i(\tilde{x}) \triangleq \frac{1}{\sqrt{i!}} (-1)^{i} \mathrm{e}^{\tilde{x}^2/2} \frac{\mathrm{d}^{i}}{\mathrm{d}\tilde{x}^{i}} \mathrm{e}^{-\tilde{x}^2/2}, \quad i = 1, 2, \dots. \qquad (10.2b)$$

Note that $\tilde{f}_i^* = \operatorname{sgn}(\rho)^{i}\, \tilde{g}_i^* = \pi_i$ satisfy (10.1b) as required, and that the features corresponding to the dominant mode ($i = 1$) are linear,[1] i.e., $\pi_1(\upsilon) = \upsilon$, as discussed in, e.g., [159], [240], and which may have been first observed by Kolomogorov.

For $K > 1$ [with (9.13)], the modal decomposition of the form (10.1) is straightforward to derive as a generalization of (10.2), and involves the corresponding multivariate Hermite polynomials—see, e.g., [148], [252]. However, if $K > 1$ and $k > 1$, then the k dominant modes need not, in general, be linear, as the following example shows.

Example 10.1. Suppose $K_X = K_Y = k = 2$, $(\tilde{X}_1, \tilde{Y}_1)$ and $(\tilde{X}_2, \tilde{Y}_2)$ are independent, and $\rho_l = \lambda_{\tilde{X}_l\tilde{Y}_l} = \mathbb{E}[\tilde{X}_l\tilde{Y}_l] > 0$ for $l = 1, 2$. Then we have the decomposition

$$\begin{aligned} P_{\tilde{X},\tilde{Y}}(\tilde{x},\tilde{y}) = P_{\tilde{X}_1}(\tilde{x}_1)\, P_{\tilde{X}_2}(\tilde{x}_2)\, P_{\tilde{Y}_1}(\tilde{y}_1)\, P_{\tilde{Y}_2}(\tilde{y}_2) \\ \cdot \left[1 + \sum_{i=1}^{\infty} \rho_1^{i}\, \pi_i(\tilde{x}_1)\, \pi_i(\tilde{y}_1) \right] \left[1 + \sum_{i=1}^{\infty} \rho_2^{i}\, \pi_i(\tilde{x}_2)\, \pi_i(\tilde{y}_2) \right], \end{aligned}$$

[1] Evidently, the maximal correlation is the magnitude of Pearson's usual correlation coefficient in this Gaussian case, i.e., $\sigma_1 = \sigma(f_1^*, g_1^*) = |\rho|$.

which when expanded is of the form (10.1). But if $\rho_1^2 > \rho_2$, then the dominant feature pair is

$$\begin{aligned}\tilde{f}_1^*(\tilde{x}) &= \pi_1(\tilde{x}_1) = \tilde{x}_1\\ \tilde{g}_1^*(\tilde{y}) &= \pi_1(\tilde{y}_1) = \tilde{y}_1,\end{aligned}$$

corresponding to singular value $\tilde{\sigma}_1 = \rho_1$, while the features for the next largest singular value are

$$\begin{aligned}\tilde{f}_2^*(\tilde{x}) &= \pi_2(\tilde{x}_1) = (\tilde{x}_1^2 - 1)/\sqrt{2}\\ \tilde{g}_2^*(\tilde{y}) &= \pi_2(\tilde{y}_1) = (\tilde{y}_1^2 - 1)/\sqrt{2},\end{aligned}$$

corresponding to singular value ρ_1^2, rather than the linear features corresponding to singular value ρ_2.

10.2 Linear Features for nonGaussian Distributions

Second, with respect to nonGaussian distributions, we first emphasize that in Proposition 9.4 (and, in turn, Proposition 9.6), only the second-moment properties of the joint distribution $P_{X,Y}$ are required to derive the optimizing linear features. As such, those results obviously apply more broadly.

Additionally, when the nonGaussian variables are defined on finite (but real-valued) alphabets, we can equivalently interpret the CCA optimization problem as that of HGR maximal correlation with the features constrained to be linear, i.e., maximizing the vector correlation (3.6b) over linear f^k and g^k. Such constraints may be practically motivated, for example. In such cases, we can relate the CCM $\tilde{\mathbf{B}}$ from the Gaussian analysis to the associated DTM $\mathbf{B}$ from the discrete analysis. In particular, we have the following theorem, whose proof is provided in Appendix G.11.

Proposition 10.2. Let $\mathcal{X} \subset \mathbb{R}^{K_X}$ and $\mathcal{Y} \subset \mathbb{R}^{K_Y}$ be finite sets with probability mass function $P_{X,Y}$ such that $\mathbb{E}[X] = 0$ and $\mathbb{E}[Y] = 0$, and let $\mathbf{B}$ be as defined in (2.8). Then

$$\boldsymbol{\Pi}^Y \, \mathbf{B} \, (\boldsymbol{\Pi}^X)^{\mathrm{T}} = \mathbf{B}_{\mathrm{G}}, \tag{10.3}$$

where $\mathbf{B}_{\mathrm{G}}$ is as defined in (9.11) (with the notation refined to distinguish it from $\mathbf{B}$), and where

$$\mathbf{\Pi}^{X} = \mathbf{X}\sqrt{\mathbf{P}_X} \quad \text{and} \quad \mathbf{\Pi}^{Y} = \mathbf{Y}\sqrt{\mathbf{P}_Y}, \tag{10.4}$$

with $\mathbf{X}$ and $\mathbf{Y}$ denoting $K_X \times |\mathcal{X}|$ and $K_Y \times |\mathcal{Y}|$ matrices whose columns are the vectors in $\mathcal{X}$ and $\mathcal{Y}$, respectively.

10.3 Features for General Continuous Variables

The extension of modal decompositions to both more general bivariate distributions over continuous alphabets—and the extraction of corresponding features—is naturally of interest in many applications. In this section, we provide a summary of some of the wide range of contributions to this area, together with representative references on the topic.

As our discussion in Section 10.1 would suggest, for bivariate distributions over continuous alphabets, many basic geometric insights carry over from the finite-alphabet case. However, the underlying Hilbert space is generally infinite-dimensional—as in (10.1)—from which subtleties arise. A natural restriction is to absolutely continuous distributions satisfying the Hilbert-Schmidt condition

$$\iint \frac{P_{X,Y}(x,y)^2}{P_X(x)\,P_Y(y)}\,\mathrm{d}x\,\mathrm{d}y < \infty, \tag{10.5}$$

which ensures compactness of the conditional expectation operators.

Among the earliest work for such scenarios is contained in the foundational contributions of Gebelein [91]. In subsequent work [231], [232], Rényi develops key aspects of the associated Hilbert space for modal decompositions, building on the work of both Hirschfeld [112] and Gebelein [91]. In turn, Csáki and Fischer [66]–[68] build on this work of Rényi, developing further aspects of the Hilbert space geometry and including some insightful examples. Aspects of the geometry also appear in Witsenhausen's analysis of common information [277], which likewise leverages the developments of Rényi [231].

In related but separate developments, modal decompositions for continuous distributions are also explored by Lancaster [160], [161],

The version of record is available at: http://dx.doi.org/10.1561/0100000107

building on the work of Hirschfeld [112].[2] Included are a variety of results on the special case of jointly Gaussian distributions. See also the related work [108]. Analysis similar to some of that of Lancaster also appears in the work of Sarmanov [238], [239].

While infinite-dimensional modal decompositions are conceptually straightforward in many respects, their practical computation is less so. In particular, in the development of the ACE algorithm in [38], Breiman and Friedman follow the framework of Rényi [232] and consider bivariate distributions over arbitrary alphabets, including continuous ones. Mild sufficient conditions for convergence of the ACE procedure are developed in this setting. However, even when these conditions are satisfied the procedure generally does not have a direct numerical implementation when applied to a continuous distribution. When using the ACE algorithm to estimate the modal decomposition from training data in such cases, [38] advocates replacing each conditional expectation step with an approximation that exploits some intuitive but heuristic data smoothing. However, in practice, the resulting features tend to depend somewhat strongly on the choice of smoothing.

The work of Buja [42], [43] on modal decompositions emphasizes continuous bivariate distributions whose features are (orthogonal) polynomials of increasing order, which includes the case of jointly Gaussian variables. The treatment leverages the work of Lancaster [162] on such distributions. As a result of the associated parametric structure, implementations of the ACE algorithm are more straightforward in this case. In the course of this development, Buja also provides a variety of other examples and observations involving modal decompositions and the ACE algorithm.

It is worth emphasizing that distributions whose modal decompositions have polynomial features have been useful in a variety of applications. For example, in [1], they are used to show that the capacity region of a degraded fading broadcast channel with Gaussian noise is not achieved by Gaussian input distribution.

[2] The treatment emphasizes distributions of bounded χ^2 mutual information—i.e., mean-square contingency (4.20); it is straightforward to verify that this condition is equivalent to (10.5).

Further results on distributions with polynomial features are developed in, e.g., [194], [195]. In particular, a characterization of such distributions in terms of their conditional moments is derived, and distributions with Laguerre, Jacobi, and Hermite polynomial features are constructed as illustrative examples.

10.4 Feature Constraints: Nonlinear CCA and PCA

In a variety of settings, it is often desirable to restrict the space of features under consideration (under the HGR maximal correlation criterion, for example), for implementational or application-specific reasons. In some cases, such constraints are straightforward to incorporate.

As a first example, when features are restricted to lie in a finite-dimensional subspace, any such feature can be expressed as finite linear combination of basis functions for this subspace. As a result, the HGR maximal correlation features over this subspace have a representation in this basis whose coefficients are the solution to a corresponding CCA problem. This is perhaps the simplest example of what is sometimes referred to as *nonlinear CCA*,[3] since the subspace will generally define a class of nonlinear features. In essence, the nonlinear feature optimization problem effectively becomes a linear feature optimization problem in this case, which can be applied to training data as a learning procedure in a straightforward manner. Accordingly, the optimization is over representations whose dimension corresponds to the size of the subspace.

In other scenarios, it is natural to restrict the features to an infinite-dimensional subspace, to which the preceding approach cannot practically be applied to solve the HGR maximal correlation problem. An alternative in this case is to use a nonparametric nonlinear CCA method typically referred to as *kernel CCA*, whereby the features are constrained to lie in a reproducing kernel Hilbert space [6], [17], [133], [157], [202]. In this case, for a given set of training data one can avoid having to perform computation directly in the underlying infinite-dimensional subspace, and instead exploit the kernel representation for the space,

[3] By extension to infinite-dimensional features spaces, sometimes HGR maximal correlation analysis is more generally referred to as nonlinear CCA in the literature.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

resulting in an optimization over representations whose dimension corresponds to the size of the training set. A variety of approaches have been developed to reduce the complexity of kernel CCA or extend its range of applicability; see, e.g., [15], [17], [20], [105], [110], [266], [288]. Beyond kernel CCA, other nonparametric approaches build more directly on modal decompositions of the form (10.1); see, e.g., [203].

Another alternative to restricting the features to a finite-dimensional subspace (corresponding to linearly parameterized classes of features) is to use nonlinearly parameterized feature classes. This approach yields yet another form of nonlinear CCA, early examples of which include [44] and [157], with the latter exploiting simple neural networks. Generalizations based on deep neural network (DNN) architectures are developed in [14], with multivariate extensions explored in [29]. For a comparative evaluation of such methods in representative application domains, see, e.g., [274].

Other methods for indirectly constraining the class of features have also been investigated. For example, in [218], the candidate features are (generally) randomized functions such that the mutual information between the data and its feature representation is constrained. In essence, this imposes that the representation be a sufficiently compressed version of the original data. Analysis of the methodology reveals connections to rate-distortion theory [63, Chapter 10], the information bottleneck method [258], and remote source coding [80], [278]. As such, the analysis therein is perhaps closest in spirit to that of this monograph.

Finally, a variety of nonlinear generalizations have also been developed for the special case of PCA. For example, kernel-based methods for PCA that are the counterparts to kernel CCA are developed in [245], [246] and referred to as *kernel PCA*. Others, such as [154], [182], are based on the use of neural network architectures and correspond to nonlinearly parameterized feature classes. As such, they are counterparts to the analogously constructed CCA methods.

11

Semi-Supervised Learning

A variety of problems deviates from the standard supervised learning model on which we have focused in previous sections. In these problems, labeled data are used in more limited ways, and instead relying more on unlabeled data in their training. These are typically referred to as *semi-supervised* learning problems, and there is a rich taxonomy and literature; see, e.g., [54] and the many references therein, including the early work [247]. While a broader development on the topic is beyond the scope of the present monograph, in this section we briefly discuss some of the most immediate implications of universal features and their analysis to some such problems.

An outline of the section is as follows. Section 11.1 describes the problem of "indirect" learning in which to carry out clustering on data, relationships to secondary data are exploited to define an appropriate measure of distance. We show, in particular, that the softmax analysis of Section 8 implies a natural procedure in which Gaussian mixture modeling is applied to the dominant features obtained from the modal decomposition with respect to the secondary data. In a different direction, Section 11.2 discusses the problem of partially-supervised learning in which features are learned in an unsupervised manner, and

The version of record is available at: http://dx.doi.org/10.1561/0100000107

labeled data is used only to obtain the classifier based on the resulting features. As an illustration of the use of universal features in this setting, an application to handwritten digit recognition using the MNIST database is described in which the relevant features are obtained via the common information between subblocks of MNIST images. A simple implementation achieves an error probability of 3.02%, close to that of a 3-layer neural net (with 300+100 hidden nodes), which yields an error probability of 3.05%.

11.1 Indirect Learning

A problem of significant interest is that of unsupervised learning, in which only unlabeled data is available to train the system. These correspond to clustering problems, and there are a number of classical approaches, originating with the work of Pearson [222]; see, e.g., [82] and the references therein for a summary.

In practice, there can be many valid clusterings of data, some more useful than others for a given target application. For instance, in the case of movies, one could cluster by any number of attributes, including time period, genre, etc. One can view these alternatives as capturing different measures of proximity in carrying out the clustering. But if one is interested in clustering movies according to the way people select movies to watch, then the measure of proximity is less straightforward to quantify.

In such cases, auxiliary labeled data can be used to effectively capture the right notion of distance for such problems, and express them in terms of universal features. To develop this notion of “indirect” learning, which has similarities in spirit to methods such as those described in [27], let

$$X \leftrightarrow Y \leftrightarrow Z$$

denote a Markov chain of discrete variables in which $Y \in \mathcal{Y}$ represents the data we seek to cluster (e.g., movies), $Z \in \mathcal{Z}$ represents the class index, and $X \in \mathcal{X}$ represents auxiliary data (e.g., people). We assume that in general $\mathcal{X}$ and $\mathcal{Y}$ are large alphabets, but that $\mathcal{Z}$ may be comparatively small, and that we have an empirical distribution $\hat{P}_{X,Y}$ obtained from

The version of record is available at: http://dx.doi.org/10.1561/0100000107

i.i.d. training data from $P_{X,Y}$ (e.g., the Netflix database), but no training samples of Z from which to directly estimate $P_{Z|Y}$, or even P_Z.

For this scenario, our universal analysis suggests the following procedure. First, for some suitably small $k \in \{1, \ldots, K-1\}$, we extract the k dominant modes in the decomposition (2.15) from $\hat{P}_{X,Y}$ (via, e.g., the ACE algorithm), then use the resulting estimate of g to define a new variable $T = g(Y) \in \mathbb{R}^k$. In turn, our softmax analysis reveals that a locally universal model for the latent variable Z is [cf. (8.4)]

$$\tilde{P}^*_{Z|T}(z|t) \propto P_Z(z) \exp\Big\{(t-\boldsymbol{\mu}_T)^{\mathrm{T}} \boldsymbol{\Lambda}_T^{-1} (\boldsymbol{\mu}_{T|Z}(z)-\boldsymbol{\mu}_T)\Big\}, \tag{11.1}$$

and unsupervised learning corresponds to fitting this model to the (induced) samples of T.

The softmax analysis further implies a rather natural model fitting procedure. In particular, as discussed in Section 8.2, the resulting distribution P_Y, matches, to first order, that of a Gaussian mixture, where $P_{T|Z}(\cdot|z)$ for $z \in \mathcal{Z}$ are the Gaussian components. Hence, this suggests that carrying out Gaussian mixture modeling on the estimate of P_T obtained from the training data—e.g., via the Expectation-Maximization (EM) algorithm [77]—to learn the parameters $\boldsymbol{\mu}_{T|Z}(t)$, $\boldsymbol{\Lambda}_{T|Z}$, and P_Z, and (soft) clustering according to the resulting $P_{Z|T}(\cdot|t)$ is locally optimal. Of course, such soft clustering can be replaced by any of a number of hard-decision alternatives if desired, such as that based on the Lloyd algorithm [92], which correspond to so-called k-means[1] clustering on the induced samples of $T = g(Y)$.

In practice, this procedure is straightforward to apply and effective. For example, applying it to, e.g., the Netflix database yields meaningful movie clusterings. For related developments and additional insights, see, e.g., [227].

11.2 Partially-Supervised Learning

Another class of learning system architectures is one in which labeled data is used to design a classifier of interest, but the design of the features themselves for such a classifier is based on unlabeled data.

[1] Note that k refers a different quantity (specifically, $|\mathcal{Z}|$) in this nomenclature than it does in our use.

These can be viewed as partially-supervised learning systems, and can provide performance close to that of fully supervised architectures while requiring significantly less labeled training data. In such cases, the feature extraction step corresponds to unsupervised dimensionality reduction, for which there are a variety of well established methods, both linear and nonlinear; see, e.g., [243].

The characterization of common information in terms of universal features, as described in Section 5.7, suggests a natural framework for nonlinear dimensionality reduction, and, in turn, constructing such partially-supervised learning systems, which we illustrate through an example involving handwritten digit recognition, using the MNIST database [167].

The MNIST database consists of a set of $n = 60\,000$ training images $x^{(1)}, \dots, x^{(n)}$ and a set of $n' = 10\,000$ test images, each depicting a single handwritten digit from the set $\mathcal{Z} \triangleq \{\mathtt{0}, \dots, \mathtt{9}\}$. We let $z^{(i)} \in \mathcal{Z}$ denote the label corresponding to training image $x^{(i)}$, and let $z \in \mathcal{Z}$ denote that for a given test image x, which are all provided in the database. Each training and test image is a black-and-white, 28×28 pixels size, and quantized to 8-bits per pixel (corresponding to intensity levels $\{0, \dots, 255\}$), so $|\mathcal{X}| = 28 \cdot 28 \cdot 256 = 200\,704$ is the image alphabet size.

Using the labeled data $\big(x^{(1)}, z^{(1)}\big), \dots, \big(x^{(n)}, z^{(n)}\big)$, we seek to train a classifier based on our framework to predict the label z of a test image x as accurately as possible.

11.2.1 Classification Architecture

The architecture we develop for this application involves three stages in a manner corresponding to a two-layer neural network. The first stage is a preprocessing step that converts the test image x to a representation $y = q(x)$ from a smaller alphabet $\mathcal{Y}$. In the second stage, we extract a low-dimensional real-valued feature $r = h(y)$ from the image representation y. Finally, in the third stage we classify the image based on this low-dimensional feature using a predictor $\varphi(\cdot)$, generating label $\hat{z} = \varphi(h(q(x)))$.

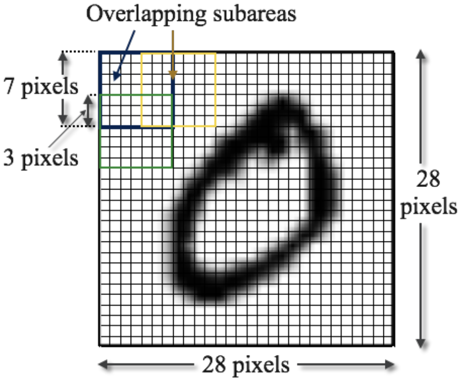

Figure 11.1: Image representation for the preprocessing stage of the semi-supervised handwritten digit classifier. Each 28×28 MNIST database image is decomposed into an array of $6 \times 6 = 36$ subimages, each of size 7×7 pixels, and each overlapping with its immediate neighbors by 3 pixels, horizontally and/or vertically.

We restrict our attention to designs based on semi-supervised learning. Specifically, q and h are designed from the unlabeled data $x^{(1)}, \dots, x^{(n)}$ in an unsupervised manner, while φ is designed in a supervised manner from the reduced labeled data

$$\big(r^{(1)}, z^{(1)}\big), \dots, \big(r^{(n)}, z^{(n)}\big), \tag{11.2a}$$

with

$$r^{(d)} = h(y^{(d)}), \qquad d = 1, \dots, n. \tag{11.2b}$$

The details of our classifier design are as follows.

Stage 1 (Preprocessing)

As depicted in Figure 11.1, we first decompose each MNIST database image into an array of $6 \times 6 = 36$ overlapping subimages, each of size 7×7 pixels, with immediately neighboring subimages overlapping by 3 pixels, horizontally and/or vertically. We denote the (i, j)th subimage by $\tilde{y}_{i,j}$, for $i, j = 1, \dots, 6$, which takes value in an alphabet of size $|\tilde{Y}| = 7 \cdot 7 \cdot 256 = 12\,544$.

Second, quantize each subimage in a lossy manner to reduce the size of the alphabet $\tilde{Y}$. For this purpose, for each (i, j), we cluster all the subimages $\tilde{y}_{i,j}^{(1)}, \dots, \tilde{y}_{i,j}^{(n)}$ extracted from the training data using the "balanced iterative reducing and clustering using hierarchies" (BIRCH) algorithm [289], which is simple and has computationally efficient (linear complexity). In particular, we use the implementation [156] with threshold parameter to $256\sqrt{3}$ and branching factor 1000. Each subimage $\tilde{y}_{i,j}$ is then represented by the cluster to which it maps, which we denote using $y_{i,j}$.[2] We further use y to denote the resulting composite image presentation, i.e.,

$$y = \begin{bmatrix} y_{1,1} & \cdots & y_{1,6} \\ \vdots & \ddots & \vdots \\ y_{6,1} & \cdots & y_{6,6} \end{bmatrix}.$$

Stage 2: Feature Extraction

We generate a k-dimensional feature from the unlabeled training data that captures as much of the common information among the subimages as possible; in our experiment we choose $k = 500$. In particular, for each of the $m = \binom{36}{2} = 630$ pairs $(y_{i,j}, y_{i',j'})$ of preprocessed subimages, we determine the k' dominant modes of the empirical pairwise distribution $\hat{P}_{Y_{i,j},Y_{i',j'}}$ generated from the reduced unlabeled training data

$$y^{(1)}, \dots, y^{(n)}.$$

In our experiment we choose $k' = 16$, and use Algorithm 1 to obtain these modes. We then order this aggregate list of $k'm = 10\,080$ modes by singular value, and construct the feature set from the subset corresponding to the overall k largest singular values, which we denote using $\hat{\sigma}_1, \dots, \hat{\sigma}_k$.

Specifically, with $\{(i_l, j_l), (i'_l, j'_l)\}$ denoting the indices of the subimage pair whose m_lth mode has singular value $\hat{\sigma}_l$, and with $\hat{f}^*_{(i_l,j_l),(i'_l,j'_l),m_l}$ and $\hat{g}^*_{(i_l,j_l),(i'_l,j'_l),m_l}$ denoting the corresponding feature functions in the decomposition of $\hat{P}_{Y_{i_l,j_l},Y_{i'_l,j'_l}}$, for a test image with representation y, we choose as our k-dimensional feature

$$r = h(y) = (h_1(y), \dots, h_k(y)) \tag{11.3a}$$

[2] The resulting alphabets $\mathcal{Y}_{i,j}$ differ in size, ranging from roughly 10 for subimages at the perimeter of the image, to roughly 500 for subimages in the middle.

with

$$h_l(y) = \hat{f}^*_{(i_l,j_l),(i'_l,j'_l),m_l}(y_{i_l,j_l}) + \hat{g}^*_{(i_l,j_l),(i'_l,j'_l),m_l}(y_{i'_l,j'_l}). \tag{11.3b}$$

We emphasize that, in accordance with our development in Section 5.7 that leads to (5.85), the elements of (11.3) are sufficient statistics for the relevant components of the common information between the associated subimage pairs.

Stage 3 (Feature Classification)

The final stage implements a low-dimensional feature classifier generated from the reduced labeled training data (11.2), with h as defined in (11.3). In particular, we choose a linear support vector machine (SVM) [62] for this purpose.

11.2.2 Performance Evaluation

When we evaluate the performance of the classifier of Section 11.2.1 on the full the set of 10 000 MNIST test images, we achieve an digit recognition error probability of 3.02%.

The classifier, which is characterized by its $k = 500$ scalar features, is naturally compared to alternatives with similar numbers of features. For example, one such alternative classifier would omit the preprocessing and feature extraction stages, and apply a linear SVM directly to the original representation of image data, corresponding to $28 \cdot 28 = 784$ scalar features. This involves training parameters of the linear classifier in fully-supervised manner, yet only achieves an error probability of 8.17% based on our experimental analysis. This reflects the importance of nonlinearities inherent in the feature formation stages of the architecture.

As another alternative, we can compare this architecture to a DNN with two hidden layers and using sigmoidal activation functions [100], and trained in a fully supervised manner. For instance, using 300 units in the first layer and 100 units in the second corresponds to a total of $300 + 100 = 400$ scalar features, and yields an error probability of 3.05% [166], [167], which is comparable to that of our classifier, which is effectively a network with a *single* hidden layer. As such, this reflects the effectiveness of the universal features extracted via our methodology, which we further emphasize are designed in an unsupervised manner—without taking into account the inference task.

As a further evaluation, in Stage 3 of our architecture, we reduced the amount of (labeled) data used to train the classifier from $n = 60\,000$ to $n/2 = 30\,000$, while still using all n *unlabeled* training samples for Stages 1 and 2. In this case, we obtain an only mildly degraded error probability of 3.4%, which is a reflection of the efficiency with which the architecture uses labeled training data, by restricting its use to the final stage.

As a final comment, we emphasize that the example in this section is not aimed at demonstrating state-of-the-art classification performance on complex data sets. Rather, it is provided simply to illustrate that useful levels of performance can be achieved in nontrivial settings even with comparatively straightforward application of the basic concepts and methodologies in the monograph.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

12

Modal Decomposition of Markov Random Fields

When the distributions of interest have additional structure, as is often the case in practice, we seek modal decompositions whose features capture this additional information. In particular, when

$$X = (X_1, \dots, X_{K_X}) \qquad \text{and} \qquad Y = (Y_1, \dots, Y_{K_Y}) \tag{12.1}$$

for some K_X and K_Y, there are often key relationships among these constituent variables to be reflected in the representation. One important example is obviously the case of jointly Gaussian structure among the variables, as developed in Section 9. In this section, we consider another important form of such structure in the form of conditional independencies, as arise when the variables involved form a Markov random field (MRF).

As we illustrate in this section, the modal decomposition analysis of Section 2 naturally extends to this case and generally expands the number of features. We focus on the case of Markov random fields characterized by pairwise relationships among the constituent variables.

Accordingly, in Section 12.1 we begin with some useful refined notation for the modal decompositions of the corresponding marginal distributions. Section 12.3 develops more detailed results for the case of tree-structured graphical models. Via the further special case of Markov

The version of record is available at: http://dx.doi.org/10.1561/0100000107

chains, connections to spectral graph theory and the graph Laplacian are identified.

12.1 Pairwise Marginal Notation

For notational convenience, we use $X_1, \dots, X_n$ to generically denote the collection of the combined constituent variables in X and Y, viz., (12.1); for example, we would relabel a collection (X_1, X_2, Y_1) as (X_1, X_2, X_3).[1] In turn, based on the results of Section 2, any pairwise marginal of the form p_{X_i,X_j} can be decomposed according to

$$p_{X_i,X_j}(x_i, x_j) = p_{X_i}(x_i)\, p_{X_j}(x_j) \left(1 + \sum_{k=1}^{K_{i,j}-1} \sigma_{i,j,k}\, f^*_{j \to i,k}(x_i)\, f^*_{i \to j,k}(x_j) \right), \tag{12.2}$$

where $K_{i,j} = \min\{|\mathcal{X}_i|, |\mathcal{X}_j|\}$ and the embeddings $f^*_{j \to i,k} \colon \mathcal{X}_i \to \mathbb{R}$ and $f^*_{i \to j,k} \colon \mathcal{X}_j \to \mathbb{R}$ have the properties

$$\mathbb{E}\big[f^*_{j \to i,k}(X_i)\big] = \mathbb{E}\big[f^*_{i \to j,k}(X_i)\big] = 0$$

and

$$\mathbb{E}\big[f^*_{j \to i,k}(X_i)\, f^*_{j \to i,k'}(X_i)\big] = \mathbb{E}\big[f^*_{i \to j,k}(X_i)\, f^*_{i \to j,k'}(X_i)\big] = \mathbb{1}_{k=k'}.$$

In particular, with $\tilde{\mathbf{B}}_{X_i,X_j}$ denoting the $|\mathcal{X}_i| \times |\mathcal{X}_j|$ CDM, whose entries are

$$\tilde{B}_{X_i,X_j}(x_i, x_j) \triangleq \frac{p_{X_i,X_j}(x_i, x_j) - p_{X_i}(x_i)\, p_{X_j}(x_j)}{\sqrt{p_{X_i}(x_i)\, p_{X_j}(x_j)}}, \tag{12.3}$$

for $x_i \in \mathcal{X}_i,\ x_j \in \mathcal{X}_j$, we can express the SVD of $\tilde{\mathbf{B}}_{X_i,X_j}$ as

$$\tilde{\mathbf{B}}_{X_i,X_j} = \sum_{k=1}^{K_{i,j}-1} \sigma_{i,j,k}\, \boldsymbol{\psi}^{X_i}_{j \to i,k} \left(\boldsymbol{\psi}^{X_j}_{i \to j,k}\right)^{\mathrm{T}}, \tag{12.4}$$

where $1 \geq \sigma_{i,j,1} \geq \cdots \geq \sigma_{i,j,K_{i,j}-1} \geq 0$ are the ordered singular values, and

$$\boldsymbol{\psi}^{X_i}_{j \to i,1}, \dots, \boldsymbol{\psi}^{X_i}_{j \to i,K_{i,j}-1} \quad \text{and} \quad \boldsymbol{\psi}^{X_j}_{i \to j,1}, \dots, \boldsymbol{\psi}^{X_j}_{i \to j,K_{i,j}-1}$$

[1] Obviously, in applications it is typically more useful to choose notation that explicitly distinguishes the variables in each of the X and Y subcollections and reflects their roles.

are the corresponding collections of (orthonormal) left and right singular vectors. Using this decomposition, the relevant embeddings are constructed according to

$$f^*_{j\to i,k'}(x_i) \triangleq \frac{\psi^{X_i}_{j\to i,k'}(x_i)}{\sqrt{p_{X_i}(x_i)}} \quad \text{and} \quad f^*_{i\to j,k'}(x_j) \triangleq \frac{\psi^{X_j}_{i\to j,k'}(x_j)}{\sqrt{p_{X_j}(x_j)}}, \tag{12.5}$$

for $k' = 1, \dots, K_{i,j}-1$, where $\psi^{X_i}_{j\to i,k}(x_i)$ and $\psi^{X_i}_{i\to j,k}(x_i)$ are, respectively, the x_ith and x_jth elements in the singular vectors $\boldsymbol{\psi}^{X_i}_{j\to i,k}$ and $\boldsymbol{\psi}^{X_j}_{i\to j,k}$.

12.2 Pairwise Markov Random Fields

Consider a MRF whose graphical representation has edges between arbitrary pairs of nodes, but where the distribution is characterized by pairwise potentials. With $\mathcal{V} = \{1, \dots, n\}$ denoting the set of nodes, and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ the set of edges, such a distribution factors according to

$$P_{X_{\mathcal{V}}}(x_{\mathcal{V}}) \propto \exp\left\{ \sum_{i\in\mathcal{V}} \tilde{\phi}_{X_i}(x_i) + \sum_{(i,j)\in\mathcal{E}} \tilde{\phi}_{X_i,X_j}(x_i, x_j) \right\}, \tag{12.6}$$

where the $\tilde{\phi}_{X_i}$ and $\tilde{\phi}_{X_i,X_j}$ are node and edge (log-)potential functions, respectively. Edge sets $\mathcal{E}$ that do not include all possible node pairs correspond to the case of structure in the distribution.

Under mild conditions, such models are equivalently characterized by their pairwise marginals P_{X_i,X_j}. To see this, note that (12.6) takes the form of an exponential family. Specifically, we have

$$P_{X_{\mathcal{V}}}(x_{\mathcal{V}}) = \frac{1}{Z} \exp\left\{ \sum_{(i,j)\in\mathcal{E}} \sum_{\substack{a_i\in\mathcal{X}_i,\\ a_j\in\mathcal{X}_j}} \underbrace{\tilde{\phi}_{X_i,X_j}(a_i, a_j)}_{\triangleq \theta^{a_i,a_j}_{i,j}} \underbrace{\mathbb{1}_{x_i=a_i, x_j=a_j}}_{\triangleq s^{a_i,a_j}_{i,j}(x_i,x_j)} \right\},$$

where

$$\theta \triangleq \{\theta^{a_i,a_j}_{i,j} : (i,j) \in \mathcal{E},\ a_i \in \mathcal{X}_i,\ a_j \in \mathcal{X}_j\} \tag{12.7}$$

are the natural parameters, and

$$s \triangleq \{s^{a_i,a_j}_{i,j} : (i,j) \in \mathcal{E},\ a_i \in \mathcal{X}_i,\ a_j \in \mathcal{X}_j\} \tag{12.8}$$

are the natural statistics, both of which we express as column vectors. In turn, $\alpha(\theta) \triangleq \log Z$ is the associated log-partition function. For any naturally-parameterized exponential family, the gradient and Hessian of the log-partition function satisfy, respectively,

$$\nabla\alpha(\theta) = \mathbb{E}[S] \tag{12.9a}$$

$$\nabla^2\alpha(\theta) = \mathbb{E}[SS^{\mathrm{T}}], \tag{12.9b}$$

where we note that the latter is positive semidefinite. But

$$\mathbb{E}[S_{i,j}^{a_i,a_j}(X_i,X_j)] = \mathbb{E}[\mathbb{1}_{X_i=a_i,X_j=a_j}] = P_{X_i,X_j}(a_i,a_j),$$

so the right-hand side of (12.9a) is

$$\mathbb{E}[S] = \{P_{X_i,X_j}(a_i,a_j)\colon (i,j)\in\mathcal{E},\ a_i\in\mathcal{X}_i,\ a_j\in\mathcal{X}_j\}. \tag{12.10}$$

Hence, when (12.9b) is strictly positive definite, then (12.9a) is invertible, and thus all the parameters of the distribution (12.6) can be recovered from (12.10).

We conclude that subject to this mild condition, the collection of modal decompositions for the constituent P_{X_i,X_J}, $(i,j)\in\mathcal{E}$, as described in Section 12.1, characterize such an MRF, and yields universal features for such distributions.

12.3 Trees and Markov Chains

The dependency of (12.6) on the pairwise marginals —and thus the CDMs (12.3)—takes a simple form in the key special case in which the graphical model corresponding to the MRF is a tree. In this case, the distribution factors according to

$$P_{X_{\mathcal{V}}}(x_{\mathcal{V}}) = \prod_{i\in\mathcal{V}} P_{X_i}(x_i) \prod_{(i,j)\in\mathcal{E}} \frac{P_{X_i,X_j}(x_i,x_j)}{P_{X_i}(x_i)\,P_{X_j}(x_j)}, \tag{12.11}$$

which using (12.3) can be equivalently expressed in the form

$$P_{X_{\mathcal{V}}}(x_{\mathcal{V}}) = \prod_{i\in\mathcal{V}} P_{X_i}(x_i) \prod_{(i,j)\in\mathcal{E}} \left(1 + \frac{\tilde{B}_{X_i,X_j}(x_i,x_j)}{\sqrt{P_{X_i}(x_i)\,P_{X_j}(x_j)}}\right). \tag{12.12}$$

A further special case of tree-structured models are Markov chains $X_1 \leftrightarrow X_2 \leftrightarrow \cdots \leftrightarrow X_n$. To obtain still further insights, let us additional constrain the class of such chains to those in which: 1) $\mathcal{X}_i = \mathcal{X}$ for some finite $\mathcal{X}$; 2) the chain is homogeneous with $P_{X_i|X_{i-1}}(\cdot|\cdot) = w(\cdot|\cdot)$ for some w; 3) the chain is irreducible; and 4) the chain is reversible. Such chains have a stationary distribution π satisfying detailed balance

$$\pi(x')\, w(x|x') = \pi(x)\, w(x'|x) \triangleq P(x, x').$$

The resulting symmetry means that the joint distribution $P_{X_i,X_{i-1}}$ for such variables can be expressed in terms of unique embeddings $f_k^* \colon \mathcal{X} \to \mathbb{R}$ that are obtained from the SVD

$$\tilde{B}(x, x') \triangleq \frac{P(x, x') - \pi(x)\,\pi(x')}{\sqrt{\pi(x)\,\pi(x')}} = \sum_{k=1}^{|\mathcal{X}|-1} \sigma_k\, \psi_k(x)\, \psi_k(x'), \tag{12.13}$$

whence the modal expansion

$$P(x, x') = \pi(x)\,\pi(x') \left(1 + \sum_{k=1}^{|\mathcal{X}|-1} \sigma_i f_k^*(x)\, f_k^*(x') \right),$$

with

$$f_k^*(x) \triangleq \frac{\psi_k(x)}{\sqrt{\pi(x)}}.$$

Specifically, we have

$$P_{X_{\mathcal{V}}}(x_{\mathcal{V}}) = \prod_{i=1}^{n} \pi(x_i) \prod_{i=2}^{n} \left(1 + \sum_{k=1}^{|\mathcal{X}|-1} \sigma_k\, f_k^*(x_i)\, f_k^*(x_{i-1}) \right). \tag{12.14}$$

Since reversible Markov chains correspond to random walks on a graph with nodes $\mathcal{V}$ and edges $\mathcal{E}$ such that edge (x, x') has weight $P(x, x')$, then the f_k^* correspond to embeddings on this graph.

For the still further special case of an unweighted random walk in which the edge traversed from any node is chosen uniformly at random, the resulting $|\mathcal{X}| \times |\mathcal{X}|$ CDM $\tilde{\mathbf{B}}$ whose entries are (12.13) is essentially equivalent to the so-called Laplacian (or admittance, or Kirchoff) matrix of graph theory, which has numerous related applications [56], [59], [72], [78], [87], [88], [150], [184], [249]. In particular, the DTM $\mathbf{B}$ for this chain, which has entries

$$B(x, x') = \frac{P(x, x')}{\sqrt{\pi(x)\,\pi(x')}}$$

specializes to

$$\mathbf{B} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}, \tag{12.15}$$

where $\mathbf{D}$ and $\mathbf{A}$ are the (diagonal) degree matrix and the adjacency matrix, respectively, associated with the graph. In turn, the (symmetric, normalized [59]) graph Laplacian is

$$\mathbf{L} \triangleq \mathbf{I} - \mathbf{B}.$$

Moreover, since the singular values of a DTM satisfy $\sigma(\mathbf{B}) \leq 1$, the eigenvalues of the (symmetric) DTM (12.15) satisfy $|\lambda(\mathbf{B})| \leq 1$, and thus $\lambda(\mathbf{L})$, whose values are referred to as the *spectrum* of the graph Laplacian, satisfy $0 \leq \lambda(\mathbf{L}) \leq 2$.

From this perspective, the diffusion maps for dimensionality reduction introduced in [61], which generate embeddings from the graph Laplacian, are equivalent to the embeddings f_k^* described above, and thus can be viewed as an instance of Hirschfeld's modal decomposition analysis. Related discussion of graph Laplacians can be found in, e.g., [191].

The version of record is available at: http://dx.doi.org/10.1561/0100000107

13

Emerging Applications and Related Developments

In recent years, there has been rapid growth in activity in the area of statistical inference and machine learning, largely motivated by the increasingly abundant computational resources available for implementing the associated methods. This, in turn, has led to an ever-expanding set of application domains, and a burgeoning literature has sought to address the corresponding challenges. Within this literature is a growing variety rooted in methods, analyses, and perspectives discussed in this monograph. In this section we provide a representative sample of just some of the many active developments in this realm.

One application involves the imposition of privacy constraints in problems of inference. For example, [189], which considers the use of χ^2 analysis in differential privacy problems, replacing mutual information with its more convenient χ^2 counterpart, corresponding to a local geometric analysis. As another example, [46] applies such χ^2 analysis to the privacy funnel problem [190], a dual of the information bottleneck [258]. Similarly, [271], [272] investigates problems of estimation under privacy constraints using χ^2 analysis, and aspects of the corresponding information bottleneck are likewise analyzed in [118], providing additional perspectives on the results in Section 5.6.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Another application area of growing interest is that of imposing fairness constraints in machine learning, using independence, separation, sufficiency, and other criteria [23]. Early uses of χ^2 analysis for this purpose appear in [19], [101], [197]. Refined approaches with low complexity are developed in [169], [170] using further HGR maximal correlation tools, with competitive performance on standard datasets. For learning problems in which abstention is allowed, [40] develops practical methods for imposing fairness in selective classification using methods based on similar HGR maximal correlation tools, with its effectiveness demonstrated on diverse datasets; a treatment of the corresponding selective regression problem is developed in [248]. Still other aspects of efficient learning with fairness constraints using χ^2 formulations are developed in [41], [242].

Another area of application for these methods has been in problems of domain adaptation in machine learning, and transfer learning. Examples of activity in this area include [171], which develops a method of transfer learning from an ensemble of pre-trained networks using a maximal correlation weighting technique, and [22], [263] which develop the use maximal correlation measures to quantify transferability in learning. Aspects of invariant learning for domain adaptation based on χ^2 analysis are also discussed in [242].

Motivated by multimodal learning and related scenarios, a variety of work explores multivariate extensions to HGR maximal correlation and the associated modal decompositions, of which those described in Sections 9 and 12 are just two. Indeed, multivariate extensions have a long history. For example, [161, Chapter XXI] considers χ^2 analysis for higher dimensional contingency tables, and [17] develops a particular multivariate extension of HGR maximal correlation. This extension can be shown to effectively correspond to a local version of Watanabe's "total correlation" generalization of mutual information to multiple variables [236]. More recently, a local version of multivariate common information based on total correlation is developed as part of [130], extending the treatment described in Section 5.7.

Multivariate extensions of the ACE algorithm (termed MACE) are developed in [175], along with associated generalizations of PCA. Motivated by problems involving multimodal data, a methodology referred to as *soft HGR* is introduced in [273], and applied to some representative problems. And [285] further considers multivariate feature extraction and related problems, including learning with side information. Other multivariate extensions arise in [85] in the context of a network model with graph structure; related results on graphs appear in [79], which includes some associated sample complexity analysis.

Other aspects of computing and robustly approximating modal decompositions are explored in, e.g., [282], which investigates the incorporation of Oja's algorithm [213], and [125], which studies a noisy version of ACE corresponding to conditional expectation operator approximation. Connections of the latter to multilayer residual learning [111] are also developed. Other aspects of robustness, including regularization methods aimed at reducing the vulnerability to adversarial attacks, appear in [177]. Complementing such work, sample complexity analysis is a similarly active area. For example, beyond the content of Section 6.2 and [193], and the specialized results in [79], [280], additional results to the topic appear in [262], [284].

A variety of problems of feature selection and learning involving limited or incomplete labeled data have been approached using the resulting tools. Examples include unsupervised learning in the form of clustering using maximal matrix norm couplings [227], as well as other forms of unsupervised feature selection and learning [128]. Other examples involve self- and semi-supervised learning [280], [286] with accompanying sample complexity analysis, and data augmentation [186]. Still others investigate methods for multilabel learning with missing labels [175]. Moreover, complementary activity in the context of distributed learning includes [261], [283].

Specific application domains continue to receive new or renewed attention. For example, there has been a variety of work in emotion recognition [185], [187], [188], [290], multimodal person recognition [181], and traffic and mobility pattern analysis [178]–[180]. Likewise, cross-modal retrieval has emerged as a potential application [176], and potential applications in communication and compression arise rather naturally. Examples include distributed source and channel coding [143], and communication of type classes [121].

Within the realm of nonlinear CCA and PCA, there continues to be significant activity, building on that listed in Section 10.4. Examples include the deep CCA architectures proposed in [119], [120] and the maximal correlation regression framework of [281]. In such directions, information theoretic interpretations of deep networks, such as those in [129], [131] that extend the analysis in Section 8, are likely to be valuable. Similarly, information-theoretic approaches to nonlinear PCA using maximal correlation tools are developed in [128], and the use of maximal correlation methods in PCA are explored in [57], [86]. Likewise, there is renewed interest in other forms of constraints on modal decompositions to expand their range of potential applications. For example, [81] arrives at monotonicity constraints on features [149] via an axiomatic approach to obtaining measures of dependence.

Another promising application of the tools of this monograph is in independent component analysis (ICA). An early use of neural networks in ICA appears in e.g., [144], [215], and kernel methods for ICA are developed in [17], where maximal correlation is used to approximate mutual information as a contrast function. Combining such work with information theoretic analyses of ICA—such as that in [172] and, for finite alphabets, [217], [219]—may lead to still further advances.

Ultimately, the examples of activity summarized in this section are only a sampling, and many more directions remain to be investigated. As such, there are abundant opportunities for further research and fruitful application across multiple communities.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Acknowledgements

The patient guidance from the Editor-in-Chief Alexander Barg, and the detailed feedback from the reviewers are gratefully acknowledged. All helped to significantly improve the manuscript in several respects. Thanks also to Dr. Yuheng Bu for his careful proofreading of an early draft.

This work was sponsored, in part, by NSF under Grant Nos. CCF-1717610 and CCF-1816209, and the MIT-IBM Watson AI Lab under Agreement No. W1771646, and supported by efforts with AFRL and the Air Force AI Accelerator, and ONR.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Appendices

A

Appendices for Section 2

A.1 Proof of Proposition 2.1

It suffices to show that the maximum eigenvalue of $\mathbf{B}\mathbf{B}^{\mathrm{T}}$ is at most unity. To this end, note that via (2.10) and (2.11) we have

$$\mathbf{B}\mathbf{B}^{\mathrm{T}} = \left[\sqrt{\mathbf{P}_Y}\right]^{-1} \mathbf{P}_{Y|X} \mathbf{P}_{X|Y} \sqrt{\mathbf{P}_Y}. \tag{A.1}$$

Now $\mathbf{P}_{Y|X}$ and $\mathbf{P}_{X|Y}$ are both column-stochastic matrices, so their product $\mathbf{P}_{Y|X}\,\mathbf{P}_{X|Y}$ is as well. As such, this product has maximum eigenvalue of unity, which follows from, e.g., [114, Theorem 8.3.4] and the fact that by definition a matrix $\mathbf{A}$ is column stochastic if $\mathbf{1}^{\mathrm{T}}\mathbf{A} = \mathbf{1}^{\mathrm{T}}$. Finally, since (A.1) is a similarity transformation of $\mathbf{P}_{Y|X}\,\mathbf{P}_{X|Y}$, it has the same eigenvalues.

Finally, (2.14b) can be verified by direct calculation using (2.3):

$$\sum_{x\in\mathcal{X}} B(x,y)\sqrt{P_X(x)} = \frac{1}{\sqrt{P_Y(y)}} \sum_{x\in\mathcal{X}} P_{X,Y}(x,y) = \sqrt{P_Y(y)}$$
$$\sum_{y\in\mathcal{Y}} \bar{B}(y,x)\sqrt{P_Y(y)} = \frac{1}{\sqrt{P_X(x)}} \sum_{y\in\mathcal{Y}} P_{X,Y}(x,y) = \sqrt{P_X(x)}.$$

■

The version of record is available at: http://dx.doi.org/10.1561/0100000107

A.2 DTM Characterization

As notation, let $\mathbf{B}(P_{X,Y})$ denote the $\mathcal{Y} \times \mathcal{X}$ dimensional DTM associated with joint distribution $P_{X,Y}$. Moreover, let $\mathcal{B}^{\mathcal{X}\times\mathcal{Y}}$ denote the set of all DTMs, i.e.,

$$\mathcal{B}^{\mathcal{X}\times\mathcal{Y}} = \mathbf{B}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}), \tag{A.2}$$

and let $\mathcal{B}_{\circ}^{\mathcal{X}\times\mathcal{Y}}$ denote the set of all DTMs corresponding to distributions with positive probabilities, i.e.,

$$\mathcal{B}_{\circ}^{\mathcal{X}\times\mathcal{Y}} = \mathbf{B}(\operatorname{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})). \tag{A.3}$$

As further notation, for a matrix $\mathbf{A}$, we use $\mathbf{A} > \mathbf{0}$ to denote that every entry of $\mathbf{A}$ is positive, and, likewise $\mathbf{A} \geq \mathbf{0}$ when all entries are nonnegative.

The following proposition characterizes $\mathcal{B}_{\circ}^{\mathcal{X}\times\mathcal{Y}}$ in (A.3).

Proposition A.1. A matrix $\mathbf{M}$ is a DTM corresponding to a joint distribution in $\operatorname{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$ if and only if $\mathbf{M} > \mathbf{0}$ and $\|\mathbf{M}\|_{\mathrm{s}} = 1$, i.e.,

$$\mathcal{B}_{\circ}^{\mathcal{X}\times\mathcal{Y}} = \{\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} \colon \mathbf{M} > \mathbf{0} \text{ and } \|\mathbf{M}\|_{\mathrm{s}} = 1\}, \tag{A.4}$$

where $\|\cdot\|_{\mathrm{s}}$ denotes the spectral norm of its argument.

Proof. The "only if" part of the claim is immediate. Indeed, since $\mathbf{M} = \mathbf{B}(P_{X,Y})$ for some positive $P_{X,Y}$, it follows that $\mathbf{M} > \mathbf{0}$. Moreover, as developed in Section 2, $\|\mathbf{B}(P_{X,Y})\|_{\mathrm{s}} = 1$.

For the "if" part of the claim, consider any matrix $\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ satisfying $\mathbf{M} > \mathbf{0}$ and $\|\mathbf{M}\|_{\mathrm{s}} = 1$. We can construct a $P_{X,Y}$ for which $\mathbf{M}$ is its DTM. To see this, first note that $\mathbf{M}^{\mathrm{T}}\mathbf{M} > \mathbf{0}$, $\mathbf{M}\mathbf{M}^{\mathrm{T}} > \mathbf{0}$, and $\lambda(\mathbf{M}^{\mathrm{T}}\mathbf{M}) = \|\mathbf{M}\|_{\mathrm{s}}^2 = 1$, where $\lambda(\cdot)$ denotes the largest eigenvalue of its argument. Then, applying the Perron-Frobenius theorem [114, Theorem 8.2.2], it follows that there exist unit-norm vectors $\boldsymbol{\psi}^X(\mathbf{M})$ and $\boldsymbol{\psi}^Y(\mathbf{M})$ with positive elements such that

$$\mathbf{M}^{\mathrm{T}}\mathbf{M}\,\boldsymbol{\psi}^X(\mathbf{M}) = \boldsymbol{\psi}^X(\mathbf{M}) \quad \text{and} \quad \mathbf{M}\mathbf{M}^{\mathrm{T}}\boldsymbol{\psi}^Y(\mathbf{M}) = \boldsymbol{\psi}^Y(\mathbf{M}).$$

In turn, this implies that $\boldsymbol{\psi}^X(\mathbf{M})$ and $\boldsymbol{\psi}^Y(\mathbf{M})$ are the right and left singular vectors corresponding to the unit principal singular value of $\mathbf{M}$, respectively, i.e.,

$$\mathbf{M}\,\boldsymbol{\psi}^X(\mathbf{M}) = \boldsymbol{\psi}^Y(\mathbf{M}) \quad \text{and} \quad \mathbf{M}^{\mathrm{T}}\boldsymbol{\psi}^Y(\mathbf{M}) = \boldsymbol{\psi}^X(\mathbf{M}). \tag{A.5}$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

We now define a $P_{X,Y}$ lying on the simplex, and show that its DTM is $\mathbf{M}$. In particular, we let

$$P_{X,Y}(x,y) \triangleq \Big[\mathbf{P}_{Y,X}\Big]_{y,x}, \quad x \in \mathcal{X},\ y \in \mathcal{Y},$$

where

$$\mathbf{P}_{Y,X} \triangleq \operatorname{diag}(\boldsymbol{\psi}^Y(\mathbf{M}))\,\mathbf{M}\operatorname{diag}(\boldsymbol{\psi}^X(\mathbf{M})), \tag{A.6}$$

with diag(·) denoting a diagonal matrix with diagonal entries specified by its (vector) argument.

That $P_{X,Y}$ is positive follows by construction, since the quantities forming $\mathbf{P}_{Y,X}$ are all positive. To verify that it sums to unity, observe that

$$\begin{aligned}\sum_{x,y} P_{X,Y}(x,y) &= \mathbf{1}^{\mathrm{T}}\mathbf{P}\,\mathbf{1}\\ &= \mathbf{1}^{\mathrm{T}}\operatorname{diag}(\boldsymbol{\psi}^Y(\mathbf{M}))\;\mathbf{M}\;\operatorname{diag}(\boldsymbol{\psi}^X(\mathbf{M}))\;\mathbf{1}\\ &= \boldsymbol{\psi}^Y(\mathbf{M})^{\mathrm{T}}\mathbf{M}\,\boldsymbol{\psi}^X(\mathbf{M})\\ &= \boldsymbol{\psi}^Y(\mathbf{M})^{\mathrm{T}}\boldsymbol{\psi}^Y(\mathbf{M})\\ &= 1.\end{aligned}$$

Moreover, applying (A.5), we obtain that the marginals take the form

$$\begin{aligned}P_Y(y) &= \mathbf{P}\,\mathbf{1}\\ &= \operatorname{diag}(\boldsymbol{\psi}^Y(\mathbf{M}))\;\mathbf{M}\;\operatorname{diag}(\boldsymbol{\psi}^X(\mathbf{M}))\;\mathbf{1}\\ &= \operatorname{diag}(\boldsymbol{\psi}^Y(\mathbf{M}))\;\mathbf{M}\,\boldsymbol{\psi}^X(\mathbf{M})\\ &= \operatorname{diag}(\boldsymbol{\psi}^Y(\mathbf{M}))\;\boldsymbol{\psi}^Y(\mathbf{M})\\ &= \boldsymbol{\psi}^Y(\mathbf{M})^2 \qquad \text{(A.7a)}\end{aligned}$$

and

$$\begin{aligned}P_X(x) &= \mathbf{P}^{\mathrm{T}}\mathbf{1}\\ &= \operatorname{diag}(\boldsymbol{\psi}^X(\mathbf{M}))\;\mathbf{M}^{\mathrm{T}}\operatorname{diag}(\boldsymbol{\psi}^Y(\mathbf{M}))\;\mathbf{1}\\ &= \operatorname{diag}(\boldsymbol{\psi}^X(\mathbf{M}))\;\mathbf{M}^{\mathrm{T}}\boldsymbol{\psi}^Y(\mathbf{M})\\ &= \operatorname{diag}(\boldsymbol{\psi}^X(\mathbf{M}))\;\boldsymbol{\psi}^X(\mathbf{M})\\ &= \boldsymbol{\psi}^X(\mathbf{M})^2, \qquad \text{(A.7b)}\end{aligned}$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

where $\boldsymbol{\psi}^X(\mathbf{M})^2$ and $\boldsymbol{\psi}^Y(\mathbf{M})^2$ are vectors whose elements are the squares of the elements of $\boldsymbol{\psi}^X(\mathbf{M})$ and $\boldsymbol{\psi}^Y(\mathbf{M})$, respectively.

Hence, using (A.7) in (A.6) we obtain

$$\begin{aligned}\mathbf{M} &= \left[\sqrt{\operatorname{diag}(\boldsymbol{\psi}^Y(\mathbf{M}))}\right]^{-1}\mathbf{P}_{Y,X}\left[\sqrt{\operatorname{diag}(\boldsymbol{\psi}^X(\mathbf{M}))}\right]^{-1}\\ &= \left[\sqrt{\mathbf{P}_Y}\right]^{-1}\mathbf{P}_{Y,X}\left[\sqrt{\mathbf{P}_X}\right]^{-1},\end{aligned} \tag{A.8}$$

where $\mathbf{P}_X$ and $\mathbf{P}_Y$ are diagonal matrices whose diagonal elements are the elements of P_X and P_Y, respectively, which are all positive. Hence, $\mathbf{M}$ is the DTM corresponding to the $P_{X,Y}$ we have constructed, i.e., $\mathbf{M} = \mathbf{B}(P_{X,Y})$. ■

The following generalization of Proposition A.1 characterizes $\mathcal{B}^{\mathcal{X}\times\mathcal{Y}}$ in (A.2).

Proposition A.2. A matrix $\mathbf{M}$ is a DTM corresponding to a joint distribution in $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ if and only if $\mathbf{M} \geq \mathbf{0}$, $\|\mathbf{M}\|_{\mathrm{s}} = 1$, and each of $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ and $\mathbf{M}\mathbf{M}^{\mathrm{T}}$ have a positive eigenvector corresponding to their unit eigenvalue, i.e.,

$$\begin{aligned}\mathcal{B}^{\mathcal{X}\times\mathcal{Y}} = \Big\{&\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} \colon \mathbf{M} \geq \mathbf{0},\ \|\mathbf{M}\|_{\mathrm{s}} = 1,\\ &\exists\, \boldsymbol{\psi}^X(\mathbf{M}) > \mathbf{0} \text{ s.t. } \mathbf{M}^{\mathrm{T}}\mathbf{M}\,\boldsymbol{\psi}^X(\mathbf{M}) = \boldsymbol{\psi}^X(\mathbf{M}),\\ &\exists\, \boldsymbol{\psi}^Y(\mathbf{M}) > \mathbf{0} \text{ s.t. } \mathbf{M}\mathbf{M}^{\mathrm{T}}\boldsymbol{\psi}^Y(\mathbf{M}) = \boldsymbol{\psi}^Y(\mathbf{M})\Big\}.\end{aligned}$$

Proof. For the "only if" part, since $\mathbf{M} = \mathbf{B}(P_{X,Y})$ for some $P_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$, it follows by construction that $\mathbf{M} \geq \mathbf{0}$, and as developed in Section 2, $\|\mathbf{B}\|_{\mathrm{s}} = 1$ with corresponding right and left principal singular vectors $\boldsymbol{\psi}^X(\mathbf{M})$ and $\boldsymbol{\psi}^Y(\mathbf{M})$ whose elements are $\{\sqrt{P_X},\ x \in \mathcal{X}\}$ and $\{\sqrt{P_Y},\ y \in \mathcal{Y}\}$, respectively, which are positive according to our assumption at the outset of this appendix. As such, these positive $\boldsymbol{\psi}^X(\mathbf{M})$ and $\boldsymbol{\psi}^Y(\mathbf{M})$ must be eigenvectors of $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ and $\mathbf{B}\mathbf{B}^{\mathrm{T}}$ corresponding to the unit eigenvalue.

The "if" part follows from the same proof as that for Proposition A.1 mutatis mutandis. However, we must be careful when applying the Perron-Frobenius theorem [114, Theorem 8.3.1] to $\mathbf{M} \geq \mathbf{0}$ as it only guarantees that the eigenvectors $\boldsymbol{\psi}^X(\mathbf{M})$ and $\boldsymbol{\psi}^Y(\mathbf{M})$ are entrywise nonnegative. If an entry of $\boldsymbol{\psi}^X(\mathbf{M})$ or $\boldsymbol{\psi}^Y(\mathbf{M})$ were zero, then the corresponding column or row of

$$\mathbf{P}_{Y,X} = \mathrm{diag}(\boldsymbol{\psi}^Y(\mathbf{M}))\, \mathbf{M}\, \mathrm{diag}(\boldsymbol{\psi}^X(\mathbf{M})),$$

which defines $P_{X,Y}$, would be zero. In turn, this would imply that $P_X(x) = 0$ for some $x \in \mathcal{X}$ or $P_Y(y) = 0$ for some $y \in \mathcal{Y}$, which would mean that $P_{X,Y} \notin \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$, so that $\mathbf{M}$ could not be a DTM. Accordingly, we add the $\boldsymbol{\psi}^X(\mathbf{M}) > \mathbf{0}$ and $\boldsymbol{\psi}^Y(\mathbf{M}) > \mathbf{0}$ conditions in the statement of the proposition. ■

Remark A.3. It is worth noting that a nonnegative square matrix $\mathbf{A} \geq \mathbf{0}$ has positive left and right eigenvectors corresponding to its Perron-Frobenius eigenvalue (or spectral radius) $\rho(\mathbf{A})$ if and only if the triangular block form of $\mathbf{A}$ is a direct sum of irreducible nonnegative square matrices whose spectral radii are also $\rho(\mathbf{A})$—see Theorem 3.14 and the preceding discussion in [32, Chapter 2, Section 3]. This means that $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ and $\mathbf{M}\mathbf{M}^{\mathrm{T}}$ have positive eigenvectors corresponding to their spectral radius of unity if and only if they have the aforementioned direct form structure after suitable similarity transformations using permutation matrices.

Finally, we establish the following.

Proposition A.4. The DTM function $\mathbf{B}\colon \mathcal{P}^{\mathcal{X}\times\mathcal{Y}} \to \mathcal{B}^{\mathcal{X}\times\mathcal{Y}}$ is bijective and continuous.

Proof. The DTM function $\mathbf{B}\colon \mathcal{P}^{\mathcal{X}\times\mathcal{Y}} \to \mathcal{B}^{\mathcal{X}\times\mathcal{Y}}$ is bijective because: 1) its range is defined to be $\mathcal{B}^{\mathcal{X}\times\mathcal{Y}}$; and 2) the proof of Proposition A.1 (and, in turn, its extension Proposition A.2) delineates the inverse function.

To prove that $\mathbf{B}\colon \mathcal{P}^{\mathcal{X}\times\mathcal{Y}} \to \mathcal{B}^{\mathcal{X}\times\mathcal{Y}}$ is continuous, consider any sequence of distributions $\{P_{X,Y}^n \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}},\ n = 1, 2, \ldots\}$ such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$

$$\lim_{n\to\infty} P_{X,Y}^n(x, y) = P_{X,Y}(x, y).$$

By the triangle inequality, we have, for all $x \in \mathcal{X}$,

$$\begin{aligned} \left|P_X^n(x) - P_X(x)\right| &= \left|\sum_{y\in\mathcal{Y}} P_{X,Y}^n(x,y) - P_{X,Y}(x,y)\right| \\ &\leq \sum_{y\in\mathcal{Y}} \left|P_{X,Y}^n(x,y) - P_{X,Y}(x,y)\right|, \end{aligned}$$

which implies that $P_X^n(x) \to P_X(x)$ as $n \to \infty$ for all $x \in \mathcal{X}$. Likewise, $P_Y^n(y) \to P_Y(y)$ as $n \to \infty$ for all $y \in \mathcal{Y}$. Hence, we have

$$\lim_{n\to\infty} \frac{P_{X,Y}^n(x,y)}{\sqrt{P_X^n(x)P_Y^n(y)}} = \frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)P_Y(y)}}$$

for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, which means that the elements of $\mathbf{B}(P_{X,Y}^n)$ converge to the elements of $\mathbf{B}(P_{X,Y})$, and where we note that the denominator terms are positive according to our assumption at the outset of this appendix. Therefore, the DTM function is continuous. ■

A.3 Conditional Expectation Operator Representations

It is reasonable to ask why it is natural to focus on the SVD of the CDM $\tilde{\mathbf{B}}$ corresponding to $\tilde{B}$, as opposed to other commonly used representations of the conditional expectation operator $P_{X|Y}$, such as simply

$$B_0(x,y) \triangleq P_{X,Y}(x,y),$$

or

$$B_1(x,y) \triangleq \frac{P_{X,Y}(x,y)}{P_X(x)\,P_Y(y)},$$

whose logarithm is the pointwise mutual information [60] (also referred to as the information density [107]). While fulling addressing this question is beyond the scope of the present development, we can show that $\tilde{B}$ generates inner product spaces with the "right" properties, and that it does so uniquely over a reasonable class of candidates.

Our characterization takes the form of the following proposition, in which $\mathbf{P}_{X|Y}$ is the representation of the conditional expectation operator $\mathbb{E}[\cdot|Y=y]$ defined in (2.11). In particular, expressing a function f as a length-$|\mathcal{X}|$ (column) vector $\mathbf{f}$,

$$\mathbb{E}[f(X)|Y=y] = \mathbf{P}_{X|Y}^{\mathrm{T}}\mathbf{f}.$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Proposition A.5. Define an inner product on $\mathbb{R}^{|\mathcal{X}|}$ using a distribution Q_X

$$\langle \mathbf{f}_1, \mathbf{f}_2 \rangle_{Q_X} \triangleq \sum_{x \in \mathcal{X}} Q_X(x)\, f_1(x)\, f_2(x),$$

yielding $\ell^2(\mathcal{X}, Q_X)$, and similarly use P_Y to convert $\mathbb{R}^{|\mathcal{Y}|}$ into $\ell^2(\mathcal{Y}, P_Y)$, i.e.,

$$\langle \mathbf{g}_1, \mathbf{g}_2 \rangle_{P_Y} \triangleq \sum_{y \in \mathcal{Y}} P_Y(y)\, g_1(y)\, g_2(y).$$

Then

$$\min_{Q_X} \max_{\mathbf{f} \in \ell^2(\mathcal{X}, Q_X)} \frac{\|\mathbf{P}_{X|Y}^{\mathrm{T}} \mathbf{f}\|_{P_Y}}{\|\mathbf{f}\|_{Q_X}} = 1,$$

and, moreover,

$$Q_X^* = P_X = P_Y\, P_{X|Y}$$

is the unique minimizer.

Proof. Note that for all Q_X we have $\mathbf{1} \in \ell^2(\mathcal{X}, Q_X)$ with $\|\mathbf{1}\|_{Q_X} = 1$. Also, $\|\mathbf{P}_{X|Y}^{\mathrm{T}} \mathbf{1}\|_{P_Y} = \|\mathbf{1}\|_{P_Y} = 1$. Hence,

$$\max_{\mathbf{f} \in \ell^2(\mathcal{X}, Q_X)} \frac{\|\mathbf{P}_{X|Y}^{\mathrm{T}} \mathbf{f}\|_{P_Y}}{\|\mathbf{f}\|_{Q_X}} \geq 1, \quad \text{for all } P_Y \text{ and } Q_X.$$

But we know that $Q_X = P_X$ achieves the lower bound, which proves the minimum. In particular, via Jensen's inequality we have

$$\mathbb{E}\Big[\mathbb{E}[f(X)|Y]^2\Big] \leq \mathbb{E}\Big[\mathbb{E}[f(X)^2|Y]\Big] = \mathbb{E}[f(X)^2].$$

To prove that P_X is the unique minimizer, suppose we use $Q_X \neq P_X$ for the inner product. Then consider the adjoint operator, which for $\mathbf{f} \in \ell^2(\mathcal{X}, Q_X)$ and $\mathbf{g} \in \ell^2(\mathcal{Y}, P_Y)$ is defined by

$$\begin{aligned} \mathbb{E}_{P_Y}\left[\mathbb{E}[f(X)|Y]\, g(Y)\right] &= \mathbb{E}_{P_{X,Y}}\left[f(X)\, g(Y)\right] \\ &= \mathbb{E}_{P_X}\left[f(X)\, \mathbb{E}[g(Y)|X]\right] \\ &= \mathbb{E}_{Q_X}\left[f(X)\, \mathbb{E}[g(Y)|X]\, \frac{P_X(X)}{Q_X(X)}\right]. \end{aligned}$$

So the adjoint operator is

$$(P_{X|Y}^*\, g)(x) = \frac{P_X(x)}{Q_X(x)}\, \mathbb{E}_{P_{Y|X}}\left[g(Y)|X = x\right].$$

Now observe that

$$(P_{X|Y}^{*}\,1)(x) = \frac{P_X(x)}{Q_X(x)},$$

so $\|\mathbf{1}\|_{P_Y} = 1$ and

$$\|P_{X|Y}^{*}\,1\|_{Q_X}^{2} = \sum_{x\in\mathcal{X}} Q_X(x)\frac{P_X(x)^2}{Q_X(x)^2} = 1+\chi^2(Q\|P) > 1,$$

where the last inequality follows because $P_X \neq Q_X$. Hence the largest singular value of $P_{X|Y}^{*}$ is greater than unity. Hence, $Q_X^{*} = P_X$ is the unique minimizer. ■

Note that Proposition A.5 shows that given $P_{X,Y}$, the *only* choice of inner products that make $P_{Y|X}$ and $P_{X|Y}$ adjoints and contractive operators (so that the data processing inequality is satisfied locally) are those with respect to P_X and P_Y. It also establishes that if we are given only $P_{X|Y}$, we are free to choose P_Y, but we must choose the corresponding P_X for the other inner product to obtain the required contraction property.

We comment that the restriction in Proposition A.5 to inner products corresponding to weighting by distribution is natural. In general, each inner product corresponds to a positive semidefinite matrix $\mathbf{A}$, i.e., $\langle \mathbf{f}, \mathbf{g}\rangle_{\mathbf{A}} = \mathbf{f}^{\mathrm{T}}\mathbf{A}\mathbf{g}$. For simplicity, we neglect the orthogonal matrices in the spectral decomposition of $\mathbf{A}$, and only consider diagonal matrices $\mathbf{A}$ with positive diagonal entries, which correspond to weighted inner products. Moreover, we restrict the diagonal entries to sum to unity to have a “well-defined” problem (indeed, allowing arbitrary scaling would make the infimum in our proposition zero).

The version of record is available at: http://dx.doi.org/10.1561/0100000107

B

Appendices for Section 4

B.1 Proof of Proposition 4.2

The first part of the proposition is immediate: from (4.7) we obtain both

$$\mathbb{E}_{P_0}[h(Z)] = \sum_{z\in\mathcal{Z}} P_0(z)\,h(z) = \frac{1}{\epsilon}\sum_{z\in\mathcal{Z}}(P(z)-P_0(z)) = 0,$$

and

$$\xi(z) = \sqrt{P_0(z)}\,h(z) = \frac{P(z)-P_0(z)}{\epsilon\sqrt{P_0(z)}} = \phi(z), \tag{B.1}$$

where we have further used (4.6), and where to obtain the last equality we have used (4.2). The second part of the proposition is trivially true when $h \equiv 0$. When $h \not\equiv 0$, it suffices to note that P in (4.10) satisfies

$$\sum_{z\in\mathcal{Z}} P(z) = \sum_{z\in\mathcal{Z}} P_0(z) + \epsilon\mathbb{E}_{P_0}[h(Z)] = 1,$$

and that $P(z)\in[0,1]$ whenever

$$\epsilon \le \min\left\{\frac{-1}{\min\limits_{z\in\mathcal{Z}} h(z)}\,,\,\frac{1-\max\limits_{z\in\mathcal{Z}} P_0(z)}{\max\limits_{z\in\mathcal{Z}} h(z)\,\max\limits_{z\in\mathcal{Z}} P_0(z)}\right\},$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

where we have used that since $h \not\equiv 0$,

$$\min_{z \in \mathcal{Z}} h(z) < 0 \quad \text{and} \quad \max_{z \in \mathcal{Z}} h(z) > 0,$$

and that since $P_0 \in \operatorname{relint}(\mathcal{P}^{\mathcal{Z}})$,

$$\max_{z \in \mathcal{Z}} P_0(z) < 1.$$

Finally, using, in turn, (4.2), (4.10), and (4.6), we obtain

$$\phi(z) = \frac{P(z) - P_0(z)}{\epsilon\sqrt{P_0(z)}} = \sqrt{P_0(z)}\, h(z) = \xi(z).$$

■

B.2 Proof of Corollary 4.3

To obtain the first part of the corollary, we have that given P there exists h such that

$$\frac{1}{\epsilon} \log \frac{P(z)}{P_0(z)} = \frac{1}{\epsilon} \log(1 + \epsilon h(z)) = h(z) + h_\epsilon(z), \tag{B.2}$$

with $h_\epsilon(z)$ denoting an $\mathcal{o}(1)$ term, where to obtain the first equality we have used (4.7), and to obtain the second equality we have used the first-order Taylor series approximation $\log(1+\omega) = \omega + \mathcal{o}(\omega)$. In turn, using (4.8) it follows that

$$h_{\text{LL}}(z) = h(z) + \tilde{h}_\epsilon(z) \tag{B.3}$$

where $\tilde{h}_\epsilon$ is a function such that

$$\tilde{h}_\epsilon(z) = \mathcal{o}(1),\ \epsilon \to 0,\ z \in \mathcal{Z} \quad \text{and} \quad \mathbb{E}_{P_0}\left[\tilde{h}_\epsilon(Z)\right] = 0. \tag{B.4}$$

Multiplying both sides of (B.3) by $\sqrt{P_0(z)}$ yields the (4.13).

To obtain the second part of the corollary, we use from (4.10) that given h satisfying (4.8) there exists P such that (B.2) holds for sufficiently small ϵ. Subtracting the mean with respect to P_0 from (B.2) then yields (4.14). ■

B.3 Proof of Lemma 4.4

We have

$$\begin{aligned}\mathbb{E}_P[h(Z)] &= \sum_{z\in\mathcal{Z}} P(z)\,h(z)\\ &= \sum_{z\in\mathcal{Z}} \left(P_0(z)+\epsilon\sqrt{P_0(z)}\;\phi(z)\right)\frac{\xi(z)}{\sqrt{P_0(z)}} \qquad \text{(B.5)}\\ &= \sum_{z\in\mathcal{Z}} \sqrt{P_0(z)}\,\xi(z)+\epsilon\sum_{z\in\mathcal{Z}}\phi(z)\,\xi(z) \qquad \text{(B.6)}\\ &= \epsilon\langle\phi,\xi\rangle, \qquad \text{(B.7)}\end{aligned}$$

where to obtain (B.5) we have used (4.2) and (4.6), and to obtain (B.7) we have used that (4.8), i.e., $\xi \in \mathcal{J}^{\mathcal{Z}}(P_0)$, implies that the first term in (B.6) is zero in accordance with (4.5). ■

B.4 Proof of Lemma 4.5

With the feature functions

$$L_i(z) \triangleq \frac{1}{\epsilon}\left(\frac{P_i(z)}{P_0(z)}-1\right)$$

we have, for $i = 1, 2$,

$$\begin{aligned}\log\frac{P_i(z)}{P_0(z)} &= \log(1+\epsilon L_i(z))\\ &= \epsilon L_i(z)-\frac{1}{2}\epsilon^2 L_i(z)^2+o(\epsilon^2) \qquad \text{(B.8)}\\ &= \epsilon\frac{\phi_i(z)}{\sqrt{P_0(z)}}-\frac{1}{2}\epsilon^2\frac{\phi_i(z)^2}{P_0(z)}+o(\epsilon^2), \qquad \text{(B.9)}\end{aligned}$$

where to obtain (B.8) we have used the second-order Taylor series approximation

$$\log(1+\omega)=\omega-\frac{1}{2}\omega^2+o(\omega^2), \quad \omega\to 0,$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

and where to obtain (B.9) we have used (4.9) from the first part of Proposition 4.2. Hence,

$$\begin{aligned}\log\frac{P_1(z)}{P_2(z)} &= \log\frac{P_1(z)}{P_0(z)} - \log\frac{P_2(z)}{P_0(z)}\\ &= \epsilon\frac{\phi_1(z)-\phi_2(z)}{\sqrt{P_0(z)}} - \frac{1}{2}\epsilon^2\frac{\phi_1(z)^2-\phi_2(z)^2}{P_0(z)} + o(\epsilon^2),\end{aligned}$$

and, in turn,

$$\begin{aligned}D(P_1\|P_2) &= \sum_{z\in\mathcal{Z}} P_1(z)\log\frac{P_1(z)}{P_2(z)}\\ &= \sum_{z\in\mathcal{Z}} P_0(z)\log\frac{P_1(z)}{P_2(z)} + \sum_{z\in\mathcal{Z}}(P_1(z)-P_0(z))\log\frac{P_1(z)}{P_2(z)}\\ &= \sum_{z\in\mathcal{Z}} P_0(z)\,\epsilon\,\frac{\phi_1(z)-\phi_2(z)}{\sqrt{P_0(z)}}\\ &\qquad - \sum_{z\in\mathcal{Z}} P_0(z)\,\frac{1}{2}\epsilon^2\,\frac{\phi_1(z)^2-\phi_2(z)^2}{P_0(z)}\\ &\qquad\qquad + \sum_{z\in\mathcal{Z}}\epsilon\sqrt{P_0(z)}\,\phi_1(z)\,\epsilon\frac{\phi_1(z)-\phi_2(z)}{\sqrt{P_0(z)}}\\ &\qquad\qquad\qquad - \sum_{z\in\mathcal{Z}}\epsilon\sqrt{P_0(z)}\,\phi_1(z)\,\frac{1}{2}\epsilon^2\,\frac{\phi_1(z)^2-\phi_2(z)^2}{P_0(z)}\\ &\qquad\qquad\qquad\quad + o(\epsilon^2)\\ &= 0 - \frac{1}{2}\epsilon^2\sum_{z\in\mathcal{Z}}(\phi_1(z)^2-\phi_2(z)^2)\\ &\qquad + \epsilon^2\sum_{z\in\mathcal{Z}}\phi_1(z)(\phi_1(z)-\phi_2(z)) + o(\epsilon^2)\\ &= \frac{1}{2}\epsilon^2[\|\phi_2\|^2 - \|\phi_1\|^2 + 2\|\phi_1\|^2 - 2\langle\phi_1,\phi_2\rangle] + o(\epsilon^2)\\ &= \frac{1}{2}\epsilon^2\|\phi_1-\phi_2\|^2 + o(\epsilon^2).\end{aligned}$$

■

The version of record is available at: http://dx.doi.org/10.1561/0100000107

B.5 Proof of Lemma 4.6

To obtain (4.19), it suffices to note that since

$$\tilde{P}_0(z) = P_0(z) + \epsilon\sqrt{P_0(z)}\,\tilde{\phi}_0(z)$$

for some $\tilde{\phi}_0(z)$ such that $\|\tilde{\phi}_0\| \le 1$, we have, for $i = 1, 2$,

$$\begin{aligned}\tilde{\phi}_i(z) &= \frac{P_i(z) - (P_0(z) + \epsilon\sqrt{P_0(z)}\,\tilde{\phi}_0(z))}{\epsilon\sqrt{P_0(z) + \epsilon\sqrt{P_0(z)}\,\tilde{\phi}_0(z)}} \\ &= \frac{(P_i(z) - P_0(z)) - \epsilon\sqrt{P_0(z)}\,\tilde{\phi}_0(z)}{\epsilon\sqrt{P_0(z)}\sqrt{1 + \epsilon\frac{\tilde{\phi}_0(z)}{\sqrt{P_0(z)}}}} \\ &= (\phi_i(z) - \tilde{\phi}_0(z))\,(1 + o(1)),\end{aligned}$$

where to obtain the last equality we have used that $(1+\omega)^{-1/2} = 1+o(1)$ as $\omega \to 0$. ■

B.6 Proof of Lemma 4.8

It suffices to note that since

$$\begin{aligned}\chi^2(P_{Z,W} \,\|\, P_Z P_W) &= \mathbb{E}_{P_Z}\Big[\chi^2(P_{W|Z}(\cdot|Z) \,\|\, P_W)\Big] \\ &= \mathbb{E}_{P_W}\Big[\chi^2(P_{Z|W}(\cdot|W) \,\|\, P_Z)\Big],\end{aligned}$$

we have for all $z \in \mathcal{Z}$,

$$\begin{aligned}&\chi^2(P_{W|Z}(\cdot|z) \,\|\, P_W)\,\min_{z'\in\mathcal{Z}} P_Z(z') \\ &\qquad \le \chi^2(P_{Z,W} \,\|\, P_Z P_W) \le \max_{z'\in\mathcal{Z}} \chi^2(P_{W|Z}(\cdot|z') \,\|\, P_W),\end{aligned}$$

and, similarly, for all $w \in \mathcal{W}$,

$$\begin{aligned}&\chi^2(P_{Z|W}(\cdot|w) \,\|\, P_Z)\,\min_{w'\in\mathcal{W}} P_W(w') \\ &\qquad \le \chi^2(P_{Z,W} \,\|\, P_Z P_W) \le \max_{w'\in\mathcal{W}} \chi^2(P_{Z|W}(\cdot|w') \,\|\, P_Z),\end{aligned}$$

where both the constituent minima are finite and nonzero as $\epsilon \to 0$ due to (4.24). ■

B.7 Proof of Lemma 4.9

The “if” part of the lemma follows from using $\mathcal{O}(\epsilon)$-dependence between Z and W in the form (4.25c) with Lemma 4.5 to obtain

$$D(P_{Z|W}(\cdot|w)\|P_Z) = \frac{\epsilon^2}{2}\|\boldsymbol{\phi}_w^{Z|W}\|^2 + o(\epsilon^2), \quad w \in \mathcal{W}, \tag{B.10}$$

where for each $w \in \mathcal{W}$,

$$\phi_w^{Z|W}(z) \triangleq \frac{P_{Z|W}(z|w) - P_Z(z)}{\epsilon\sqrt{P_Z(z)}}, \quad z \in \mathcal{Z}$$

is the information vector associated with $P_{Z|W}(\cdot|w)$, and for which $\|\boldsymbol{\phi}_w^{Z|W}\| \leq 1$. Alternatively, the inequality

$$D(p\|q) \leq \log(1 + \chi^2(p\|q)) \leq \chi^2(p\|q),$$

valid for all finite $\mathcal{Z}$ and $p, q \in \mathcal{P}^{\mathcal{Z}}$, which is derived in, e.g., [93, Theorem 5], is sufficient to obtain this part of the lemma, using $p = P_{Z,W}$ and $q = P_Z P_W$.

To obtain the “only if” part of the lemma, note that for any finite $\mathcal{Z}$ and $p, q \in \mathcal{P}^{\mathcal{Z}}$, we have, using Pinsker's inequality [70],

$$\begin{aligned} \chi^2(p\|q) &\leq \frac{1}{\min_{z\in\mathcal{Z}} q(z)} \sum_{z\in\mathcal{Z}} (p(z) - q(z))^2 \\ &\leq \frac{1}{\min_{z\in\mathcal{Z}} q(z)} \left(\sum_{z\in\mathcal{Z}} |p(z) - q(z)|\right)^2 \\ &\leq \frac{2D(p\|q)}{\min_{z\in\mathcal{Z}} q(z)}. \end{aligned} \tag{B.11}$$

The result then follows setting, again, $p = P_{Z,W}$ and $q = P_Z P_W$, since the minimum in (B.11) is finite and nonzero due to (4.24). ■

B.8 Proof of Lemma 4.10

It suffices to note that since

$$\begin{aligned} I(Z;W) &\triangleq D(P_{Z,W} \parallel P_Z P_W) \\ &= \mathbb{E}_{P_Z}\Big[D(P_{W|Z}(\cdot|Z) \parallel P_W)\Big] \\ &= \mathbb{E}_{P_W}\Big[D(P_{Z|W}(\cdot|W) \parallel P_Z)\Big], \end{aligned}$$

we have for all $z \in \mathcal{Z}$,

$$\begin{aligned} D(P_{W|Z}(\cdot|z) \,\|\, P_W) \min_{z' \in \mathcal{Z}} P_Z(z') \\ \leq I(Z;W) \leq \max_{z' \in \mathcal{Z}} D(P_{W|Z}(\cdot|z') \,\|\, P_W), \end{aligned}$$

and, similarly, for all $w \in \mathcal{W}$,

$$\begin{aligned} D(P_{Z|W}(\cdot|w) \,\|\, P_Z) \min_{w' \in \mathcal{W}} P_W(w') \\ \leq I(Z;W) \leq \max_{w' \in \mathcal{W}} D(P_{Z|W}(\cdot|w') \,\|\, P_Z), \end{aligned}$$

where both the constituent minima are finite and nonzero as $\epsilon \to 0$ due to (4.24). ■

B.9 Proof of Lemma 4.12

Let $Q_i(\ell^k)$ denote the probability of obtaining a given value ℓ^k in (4.31) when the sequence z_1^m from which ℓ^k is formed is generated from P_i, for $i \in \{1,2\}$. Then the optimum error exponent is obtained by a rule that decides based on comparing

$$\frac{1}{m} \log \frac{Q_1(\ell^k)}{Q_2(\ell^k)} \tag{B.12}$$

to a threshold.

For $i \in \{1,2\}$, we have

$$\begin{aligned} \lim_{m\to\infty} \frac{1}{m} \log Q_i(\ell^k) &= \min_{\{\hat{P}\colon \mathbb{E}_{\hat{P}}\left[h^k(Z)\right]=\ell^k\}} D(\hat{P}\|P_i) && \text{(B.13)} \\ &= \min_{\substack{\{\hat{\phi}\colon \epsilon\langle\hat{\phi},\xi_l\rangle=\ell_l, \\ l=1,\dots,k\}}} \frac{\epsilon^2}{2} \|\hat{\phi} - \phi_i\|^2 + o(\epsilon^2) && \text{(B.14)} \\ &= \frac{\epsilon^2}{2} \|\hat{\phi}_* - \phi_i\|^2 + o(\epsilon^2) && \text{(B.15)} \\ &= \frac{1}{2} \sum_{l=1}^{k} \left(\ell_l - \epsilon\langle\phi_i, \xi_l\rangle\right)^2 + o(\epsilon^2), && \text{(B.16)} \end{aligned}$$

where to obtain (B.13) we have used Sanov's Theorem [76], to obtain (B.14) we have used both Lemma 4.5 and that

$$\mathbb{E}_{P_i}[h_l(Z)] = \epsilon\langle\phi_i,\xi_l\rangle, \qquad i=1,2 \text{ and } l=1,\dots,k, \tag{B.17}$$

which follows from Lemma 4.4 since (4.32a) holds, and to obtain (B.15) and, in turn, (B.16), we have used

$$\hat{\phi}_* = \operatorname*{arg\,min}_{\substack{\{\hat{\phi}:\ \epsilon\langle\hat{\phi},\xi_l\rangle=\ell_l,\\ l=1,\dots,k\}}} \|\hat{\phi}-\phi_i\|^2 = \phi_i - \frac{1}{\epsilon}\sum_{l=1}^{k}(\ell_l-\epsilon\langle\phi_i,\xi_l\rangle)\xi_l,$$

the last equality of which is obtained using, e.g., Lagrange multipliers, together with (4.33).

Using (B.16) in (B.12) then yields

$$\begin{aligned}\lim_{m\to\infty}\frac{1}{m}\log\frac{Q_1(\ell^k)}{Q_2(\ell^k)} &= \sum_{l=1}^{k}\ell_l\epsilon\langle\phi_2-\phi_1,\xi_l\rangle + \epsilon^2\sum_{l=1}^{k}\Big(\langle\phi_1,\xi_l\rangle^2-\langle\phi_2,\xi_l\rangle^2\Big)\\ &= \sum_{l=1}^{k}\ell_l\epsilon\langle\phi_2-\phi_1,\xi_l\rangle + o(\epsilon^2), \end{aligned} \tag{B.18}$$

where (B.18) is obtained as follows: via the triangle inequality, $\tilde{P}_0 = (P_1+P_2)/2 \in \mathcal{N}_\epsilon^{\mathcal{Z}}(P_0)$, so according to Lemma 4.6 we have

$$0 = \tilde{\phi}_1(z) - \tilde{\phi}_2(z) = \big(\phi_1(z)-\phi_2(z)\big)\big(1+o(1)\big),$$

whence $\phi_1(z)-\phi_2(z) = o(1)$.

Using (B.18) with (B.17), we obtain that an asymptotically optimal decision rule compares the projection

$$\sum_{l=1}^{k}\ell_l\left(\mathbb{E}_{P_1}[h_l(Z)] - \mathbb{E}_{P_2}[h_l(Z)]\right)$$

to a threshold. Accordingly, via Cramér's Theorem [76] the error exponent under P_i is

$$E_i(\lambda) = \min_{P\in\mathcal{S}(\lambda)} D(P\|P_i), \tag{B.19}$$

where

$$\begin{aligned}\mathcal{S}(\lambda) \triangleq \Big\{P \in \mathcal{P}^{\mathcal{Z}}:&\\ &\sum_{l=1}^{k} \mathbb{E}_P[h_l(Z)]\Big(\mathbb{E}_{P_1}[h_l(Z)] - \mathbb{E}_{P_2}[h_l(Z)]\Big)\\ &\quad = \sum_{l=1}^{k}\Big(\lambda\,\mathbb{E}_{P_1}[h_l(Z)] + (1-\lambda)\,\mathbb{E}_{P_2}[h_l(Z)]\Big)\\ &\qquad\qquad\qquad \cdot \Big(\mathbb{E}_{P_1}[h_l(Z)] - \mathbb{E}_{P_2}[h_l(Z)]\Big)\Big\}.\end{aligned} \tag{B.20}$$

Using (B.17), the constraint (B.20) is expressed in information space as

$$\sum_{l=1}^{k}\langle\phi,\xi_l\rangle\langle\phi_1-\phi_2,\xi_l\rangle = \sum_{l=1}^{k}\langle\lambda\,\phi_1+(1-\lambda)\,\phi_2,\xi_l\rangle\langle\phi_1-\phi_2,\xi_l\rangle,$$

i.e.,

$$\Big\langle \phi - (\lambda\,\phi_1 + (1-\lambda)\,\phi_2), \sum_{l=1}^{k}\langle\phi_1-\phi_2,\xi_l\rangle\,\xi_l\Big\rangle = 0 \tag{B.21}$$

In turn, the optimizing P in (B.19), which we denote by P^*, lies in the exponential family through P_i with natural statistic

$$\sum_{l=1}^{k}\langle\phi_1-\phi_2,\xi_l\rangle\,h_l(z),$$

i.e., the family whose members are of the form

$$\log\tilde{P}_\theta(z) = \theta\sum_{l=1}^{k}\langle\phi_1-\phi_2,\xi_l\rangle\,h_l(z) + \log P_i(z) - \alpha(\theta),$$

for which the associated information vector is

$$\epsilon\,\tilde{\phi}_\theta(z) = \theta\sum_{l'=1}^{k}\langle\phi_1-\phi_2,\xi_{l'}\rangle\,\xi_{l'}(z) + \epsilon\phi_i(z) - \alpha(\theta)\sqrt{P_0(z)} + o(\epsilon),$$

so

$$\epsilon\,\langle\tilde{\phi}_\theta,\xi_l\rangle = \theta\langle\phi_1-\phi_2,\xi_l\rangle + \epsilon\,\langle\phi_i,\xi_l\rangle + o(\epsilon),$$

where we have used (4.33). Hence, via (B.21) we obtain that the intersection with the linear family (B.20) is at $P^* = P_{\theta^*}$ with

$$\theta^* = \epsilon \, \frac{\sum_{l=1}^k \langle \phi_1 - \phi_2, \xi_l \rangle \langle \lambda \, \phi_1 + (1-\lambda) \, \phi_2 - \phi_i, \xi_l \rangle}{\sum_{l=1}^k \langle \phi_1 - \phi_2, \xi_l \rangle^2} + o(\epsilon),$$

and thus

$$\begin{aligned} E_i(\lambda) &= D(P^* \| P_i) \\ &= \frac{1}{2} \left\| \theta^* \sum_{l=1}^k \langle \phi_1 - \phi_2, \xi_l \rangle \, \xi_l \right\|^2 + \frac{1}{2} \alpha(\theta^*)^2 + o(\epsilon^2) \\ &= \frac{(\theta^*)^2}{2} \sum_{l=1}^k \langle \phi_1 - \phi_2, \xi_l \rangle^2 + \frac{1}{2} \alpha(\theta^*)^2 + o(\epsilon^2) \\ &= \frac{\epsilon^2}{2} \frac{\left(\sum_{l=1}^k \langle \phi_1 - \phi_2, \xi_l \rangle \langle \lambda \, \phi_1 + (1-\lambda) \, \phi_2 - \phi_i, \xi_l \rangle \right)^2}{\sum_{l=1}^k \langle \phi_1 - \phi_2, \xi_l \rangle^2} + o(\epsilon^2), \end{aligned} \tag{B.22}$$

where to obtain the penultimate equality we have again exploited (4.33), and where to obtain the last equality we have used that

$$\alpha(\theta^*) = o(\epsilon^2)$$

since $\theta^* = \mathcal{O}(\epsilon)$, $\alpha(0) = 0$, and

$$\nabla \alpha(0) = \mathbb{E}_{P_i} \left[\sum_{l=1}^k \langle \phi_1 - \phi_2, \xi_l \rangle \, h_l(Z) \right] = \epsilon \sum_{l=1}^k \langle \phi_1 - \phi_2, \xi_l \rangle \, \langle \phi_i, \xi_l \rangle = \mathcal{O}(\epsilon).$$

Finally, $E_1(\lambda) = E_2(\lambda)$ when $\lambda = 1/2$, so the overall error probability has exponent (4.35). ■

The version of record is available at: http://dx.doi.org/10.1561/0100000107

C

Appendices for Section 5

C.1 Proof of Lemma 5.4

Via (5.13), we have

$$\begin{aligned}\sum_{z\in\mathcal{Z}} \phi_{w_i}^{Z|W_i}(z)\,\phi_{w_j}^{Z|W_j}(z) \\ &= \frac{1}{\epsilon^2}\Bigg[\sum_{z\in\mathcal{Z}} \frac{P_{Z|W_i}(z|w_i)\,P_{Z|W_j}(z|w_j)}{P_Z(z)} \\ &\qquad - \sum_{z\in\mathcal{Z}} \left(P_{Z|W_i}(z|w_i) + P_{Z|W_j}(z|w_j)\right) \\ &\qquad\qquad + \sum_{z\in\mathcal{Z}} P_Z(z)\Bigg] = 0,\end{aligned}$$

where the first sum within the brackets is 1 since, using the pairwise marginal and conditional independencies,

$$\begin{aligned}\frac{P_{Z|W_i}(z|w_i)\,P_{Z|W_j}(z|w_j)}{P_Z(z)} &= \frac{P_{W_i|Z}(w_i|z)\,P_{W_j|Z}(w_j|z)\,P_Z(z)}{P_{W_i}(w_i)\,P_{W_j}(w_j)} \\ &= \frac{P_{W_i,W_j|Z}(w_i,w_j|z)\,P_Z(z)}{P_{W_i,W_j}(w_i,w_j)}\end{aligned}$$

$$= P_{Z|W_i,W_j}(z|w_i, w_j).$$

■

C.2 Proof of Lemma 5.5

Due to the conditional independence among the W^k,

$$\begin{aligned} P_{Z|W^k}(z|w^k) &= \frac{P_Z(z)}{P_{W^k}(w^k)} \prod_{i=1}^{k} P_{W_i|Z}(w_i|z) \\ &= \frac{P_Z(z)}{\pi(w^k)} \prod_{i=1}^{k} \frac{P_{Z|W_i}(z|w_i)}{P_Z(z)}, \end{aligned} \tag{C.1}$$

with

$$\pi(w^k) = \frac{P_{W^k}(w^k)}{\prod_{i=1}^{k} P_{W_i}(w_i)} = \sum_{z'} P_Z(z') \prod_{i=1}^{k} \frac{P_{Z|W_i}(z'|w_i)}{P_Z(z')}.$$

Moreover,

$$\begin{aligned} P_Z(z) \prod_{i=1}^{k} \frac{P_{Z|W_i}(z|w_i)}{P_Z(z)} &= P_Z(z) \prod_{i=1}^{k} \left(1 + \frac{\epsilon}{\sqrt{P_Z(z)}}\, \phi_{w_i}^{Z|W_i}\right) \\ &= P_Z(z) + \epsilon \sqrt{P_Z(z)} \sum_{i=1}^{k} \phi_{w_i}^{Z|W_i} + o(\epsilon), \end{aligned} \tag{C.2}$$

where to obtain (C.2) we have used Fact 5.6. In turn, summing (C.2) over z we obtain

$$\pi(w^k) = 1 + o(\epsilon), \tag{C.3}$$

where we have used that since $\phi_{w_i}^{Z|W_i} \in \mathcal{J}^{\mathcal{Z}}$,

$$\sum_{z} \sqrt{P_Z(z)}\, \phi_{w_i}^{Z|W_i}(z) = 0.$$

Hence, using (C.2) and (C.3) with (5.14) in (C.1), we obtain (5.15).

■

The version of record is available at: http://dx.doi.org/10.1561/0100000107

C.3 Proof of Lemma 5.8

First, note that the (i,j)th entry of $\mathbf{A}_1^{\mathrm{T}}\mathbf{Z}\,\mathbf{A}_2$ is $\mathbf{a}_{1,i}^{\mathrm{T}}\mathbf{Z}\mathbf{a}_{2,j}$, where $\mathbf{a}_{1,i}$ and $\mathbf{a}_{2,j}$ denote the ith and jth columns of $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively. Hence,

$$\mathbb{E}\Big[\big\|\mathbf{A}_1^{\mathrm{T}}\mathbf{Z}\,\mathbf{A}_2\big\|_{\mathrm{F}}^2\Big] = \mathbb{E}\Big[\sum_{i,j}\big(\mathbf{a}_{1,i}^{\mathrm{T}}\,\mathbf{Z}\,\mathbf{a}_{2,j}\big)^2\Big] = \sum_{i,j}\mathbb{E}\Big[\big(\mathbf{a}_{1,i}^{\mathrm{T}}\mathbf{Z}\,\mathbf{a}_{2,j}\big)^2\Big]. \tag{C.4}$$

Next, with Z_{ij} denoting the (i,j)th element of $\mathbf{Z}$, note that

$$\mathbb{E}\Big[\big(\mathbf{a}_{1,i}^{\mathrm{T}}\mathbf{Z}\,\mathbf{a}_{2,j}\big)^2\Big] = \mathbb{E}\Big[\big(\underbrace{\mathbf{a}_{1,i}^{\mathrm{T}}\mathbf{Q}_{1,i}^{\mathrm{T}}}_{\triangleq\tilde{\mathbf{a}}_{1,i}^{\mathrm{T}}}\ \mathbf{Z}\ \underbrace{\mathbf{Q}_{2,j}\,\mathbf{a}_{2,j}}_{\triangleq\tilde{\mathbf{a}}_{2,j}}\big)^2\Big] \tag{C.5}$$

$$= \big\|\mathbf{a}_{1,i}\big\|^2\,\big\|\mathbf{a}_{2,j}\big\|^2\,\mathbb{E}\Big[\big(\underbrace{\mathbf{e}_1^{\mathrm{T}}\mathbf{Z}\,\mathbf{e}_1}_{=Z_{11}}\big)^2\Big], \tag{C.6}$$

where to obtain (C.5) we have used Definition 5.7 with orthogonal matrices $\mathbf{Q}_{1,i}$ and $\mathbf{Q}_{2,j}$, and to obtain (C.6) we have chosen $\mathbf{Q}_{1,i}$ and $\mathbf{Q}_{2,j}$ so that

$$\tilde{\mathbf{a}}_{1,i} = \big\|\mathbf{a}_{1,i}\big\|\,\mathbf{e}_1 \qquad \text{and} \qquad \tilde{\mathbf{a}}_{2,j} = \big\|\mathbf{a}_{2,j}\big\|\,\mathbf{e}_1.$$

In turn, substituting (C.6) into (C.4) yields

$$\begin{aligned}\mathbb{E}\Big[\big\|\mathbf{A}_1^{\mathrm{T}}\mathbf{Z}\,\mathbf{A}_2\big\|_{\mathrm{F}}^2\Big] &= \mathbb{E}\Big[Z_{11}^2\Big]\,\sum_{i,j}\big\|\mathbf{a}_{1,i}\big\|^2\,\big\|\mathbf{a}_{2,j}\big\|^2\\ &= \mathbb{E}\Big[Z_{11}^2\Big]\,\sum_{i}\big\|\mathbf{a}_{1,i}\big\|^2\,\sum_{j}\big\|\mathbf{a}_{2,j}\big\|^2\\ &= \mathbb{E}\Big[Z_{11}^2\Big]\,\big\|\mathbf{A}_1\big\|_{\mathrm{F}}^2\,\big\|\mathbf{A}_2\big\|_{\mathrm{F}}^2. \end{aligned}\tag{C.7}$$

Finally, with $\tilde{\mathbf{Q}}_l$ denoting the permutation matrix that interchanges the first and lth columns of the identity matrix, it follows from Definition 5.7 with $\mathbf{Q}_1 = \tilde{\mathbf{Q}}_i$ and $\mathbf{Q}_2 = \tilde{\mathbf{Q}}_j$ that $Z_{ij}\overset{\mathrm{d}}{=}Z_{11}$, and thus

$$\mathbb{E}\Big[\big\|\mathbf{Z}\big\|_{\mathrm{F}}^2\Big] = k_1k_2\,\mathbb{E}\Big[Z_{11}^2\Big],$$

which when used in conjunction with (C.7) yields (5.28). ■

C.4 Proof of Proposition 5.10

Without loss of generality, as discussed in Section 4.4 we assume that f^k and g^k are normalized according to (3.6c) and (3.6d), so the associated feature vectors $\boldsymbol{\Xi}^X$ and $\boldsymbol{\Xi}^Y$, respectively, satisfy (3.13).

We first analyze the error probability in decisions about the value of V based on S^k. To begin, we have

$$\begin{aligned}\bar{E}^{V|S}(f^k) &= \lim_{m\to\infty} \frac{-\mathbb{E}_{\mathrm{RIE}}\left[\log p_{\mathrm{e}}^{V|S}(\mathcal{C}_\epsilon^Y(P_Y), f^k)\right]}{m} \\ &= \mathbb{E}_{\mathrm{RIE}}\left[\lim_{m\to\infty} \frac{-\log p_{\mathrm{e}}^{V|S}(\mathcal{C}_\epsilon^Y(P_Y), f^k, v_*, v_*')}{m}\right] \qquad (\mathrm{C.8}) \\ &= \mathbb{E}_{\mathrm{RIE}}\left[E^{V|S}(\mathcal{C}_\epsilon^Y(P_Y), f^k, v_*, v_*')\right], \qquad (\mathrm{C.9})\end{aligned}$$

where to obtain the (C.8) we have used standard pairwise exponent analysis. Specifically, (with slight abuse of notation) $p_{\mathrm{e}}^{V|S}(\mathcal{C}_\epsilon^Y(P_Y), f^k, v, v')$ denotes the pairwise error probability distinguishing distinct v and v' in $\mathcal{V}$ based on s^k, and

$$(v_*, v_*') = \operatorname*{arg\,min}_{\{v,v'\in\mathcal{V}:\ v\neq v'\}} p_{\mathrm{e}}^{V|S}(\mathcal{C}_\epsilon^Y(P_Y), f^k, v, v'), \qquad (\mathrm{C.10})$$

whose dependence on f^k and $\mathcal{C}_\epsilon^X(P_X)$ we leave implicit in our notation. Finally, in (C.9) we have used the notation

$$E^{V|S}(\mathcal{C}_\epsilon^Y(P_Y), f^k, v, v') \triangleq \lim_{m\to\infty} \frac{-\log p_{\mathrm{e}}^{V|S}(\mathcal{C}_\epsilon^Y(P_Y), f^k, v, v')}{m}$$

for any distinct v and v'.

Now

$$\begin{aligned}&E^{V|S}(\mathcal{C}_\epsilon^Y(P_Y), f^k, v, v') \\ &\quad= \frac{\epsilon^2}{8}\sum_{i=1}^{k}\left((\boldsymbol{\phi}_v^{X|V} - \boldsymbol{\phi}_{v'}^{X|V})^{\mathrm{T}}\boldsymbol{\xi}_i^X\right)^2 + o(\epsilon^2) \qquad (\mathrm{C.11}) \\ &\quad= \frac{\epsilon^2}{8}\left\|(\boldsymbol{\Xi}^X)^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}(\boldsymbol{\phi}_v^{Y|V} - \boldsymbol{\phi}_{v'}^{Y|V})\right\|^2 + o(\epsilon^2) \qquad (\mathrm{C.12}) \\ &\quad= \frac{\epsilon^2}{8}\left\|(\boldsymbol{\Xi}^X)^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\boldsymbol{\Phi}^{Y|V}(\mathbf{e}_v - \mathbf{e}_{v'})\right\|^2 + o(\epsilon^2), \qquad (\mathrm{C.13})\end{aligned}$$

where to obtain (C.11) we have used Lemma 4.12 with $\phi_v^{Y|V}$ and $\phi_{v'}^{Y|V}$ as defined in (5.23a), to obtain (C.12) we have used (5.24), and in (C.13) we have exploited elementary vector notation (with an abuse of notation as discussed in footnote 5). Moreover, for fixed v and v',

$$\begin{aligned}
&\mathbb{E}_{\mathrm{RIE}}\left[E^{V|S}(\mathcal{C}_\epsilon^Y(P_Y), f^k, v, v')\right] \\
&\quad= \mathbb{E}_{\mathrm{RIE}}\left[\frac{\epsilon^2}{8}\left\|(\boldsymbol{\Xi}^X)^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\boldsymbol{\Phi}^{Y|V}(\mathbf{e}_v - \mathbf{e}_{v'})\right\|^2\right] + o(\epsilon^2) \\
&\quad= \frac{\epsilon^2\,\mathbb{E}_{\mathrm{RIE}}\left[\left\|\boldsymbol{\Phi}^{Y|V}\right\|_{\mathrm{F}}^2\right]}{4\,|\mathcal{Y}|\,|\mathcal{V}|}\left\|\mathbf{B}\,\boldsymbol{\Xi}^X\right\|_{\mathrm{F}}^2 + o(\epsilon^2), \qquad \text{(C.14)}
\end{aligned}$$

where to obtain (C.14) we have used Lemma 5.8.

Then since (C.14) does not depend on v or v', it follows from the law of total expectation that (C.9) satisfies

$$\begin{aligned}
&\mathbb{E}_{\mathrm{RIE}}\left[E^{V|S}(\mathcal{C}_\epsilon^Y(P_Y), f^k, v_*, v_*')\right] \\
&\quad= \epsilon^2\underbrace{\frac{\mathbb{E}_{\mathrm{RIE}}\left[\left\|\boldsymbol{\Phi}^{Y|V}\right\|_{\mathrm{F}}^2\right]}{4\,|\mathcal{Y}|\,|\mathcal{V}|}}_{\triangleq \bar{E}_0^{Y|V}}\left\|\mathbf{B}\,\boldsymbol{\Xi}^X\right\|_{\mathrm{F}}^2 + o(\epsilon^2) \qquad \text{(C.15)} \\
&\quad\le \bar{E}_0^{Y|V}\,\epsilon^2\sum_{i=1}^{k}\sigma_i^2 + o(\epsilon^2), \qquad \text{(C.16)}
\end{aligned}$$

where to obtain (C.16) we have used Lemma 3.1 with the relevant constraint in (3.13) induced by our choice of normalization (3.6c). Moreover, the inequality in (C.16) holds with equality when we choose $\boldsymbol{\Xi}^X$ according to (3.14a), i.e., the optimal features are $f^k = f_*^k$.

We analyze the error probability in decisions about V based on T^k similarly. In particular, we obtain

$$\begin{aligned}
\bar{E}^{V|T}(g^k) &= \lim_{m\to\infty}\frac{-\mathbb{E}_{\mathrm{RIE}}\left[\log p_{\mathrm{e}}^{V|T}(\mathcal{C}_\epsilon^Y(P_Y), g^k)\right]}{m} \\
&= \frac{\epsilon^2\,\mathbb{E}_{\mathrm{RIE}}\left[\left\|\boldsymbol{\Phi}^{Y|V}\right\|_{\mathrm{F}}^2\right]}{4\,|\mathcal{Y}|\,|\mathcal{V}|}\left\|\boldsymbol{\Xi}^Y\right\|_{\mathrm{F}}^2 + o(\epsilon^2) \qquad \text{(C.17)} \\
&= \bar{E}_0^{Y|V}\,\epsilon^2\,k + o(\epsilon^2), \qquad \text{(C.18)}
\end{aligned}$$

for any $\boldsymbol{\Xi}^Y$ satisfying (3.13), i.e., any choice of (normalized) g^k.

We obtain the error probability in decisions about the value of U from symmetry considerations. In particular, it suffices to interchange the roles of U and V, and X and Y—so B is replaced with its adjoint—in the preceding analysis, which yields that

$$\bar{E}^{U|T}(g^k) \le \underbrace{\frac{\mathbb{E}_{\mathrm{RIE}}\Big[\big\|\boldsymbol{\Phi}^{X|U}\big\|_{\mathrm{F}}^2\Big]}{4\,|\mathcal{X}|\,|\mathcal{U}|}}_{\triangleq \bar{E}_0^{X|U}}\;\epsilon^2 \sum_{i=1}^{k} \sigma_i^2 + o(\epsilon^2),$$

with equality when $g^k = g_*^k$, and

$$\bar{E}^{U|S}(f^k) = \bar{E}_0^{X|U}\,\epsilon^2\,k + o(\epsilon^2),$$

for any choice of (normalized) f^k.

It follows that the inequalities in (5.32) simultaneously all hold with equality for the choices $f^k = f_*^k$ and $g^k = g_*^k$. ■

C.5 Proof of Proposition 5.13

Without loss of generality, as discussed in Section 4.4 we assume that f^k and g^k are normalized according to (3.6c) and (3.6d), so the associated feature vectors $\boldsymbol{\Xi}^X$ and $\boldsymbol{\Xi}^Y$, respectively, satisfy (3.13).

We first analyze the error probability in decisions about the value of V^k based on S^k. To begin, the error exponent in decisions about V_i satisfies

$$\begin{align}
E^{V_i|S}(\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y), f^k) &\le E^{V_i|S}(\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y), f^k, v_i, v_i') \tag{C.19}\\
&= \frac{\epsilon^2}{8}\,\big\|(\boldsymbol{\Xi}^X)^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}(\boldsymbol{\phi}_{v_i}^{Y|V_i} - \boldsymbol{\phi}_{v_i'}^{Y|V_i})\big\|^2 + o(\epsilon^2) \tag{C.20}\\
&\le \frac{\epsilon^2}{2}\max_{v_i\in\mathcal{V}_i}\big\|(\boldsymbol{\Xi}^X)^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\boldsymbol{\phi}_{v_i}^{Y|V_i}\big\|^2 + o(\epsilon^2) \tag{C.21}\\
&\le \frac{\epsilon^2}{2}\,\big\|\boldsymbol{\Xi}^X\big\|_{\mathrm{s}}^2\max_{v_i\in\mathcal{V}_i}\big\|\mathbf{B}^{\mathrm{T}}\boldsymbol{\phi}_{v_i}^{Y|V_i}\big\|^2 + o(\epsilon^2) \tag{C.22}\\
&\le \frac{\epsilon^2}{2}\max_{v_i\in\mathcal{V}_i}\big\|\mathbf{B}^{\mathrm{T}}\boldsymbol{\phi}_{v_i}^{Y|V_i}\big\|^2 + o(\epsilon^2) \tag{C.23}\\
&= \frac{\epsilon^2}{2}\max_{v_i\in\mathcal{V}_i}\big\|\mathbf{B}^{\mathrm{T}}\tilde{\boldsymbol{\phi}}_{v_i}^{Y|V_i}\big\|^2(\delta_{v_i}^{Y|V_i})^2 + o(\epsilon^2) \tag{C.24}
\end{align}$$

$$\leq \frac{\epsilon^2}{2} \max_{v_i \in \mathcal{V}_i} \left\| \mathbf{B}^{\mathrm{T}} \tilde{\boldsymbol{\phi}}_{v_i}^{Y|V_i} \right\|^2 + o(\epsilon^2) \tag{C.25}$$

$$= \frac{\epsilon^2}{2} \left\| \mathbf{B}^{\mathrm{T}} \tilde{\boldsymbol{\phi}}_{*}^{Y|V_i} \right\|^2 + o(\epsilon^2), \tag{C.26}$$

where (C.19) follows from standard pairwise error analysis with the pairwise error exponent in distinguishing distinct v_i and v_i' in $\mathcal{V}_i$ denoted using $E^{V_i|S}(\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y), f^k, v_i, v_i')$, to obtain (C.20) we have adapted (C.12) in the proof of Proposition 5.10, to obtain (C.21) we have used the triangle inequality, to obtain (C.22) we have used Fact 5.12, to obtain (C.23) we have used that $\left\|\boldsymbol{\Xi}^X\right\|_{\mathrm{s}} = \left\|\boldsymbol{\Xi}^Y\right\|_{\mathrm{s}} = 1$, and to obtain (C.24) and (C.25) we have used the decomposition

$$\boldsymbol{\phi}_{v_i}^{Y|V_i} = \tilde{\boldsymbol{\phi}}_{v_i}^{Y|V_i}\, \delta_{v_i}^{Y|V_i},$$

where $\left\|\tilde{\boldsymbol{\phi}}_{v_i}^{Y|V_i}\right\| = 1$ and $\left|\delta_{v_i}^{Y|V_i}\right| \leq 1$. In (C.26), we have introduced the notation

$$\tilde{\boldsymbol{\phi}}_{*}^{Y|V_i} \triangleq \operatorname*{arg\,max}_{\left\{\tilde{\boldsymbol{\phi}}_{v_i}^{Y|V_i} \colon v_i \in \mathcal{V}_i\right\}} \left\| \mathbf{B}^{\mathrm{T}} \tilde{\boldsymbol{\phi}}_{v_i}^{Y|V_i} \right\|^2,$$

and note that since V^k is a multi-attribute, by Lemma 5.4 the matrix

$$\tilde{\boldsymbol{\Phi}}_{*}^{Y|V^k} \triangleq \begin{bmatrix} \tilde{\boldsymbol{\phi}}_{*}^{Y|V_1} & \cdots & \tilde{\boldsymbol{\phi}}_{*}^{Y|V_k} \end{bmatrix} \tag{C.27a}$$

has orthogonal columns, so

$$\left(\tilde{\boldsymbol{\Phi}}_{*}^{Y|V^k}\right)^{\mathrm{T}} \tilde{\boldsymbol{\Phi}}_{*}^{Y|V^k} = \mathbf{I}. \tag{C.27b}$$

Hence, for each $i \in \{1, \dots, k\}$,

$$\max_{\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y), f^k} \min_{j \leq i} E^{V_j|S}(\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y), f^k)$$

$$\leq \max_{\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y), f^k} \min_{j \leq i} \frac{\epsilon^2}{2} \left\| \mathbf{B}^{\mathrm{T}} \tilde{\boldsymbol{\phi}}_{\mathbf{B}}^{Y|V_j} \right\|^2 + o(\epsilon^2) \tag{C.28}$$

$$= \max_{\substack{\tilde{\boldsymbol{\Phi}}_{\mathbf{B}}^{Y|V^i} \colon \\ (\tilde{\boldsymbol{\Phi}}_{\mathbf{B}}^{Y|V^i})^{\mathrm{T}} \tilde{\boldsymbol{\Phi}}_{\mathbf{B}}^{Y|V^i} = \mathbf{I}}} \min_{j \leq i} \frac{\epsilon^2}{2} \left\| \mathbf{B}^{\mathrm{T}} \tilde{\boldsymbol{\phi}}_{\mathbf{B}}^{Y|V_j} \right\|^2 + o(\epsilon^2) \tag{C.29}$$

$$= \max_{\mathcal{S} \subset \mathbb{R}^{K_Y} \colon \dim(\mathcal{S}) = i} \min_{\tilde{\boldsymbol{\phi}} \in \mathcal{S} \colon \|\tilde{\boldsymbol{\phi}}\| = 1} \frac{\epsilon^2}{2} \left\| \mathbf{B}^{\mathrm{T}} \tilde{\boldsymbol{\phi}} \right\|^2 + o(\epsilon^2) \tag{C.30}$$

$$= \frac{\epsilon^2}{2}\,\sigma_i^2 + o(\epsilon^2), \tag{C.31}$$

where to obtain (C.28) we have used (C.26), to obtain (C.29) we have used (C.27), to obtain (C.30) we have used the definition of a subspace, and to obtain (C.31) we have used Lemma 5.11.

We further note that the inequalities leading to the right-hand side of (C.31) hold with equality for all $i \in \{1, \dots, k\}$ when we choose

$$\mathcal{V}_i = \{+1, -1\} \quad \text{and} \quad \boldsymbol{\phi}_{+1}^{Y|V_i} = -\boldsymbol{\phi}_{-1}^{Y|V_i} = \boldsymbol{\psi}_i^Y,$$

for $i = 1, \dots, k$ (so $P_{V_i}(+1) = P_{V_i}(-1) = 1/2$), and

$$\boldsymbol{\Xi}^X = \boldsymbol{\Psi}_{(k)}^X,$$

with $\boldsymbol{\Psi}_{(k)}^X$ as defined in (3.15). In particular, the equalities leading to (C.26) all hold with equality with these choices so (C.28) holds with equality, and, via Lemma 5.11, (C.31) holds when $\mathcal{S}$ is the space spanned by the columns of $\boldsymbol{\Psi}_{(i)}^Y$ and $\tilde{\boldsymbol{\phi}} = \boldsymbol{\psi}_i^Y$.

We similarly analyze the error probability in decisions about U^k based on S^k. In particular, the error exponent in decisions about U_i satisfies

$$E^{U_i|S}(\mathcal{C}_\epsilon^{\mathcal{X},k}(P_X), f^k) \leq \frac{\epsilon^2}{2}\left\|\tilde{\boldsymbol{\phi}}_*^{X|U_i}\right\|^2 + o(\epsilon^2), \tag{C.32}$$

where we have used the decomposition

$$\boldsymbol{\phi}_{u_i}^{X|U_i} = \tilde{\boldsymbol{\phi}}_{u_i}^{X|U_i}\,\delta_{u_i}^{X|U_i}$$

with $\|\tilde{\boldsymbol{\phi}}_{u_i}^{X|U_i}\| = 1$ and $|\delta_{u_i}^{X|U_i}| \leq 1$, and where

$$\tilde{\boldsymbol{\phi}}_*^{X|U_i} \triangleq \operatorname*{arg\,max}_{\left\{\tilde{\boldsymbol{\phi}}_{u_i}^{X|U_i} :\, u_i \in \mathcal{U}_i\right\}} \left\|\tilde{\boldsymbol{\phi}}_{u_i}^{X|U_i}\right\|^2.$$

Analogously, we note that since U^k is a multi-attribute, by Lemma 5.4 the matrix

$$\tilde{\boldsymbol{\Phi}}_*^{X|U^k} \triangleq \begin{bmatrix} \tilde{\boldsymbol{\phi}}_*^{X|U_1} & \cdots & \tilde{\boldsymbol{\phi}}_*^{X|U_k} \end{bmatrix} \tag{C.33a}$$

has orthogonal columns, so

$$(\tilde{\boldsymbol{\Phi}}_*^{X|U^k})^{\mathrm{T}}\tilde{\boldsymbol{\Phi}}_*^{X|U^k} = \mathbf{I}. \tag{C.33b}$$

Hence, for each $i \in \{1, \dots, k\}$,

$$\max_{\mathcal{C}_\epsilon^{\mathcal{X},k}(P_X), f^k} \; \min_{j \leq i} \; E^{U_j|S}(\mathcal{C}_\epsilon^{\mathcal{X},k}(P_X), f^k)$$

$$\leq \max_{\mathcal{S} \subset \mathbb{R}^{K_Y} :\; \dim(\mathcal{S}) = i} \; \min_{\tilde{\boldsymbol{\phi}} \in \mathcal{S} :\; \|\tilde{\boldsymbol{\phi}}\| = 1} \; \frac{\epsilon^2}{2} \|\tilde{\boldsymbol{\phi}}\|^2 + o(\epsilon^2) \tag{C.34}$$

$$= \frac{\epsilon^2}{2} + o(\epsilon^2). \tag{C.35}$$

In this case, it is straightforward to verify that the corresponding inequalities leading to (C.35)—and so to (C.32) as well—all hold with equality when

$$\mathcal{U}_i = \{+1, -1\} \quad \text{and} \quad \boldsymbol{\phi}_{+1}^{X|U_i} = -\boldsymbol{\phi}_{-1}^{X|U_i} = \tilde{\boldsymbol{\phi}}_*^{X|U_i},$$

for any $\tilde{\boldsymbol{\Phi}}_*^{X|U^k}$ satisfying (C.33b), and when $\boldsymbol{\Xi}^X = \tilde{\boldsymbol{\Phi}}_*^{X|U^k}$.

The associated error probabilities for decisions about U^k and V^k based on T^k follow from symmetry considerations. In particular, it suffices to interchange the roles of U and V, and X and Y—so B is replaced with its adjoint—in the preceding analysis. This immediately yields that for $i \in \{1, \dots, k\}$,

$$\max_{\mathcal{C}_\epsilon^{\mathcal{X},k}(P_X), g^k} \; \min_{j \leq i} \; E^{U_j|T}(\mathcal{C}_\epsilon^{\mathcal{X},k}(P_X), g^k) = \frac{\epsilon^2}{2} \sigma_i^2 + o(\epsilon^2), \tag{C.36}$$

which is achieved by the choices

$$\mathcal{U}_i = \{+1, -1\} \quad \text{and} \quad \boldsymbol{\phi}_{+1}^{X|U_i} = -\boldsymbol{\phi}_{-1}^{X|U_i} = \boldsymbol{\psi}_i^X,$$

for $i = 1, \dots, k$ (so $P_{U_i}(+1) = P_{U_i}(-1) = 1/2$), and

$$\boldsymbol{\Xi}^Y = \boldsymbol{\Psi}_{(k)}^Y,$$

with $\boldsymbol{\Psi}_{(k)}^Y$ as defined in (3.15).

And it likewise yields, also for $i \in \{1, \dots, k\}$, that

$$\max_{\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y), g^k} \; \min_{j \leq i} \; E^{V_j|T}(\mathcal{C}_\epsilon^{\mathcal{Y},k}(P_Y), g^k) = \frac{\epsilon^2}{2} + o(\epsilon^2), \tag{C.37}$$

which is achieved by the choices

$$\mathcal{V}_i = \{+1, -1\} \quad \text{and} \quad \boldsymbol{\phi}_{+1}^{Y|V_i} = -\boldsymbol{\phi}_{-1}^{Y|V_i} = \tilde{\boldsymbol{\phi}}_{*}^{Y|V_i},$$

for any $\tilde{\boldsymbol{\Phi}}_{*}^{Y|V^k}$ satisfying (C.27b), and when $\boldsymbol{\Xi}^Y = \tilde{\boldsymbol{\Phi}}_{*}^{Y|V^k}$.

It follows that the inequalities in (5.36) simultaneously all hold with equality for the choices $f^k = f_*^k$ and $g^k = g_*^k$. and (5.37). ■

C.6 Proof of Corollary 5.15

First, note that

$$\begin{aligned}
P_{U^k|X^m,Y^m}(u^k|x^m, y^m) &= P_{U^k|X^m}(u^k|x^m) \qquad &(\text{C.38})\\
&= \prod_{i=1}^{k} P_{U_i|X^m}(u_i|x^m) \qquad &(\text{C.39})\\
&= \prod_{i=1}^{k} P_{U_i}(u_i)\, \frac{P_{X^m|U_i}(x^m|u_i)}{P_{X^m}(x^m)}\\
&= \prod_{i=1}^{k} P_{U_i}(u_i) \prod_{j=1}^{m} \frac{P_{X|U_i}(x_j|u_i)}{P_X(x_j)} \qquad &(\text{C.40})\\
&= \left(\frac{1}{2}\right)^k \prod_{i=1}^{k}\prod_{j=1}^{m}(1 + \epsilon u_i\, f_i^*(x_j)) \qquad &(\text{C.41})\\
&= \left(\frac{1}{2}\right)^k \left(1 + \epsilon \sum_{i=1}^{k} u_i \sum_{j=1}^{m} f_i^*(x_j)\right) + o(\epsilon) \qquad &(\text{C.42})\\
&= \left(\frac{1}{2}\right)^k \left(1 + \epsilon m \sum_{i=1}^{k} u_i\, s_i *\right) + o(\epsilon), \qquad &(\text{C.43})
\end{aligned}$$

where to obtain (C.38) we have used the Markov structure (5.8a), to obtain (C.39) we have used that U^k is a multi-attribute of X^m, to obtain (C.40) we have used (5.8b) and (5.8d) with (5.18b), to obtain (C.41) we have used (5.37a), and to obtain (C.42) we have used Fact 5.6.

Next, from symmetry considerations, we obtain the analogous result

$$P_{V^k|X^m,Y^m}(v^k|x^m, y^m) = \left(\frac{1}{2}\right)^k \left(1 + \epsilon m \sum_{i=1}^{k} v_i\, t_i^*\right) + o(\epsilon), \tag{C.44}$$

We then obtain (5.41) by substituting (C.43) and (C.44) into

$$P_{U^k,V^k|X^m,Y^m}(u^k, v^k|x^m, y^m) \\ = P_{U^k|X^m,Y^m}(u^k|x^m, y^m)\, P_{V^k|X^m,Y^m}(v^k|x^m, y^m),$$

which is a consequence of the Markov structure (5.8a).

Finally, using the preceding results we have

$$\begin{aligned} P_{U^k|S_*^k,T_*^k,V^k}(u^k|s_*^k, t_*^k, v^k) &= \frac{P_{U^k,V^k|S_*^k,T_*^k}(u^k, v^k|s_*^k, t_*^k)}{P_{V^k|S_*^k,T_*^k}(v^k|s_*^k, t_*^k)} \\ &= \frac{P_{U^k,V^k|X^m,Y^m}(u^k, v^k|x^m, y^m) + o(\epsilon)}{P_{V^k|Y^m}(v^k|y^m) + o(\epsilon)} \\ &= \frac{P_{U^k|X^m}(u^k|x^m)\, P_{V^k|Y^m}(v^k|y^m) + o(\epsilon)}{P_{V^k|Y^m}(v^k|y^m) + o(\epsilon)} \\ &= P_{U^k|X^m}(u^k|x^m) + o(\epsilon), \end{aligned}$$

which combined with (C.43) verifies (5.43a), and (5.43b) follows from symmetry considerations. ■

C.7 Proof of Proposition 5.16

The following lemma is useful in our proof.

Lemma C.1. Given $\epsilon > 0$ and configurations $\mathcal{C}_\epsilon^{\mathcal{X}}(P_X)$ and $\mathcal{C}_\epsilon^{\mathcal{Y}}(P_Y)$ for ϵ-attributes U and V, respectively, we have

$$\frac{P_{U,V}(u, v)}{P_U(u)\, P_V(v)} = 1 + \epsilon^2\, \tilde{\sigma}(u, v), \tag{C.45}$$

and

$$I(U; V) = \frac{\epsilon^4}{2} \sum_{u\in\mathcal{U}, v\in\mathcal{V}} P_U(u)\, P_V(v)\, \tilde{\sigma}(u, v)^2 + o(\epsilon^4), \tag{C.46}$$

where

$$\tilde{\sigma}(u, v) \triangleq (\boldsymbol{\phi}_v^{Y|V})^{\mathrm{T}} \mathbf{B}\, \boldsymbol{\phi}_u^{X|U}. \tag{C.47}$$

Proof of Lemma C.1. First, we obtain (C.45) via

$$\begin{aligned}
\frac{P_{U,V}(u,v)}{P_U(u)\,P_V(v)} &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \frac{P_{Y|V}(y|v)\,P_{Y|X}(y|x)\,P_{X|U}(x|u)}{P_Y(y)} \\
&= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \Bigg[P_{Y,X|U}(y,x|u) + P_{X,Y|V}(x,y|v) - P_{X,Y}(x,y) \\
&\qquad\qquad + \frac{P_{Y|V}(y|v) - P_Y(y)}{\sqrt{P_Y(y)}} \\
&\qquad\qquad\qquad \cdot \frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)}\,\sqrt{P_Y(y)}} \\
&\qquad\qquad\qquad\qquad \cdot \frac{P_{X|U}(x|u) - P_X(x)}{\sqrt{P_X(x)}} \Bigg] \\
&= 1 + \epsilon^2\,\tilde{\sigma}(u,v), \qquad\qquad (C.48)
\end{aligned}$$

with $\tilde{\sigma}(u,v)$ as defined in (C.47). In turn, we obtain (C.46) via

$$\begin{aligned}
I(U;V) &= D(P_{U,V}\|P_U P_V) \\
&= \sum_{u\in\mathcal{U},v\in\mathcal{V}} P_{U,V}(u,v)\log\frac{P_{U,V}(u,v)}{P_U(u)\,P_V(v)} \\
&= \sum_{u\in\mathcal{U},v\in\mathcal{V}} P_{U,V}(u,v)\left[\epsilon^2\tilde{\sigma}(u,v) - \frac{\epsilon^4}{2}\tilde{\sigma}(u,v)^2 + o(\epsilon^4)\right] \qquad (C.49) \\
&= \sum_{u\in\mathcal{U},v\in\mathcal{V}} P_U(u)\,P_V(v)\,[1+\epsilon^2\tilde{\sigma}(u,v)] \\
&\qquad\qquad \cdot \left[\epsilon^2\tilde{\sigma}(u,v) - \frac{\epsilon^4}{2}\tilde{\sigma}(u,v)^2 + o(\epsilon^4)\right] \qquad (C.50) \\
&= \frac{\epsilon^4}{2}\sum_{u\in\mathcal{U},v\in\mathcal{V}} P_U(u)\,P_V(v)\,\tilde{\sigma}(u,v)^2 + o(\epsilon^4), \qquad (C.51)
\end{aligned}$$

where to obtain (C.49) we have used (C.45) and the Taylor series expansion $\log(1+\omega) = \omega - \omega^2/2 + o(\omega^2)$, where to obtain (C.50) we have again used (C.45), and where to obtain (C.51) we have used that

$$\sum_{u\in\mathcal{U},v\in\mathcal{V}} P_U(u)\,P_V(v)\,\tilde{\sigma}(u,v) = 0$$

as a consequence of (5.25), since $\tilde{\sigma}(u,v)$ takes the form (C.47). ■

Proceeding to the proof of Proposition 5.16, we have

$$\begin{aligned}
I(U^k;V^k) &= \frac{\epsilon^4}{2} \sum_{u^k,v^k} P_{U^k}(u^k)\, P_{V^k}(v^k)\, \tilde{\sigma}(u^k,v^k)^2 + o(\epsilon^4) && \text{(C.52)}\\
&\leq \frac{\epsilon^4}{2} \max_{u^k,v^k} \tilde{\sigma}(u^k,v^k)^2 + o(\epsilon^4)\\
&= \frac{\epsilon^4}{2} \Big((\boldsymbol{\phi}^{Y|V^k})^{\mathrm{T}} \mathbf{B}\, \boldsymbol{\phi}^{X|U^k} \Big)^2 + o(\epsilon^4) && \text{(C.53)}\\
&= \frac{\epsilon^4}{2} \Big(\sum_{i=1}^{k} \sum_{j=1}^{k} (\boldsymbol{\phi}^{Y|V_i})^{\mathrm{T}} \mathbf{B}\, \boldsymbol{\phi}^{X|U_j} \Big)^2 + o(\epsilon^4) && \text{(C.54)}\\
&= \frac{\epsilon^4}{2} \left\| (\boldsymbol{\Phi}^{Y|V^k})^{\mathrm{T}} \mathbf{B}\, \boldsymbol{\Phi}^{X|U^k} \right\|_{\mathrm{F}}^2 + o(\epsilon^4) && \text{(C.55)}\\
&= \frac{\epsilon^4}{2} \left\| (\tilde{\boldsymbol{\Phi}}^{Y|V^k} \boldsymbol{\Delta}^{Y|V^k})^{\mathrm{T}} \mathbf{B}\, \tilde{\boldsymbol{\Phi}}^{X|U^k} \boldsymbol{\Delta}^{X|U^k} \right\|_{\mathrm{F}}^2 + o(\epsilon^4) && \text{(C.56)}\\
&\leq \frac{\epsilon^4}{2} \left(\max_i \left\| \boldsymbol{\phi}^{Y|V_i} \right\|^2 \right) \left(\max_i \left\| \boldsymbol{\phi}^{X|U_i} \right\|^2 \right)\\
&\qquad\qquad \cdot \left\| \tilde{\boldsymbol{\Phi}}^{Y|V^k} \right\|_{\mathrm{s}}^2 \left\| \mathbf{B}\, \tilde{\boldsymbol{\Phi}}^{X|U^k} \right\|_{\mathrm{F}}^2 + o(\epsilon^4) && \text{(C.57)}\\
&\leq \frac{\epsilon^4}{2} \left\| \mathbf{B}\, \tilde{\boldsymbol{\Phi}}^{X|U^k} \right\|_{\mathrm{F}}^2 + o(\epsilon^4) && \text{(C.58)}\\
&\leq \frac{\epsilon^4}{2} \sum_{i=1}^{k} \sigma_i^2 + o(\epsilon^4), && \text{(C.59)}
\end{aligned}$$

where to obtain (C.52) we have used (C.46) of Lemma C.1 with $U = U^k$ and $V = V^k$ so

$$\tilde{\sigma}(u^k,v^k) \triangleq (\boldsymbol{\phi}^{Y|V^k}_{v^k})^{\mathrm{T}} \mathbf{B}\, \boldsymbol{\phi}^{X|U^k}_{u^k}, \tag{C.60}$$

in (C.53) we have introduced the notation

$$\boldsymbol{\phi}^{X|U^k} \triangleq \boldsymbol{\phi}^{X|U^k}_{u^k_{\max}} \quad \text{and} \quad \boldsymbol{\phi}^{Y|V^k} \triangleq \boldsymbol{\phi}^{Y|V^k}_{v^k_{\max}},$$

where

$$(u^k_{\max}, v^k_{\max}) \triangleq \operatorname*{arg\,max}_{u^k,v^k} \tilde{\sigma}(u^k,v^k)^2.$$

To obtain (C.54) we have used Lemma 5.5 together with the notation

$$\boldsymbol{\phi}^{X|U_i} \triangleq \boldsymbol{\phi}^{X|U_i}_{u_i^{\max}} \quad \text{and} \quad \boldsymbol{\phi}^{Y|V_i} \triangleq \boldsymbol{\phi}^{Y|V_i}_{v_i^{\max}}$$

with

$$\begin{aligned}u^k_{\max} &= (u_1^{\max}, \dots, u_k^{\max})\\ v^k_{\max} &= (v_1^{\max}, \dots, v_k^{\max}),\end{aligned}$$

and in (C.55) we have introduced the notation

$$\begin{aligned}\mathbf{\Phi}^{X|U^k} &\triangleq \begin{bmatrix}\boldsymbol{\phi}^{X|U_1} & \cdots & \boldsymbol{\phi}^{X|U_k}\end{bmatrix}\\ \mathbf{\Phi}^{Y|V^k} &\triangleq \begin{bmatrix}\boldsymbol{\phi}^{Y|V_1} & \cdots & \boldsymbol{\phi}^{Y|V_k}\end{bmatrix}.\end{aligned}$$

To obtain (C.56) have used the factorizations

$$\begin{aligned}\mathbf{\Phi}^{X|U^k} &= \tilde{\mathbf{\Phi}}^{X|U^k} \mathbf{\Delta}^{X|U^k}\\ \mathbf{\Phi}^{Y|V^k} &= \tilde{\mathbf{\Phi}}^{Y|V^k} \mathbf{\Delta}^{Y|V^k},\end{aligned}$$

where, due to Lemma 5.4,

$$(\tilde{\mathbf{\Phi}}^{X|U^k})^{\mathrm{T}} \tilde{\mathbf{\Phi}}^{X|U^k} = (\tilde{\mathbf{\Phi}}^{Y|V^k})^{\mathrm{T}} \tilde{\mathbf{\Phi}}^{Y|V^k} = \mathbf{I},$$

and where $\mathbf{\Delta}^{X|U^k}$ and $\mathbf{\Delta}^{Y|V^k}$ are diagonal matrices, to obtain (C.57) we have repeatedly used Fact 5.12, to obtain (C.58) we have used the properties of the factorization, and to obtain (C.59) we have used Lemma 3.1. Finally, it is straightforward to verify that the inequalities leading to (C.59) all hold with equality when we choose the particular configurations (5.37), i.e., when

$$\boldsymbol{\phi}^{X|U_i}_{u_i} = u_i\, \boldsymbol{\psi}^X_i, \quad u_i \in \{+1, -1\} \tag{C.61a}$$

$$\boldsymbol{\phi}^{Y|V_i}_{v_i} = v_i\, \boldsymbol{\psi}^Y_i, \quad v_i \in \{+1, -1\}. \tag{C.61b}$$

To obtain (5.46), first note that, starting from (C.60),

$$\begin{aligned}\tilde{\sigma}(u^k, v^k) &= (\boldsymbol{\phi}^{Y|V^k}_{v^k})^{\mathrm{T}} \mathbf{B}\, \boldsymbol{\phi}^{X|U^k}_{u^k}\\ &= \left(\sum_{i=1}^k \boldsymbol{\phi}^{Y|V_i}_{v_i}\right)^{\mathrm{T}} \mathbf{B} \left(\sum_{j=1}^k \boldsymbol{\phi}^{X|U_j}_{u_j}\right) + o(1) && \text{(C.62)}\\ &= \left(\sum_{i=1}^k v_i\, \boldsymbol{\psi}^Y_i\right)^{\mathrm{T}} \mathbf{B} \left(\sum_{j=1}^k u_j\, \boldsymbol{\psi}^X_j\right) + o(1) && \text{(C.63)}\end{aligned}$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

$$
\begin{aligned}
&= \sum_{i=1}^{k}\sum_{j=1}^{k} u_j\, v_i\, (\boldsymbol{\psi}_i^Y)^{\mathrm{T}} \mathbf{B}\, \boldsymbol{\psi}_j^X + o(1) \\
&= \sum_{i=1}^{k} u_i\, v_i\, \sigma_i + o(1), \qquad \text{(C.64)}
\end{aligned}
$$

where to obtain (C.62) we have used Lemma 5.5 to obtain (C.63) we have used (C.61), and to obtain (C.64) we have used (2.12a).

Hence, using (C.45) in Lemma C.1 with $U = U^k$ and $V = V^k$ and substituting (C.64), we obtain

$$
\begin{aligned}
P_{U^k,V^k}(u^k, v^k) &= P_{U^k}(u^k)\, P_{V^k}(v^k)\, (1 + \epsilon^2\, \tilde{\sigma}(u^k, v^k)) \\
&= \frac{1}{4^k}\left(1 + \epsilon^2 \sum_{i=1}^{k} u_i\, v_i\, \sigma_i\right) + o(\epsilon^2),
\end{aligned}
$$

viz., (5.46). ■

C.8 Proof of Proposition 5.18

First, we note that, in accordance with the discussion of Section 4.2, the conditions of the proposition imply that U^k has a configuration of the form $\mathcal{C}^{\mathcal{X},}_{\epsilon(1+o(1))}$. Next, we have

$$
\begin{aligned}
I(U^k; Y) &= \frac{1}{2^k} \sum_{u^k} D(P_{Y|U^k}(\cdot|u^k) \| P_Y) && \text{(C.65)} \\
&= \frac{\epsilon^2}{2^{k+1}} \sum_{u^k} \|\boldsymbol{\phi}_{u^k}^{Y|U^k}\|^2 + o(\epsilon^2) && \text{(C.66)} \\
&= \frac{\epsilon^2}{2^{k+1}} \sum_{u^k} \|\mathbf{B}\, \boldsymbol{\phi}_{u^k}^{X|U^k}\|^2 + o(\epsilon^2) && \text{(C.67)} \\
&= \frac{\epsilon^2}{2^{k+1}} \sum_{u^k} \left\| \mathbf{B} \sum_{i=1}^{k} \boldsymbol{\phi}_{u_i}^{X|U_i} \right\|^2 + o(\epsilon^2) && \text{(C.68)} \\
&= \frac{\epsilon^2}{2} \sum_{i=1}^{k} \|\mathbf{B}\, \boldsymbol{\phi}^{X|U_i}\|^2 + o(\epsilon^2) && \text{(C.69)} \\
&= \frac{\epsilon^2}{2} \|\mathbf{B}\, \boldsymbol{\Phi}^{X|U^k}\|_{\mathrm{F}}^2 + o(\epsilon^2)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\epsilon^2}{2} \| \mathbf{B} \, \tilde{\mathbf{\Phi}}^{X|U^k} \, \mathbf{\Delta}^{X|U^k} \|_{\mathrm{F}}^2 + o(\epsilon^2) && \text{(C.70)} \\
&\leq \frac{\epsilon^2}{2} \| \mathbf{B} \, \tilde{\mathbf{\Phi}}^{X|U^k} \|_{\mathrm{F}}^2 \, \| \mathbf{\Delta}^{X|U^k} \|_{\mathrm{s}}^2 + o(\epsilon^2) && \text{(C.71)} \\
&\leq \frac{\epsilon^2}{2} \| \mathbf{B} \, \tilde{\mathbf{\Phi}}^{X|U^k} \|_{\mathrm{F}}^2 + o(\epsilon^2) \\
&\leq \frac{\epsilon^2}{2} \sum_{i=1}^{k} \sigma_i^2 + o(\epsilon^2), && \text{(C.72)}
\end{aligned}
$$

where to obtain (C.65) we have used that all configurations u^k are equiprobable due to condition 1, to obtain (C.66) we have used Lemma 4.5, to obtain (C.67) we have used (5.22) with $U = U^k$, to obtain (C.68) we have used Lemma 5.5, and to obtain (C.69) we have used that constraint 2 implies that

$$
\boldsymbol{\phi}_{+1}^{X|U_i} = -\boldsymbol{\phi}_{-1}^{X|U_i} \triangleq \boldsymbol{\phi}^{X|U_i}, \tag{C.73}
$$

for $i = 1, \ldots, K-1$, since

$$
\sum_{x} P_{U_i}(u) \, \phi_u^{X|U_i} = 0.
$$

To obtain (C.70) we have used the notation

$$
\mathbf{\Phi}^{X|U^k} \triangleq \begin{bmatrix} \boldsymbol{\phi}^{X|U_1} & \cdots & \boldsymbol{\phi}^{X|U_k} \end{bmatrix}
$$

with factorization

$$
\mathbf{\Phi}^{X|U^k} = \tilde{\mathbf{\Phi}}^{X|U^k} \mathbf{\Delta}^{X|U^k}
$$

where, due to Lemma 5.4,

$$
(\tilde{\mathbf{\Phi}}^{X|U^k})^{\mathrm{T}} \tilde{\mathbf{\Phi}}^{X|U^k} = \mathbf{I} \tag{C.74}
$$

and $\mathbf{\Delta}^{X|U^k}$ is a diagonal matrix whose ith diagonal entry is $\|\boldsymbol{\phi}^{X|U_i}\| \leq 1 + o(1)$. To obtain (C.71) we have used Fact 5.12, and to obtain (C.72) we have used Lemma 3.1 with the constraint (C.74). Furthermore, the inequalities leading to (C.72) hold with equality when we choose

$$
\mathbf{\Phi}^{X|U^k} = \mathbf{\Psi}_{(k)}^{X},
$$

with $\mathbf{\Psi}_{(k)}^{X}$ as defined in (3.15), so the optimum configuration is (5.37a).

The corresponding result, including (5.37b), for the maximization of $I(V^k; X)$ subject to $I(V_i; Y) \leq \epsilon^2/2$ and the other corresponding constraints follows immediately from symmetry considerations. ■

C.9 Proof of Proposition 5.21

First, without loss of generality let us choose $\delta(\cdot)$ such that

$$\delta(\epsilon) \geq \epsilon, \tag{C.75}$$

in which case, we have, for all ϵ sufficiently small,

$$\begin{align} P_{X,Y} &\in \bar{\mathcal{N}}_{\epsilon}^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y) \tag{C.76}\\ &\subset \mathcal{N}_{\epsilon}^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y) \tag{C.77}\\ &\subset \mathcal{N}_{\delta(\epsilon)}^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y) \tag{C.78}\\ &\subset \mathcal{N}_{\sqrt{2\delta(\epsilon)}}^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y), \tag{C.79}\end{align}$$

where (C.76) is given, where (C.77) follows from (5.53), where (C.78) follows from (C.75), and where (C.79) holds when $\epsilon \leq 2$.

Next, from (5.64) and (5.63) it follows that for $w \in \mathcal{W}$,

$$\tilde{\phi}_w^{X,Y|W}(x,y) = \check{\phi}_w^{X,Y|W}(x,y) + o(1), \quad x \in \mathcal{X},\ y \in \mathcal{Y}, \tag{C.80}$$

wherein

$$\begin{align} \|\check{\boldsymbol{\phi}}_w^{X,Y|W}\|^2 &= \frac{1}{2}\sum_{x,y}\left(\sqrt{P_Y(y)}\,\phi_w^{X|W}(x) + \sqrt{P_X(x)}\,\phi_w^{Y|W}(y)\right)^2 \nonumber\\ &= \frac{1}{2}\left(\sum_{x,y} P_Y(y)\,\phi_w^{X|W}(x)^2 + \sum_{x,y} P_X(x)\,\phi_w^{Y|W}(y)^2\right) \nonumber\\ &= \frac{1}{2}\left(\|\boldsymbol{\phi}_w^{X|W}\|^2 + \|\boldsymbol{\phi}_w^{Y|W}\|^2\right) \tag{C.81}\\ &\leq 1. \tag{C.82}\end{align}$$

Hence,

$$P_{X,Y|W}(\cdot,\cdot|w) \in \mathcal{N}_{\sqrt{2\delta(\epsilon)}(1+o(1))}^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y), \quad w \in \mathcal{W}. \tag{C.83}$$

Furthermore, due to (C.79) and (C.83), we may apply Lemma 4.6 with $P_1 = P_{X,Y|W}(\cdot,\cdot|w)$, $P_2 = P_{X,Y}$, $P_0 = P_X P_Y$, and $\tilde{P}_0 = P_{X,Y}$ to

$$\phi_w^{X,Y|W}(x,y) \triangleq \frac{P_{X,Y|W}(x,y|w) - P_{X,Y}(x,y)}{\sqrt{2\delta(\epsilon)}\,\sqrt{P_{X,Y}(x,y)}} \tag{C.84}$$

yielding

$$\begin{aligned}\phi_w^{X,Y|W}(x,y) &= \tilde{\phi}_w^{X,Y|W}(x,y) - \frac{\tilde{B}(y,x)}{\sqrt{2\delta(\epsilon)}} + o(1)\\ &= \check{\phi}_w^{X,Y|W}(x,y) + o(1), \qquad \text{(C.85)}\end{aligned}$$

where to obtain (C.85) we have used both (C.80) and that (C.78) implies $|\tilde{B}(y,x)| \leq \delta(\epsilon)$. Combining (C.85) with (C.82), we conclude

$$P_{X,Y|W}(\cdot,\cdot|w) \in \mathcal{N}^{\mathcal{X}\times\mathcal{Y}}_{\sqrt{2\delta(\epsilon)}(1+o(1))}(P_{X,Y}), \quad w \in \mathcal{W}. \qquad \text{(C.86)}$$

As a result, for all $w \in \mathcal{W}$,

$$\begin{aligned}D(P_{X,Y|W}(\cdot,\cdot|w)\|P_{X,Y}) &= \delta(\epsilon)\left\|\boldsymbol{\phi}_w^{X,Y|W}\right\|^2 + o(\delta(\epsilon)) \qquad \text{(C.87)}\\ &= \delta(\epsilon)\left\|\check{\boldsymbol{\phi}}_w^{X,Y|W}\right\|^2 + o(\delta(\epsilon)), \qquad \text{(C.88)}\end{aligned}$$

where to obtain (C.87) we have used the special case of Lemma 4.5, and to obtain (C.88) we have used (C.85). In turn,

$$\begin{aligned}I(W;X,Y) &= \sum_{w\in\mathcal{W}} P_W(w)\, D(P_{X,Y|W}(\cdot,\cdot|w)\|P_{X,Y})\\ &= \delta(\epsilon)\sum_{w\in\mathcal{W}} P_W(w)\left\|\check{\boldsymbol{\phi}}_w^{X,Y|W}\right\|^2 + o(\delta(\epsilon)) \qquad \text{(C.89)}\\ &= \frac{\delta(\epsilon)}{2}\sum_{w\in\mathcal{W}} P_W(w)\Big(\|\boldsymbol{\phi}_w^{X|W}\|^2 + \|\boldsymbol{\phi}_w^{Y|W}\|^2\Big) + o(\delta(\epsilon)), \qquad \text{(C.90)}\end{aligned}$$

where to obtain (C.89) we have used (C.88), and where to obtain (C.90) we have used (C.81).

Hence, we seek to minimize (C.90) subject to the constraint (5.65). To this end, let us define

$$\tilde{\boldsymbol{\Phi}}^{X|W} \triangleq \sqrt{\delta(\epsilon)}\,\boldsymbol{\Phi}^{X|W}\sqrt{\mathbf{P}_W} \qquad \text{(C.91)}$$

$$\tilde{\boldsymbol{\Phi}}^{Y|W} \triangleq \sqrt{\delta(\epsilon)}\,\boldsymbol{\Phi}^{Y|W}\sqrt{\mathbf{P}_W}, \qquad \text{(C.92)}$$

where, consistent with the notation and terminology in Definition 5.9, $\boldsymbol{\Phi}^{X|W}$ is a $|\mathcal{X}| \times |\mathcal{W}|$ matrix whose wth column is $\boldsymbol{\phi}_w^{X|W}$, where $\boldsymbol{\Phi}^{Y|W}$ is a $|\mathcal{Y}| \times |\mathcal{W}|$ matrix whose wth column is $\boldsymbol{\phi}_w^{Y|W}$, and where $\mathbf{P}_W$ is a

$|\mathcal{W}| \times |\mathcal{W}|$ diagonal matrix whose wth diagonal entry is $P_W(w)$. Then we can equivalently express the constraint (5.65) in the form

$$\tilde{\mathbf{B}} = \tilde{\mathbf{\Phi}}^{Y|W} \left(\tilde{\mathbf{\Phi}}^{X|W}\right)^{\mathrm{T}}, \tag{C.93}$$

and the objective function (C.90) as

$$I(W;X,Y) = \frac{1}{2}\left(\|\tilde{\mathbf{\Phi}}^{X|W}\|_{\mathrm{F}}^2 + \|\tilde{\mathbf{\Phi}}^{Y|W}\|_{\mathrm{F}}^2\right) + o(\delta(\epsilon)) \geq \|\tilde{\mathbf{B}}\|_* + o(\delta(\epsilon)), \tag{C.94}$$

where to obtain the inequality we have used Lemma 5.20 with (C.93).

It is straightforward to verify that the inequality in (C.94) holds with equality subject to the constraints in (5.62) when we choose the configuration for W according to

$$\mathcal{W} = \{\pm 1, \dots, \pm(K-1)\} \tag{C.95a}$$

$$\boldsymbol{\phi}_i^{X|W} = -\boldsymbol{\phi}_{-i}^{X|W} = \sqrt{\frac{\sigma_i}{\tilde{\sigma}_i\,\delta(\epsilon)}}\,\boldsymbol{\psi}_i^X, \quad i = 1, \dots, K-1 \tag{C.95b}$$

$$\boldsymbol{\phi}_i^{Y|W} = -\boldsymbol{\phi}_{-i}^{Y|W} = \sqrt{\frac{\sigma_i}{\tilde{\sigma}_i\,\delta(\epsilon)}}\,\boldsymbol{\psi}_i^Y, \quad i = 1, \dots, K-1 \tag{C.95c}$$

$$P_W(i) = P_W(-i) = \frac{1}{2}\tilde{\sigma}_i, \quad i = 1, \dots, K-1, \tag{C.95d}$$

where

$$\tilde{\sigma}_i \triangleq \frac{\sigma_i}{\sum_{i'=1}^{K-1} \sigma_{i'}}. \tag{C.95e}$$

Specifically, with the choices (C.95) we have

$$\|\boldsymbol{\phi}_w^{X|W}\| \leq 1 \quad \text{and} \quad \|\boldsymbol{\phi}_w^{Y|W}\| \leq 1$$

since

$$\sum_{i=1}^{K-1} \sigma_i = \|\tilde{\mathbf{B}}\|_* \leq \epsilon \leq \delta(\epsilon)$$

from $P_{X,Y} \in \bar{\mathcal{N}}_\epsilon^{\mathcal{X}\times\mathcal{Y}}(P_X P_Y)$ as given and the choice (C.75), so the constraints (5.63) are satisfied. Moreover, we satisfy constraints(5.66) by the symmetric construction of our information vector sets. And we satisfy the constraint (5.65) by construction since

$$\delta(\epsilon) \sum_{w\in\mathcal{W}} P_W(w)\,\boldsymbol{\phi}_w^{Y|W} \left(\boldsymbol{\phi}_w^{X|W}\right)^{\mathrm{T}} = \delta(\epsilon) \sum_{i=1}^{K-1} \frac{\sigma_i}{\delta(\epsilon)}\,\boldsymbol{\psi}_i^Y \left(\boldsymbol{\psi}_i^X\right)^{\mathrm{T}} = \tilde{\mathbf{B}},$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

where last equality follows from (2.30). Finally, evaluating the leading term in (C.90) we obtain

$$\frac{\delta(\epsilon)}{2}\sum_{w\in\mathcal{W}} P_W(w)(\|\phi_w^{X|W}\|^2+\|\phi_w^{Y|W}\|^2) = \frac{\delta(\epsilon)}{2}\left(2\sum_{i=1}^{K-1}\frac{\sigma_i}{\delta(\epsilon)}\right) = \sum_{i=1}^{K-1}\sigma_i,$$

so the inequality in (C.94) is achieved with equality.

It remains only to choose $\delta(\cdot)$ satisfying (C.75) and $\lim_{\epsilon\to 0}\delta(\epsilon)=0$. The leading term in (C.90) with the configuration (C.95) does not depend on this choice, so we focus on the $o(\delta(\epsilon))$ term, which is minimized by the choice $\delta(\epsilon)=\epsilon$, yielding (5.72). In turn, (5.73) is obtained by rewriting (C.95) using (5.61a)–(5.61b) and (2.17). ■

C.10 Proof of Corollary 5.22

We have, for the extended model (5.74),

$$\begin{aligned}
&P_{X^m,Y^m|W}(x^m,y^m|w)\\
&\quad= P_{X^m}(x^m)\,P_{Y_m}(y^m)\\
&\qquad\cdot\prod_{l=1}^{m}\Big(1+\operatorname{sgn}(w)\,\|\tilde{\mathbf{B}}\|_*^{1/2}\,f_{|w|}^*(x_l)\Big)\Big(1+\operatorname{sgn}(w)\,\|\tilde{\mathbf{B}}\|_*^{1/2}\,g_{|w|}^*(y_l)\Big) &&\text{(C.96)}\\
&\quad= P_{X^m}(x^m)\,P_{Y_m}(y^m)\Big(1+m\,\operatorname{sgn}(w)\,\|\tilde{\mathbf{B}}\|_*^{1/2}r_{|w|}^*\Big)+o(\sqrt{\epsilon}) &&\text{(C.97)}\\
&\quad= P_{X^m,Y^m}(x^m,y^m)\Big(1+m\,\operatorname{sgn}(w)\,\|\tilde{\mathbf{B}}\|_*^{1/2}r_{|w|}^*\Big)+o(\sqrt{\epsilon}), &&\text{(C.98)}
\end{aligned}$$

where to obtain (C.96) we have used (5.73) and (5.74), to obtain (C.97) we have used that X, Y are sub-ϵ dependent so (5.51) holds, together with Fact 5.6 and (5.78), and to obtain (C.98) we have used that since sub-ϵ dependence implies $|\tilde{B}(y,x)|\le\epsilon$ for all $x\in\mathcal{X}$ and $y\in\mathcal{Y}$,

$$P_{X,Y}(x,y) = P_X(x)\,P_Y(y)+\mathcal{O}(\epsilon) = P_X(x)\,P_Y(y)+o(\sqrt{\epsilon}),$$

whence

$$P_{X^m,Y^m}(x^m,y^m) = P_{X^m}(x^m)\,P_{Y^m}(y^m)+o(\sqrt{\epsilon}).$$

Finally, substituting P_W from (5.73) and using (C.98) in

$$P_{W|X^m,Y^m}(w|x^m,y^m) = \frac{P_{X^m,Y^m|W}(x^m,y^m|w)\,P_W(w)}{P_{X^m,Y^m}(x^m,y^m)}$$

yields (5.77). ■

C.11 Proof of Corollary 5.23

For the first part of the corollary, note that (5.83) and (5.72) together imply (5.82). To show (5.83), we begin by defining

$$\phi_{w_i}^{X,Y|W_i} \triangleq \frac{P_{X,Y|W_i}(x,y|w_i) - P_{X,Y}(x,y)}{\sqrt{2\epsilon}\sqrt{P_{X,Y}(x,y)}}, \quad w_i \in \mathcal{W}_\circ. \tag{C.99}$$

Then, since

$$\begin{aligned} P_{X,Y|W_i}(x,y|+1) &= P_{X,Y|W}(x,y|i) \\ P_{X,Y|W_i}(x,y|-1) &= P_{X,Y|W}(x,y|-i). \end{aligned}$$

we have

$$\phi_{+1}^{X,Y|W_i} = \phi_i^{X,Y|W} = \check{\phi}_i^{X,Y|W} + o(1) \tag{C.100a}$$

$$\phi_{-1}^{X,Y|W_i} = \phi_{-i}^{X,Y|W} = \check{\phi}_{-i}^{X,Y|W} + o(1), \tag{C.100b}$$

where $\phi_w^{X,Y|W}$ and $\check{\phi}_w^{X,Y|W}$ are as defined in (C.84) (setting $\delta(\epsilon) = \epsilon$) and (5.64b), respectively.

Next, note that

$$\begin{aligned} &P_{X,Y|W_i}(x,y|0) \\ &\quad = \frac{1}{P_{W_i}(0)} \sum_{\{j\colon j\neq i\}} (1 - P_{W_j}(0)) P_{X,Y|\{W_j\neq 0\}}(x,y) && \text{(C.101)} \\ &\quad = \frac{1}{(1-\tilde{\sigma}_i)} \sum_{\{j\colon j\neq i\}} \tilde{\sigma}_j \, P_{X,Y|\{W_j\neq 0\}}(x,y) && \text{(C.102)} \\ &\quad = \frac{1}{2(1-\tilde{\sigma}_i)} \sum_{\{j\colon j\neq i\}} \tilde{\sigma}_j \left(P_{X,Y|W}(x,y|j) + P_{X,Y|W}(x,y|-j)\right), && \text{(C.103)} \end{aligned}$$

where to obtain (C.101) we have used that the events $\{W_i \neq 0\}$, $i = 1,\dots,K-1$ form a partition of sample space, to obtain (C.102) we have used (C.95d), and to obtain (C.103) we have used that

$$\begin{aligned}&P_{X,Y|\{W_j\neq 0\}}(x,y)\\&\quad=\frac{P_{X,Y|W_j}(x,y|{+1})\,P_W(j)+P_{X,Y|W_j}(x,y|{-1})\,P_W(-j)}{P_W(j)+P_W(-j)}\\&\quad=\frac{1}{2}\Big(P_{X,Y|W}(x,y|j)+P_{X,Y|W}(x,y|{-j})\Big).\end{aligned}$$

Hence,

$$\begin{aligned}\boldsymbol{\phi}_0^{X,Y|W_i}&=\frac{1}{2(1-\tilde{\sigma}_i)}\sum_{\{j\colon j\neq i\}}\tilde{\sigma}_j\left(\boldsymbol{\phi}_j^{X,Y|W}+\boldsymbol{\phi}_{-j}^{X,Y|W}\right) &\text{(C.104)}\\&=\frac{1}{2(1-\tilde{\sigma}_i)}\sum_{\{j\colon j\neq i\}}\tilde{\sigma}_j\left(\check{\boldsymbol{\phi}}_j^{X,Y|W}+\check{\boldsymbol{\phi}}_{-j}^{X,Y|W}\right)+o(1) &\text{(C.105)}\\&=o(1), &\text{(C.106)}\end{aligned}$$

where to obtain (C.104) we have used (C.103) with (C.84) (setting $\delta(\epsilon)=\epsilon$) and (C.99), to obtain (C.105) we have used (C.100), and to obtain (C.106) we have used (5.64b) with (C.95b)–(C.95c) to conclude that the term in parentheses is zero, since for $w\in\mathcal{W}$,

$$\check{\boldsymbol{\phi}}_w^{X,Y|W}=\operatorname{sgn}(w)\sqrt{\frac{\sigma_{|w|}}{2\tilde{\sigma}_{|w|}\,\epsilon}}\left(\sqrt{P_Y(y)}\,\psi_{|w|}^Y+\sqrt{P_X(x)}\,\psi_{|w|}^X\right).\qquad\text{(C.107)}$$

From (C.100) with (C.82), and from (C.106), we conclude

$$P_{X,Y|W_i}(\cdot,\cdot|w_i)\in\mathcal{N}_{\sqrt{2\epsilon}(1+o(1))}^{\mathcal{X}\times\mathcal{Y}}(P_{X,Y}),\quad w_i\in\mathcal{W}_\circ,$$

whence

$$\begin{aligned}D(P_{X,Y|W_i}(\cdot,\cdot|j)\|P_{X,Y})&=\epsilon\left\|\boldsymbol{\phi}_j^{X,Y|W_i}\right\|^2+o(\epsilon)\\&=\begin{cases}\epsilon\left\|\check{\boldsymbol{\phi}}_i^{X,Y|W}\right\|^2+o(\epsilon) & j=+1\\ \epsilon\left\|\check{\boldsymbol{\phi}}_{-i}^{X,Y|W}\right\|^2+o(\epsilon) & j=-1\\ o(\epsilon) & j=0,\end{cases}\qquad\text{(C.108)}\end{aligned}$$

where to obtain the first equality we have used the special case of Lemma 4.5, and to obtain the second equality we have used (C.100) and (C.106). In turn, we obtain (5.83) via

$$\begin{aligned} I(W_i; X, Y) &= \sum_{j \in \mathcal{W}_\circ} P_{W_i}(j)\, D(P_{X,Y|W_i}(\cdot,\cdot|j) \| P_{X,Y}) \\ &= 2\,\epsilon P_W(i) \left\| \check{\boldsymbol{\phi}}_i^{X,Y|W} \right\|^2 + o(\epsilon) && \text{(C.109)} \\ &= \sigma_i + o(\epsilon), && \text{(C.110)} \end{aligned}$$

where to obtain the (C.109) we have used the first equality in (C.95d), and to obtain (C.110) we have used the second equality in (C.95d) and that

$$\begin{aligned} \left\| \check{\boldsymbol{\phi}}_w^{X,Y|W} \right\|^2 &= \frac{\sigma_{|w|}}{2\tilde{\sigma}_{|w|}\epsilon} \sum_{x,y} \left(\sqrt{P_Y(y)}\, \psi_{|w|}^X(x) + \sqrt{P_X(x)}\, \psi_{|w|}^Y(y) \right)^2 && \text{(C.111)} \\ &= \frac{\sigma_{|w|}}{\tilde{\sigma}_{|w|}\,\epsilon}. && \text{(C.112)} \end{aligned}$$

To obtain (C.111) we have used (C.107), and to obtain (C.112) we have used (5.63).

Turning now to the second part of the corollary, consistent with Definition 5.19,

$$C_\epsilon(U_i, V_i) = \min_{P_{\tilde{W}_i|U_i,V_i}\,:\, U_i \leftrightarrow \tilde{W}_i \leftrightarrow V_i} I(\tilde{W}_i; U_i, V_i). \tag{C.113}$$

by adapting the analysis of Proposition 5.21 we used to obtain $C_\epsilon(X, Y)$. In particular, from (5.47) we have that the counterpart to (2.3) is

$$B_i(u_i, v_i) \triangleq \frac{P_{U_i,V_i}(u_i, v_i)}{\sqrt{P_{U_i}(u_i)}\,\sqrt{P_{V_i}(v_i)}} = \frac{1}{2}(1 + \tilde{\epsilon}^2\,\sigma_i\,u_i\,v_i) + o(\tilde{\epsilon}^2)$$

for $u_i, v_i \in \{-1, +1\}$, and that to (2.8) is

$$\begin{aligned} \mathbf{B}_i &\triangleq \left[\sqrt{\mathbf{P}_{V_i}}\right]^{-1} \mathbf{P}_{V_i,U_i} \left[\sqrt{\mathbf{P}_{U_i}}\right]^{-1} \\ &= \underbrace{\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}}_{\triangleq \boldsymbol{\Psi}^{V_i}} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{\epsilon}^2\,\sigma_i \end{bmatrix} \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \frac{1}{\sqrt{2}}}_{\triangleq \left(\boldsymbol{\Psi}^{U_i}\right)^{\mathrm{T}}} + o(\tilde{\epsilon}^2). \end{aligned}$$

Note too that the counterpart to (2.29), i.e.,

$$\tilde{\mathbf{B}}_i \triangleq \left[\sqrt{\mathbf{P}_{V_i}}\right]^{-1} \left[\mathbf{P}_{V_i,U_i} - \mathbf{P}_{V_i}\,\mathbf{P}_{U_i}\right] \left[\sqrt{\mathbf{P}_{U_i}}\right]^{-1}$$

satisfies

$$\|\tilde{\mathbf{B}}_i\|_* = \tilde{\epsilon}^2\,\sigma_i + o(\tilde{\epsilon}^2) \le \tilde{\epsilon}^2\,\|\tilde{\mathbf{B}}\|_* + o(\tilde{\epsilon}^2) \le \tilde{\epsilon}^2\,\epsilon + o(\tilde{\epsilon}^2),$$

so

$$P_{U_i,V_i} \in \bar{\mathcal{N}}^{\mathcal{U}_i\times\mathcal{V}_i}_{\tilde{\epsilon}^2\epsilon+o(\tilde{\epsilon}^2)}(P_{U_i}P_{V_i}).$$

Hence, the counterpart of (5.72) in Proposition 5.21 for the new variables (U_i, V_i) is

$$C_\epsilon(U_i, V_i) = \tilde{\epsilon}^2\,\sigma_i + o(\tilde{\epsilon}^2\epsilon).$$

as $\epsilon, \tilde{\epsilon} \to 0$, and thus (5.84) follows. ■

C.12 Proof of Corollary 5.24

Since the event $W_i = j$ is the event $W = ji$ for $j \in \{-1, +1\}$, the cases $w_i = \pm 1$ in (5.86) follow immediately from Corollary 5.22. The case $w_i = 0$ in (5.86) is then determined by the constraint that the result is a distribution. ■

C.13 Proof of Corollary 5.25

First note that (5.88b) is readily obtained from (5.44b), exploiting that u_i and v_i are uniquely determined in the cases $z_i = \pm 2$. Second, note that

$$P_{\tilde{W}_i|Z_i,X^m,Y^m}(\tilde{w}_i|z_i,x^m,y^m)$$
$$= \sum_{\substack{\{(u_i,v_i):\\ u_i+v_i=z_i\}}} P_{\tilde{W}_i|U_i,V_i}(\tilde{w}_i|u_i,v_i)\,P_{U_i,V_i|Z_i,X^m,Y^m}(u_i,v_i|z_i,x^m,y^m), \tag{C.114}$$

which is obtained by exploiting (5.87) and the structure in $\tilde{W}_i$ implicit in (5.80). To obtain (5.88a) from (C.114) we use that

$$P_{\tilde{W}_i|U_i,V_i}(\tilde{w}_i|u_i,v_i) = \frac{1}{2}\left(1 + \operatorname{sgn}(\tilde{w}_i\,z_i)\,\sqrt{\sigma_i}\right) + o(\tilde{\epsilon}\sqrt{\epsilon}),$$

which is obtained by adapting Corollary 5.22, and that the second factor in the summation in (C.114) is unity when $z_i = \pm 2$ since u_i and v_i are uniquely determined in these cases. ■

C.14 Proof of Lemma 5.26

First, note that with

$$\rho(\omega) \triangleq (1+\omega)\log(1+\omega), \qquad \omega \in [-1,\infty) \tag{C.115}$$

we have

$$I(X;Y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_X(x)\,P_Y(y)\,\rho\left(\sum_{i=1}^{k}\sigma_i\,f_i^*(x)\,g_i^*(y)\right), \tag{C.116}$$

which is convex in $\sigma^k = (\sigma_1,\dots,\sigma_k)$ since ρ is convex. In turn, we have

$$\begin{aligned}
&\max_{\sigma^k\in[0,1]^k} I(X;Y)\\
&\qquad = \max_{\sigma^k\in\{0,1\}^k} I(X;Y) && \text{(C.117)}\\
&\qquad = \max_{\mathcal{S}} \sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_X(x)\,P_Y(y)\left(\prod_{i\in\mathcal{S}}\mathbb{1}_{f_i^*(x)=g_i^*(y)}\right)\rho\left(\sum_{i\in\mathcal{S}} f_i^*(x)^2\right) && \text{(C.118)}\\
&\qquad = \max_{\mathcal{S}} \sum_{x\in\mathcal{X}} P_X(x)\log\left(1+\sum_{i\in\mathcal{S}} f_i^*(x)^2\right) && \text{(C.119)}\\
&\qquad = \sum_{x\in\mathcal{X}} P_X(x)\log\left(1+\sum_{i=1}^{k} f_i^*(x)^2\right), && \text{(C.120)}
\end{aligned}$$

where to obtain (C.117) we have used that the left-hand side is the maximum of a convex function over a convex set, so achieved on the boundary of the set, to obtain (C.118) we let

$$\mathcal{S} \triangleq \{i\in\{1,\dots,k\}\colon \sigma_i = 1\},$$

and use that when $\sigma_i = \mathbb{E}[f_i^*(X)\,g_i^*(Y)] = 1$ we have $f_i^*(X) = g_i^*(Y)$ with probability one, so

$$\left[\sigma_i f_i^*(X)\,g_i^*(Y)\right]\Big|_{\sigma_i=1} = f_i^*(X)^2$$

with probability one, to obtain (C.119) we have used that

$$\sum_{y\in\mathcal{Y}} P_Y(y)\prod_{i\in\mathcal{S}}\mathbb{1}_{f_i^*(x)=g_i^*(y)} = \left(1+\sum_{i\in\mathcal{S}} f_i^*(x)^2\right)^{-1}$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

since $\sum_{y \in \mathcal{Y}} P_{X,Y}(x,y) = P_X(x)$, and to obtain (C.120) we have used the monotonicity of $\log(\cdot)$, so $I(X;Y)$ is maximized when $\mathcal{S}$ is as large as possible, i.e., $\mathcal{S} = \{1, \dots, k\}$. Finally, to obtain the first equality in (5.91), note that when $\sigma_1 = \cdots = \sigma_k = 1$,

$$\begin{aligned} I(X;Y) &= I(f_*^k(X); g_*^k(Y)) && \text{(C.121)} \\ &= H(f_*^k(X)) - \underbrace{H(f_*^k(X)|g_*^k(Y))}_{=0} = H(f_*^k(X)), && \text{(C.122)} \end{aligned}$$

where to obtain (C.121) we have used (the analysis yielding) (2.24), and to obtain (C.122) we have used that $f_*^k(X) = g_*^k(Y)$ with probability one. ■

C.15 Proof of Proposition 5.27

To obtain (5.92), first note that in (5.90) we have

$$I(W;X,Y) = I(W;X) + \underbrace{I(W;Y|X)}_{=0} = I(W;X) \tag{C.123}$$

due to the first Markov constraint. Next, we define the additional CDMs $\tilde{\mathbf{B}}_{X,W}$ and $\tilde{\mathbf{B}}_{W,Y}$ with entries [cf. (2.28) and (2.29)]

$$\tilde{B}_{X,W}(x,w) = \frac{P_{X,W}(x,w) - P_X(w)\,P_W(w)}{\sqrt{P_X(x)}\sqrt{P_W(w)}}, \quad x \in \mathcal{X}, \; w \in \mathcal{W}$$

and

$$\tilde{B}_{Y,W}(y,w) = \frac{P_{Y,W}(y,w) - P_Y(y)\,P_W(w)}{\sqrt{P_W(w)}\sqrt{P_Y(y)}}, \quad w \in \mathcal{W}, \; y \in \mathcal{Y},$$

respectively, and note that that the Markov constraints in (5.90) can be expressed in the matrix form

$$\tilde{\mathbf{B}}_{Y,W} = \tilde{\mathbf{B}}\,\tilde{\mathbf{B}}_{X,W} \qquad \text{and} \qquad \tilde{\mathbf{B}}_{X,W} = \tilde{\mathbf{B}}^{\mathrm{T}}\tilde{\mathbf{B}}_{Y,W},$$

whence

$$\underbrace{(\mathbf{I} - \tilde{\mathbf{B}}^{\mathrm{T}}\tilde{\mathbf{B}})}_{\triangleq \mathbf{A}}\tilde{\mathbf{B}}_{X,W} = \mathbf{0}. \tag{C.124}$$

Since $\tilde{\mathbf{B}}$ has SVD [cf. (2.30), (3.15), and (3.16)]

$$\tilde{\mathbf{B}} = \mathbf{\Psi}_{(K-1)}^{Y}\mathbf{\Sigma}_{(K-1)}\left(\mathbf{\Psi}_{(K-1)}^{X}\right)^{\mathrm{T}},$$

it follows that $\mathbf{A}$ has SVD

$$\mathbf{A} = \boldsymbol{\Psi}^X_{(K-1)}(\mathbf{I} - \boldsymbol{\Sigma}^2_{(K-1)})\big(\boldsymbol{\Psi}^X_{(K-1)}\big)^{\mathrm{T}},$$

and, thus, singular values $1-\sigma_1^2, \dots, 1-\sigma_{K-1}^2$.

Accordingly, if $\sigma_1 < 1$, then $\mathbf{A}$ has full rank and (C.124) is satisfied if and only if $\tilde{\mathbf{B}}_{X,W} = \mathbf{0}$, i.e., X and W are independent, in which case $\bar{C}(X,Y) = I(W;X) = 0$, as the second case of (5.92) reflects. But if, instead, $\sigma_1 = 1$, the columns of $\tilde{\mathbf{B}}_{X,W}$ must lie in the nullspace of $\mathbf{A}$, i.e., $\tilde{\mathbf{B}}_{X,W}$ must have an SVD of the form

$$\tilde{\mathbf{B}}_{X,W} = \boldsymbol{\Psi}^X_{(k)}\boldsymbol{\Lambda}\big(\boldsymbol{\Psi}^W\big)^{\mathrm{T}}, \tag{C.125}$$

where $\boldsymbol{\Lambda}$ is an $k \times k$ diagonal matrix with diagonal entries $\lambda_1, \dots, \lambda_k \in [0,1]$, with k defined as in (5.92), and $\boldsymbol{\Psi}^W$ is an orthogonal matrix.[1]

It remains only to determine the optimizing choices of $\lambda_1, \dots, \lambda_k$ in (5.90) for this case. For this, applying Lemma 5.26 with W replacing Y, we conclude that the optimizing $\lambda_1, \dots, \lambda_k$ in (C.125) are all unity and

$$I(X;W) = H\big(f_*^k(X)\big) = \sum_{x\in\mathcal{X}} P_X(x)\log\left(1 + \sum_{i=1}^{k} f_i^*(x)^2\right).$$

■

[1] Note, in particular, this implies $|\mathcal{W}| = k$ suffices.

D

Appendices for Section 6

D.1 Proof of Proposition 6.1

Our proof makes use of two lemmas. The first is the following vector generalization of Bernstein's inequality [51, Theorem 2.4].

Lemma D.1 (Bernstein Inequality (Vector Version)). For some dimension d, let $\tilde{Z}_1, \dots, \tilde{Z}_n \in \mathbb{R}^d$ be independent zero-mean random vectors such that for some constant $c > 0$,

$$\mathbb{P}\Big(\|\tilde{Z}_i\| \leq c\Big) = 1, \qquad i = 1, \dots, n.$$

Moreover, let $\bar{c} \in (0, c^2]$ be a constant such that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Big[\|\tilde{Z}_i\|^2\Big] \leq \bar{c}.$$

Then, for all $0 \leq \delta \leq \bar{c}/c$,[1]

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} \tilde{Z}_i\right\| \geq \delta\right) \leq \exp\left\{\frac{1}{4} - \frac{\delta^2 n}{8\bar{c}}\right\}.$$

The second lemma is the following.

[1] As noted in [51], this bound does not depend on d.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Lemma D.2. Given dimensions k_1 and k_2 and any matrices $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{k_1 \times k_2}$, we have, for every $k \in \{1, \dots, \min\{k_1, k_2\}\}$,

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

Proof of Lemma D.2. We have

$$\begin{aligned}\sum_{i=1}^{k} \left|\sigma_i(\mathbf{A}_1) - \sigma_i(\mathbf{A}_2)\right| &\leq \sum_{i=1}^{k} \sigma_i(\mathbf{A}_1 - \mathbf{A}_2) && \text{(D.2)}\\ &\leq \sqrt{k} \sqrt{\sum_{i=1}^{k} \sigma_i(\mathbf{A}_1 - \mathbf{A}_2)^2} && \text{(D.3)}\\ &\leq \sqrt{k} \sqrt{\sum_{i=1}^{\min\{k_1,k_2\}} \sigma_i(\mathbf{A}_1 - \mathbf{A}_2)^2}\\ &= \sqrt{k}\, \|\mathbf{A}_1 - \mathbf{A}_2\|_{\mathrm{F}}, && \text{(D.4)}\end{aligned}$$

where to obtain (D.2) we use the following standard inequality (see, e.g., [113, Theorem 3.4.5]):

Lemma D.3 (Lidskii Inequality)**.** Given dimensions k_1 and k_2 and any matrices $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{k_1 \times k_2}$, we have, for every $k \in \{1, \min\{k_1, k_2\}\}$ and $1 \leq i_1 < i_2 < \cdots < i_k \leq \min\{k_1, k_2\}$,

$$\sum_{j=1}^{k} \left|\sigma_{i_j}(\mathbf{A}_1) - \sigma_{i_j}(\mathbf{A}_2)\right| \leq \|\mathbf{A}_1 - \mathbf{A}_2\|_{(k)},$$

where $\sigma_1(\cdot) \geq \cdots \geq \sigma_{\min\{k_1,k_2\}}(\cdot)$ denote the ordered singular values of its (matrix) argument.

In turn, to obtain (D.3) we use the Cauchy-Schwarz inequality, and to obtain (D.4) we use the definition of the Frobenius norm. ■

Our proof of Proposition 6.1 proceeds as follows. First, for each $i \in \{1, \dots, n\}$ let $\mathbf{Z}_i$ denote a random $|\mathcal{Y}| \times |\mathcal{X}|$ matrix with (y, x)th element

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

Accordingly, the $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ are i.i.d. and $\mathbb{E}[\mathbf{Z}_i] = \tilde{\mathbf{B}}$. Now

$$\tilde{\mathbf{Z}}_i \triangleq \mathbf{Z}_i - \mathbb{E}[\mathbf{Z}_i] = \mathbf{Z}_i - \tilde{\mathbf{B}} \tag{D.5}$$

satisfies

$$\begin{aligned}
\|\tilde{\mathbf{Z}}_i\|_{\mathrm{F}}^2 &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \frac{\left(\mathbb{1}_{\{X_i=x,\,Y_i=y\}} - P_{X,Y}(x,y)\right)^2}{P_X(x)\,P_Y(y)} \\
&\le \frac{1}{p_0^2} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \left(\mathbb{1}_{\{X_i=x,\,Y_i=y\}} - P_{X,Y}(x,y)\right)^2 &\text{(D.6)}\\
&= \frac{1}{p_0^2} \Bigg[\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \mathbb{1}_{\{X_i=x,\,Y_i=y\}} \\
&\qquad\qquad - 2 \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \mathbb{1}_{\{X_i=x,\,Y_i=y\}}\, P_{X,Y}(x,y) \\
&\qquad\qquad\quad + \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{X,Y}(x,y)^2 \Bigg] &\text{(D.7)}\\
&\le \frac{2}{p_0^2} \triangleq c^2, &\text{(D.8)}
\end{aligned}$$

where to obtain (D.6) we have used (6.16), and where to obtain (D.8) we have used that in (D.7) the first term within the brackets is unity, the second is upper bounded by zero, and the third term is upper bounded by unity since

$$\sum_{z\in\mathcal{Z}} q(z)^2 \le \sum_{z\in\mathcal{Z}} q(z) = 1, \ \text{any (countable) } \mathcal{Z} \text{ and } q \in \mathcal{P}^{\mathcal{Z}}. \tag{D.9}$$

Moreover,

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\|\tilde{\mathbf{Z}}_i\|_{\mathrm{F}}^2\right] &= \mathbb{E}\left[\|\tilde{\mathbf{Z}}_1\|_{\mathrm{F}}^2\right] &\text{(D.10)}\\
&\le \frac{1}{p_0^2} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \operatorname{var}\left[\mathbb{1}_{\{X_1=x,\,Y_1=y\}}\right] &\text{(D.11)}\\
&= \frac{1}{p_0^2} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \left[P_{X,Y}(x,y) - P_{X,Y}(x,y)^2\right] &\text{(D.12)}\\
&\le \frac{1}{p_0^2} \triangleq \bar{c}, &\text{(D.13)}
\end{aligned}$$

where to obtain (D.10) we have used that $\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_n$ are i.i.d., to obtain (D.11) we take the expectation of (D.6), to obtain (D.12) we have used that

$$\operatorname{var}\left[\mathbb{1}_{\{X_1=x,\, Y_1=y\}}\right] = P_{X,Y}(x,y) - P_{X,Y}(x,y)^2$$

since $\mathbb{1}_{\{X_1=x,\, Y_1=y\}}$ is a Bernoulli random variable, and to obtain (D.13) we have used that the second term in (D.12) is upper bounded by zero.

Finally, for $0 \le \delta \le \sqrt{k/2}/p_0$ we have

$$\mathbb{P}\left(\sum_{i=1}^{k} |\hat{\sigma}_i - \sigma_i| \ge \delta\right) \le \mathbb{P}\left(\|\hat{\mathbf{B}} - \tilde{\mathbf{B}}\|_{\mathrm{F}} \ge \frac{\delta}{\sqrt{k}}\right) \tag{D.14}$$

$$= \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{Z}}_i\right\|_{\mathrm{F}} \ge \frac{\delta}{\sqrt{k}}\right) \tag{D.15}$$

$$\le \exp\left\{\frac{1}{4} - \frac{p_0^2\, \delta^2 n}{8k}\right\}, \tag{D.16}$$

where to obtain (D.14) we have used Lemma D.2, to obtain (D.15) we have used that

$$\hat{\mathbf{B}} - \tilde{\mathbf{B}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{Z}}_i, \tag{D.17}$$

and to obtain (D.16) we have used Lemma D.1 with [cf. (D.8)] $c = \sqrt{2}/p_0$ and [cf. (D.13)] $\bar{c} = 1/p_0^2$ (and construed the associated $\tilde{\mathbf{Z}}_i$ as vectors). ■

D.2 Proof of Corollary 6.2

First, we have

$$\sum_{i=1}^{k} |\hat{\sigma}_i - \sigma_i| \le \sum_{i=1}^{k} (\hat{\sigma}_i + \sigma_i) \tag{D.18}$$

$$= \|\hat{\mathbf{B}}\|_{(k)} + \|\tilde{\mathbf{B}}\|_{(k)}$$

$$\le k\left(1 + \|\hat{\mathbf{B}}\|_{\mathrm{s}}\right) \tag{D.19}$$

$$\le k\left(1 + \|\hat{\mathbf{B}}\|_{\mathrm{F}}\right), \tag{D.20}$$

where to obtain (D.18) we have used the triangle inequality, to obtain (D.19) we have used that $\|\mathbf{A}\|_{(k)} \le k\|\mathbf{A}\|_{\mathrm{s}}$ for any matrix $\mathbf{A} \in \mathbb{R}^{k_1 \times k_2}$

and $k \in \{1, \ldots, \min\{k_1, k_2\}\}$, and to obtain (D.20) we have used the standard inequality

$$\|\mathbf{A}\|_{\mathrm{s}} \leq \|\mathbf{A}\|_{\mathrm{F}} \quad \text{for any matrix } \mathbf{A}. \tag{D.21}$$

In turn,

$$\begin{aligned}
\|\hat{\mathbf{B}}\|_{\mathrm{F}}^2 &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \frac{\left(\hat{P}_{X,Y}(x,y) - P_X(x)\,P_Y(y)\right)^2}{P_X(x)\,P_Y(y)} \\
&\leq \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \left[\frac{\hat{P}_{X,Y}(x,y)^2}{P_X(x)\,P_Y(y)} + P_X(x)\,P_Y(y)\right] \\
&\leq \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \left[\frac{\hat{P}_{X,Y}(x,y)^2}{p_0^2} + P_X(x)\,P_Y(y)\right] && \text{(D.22)} \\
&\leq \frac{1}{p_0^2} + 1 && \text{(D.23)} \\
&\leq \frac{2}{p_0^2}, && \text{(D.24)}
\end{aligned}$$

where to obtain (D.22) we have used (6.16), and to obtain (D.23) we have used (D.9).

Next, with the event

$$\mathcal{E}_\delta \triangleq \left\{\sum_{i=1}^{k} |\hat{\sigma}_i - \sigma_i| \geq \delta\right\}, \qquad 0 \leq \delta \leq \frac{1}{p_0}\sqrt{\frac{k}{2}},$$

we have[2]

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{i=1}^{k} |\hat{\sigma}_i - \sigma_i|\right)^2\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{k} |\hat{\sigma}_i - \sigma_i|\right)^2 \middle| \mathcal{E}_\delta^{\mathrm{c}}\right] \mathbb{P}(\mathcal{E}_\delta^{\mathrm{c}}) \\
&\quad + \mathbb{E}\left[\left(\sum_{i=1}^{k} |\hat{\sigma}_i - \sigma_i|\right)^2 \middle| \mathcal{E}_\delta\right] \mathbb{P}(\mathcal{E}_\delta) \\
&\leq \delta^2 + k^2 \left(1 + \frac{\sqrt{2}}{p_0}\right)^2 \exp\left\{\frac{1}{4} - \frac{p_0^2\,\delta^2 n}{8k}\right\}, && \text{(D.25)}
\end{aligned}$$

[2] We use $(\cdot)^{\mathrm{c}}$ to denote set complement.

where to obtain the inequality we have used that $\mathbb{P}(\mathcal{E}_\delta^{\mathrm{c}}) \leq 1$, (D.20) with (D.24), and (6.17) in Proposition 6.1.

To obtain the tightest bound, we optimize (D.25) over δ, yielding (6.18). In particular, we have

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{i=1}^{k}|\hat{\sigma}_i - \sigma_i|\right)^2\right] &\leq \min_{\delta}\left(\delta^2 + k^2\left(1+\frac{\sqrt{2}}{p_0}\right)^2 \exp\left\{\frac{1}{4} - \frac{p_0^2\,\delta^2 n}{8k}\right\}\right) \\
&= \frac{8k}{p_0^2 n}\left(1+\log\left[k^2\left(1+\frac{\sqrt{2}}{p_0}\right)^2 \mathrm{e}^{1/4}\,\frac{p_0^2 n}{8k}\right]\right) \qquad \text{(D.26)} \\
&\leq \frac{8k}{p_0^2 n}\left(\frac{3}{4} + \log(kn)\right) \qquad \text{(D.27)} \\
&= \frac{6k + 8k\log(kn)}{p_0^2 n}, \qquad \text{(D.28)}
\end{aligned}$$

where to obtain (D.26) we recognize that the right-hand side of (D.25) takes the form of (6.19) with the mappings

$$a = k^2\left(1+\frac{\sqrt{2}}{p_0}\right)^2 \mathrm{e}^{1/4}, \qquad b = \frac{np_0^2}{8k}, \qquad \omega = \delta^2, \qquad \text{(D.29)}$$

and apply Lemma 6.3, and to obtain (D.27) we have used that $p_0 + \sqrt{2} \leq 2$ since $p_0 \leq 1/2$ as $\min\{|\mathcal{X}|, |\mathcal{Y}|\} \geq 2$, and that $\log(2) \geq 1/2$.

It remains to determine conditions under which

$$\omega_* \triangleq \delta_*^2 = \frac{8k}{p_0^2 n}\log\left[k^2\left(1+\frac{\sqrt{2}}{p_0}\right)^2 \mathrm{e}^{1/4}\,\frac{p_0^2 n}{8k}\right] \qquad \text{(D.30)}$$

satisfies the conditions of Proposition 6.1, viz.,

$$0 \leq \omega_* \leq \frac{k}{2p_0^2}. \qquad \text{(D.31)}$$

Proceeding, since from (D.30) we have

$$\omega_* \leq \frac{8k}{p_0^2 n}\left(\log(4kn) + \underbrace{\frac{1}{4} - \log(8)}_{<0}\right) < \frac{k}{2p_0^2}\left[\frac{16}{n}\log(4kn)\right],$$

where to obtain the first inequality we have again used that $p_0+\sqrt{2}\leq 2$. Hence, the second inequality in (D.31) is satisfied when n is sufficiently large that $n\geq 16\log(4kn)$.

Moreover, since from (D.30) we also have

$$\omega_* = \frac{8k}{p_0^2 n}\log\left[k(p_0+\sqrt{2})^2 \mathrm{e}^{1/4}\frac{n}{8}\right] > \frac{8k}{p_0^2 n}\log\left(\frac{n}{4}\right),$$

where to obtain the inequality we have used that $k\geq 1$, $p_0>0$, and $\mathrm{e}^{1/4}>1$. Hence, the first inequality in (D.31) is satisfied when $n\geq 4$, which we note is satisfied when our condition for satisfying the second inequality in (D.31) is. Indeed, satisfying $n\geq 16\log(4kn)$ even for $k=1$ requires $n\geq 96$. ■

D.3 Proof of Corollary 6.4

First, note that

$$\begin{aligned}\left|\frac{1}{2}\sum_{i=1}^{k}(\hat{\sigma}_i^2-\sigma_i^2)\right| &\leq \frac{1}{2}\sum_{i=1}^{k}|\hat{\sigma}_i^2-\sigma_i^2| \\ &= \frac{1}{2}\sum_{i=1}^{k}|\hat{\sigma}_i-\sigma_i|(\hat{\sigma}_i+\sigma_i) \\ &\leq \frac{1}{2}(\hat{\sigma}_1+\sigma_1)\sum_{i=1}^{k}|\hat{\sigma}_i-\sigma_i|,\end{aligned}$$

where

$$\begin{aligned}\sigma_1+\hat{\sigma}_1 &= \|\tilde{\mathbf{B}}\|_{\mathrm{s}}+\|\hat{\mathbf{B}}\|_{\mathrm{s}} \\ &\leq 1+\|\hat{\mathbf{B}}\|_{\mathrm{F}} &\text{(D.32)} \\ &\leq 1+\frac{\sqrt{2}}{p_0} = \frac{p_0+\sqrt{2}}{p_0} &\text{(D.33)} \\ &\leq \frac{2}{p_0}, &\text{(D.34)}\end{aligned}$$

whence

$$\left|\frac{1}{2}\sum_{i=1}^{k}(\hat{\sigma}_i^2-\sigma_i^2)\right| \leq \frac{1}{p_0}\sum_{i=1}^{k}|\hat{\sigma}_i-\sigma_i|. \tag{D.35}$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

To obtain (D.32) we have used (D.21) and that $\tilde{\mathbf{B}}$ is contractive, and to obtain (D.33) we have used (D.24) in the proof of Corollary 6.2, and to obtain (D.34) we have used that $p_0 \leq 1/2$ as $\min\{|\mathcal{X}|, |\mathcal{Y}|\} \geq 2$.

Hence, we obtain (6.25) from (D.35) via

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

where to obtain the final inequality we have used (6.17), which holds for $0 \leq p_0\,\delta \leq \sqrt{k/2}/p_0$. Moreover, we obtain (6.26) from (D.35) via

$$\mathbb{E}\left[\left|\frac{1}{2}\sum_{i=1}^{k}(\hat{\sigma}_i^2 - \sigma_i^2)\right|^2\right] \leq \mathbb{E}\left[\frac{1}{p_0^2}\left(\sum_{i=1}^{k}|\hat{\sigma}_i - \sigma_i|\right)^2\right] \leq \frac{6k + 8k\log(nk)}{p_0^4 n},$$

where to obtain the final equality we have used (6.18). ■

D.4 Proof of Proposition 6.5

Our proof makes use of two lemmas. The first is the following matrix generalization of Bernstein's inequality [265, Theorem 1.6].

Lemma D.4 (Bernstein Inequality (Matrix Version)). For some dimensions d_1 and d_2, let $\tilde{\mathbf{Z}}_1, \dots, \tilde{\mathbf{Z}}_n \in \mathbb{R}^{d_1 \times d_2}$ be independent, zero-mean random matrices such that for some constant $c > 0$,

$$\mathbb{P}\left(\|\tilde{\mathbf{Z}}_i\|_{\mathrm{s}} \leq c\right) = 1, \qquad i = 1, \dots, n.$$

Moreover, let $\bar{c} \in (0, c^2]$ be a constant such that

$$\max\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\mathrm{cov}(\tilde{\mathbf{Z}}_i)\right\|_{\mathrm{s}}, \left\|\frac{1}{n}\sum_{i=1}^{n}\mathrm{cov}(\tilde{\mathbf{Z}}_i^{\mathrm{T}})\right\|_{\mathrm{s}}\right\} \leq \bar{c},$$

where for an arbitrary random matrix $\mathbf{W}$

$$\mathrm{cov}(\mathbf{W}) \triangleq \mathbb{E}[(\mathbf{W} - \mathbb{E}[\mathbf{W}])(\mathbf{W} - \mathbb{E}[\mathbf{W}])^{\mathrm{T}}].$$

Then, for all $0 \leq \delta \leq \bar{c}/c$,

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{Z}}\right\|_{\mathrm{s}} \geq \delta\right) \leq (d_1 + d_2)\exp\left\{-\frac{3\delta^2 n}{8\bar{c}}\right\}.$$

The second of these lemmas is as follows.

Lemma D.5. Given $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{k_1 \times k_2}$ and $k \in \{1, \ldots, \min\{k_1, k_2\}\}$, we have

$$0 \leq \|\mathbf{A}_1\mathbf{\Psi}_{(k)}^{\mathbf{A}_1}\|_{\mathrm{F}}^2 - \|\mathbf{A}_1\mathbf{\Psi}_{(k)}^{\mathbf{A}_2}\|_{\mathrm{F}}^2 \leq 4k\,\|\mathbf{A}_1\|_{\mathrm{s}}\,\|\mathbf{A}_1 - \mathbf{A}_2\|_{\mathrm{s}}, \tag{D.36}$$

where $\boldsymbol{\psi}_i^{\mathbf{A}}$ denotes the right singular vector of $\mathbf{A}$ corresponding to $\sigma_i(\mathbf{A})$, and

$$\mathbf{\Psi}_{(k)}^{\mathbf{A}} \triangleq \begin{bmatrix}\boldsymbol{\psi}_1^{\mathbf{A}} & \cdots & \boldsymbol{\psi}_k^{\mathbf{A}}\end{bmatrix},$$

which has orthonormal columns.

Proof of Lemma D.5. The left-hand inequality follows immediately from Lemma 3.1. For the right-hand inequality, we have

$$\begin{aligned}
&\|\mathbf{A}_1\mathbf{\Psi}_{(k)}^{\mathbf{A}_1}\|_{\mathrm{F}}^2 - \|\mathbf{A}_1\mathbf{\Psi}_{(k)}^{\mathbf{A}_2}\|_{\mathrm{F}}^2 \\
&\qquad= \sum_{i=1}^{k}\left(\|\mathbf{A}_1\boldsymbol{\psi}_i^{\mathbf{A}_1}\|^2 - \|\mathbf{A}_1\boldsymbol{\psi}_i^{\mathbf{A}_2}\|^2\right) \\
&\qquad\leq \sum_{i=1}^{k}\left|\|\mathbf{A}_1\boldsymbol{\psi}_i^{\mathbf{A}_1}\|^2 - \|\mathbf{A}_1\boldsymbol{\psi}_i^{\mathbf{A}_2}\|^2\right| &\text{(D.37)} \\
&\qquad= \sum_{i=1}^{k}\left|\|\mathbf{A}_1\boldsymbol{\psi}_i^{\mathbf{A}_1}\| - \|\mathbf{A}_1\boldsymbol{\psi}_i^{\mathbf{A}_2}\|\right|\left(\|\mathbf{A}_1\boldsymbol{\psi}_i^{\mathbf{A}_1}\| + \|\mathbf{A}_1\boldsymbol{\psi}_i^{\mathbf{A}_2}\|\right) \\
&\qquad\leq 2\,\|\mathbf{A}_1\|_{\mathrm{s}}\sum_{i=1}^{k}\left|\|\mathbf{A}_1\boldsymbol{\psi}_i^{\mathbf{A}_1}\| - \|\mathbf{A}_1\boldsymbol{\psi}_i^{\mathbf{A}_2}\|\right| &\text{(D.38)} \\
&\qquad\leq 2\,\|\mathbf{A}_1\|_{\mathrm{s}}\sum_{i=1}^{k}\Bigg(\left|\|\mathbf{A}_1\boldsymbol{\psi}_i^{\mathbf{A}_1}\| - \|\mathbf{A}_2\boldsymbol{\psi}_i^{\mathbf{A}_2}\|\right| \\
&\qquad\qquad\qquad\qquad + \left|\|\mathbf{A}_2\boldsymbol{\psi}_i^{\mathbf{A}_2}\| - \|\mathbf{A}_1\boldsymbol{\psi}_i^{\mathbf{A}_2}\|\right|\Bigg) &\text{(D.39)} \\
&\qquad\leq 2\,\|\mathbf{A}_1\|_{\mathrm{s}}\sum_{i=1}^{k}\left(|\sigma_i(\mathbf{A}_1) - \sigma_i(\mathbf{A}_2)| + \|(\mathbf{A}_1 - \mathbf{A}_2)\boldsymbol{\psi}_i^{\mathbf{A}_2}\|\right) &\text{(D.40)}
\end{aligned}$$

$$\leq 2 \left\|\mathbf{A}_1\right\|_{\mathrm{s}} \sum_{i=1}^{k} \Big(\left|\sigma_i(\mathbf{A}_1) - \sigma_i(\mathbf{A}_2)\right| + \left\|\mathbf{A}_1 - \mathbf{A}_2\right\|_{\mathrm{s}} \Big) \tag{D.41}$$

$$\leq 4k \left\|\mathbf{A}_1\right\|_{\mathrm{s}} \left\|\mathbf{A}_2 - \mathbf{A}_1\right\|_{\mathrm{s}}, \tag{D.42}$$

where to obtain (D.37) we have used the triangle inequality, to obtain (D.38) we have used Fact 5.12, to obtain (D.39) we have again used the triangle inequality, to obtain (D.40) we have used that $\|\mathbf{A}_1 \boldsymbol{\psi}_i^{\mathbf{A}_1}\| = \sigma_i(\mathbf{A}_1)$ and $\|\mathbf{A}_2 \boldsymbol{\psi}_i^{\mathbf{A}_2}\| = \sigma_i(\mathbf{A}_2)$, and the (reverse) triangle inequality, to obtain (D.41) we have again used Fact 5.12, and to obtain (D.42) we have used the following standard inequality [114, Corollary 7.3.5(a)] [257, Theorem 1]:

Lemma D.6 (Weyl Inequality). For every $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{k_1 \times k_2}$, we have, with $K \triangleq \min\{k_1, k_2\}$,

$$\max_{1 \leq i \leq K} \left|\sigma_i(\mathbf{A}_1) - \sigma_i(\mathbf{A}_2)\right| \leq \left\|\mathbf{A}_1 - \mathbf{A}_2\right\|_{\mathrm{s}}. \tag{D.43}$$

■

Our proof of Proposition 6.5 proceeds as follows. First, with $\check{f}_i$ as defined in (6.10a) and $\hat{\boldsymbol{\psi}}_i^X$ as defined via (6.14a), we have

$$\begin{aligned}
\mathbb{E}_{P_Y}\left[\left\|\mathbb{E}_{P_{X|Y}}\left[f_*^k(X)\right]\right\|^2 - \left\|\mathbb{E}_{P_{X|Y}}\left[\check{f}_*^k(X)\right]\right\|^2 \right]& \\
&= \left\|\tilde{\mathbf{B}}\, \boldsymbol{\Psi}_{(k)}^X\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\, \hat{\boldsymbol{\Psi}}_{(k)}^X\right\|_{\mathrm{F}}^2 && \text{(D.44)}\\
&\leq 4k \left\|\tilde{\mathbf{B}}\right\|_{\mathrm{s}} \left\|\tilde{\mathbf{B}} - \hat{\mathbf{B}}\right\|_{\mathrm{s}} && \text{(D.45)}\\
&\leq 4k \left\|\tilde{\mathbf{B}} - \hat{\mathbf{B}}\right\|_{\mathrm{s}}, && \text{(D.46)}
\end{aligned}$$

where to obtain (D.44) we have used (6.28), to obtain (D.45) we have used Lemma D.5, and to obtain (D.46) we have used that $\|\tilde{\mathbf{B}}\|_{\mathrm{s}} \leq 1$.

Now let $\hat{\mathbf{B}}_i$ denote an $|\mathcal{Y}| \times |\mathcal{X}|$ matrix with (y, x)th entry

$$\hat{B}_i(x, y) \triangleq \frac{\mathbb{1}_{X_i = x,\ Y_i = y}}{\sqrt{P_X(x)\, P_Y(y)}},$$

and let

$$\tilde{\mathbf{Z}}_i \triangleq \hat{\mathbf{B}}_i - \mathbf{B},$$

which we note is consistent with the definition (D.5) in the proof of Proposition 6.1 and thus $\mathbb{E}\Big[\tilde{\mathbf{Z}}_i\Big] = \mathbf{0}$. Then we have

$$\begin{aligned}\big\|\tilde{\mathbf{Z}}_i\big\|_{\mathrm{s}} &= \big\|\hat{\mathbf{B}}_i - \mathbf{B}\big\|_{\mathrm{s}} \\ &\leq \big\|\mathbf{B}\big\|_{\mathrm{s}} + \big\|\hat{\mathbf{B}}_i\big\|_{\mathrm{s}} \qquad (D.47)\\ &= 1 + \frac{1}{\sqrt{P_X(X_i)\, P_Y(Y_i)}} \qquad (D.48)\\ &\leq 1 + \frac{1}{p_0} \triangleq c, \qquad (D.49)\end{aligned}$$

where to obtain (D.47) we have used the spectral norm triangle inequality, to obtain (D.48) we have used that $\|\mathbf{B}\|_{\mathrm{s}} = 1$ and $\hat{\mathbf{B}}_i$ has a single nonzero entry so (with the usual abuse of notation as discussed in footnote 5) $\mathbf{e}_{Y_i}$ and $\mathbf{e}_{X_i}$ are its principal left and right singular vectors, respectively, and to obtain (D.49) we have used the definition of p_0.

Next, we have

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n} \operatorname{cov}(\tilde{\mathbf{Z}}_i)\right\|_{\mathrm{s}} &= \big\|\operatorname{cov}(\tilde{\mathbf{Z}}_1)\big\|_{\mathrm{s}} \qquad (D.50)\\ &= \big\|\mathbb{E}\big[(\hat{\mathbf{B}}_1 - \mathbf{B})(\hat{\mathbf{B}}_1 - \mathbf{B})^{\mathrm{T}}\big]\big\|_{\mathrm{s}} \\ &= \big\|\mathbb{E}\big[\hat{\mathbf{B}}_1\hat{\mathbf{B}}_1^{\mathrm{T}}\big] - \mathbf{B}\mathbf{B}^{\mathrm{T}}\big\|_{\mathrm{s}} \\ &\leq \big\|\mathbf{B}\mathbf{B}^{\mathrm{T}}\big\|_{\mathrm{s}} + \big\|\mathbb{E}\big[\hat{\mathbf{B}}_1\hat{\mathbf{B}}_1^{\mathrm{T}}\big]\big\|_{\mathrm{s}} \qquad (D.51)\\ &= 1 + \max_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \frac{P_{X|Y}(x|y)}{P_X(x)} \qquad (D.52)\\ &\leq 1 + \frac{1}{p_0} \max_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} P_{X|Y}(x|y) \qquad (D.53)\\ &= 1 + \frac{1}{p_0}, \qquad (D.54)\end{aligned}$$

where to obtain (D.50) we have used that the $\tilde{\mathbf{Z}}_1, \dots, \tilde{\mathbf{Z}}_n$ are identically distributed, to obtain (D.51) we have again used the triangle inequality, to obtain (D.52) we have used that $\sigma_0^2 = 1$ is the principal singular value of $\mathbf{B}\mathbf{B}^{\mathrm{T}}$, and that $\hat{\mathbf{B}}_1\hat{\mathbf{B}}_1^{\mathrm{T}}$ is a diagonal matrix whose (y, y)th entry is $\mathbb{1}_{Y_1=y}/(P_X(X_1)\, P_Y(y))$, so $\mathbb{E}[\hat{\mathbf{B}}_1\hat{\mathbf{B}}_1^{\mathrm{T}}]$ has (y, y)th entry

$$\mathbb{E}\left[\frac{\mathbb{1}_{Y_1=y}}{P_X(X_1)\, P_Y(y)}\right] = \sum_{x'\in\mathcal{X},\, y'\in\mathcal{Y}} P_{X,Y}(x', y')\frac{\mathbb{1}_{y'=y}}{P_X(x')\, P_Y(y')}$$

$$= \sum_{x\in\mathcal{X}} \frac{P_{X,Y}(x,y)}{P_X(x)\,P_Y(y)}$$

$$= \sum_{x\in\mathcal{X}} \frac{P_{X|Y}(x|y)}{P_X(x)},$$

and to obtain (D.53) we have again used the definition of p_0. Moreover, interchanging the roles of x and y, we have, by symmetry,

$$\left\|\frac{1}{n}\sum_{i=1}^{n} \mathrm{cov}(\tilde{\mathbf{Z}}_i^{\mathrm{T}})\right\|_{\mathrm{s}} = \left\|\frac{1}{n}\sum_{i=1}^{n} \mathrm{cov}(\tilde{\mathbf{Z}}_i)\right\|_{\mathrm{s}} = 1+\frac{1}{p_0} \triangleq \bar{c}, \tag{D.55}$$

where to obtain the second equality we have used (D.54).

Finally, we have

$$\mathbb{P}_{\check{f}_*^k}\left(\mathbb{E}_{P_Y}\left[\left\|\mathbb{E}_{P_{X|Y}}\left[f_*^k(X)\right]\right\|^2 - \left\|\mathbb{E}_{P_{X|Y}}\left[\check{f}_*^k(X)\right]\right\|^2\right] \geq \delta\right)$$

$$\leq \mathbb{P}\left(\left\|\tilde{\mathbf{B}}-\hat{\mathbf{B}}\right\|_{\mathrm{s}} \geq \frac{\delta}{4k}\right) \tag{D.56}$$

$$\leq \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{Z}}_i\right\|_{\mathrm{s}} \geq \frac{\delta}{4k}\right) \tag{D.57}$$

$$\leq (|\mathcal{X}|+|\mathcal{Y}|)\exp\left\{-\frac{3n}{8}\left(\frac{1}{1+1/p_0}\right)\left(\frac{\delta}{4k}\right)^2\right\} \tag{D.58}$$

$$\leq (|\mathcal{X}|+|\mathcal{Y}|)\exp\left\{-\frac{p_0\,\delta^2\,n}{64k^2}\right\}, \tag{D.59}$$

where to obtain (D.56) we have used (D.46), to obtain (D.57) we have used (D.17), to obtain (D.58) we have used Lemma D.4, and to obtain (D.59) we have again used that $p_0 \leq 1/2$ since $\min\{|\mathcal{X}|,|\mathcal{Y}|\} \geq 2$. ■

D.5 Proof of Corollary 6.6

First, adapting our notation from (6.27) for convenience,

$$\tilde{\mu}_2(\check{f}_*^k) \triangleq \mathbb{E}_{P_Y}\left[\left\|\mathbb{E}_{P_{X|Y}}\left[f_*^k(X)\right]\right\|^2 - \left\|\mathbb{E}_{P_{X|Y}}\left[\check{f}_*^k(X)\right]\right\|^2\right]$$

$$= \left\|\tilde{\mathbf{B}}\boldsymbol{\Psi}_{(k)}^X\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Psi}}_{(k)}^X\right\|_{\mathrm{F}}^2 \tag{D.60}$$

$$\leq \left\|\tilde{\mathbf{B}}\boldsymbol{\Psi}_{(k)}^X\right\|_{\mathrm{F}}^2$$

$$\leq \left\|\tilde{\mathbf{B}}\right\|_{\mathrm{s}}^2 \left\|\boldsymbol{\Psi}_{(k)}^X\right\|_{\mathrm{F}}^2 \tag{D.61}$$

$$
\begin{align}
&\leq \sum_{i=1}^{k} \left\| \boldsymbol{\psi}_i^X \right\|^2 \tag{D.62}\\
&= k, \tag{D.63}
\end{align}
$$

where to obtain (D.60) we have used (D.44), to obtain (D.61) we have used Fact 5.12, to obtain (D.62) we have used that $\|\tilde{\mathbf{B}}\|_{\mathrm{s}} \leq 1$, and to obtain (D.63) we have used that the singular vectors have unit norm.

Next, with the event

$$\mathcal{E}_\delta \triangleq \left\{ \tilde{\mu}_2(\check{f}_*^k) \geq \delta \right\}, \qquad 0 \leq \delta \leq 4k,$$

we have that the left-hand side of (6.32) is bounded according to

$$
\begin{align}
\mathbb{E}_{\check{f}_*^k}\left[\tilde{\mu}_2(\check{f}_*^k)^2 \right] &= \mathbb{E}_{\check{f}_*^k}\left[\tilde{\mu}_2(\check{f}_*^k)^2 \middle| \mathcal{E}_\delta^{\mathrm{c}} \right] \mathbb{P}(\mathcal{E}_\delta^{\mathrm{c}}) + \mathbb{E}_{\check{f}_*^k}\left[\tilde{\mu}_2(\check{f}_*^k)^2 \middle| \mathcal{E}_\delta \right] \mathbb{P}(\mathcal{E}_\delta) \nonumber\\
&\leq \delta^2 + k^2 (|\mathcal{X}| + |\mathcal{Y}|) \exp\left\{ -\frac{p_0\, \delta^2 n}{64k^2} \right\}, \tag{D.64}
\end{align}
$$

where to obtain the inequality we have used that $\mathbb{P}(\mathcal{E}_\delta^{\mathrm{c}}) \leq 1$, (D.63), and Proposition 6.5.

To obtain the tightest bound, we optimize (D.64) over δ, yielding (6.32). In particular, we have

$$
\begin{align}
\mathbb{E}_{\check{f}_*^k}\left[\tilde{\mu}_2(\check{f}_*^k)^2 \right] &\leq \min_{\delta} \left(\delta^2 + k^2 (|\mathcal{X}| + |\mathcal{Y}|) \exp\left\{ -\frac{p_0\, \delta^2 n}{64k^2} \right\} \right) \nonumber\\
&= \frac{64k^2}{p_0 n} \left[1 + \log\left(k^2 (|\mathcal{X}| + |\mathcal{Y}|) \frac{p_0 n}{64k^2} \right) \right] \tag{D.65}\\
&= \frac{64k^2}{p_0 n} \Big(\log\big[(|\mathcal{X}| + |\mathcal{Y}|) p_0 n \big] + \big[1 - \log(64) \big] \Big) \nonumber\\
&\leq \frac{64k^2}{p_0 n} \Big(\log\big[(|\mathcal{X}| + |\mathcal{Y}|) p_0 n \big] - 3 \Big), \tag{D.66}
\end{align}
$$

where to obtain (D.65) we recognize that the right-hand side of (D.64) takes the form of (6.19) with the mappings

$$a = k^2 (|\mathcal{X}| + |\mathcal{Y}|), \qquad b = \frac{p_0 n}{64k^2}, \qquad \omega = \delta^2, \tag{D.67}$$

and apply Lemma 6.3, and to obtain the last inequality we have used that $\log(64) \geq 4$.

It remains to impose the constraints $0 \leq \delta_* \leq 4k$ on the minimizer δ_*, which we equivalently express in the form $0 \leq \omega_* \leq 16k^2$ using (D.67). Substituting (6.20) from Lemma 6.3 for ω_* and using a and b from (D.67), the constraint $\omega_* \geq 0$ imposes (6.31a), viz.,

$$\frac{p_0 n}{64} \geq \frac{1}{(|\mathcal{X}| + |\mathcal{Y}|)}.$$

Meanwhile, the constraint $\omega_* \leq 16k^2$ imposes (6.31b), viz.,

$$\frac{p_0 n}{4} \geq \log\left(\frac{p_0 n}{64}(|\mathcal{X}| + |\mathcal{Y}|)\right).$$

■

D.6 Proof of Proposition 6.7

To obtain Proposition 6.7, we adapt the proof of Proposition 6.1, replacing the use of the Frobenius norm of Lemma D.2 with the following spectral norm bound.

Lemma D.7. Given dimensions k_1 and k_2 and any matrices $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{k_1 \times k_2}$, we have, for every $k \in \{1, \dots, \min\{k_1, k_2\}\}$,

$$\sum_{i=1}^{k} |\sigma_i(\mathbf{A}_1) - \sigma_i(\mathbf{A}_2)| \leq k \, \|\mathbf{A}_1 - \mathbf{A}_2\|_{\mathrm{s}}. \tag{D.68}$$

Proof of Lemma D.7. We have

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

where the second inequality follows from Lemma D.3. ■

In particular, to establish Proposition 6.7, we replace (D.14)–(D.16) in Appendix D.1 with

$$\mathbb{P}\left(\sum_{i=1}^{k}|\hat{\sigma}_i - \sigma_i| \geq \delta\right) \leq \mathbb{P}\left(\|\hat{\mathbf{B}} - \tilde{\mathbf{B}}\|_{\mathrm{s}} \geq \frac{\delta}{k}\right) \tag{D.69}$$

$$= \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{Z}}_i\right\|_{\mathrm{s}} \geq \frac{\delta}{k}\right) \tag{D.70}$$

$$\leq (|\mathcal{X}| + |\mathcal{Y}|)\exp\left\{-\frac{p_0\,\delta^2 n}{4k^2}\right\}, \quad 0 \leq \delta \leq k, \tag{D.71}$$

where to obtain (D.69) we use Lemma D.7, then, as in the proof of Proposition 6.5, to obtain (D.70) we use (D.17), and to obtain (D.71) we use Lemma D.4 with (D.49) and (D.55) providing c and $\bar{c}$, respectively, and that $p_0 \leq 1/2$. ■

D.7 Proof of Proposition 6.8

We adapt the proof of Proposition 6.5, replacing the use of the spectral norm bound of Lemma D.5 with the following Frobenius norm bound.

Lemma D.8. Given $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{k_1 \times k_2}$ and $k \in \{1, \dots, \min\{k_1, k_2\}\}$, we have

$$0 \leq \|\mathbf{A}_1 \mathbf{\Psi}_{(k)}^{\mathbf{A}_1}\|_{\mathrm{F}}^2 - \|\mathbf{A}_1 \mathbf{\Psi}_{(k)}^{\mathbf{A}_2}\|_{\mathrm{F}}^2 \leq 4\sqrt{k}\,\|\mathbf{A}_1\|_{\mathrm{s}}\,\|\mathbf{A}_1 - \mathbf{A}_2\|_{\mathrm{F}}, \tag{D.72}$$

where $\mathbf{\Psi}_{(k)}^{\mathbf{A}}$ is as defined in Lemma D.5.

Proof of Lemma D.8. First, reproducing (D.40) from the proof of Lemma D.5, we have

$$\|\mathbf{A}_1 \mathbf{\Psi}_{(k)}^{\mathbf{A}_1}\|_{\mathrm{F}}^2 - \|\mathbf{A}_1 \mathbf{\Psi}_{(k)}^{\mathbf{A}_2}\|_{\mathrm{F}}^2$$
$$\leq 2\,\|\mathbf{A}_1\|_{\mathrm{s}}\left(\sum_{i=1}^{k}|\sigma_i(\mathbf{A}_1) - \sigma_i(\mathbf{A}_2)| + \sum_{i=1}^{k}\|(\mathbf{A}_1 - \mathbf{A}_2)\boldsymbol{\psi}_i^{\mathbf{A}_2}\|\right). \tag{D.73}$$

For the second sum in (D.73), we have

$$\begin{aligned}\left(\sum_{i=1}^{k}\left\|(\mathbf{A}_1-\mathbf{A}_2)\boldsymbol{\psi}_i^{\mathbf{A}_2}\right\|\right)^2 &\leq k\sum_{i=1}^{k}\left\|(\mathbf{A}_1-\mathbf{A}_2)\boldsymbol{\psi}_i^{\mathbf{A}_2}\right\|^2 &\text{(D.74)}\\ &= k\left\|(\mathbf{A}_1-\mathbf{A}_2)\boldsymbol{\Psi}_{(k)}^{\mathbf{A}_2}\right\|_{\mathrm{F}}^2\\ &\leq k\left\|\mathbf{A}_1-\mathbf{A}_2\right\|_{\mathrm{F}}^2, &\text{(D.75)}\end{aligned}$$

where to obtain (D.74) we use the Cauchy-Schwarz inequality, and to obtain (D.75) we use Lemma 3.1, recognizing that the right-hand side of (3.1) is upper bounded by $\|\mathbf{A}\|_{\mathrm{F}}^2$. Hence, using Lemma D.2 to bound the first term in (D.73), and (D.75) to bound the second, we obtain (D.72). ■

To establish Proposition 6.8, starting from (D.44) in Appendix D.4, but using Lemma D.8 instead of Lemma D.5, the bound (D.46) becomes

$$\begin{aligned}&\mathbb{E}_{P_Y}\left[\left\|\mathbb{E}_{P_{X|Y}}\left[f_*^k(X)\right]\right\|^2-\left\|\mathbb{E}_{P_{X|Y}}\left[\check{f}_*^k(X)\right]\right\|^2\right]\\ &\qquad= \left\|\tilde{\mathbf{B}}\,\boldsymbol{\Psi}_{(k)}^X\right\|_{\mathrm{F}}^2-\left\|\tilde{\mathbf{B}}\,\hat{\boldsymbol{\Psi}}_{(k)}^X\right\|_{\mathrm{F}}^2\\ &\qquad\leq 4\sqrt{k}\left\|\tilde{\mathbf{B}}-\hat{\mathbf{B}}\right\|_{\mathrm{F}}. \qquad\text{(D.76)}\end{aligned}$$

In turn, (D.56)–(D.59) then becomes [cf. (6.29)]

$$\begin{aligned}&\mathbb{P}_{\check{f}_*^k}\left(\mathbb{E}_{P_Y}\left[\left\|\mathbb{E}_{P_{X|Y}}\left[f_*^k(X)\right]\right\|^2-\left\|\mathbb{E}_{P_{X|Y}}\left[\check{f}_*^k(X)\right]\right\|^2\right]\geq\delta\right)\\ &\quad\leq\mathbb{P}\left(\left\|\tilde{\mathbf{B}}-\hat{\mathbf{B}}\right\|_{\mathrm{F}}\geq\frac{\delta}{4\sqrt{k}}\right) &\text{(D.77)}\\ &\quad=\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{Z}}_i\right\|_{\mathrm{s}}\geq\frac{\delta}{4\sqrt{k}}\right) &\text{(D.78)}\\ &\quad\leq\exp\left\{\frac{1}{4}-\frac{p_0^2\,\delta^2 n}{128k}\right\},\quad 0\leq\delta\leq(4/p_0)\sqrt{k/2}, &\text{(D.79)}\end{aligned}$$

where, as in the proof of Proposition 6.1, to obtain (D.78) we use (D.17), and to obtain (D.79) we use Lemma D.1, with (D.8) and (D.13) again providing c and $\bar{c}$, respectively. ■

D.8 Analysis of the Feature Quality Measure (6.35)

To begin, we have

$$
\begin{aligned}
&\big\|\mathbb{E}[f_*^k(X)\,g_*^k(Y)^{\mathrm{T}}] - \mathbb{E}[\check{f}_*^k(X)\,\check{g}_*^k(Y)^{\mathrm{T}}]\big\|_{\mathrm{F}} \\
&\quad= \big\|\big(\boldsymbol{\Psi}_{(k)}^Y\big)^{\mathrm{T}}\tilde{\mathbf{B}}\boldsymbol{\Psi}_{(k)}^X - \big(\hat{\boldsymbol{\Psi}}_{(k)}^Y\big)^{\mathrm{T}}\tilde{\mathbf{B}}\hat{\boldsymbol{\Psi}}_{(k)}^X\big\|_{\mathrm{F}} \\
&\quad= \big\|\big(\boldsymbol{\Psi}_{(k)}^Y\big)^{\mathrm{T}}\tilde{\mathbf{B}}\boldsymbol{\Psi}_{(k)}^X - \big(\hat{\boldsymbol{\Psi}}_{(k)}^Y\big)^{\mathrm{T}}\hat{\mathbf{B}}\hat{\boldsymbol{\Psi}}_{(k)}^X \\
&\qquad\qquad - \big(\hat{\boldsymbol{\Psi}}_{(k)}^Y\big)^{\mathrm{T}}\tilde{\mathbf{B}}\hat{\boldsymbol{\Psi}}_{(k)}^X + \big(\hat{\boldsymbol{\Psi}}_{(k)}^Y\big)^{\mathrm{T}}\hat{\mathbf{B}}\hat{\boldsymbol{\Psi}}_{(k)}^X\big\|_{\mathrm{F}} \\
&\quad= \big\|\big(\boldsymbol{\Sigma}_{(k)} - \hat{\boldsymbol{\Sigma}}_{(k)}\big) - \big(\hat{\boldsymbol{\Psi}}_{(k)}^Y\big)^{\mathrm{T}}(\tilde{\mathbf{B}} - \hat{\mathbf{B}})\hat{\boldsymbol{\Psi}}_{(k)}^X\big\|_{\mathrm{F}},
\end{aligned}
\tag{D.80}
$$

where $\hat{\boldsymbol{\Sigma}}_{(k)}$ is a diagonal matrix whose diagonal elements are $\hat{\sigma}_1, \dots, \hat{\sigma}_k$.

This measure exhibits similar sample complexity behavior to that obtained in Section 6.2.2. To see this, note that in this case we have

$$
\begin{aligned}
&\big\|\big(\boldsymbol{\Sigma}_{(k)} - \hat{\boldsymbol{\Sigma}}_{(k)}\big) - \big(\hat{\boldsymbol{\Psi}}_{(k)}^Y\big)^{\mathrm{T}}(\tilde{\mathbf{B}} - \hat{\mathbf{B}})\hat{\boldsymbol{\Psi}}_{(k)}^X\big\|_{\mathrm{F}} \\
&\quad\le \big\|\boldsymbol{\Sigma}_{(k)} - \hat{\boldsymbol{\Sigma}}_{(k)}\big\|_{\mathrm{F}} + \underbrace{\big\|\hat{\boldsymbol{\Psi}}_{(k)}^Y\big\|_{\mathrm{s}}}_{=1}\big\|(\tilde{\mathbf{B}} - \hat{\mathbf{B}})\hat{\boldsymbol{\Psi}}_{(k)}^X\big\|_{\mathrm{F}},
\end{aligned}
\tag{D.81}
$$

where we have used, in turn, the triangle inequality for the Frobenius norm, and Fact 5.12, and where the spectral norm is unity because $\hat{\boldsymbol{\Psi}}_{(k)}^Y$ has orthonormal columns. Moreover, the first term in (D.81) satisfies, using Lemma D.7,

$$
\big\|\boldsymbol{\Sigma}_{(k)} - \hat{\boldsymbol{\Sigma}}_{(k)}\big\|_{\mathrm{F}} \le \sum_{i=1}^{k} |\sigma_i - \hat{\sigma}_i| \le k\,\big\|\tilde{\mathbf{B}} - \hat{\mathbf{B}}\big\|_{\mathrm{s}},
\tag{D.82}
$$

while the remaining term satisfies

$$
\big\|(\tilde{\mathbf{B}} - \hat{\mathbf{B}})\hat{\boldsymbol{\Psi}}_{(k)}^X\big\|_{\mathrm{F}} \le \sqrt{\sum_{i=1}^{k} \sigma_i(\tilde{\mathbf{B}} - \hat{\mathbf{B}})^2}
\tag{D.83}
$$

$$
\le \sum_{i=1}^{k} \sigma_i(\tilde{\mathbf{B}} - \hat{\mathbf{B}})
\tag{D.84}
$$

$$
\le k\,\big\|\tilde{\mathbf{B}} - \hat{\mathbf{B}}\big\|_{\mathrm{s}},
\tag{D.85}
$$

where to obtain (D.83) we have used Lemma 3.1, and to obtain (D.84) we have used (6.23). Using (D.82) and (D.85) in (D.81), and, in turn, (D.80) yields

$$\left\|\mathbb{E}[f_*^k(X)\, g_*^k(Y)^{\mathrm{T}}] - \mathbb{E}[\check{f}_*^k(X)\, \check{g}_*^k(Y)^{\mathrm{T}}]\right\|_{\mathrm{F}} \leq 2k \left\|\tilde{\mathbf{B}} - \hat{\mathbf{B}}\right\|_{\mathrm{s}}. \tag{D.86}$$

Thus, we obtain a bound of the same form (to within a factor of two) as that for the measure (6.27), for which we obtained

$$\mathbb{E}_{P_Y}\left[\left\|\mathbb{E}_{P_{X|Y}}[f_*^k(X)]\right\|^2 - \left\|\mathbb{E}_{P_{X|Y}}[\check{f}_*^k(X)]\right\|^2\right] \leq 4k \left\|\tilde{\mathbf{B}} - \hat{\mathbf{B}}\right\|_{\mathrm{s}}. \tag{D.87}$$

As such analogous sample complexity bounds follow.

Finally, as in Section 6.2.3, we can similarly replace the use of the spectral norm with the Frobenius norm. In particular, (D.82) can be replaced with

$$\left\|\mathbf{\Sigma}_{(k)} - \hat{\mathbf{\Sigma}}_{(k)}\right\|_{\mathrm{F}} \leq \sum_{i=1}^{k} |\sigma_i - \hat{\sigma}_i| \leq \sqrt{k}\left\|\tilde{\mathbf{B}} - \hat{\mathbf{B}}\right\|_{\mathrm{F}}, \tag{D.88}$$

where we now use Lemma D.2 instead of Lemma D.7. Using (D.88), and the simple upper bound

$$\left\|(\tilde{\mathbf{B}} - \hat{\mathbf{B}})\hat{\mathbf{\Psi}}_{(k)}^{X}\right\|_{\mathrm{F}} \leq \left\|\tilde{\mathbf{B}} - \hat{\mathbf{B}}\right\|_{\mathrm{F}} \underbrace{\left\|\hat{\mathbf{\Psi}}_{(k)}^{X}\right\|_{\mathrm{s}}}_{=1} \leq \left\|\tilde{\mathbf{B}} - \hat{\mathbf{B}}\right\|_{\mathrm{F}} \tag{D.89}$$

instead of (D.85), in (D.81) yields

$$\begin{aligned}\left\|\mathbb{E}[f_*^k(X)\, g_*^k(Y)^{\mathrm{T}}] - \mathbb{E}[\check{f}_*^k(X)\, \check{g}_*^k(Y)^{\mathrm{T}}]\right\|_{\mathrm{F}} &\leq (1+\sqrt{k})\left\|\tilde{\mathbf{B}} - \hat{\mathbf{B}}\right\|_{\mathrm{F}} \\ &\leq 2\sqrt{k}\left\|\tilde{\mathbf{B}} - \hat{\mathbf{B}}\right\|_{\mathrm{F}},\end{aligned} \tag{D.90}$$

the second (looser) inequality of which matches (to within a factor of two) that for the measure (6.27), for which we obtained

$$\mathbb{E}_{P_Y}\left[\left\|\mathbb{E}_{P_{X|Y}}[f_*^k(X)]\right\|^2 - \left\|\mathbb{E}_{P_{X|Y}}[\check{f}_*^k(X)]\right\|^2\right] \leq 4\sqrt{k}\left\|\tilde{\mathbf{B}} - \hat{\mathbf{B}}\right\|_{\mathrm{F}}. \tag{D.91}$$

As such analogous sample complexity bounds follow in this form too.

D.9 Proof of Proposition 6.9

First, note that $\mathcal{S}_\delta^{\mathrm{F}}$, $\mathcal{S}_\delta^{\mathrm{s}}$, and $\mathcal{S}_\delta^{k}$ are non-empty as they contain $P_{X,Y}$, and bounded since $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ is bounded in $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$. In addition, our proof makes use of the following lemma.

Lemma D.9. For any $P_{X,Y} \in \operatorname{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$ and $\delta > 0$, let $\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y})$, $\mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y})$, and $\mathcal{S}_\delta^{k}(P_{X,Y})$ be as defined in (6.39). Then

P1. $\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y}), \mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y}) \subseteq \operatorname{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$ are compact sets for every $0 < \delta < B_{\min}(P_{X,Y})$, with $B_{\min}(\cdot)$ as defined in (6.41).

P2. $\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y}) \subseteq \mathcal{S}_{4\delta\sqrt{k}}^{k}(P_{X,Y})$ for every $\delta > 0$.

P3. $\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y}) \subseteq \mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y}) \subseteq \mathcal{S}_{4\delta k}^{k}(P_{X,Y})$ for every $\delta > 0$.

Proof of Lemma D.9. To establish property P1 for $\mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y})$, fix any $0 < \delta < B_{\min}(P_{X,Y})$, and consider the set[3]

$$\mathcal{M}_\delta^{\mathrm{s}}(P_{X,Y}) \triangleq \left\{ \mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} : \mathbf{M} \geq \mathbf{0},\ \|\mathbf{M}\|_{\mathrm{s}} = 1,\ \|\mathbf{M} - \mathbf{B}\|_{\mathrm{s}} \leq \delta \right\}.$$

We first show that $\mathcal{M}_\delta^{\mathrm{s}}(P_{X,Y})$ is closed. To this end, take any sequence $\{\mathbf{M}_n \in \mathcal{M}_\delta^{\mathrm{s}}(P_{X,Y}),\ n = 1, 2, \ldots\}$ such that $\mathbf{M}_n \to \mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ as $n \to \infty$. Then, clearly $\mathbf{M} \geq \mathbf{0}$ and $\|\mathbf{M}\|_{\mathrm{s}} = 1$ (by continuity of the spectral norm). Moreover, we have

$$\begin{aligned} \|\mathbf{M} - \mathbf{B}\|_{\mathrm{s}} &\leq \|\mathbf{M} - \mathbf{M}_n\|_{\mathrm{s}} + \|\mathbf{M}_n - \mathbf{B}\|_{\mathrm{s}} \\ &\leq \lim_{n\to\infty} \|\mathbf{M} - \mathbf{M}_n\|_{\mathrm{s}} + \delta \\ &\leq \delta, \end{aligned}$$

where the first inequality is the triangle inequality, the second inequality follows from using the fact that $\mathbf{M}_n \in \mathcal{M}_\delta^{\mathrm{s}}(P_{X,Y})$ and then letting $n \to \infty$, and the final inequality holds because $\mathbf{M}_n \to \mathbf{M}$. Hence, $\mathcal{M}_\delta^{\mathrm{s}}(P_{X,Y})$ is closed.

Next, we show that $\mathcal{M}_\delta^{\mathrm{s}}(P_{X,Y}) \subseteq \mathcal{B}_\circ^{\mathcal{X}\times\mathcal{Y}}$, with $\mathcal{B}_\circ^{\mathcal{X}\times\mathcal{Y}}$ as defined in (A.3). Due to (A.4), it suffices to show that $\mathbf{M} > \mathbf{0}$ for every $\mathbf{M} \in \mathcal{M}_\delta^{\mathrm{s}}(P_{X,Y})$, which we obtain by noting that for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have, with $M(x,y)$ denoting the (y,x)th entry of $\mathbf{M}$,

$$\begin{aligned} M(x,y) &\geq B(x,y) - |M(x,y) - B(x,y)| && \text{(D.92)} \\ &\geq B(x,y) - \|\mathbf{M} - \mathbf{B}\|_{\mathrm{s}} && \text{(D.93)} \\ &\geq B_{\min}(P_{X,Y}) - \delta && \text{(D.94)} \\ &> 0, && \text{(D.95)} \end{aligned}$$

[3] As in Appendix A.2, we use $\mathbf{A} \geq 0$ to denote that all the entries of $\mathbf{A}$ are nonnegative.

where to obtain (D.92) we have used the triangle inequality, to obtain (D.93) we have used that for an arbitrary matrix $\mathbf{A}$ with entries $a_{i,j}$,

$$|a_{i,j}| = |\mathbf{e}_i^{\mathrm{T}} \mathbf{A} \mathbf{e}_j| \leq \|\mathbf{e}_i\| \|\mathbf{A}\|_{\mathrm{s}} \|\mathbf{e}_j\| = \|\mathbf{A}\|_{\mathrm{s}}, \quad \text{all } i, j,$$

with the inequality due to Lemma 3.2, where to obtain (D.94) we have used that $\mathbf{M} \in \mathcal{M}_\delta^{\mathrm{s}}(P_{X,Y})$ and (6.41), and where to obtain (D.95) we have used the given constraint on δ.

Now via Proposition A.1 it follows that $\mathcal{S}_\delta(P_{X,Y})$ is the preimage of $\mathcal{M}_\delta^{\mathrm{s}}(P_{X,Y}) \subseteq \mathcal{B}_\circ^{\mathcal{X}\times\mathcal{Y}}$ under the DTM function $\mathbf{B}(\cdot)$, so $\mathcal{S}_\delta(P_{X,Y}) \subseteq \operatorname{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$. Furthermore, since, as shown in Proposition A.4, the restricted DTM function $\mathbf{B}\colon \operatorname{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}) \to \mathcal{B}_\circ^{\mathcal{X}\times\mathcal{Y}}$ is continuous, $\mathcal{S}_\delta(P_{X,Y})$ is closed because $\mathcal{M}_\delta^{\mathrm{s}}(P_{X,Y})$ is closed [235, Corollary, p. 87]. Since $\mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y})$ is also bounded, it is compact [235, Theorem 2.41].

Property P1 for $\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y})$ is obtained in a directly analogous manner. In particular, since the Frobenius norm is also continuous and satisfies the triangle inequality, it suffices to follow the same analysis, but now with respect to the set

$$\mathcal{M}_\delta^{\mathrm{F}}(P_{X,Y}) \triangleq \left\{\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} \colon \mathbf{M} \geq \mathbf{0},\ \|\mathbf{M}\|_{\mathrm{s}} = 1,\ \|\mathbf{M} - \mathbf{B}\|_{\mathrm{F}} \leq \delta\right\}.$$

To obtain property P2, we use the fact that [cf. (D.76)]

$$\left| \|\mathbf{B}\,\mathbf{\Psi}_{(k)}^X\|_{\mathrm{F}}^2 - \|\mathbf{B}\,\hat{\mathbf{\Psi}}_{(k)}^X\|_{\mathrm{F}}^2 \right| \leq 4\sqrt{k}\,\|\mathbf{B} - \hat{\mathbf{B}}\|_{\mathrm{F}},$$

which is obtained in precisely the same manner as (D.76), from which it follows immediately that $\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y}) \subseteq \mathcal{S}_{4\delta\sqrt{k}}^k(P_{X,Y})$.

Analogously, to obtain property P3, we use the fact that [cf. (D.46)]

$$\left| \|\mathbf{B}\,\mathbf{\Psi}_{(k)}^X\|_{\mathrm{F}}^2 - \|\mathbf{B}\,\hat{\mathbf{\Psi}}_{(k)}^X\|_{\mathrm{F}}^2 \right| \leq 4k\,\|\mathbf{B} - \hat{\mathbf{B}}\|_{\mathrm{s}},$$

which is obtained in precisely the same manner as (D.46), from which it follows immediately that $\mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y}) \subseteq \mathcal{S}_{4\delta k}^k(P_{X,Y})$. Finally, for the remaining part of property P3, we use the standard norm inequality $\|\mathbf{A}\|_{\mathrm{s}} \leq \|\mathbf{A}\|_{\mathrm{F}}$, for any matrix $\mathbf{A}$, obtaining $\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y}) \subseteq \mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y})$. ■

Proceeding to the proof of Proposition 6.9, with the notation (6.43)–

(6.45) we have, via Sanov's theorem [76, Theorem 2.1.10],

$$E_{-}(\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y})) = E_{*}(\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y})) \tag{D.96}$$

$$E_{-}(\mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y})) = E_{*}(\mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y})) \tag{D.97}$$

$$E_{-}(\mathcal{S}_\delta^{k}(P_{X,Y})) \leq E(\mathcal{S}_\delta^{k}(P_{X,Y})), \tag{D.98}$$

for $0 < \delta < B_{\min}(P_{X,Y})$ with $B_{\min}(\cdot)$ as defined in (6.41), where to obtain (D.96) and (D.97) we have used that $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}\backslash\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y})$ and $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}\backslash\mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y})$, respectively, are open sets (with respect to $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$), since $\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y})$ and $\mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y})$ are closed according to property P1 of Lemma D.9. Hence, (6.42a) follows according to

$$E(\mathcal{S}_{4\delta\sqrt{k}}^{k}(P_{X,Y})) \geq E_{-}(\mathcal{S}_{4\delta\sqrt{k}}^{k}(P_{X,Y})) \tag{D.99}$$

$$\geq E_{-}(\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y})) \tag{D.100}$$

$$= E_{*}(\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y})), \tag{D.101}$$

where to obtain (D.99) we have used (D.98), to obtain (D.100) we have used that

$$\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}\backslash\mathcal{S}_{4\delta\sqrt{k}}^{k} \subseteq \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}\backslash\mathcal{S}_\delta^{\mathrm{F}},$$

which follows from property P2 of Lemma D.9, and to obtain (D.101) we have used (D.96). Analogously, (6.42b) follows according to

$$E(\mathcal{S}_{4\delta k}^{k}(P_{X,Y})) \geq E_{-}(\mathcal{S}_{4\delta k}^{k}(P_{X,Y})) \tag{D.102}$$

$$\geq E_{-}(\mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y})) \tag{D.103}$$

$$= E_{*}(\mathcal{S}_\delta^{\mathrm{s}}(P_{X,Y})) \tag{D.104}$$

$$\geq E_{*}(\mathcal{S}_\delta^{\mathrm{F}}(P_{X,Y})), \tag{D.105}$$

where to obtain (D.102) we have used (D.98), to obtain (D.103) we have used the first subset relation in

$$\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}\backslash\mathcal{S}_{4\delta k}^{k} \subseteq \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}\backslash\mathcal{S}_\delta^{\mathrm{s}} \subseteq \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}\backslash\mathcal{S}_\delta^{\mathrm{F}}, \tag{D.106}$$

which follows from property P3 of Lemma D.9, to obtain (D.104) we have used (D.97), and to obtain (D.105) we have used (D.96) and the second subset relation in (D.106). ■

The version of record is available at: http://dx.doi.org/10.1561/0100000107

D.10 Proof of Lemma 6.10

Our proof makes use of the following special case of Sanov's Theorem [76, Theorem 2.1.10, Exercise 2.1.19], [71, Theorem 2.1]:

Lemma D.10. For every distribution $P_Z \in \mathcal{P}^{\mathcal{Z}}$, and every closed and convex subset $\mathcal{S} \subseteq \mathcal{P}^{\mathcal{Z}}$ of probability distributions that has non-empty interior, we have that the empirical distribution $\hat{P}_Z$ formed from n i.i.d. samples of P_Z satisfies

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\hat{P}_Z \in \mathcal{S}\Big) = -\min_{Q_Z \in \mathcal{S}} D(Q_Z \| P_Z),$$

where the minimum is achieved by a unique distribution.

Without loss of generality we may restrict our attention to the case in which $\mathbb{E}[h(Z)] > 0$. For any $\gamma > 0$, define the sets

$$\mathcal{S}_\gamma^+ \triangleq \Big\{Q_Z \in \mathcal{P}^{\mathcal{Z}} \colon \mathbb{E}_{Q_Z}[h(Z)] \geq (1+\gamma)\,\mathbb{E}[h(Z)]\Big\}$$
$$\mathcal{S}_\gamma^- \triangleq \Big\{Q_Z \in \mathcal{P}^{\mathcal{Z}} \colon \mathbb{E}_{Q_Z}[h(Z)] \leq (1-\gamma)\,\mathbb{E}[h(Z)]\Big\},$$

where

$$\mathbb{E}_{Q_Z}[h(Z)] \triangleq \sum_{z\in\mathcal{Z}} Q_Z(z)\,h(z).$$

Furthermore, since we will eventually let $\gamma \to 0$, we may assume that

$$0 < \gamma < \min\left\{\left(\frac{\max_{z\in\mathcal{Z}} h(z)}{\mathbb{E}[h(Z)]} - 1\right), \left(1 - \frac{\min_{z\in\mathcal{Z}} h(z)}{\mathbb{E}[h(Z)]}\right)\right\},$$

so that

$$\min_{z\in\mathcal{Z}} h(z) < (1-\gamma)\,\mathbb{E}[h(Z)] < (1+\gamma)\,\mathbb{E}[h(Z)] < \max_{z\in\mathcal{Z}} h(z),$$

where $\min_{z\in\mathcal{Z}} h(z) < \max_{z\in\mathcal{Z}} h(z)$ because $\mathrm{var}[h(Z)] > 0$. Hence, $\mathcal{S}_\gamma^+$ and $\mathcal{S}_\gamma^-$ are closed and convex sets that have non-empty interiors. Using Lemma D.10, we have

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\hat{P}_Z \in \mathcal{S}_\gamma^+\Big) = -\min_{Q_Z \in \mathcal{S}_\gamma^+} D(Q_Z \| P_Z) = -D(Q_Z^+ \| P_Z) \tag{D.107}$$

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\hat{P}_Z \in \mathcal{S}_\gamma^-\Big) = -\min_{Q_Z \in \mathcal{S}_\gamma^-} D(Q_Z \| P_Z) = -D(Q_Z^- \| P_Z), \tag{D.108}$$

where the (unique) minimizing distributions $Q_Z^+ \in \mathcal{S}_\gamma^+$ and $Q_Z^- \in \mathcal{S}_\gamma^-$ are members of the exponential family

$$Q_Z(z;\theta) = P_Z(z)\exp\{\theta\, h(z) - \alpha(\theta)\}, \quad z \in \mathcal{Z},$$

with natural parameter $\theta \in \mathbb{R}$. Recall that the (infinitely differentiable) log-partition function

$$\alpha(\theta) \triangleq \log(\mathbb{E}[\exp\{\theta\, h(Z)\}])$$

has derivatives[4]

$$\alpha'(\theta) = \mathbb{E}_{Q_Z(\cdot;\theta)}[h(Z)] \quad \text{and} \quad \alpha''(\theta) = \mathrm{var}_{Q_Z(\cdot;\theta)}[h(Z)] > 0,$$

where the second derivative is positive because every element of $\mathcal{Z}$ has positive probability under $Q_Z(\cdot;\theta)$. The minimizing distributions are $Q_Z^+ = Q_Z(\cdot;\theta_+)$ and $Q_Z^- = Q_Z(\cdot;\theta_-)$, where the optimal parameters $\theta_+ > 0$ and $\theta_- < 0$ are chosen to satisfy (cf. [71, Example 2.1])

$$\begin{aligned}\alpha'(\theta_+) &= \mathbb{E}_{Q_Z(\cdot;\theta_+)}[h(Z)] = (1+\gamma)\,\mathbb{E}[h(Z)],\\ \alpha'(\theta_-) &= \mathbb{E}_{Q_Z(\cdot;\theta_-)}[h(Z)] = (1-\gamma)\,\mathbb{E}[h(Z)].\end{aligned}$$

Now assume that

$$\lim_{\gamma\to 0^+} \frac{D(Q_Z^+ \| P_Z)}{\gamma^2} = \lim_{\gamma\to 0^+} \frac{D(Q_Z^- \| P_Z)}{\gamma^2} = \frac{1}{2}\,\frac{(\mathbb{E}[h(Z)])^2}{\mathrm{var}[h(Z)]} \tag{D.109}$$

and define

$$\mathcal{S}_\gamma^\pm \triangleq \mathcal{S}_\gamma^+ \cup \mathcal{S}_\gamma^- = \left\{ Q_Z \in \mathcal{P}^{\mathcal{Z}} \colon \left| \frac{\mathbb{E}_{Q_Z}[h(Z)]}{\mathbb{E}[h(Z)]} - 1 \right| \geq \gamma \right\}. \tag{D.110}$$

Since $\mathcal{S}_\gamma^+$ and $\mathcal{S}_\gamma^-$ are disjoint, we have

$$\mathbb{P}\Big(\hat{P}_Z \in \mathcal{S}_\gamma^\pm\Big) = \mathbb{P}\Big(\hat{P}_Z \in \mathcal{S}_\gamma^+\Big) + \mathbb{P}\Big(\hat{P}_Z \in \mathcal{S}_\gamma^-\Big).$$

Hence, via the Laplace principle it follows that

$$-\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\hat{P}_Z \in \mathcal{S}_\gamma^\pm\Big) = \min\Big\{ D(Q_Z^+ \| P_Z)\,,\, D(Q_Z^- \| P_Z) \Big\}, \tag{D.111}$$

[4] In this section we use $'$, $''$, and $'''$ as notation for first, second, and third derivatives.

where we have used (D.107) and (D.108). Applying (D.109) to (D.111), and recognizing (D.110), we obtain (6.46) as desired.

Thus, it remains only to show (D.109). To this end, consider the function

$$d(\theta) \triangleq D(Q_Z(\cdot;\theta)\|P_Z), \quad \theta \in \mathbb{R}.$$

It is straightforward to verify that

$$\begin{aligned} d(\theta) &= \theta\,\alpha'(\theta) - \alpha(\theta) \\ d'(\theta) &= \theta\,\alpha''(\theta) \\ d''(\theta) &= \alpha''(\theta) + \theta\,\alpha'''(\theta), \end{aligned}$$

which means that $d(0) = d'(0) = 0$, and $d''(0) = \alpha''(0) = \operatorname{var}[h(Z)]$. Hence, by Taylor's theorem we have

$$\lim_{\theta\to 0} \frac{d(\theta)}{\theta^2} = \frac{1}{2}\operatorname{var}[h(Z)]. \tag{D.112}$$

Now given any $\tau \in \mathbb{R}$, there exists a unique θ_τ such that

$$\alpha'(\theta_\tau) = \mathbb{E}_{Q_Z(\cdot;\theta_\tau)}[h(Z)] = (1+\tau)\,\mathbb{E}[h(Z)],$$

since α' is increasing (since α'' is positive). Next, observe that

$$\begin{aligned} \lim_{\tau\to 0} \frac{d(\theta_\tau)}{\tau^2} &= \lim_{\tau\to 0} \frac{d(\theta_\tau)}{\theta_\tau^2} \lim_{\tau\to 0} \frac{\theta_\tau^2}{\tau^2} \\ &= \frac{\operatorname{var}[h(Z)]}{2}\left(\lim_{\tau\to 0}\frac{\theta_\tau}{\tau}\right)^2 \\ &= \frac{\operatorname{var}[h(Z)]}{2}\left(\left.\frac{\mathrm{d}\theta_\tau}{\mathrm{d}\tau}\right|_{\tau=0}\right)^2 \\ &= \frac{\operatorname{var}[h(Z)]}{2}\left(\frac{\mathbb{E}[h(Z)]}{\alpha''(0)}\right)^2 \\ &= \frac{\mathbb{E}[h(Z)]^2}{2\operatorname{var}[h(Z)]}, \end{aligned} \tag{D.113}$$

where the second equality follows from (D.112), the fact that $\theta_\tau \to 0$ as $\tau \to 0$ (by the continuity of the inverse of $\alpha'(\cdot)$), and the continuity of $t \to t^2$, where the third equality follows from the definition of derivative and the fact that $\theta_0 = 0$ (since $\alpha'(0) = \mathbb{E}[h(Z)]$), where the fourth equality holds because

$$\left.\frac{\mathrm{d}\theta_\tau}{\mathrm{d}\tau}\right|_{\tau=0} = \left(\left.\frac{\mathrm{d}\tau_\theta}{\mathrm{d}\theta}\right|_{\theta=0} \right)^{-1}$$

with

$$\tau_\theta = \frac{\alpha'(\theta)}{\mathbb{E}[h(Z)]} - 1,$$

and where the fifth equality holds because $\alpha''(0) = \mathrm{var}[h(Z)]$.

In turn, setting $\tau = \gamma > 0$ and $\theta_\tau = \theta_+ > 0$ yields

$$\lim_{\gamma \to 0^+} \frac{D(Q_Z^+ \| P_Z)}{\gamma^2} = \lim_{\tau \to 0^+} \frac{d(\theta_\tau)}{\tau^2}, \tag{D.114a}$$

and setting $\tau = -\gamma < 0$ and $\theta_\tau = \theta_- < 0$ yields

$$\lim_{\gamma \to 0^+} \frac{D(Q_Z^- \| P_Z)}{\gamma^2} = \lim_{\tau \to 0^-} \frac{d(\theta_\tau)}{\tau^2}. \tag{D.114b}$$

Finally, replacing the right-hand sides of (D.114) with (D.113) yields (D.109). ■

E

Appendices for Section 7

E.1 Proof of Proposition 7.1

From (7.2), it follows immediately that

$$\begin{aligned} \mathbb{E}[M] &= \sum_{y \in \hat{\mathcal{Y}}(x)} \mathbb{P}(\mathcal{E}_y(x)) && \text{(E.1)} \\ &\leq \max_{\{\hat{\mathcal{Y}}(x) \subset \mathcal{Y}:\, |\hat{\mathcal{Y}}(x)| = l\}} \sum_{y \in \hat{\mathcal{Y}}(x)} \mathbb{P}(\mathcal{E}_y(x)) && \text{(E.2)} \\ &= \sum_{y \in \hat{\mathcal{Y}}^*(x)} \mathbb{P}(\mathcal{E}_y(x)), && \text{(E.3)} \end{aligned}$$

where

$$y_1^*(x) = \arg\max_{y \in \mathcal{Y}} \mathbb{P}(\mathcal{E}_y(x)) \tag{E.4a}$$

$$y_i^*(x) = \arg\max_{y \in \mathcal{Y} \setminus \{y_1^*(x), \ldots, y_{i-1}^*(x)\}} \mathbb{P}(\mathcal{E}_y(x)), \quad i \in \{2, \ldots, l\}. \tag{E.4b}$$

It remains only to evaluate the constituent event probabilities, which are obtained as follows:

$$\mathbb{P}(\mathcal{E}_y(x)) = \mathbb{P}\Big(V^k(y) = V_\circ^k(x)\Big)$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

$$
\begin{aligned}
&= \sum_{\{v^k, v_\circ^k :\, v^k = v_\circ^k\}} P_{V^k|Y}(v^k|y)\, P_{V^k|X}(v_\circ^k|x) \\
&= \sum_{v^k} P_{V^k|Y}(v^k|y)\, P_{V^k|X}(v^k|x),
\end{aligned}
$$

wherein, using (5.37b) and (5.38a),

$$
\begin{aligned}
&P_{V^k|Y}(v^k|y)\, P_{V^k|X}(v^k|x) \\
&\qquad = \prod_{i=1}^{k} \Big(P_{V_i|Y}(v_i|y)\, P_{V_i|X}(v_i|x) \Big) \\
&\qquad = \prod_{i=1}^{k} \left(\frac{P_{Y|V_i}(y|v_i)\, P_{V_i}(v_i)}{P_Y(y)}\, \frac{P_{X|V_i}(x|v_i)\, P_{V_i}(v_i)}{P_X(x)} \right) \\
&\qquad = \frac{1}{2^{2k}} \prod_{i=1}^{k} \Big((1 + \epsilon v_i\, g_i^*(y))\, (1 + \epsilon v_i\, \sigma_i\, f_i^*(x)) \Big) \\
&\qquad = \frac{1}{2^{2k}} \prod_{i=1}^{k} \Big(1 + \epsilon\, v_i (\sigma_i\, f_i^*(x) + g_i^*(y)) + \epsilon^2 \sigma_i\, f_i^*(x)\, g_i^*(y) \Big).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\mathbb{P}(\mathcal{E}_y(x)) &= \frac{1}{2^{2k}} \prod_{i=1}^{k} \sum_{v_i} \Big(1 + \epsilon\, v_i (\sigma_i\, f_i^*(x) + g_i^*(y)) + \epsilon^2 \sigma_i\, f_i^*(x)\, g_i^*(y) \Big) \\
&= \frac{1}{2^k} \prod_{i=1}^{k} (1 + \epsilon^2 \sigma_i\, f_i^*(x)\, g_i^*(y)) \\
&= \frac{1}{2^k} \left(1 + \epsilon^2 \sum_{i=1}^{k} \sigma_i\, f_i^*(x)\, g_i^*(y) \right) + o(\epsilon^2),
\end{aligned}
$$

the nonvanishing term of which we note is a monotonic function of the quantity being maximized in (7.6). ■

F

Appendices for Section 8

F.1 Proof of Proposition 8.1

First, without loss of generality we impose the constraints

$$\mathbb{E}\big[g(Y)\big] = 0 \qquad \text{and} \qquad \mathbb{E}\big[\beta(y)\big] = 0, \tag{F.1}$$

since other solutions are simple reparameterizations.

It is convenient to first establish the following special case of Proposition 8.1.

Lemma F.1. Let the hypotheses of Proposition 8.1 be satisfied, together with the further constraints

$$\boldsymbol{\mu}_S = \mathbf{0} \qquad \text{and} \qquad \boldsymbol{\Lambda}_S = \mathbf{I}. \tag{F.2}$$

Then

$$\begin{aligned}\min_{\tilde{P}_{Y|S}(\cdot|s)\in\tilde{\mathcal{P}}^{\mathcal{Y}}_s(P_Y)} &\sum_{s\in\mathcal{S}} P_S(s)\, D\big(P_{Y|S}(\cdot|s) \,\|\, \tilde{P}_{Y|S}(\cdot|s)\big) \\ &= I(Y;S) - \frac{1}{2}\mathbb{E}\Big[\big\|\boldsymbol{\mu}_{S|Y}(Y)\big\|^2\Big] + o(\epsilon^2),\end{aligned} \tag{F.3}$$

and is achieved by the parameters

$$g(y) = g_{*,S}(y) \triangleq \boldsymbol{\mu}_{S|Y}(y) + o(\epsilon) \text{ and } \beta(y) = \beta_{*,S}(y) \triangleq o(\epsilon), \tag{F.4a}$$

i.e.,

$$\tilde{P}^{*}_{Y|S}(y|s) \propto P_Y(y)\exp\Bigl\{s^{\mathrm{T}}\boldsymbol{\mu}_{S|Y}(y)\Bigr\}(1+o(1)). \tag{F.4b}$$

Using Lemma F.1, we establish Proposition 8.1 as follows. First, let us assume $\boldsymbol{\Lambda}_S$ is nonsingular, and let

$$\tilde{s} \triangleq \boldsymbol{\Lambda}_S^{-1/2}(s-\boldsymbol{\mu}_S) \tag{F.5}$$

so

$$\boldsymbol{\mu}_{\tilde{S}} = \mathbf{0} \qquad \text{and} \qquad \boldsymbol{\Lambda}_{\tilde{S}} = \mathbf{I}.$$

Then we may rewrite $\tilde{P}^{g,\beta}_{Y|S}(y|s)$ in the form

$$\tilde{P}^{g,\beta}_{Y|S}(y|s) = P_Y(y)\exp\Bigl\{\tilde{s}^{\mathrm{T}}\tilde{g}(y)+\tilde{\beta}(y)-\tilde{\alpha}(\tilde{s})\Bigr\} \tag{F.6}$$

$$\triangleq \tilde{P}^{\tilde{g},\tilde{\beta}}_{Y|\tilde{S}}(y|\tilde{s}) \tag{F.7}$$

where

$$\tilde{g}(y) \triangleq \boldsymbol{\Lambda}_S^{1/2} g(y) \tag{F.8}$$

$$\tilde{\beta}(y) \triangleq \boldsymbol{\mu}_S^{\mathrm{T}} g(y) + \beta(y) \tag{F.9}$$

$$\tilde{\alpha}(\tilde{s}) \triangleq \alpha(\boldsymbol{\mu}_S + \boldsymbol{\Lambda}_S^{1/2}\tilde{s}), \tag{F.10}$$

and for which $\mathbb{E}\bigl[\tilde{g}(Y)\bigr] = \mathbf{0}$ and $\mathbb{E}\bigl[\tilde{\beta}(Y)\bigr] = 0$.

Using these definitions, we have

$$\tilde{g}_{*,S}(y) = \mathbb{E}_{P_{\tilde{S}|Y}(\cdot|y)}\bigl[\tilde{S}\bigr] + o(\epsilon) = \boldsymbol{\Lambda}_S^{-1/2}(\boldsymbol{\mu}_{S|Y}(y)-\boldsymbol{\mu}_S) + o(\epsilon), \tag{F.11}$$

where to obtain the first equality we have used Lemma F.1, and to obtain (F.11) we have used (F.5). Combining (F.11) with (F.8) yields (8.4a). Similarly, via Lemma F.1 we obtain

$$\tilde{\beta}_{*,S}(y) = o(\epsilon),$$

which when combined with (F.9) yields (8.4b). In turn, we obtain (8.2) via

$$\begin{aligned}\min_{\tilde{P}_{Y|\tilde{S}}(\cdot|s)\in\tilde{\mathcal{P}}^{\mathcal{Y}}_{\tilde{s}}(P_Y)} & \sum_{\tilde{s}\in\mathcal{S}} P_{\tilde{S}}(\tilde{s})\, D(P_{Y|\tilde{S}}(\cdot|\tilde{s}) \,\|\, \tilde{P}_{Y|\tilde{S}}(\cdot|\tilde{s})) \\ &= I(Y;\tilde{S}) - \frac{1}{2}\mathbb{E}\Bigl[\bigl\|\boldsymbol{\mu}_{\tilde{S}|Y}(Y)\bigr\|^2\Bigr] + o(\epsilon^2) \\ &= I(Y;S) - \frac{1}{2}\mathbb{E}\Bigl[\bigl\|\boldsymbol{\Lambda}_S^{-1/2}(\boldsymbol{\mu}_{S|Y}(Y)-\boldsymbol{\mu}_S)\bigr\|^2\Bigr] + o(\epsilon^2),\end{aligned}$$

where to obtain the first equality we have used Lemma F.1, and to obtain the second we have used (F.5) and the invariance of mutual information to coordinate transformations.

It remains only to establish Lemma F.1.

Proof of Lemma F.1. First, note that

$$D(P_{Y,S}\|\tilde{P}_{Y|S}P_S) = I(Y;S) - \underbrace{\mathbb{E}\big[S^{\mathrm{T}}g(Y) - \alpha(S)\big]}_{\triangleq \tilde{\ell}(g,\beta)}, \tag{F.12}$$

so we seek to maximize $\tilde{\ell}(g,\beta)$. Moreover, note that since χ^2-divergence is an f-divergence, it satisfies a data processing inequality [69], so S, Y are ϵ-dependent for any choice of f that induces S.

Fixing $s \in \mathcal{S}$, note that $P_{Y|S}(\cdot|s) \in \mathcal{N}_\epsilon^{\mathcal{Y}}(P_Y)$ and, in addition, $\tilde{P}_{Y|S}^{0,0}(\cdot|s) = P_Y$. As a result, it follows that the optimizing $\tilde{P}_{Y|S}^{g,\beta}(\cdot|s)$ is such that $\tilde{P}_{Y|S}^{g,\beta}(\cdot|s) \in \mathcal{N}_\epsilon^{\mathcal{Y}}(P_Y)$, and thus we may restrict our search to parameters (g,β) in this neighborhood.

In turn, defining

$$\tilde{\phi}_s^{Y|S}(y) \triangleq \frac{\tilde{P}_{Y|S}^{g,\beta}(y|s) - P_Y(y)}{\epsilon\sqrt{P_Y(y)}},$$

it follows from (8.1) that

$$\epsilon^2 \sum_{y\in\mathcal{Y}} \tilde{\phi}_s^{Y|S}(y)^2 = \sum_{y\in\mathcal{Y}} P_Y(y)\Big(\exp\Big\{s^{\mathrm{T}}g(y) + \beta(y) - \alpha(s)\Big\} - 1\Big)^2 \le \epsilon^2,$$

so

$$s^{\mathrm{T}}g(y) + \beta(y) - \alpha(s) = o(1).$$

Hence, for the Taylor series expansions[1] (in ϵ)

$$g(y) = \sum_{i=0}^{2} \epsilon^i g^{(i)}(y) + o(\epsilon^2) \tag{F.13a}$$

$$\beta(y) = \sum_{i=0}^{2} \epsilon^i \beta^{(i)}(y) + o(\epsilon^2) \tag{F.13b}$$

$$\alpha(s) = \sum_{i=0}^{2} \epsilon^i \alpha^{(i)}(s) + o(\epsilon^2), \tag{F.13c}$$

[1] In this analysis we assume the existence of these Taylor series. Moreover, we use superscript notation $^{(i)}$ to denote the ith derivative (with respect to ϵ.)

wherein $g^{(i)}(y)$, $\beta^{(i)}(y)$, and $\alpha^{(i)}(s)$ for $i \in \{0, 1, 2\}$ do not depend on ϵ, it follows that

$$s^{\mathrm{T}} g^{(0)}(y) + \beta^{(0)}(y) = \alpha^{(0)}(s). \tag{F.14}$$

But due to (F.1), in the Taylor series (F.13) we must also have

$$\mathbb{E}[g^{(i)}(Y)] = 0 \quad \text{and} \quad \mathbb{E}[\beta^{(i)}(Y)] = 0, \qquad i \in \{0, 1, 2\}. \tag{F.15}$$

Taking the expectation of both sides of (F.14) with respect to P_Y then yields that

$$\alpha^{(0)}(s) = 0. \tag{F.16}$$

Next, with

$$\tau_s^{(i)}(y) \triangleq s^{\mathrm{T}} g^{(i)}(y) + \beta^{(i)}(y), \qquad i = 1, 2, \tag{F.17}$$

and using the Taylor series

$$\mathrm{e}^{\omega} = \sum_{j=0}^{l} \frac{1}{j!}\, \omega^j + o(\omega^l),$$

we obtain that $Z(s) \triangleq \mathrm{e}^{\alpha(s)}$, via (8.1), can be expressed in the form

$$\begin{aligned} Z(s) &= \sum_{y \in \mathcal{Y}} P_Y(y) \exp\Big\{ s^{\mathrm{T}} g(y) + \beta(y) \Big\} \\ &= \mathbb{E}_{P_Y}\left[\exp\left\{ \sum_{i=1}^{2} \epsilon^i \tau_s^{(i)}(Y) + o(\epsilon^2) \right\} \right] \\ &= \mathbb{E}_{P_Y}\left[\left(\sum_{j=0}^{2} \epsilon^j \frac{1}{j!} \tau_s^{(1)}(Y)^j + o(\epsilon^2) \right) \right. \\ &\qquad \left. \cdot \left(\sum_{j=0}^{1} \epsilon^{2j} \frac{1}{j!} \tau_s^{(2)}(Y)^j + o(\epsilon^2) \right) \Big(1 + o(\epsilon^2) \Big) \right] \\ &= 1 + \sum_{i=1}^{2} \epsilon^i \upsilon_i(s) + o(\epsilon^2), \end{aligned}$$

with

$$\upsilon_1(s) \triangleq \mathbb{E}_{P_Y}\big[\tau_s^{(1)}(Y)\big] = 0 \tag{F.18a}$$

$$\upsilon_2(s) \triangleq \mathbb{E}_{P_Y}\left[\frac{1}{2}\, \tau_s^{(1)}(Y)^2 + \tau_s^{(2)}(Y) \right] = \frac{1}{2}\, \mathbb{E}_{P_Y}\big[\tau_s^{(1)}(Y)^2\big], \tag{F.18b}$$

where we have used (F.14) with (F.16) to conclude $\tau_s^{(0)} = 0$, and that

$$\mathbb{E}_{P_Y}\big[\tau_s^{(i)}(Y)\big] = 0, \qquad i = 1, 2, \dots$$

due to (F.15).

Next, using the Taylor series

$$\log(1+\omega) = \omega - \frac{1}{2}\,\omega^2 + o(\omega^2),$$

we obtain that $\alpha(s) = \log Z(s)$ is of the form

$$\alpha(s) = \left(\sum_{i=1}^{2} \epsilon^i \upsilon_i(s)\right) - \frac{1}{2}\left(\sum_{i=1}^{2} \epsilon^i \upsilon_i(s)\right)^2 + o(\epsilon^2).$$

So in the Taylor series (F.13c) for $\alpha(s)$, we obtain

$$\alpha^{(1)}(s) = \upsilon_1(s) = 0 \tag{F.19a}$$

$$\alpha^{(2)}(s) = \upsilon_2(s) - \frac{1}{2}\,\upsilon_1(s)^2 = \upsilon_2(s). \tag{F.19b}$$

We write $\tilde{\ell}(g, \beta)$ in (F.12), which we seek to maximize, in the form

$$\tilde{\ell}(g,\beta) = \sum_{i=0}^{2} \epsilon^i\, \mathbb{E}\big[S^{\mathrm{T}} g^{(i)}(Y) - \alpha^{(i)}(S)\big] + o(\epsilon^2) \tag{F.20}$$

$$= \sum_{i=1}^{2} \epsilon^i\, \mathbb{E}\big[S^{\mathrm{T}} g^{(i)}(Y) - \alpha^{(i)}(S)\big] + o(\epsilon^2) \tag{F.21}$$

$$\begin{aligned} &= \epsilon\, \mathbb{E}\big[S^{\mathrm{T}} g^{(1)}(Y)\big] \\ &\qquad + \epsilon^2\, \mathbb{E}\big[S^{\mathrm{T}} g^{(2)}(Y)\big] - \epsilon^2\, \mathbb{E}\big[\alpha^{(2)}(S)\big] + o(\epsilon^2) \end{aligned} \tag{F.22}$$

$$= \epsilon \mathbb{E}\big[S^{\mathrm{T}} g^{(1)}(Y)\big] - \epsilon^2\, \mathbb{E}\big[\alpha^{(2)}(S)\big] + o(\epsilon^2), \tag{F.23}$$

where to obtain (F.20) we have used (F.13), to obtain (F.21) we have used that

$$\mathbb{E}\big[S^{\mathrm{T}} g^{(0)}(Y) - \alpha^{(0)}(S)\big] = -\mathbb{E}\big[\beta^{(0)}(Y)\big] = 0,$$

due to (F.14) and (F.15), to obtain (F.22) we have used (F.19a), and to obtain (F.23) we have used that the second term in (F.22) is $o(\epsilon^2)$, which follows from the fact that for any i

$$\mathbb{E}\big[S^{\mathrm{T}} g^{(i)}(Y)\big] = \mathbb{E}\big[S^{\mathrm{T}}\big]\, \underbrace{\mathbb{E}\big[g^{(i)}(Y)\big]}_{=0} + \mathcal{O}(\epsilon) \in \mathcal{O}(\epsilon),$$

since $P_{S,Y} \in \mathcal{N}_{\epsilon}^{\mathcal{X}\times\mathcal{Y}}(P_S P_Y)$.

Hence, we write (F.23) in the form

$$\tilde{\ell}(g,\beta) = \tilde{\ell}_2(g^{(1)},\beta^{(1)}) + o(\epsilon^2), \tag{F.24a}$$

with

$$\tilde{\ell}_2(g^{(1)},\beta^{(1)}) \triangleq \epsilon\mathbb{E}[S^{\mathrm{T}} g^{(1)}(Y)] - \epsilon^2\,\mathbb{E}[\alpha^{(2)}(S)], \tag{F.24b}$$

where we note $\tilde{\ell}_2(g^{(1)},\beta^{(1)}) \in \mathcal{O}(\epsilon^2)$. In addition, we note that there is no dependence on $g^{(0)}$ and $\beta^{(0)}$ in (F.24b). Indeed, they can be freely chosen subject to the constraints (F.15), and those choices have no effect on the resulting $\tilde{P}_{Y|S}^{g,\beta}(\cdot|s)$; for example, we may choose

$$g^{(0)}(y) = 0 \quad \text{and} \quad \beta^{(0)}(y) = 0, \qquad \text{all } y \in \mathcal{Y}.$$

Proceeding, to express the second term in (F.24b) in terms of $g^{(1)}(y)$ and $\beta^{(1)}(y)$, note that, using (F.18b) and (F.17),

$$\begin{aligned} \upsilon_2(s) &= \frac{1}{2}\,\mathbb{E}_{P_Y}\left[\tau_s^{(1)}(Y)^2\right] \\ &= \frac{1}{2}\,\mathbb{E}_{P_Y}\left[(s^{\mathrm{T}} g^{(1)}(Y) + \beta^{(1)}(Y))^2\right], \end{aligned} \tag{F.25}$$

so

$$\epsilon^2\,\mathbb{E}_{P_S}[\alpha^{(2)}(S)] = \epsilon^2\,\mathbb{E}_{P_S}[\upsilon_2(S)] \tag{F.26}$$

$$= \frac{\epsilon^2}{2}\,\mathbb{E}_{P_Y}\left[\mathbb{E}_{P_S}\left[(S^{\mathrm{T}} g^{(1)}(Y))^2\right] + \beta^{(1)}(Y)^2\right] \tag{F.27}$$

$$= \frac{\epsilon^2}{2}\,\mathbb{E}_{P_Y}\left[\|g^{(1)}(Y)\|^2\right] + \frac{\epsilon^2}{2}\,\mathbb{E}_{P_Y}\left[\beta^{(1)}(Y)^2\right], \tag{F.28}$$

where to obtain (F.26) we have used (F.19b), to obtain (F.27) we have used (F.25) and (F.2), and where to obtain (F.28) we have used Lemma 5.8 with (F.2) (and $k_1 = k$ and $k_2 = 1$).

Since $\beta^{(1)}(y)$ only appears in (F.24b) through the second term in (F.28), we conclude that its optimum value is

$$\beta_{*,S}^{(1)}(y) \equiv 0. \tag{F.29}$$

Combining the remainder of (F.28) with the first term in (F.24b), we then have, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\tilde{\ell}_2(g^{(1)}, \beta^{(1)}_{*,S}) &= \epsilon \mathbb{E}_{P_{S,Y}}\left[\left(S - \frac{\epsilon}{2}\, g^{(1)}(Y)\right)^{\mathrm{T}} g^{(1)}(Y)\right] \\ &= \epsilon \mathbb{E}_{P_Y}\left[\mathbb{E}_{P_{S|Y}}\left[S - \frac{\epsilon}{2}\, g^{(1)}(Y)\right]^{\mathrm{T}} g^{(1)}(Y)\right] \\ &\leq \epsilon \sqrt{\mathbb{E}_{P_Y}\left[\left\|\mathbb{E}_{P_{S|Y}}\left[S - \frac{\epsilon}{2}\, g^{(1)}(Y)\right]\right\|^2\right]} \sqrt{\mathbb{E}_{P_Y}\left[\|g^{(1)}(Y)\|^2\right]},\end{aligned} \tag{F.30}$$

where the inequality holds with equality when

$$g^{(1)}(Y) \propto \mathbb{E}_{P_{S|Y}}\left[S - \frac{\epsilon}{2}\, g^{(1)}(Y)\right] = \mathbb{E}_{P_{S|Y}}[S] - \frac{\epsilon}{2}\, g^{(1)}(Y), \tag{F.31}$$

for some nonnegative constant of proportionality, i.e., when

$$g^{(1)}(Y) = c\,\mathbb{E}_{P_{S|Y}}[S] \tag{F.32}$$

for $0 \leq c \leq 2/\epsilon$. In this case,

$$\begin{aligned}\tilde{\ell}_2(g^{(1)}, \beta^{(1)}_{*,S}) &= \epsilon c\left(1 - \frac{\epsilon}{2} c\right) \mathbb{E}_{P_Y}\left[\|\mathbb{E}_{P_{S|Y}}[S]\|^2\right] \\ &= \frac{1}{2}\,(1 - (1 - \epsilon c)^2) \mathbb{E}_{P_Y}\left[\|\mathbb{E}_{P_{S|Y}}[S]\|^2\right] \\ &\leq \frac{1}{2}\, \mathbb{E}_{P_Y}\left[\|\mathbb{E}_{P_{S|Y}}[S]\|^2\right],\end{aligned} \tag{F.33}$$

where equality is achieved when $c = 1/\epsilon$. Hence, the optimum value of $g^{(1)}(y)$ is

$$g^{(1)}_{*,S}(y) = \frac{1}{\epsilon}\, \boldsymbol{\mu}_{S|Y}(y), \tag{F.34}$$

which we note has $\mathbb{E}[g^{(1)}_{*,S}(Y)] = 0$, as our constraints (F.15) dictate. In turn, substituting the right-hand side of (F.33) into (F.12) via (F.24), we obtain (F.3) as desired.

Moreover, the corresponding $g_{*,S}(y)$ and $\beta_{*,S}(y)$ satisfy (F.4a) as desired, i.e., $\tilde{P}^*_{Y|S}(y|s)$ takes the form (F.4b). ■

F.2 Proof of Corollary 8.2

First, without loss of generality we impose on f the constraints (3.6c), so

$$\boldsymbol{\mu}_S = \mathbf{0} \qquad \text{and} \qquad \boldsymbol{\Lambda}_S = \mathbf{I}, \tag{F.35}$$

Next, note that since f is injective, by the invariance of mutual information to coordinate transformations, the first term on the right-hand side of (8.2) is $I(X;Y)$, which doesn't depend on f. Accordingly, we have, specializing the second term on the right-hand side of (8.2) to the case (F.35),

$$f_* = \operatorname*{arg\,max}_{f \in \mathcal{F}_k} \mathbb{E}_{P_Y}\left[\left\| \mathbb{E}_{\hat{P}_{X|Y}}[f(X)] \right\|^2 \right] = \operatorname*{arg\,max}_{\boldsymbol{\Xi}^X} \| \tilde{\mathbf{B}}\, \boldsymbol{\Xi}^X \|_{\mathrm{F}}^2,$$

where $\tilde{\mathbf{B}}$ is as defined in (2.29), $\boldsymbol{\Xi}^X$ is the $|\mathcal{X}| \times k$ matrix whose ith column is the feature vector associated with f_i, the ith element of f, and the maximization with respect to $\boldsymbol{\Xi}^X$ is subject to the constraint

$$\left(\boldsymbol{\Xi}^X\right)^{\mathrm{T}} \boldsymbol{\Xi}^X = \mathbf{I},$$

which corresponds to (3.6c). Accordingly, applying Lemma 3.1, we obtain

$$\boldsymbol{\Xi}^X = \boldsymbol{\Psi}_{(k)}^X,$$

i.e.,

$$f_i(x) = f_i^*(x), \quad i = 1, \dots, k,$$

Finally, to obtain (8.10a) we use (2.26b) in (8.4a) with (F.35), and (8.10b) follows immediately via (8.4b). ■

G

Appendices for Section 9

G.1 Proof of Fact 9.3

Suppose $\mathbf{M}$ is $k_1 \times k_2$. Let $\boldsymbol{\psi}_i^{\mathrm{L}}$ and $\boldsymbol{\psi}_i^{\mathrm{R}}$ denote the left and right singular vectors of $\mathbf{M}$ corresponding to $\sigma_i(\mathbf{M})$, for $i = 1, \ldots, \min\{k_1, k_2\}$. Then

$$\begin{bmatrix} (\boldsymbol{\psi}_i^{\mathrm{L}})^{\mathrm{T}} & (\boldsymbol{\psi}_i^{\mathrm{R}})^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \qquad \text{and} \qquad \begin{bmatrix} (\boldsymbol{\psi}_i^{\mathrm{L}})^{\mathrm{T}} & -(\boldsymbol{\psi}_i^{\mathrm{R}})^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$$

are eigenvectors of $\boldsymbol{\Lambda}$ with eigenvalues $1 + \sigma_i(\mathbf{M})$ and $1 - \sigma_i(\mathbf{M})$, respectively. The remaining $\max\{k_1, k_2\} - \min\{k_1, k_2\}$ eigenvalues are 1 and, if $k_2 > k_1$, correspond to eigenvectors

$$\begin{bmatrix} \mathbf{0} & (\boldsymbol{\psi}_j^{\mathrm{R}})^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}},$$

for $j = k_1 + 1, \ldots, k_2$. Likewise, if $k_1 > k_2$ these unity eigenvalues correspond to eigenvectors

$$\begin{bmatrix} (\boldsymbol{\psi}_j^{\mathrm{L}})^{\mathrm{T}} & \mathbf{0} \end{bmatrix}^{\mathrm{T}},$$

for $j = k_2 + 1, \ldots, k_1$. Hence, we conclude that the eigenvalues are nonnegative if and only if $\sigma_i(\mathbf{M}) \leq 1$ for $i = 1, \ldots, \min\{k_1, k_2\}$. ■

The version of record is available at: http://dx.doi.org/10.1561/0100000107

G.2 Proof of Corollary 9.5

Applying (9.18) we have

$$\begin{aligned}
\mathbb{E}[\mathbf{g}^*(Y)|X] &= \mathbb{E}[(\mathbf{G}^*)^{\mathrm{T}} Y|X] \\
&= (\mathbf{G}^*)^{\mathrm{T}} \mathbf{\Gamma}_{Y|X} X \\
&= (\mathbf{G}^*)^{\mathrm{T}} \mathbf{\Lambda}_{YX} \mathbf{\Lambda}_X^{-1} X \\
&= (\mathbf{G}^*)^{\mathrm{T}} (\mathbf{G}^*)^{-\mathrm{T}} \mathbf{\Sigma}\, (\mathbf{F}^*)^{-1} \mathbf{\Lambda}_X^{-1} X \\
&= \mathbf{\Sigma}\, (\mathbf{\Psi}^X)^{\mathrm{T}} \mathbf{\Lambda}_X^{1/2} \mathbf{\Lambda}_X^{-1} X \\
&= \mathbf{\Sigma}\, (\mathbf{F}^*)^{\mathrm{T}} X \\
&= \mathbf{\Sigma}\, \mathbf{f}^*(X)
\end{aligned}$$

and, analogously,

$$\begin{aligned}
\mathbb{E}[\mathbf{f}^*(X)|Y] &= \mathbb{E}[(\mathbf{F}^*)^{\mathrm{T}} X|Y] \\
&= (\mathbf{F}^*)^{\mathrm{T}} \mathbf{\Gamma}_{X|Y} Y \\
&= (\mathbf{F}^*)^{\mathrm{T}} \mathbf{\Lambda}_{XY} \mathbf{\Lambda}_Y^{-1} Y \\
&= (\mathbf{F}^*)^{\mathrm{T}} (\mathbf{F}^*)^{-\mathrm{T}} \mathbf{\Sigma}\, (\mathbf{G}^*)^{-1} \mathbf{\Lambda}_Y^{-1} Y \\
&= \mathbf{\Sigma}\, (\mathbf{\Psi}^Y)^{\mathrm{T}} \mathbf{\Lambda}_Y^{1/2} \mathbf{\Lambda}_Y^{-1} Y \\
&= \mathbf{\Sigma}\, (\mathbf{G}^*)^{\mathrm{T}} Y \\
&= \mathbf{\Sigma}\, \mathbf{g}^*(Y).
\end{aligned}$$

■

G.3 Proof of Lemma 9.8

First, we have

$$\begin{aligned}
&(\mathbf{A}\mu_P + \mathbf{c} - (\mathbf{A}\mu_Q + \mathbf{c}))^{\mathrm{T}} (\mathbf{A}\mathbf{\Lambda}_Q \mathbf{A}^{\mathrm{T}})^{-1} (\mathbf{A}\mu_P + \mathbf{c} - (\mathbf{A}\mu_Q + \mathbf{c})) \\
&\quad = (\mu_P - \mu_Q)^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{A}^{-\mathrm{T}} \mathbf{\Lambda}_Q^{-1} \mathbf{A}^{-1} \mathbf{A} (\mu_P - \mu_Q) \\
&\quad = (\mu_P - \mu_Q)^{\mathrm{T}} \mathbf{\Lambda}_Q^{-1} (\mu_P - \mu_Q). \qquad \text{(G.1)}
\end{aligned}$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

Second, we have

$$\begin{align}
&\|(\mathbf{A}\boldsymbol{\Lambda}_Q\mathbf{A}^{\mathrm{T}})^{-1/2}(\mathbf{A}\boldsymbol{\Lambda}_P\mathbf{A}^{\mathrm{T}}-\mathbf{A}\boldsymbol{\Lambda}_Q\mathbf{A}^{\mathrm{T}})(\mathbf{A}\boldsymbol{\Lambda}_Q\mathbf{A}^{\mathrm{T}})^{-1/2}\|_{\mathrm{F}}^2 \notag\\
&\quad = \mathrm{tr}\Big((\mathbf{A}\boldsymbol{\Lambda}_Q\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{A}(\boldsymbol{\Lambda}_P-\boldsymbol{\Lambda}_Q)\mathbf{A}^{\mathrm{T}}(\mathbf{A}\boldsymbol{\Lambda}_Q\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{A}(\boldsymbol{\Lambda}_P-\boldsymbol{\Lambda}_Q)\mathbf{A}^{\mathrm{T}}\Big) \tag{G.2}\\
&\quad = \mathrm{tr}\Big(\mathbf{A}^{-\mathrm{T}}\boldsymbol{\Lambda}_Q^{-1}(\boldsymbol{\Lambda}_P-\boldsymbol{\Lambda}_Q)\boldsymbol{\Lambda}_Q^{-1}(\boldsymbol{\Lambda}_P-\boldsymbol{\Lambda}_Q)\mathbf{A}^{\mathrm{T}}\Big) \notag\\
&\quad = \mathrm{tr}\Big(\boldsymbol{\Lambda}_Q^{-1}(\boldsymbol{\Lambda}_P-\boldsymbol{\Lambda}_Q)\boldsymbol{\Lambda}_Q^{-1}(\boldsymbol{\Lambda}_P-\boldsymbol{\Lambda}_Q)\Big) \tag{G.3}\\
&\quad = \|\boldsymbol{\Lambda}_Q^{-1/2}(\boldsymbol{\Lambda}_P-\boldsymbol{\Lambda}_Q)\boldsymbol{\Lambda}_Q^{-1/2}\|_{\mathrm{F}}^2, \tag{G.4}
\end{align}$$

where to obtain (G.2) we have used Lemma 9.9, to obtain (G.3) we have used the invariance of the trace operator to cyclic permutations, and to obtain (G.4) we have again used Lemma 9.9. Combining (G.1) and (G.4) with (9.35b), we obtain (9.36). ■

G.4 Proof of Lemma 9.11

With

$$\tilde{C} \triangleq \begin{bmatrix} \tilde{Z} \\ \tilde{W} \end{bmatrix}, \tag{G.5}$$

via (9.3), we have

$$\boldsymbol{\Lambda}_{\tilde{C}} = \mathbb{E}[\tilde{C}\tilde{C}^{\mathrm{T}}] = \begin{bmatrix} \mathbf{I} & \epsilon\boldsymbol{\Phi}^{Z|W} \\ \epsilon\boldsymbol{\Phi}^{Z|W} & \mathbf{I} \end{bmatrix},$$

Thus,

$$\begin{align}
\bar{D}(\mathtt{N}(\mathbf{0},\boldsymbol{\Lambda}_{\tilde{C}}) \parallel \mathtt{N}(\mathbf{0},\mathbf{I})) &= \frac{1}{2}\|\boldsymbol{\Lambda}_{\tilde{C}} - \mathbf{I}\|_{\mathrm{F}}^2 \notag\\
&= \frac{1}{2}\left\|\begin{bmatrix} \mathbf{0} & \epsilon\boldsymbol{\Phi}^{Z|W} \\ \epsilon(\boldsymbol{\Phi}^{Z|W})^{\mathrm{T}} & \mathbf{0} \end{bmatrix}\right\|_{\mathrm{F}}^2 \notag\\
&= \epsilon^2\,\|\boldsymbol{\Phi}^{Z|W}\|_{\mathrm{F}}^2. \tag{G.6}
\end{align}$$

■

G.5 Proof of Lemma 9.12

We obtain

$$
\begin{aligned}
\bar{D}(P_{Z|W}(\cdot|w)\|P_Z) &= \bar{D}(P_{\tilde{Z}|\tilde{W}}(\cdot|\tilde{w})\|P_{\tilde{Z}}) && \text{(G.7)}\\
&= \epsilon^2\,\tilde{w}^{\mathrm{T}}\big(\boldsymbol{\Phi}^{Z|W}\big)^{\mathrm{T}}\boldsymbol{\Phi}^{Z|W}\tilde{w}\\
&\qquad\qquad + \frac{1}{2}\Big\|\Big(\mathbf{I}-\epsilon^2\boldsymbol{\Phi}^{Z|W}\big(\boldsymbol{\Phi}^{Z|W}\big)^{\mathrm{T}}\Big)-\mathbf{I}\Big\|_{\mathrm{F}}^2 && \text{(G.8)}\\
&= \epsilon^2\big\|\boldsymbol{\Phi}^{Z|W}\tilde{w}\big\|^2+\frac{\epsilon^4}{2}\big\|\boldsymbol{\Phi}^{Z|W}\big(\boldsymbol{\Phi}^{Z|W}\big)^{\mathrm{T}}\big\|_{\mathrm{F}}^2\\
&= \epsilon^2\big\|\boldsymbol{\Phi}^{Z|W}\tilde{w}\big\|^2+o(\epsilon^2),
\end{aligned}
$$

where to obtain (G.7) we have used Lemma 9.8, and to obtain (G.8) we have used (9.35b) and the fact that

$$\tilde{Z} = \epsilon\boldsymbol{\Phi}^{Z|W}\tilde{W}+\nu_{\tilde{W}\to\tilde{Z}},$$

where $\boldsymbol{\Phi}^{Z|W}$ is as defined in (9.40) and

$$\mathbb{E}\big[\nu_{\tilde{W}\to\tilde{Z}}\nu_{\tilde{W}\to\tilde{Z}}^{\mathrm{T}}\big] = \mathbf{I}-\epsilon^2\boldsymbol{\Phi}^{Z|W}\big(\boldsymbol{\Phi}^{Z|W}\big)^{\mathrm{T}}.$$

■

G.6 Proof of Lemma 9.13

We have

$$
\begin{aligned}
\mathbb{E}_{P_W}\big[\bar{D}(P_{Z|W}(\cdot|W)\|P_Z)\big] &= \epsilon^2\,\mathbb{E}\Big[\big\|\boldsymbol{\Phi}^{Z|W}\tilde{W}\big\|^2\Big]+o(\epsilon^2) && \text{(G.9)}\\
&= \epsilon^2\,\big\|\boldsymbol{\Phi}^{Z|W}\big\|_{\mathrm{F}}^2+o(\epsilon^2) && \text{(G.10)}\\
&= \bar{D}(P_{Z,W}\|P_ZP_W)(1+o(1)), && \text{(G.11)}
\end{aligned}
$$

where to obtain (G.9) we have used Lemma 9.12, to obtain (G.10) we have used Lemma 5.8 (with $k_1 = K_W$ and $k_2 = 1$) since $\tilde{W}$ is spherically symmetric, and to obtain (G.11) we have used Lemma 9.11. ■

G.7 Proof of Fact 9.14

Let the ith singular value of $\mathbf{A}$ be λ_i. Then it suffices to note that

$$\begin{aligned}\log\left|\mathbf{I}-\epsilon^2\mathbf{A}\,\mathbf{A}^{\mathrm{T}}\right| &= \log\prod_i(1-\epsilon^2\lambda_i^2)\\ &= \log\left(1-\epsilon^2\sum_i\lambda_i^2+o(\epsilon^2)\right)\\ &= -\epsilon^2\sum_i\lambda_i^2+o(\epsilon^2)\\ &= -\epsilon^2\|\mathbf{A}\|_{\mathrm{F}}^2+o(\epsilon^2).\end{aligned}$$

■

G.8 Proof of Lemma 9.15

We have

$$\begin{aligned}D\big(P_{Z|W}(\cdot|w)\,\|\,P_Z\big) &= D\big(P_{\tilde{Z}|\tilde{W}}(\cdot|\tilde{w})\,\|\,P_{\tilde{Z}}\big) &\text{(G.12)}\\ &= \frac{1}{2}\bigg[\epsilon^2\,\left\|\mathbf{\Phi}^{Z|W}\tilde{w}\right\|^2\\ &\qquad+\mathrm{tr}\Big(\Big(\mathbf{I}-\epsilon^2\mathbf{\Phi}^{Z|W}\big(\mathbf{\Phi}^{Z|W}\big)^{\mathrm{T}}\Big)-\mathbf{I}\Big)\\ &\qquad\qquad-\log\Big|\mathbf{I}-\epsilon^2\mathbf{\Phi}^{Z|W}\big(\mathbf{\Phi}^{Z|W}\big)^{\mathrm{T}}\Big|\bigg]\\ &= \frac{\epsilon^2}{2}\left\|\mathbf{\Phi}^{Z|W}\tilde{w}\right\|^2+o(\epsilon^2) &\text{(G.13)}\\ &= \frac{1}{2}\,\bar{D}(P_{Z|W}(\cdot|w)\|P_Z)+o(\epsilon^2), &\text{(G.14)}\end{aligned}$$

where to obtain (G.12) we have used the coordinate invariance of KL divergence, to obtain (G.13) we have used Fact 9.14, and to obtain (G.14) we have used Lemma 9.12. ■

G.9 Proof of Corollary 9.16

We have

$$I(Z;W)=\mathbb{E}_{P_W}\left[D\big(P_{Z|W}(\cdot|W)\,\|\,P_Z\big)\right]$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

$$= \frac{1}{2}\,\mathbb{E}_{P_W}\left[\bar{D}(P_{Z|W}(\cdot|W) \parallel P_Z)\right] + o(\epsilon^2) \tag{G.15}$$

$$= \frac{1}{2}\,\bar{D}(P_{Z,W} \parallel P_Z P_W) + o(\epsilon^2), \tag{G.16}$$

where to obtain (G.15) we have used Lemma 9.15, and to obtain (G.16) we have used Lemma 9.13. ■

G.10 Proof of Lemma 9.17

First, via (9.46) and the invariance of divergence to coordinate transformations we have

$$D(P_{X,Y} \| Q_{X,Y}) = \frac{1}{2}\Big[\operatorname{tr}\Big((\mathbf{\Lambda}^{Q}_{\tilde{X},\tilde{Y}})^{-1}\,\mathbf{\Lambda}^{P}_{\tilde{X},\tilde{Y}} - \mathbf{I}\Big) - \log\big|\mathbf{\Lambda}^{P}_{\tilde{X},\tilde{Y}}\,(\mathbf{\Lambda}^{Q}_{\tilde{X},\tilde{Y}})^{-1}\big|\Big], \tag{G.17}$$

where

$$\mathbf{\Lambda}^{P}_{\tilde{X},\tilde{Y}} = \begin{bmatrix} \mathbf{I} & \tilde{\mathbf{B}}_P^{\mathrm{T}} \\ \tilde{\mathbf{B}}_P & \mathbf{I} \end{bmatrix} \quad \text{and} \quad \mathbf{\Lambda}^{Q}_{\tilde{X},\tilde{Y}} = \begin{bmatrix} \mathbf{I} & \tilde{\mathbf{B}}_Q^{\mathrm{T}} \\ \tilde{\mathbf{B}}_Q & \mathbf{I} \end{bmatrix}. \tag{G.18}$$

Next, using (G.18) with (9.54) we obtain

$$\begin{aligned}(\mathbf{\Lambda}^{Q}_{\tilde{X},\tilde{Y}})^{-1}\,\mathbf{\Lambda}^{P}_{\tilde{X},\tilde{Y}} &= \begin{bmatrix} \mathbf{I} + \tilde{\mathbf{B}}_Q^{\mathrm{T}}\tilde{\mathbf{B}}_Q & -\tilde{\mathbf{B}}_Q^{\mathrm{T}} \\ -\tilde{\mathbf{B}}_Q & \mathbf{I} + \tilde{\mathbf{B}}_Q\tilde{\mathbf{B}}_Q^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \tilde{\mathbf{B}}_P^{\mathrm{T}} \\ \tilde{\mathbf{B}}_P & \mathbf{I} \end{bmatrix} + o(\epsilon^2) \\ &= \begin{bmatrix} \mathbf{I} + \tilde{\mathbf{B}}_Q^{\mathrm{T}}(\tilde{\mathbf{B}}_Q - \tilde{\mathbf{B}}_P) & (\tilde{\mathbf{B}}_P - \tilde{\mathbf{B}}_Q)^{\mathrm{T}} \\ \tilde{\mathbf{B}}_P - \tilde{\mathbf{B}}_Q & \mathbf{I} + \tilde{\mathbf{B}}_Q(\tilde{\mathbf{B}}_Q - \tilde{\mathbf{B}}_P)^{\mathrm{T}} \end{bmatrix} + o(\epsilon^2),\end{aligned}$$

so

$$\begin{aligned}\operatorname{tr}\Big((\mathbf{\Lambda}^{Q}_{\tilde{X},\tilde{Y}})^{-1}\mathbf{\Lambda}^{P}_{\tilde{X},\tilde{Y}} - \mathbf{I}\Big) &= 2\operatorname{tr}\Big(\tilde{\mathbf{B}}_Q(\tilde{\mathbf{B}}_Q - \tilde{\mathbf{B}}_P)^{\mathrm{T}}\Big) + o(\epsilon^2) \\ &= 2\,\|\tilde{\mathbf{B}}_Q\|_{\mathrm{F}}^2 - 2\operatorname{tr}(\tilde{\mathbf{B}}_Q\tilde{\mathbf{B}}_P^{\mathrm{T}}) + o(\epsilon^2). \end{aligned} \tag{G.19}$$

Finally, using Fact 9.14 and the block matrix determinant identity we obtain

$$\log|\mathbf{\Lambda}^{P}_{\tilde{X}\tilde{Y}}| = \log|\mathbf{I} - \tilde{\mathbf{B}}_P^{\mathrm{T}}\tilde{\mathbf{B}}_P| = -\|\tilde{\mathbf{B}}_P\|_{\mathrm{F}}^2 + o(\epsilon^2) \tag{G.20a}$$

$$\log|\mathbf{\Lambda}^{Q}_{\tilde{X}\tilde{Y}}| = \log|\mathbf{I} - \tilde{\mathbf{B}}_Q^{\mathrm{T}}\tilde{\mathbf{B}}_Q| = -\|\tilde{\mathbf{B}}_Q\|_{\mathrm{F}}^2 + o(\epsilon^2), \tag{G.20b}$$

so substituting (G.19) and (G.20) in (G.17) yields (9.49). ■

G.11 Proof of Proposition 10.2

Given $\mathcal{X} = \{x_1, \dots, x_{|\mathcal{X}|}\} \subset \mathbb{R}^{K_X}$ and $\mathcal{Y} = \{y_1, \dots, y_{|\mathcal{Y}|}\} \subset \mathbb{R}^{K_Y}$, consider (without loss of generality) arbitrary zero-mean, unit-variance features $f\colon \mathcal{X} \to \mathbb{R}$ and $g\colon \mathcal{Y} \to \mathbb{R}$, whose corresponding feature vectors are $\boldsymbol{\xi}^X \in \mathcal{I}^{\mathcal{X}}$ and $\boldsymbol{\xi}^Y \in \mathcal{I}^{\mathcal{Y}}$, respectively.

In addition, let $f_{\mathrm{L}}(x) = (\boldsymbol{\xi}_{\mathrm{G}}^X)^{\mathrm{T}} x$ and $g_{\mathrm{L}}(y) = (\boldsymbol{\xi}_{\mathrm{G}}^Y)^{\mathrm{T}} y$ denote the linear MMSE approximations to f and g, respectively, i.e.,

$$\boldsymbol{\xi}_{\mathrm{G}}^X = \operatorname*{arg\,min}_{\boldsymbol{\xi}_{\mathrm{G}} \in \mathbb{R}^{K_X}} \mathbb{E}\left[\left(\boldsymbol{\xi}_{\mathrm{G}}^{\mathrm{T}} X - f(X)\right)^2\right] \tag{G.21a}$$

$$\boldsymbol{\xi}_{\mathrm{G}}^Y = \operatorname*{arg\,min}_{\boldsymbol{\xi}_{\mathrm{G}} \in \mathbb{R}^{K_Y}} \mathbb{E}\left[\left(\boldsymbol{\xi}_{\mathrm{G}}^{\mathrm{T}} Y - g(Y)\right)^2\right]. \tag{G.21b}$$

Without loss of generality, we assume that $\boldsymbol{\Lambda}_X = \mathbf{I}$ and $\boldsymbol{\Lambda}_Y = \mathbf{I}$, since if not they can be converted to this form by linear transformation. Then, from standard linear (MMSE) estimation theory it follows that

$$\boldsymbol{\xi}_{\mathrm{G}}^X = \mathbb{E}[f(X)\, X^{\mathrm{T}}] = \underbrace{\mathbf{X}\sqrt{\mathbf{P}_X}}_{\triangleq \boldsymbol{\Pi}^X} \boldsymbol{\xi}^X \tag{G.22a}$$

$$\boldsymbol{\xi}_{\mathrm{G}}^Y = \mathbb{E}[\, g(Y)\, Y^{\mathrm{T}}] = \underbrace{\mathbf{Y}\sqrt{\mathbf{P}_Y}}_{\triangleq \boldsymbol{\Pi}^Y} \boldsymbol{\xi}^Y, \tag{G.22b}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are matrices whose columns are the vectors in $\mathcal{X}$ and $\mathcal{Y}$, respectively, and where the matrices $\boldsymbol{\Pi}^X$ and $\boldsymbol{\Pi}^Y$ characterize the associated projections.

Next, note that if f is linear, i.e.,

$$f(x) = \boldsymbol{\zeta}_{\mathrm{G}}^{\mathrm{T}} x \tag{G.23}$$

for some $\boldsymbol{\zeta}_{\mathrm{G}}$, then

$$(\boldsymbol{\Pi}^X)^{\mathrm{T}} \boldsymbol{\zeta}_{\mathrm{G}} = \sqrt{\mathbf{P}_X}\, \mathbf{X}^{\mathrm{T}} \boldsymbol{\zeta}_{\mathrm{G}} = \boldsymbol{\xi}^X$$

since $\mathbf{X}^{\mathrm{T}} \boldsymbol{\zeta}_{\mathrm{G}}$ is a vector whose xth element is $f(x)$.

Now for any feature f with feature vector ξ^X, the feature vector $\boldsymbol{\xi}^Y \triangleq \mathbf{B}\, \boldsymbol{\xi}^X$ corresponds to the feature $g(y) = \mathbb{E}[f(X)|Y = y]$. To see this, it suffices to note that $\boldsymbol{\xi}^Y$ has elements

$$\begin{aligned}\xi^Y(y) &= \sum_{x\in\mathcal{X}} B(x,y)\,\xi^X(x)\\ &= \sum_{x\in\mathcal{X}} \frac{1}{\sqrt{P_Y(y)}}\,P_{Y|X}(y|x)\,\sqrt{P_X(x)}\,\sqrt{P_X(x)}\,f(x)\\ &= \sqrt{P_Y(y)}\,\underbrace{\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)\,f(x)}_{=g(y)}.\end{aligned}$$

In turn, specializing f to the case (G.23) yields

$$\mathbf{B}\,(\boldsymbol{\Pi}^X)^{\mathrm{T}}\boldsymbol{\zeta}_{\mathrm{G}} = \sqrt{\mathbf{P}_Y}\,\mathbb{E}[\boldsymbol{\zeta}_{\mathrm{G}}^{\mathrm{T}}\,X|Y],$$

and specializing the resulting g in (G.22b) with (G.21b) yields

$$\begin{aligned}\boldsymbol{\Pi}^Y\mathbf{B}\,(\boldsymbol{\Pi}^X)^{\mathrm{T}}\boldsymbol{\zeta}_{\mathrm{G}} &= \underset{\boldsymbol{\xi}_{\mathrm{G}}\in\mathbb{R}^{K_Y}}{\arg\min}\,\mathbb{E}\Big[(\boldsymbol{\xi}_{\mathrm{G}}^{\mathrm{T}}\,Y - \mathbb{E}[\boldsymbol{\zeta}_{\mathrm{G}}^{\mathrm{T}}\,X|Y])^2\Big]\\ &= \underset{\boldsymbol{\xi}_{\mathrm{G}}\in\mathbb{R}^{K_Y}}{\arg\min}\,\mathbb{E}\Big[(\boldsymbol{\xi}_{\mathrm{G}}^{\mathrm{T}}\,Y - \boldsymbol{\zeta}_{\mathrm{G}}^{\mathrm{T}}\,X)^2\Big] && \text{(G.24)}\\ &= \boldsymbol{\Lambda}_Y^{-1}\,\boldsymbol{\Lambda}_{YX}\,\boldsymbol{\zeta}_{\mathrm{G}} && \text{(G.25)}\\ &= \mathbf{B}_{\mathrm{G}}\,\boldsymbol{\zeta}_{\mathrm{G}}. && \text{(G.26)}\end{aligned}$$

To obtain (G.24) we have used

$$\begin{aligned}\mathbb{E}\Big[(\boldsymbol{\xi}_{\mathrm{G}}^{\mathrm{T}}\,Y - \boldsymbol{\zeta}_{\mathrm{G}}^{\mathrm{T}}\,X)^2\Big] = \mathbb{E}\Big[(\boldsymbol{\xi}_{\mathrm{G}}^{\mathrm{T}}\,Y - \mathbb{E}[\boldsymbol{\zeta}_{\mathrm{G}}^{\mathrm{T}}\,X|Y])^2\Big]&\\ + \mathbb{E}\Big[(\boldsymbol{\zeta}_{\mathrm{G}}^{\mathrm{T}}\,X - \mathbb{E}[\boldsymbol{\zeta}_{\mathrm{G}}^{\mathrm{T}}\,X|Y])^2\Big]&\\ - 2\,\mathbb{E}\Big[(\boldsymbol{\zeta}_{\mathrm{G}}^{\mathrm{T}}\,X - \mathbb{E}[\boldsymbol{\zeta}_{\mathrm{G}}^{\mathrm{T}}\,X|Y])&\\ \cdot(\boldsymbol{\xi}_{\mathrm{G}}^{\mathrm{T}}\,Y - \mathbb{E}[\boldsymbol{\zeta}_{\mathrm{G}}^{\mathrm{T}}\,X|Y])\Big]&,\end{aligned}$$

where we note that the second term does not depend on $\boldsymbol{\xi}_{\mathrm{G}}$, and that the last term is zero due to the orthogonality property of MMSE estimators. In turn, to obtain (G.25) we recognize (G.24) as a linear MMSE estimation problem, whose solution depends only on the first and second moments of (X,Y), and to obtain (G.26) we use (9.11). Finally, since (G.26) holds for all $\boldsymbol{\zeta}_{\mathrm{G}}$, we obtain (10.3) as desired. ■

G.12 Proof of Fact 9.22

It suffices to note that

$$\begin{aligned}
\mathbf{\Lambda}_{\tilde{Z}_1\tilde{Z}_3} &= \mathbb{E}\Big[\mathbb{E}\big[\tilde{Z}_1\tilde{Z}_3^{\mathrm{T}}\big|\tilde{Z}_2\big]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\big[\tilde{Z}_1\big|\tilde{Z}_2\big]\,\mathbb{E}\big[\tilde{Z}_3\big|\tilde{Z}_2\big]^{\mathrm{T}}\Big] \\
&= \mathbb{E}\Big[\mathbf{\Lambda}_{\tilde{Z}_1\tilde{Z}_2}\,\tilde{Z}_2\,\tilde{Z}_2^{\mathrm{T}}\mathbf{\Lambda}_{\tilde{Z}_3\tilde{Z}_2}^{\mathrm{T}}\Big] \\
&= \mathbf{\Lambda}_{\tilde{Z}_1\tilde{Z}_2}\,\mathbf{\Lambda}_{\tilde{Z}_2}\,\mathbf{\Lambda}_{\tilde{Z}_2\tilde{Z}_3} \\
&= \mathbf{\Lambda}_{\tilde{Z}_1\tilde{Z}_2}\,\mathbf{\Lambda}_{\tilde{Z}_2\tilde{Z}_3}.
\end{aligned}$$

■

G.13 Proof of Proposition 9.24

Without loss of generality, we restrict $S_{(k)}$ and $T_{(k)}$ so that they are normalized with respect to P_X and P_Y, respectively, i.e., $(\mathbf{F}_{(k)}, \mathbf{G}_{(k)}) \in \mathcal{L}$ as defined in (9.27). Accordingly, we have the representations (9.29) in which $\mathbf{\Xi}^X$ and $\mathbf{\Xi}^Y$ satisfy (9.31).

For the MMSE estimation of V based on $S_{(k)}$, the MSE is

$$\begin{aligned}
\lambda_{\mathrm{e}}^{V|S}(\mathcal{C}_{\epsilon_Y}^{K_Y}(\mathbf{\Lambda}_Y), \mathbf{F}_{(k)}) &= \mathrm{tr}\big(\mathbf{\Lambda}_V^{1/2}\mathbf{\Lambda}_{\tilde{V}|S_{(k)}}\mathbf{\Lambda}_V^{1/2}\big) \\
&= \mathrm{tr}(\mathbf{\Lambda}_V) - \epsilon_Y^2\,\big\|\mathbf{\Lambda}_V^{1/2}\,(\mathbf{\Phi}^{Y|V})^{\mathrm{T}}\tilde{\mathbf{B}}\,\mathbf{\Xi}^X\big\|_{\mathrm{F}}^2, \qquad \text{(G.27)}
\end{aligned}$$

where we have used (9.65b), so

$$\begin{aligned}
\bar{\lambda}_{\mathrm{e}}^{V|S}(\mathbf{F}_{(k)}) &= \mathbb{E}_{\mathrm{RIE}}\big[\lambda_{\mathrm{e}}^{V|S}(\mathcal{C}_{\epsilon_Y}^{K_Y}(\mathbf{\Lambda}_Y), \mathbf{F}_{(k)})\big] \\
&= \mathrm{tr}(\mathbf{\Lambda}_V) - \frac{\epsilon_Y^2}{K_V K_Y}\,\mathrm{tr}(\mathbf{\Lambda}_V)\,\mathbb{E}\Big[\big\|\mathbf{\Phi}^{Y|V}\big\|_{\mathrm{F}}^2\Big]\,\big\|\tilde{\mathbf{B}}\,\mathbf{\Xi}^X\big\|_{\mathrm{F}}^2 \qquad \text{(G.28)} \\
&\geq \mathrm{tr}(\mathbf{\Lambda}_V)\left[1 - \frac{\epsilon_Y^2}{K_V K_Y}\,\mathbb{E}\Big[\big\|\mathbf{\Phi}^{Y|V}\big\|_{\mathrm{F}}^2\Big]\,\sum_{i=1}^{k}\sigma_i^2\right], \qquad \text{(G.29)}
\end{aligned}$$

where to obtain (G.28) we have used Lemma 5.8, and to obtain (G.29) we have used Lemma 3.1. It further follows from Lemma 3.1 that the inequality (G.29) holds with equality when we choose $\mathbf{\Xi}^X$ according to (9.32a).

For the MMSE estimation of U based on $S_{(k)}$, the MSE is

$$\begin{aligned}\lambda_{\mathrm{e}}^{U|S}\big(\mathcal{C}_{\epsilon_X}^{K_X}(\mathbf{\Lambda}_X), \mathbf{F}_{(k)}\big) &= \mathrm{tr}\big(\mathbf{\Lambda}_U^{1/2}\mathbf{\Lambda}_{\tilde{U}|S_{(k)}}\mathbf{\Lambda}_U^{1/2}\big) \\ &= \mathrm{tr}(\mathbf{\Lambda}_U) - \epsilon_X^2 \left\| \mathbf{\Lambda}_U^{1/2} \left(\mathbf{\Phi}^{X|U}\right)^{\mathrm{T}} \mathbf{\Xi}^X \right\|_{\mathrm{F}}^2, \end{aligned} \tag{G.30}$$

so

$$\begin{aligned}\bar{\lambda}_{\mathrm{e}}^{U|S}(\mathbf{F}_{(k)}) &= \mathbb{E}_{\mathrm{RIE}}\big[\lambda_{\mathrm{e}}^{U|S}\big(\mathcal{C}_{\epsilon_X}^{K_X}(\mathbf{\Lambda}_X), \mathbf{F}_{(k)}\big)\big] \\ &= \mathrm{tr}(\mathbf{\Lambda}_U) - \frac{\epsilon_X^2}{K_U K_X}\, \mathrm{tr}(\mathbf{\Lambda}_U)\, \mathbb{E}\Big[\big\|\mathbf{\Phi}^{X|U}\big\|_{\mathrm{F}}^2\Big] \left\|\mathbf{\Xi}^X\right\|_{\mathrm{F}}^2 && \text{(G.31)} \\ &= \mathrm{tr}(\mathbf{\Lambda}_U)\left[1 - \frac{\epsilon_X^2\, k}{K_U K_X}\, \mathbb{E}\Big[\big\|\mathbf{\Phi}^{X|U}\big\|_{\mathrm{F}}^2\Big]\right] && \text{(G.32)} \end{aligned}$$

for any (admissible) choice of $\mathbf{F}_{(k)}$, where to obtain (G.31) we have used Lemma 5.8, and to obtain (G.32) we have used that $\left\|\mathbf{\Xi}^X\right\|_{\mathrm{F}}^2 = k$ due to (9.31). Hence, the unique Pareto optimal choice of $\mathbf{F}_{(k)}$ in (9.71) is as given by (9.32a).

Via a symmetry argument (corresponding to interchanging the roles of X and Y, and U and V, and noting that $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{B}}^{\mathrm{T}}$ share the same singular values), it follows that

$$\begin{aligned}\bar{\lambda}_{\mathrm{e}}^{V|T}(\mathbf{G}_{(k)}) &= \mathrm{tr}(\mathbf{\Lambda}_V)\left[1 - \frac{\epsilon_Y^2\, k}{K_V K_Y}\, \mathbb{E}\Big[\big\|\mathbf{\Phi}^{Y|V}\big\|_{\mathrm{F}}^2\Big]\right] \\ \bar{\lambda}_{\mathrm{e}}^{U|T}(\mathbf{G}_{(k)}) &\geq \mathrm{tr}(\mathbf{\Lambda}_U)\left[1 - \frac{\epsilon_X^2}{K_U K_X}\, \mathbb{E}\Big[\big\|\mathbf{\Phi}^{X|U}\big\|_{\mathrm{F}}^2\Big] \sum_{i=1}^{k} \sigma_i^2\right], \end{aligned}$$

with equality in the latter when $\mathbf{G}_{(k)}$ is given by (9.32b). Hence, the unique Pareto optimal choice of $\mathbf{G}_{(k)}$ in (9.71) is as given by (9.32a).

Finally, we note that we obtain (9.72) by recognizing that

$$\bar{E}^{X|U} = \frac{\mathbb{E}\Big[\big\|\mathbf{\Phi}^{X|U}\big\|_{\mathrm{F}}^2\Big]}{K_U K_X} \quad \text{and} \quad \bar{E}^{Y|V} = \frac{\mathbb{E}\Big[\big\|\mathbf{\Phi}^{Y|V}\big\|_{\mathrm{F}}^2\Big]}{K_V K_Y}.$$

■

G.14 Proof of Proposition 9.25

For the MMSE estimation of V based on S, the MSE is, starting from (G.27),

$$\begin{aligned}
&\lambda_{\mathrm{e}}^{V|S}(\mathcal{C}_{\epsilon_Y}^{K_Y}(\mathbf{\Lambda}_Y), \mathbf{F}_{(k)},) \\
&\quad = \operatorname{tr}(\mathbf{\Lambda}_V) - \epsilon_Y^2 \, \|\mathbf{\Lambda}_V^{1/2} \, (\mathbf{\Phi}^{Y|V})^{\mathrm{T}} \tilde{\mathbf{B}} \, \mathbf{\Xi}^X \|_{\mathrm{F}}^2 \\
&\quad = \operatorname{tr}(\mathbf{\Lambda}_V) - \epsilon_Y^2 \, \|\mathbf{\Lambda}_V^{1/2} \, \mathbf{Q}^{Y|V} \, \mathbf{\Delta}^{Y|V} \, (\tilde{\mathbf{\Phi}}^{Y|V})^{\mathrm{T}} \tilde{\mathbf{B}} \, \mathbf{\Xi}^X \|_{\mathrm{F}}^2 && \text{(G.33)} \\
&\quad \geq \operatorname{tr}(\mathbf{\Lambda}_V) - \epsilon_Y^2 \, \|\mathbf{\Lambda}_V^{1/2}\|_{\mathrm{s}}^2 \, \|\mathbf{\Delta}^{Y|V}\|_{\mathrm{s}}^2 \, \|(\tilde{\mathbf{\Phi}}^{Y|V})^{\mathrm{T}} \tilde{\mathbf{B}}\|_{\mathrm{F}}^2 \, \|\mathbf{\Xi}^X\|_{\mathrm{s}}^2 && \text{(G.34)} \\
&\quad \geq \operatorname{tr}(\mathbf{\Lambda}_V) - \epsilon_Y^2 \, \|\mathbf{\Lambda}_V^{1/2}\|_{\mathrm{s}}^2 \left(\frac{K_Y + k}{k}\right) \sum_{i=1}^k \sigma_i^2 && \text{(G.35)} \\
&\quad \geq k - \epsilon_Y^2 \left(\frac{K_Y + k}{k}\right) \sum_{i=1}^k \sigma_i^2, && \text{(G.36)}
\end{aligned}$$

where to obtain (G.33) we have expressed $\mathbf{\Phi}^{Y|V}$ in terms of its SVD

$$\mathbf{\Phi}^{Y|V} = \tilde{\mathbf{\Phi}}^{Y|V} \mathbf{\Delta}^{Y|V} \, (\mathbf{Q}^{Y|V})^{\mathrm{T}}, \tag{G.37}$$

where the $K_Y \times k$ matrix $\tilde{\mathbf{\Phi}}^{Y|V}$ has orthonormal columns, i.e.,

$$(\tilde{\mathbf{\Phi}}^{Y|V})^{\mathrm{T}} \tilde{\mathbf{\Phi}}^{Y|V} = \mathbf{I}, \tag{G.38}$$

$\mathbf{\Delta}^{Y|V}$ is a $k \times k$ diagonal matrix, and $\mathbf{Q}^{Y|V}$ is $k \times k$ orthogonal matrix. To obtain (G.34) we have (repeatedly) used Fact 5.12 and the fact that $\mathbf{Q}^{Y|V}$ is orthogonal, and to obtain (G.35) we have used that $\mathbf{\Xi}^X$ satisfies (9.31), that

$$\|\mathbf{\Delta}^{Y|V}\|_{\mathrm{s}}^2 = \|\mathbf{\Phi}^{Y|V}\|_{\mathrm{s}}^2 \leq \frac{K_Y + k}{k} \tag{G.39}$$

since V is a Gaussian multi-attribute and so satisfies the property of Definition 9.20, and applied Lemma 3.1 with the constraint (G.38). And to obtain (G.36) we have used the last constraint in (9.74), which implies that none of the singular values of $\mathbf{\Lambda}_V$ are smaller than unity. Finally, it is straightforward to verify that the inequalities leading to the right-hand side of (G.36) hold with equality when

$$\mathbf{\Lambda}_V = \mathbf{I} \tag{G.40a}$$

$$\mathbf{\Phi}^{Y|V} = \sqrt{\frac{K_Y + k}{k}} \, \mathbf{\Psi}_{(k)}^Y \tag{G.40b}$$

The version of record is available at: http://dx.doi.org/10.1561/0100000107

and

$$\boldsymbol{\Xi}^X = \boldsymbol{\Psi}^X_{(k)}. \tag{G.41}$$

For the MMSE estimation of U based on S, the MSE is, starting from (G.30),

$$\begin{aligned}
&\lambda_{\mathrm{e}}^{U|S}(\mathcal{C}_{\epsilon_X}^{K_X}(\boldsymbol{\Lambda}_X), \mathbf{F}_{(k)}) \\
&\quad = \mathrm{tr}(\boldsymbol{\Lambda}_U) - \epsilon_X^2 \left\| \boldsymbol{\Lambda}_U^{1/2} \, (\boldsymbol{\Phi}^{X|U})^{\mathrm{T}} \boldsymbol{\Xi}^X \right\|_{\mathrm{F}}^2 \\
&\quad \geq \mathrm{tr}(\boldsymbol{\Lambda}_U) - \epsilon_X^2 \left\| \boldsymbol{\Lambda}_U^{1/2} \right\|_{\mathrm{s}}^2 \left\| \boldsymbol{\Delta}^{X|U} \right\|_{\mathrm{s}}^2 \left\| (\tilde{\boldsymbol{\Phi}}^{X|U})^{\mathrm{T}} \boldsymbol{\Xi}^X \right\|_{\mathrm{F}}^2 && \text{(G.42)} \\
&\quad \geq \mathrm{tr}(\boldsymbol{\Lambda}_U) - \epsilon_X^2 \left\| \boldsymbol{\Lambda}_U^{1/2} \right\|_{\mathrm{s}}^2 (K_X + k) && \text{(G.43)} \\
&\quad \geq k - \epsilon_X^2 \, (K_X + k), && \text{(G.44)}
\end{aligned}$$

where to obtain (G.42) we have used Fact 5.12, expressing $\boldsymbol{\Phi}^{X|U}$ in terms of its SVD

$$\boldsymbol{\Phi}^{X|U} = \tilde{\boldsymbol{\Phi}}^{X|U} \boldsymbol{\Delta}^{X|U} \, (\mathbf{Q}^{X|U})^{\mathrm{T}}, \tag{G.45}$$

where the $K_Y \times k$ matrix $\tilde{\boldsymbol{\Phi}}^{X|U}$ has orthonormal columns, i.e.,

$$(\tilde{\boldsymbol{\Phi}}^{X|U})^{\mathrm{T}} \tilde{\boldsymbol{\Phi}}^{X|U} = \mathbf{I}, \tag{G.46}$$

$\boldsymbol{\Delta}^{X|U}$ is a $k \times k$ diagonal matrix, and $\mathbf{Q}^{X|U}$ is $k \times k$ orthogonal matrix. To obtain (G.43) we have used that $\tilde{\boldsymbol{\Phi}}^{X|U}$ satisfies (G.46) and $\boldsymbol{\Xi}^X$ satisfies (9.31), Lemma 3.1, and that

$$\left\| \boldsymbol{\Delta}^{X|U} \right\|_{\mathrm{s}}^2 = \left\| \boldsymbol{\Phi}^{X|U} \right\|_{\mathrm{s}}^2 \leq \frac{K_X + k}{k} \tag{G.47}$$

since U is a Gaussian multi-attribute and so satisfies the property of Definition 9.20. To obtain (G.44) we have used the penultimate constraint in (9.74), which implies that none of the singular values of $\boldsymbol{\Lambda}_U$ are smaller than unity. Finally, the inequalities leading to the right-hand side of (G.44) hold with equality when, for example,

$$\boldsymbol{\Lambda}_U = \mathbf{I} \tag{G.48a}$$

$$\boldsymbol{\Phi}^{X|U} = \sqrt{\frac{K_X + k}{k}} \, \boldsymbol{\Psi}^X_{(k)} \tag{G.48b}$$

and $\boldsymbol{\Xi}^X$ is chosen according to (G.41).

Via a symmetry argument (corresponding to interchanging the roles of X and Y, and U and V, and noting that $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{B}}^{\mathrm{T}}$ share the same singular values), it follows that $\lambda_{\mathrm{e}}^{U|T}(\mathcal{C}_{\epsilon_X}^{K_X}(\mathbf{\Lambda}_X), \mathbf{G}_{(k)})$ is minimized by choosing (G.48) and

$$\mathbf{\Xi}^Y = \mathbf{\Psi}^Y_{(k)}, \tag{G.49}$$

and that $\lambda_{\mathrm{e}}^{V|T}(\mathcal{C}_{\epsilon_Y}^{K_Y}(\mathbf{\Lambda}_Y), \mathbf{G}_{(k)})$ is minimized by choosing, for example, (G.40) and (G.49). The unique Pareto optimality of the choices then follows. Finally, (9.75) is obtained by combining the characterizations

$$\mathbf{\Lambda}_{XU} = \epsilon_X\, \mathbf{\Lambda}_X^{1/2}\, \mathbf{\Phi}^{X|U} \quad \text{and} \quad \mathbf{\Lambda}_{YV} = \epsilon_Y\, \mathbf{\Lambda}_Y^{1/2}\, \mathbf{\Phi}^{Y|V} \tag{G.50}$$

with (G.48b), (G.40b), and (9.34). ■

G.15 Proof of Proposition 9.26

First, since mutual information is invariant to coordinate transformations, without loss of generality we may choose $\mathbf{\Lambda}_U = \mathbf{\Lambda}_V = \mathbf{I}$, in which case $\tilde{U} = U$ and $\tilde{V} = V$. Then, using the conditional independencies associated with the Markov chain (9.63) we have

$$\begin{aligned}
\mathbf{\Lambda}_{UV} &= \mathbb{E}\Big[\mathbb{E}\big[U\, V^{\mathrm{T}} \big| \tilde{X}, \tilde{Y}\big]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\big[U \big| \tilde{X}\big]\, \mathbb{E}\big[V^{\mathrm{T}} \big| \tilde{Y}\big]\Big] \\
&= \epsilon_X \epsilon_Y\, \mathbb{E}\Big[\big(\mathbf{\Phi}^{X|U}\big)^{\mathrm{T}} \tilde{X}\, \tilde{Y}^{\mathrm{T}} \mathbf{\Phi}^{Y|V}\Big] \\
&= \epsilon_X \epsilon_Y\, \big(\mathbf{\Phi}^{X|U}\big)^{\mathrm{T}} \tilde{\mathbf{B}}^{\mathrm{T}} \mathbf{\Phi}^{Y|V}.
\end{aligned} \tag{G.51}$$

Hence, with

$$\tilde{C} \triangleq \begin{bmatrix} U \\ V \end{bmatrix} \quad \text{so} \quad \mathbf{\Lambda}_{\tilde{C}} = \begin{bmatrix} \mathbf{I} & \mathbf{\Lambda}_{UV} \\ \mathbf{\Lambda}_{UV}^{\mathrm{T}} & \mathbf{I} \end{bmatrix},$$

we have

$$\begin{aligned}
I(U;V) &= -\frac{1}{2} \log |\mathbf{\Lambda}_{\tilde{C}}| \\
&= -\frac{1}{2} \log \big|\mathbf{I} - \mathbf{\Lambda}_{UV} \mathbf{\Lambda}_{UV}^{\mathrm{T}}\big| \\
&= \frac{\epsilon_X^2 \epsilon_Y^2}{2} \, \big\|\big(\mathbf{\Phi}^{Y|V}\big)^{\mathrm{T}}\, \tilde{\mathbf{B}}\, \mathbf{\Phi}^{X|U}\big\|_{\mathrm{F}}^2 + o(\epsilon_X^2 \epsilon_Y^2)
\end{aligned} \tag{G.52}$$

$$\leq \frac{\epsilon_X^2 \epsilon_Y^2}{2} \left\| \tilde{\mathbf{B}}\, \tilde{\mathbf{\Phi}}^{X|U} \right\|_{\mathrm{F}}^2 \left\| \tilde{\mathbf{\Phi}}^{Y|V} \right\|_{\mathrm{s}}^2 \left\| \mathbf{\Delta}^{X|U} \right\|_{\mathrm{s}}^2 \left\| \mathbf{\Delta}^{Y|V} \right\|_{\mathrm{s}}^2 + o(\epsilon_X^2 \epsilon_Y^2) \tag{G.53}$$

$$\leq \frac{\epsilon_X^2 \epsilon_Y^2}{2} \left(\frac{K_X + k}{k} \right) \left(\frac{K_Y + k}{k} \right) \sum_{i=1}^{k} \sigma_i^2 + o(\epsilon_X^2 \epsilon_Y^2), \tag{G.54}$$

where to obtain (G.52) we have used (G.51) with Fact 9.14, to obtain (G.53) we have used the SVDs (G.45) and (G.37) with (repeatedly) Fact 5.12, and to obtain (G.54) we have used both Lemma 3.1 with (G.46), and both (G.47) and (G.39). Finally, it is straightforward to verify that the inequality (G.54) is achieved with equality when $\mathbf{\Phi}^{Y|V}$ and $\mathbf{\Phi}^{X|U}$ are chosen according to (G.40b) and (G.48b), respectively, which correspond to (9.75b)–(9.75c). With these choices, (G.51) specializes to

$$\begin{aligned} \mathbf{\Lambda}_{UV} &= \epsilon_X \epsilon_Y \sqrt{\frac{K_X + k}{k}} \sqrt{\frac{K_Y + k}{k}} \left(\mathbf{\Psi}_{(k)}^X \right)^{\mathrm{T}} \tilde{\mathbf{B}}^{\mathrm{T}} \mathbf{\Psi}_{(k)}^Y \\ &= \epsilon_X \epsilon_Y \sqrt{\frac{K_X + k}{k}} \sqrt{\frac{K_Y + k}{k}} \mathbf{\Sigma}_{(k)}, \end{aligned}$$

where we have used (9.57b). ■

G.16 Proof of Corollary 9.27

It suffices to note from Corollary 9.16 that U and V so constrained correspond to $\epsilon_X(1 + o(1))$- and $\epsilon_Y(1 + o(1))$-multi-attributes with

$$\epsilon_X \triangleq \epsilon \sqrt{\frac{k}{K_X + k}} \quad \text{and} \quad \epsilon_Y \triangleq \epsilon \sqrt{\frac{k}{K_Y + k}}. \tag{G.55}$$

In particular, the multi-attribute property (9.61) specialized to U is obtained via

$$\begin{aligned} \left\| \mathbf{\Phi}^{X|U} \right\|_{\mathrm{s}}^2 &= \lambda_{\max}\left(\left(\mathbf{\Phi}^{X|U} \right)^{\mathrm{T}} \mathbf{\Phi}^{X|U} \right) \\ &= \frac{1}{\epsilon_X^2} \lambda_{\max}\left(\mathbf{\Lambda}_{XU}^{\mathrm{T}} \mathbf{\Lambda}_X^{-1} \mathbf{\Lambda}_{XU} \right) \\ &= \max_{i \in \{1,\dots,k\}} \left\| \boldsymbol{\phi}^{X|U_i} \right\|^2 \\ &\leq \frac{K_X + k}{k} (1 + o(1)), \end{aligned}$$

where to obtain the second equality we have used that $\mathbf{\Lambda}_U = \mathbf{I}$, and to obtain third equality we have used that $\mathbf{\Lambda}_{XU}^{\mathrm{T}}\mathbf{\Lambda}_X^{-1}\mathbf{\Lambda}_{XU}$ is diagonal. To obtain the inequality, note that with ϵ_X as defined,

$$I(U_i;X) = \epsilon_X^2 \|\boldsymbol{\phi}^{X|U_i}\|^2 + o(\epsilon_X^2) \le \epsilon^2 = \epsilon_X^2 \left(\frac{K_X + k}{k}\right),$$

whence

$$\|\boldsymbol{\phi}^{X|U_i}\|^2 \le \frac{K_X + k}{k}(1 + o(1)).$$

The corresponding result for V follows from symmetry. Finally, substituting for ϵ_X and ϵ_Y in the right-hand side of (9.77) yields (9.79), ■

G.17 Proof of Corollary 9.28

First, (9.84) holds from the conditional independence in the definition of an attribute. Next, since the variables are jointly Gaussian, it suffices to obtain the associated second moment characterization. In particular, using that $\mathbf{\Phi}^{X|U} = \mathbf{\Psi}_{(k)}^X$ and (9.24a) we have

$$\mathbb{E}[U|X] = \epsilon \left(\mathbf{\Phi}^{X|U}\right)^{\mathrm{T}} \tilde{X} = \epsilon S_{(k)}^* \tag{G.56}$$

and

$$\mathbf{\Lambda}_{U|X} = \mathbf{I} - \epsilon^2 \left(\mathbf{\Phi}^{X|U}\right)^{\mathrm{T}} \mathbf{\Phi}^{X|U} = \mathbf{I} - \epsilon^2 \left(\mathbf{F}_{(k)}^*\right)^{\mathrm{T}} \mathbf{\Lambda}_X \mathbf{F}_{(k)}^* = (1 - \epsilon^2)\,\mathbf{I}, \tag{G.57}$$

whence (9.85a). And via a symmetry argument (corresponding to interchanging the roles of X and Y, and U and V), we obtain (9.85b) from (9.85a).

Next,

$$\begin{aligned} P_{U|S_{(k)}^*,T_{(k)}^*,V}(u|s_{(k)}^*, t_{(k)}^*, v) &= \frac{P_{U,V|S_{(k)}^*,T_{(k)}^*}(u,v|s_{(k)}^*, t_{(k)}^*)}{P_{V|S_{(k)}^*,T_{(k)}^*}(v|s_{(k)}^*, t_{(k)}^*)} \\ &= \frac{P_{U,V|X,Y}(u,v|x,y)}{P_{V|X,Y}(v|x,y)} \\ &= P_{U|X}(u|x) \\ &= P_{U|S_{(k)}^*}(u|s_{(k)}^*). \end{aligned} \tag{G.58}$$

Verifying (9.86a), and (9.86b) follows from symmetry considerations.

Finally, to obtain (9.87a) we have

$$\begin{aligned}\mathbb{E}[V|X] &= \epsilon\big(\mathbf{\Phi}^{Y|V}\big)^{\mathrm{T}}\tilde{\mathbf{B}}\,\tilde{X}\\ &= \epsilon\big(\mathbf{\Psi}^{Y}_{(k)}\big)^{\mathrm{T}}\tilde{\mathbf{B}}\,\tilde{X}\\ &= \epsilon\mathbf{\Sigma}_{(k)}\,\big(\mathbf{\Psi}^{X}_{(k)}\big)^{\mathrm{T}}\tilde{X}\\ &= \epsilon\mathbf{\Sigma}_{(k)}\,S^{*}_{(k)},\end{aligned}$$

and

$$\begin{aligned}\mathbf{\Lambda}_{V|X} &= \mathbf{I} - \epsilon^2\,\big(\mathbf{\Phi}^{Y|V}\big)^{\mathrm{T}}\tilde{\mathbf{B}}\,\tilde{\mathbf{B}}^{\mathrm{T}}\mathbf{\Phi}^{Y|V}\\ &= \mathbf{I} - \epsilon^2\,\big(\mathbf{\Psi}^{Y}\big)^{\mathrm{T}}\tilde{\mathbf{B}}\,\tilde{\mathbf{B}}^{\mathrm{T}}\mathbf{\Psi}^{Y}\\ &= \mathbf{I} - \epsilon^2\,\mathbf{\Sigma}_{(k)}\,\big(\mathbf{\Psi}^{X}\big)^{\mathrm{T}}\mathbf{\Psi}^{X}\,\mathbf{\Sigma}_{(k)}\\ &= \mathbf{I} - \epsilon^2\,\mathbf{\Sigma}^2_{(k)}.\end{aligned}$$

Via a symmetry argument, we obtain (9.87b) from (9.87a). ■

G.18 Proof of Proposition 9.29

First, since the constraints on U and V coincide with those of Corollary 9.27, from the proof of the latter we obtain that U and V are $\epsilon_X(1+o(1))$- and $\epsilon_Y(1+o(1))$-multi-attributes with ϵ_X and ϵ_Y as given by (G.55).

As such, for the maximization of $I(U;Y)$ we have

$$\begin{aligned}I(Y;U) &= I(\tilde{Y};U)\\ &= \frac{\epsilon_X^2}{2}\|\tilde{\mathbf{B}}\,\mathbf{\Phi}^{X|U}\|_{\mathrm{F}}^2 + o(\epsilon_X^2) && \text{(G.59)}\\ &= \frac{\epsilon_X^2}{2}\|\tilde{\mathbf{B}}\,\tilde{\mathbf{\Phi}}^{X|U}\mathbf{\Delta}^{X|U}\|_{\mathrm{F}}^2 + o(\epsilon_X^2) && \text{(G.60)}\\ &\le \frac{\epsilon_X^2}{2}\|\tilde{\mathbf{B}}\,\tilde{\mathbf{\Phi}}^{X|U}\|_{\mathrm{F}}^2\|\mathbf{\Delta}^{X|U}\|_{\mathrm{s}}^2 + o(\epsilon_X^2) && \text{(G.61)}\\ &\le \frac{\epsilon_X^2}{2}\|\tilde{\mathbf{B}}\,\tilde{\mathbf{\Phi}}^{X|U}\|_{\mathrm{F}}^2\left(\frac{K_X+k}{k}\right) + o(\epsilon_X^2) && \text{(G.62)}\\ &\le \frac{\epsilon^2}{2}\sum_{i=1}^{k}\sigma_i^2 + o(\epsilon^2), && \text{(G.63)}\end{aligned}$$

where to obtain (G.59) we have used Lemma 9.11 with Corollary 9.16, to obtain (G.60) we have used the SVD (G.45), to obtain (G.61) we have used Fact 5.12, to obtain (G.62) we have used the (G.47), and to obtain (G.63) we have used Lemma 3.1 with (G.46). Finally, it is straightforward to verify that the inequalities all hold with equality when $\mathbf{\Phi}^{X|U}$ is chosen according to (G.48b), which corresponds to (9.80a).

Via a symmetry argument (corresponding to interchanging the roles of X and Y, and U and V, and noting that $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{B}}^{\mathrm{T}}$ share the same singular values), it follows that

$$I(V;X) \le \frac{\epsilon^2}{2}\sum_{i=1}^{k}\sigma_i^2 + \mathrm{o}(\epsilon^2),$$

where the inequality holds with equality when $\mathbf{\Phi}^{Y|V}$ is chosen according to (G.40b), which corresponds (9.80b). ■

G.19 Proof of Proposition 9.30

To simplify the exposition, we first consider the case $K_X = K_Y = K$. First, as defined in (9.24), $S^* \triangleq S^*_{(K)}$ and $T^* \triangleq S^*_{(K)}$ are invertible transformations of X and Y, respectively, with [cf. (9.17)]

$$\mathbf{\Lambda}_{S^*} = \mathbf{\Lambda}_{T^*} = \mathbf{I} \quad \text{and} \quad \mathbf{\Lambda}_{T^* S^*} = \mathbf{\Sigma}$$

and $S^* \leftrightarrow X \leftrightarrow W \leftrightarrow Y \leftrightarrow T^*$, a special case of which is

$$S_i^* \leftrightarrow W \leftrightarrow T_i^*, \qquad i = 1, \dots, K. \tag{G.64}$$

Then

$$\begin{aligned}
&I(W;X,Y)\\
&\quad = I(W;S^*,T^*) && \text{(G.65)}\\
&\quad = \sum_{i=1}^{K} I(W;S_i^*,T_i^* | (S^*)^{i-1},(T^*)^{i-1}) && \text{(G.66)}\\
&\quad = \sum_{i=1}^{K} I(S_i^*,T_i^*;W,(S^*)^{i-1},(T^*)^{i-1}) - \underbrace{I(S_i^*,T_i^*;(S^*)^{i-1},(T^*)^{i-1})}_{=0} && \text{(G.67)}
\end{aligned}$$

$$= \sum_{i=1}^{K} I(W; S_i^*, T_i^*) + \underbrace{I((S^*)^{i-1}, (T^*)^{i-1}; S_i^*, T_i^*|W)}_{\geq 0} \tag{G.68}$$

$$\geq \sum_{i=1}^{K} I(W; S_i^*, T_i^*) \tag{G.69}$$

$$= \sum_{i=1}^{K} I(W; S_i^*) + I(W; , T_i^*|S_i^*) \tag{G.70}$$

$$= \sum_{i=1}^{K} I(W; S_i^*) + (I(S_i^*, W; T_i^*) - I(S_i^*; T_i^*)) \tag{G.71}$$

$$= \sum_{i=1}^{K} I(W; S_i^*) + I(W; T_i^*) - I(S_i^*; T_i^*) \tag{G.72}$$

$$\geq \sum_{i=1}^{K} I(\mathbb{E}[S_i^*|W]; S_i^*) + I(\mathbb{E}[T_i^*|W]; T_i^*) - I(S_i^*; T_i^*) \tag{G.73}$$

$$\geq \sum_{i=1}^{K} R_{S_i^*}(D_{S_i^*}) + R_{T_i^*}(D_{T_i^*}) - I(S_i^*; T_i^*) \tag{G.74}$$

$$= \frac{1}{2} \sum_{i=1}^{K} \log \frac{1 - \sigma_i^2}{D_{S_i^*} D_{T_i^*}}, \tag{G.75}$$

where to obtain (G.65) we have used the invariance of mutual information to coordinate transformation, to obtain (G.66) have used chain rule of mutual information, to obtain (G.67) we have used the chain rule of mutual information and note that the second term is zero since the S_i^*, T_i^* are independent of $(S^*)^{i-1}, (T^*)^{i-1}$, to obtain (G.68) we have used the chain rule of mutual information on the first term in (G.67), to obtain (G.69) we neglect the second term in (G.68), to obtain (G.70) we use the chain rule of mutual information, to obtain (G.71) we use the chain rule of information on the second term in (G.70), to obtain (G.72) we have used (G.64), to obtain (G.73) we have used the data processing inequality, and to obtain (G.74) we have used the definition of the (Gaussian) rate-distortion function (see, e.g., [63, Chapter 10]).

Now

$$D_{S_i^*} = \mathbb{E}\Big[\big(S_i^* - \mathbb{E}[S_i^*|W]\big)^2\Big] = 1 - \underbrace{\mathbb{E}\Big[\mathbb{E}[S_i^*|W]^2\Big]}_{\triangleq \delta_{S_i^*}} \tag{G.76}$$

$$D_{T_i^*} = \mathbb{E}\Big[\big(T_i^* - \mathbb{E}[T_i^*|W]\big)^2\Big] = 1 - \underbrace{\mathbb{E}\Big[\mathbb{E}[T_i^*|W]^2\Big]}_{\triangleq \delta_{T_i^*}} \tag{G.77}$$

and

$$\begin{aligned} \sigma_i^2 &= \mathbb{E}[\mathbb{E}[S_i^* T_i^*|W]]^2 \\ &= \mathbb{E}[\mathbb{E}[S_i^*|W]\,\mathbb{E}[T_i^*|W]]^2 \\ &\le \mathbb{E}\Big[\mathbb{E}[S_i^*|W]^2\Big]\,\mathbb{E}\Big[\mathbb{E}[T_i^*|W]^2\Big] \\ &= \delta_{S_i^*}\delta_{T_i^*}, \end{aligned} \tag{G.78}$$

where we have used the Cauchy-Schwarz inequality.

Hence, the lower bound (G.75) is minimized by maximizing $(1 - \delta_{S_i^*})(1 - \delta_{T_i^*})$ for each $i \in \{1, \ldots, K\}$, subject to the constraint (G.78), which is a straightforward exercise, yielding $\delta_{S_i^*} = \delta_{T_i^*} = \sigma_i$, whence

$$I(W; X, Y) \ge \frac{1}{2}\sum_{i=1}^{K} \log\left(\frac{1+\sigma_i}{1-\sigma_i}\right).$$

To show that the lower bound is achieved, choose W such that W, X, Y are jointly Gaussian, and let $W \in \mathbb{R}^K$ be zero-mean with $\mathbf{\Lambda}_W = \mathbf{I}$. Finally, choose

$$\mathbf{\Lambda}_{S^*W} = \mathbf{\Lambda}_{T^*W} = \mathbf{\Sigma}^{1/2}. \tag{G.79}$$

Then using Fact 9.22, we confirm that with W so defined, $S^* \leftrightarrow W \leftrightarrow T^*$ since

$$\mathbf{\Lambda}_{S^*W}\,\mathbf{\Lambda}_{T^*W}^{\mathrm{T}} = \mathbf{\Sigma} = \mathbf{\Lambda}_{S^*T^*}.$$

Finally, exploiting the resulting conditional independence structure, we

The version of record is available at: http://dx.doi.org/10.1561/0100000107

have

$$\begin{aligned}
I(W;S^*,T^*) &= \sum_{i=1}^{K} I(W_i;S_i^*,T_i^*) \\
&= \sum_{i=1}^{K} I(W_i;S_i^*) + I(W_i;T_i^*) - I(S_i^*;T_i^*) \\
&= -\frac{1}{2}\sum_{i=1}^{K}\Big[\log(1-\lambda_{S_i^*W_i}) + \log(1-\lambda_{T_i^*W_i}) \\
&\qquad\qquad\qquad + \log(1-\lambda_{S_i^*T_i^*})\Big] \\
&= \frac{1}{2}\sum_{i=1}^{K}\log\frac{1-\sigma_i^2}{(1-\sigma_i)(1-\sigma_i)} \\
&= \frac{1}{2}\sum_{i=1}^{K}\log\frac{1+\sigma_i}{1-\sigma_i}.
\end{aligned}$$

The extension to the case $K_X \neq K_Y$ is straightforward. In particular, when $K_X \leq K_Y$, we write

$$\tilde{\mathbf{B}} = \mathbf{\Psi}_-^Y\,\mathbf{\Sigma}_-(\mathbf{\Psi}^X)^{\mathrm{T}},$$

where the $K_Y \times K_X$ matrix $\mathbf{\Psi}_-^Y$ is formed from the first K_X columns of $\mathbf{\Psi}^Y$, and where the $K_X \times K_X$ matrix $\mathbf{\Sigma}_-$ is formed from the first K_X rows of $\mathbf{\Sigma}$ (the rest being all zeros). In the associated analysis, we then replace (G.79) with

$$\begin{aligned}
\mathbf{\Lambda}_{S^*W} &= \mathbf{\Sigma}_-^{1/2} \\
\mathbf{\Lambda}_{T^*W} &= \mathbf{\Sigma}^{1/2}.
\end{aligned}$$

Likewise, when $K_X > K_Y$ we write

$$\tilde{\mathbf{B}} = \mathbf{\Psi}^Y\,\tilde{\mathbf{\Sigma}}_-(\mathbf{\Psi}_-^X)^{\mathrm{T}},$$

where the $K_X \times K_Y$ matrix $\mathbf{\Psi}_-^X$ is formed from the first K_Y columns of $\mathbf{\Psi}^X$, and where the $K_Y \times K_Y$ matrix $\tilde{\mathbf{\Sigma}}_-$ is formed from the first K_Y columns of $\mathbf{\Sigma}$ (the rest being all zeros). In this case, the associated analysis replaces (G.79) with

$$\begin{aligned}
\mathbf{\Lambda}_{S^*W} &= \mathbf{\Sigma}^{1/2} \\
\mathbf{\Lambda}_{T^*W} &= \tilde{\mathbf{\Sigma}}_-^{1/2}.
\end{aligned}$$

The result in both cases can be expressed in the form

$$\begin{aligned}\mathbf{\Lambda}_{XW} &= \mathbf{\Lambda}_X^{1/2}\,\mathbf{\Psi}_{(K)}^X\,\mathbf{\Sigma}_{(K)}^{1/2}\\ \mathbf{\Lambda}_{YW} &= \mathbf{\Lambda}_Y^{1/2}\,\mathbf{\Psi}_{(K)}^Y\,\mathbf{\Sigma}_{(K)}^{1/2},\end{aligned}$$

and, in turn, (9.90). ■

G.20 Proof of Corollary 9.31

With

$$A^* \triangleq \begin{bmatrix} S^* \\ T^* \end{bmatrix}$$

we have, as a straightforward exercise in linear algebra,

$$\begin{aligned}\mathbb{E}[W|X,Y] &= \mathbb{E}[W|S^*,T^*]\\ &= \mathbf{\Lambda}_{W,A^*}\,\mathbf{\Lambda}_{A^*}^{-1}A^*\\ &= \begin{bmatrix}\mathbf{\Sigma}_{(K)}^{1/2} & \mathbf{\Sigma}_{(K)}^{1/2}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{\Sigma}_{(K)}\\ \mathbf{\Sigma}_{(K)} & \mathbf{I}\end{bmatrix}^{-1}\begin{bmatrix}S_{(K)}^*\\ T_{(K)}^*\end{bmatrix}\\ &= \mathbf{\Sigma}_{(K)}^{1/2}\big(\mathbf{I}+\mathbf{\Sigma}_{(K)}\big)^{-1}\begin{bmatrix}\mathbf{I} & \mathbf{I}\end{bmatrix}\begin{bmatrix}S_{(K)}^*\\ T_{(K)}^*\end{bmatrix},\end{aligned}$$

and

$$\begin{aligned}\mathbf{\Lambda}_{W|X,Y} &= \mathbf{\Lambda}_{W|S^*,T^*}\\ &= \mathbf{I} - \mathbf{\Lambda}_{W|A^*}^{\mathrm{T}}\mathbf{\Lambda}_{A^*}^{-1}\mathbf{\Lambda}_{W|A^*}\\ &= \mathbf{I} - 2\mathbf{\Sigma}_{(K)}\big(\mathbf{I}+\mathbf{\Sigma}_{(K)}\big)^{-1}\\ &= \big(\mathbf{I}-\mathbf{\Sigma}_{(K)}\big)\big(\mathbf{I}+\mathbf{\Sigma}_{(K)}\big)^{-1}.\end{aligned}$$

■

G.21 Proof of Corollary 9.32

It suffices to verify that $\check{W}$ so defined has $\mathbf{\Lambda}_{\check{W}X}$ and $\mathbf{\Lambda}_{\check{W}Y}$ matching (9.90), i.e., that

$$\begin{bmatrix}\mathbf{\Lambda}_{\check{W}X} & \mathbf{\Lambda}_{\check{W}Y}\end{bmatrix} = \begin{bmatrix}\mathbf{\Lambda}_{\check{W}U} & \mathbf{\Lambda}_{\check{W}V}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{\Lambda}_{UV}\\ \mathbf{\Lambda}_{UV}^{\mathrm{T}} & \mathbf{I}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{\Lambda}_{UX} & \mathbf{\Lambda}_{UY}\\ \mathbf{\Lambda}_{VX} & \mathbf{\Lambda}_{VY}\end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{\Lambda}_{WX} & \mathbf{\Lambda}_{WY} \end{bmatrix} \tag{G.80}$$

holds. But from (9.80) and (9.81) with $\epsilon = 1$, and using Fact 9.22 with (9.18), we have

$$\begin{bmatrix} \mathbf{\Lambda}_{UX} & \mathbf{\Lambda}_{UY} \\ \mathbf{\Lambda}_{VX} & \mathbf{\Lambda}_{VY} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{\Sigma}_{(K)} \\ \mathbf{\Sigma}_{(K)} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \left(\mathbf{F}^*_{(K)}\right)^{\mathrm{T}} \mathbf{\Lambda}_X & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{G}^*_{(K)}\right)^{\mathrm{T}} \mathbf{\Lambda}_Y \end{bmatrix}$$

and $\mathbf{\Lambda}_{UV} = \mathbf{\Sigma}_{(K)}$, and from (9.90) we have

$$\begin{bmatrix} \mathbf{\Lambda}_{WX} & \mathbf{\Lambda}_{WY} \end{bmatrix} = \mathbf{\Sigma}_{(K)}^{1/2} \begin{bmatrix} \left(\mathbf{F}^*_{(K)}\right)^{\mathrm{T}} \mathbf{\Lambda}_X & \left(\mathbf{G}^*_{(K)}\right)^{\mathrm{T}} \mathbf{\Lambda}_Y \end{bmatrix}.$$

Hence, with the choices (9.96), it follows that (G.80) holds. ■

G.22 Proof of Lemma 9.35

First, note that

$$\begin{aligned} &\mathbb{E}\left[\left\| V(y) - V_\circ(x) \right\|^2\right] \\ &\quad = \mathbb{E}\left[\operatorname{tr}\left(\left(V(y) - V_\circ(x)\right)\left(V(y) - V_\circ(x)\right)^{\mathrm{T}}\right)\right] \\ &\quad = \operatorname{tr} \mathbf{\Lambda}_{V(y) - V_\circ(x)} + \left\| \mathbb{E}[V(y) - V_\circ(x)] \right\|^2 && \text{(G.81)} \\ &\quad = \operatorname{tr}(\mathbf{\Lambda}_{V|Y}) + \operatorname{tr}(\mathbf{\Lambda}_{V|X}) + \left\| \mathbb{E}[V|Y = y] - \mathbb{E}[V|X = x] \right\|^2, && \text{(G.82)} \end{aligned}$$

where to obtain (G.81) we have used the trivial identity $\mathbb{E}[ZZ^{\mathrm{T}}] = \mathbf{\Lambda}_Z + \mathbb{E}[Z]\mathbb{E}[Z]^{\mathrm{T}}$, and to obtain (G.82) we have used that $V(y)$ and $V_\circ(x)$ are independent and distributed according to $P_{V|Y}(\cdot|y)$ and $P_{V|X}(\cdot|x)$, respectively.

Finally, substituting (9.85b) and (9.87a) from Corollary 9.28 into (G.82), we obtain

$$\mathbb{E}\left[\left\| V(y) - V_\circ(x) \right\|^2\right] = 2k - k\,\epsilon^2 - \epsilon^2 \sum_{i=1}^{k} \sigma_i^2 + \epsilon^2 \sum_{i=1}^{k} \left(g_i(y) - \sigma_i\, f_i(x)\right)^2$$

and the lemma follows. ■

G.23 Proof of Proposition 9.36

First, we rewrite (9.111) in the equivalent form

$$\left(\boldsymbol{\Psi}^Y_{(k)}\right)^{\mathrm{T}} \tilde{y} = \boldsymbol{\Sigma}_{(k)} \left(\boldsymbol{\Psi}^X_{(k)}\right)^{\mathrm{T}} \tilde{x}. \tag{G.83}$$

Next, we note that the objective function $P_Y(y)$ in (9.112) is a monotonically decreasing function of $\|\tilde{y}\|$, and thus we seek to find the minimum norm solution $\tilde{y}^*(x)$ to (G.83). Via familiar linear algebra results—see, e.g., [114, Problem 7.3.P9]—the solution follows as

$$\tilde{y}^*(x) = \left(\boldsymbol{\Psi}^Y_{(k)}\right)^{\dagger\mathrm{T}} \boldsymbol{\Sigma}_{(k)} \left(\boldsymbol{\Psi}^X_{(k)}\right)^{\mathrm{T}} x = \boldsymbol{\Psi}^Y_{(k)} \boldsymbol{\Sigma}_{(k)} \left(\boldsymbol{\Psi}^X_{(k)}\right)^{\mathrm{T}} \tilde{x}. \tag{G.84}$$

Rewriting (G.84) in terms of y and x and using standard pseudoinverse properties then yields (9.113). ■

G.24 Proof of Proposition 9.39

We have

$$\begin{aligned} D(P_{X,Y} \,\|\, P^{(k)}_{X,Y}) &= \|\tilde{\mathbf{B}} - \tilde{\mathbf{B}}^{(k)}\|_{\mathrm{F}}^2 + o(\epsilon^2) & \text{(G.85)} \\ &\geq \sum_{i=k+1}^{K} \sigma_i^2 + o(\epsilon^2), & \text{(G.86)} \end{aligned}$$

where to obtain (G.85) we have used Lemma 9.17 with $\tilde{\mathbf{B}}$ as defined in (9.11) and

$$\tilde{\mathbf{B}}^{(k)} \triangleq \boldsymbol{\Lambda}_Y^{-1/2} \boldsymbol{\Lambda}^{(k)}_{YX} \boldsymbol{\Lambda}_X^{-1/2}, \tag{G.87}$$

and to obtain (G.86) we have used Lemma 7.3 together with the fact that $\mathrm{rank}(\tilde{\mathbf{B}}^{(k)}) \leq k$ since $\mathrm{rank}(\boldsymbol{\Lambda}^{(k)}_{XY}) \leq k$ and $\boldsymbol{\Lambda}_X$ and $\boldsymbol{\Lambda}_Y$ are positive definite. Finally, it is straightforward to verify that the inequality (G.86) holds with equality when $\tilde{\mathbf{B}}^{(k)} = \tilde{\mathbf{B}}^{(k)}_*$, with $\tilde{\mathbf{B}}^{(k)}_*$ as given in (9.57). ■

G.25 Proof of Proposition 9.40

First, with the gain matrix [cf. (9.5)] $\boldsymbol{\Gamma}_{Y|X} = \boldsymbol{\Lambda}_{YX}\, \boldsymbol{\Lambda}_X^{-1}$ we rewrite the optimization in (9.121a) as

$$\underset{\tilde{\boldsymbol{\Gamma}}_{Y|X}:\ \mathrm{rank}(\tilde{\boldsymbol{\Gamma}}_{Y|X}) \leq k}{\arg\min} \quad \mathbb{E}_{P_{X,Y}}\left[\left\|Y - \tilde{\boldsymbol{\Gamma}}_{Y|X} X\right\|^2\right]$$

$$\begin{aligned}
&= \operatorname*{arg\,min}_{\tilde{\mathbf{\Gamma}}_{Y|X}:\ \mathrm{rank}(\tilde{\mathbf{\Gamma}}_{Y|X})\leq k} \Big(\mathbb{E}_{P_{X,Y}}\left[\|Y - \mathbf{\Gamma}_{Y|X}X\|^2\right] \\
&\qquad\qquad + \mathbb{E}_{P_X}\left[\|(\mathbf{\Gamma}_{Y|X} - \tilde{\mathbf{\Gamma}}_{Y|X})X\|^2\right]\Big) \qquad \text{(G.88)} \\
&= \operatorname*{arg\,min}_{\tilde{\mathbf{\Gamma}}_{Y|X}:\ \mathrm{rank}(\tilde{\mathbf{\Gamma}}_{Y|X})\leq k} \mathbb{E}_{P_X}\left[\|(\mathbf{\Gamma}_{Y|X} - \tilde{\mathbf{\Gamma}}_{Y|X})X\|^2\right] \qquad \text{(G.89)} \\
&= \operatorname*{arg\,min}_{\tilde{\mathbf{\Gamma}}_{Y|X}:\ \mathrm{rank}(\tilde{\mathbf{\Gamma}}_{Y|X})\leq k} \|(\mathbf{\Gamma}_{Y|X} - \tilde{\mathbf{\Gamma}}_{Y|X})\mathbf{\Lambda}_X^{1/2}\|_{\mathrm{F}}^2, \qquad \text{(G.90)}
\end{aligned}$$

where to obtain (G.88) we have used the orthogonal properties of the error in the MMSE estimate, and to obtain (G.89) we have used that the first term does not depend on $\tilde{\mathbf{\Gamma}}_{Y|X}$.

In turn, with

$$\mathbf{A} \triangleq \mathbf{\Gamma}_{Y|X}\mathbf{\Lambda}_X^{1/2} = \mathbf{\Lambda}_{YX}\,\mathbf{\Lambda}_X^{-1/2} \quad \text{and} \quad \tilde{\mathbf{A}} \triangleq \tilde{\mathbf{\Gamma}}_{Y|X}\mathbf{\Lambda}_X^{1/2}, \tag{G.91}$$

we can rewrite (G.90) in the form

$$\min_{\tilde{\mathbf{\Gamma}}_{Y|X}:\ \mathrm{rank}(\tilde{\mathbf{\Gamma}}_{Y|X})\leq k} \|(\mathbf{\Gamma}_{Y|X} - \tilde{\mathbf{\Gamma}}_{Y|X})\mathbf{\Lambda}_X^{1/2}\|_{\mathrm{F}}^2 = \min_{\tilde{\mathbf{A}}:\ \mathrm{rank}(\tilde{\mathbf{A}})\leq k} \|\mathbf{A} - \tilde{\mathbf{A}}\|_{\mathrm{F}}^2, \tag{G.92}$$

since $\mathrm{rank}(\tilde{\mathbf{A}}) \leq k$ if and only if $\mathrm{rank}(\tilde{\mathbf{\Gamma}}_{Y|X}) \leq k$ since $\mathbf{\Lambda}_X$ is nonsingular. Hence, with the SVD for $\mathbf{A}$ expressed in the form (9.121c), it follows form Lemma 7.3 that the minimum on the right-hand side of (G.92) is achieved by the choice

$$\tilde{\mathbf{A}} = \tilde{\mathbf{\Psi}}_{(k)}^{Y}\tilde{\mathbf{\Sigma}}_{(k)}\big(\tilde{\mathbf{\Psi}}_{(k)}^{X}\big)^{\mathrm{T}} = \mathbf{A}\,\tilde{\mathbf{\Psi}}_{(k)}^{X}\,\big(\tilde{\mathbf{\Psi}}_{(k)}^{X}\big)^{\mathrm{T}},$$

i.e., using (G.91),

$$\begin{aligned}
\tilde{\mathbf{\Gamma}}_{Y|X} &= \mathbf{\Lambda}_{YX}\mathbf{\Lambda}_X^{-1/2}\,\tilde{\mathbf{\Psi}}_{(k)}^{X}\,\big(\tilde{\mathbf{\Psi}}_{(k)}^{X}\big)^{\mathrm{T}}\mathbf{\Lambda}_X^{-1/2} \\
&= \mathbf{\Lambda}_{YX}\Big(\big(\tilde{\mathbf{\Psi}}_{(k)}^{X}\big)^{\mathrm{T}}\mathbf{\Lambda}_X^{1/2}\Big)^{\dagger}\,\big(\tilde{\mathbf{\Psi}}_{(k)}^{X}\big)^{\mathrm{T}}\mathbf{\Lambda}_X^{-1/2} \\
&= \Big(\mathbf{\Lambda}_X^{-1/2}\,\tilde{\mathbf{\Psi}}_{(k)}^{X}\,\big(\mathbf{\Lambda}_X^{1/2}\tilde{\mathbf{\Psi}}_{(k)}^{X}\big)^{\dagger}\mathbf{\Lambda}_{XY}\Big)^{\mathrm{T}},
\end{aligned}$$

where we obtain the second and third equalities using standard pseudoinverse properties. Finally, since

$$\mathbf{\Lambda}_{XY}^{(k)\circ} = \mathbf{\Lambda}_X\tilde{\mathbf{\Gamma}}_{Y|X}^{\mathrm{T}},$$

we obtain (9.121). ■

References

[1] E. Abbe and L. Zheng, "A coordinate system for Gaussian networks," *IEEE Trans. Inform. Theory*, vol. 58, no. 2, Feb. 2012, pp. 721–733.

[2] A. Achille and S. Soatto, "Emergence of invariance and disentangling in deep representations," *J. Mach. Learn. Res.*, vol. 19, no. 1, Jan. 2018, pp. 1947–1980.

[3] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the Markov operator," *Ann. Prob.*, vol. 4, no. 6, Dec. 1976, pp. 925–939.

[4] R. Ahlswede and G. Körner, "On common information and related characteristics of correlated information sources," in *Proc. Prague Conf. Inform. Theory*, Prague, Czechoslovakia, Sep. 1974.

[5] R. Ahlswede and G. Körner, "Appendix: On common information and related characteristics of correlated information sources," in *General Theory of Information Transfer and Combinatorics*, R. Ahlswede, L. Bäumer, N. Cai, H. Aydinian, V. Blinovsky, C. Deppe, and H. Mashurian, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 664–677.

[6] S. Akaho, "A kernel method for canonical correlation analysis," in *Proc. Int. Meeting Psychometric Soc. (IMPS)*, Osaka, Japan, Jul. 2001.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[7] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Proc. Int. Conf. Learning Repr. (ICLR)*, Toulon, France, Apr. 2017.

[8] S.-I. Amari, *Information Geometry and Its Applications.* Tokyo, Japan: Springer, 2016.

[9] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry.* Oxford, UK: Oxford University Press, 2000.

[10] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On hypercontractivity and a data processing inequality," in *Proc. Int. Symp. Inform. Theory (ISIT)*, Honolulu, HI, Jun. 2014.

[11] V. Anantharam, A. A. Gohari, S. Kamath, and C. Nair, "On hypercontractivity and the mutual information between Boolean functions," in *Proc. Allerton Conf. Commun., Contr., Computing*, Monticello, IL, Oct. 2013.

[12] V. Anantharam, A. A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover," *CoRR*, vol. abs/1304.6133, 2013. arXiv: 1304.6133. URL: http://arxiv.org/abs/1304.6133.

[13] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Hoboken, NJ: Wiley, 2003.

[14] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Machine Learning (ICML)*, ser. Proc. Mach. Learn. Res. (PMLR), vol. 28, pp. 1247–1255, Atlanta, GA, Jun. 2013.

[15] R. Arora and K. Livescu, "Kernel CCA for multi-view learning of acoustic features using articulatory measurements," in *Proc. Symp. Mach. Learning Speech, Lang. Process.*, pp. 34–37, Portland, OR, Sep. 2012.

[16] F. Bach, J. Mairal, and J. Ponce, "Convex sparse matrix factorizations," CNRS, Paris, France, Tech. Rep. HAL-00345747, 2008.

[17] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *J. Mach. Learn. Res.*, vol. 3, Jul. 2002, pp. 1–48.

[18] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Dept. Statist., Univ. Calif., Berkeley, CA, Tech. Rep. 688, Apr. 2005.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[19] S. Baharlouei, M. Nouiehed, A. Beirami, and M. Razaviyayn, “Rényi fair inference,” in *Proc. Int. Conf. Learning Repr. (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020.

[20] S. Balakrishnan, K. Puniyani, and J. Lafferty, “Sparse additive functional and kernel CCA,” in *Proc. Int. Conf. Machine Learning (ICML)*, pp. 911–918, Edinburgh, Scotland, Jun. 2012.

[21] P. Baldi and K. Hornik, “Neural networks and principal component analysis: Learning from examples without local minima,” *Neural Netw.*, vol. 2, no. 1, 1989, pp. 53–58.

[22] Y. Bao, Y. Li, S.-L. Huang, L. Zhang, L. Zheng, A. R. Zamir, and L. Guibas, “An information-theoretic approach to transferability in task transfer learning,” in *Proc. Int. Conf. Image Processing (ICIP)*, Taipei, Taiwan, Sep. 2019.

[23] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities.* fairmlbook.org, 2019.

[24] A. R. Barron, “Universal approximation bounds for superpositions of a sigmoidal function,” *IEEE Trans. Inform. Theory*, vol. 39, no. 3, May 1993, pp. 930–945.

[25] A. R. Barron and C.-W. Sheu, “Approximation of density functions by sequences of exponential families,” *Ann. Stat.*, vol. 19, no. 3, Sep. 1991, pp. 1347–1369.

[26] A. Bastlevsky, *Statistical Factor Analysis and Related Methods.* Hoboken, NJ: Wiley, 1994.

[27] S. Basu, M. Bilenko, A. Banerjee, and R. Mooney, “Probabilistic semi-supervised clustering with constraints,” in *Semi-Supervised Learning*, O. Chapelle, B. Schölkopf, and A. Zien, Eds., Cambridge, MA: MIT Press, 2006, ch. 5, pp. 73–102.

[28] J. Bennett and S. Lanning, “The Netflix Prize,” in *Proc. KDD Cup and Workshop*, San Jose, CA, Aug. 2007.

[29] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, and R. Arora, “Deep generalized canonical correlation analysis,” in *Proc. Workshop Represent. Learning NLP (RePL4NLP)*, pp. 1–6, Florence, Italy, Aug. 2019.

[30] J. P. Benzécri, *L’Analyse des Données, Tôme 2: L’Analyse des Correspondances.* Paris, France: Dunod, 1973.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[31] J. P. Benzécri, *Correspondence Analysis Handbook.* New York, NY: Marcel Dekker, 1992.

[32] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, ser. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), 1994.

[33] C. M. Bishop, “Latent variable models,” in *Learning in Graphical Models*, M. I. Jordan, Ed., Cambridge, MA: MIT Press, 1999, pp. 371–403.

[34] C. M. Bishop, *Pattern Recognition and Machine Learning.* New York, NY: Springer, 2006.

[35] P. Biswal, “Hypercontractivity and its applications,” *CoRR*, vol. abs/1101.2913, 2011. arXiv: 1101.2913. URL: http://arxiv.org/abs/1101.2913.

[36] S. Borade and L. Zheng, “Euclidean information theory,” in *Proc. Int. Zurich Seminar Commun. (IZS)*, Zurich, Switzerland, Mar. 2008.

[37] D. Braess and T. Sauer, “Bernstein polynomials and learning theory,” *J. Approx. Theory*, vol. 128, no. 2, 2004, pp. 187–206.

[38] L. Breiman and J. H. Friedman, “Estimating optimal transformations for multiple regression and correlation,” *J. Am. Stat. Assoc.*, vol. 80, no. 391, Sep. 1985, pp. 580–598.

[39] D. R. Brillinger, *Time Series: Data Analysis and Theory.* New York, NY: Holt Rinehart and Winston, 1975.

[40] Y. Bu, J. K.-W. Lee, S. Das, R. Panda, D. Rajan, P. Sattigeri, and G. W. Wornell, “Fair selective classification via sufficiency,” in *Proc. Int. Conf. Machine Learning (ICML)*, ser. Proc. Mach. Learn. Res. (PMLR), vol. 139, pp. 6076–6086, (virtual), Jul. 2021.

[41] Y. Bu, T. T. Wang, and G. W. Wornell, “SDP methods for sensitivity-constrained privacy funnel and information bottleneck problems,” in *Proc. Int. Symp. Inform. Theory (ISIT)*, Melbourne, Australia, Jul. 2021.

[42] A. Buja, “Theory of bivariate ACE,” Dept. Statistics, University of Washington, Seattle, WA, Tech. Rep. 74, Dec. 1985.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[43] A. Buja, "Remarks on functional canonical variates, alternating least squares methods and ACE," *Ann. Stat.*, vol. 18, no. 3, 1990, pp. 1032–1069.

[44] E. van der Burg and J. de Leeuw, "Nonlinear canonical correlation," *Br. J. Math. Stat. Psychol.*, vol. 36, no. 1, May 1983, pp. 54–80.

[45] T. Cacoullos, "Estimation of a multivariate density," *Ann. Inst. Statist. Math.*, vol. 18, no. 1, Dec. 1966, pp. 179–189.

[46] F. P. Calmon, A. Makhdoumi, and M. Médard, "Fundamental limits of perfect privacy," in *Proc. Int. Symp. Inform. Theory (ISIT)*, Hong Kong, China, Jun. 2015.

[47] F. P. Calmon, A. Makhdoumi, M. Médard, M. Varia, M. Christiansen, and K. R. Duffy, "Principal inertia components and applications," *IEEE Trans. Inform. Theory*, vol. 63, no. 8, Aug. 2017, pp. 5011–5038.

[48] F. P. Calmon, M. Varia, and M. Médard, "An exploration of the role of principal inertia components in information theory," in *Proc. Inform. Theory Workshop (ITW)*, Hobart, TAS, Australia, Nov. 2014.

[49] F. P. Calmon, M. Varia, M. Médard, M. M. Christiansen, K. R. Duffy, and S. Tessaro, "Bounds on inference," in *Proc. Allerton Conf. Commun., Contr., Computing*, Monticello, IL, Oct. 2013.

[50] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, Jun. 2010, pp. 925–936.

[51] E. J. Candès and Y. Plan, "A probabilistic and RIPless theory of compressed sensing," *IEEE Trans. Inform. Theory*, vol. 57, no. 11, Nov. 2011, pp. 7235–7254.

[52] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations Comput. Math.*, vol. 9, no. 6, 2009, pp. 717–772.

[53] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inform. Theory*, vol. 56, no. 5, May 2010, pp. 2054–2080.

[54] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning.* Cambridge, MA: MIT Press, 2006.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[55] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, “Information bottleneck for Gaussian variables,” *J. Mach. Learn. Res.*, vol. 6, May 2005, pp. 165–188.

[56] J. Cheeger, “A lower bound for the smallest eigenvalue of the Laplacian,” in *Problems in Analysis*, R. C. Gunning, Ed., Princeton, NJ: Princeton University Press, 1970, pp. 195–199.

[57] H. Chen, J. Li, J. Gao, Y. Sun, Y. Hu, and B. Yin, “Maximally correlated principal component analysis based on deep parameterization learning,” *ACM Trans. Knowl. Discov. Data*, vol. 13, no. 4, Jul. 2019, pp. 1–17.

[58] M. A. Chmielewski, “Elliptically symmetric distributions: A review and bibliography,” *Int. Stat. Review*, vol. 49, no. 1, Apr. 1981, pp. 67–74.

[59] F. R. K. Chung, *Spectral Graph Theory*, ser. Regional Conference Series in Mathematics 92. Providence, RI: Am. Math. Soc., 1997.

[60] K. W. Church and P. Hanks, “Word association norms, mutual information, and lexicography,” *J. Comput. Linguist.*, vol. 16, no. 1, Mar. 1990, pp. 22–29.

[61] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, “Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps,” *Proc. Nat. Acad. Sci.*, vol. 102, no. 21, 2005, pp. 7426–7431.

[62] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, Sep. 1995, pp. 273–297.

[63] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY: John Wiley and Sons, 2006.

[64] D. R. Cox, “The regression analysis of binary sequences (with discussion),” *J. Roy. Stat. Soc., Ser. B*, vol. 20, no. 2, 1958, pp. 215–242.

[65] B. R. Crain, “Estimation of distributions using orthogonal expansions,” *Ann. Stat.*, vol. 2, no. 3, May 1974, pp. 454–463.

[66] P. Csáki and J. Fischer, “Contributions to the problem of maximal correlation,” *Publ. Math. Inst. Hung. Acad. Sci.*, vol. 5, 1960, pp. 325–337.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[67] P. Csáki and J. Fischer, "On bivariate stochastic connection," *Publ. Math. Inst. Hung. Acad. Sci.*, vol. 5, 1960, pp. 311–323.

[68] P. Csáki and J. Fischer, "On the general notion of maximum correlation," *Magyar Tud. Akad. Mat. Kutató Int Közl*, vol. 8, 1963, pp. 27–51.

[69] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Math. Hungarica*, vol. 2, no. 1–4, 1972, pp. 191–213.

[70] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge, UK: Cambridge University Press, 2011.

[71] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, 2004, pp. 417–528.

[72] D. M. Cvetković, M. Doob, and H. Sachs, *Spectra of Graphs, Theory and Application.* New York, NY: Academic Press, 1980.

[73] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control, Signals, Systems*, vol. 2, no. 4, 1989, pp. 303–314.

[74] A. P. Dawid, "Spherical matrix distributions and a multivariate model," *J. Roy. Stat. Soc., Ser. B*, vol. 39, 1977, pp. 254–261.

[75] A. Dembo, A. Kagan, and L. A. Shepp, "Remarks on the maximum correlation coefficient," *Bernoulli*, vol. 7, no. 2, 2001, pp. 343–350.

[76] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, vol. 38, 2nd ed., ser. Stochastic Modelling and Applied Probability. New York, NY: Springer, 1998.

[77] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc., B*, vol. 39, no. 1, 1977, pp. 1–38.

[78] P. Diaconis and D. W. Strook, "Geometric bounds for eigenvalues of Markov chains," *Ann. Appl. Prob.*, vol. 1, no. 1, 1991, pp. 36–61.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[79] M. Diaz, H. Wang, F. P. Calmon, and L. Sankar, “On the robustness of information-theoretic privacy measures and mechanisms,” *IEEE Trans. Inform. Theory*, vol. 66, no. 4, Apr. 2020, pp. 1949–1978.

[80] R. Dobrushin and B. Tsybakov, “Information transmission with additional noise,” *IEEE Trans. Inform. Theory*, vol. 8, no. 5, Sep. 1962, pp. 293–304.

[81] E. Domanovitz and U. Erez, “On the importance of asymmetry and monotonicity constraints in maximal correlation analysis,” in *Proc. Int. Symp. Inform. Theory (ISIT)*, pp. 3112–3116, Paris, France, Jul. 2019.

[82] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY: Wiley, 2000.

[83] G. K. Dziugaite and D. M. Roy, “Neural network matrix factorization,” *CoRR*, vol. abs/1511.06443, 2015. arXiv: 1511.06443. URL: http://arxiv.org/abs/1511.06443.

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

[85] S. Feizi, A. Makhdoumi, K. Duffy, M. Kellis, and M. Médard, “Network maximal correlation,” *IEEE Trans. Netw. Sci., Eng.*, vol. 4, no. 4, Oct. 2017, pp. 229–247.

[86] S. Feizi and D. Tse, “Maximally correlated principal component analysis,” *CoRR*, vol. abs/1702.05471, 2017. arXiv: 1702.05471. URL: https://arxiv.org/abs/1702.05471.

[87] M. Fiedler, “Algebraic connectivity of graphs,” *Czechoslovak Math. J.*, vol. 23, no. 98, 1973, pp. 298–305.

[88] M. Fiedler, “A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory,” *Czech. Math. J.*, vol. 25, no. 100, 1975, pp. 619–633.

[89] P. Gács and J. Körner, “Common information is far less than mutual information,” *Probl. Contr. Inform. Theory*, vol. 2, no. 2, 1973, pp. 149–162.

[90] A. Ganesh, J. Wright, X. Li, E. Candès, and Y. Ma, "Dense error correction for low-rank matrices via principal component pursuit," in *Proc. Int. Symp. Inform. Theory (ISIT)*, Austin, TX, Jun. 2010.

[91] H. Gebelein, "Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung," *Z. Angewandte Math., Mech.*, vol. 21, no. 6, 1941, pp. 364–379.

[92] A. Gersho and R. Gray, *Vector Quantization and Signal Compression.* Boston, MA: Kluwer Academic Press, 1991.

[93] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *Int. Stat. Rev.*, vol. 70, no. 3, 2002, pp. 419–435.

[94] A. Gifi, *Nonlinear Multivariate Analysis.* Chichester, UK: Wiley, 1990.

[95] D. V. Gokhale, "Iterative maximum likelihood estimation for discrete distributions," *Sankhyā B*, vol. 35, no. 3, Sep. 1973, pp. 293–298.

[96] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Commun. ACM*, vol. 35, no. 12, Dec. 1992, pp. 61–70.

[97] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 4th ed. Baltimore, MD: Johns Hopkins University Press, 2012.

[98] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3–4, Dec. 1953, pp. 237–264.

[99] I. J. Good, "Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables," *Ann. Math. Stat.*, vol. 34, no. 3, Sep. 1963, pp. 911–934.

[100] I. Goodfellow, J. Bengio, and A. Courville, *Deep Learning.* Cambridge, MA: MIT Press, 2017.

[101] V. Grari, S. Lamprier, and M. Detyniecki, "Fairness-aware neural Rényi minimization for continuous features," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI-20)*, pp. 2262–2268, (virtual), Jan. 2021.

[102] M. Greenacre, *Theory and Applications of Correspondence Analysis.* London, UK: Academic Press, 1984.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[103] M. Greenacre, *Correspondence Analysis in Practice*, 3rd ed. New York, NY: Chapman & Hall/CRC, 2016.

[104] M. Haber, “Maximum likelihood methods for linear and log-linear models in categorical data,” *Comp. Stat., Data Anal.*, vol. 3, May 1985, pp. 1–10.

[105] N. Halko, P. G. Martinsson, and J. A. Tropp, “Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions,” *SIAM Rev.*, vol. 53, no. 2, 2011, pp. 217–288.

[106] W. J. Hall, “On characterizing dependence in joint distributions,” in *Essays in Probability and Statistics*, R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao, and K. J. C. Smith, Eds., Chapel Hill, NC: University of North Carolina Press, 1970, pp. 339–376.

[107] T. S. Han and S. Verdú, “Approximation theory of output statistics,” *IEEE Trans. Inform. Theory*, vol. 39, no. 3, May 1993, pp. 752–772.

[108] E. J. Hannan, “The general theory of canonical correlation and its relation to functional analysis,” *J. Aust. Math. Soc.*, vol. 2, no. 2, Oct. 1960, pp. 229–242.

[109] W. K. Härdle and L. Simar, “Canonical correlation analysis,” in *Applied Multivariate Statistical Analysis*, 4th ed., Berlin, Germany: Springer, 2015, pp. 443–454.

[110] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods,” *Neural Comput.*, vol. 16, no. 12, Dec. 2004, pp. 2639–2664.

[111] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. Conf. Comp. Vision, Pattern Recog. (CVPR)*, pp. 770–778, 2016.

[112] H. O. Hirschfeld, “A connection between correlation and contingency,” *Proc. Cambridge Phil. Soc.*, vol. 31, 1935, pp. 520–524.

[113] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge, UK: Cambridge University Press, 1991.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[114] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. Cambridge, UK: Cambridge Uiversity Press, 2012.

[115] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural Netw.*, vol. 2, no. 5, 1989, pp. 359–366.

[116] H. Hotelling, “Analysis of a complex of statistical variables into principal components,” *J. Ed. Psych.*, vol. 24, 1933, pp. 417–441, 498–520.

[117] H. Hotelling, “Relations between two sets of variates,” *Biometrika*, vol. 28, 1936, pp. 321–377.

[118] H. Hsu, S. Asoodeh, S. Salamatian, and F. P. Calmon, “Generalizing bottleneck problems,” in *Proc. Int. Symp. Inform. Theory (ISIT)*, pp. 531–535, Vail, Colorado, Jun. 2018.

[119] H. Hsu, S. Salamatian, and F. P. Calmon, “Correspondence analysis using neural networks,” in *Proc. Int. Conf. Artif. Intell., Stat. (AISTATS)*, ser. Proc. Mach. Learn. Res. (PMLR), vol. 89, pp. 2671–2680, Naha, Japan, Apr. 2019.

[120] H. Hsu, S. Salamatian, and F. P. Calmon, “Generalizing correspondence analysis for applications in machine learning,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 44, no. 12, Dec. 2022, pp. 9347–9362.

[121] S.-L. Huang, “Communicating type classes through channels: An information geometric view,” in *Proc. Inform. Theory Workshop (ITW)*, Mumbai, India, Nov. 2022.

[122] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng, “An information-theoretic view of learning in high dimensions: Universal features, maximal correlations, bottlenecks, and common information,” in *Proc. Inform. Theory Appl. Workshop (ITA)*, San Diego, CA, Feb. 2018.

[123] S.-L. Huang, A. Makur, L. Zheng, and G. W. Wornell, “An information-theoretic approach to universal feature selection in high-dimensional inference,” in *Proc. Int. Symp. Inform. Theory (ISIT)*, Aachen, Germany, Jun. 2017.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[124] S.-L. Huang, G. W. Wornell, and L. Zheng, “Gaussian universal features, canonical correlations, and common information,” in *Proc. Inform. Theory Workshop (ITW)*, Guangzhou, China, Nov. 2018.

[125] S.-L. Huang and X. Xu, “On the robustness of noisy ACE algorithm and multi-layer residual learning,” in *Proc. Int. Symp. Inform. Theory (ISIT)*, Paris, France, Jul. 2019.

[126] S.-L. Huang and X. Xu, “On the sample complexity of HGR maximal correlation functions,” in *Proc. Inform. Theory Workshop (ITW)*, Visby, Sweden, Aug. 2019.

[127] S.-L. Huang and X. Xu, “On the sample complexity of HGR maximal correlation functions for large datasets,” *IEEE Trans. Inform. Theory*, vol. 67, no. 3, Mar. 2021, pp. 1951–1980.

[128] S.-L. Huang, X. Xu, and L. Zheng, “An information-theoretic approach to unsupervised feature selection for high-dimensional data,” *IEEE J. Select. Areas Inform. Theory*, vol. 1, no. 1, May 2020, pp. 157–166.

[129] S.-L. Huang, X. Xu, L. Zheng, and G. W. Wornell, “An information theoretic interpretation to deep neural networks,” in *Proc. Int. Symp. Inform. Theory (ISIT)*, Paris, France, Jul. 2019.

[130] S.-L. Huang, X. Xu, L. Zheng, and G. W. Wornell, “A local characterization for Wyner common information,” in *Proc. Int. Symp. Inform. Theory (ISIT)*, Los Angeles, CA, Jun. 2020.

[131] S.-L. Huang, X. Xu, L. Zheng, and G. W. Wornell, “An information theoretic interpretation to deep neural networks,” *Entropy*, vol. 24, no. 1, Jan. 2022.

[132] S.-L. Huang and L. Zheng, “Linear information coupling problems,” in *Proc. Int. Symp. Inform. Theory (ISIT)*, Cambridge, MA, Jul. 2012.

[133] S. Y. Huang, M.-H. Lee, and C. K. Hsiao, “Nonlinear measures of association with kernel canonical correlation analysis and applications,” *J. Stat. Planning, Inference*, vol. 139, no. 7, Jul. 2009, pp. 2162–2174.

[134] C. Ireland and S. Kullback, “Contingency tables with given marginals,” *Biometrica*, vol. 55, 1968, pp. 179–188.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[135] A. J. Izenman, “Reduced-rank regression for the multivariate linear model,” *J. Multivariate Anal.*, vol. 5, 1975, pp. 248–264.

[136] W. James and C. Stein, “Estimation with quadratic loss,” in *Proc. Berkeley Symp. Math. Statist. Prob.*, pp. 361–379, Berkeley, CA, 1961.

[137] H. Jeffreys, *Theory of Probability*, 2nd ed. Oxford, UK: Clarendon Press, 1948.

[138] F. Jelinek and R. L. Mercer, “Interpolated estimation of Markov source parameters from sparse data,” in *Proc. Workshop, Patt. Recogn. Practice*, pp. 381–397, Amsterdam, Netherlands, May 1980.

[139] W. E. Johnson, “Probability: Deductive and inductive problems,” *Mind*, vol. 41, no. 164, Oct. 1932, pp. 409–423.

[140] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY: Springer, 2002.

[141] S. Kamath and V. Anantharam, “A new dual to the Gács-Körner common information defined via the Gray-Wyner system,” in *Proc. Allerton Conf. Commun., Contr., Computing*, Monticello, IL, Sep. 2010.

[142] S. Kamath and V. Anantharam, “Non-interactive simulation of joint distributions: The Hirschfeld-Gebelein-Rényi maximal correlation and the hypercontractivity ribbon,” in *Proc. Allerton Conf. Commun., Contr., Computing*, Monticello, IL, Sep. 2012.

[143] W. Kang and S. Ulukus, “A new data processing inequality and its applications in distributed source and channel coding,” *IEEE Trans. Inform. Theory*, vol. 57, no. 1, Jan. 2010, pp. 56–69.

[144] J. Karhunen, E. Oja, L. Wang, R. Vigário, and J. Joutsensalo, “A class of neural networks for independent component analysis,” *IEEE Trans. Neural Netw.*, vol. 8, no. 3, May 1997, pp. 486–504.

[145] S. M. Katz, “Estimation of probabilities from sparse data for the language model component of a speech recognizer,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 3, Mar. 1984, pp. 400–401.

[146] J. Kay, “Canonical correlation analysis using a neural network,” in *Proc. Symp. Comp. Stat. (COMPSTAT)*, pp. 305–308, Neuchâtel, Switzerland, Aug. 1992.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[147] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inform. Theory*, vol. 56, no. 6, Jun. 2010, pp. 2980–2998.

[148] W. F. Kibble, "An extension of a theorem of Mehler's on Hermite polynomials," *Proc. Cambridge Phil. Soc.*, vol. 41, no. 1, Jun. 1945, pp. 12–15.

[149] G. Kimeldorf and A. R. Sampson, "Monotone dependence," *Ann. Stat.*, vol. 6, no. 4, Jul. 1978, pp. 895–903.

[150] G. Kirchoff, "Uber die auflosung der gleichungen, auf welche man bei der untersuchung der linearen verteilung galvanischer strome gefuhrt wird," *Ann. Phys. Chem.*, vol. 72, no. 12, 1847, pp. 497–508.

[151] A. Klami, S. Virtanen, and S. Kaski, "Bayesian canonical correlation analysis," *J. Mach. Learn. Res.*, vol. 14, no. 30, Apr. 2013, pp. 965–1003.

[152] R. Kneser and H. Ney, "Improved backing-off for M-gram language modeling," in *Proc. Int. Conf. Acoust. Speech, Signal Processing (ICASSP)*, pp. 181–184, Detroit, MI, May 1995.

[153] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, Aug. 2009, pp. 30–37.

[154] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *Am. Inst. Chem. Eng. (AIChE) J.*, vol. 37, no. 2, Feb. 1991, pp. 233–243.

[155] G. R. Kumar and T. A. Courtade, "Which Boolean functions are most informative?" In *Proc. Int. Symp. Inform. Theory (ISIT)*, Istanbul, Turkey, Jul. 2013.

[156] M. Kumar, A. Gramfort, and J. Nothman, *Machine Learning in Python code for BIRCH scikit-learn.sklearn.cluster.birch.* URL: https://github.com/scikit-learn/scikit-learn/blob/a24c8b46/sklearn/cluster/birch.py.

[157] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *Int. J. Neural Syst.*, vol. 10, no. 5, 2000, pp. 365–377.

[158] H. O. Lancaster, "A reconciliation of χ^2, considered from metrical and enumerative aspects," *Sankhyā*, vol. 13, no. 1–2, Dec. 1953, pp. 1–10.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[159] H. O. Lancaster, "Some properties of the bivariate normal distribution considered in the form of a contingency table," *Biometrika*, vol. 44, no. 1–2, Jun. 1957, pp. 289–292.

[160] H. O. Lancaster, "The structure of bivariate distributions," *Ann. Math. Stat.*, vol. 29, 1958, pp. 719–736.

[161] H. O. Lancaster, *The Chi-Squared Distribution*. New York, NY: Wiley, 1969.

[162] H. O. Lancaster, "Joint probability distributions in the Meixner classes," *J. Roy. Stat. Soc., Ser. B*, vol. 37, no. 3, 1975, pp. 434–443.

[163] P. S. Laplace, *Essai Philosophique sur les Probabilités*, 5th ed. Paris, France: Courcier, 1814.

[164] B. Le Roux and H. Rouanet, *Geometric Data Analysis: From Correspondence Analysis to Structured Data*. Dordrecht, Netherlands: Kluwer, 2004.

[165] L. Lebart, A. Morineau, and K. Warwick, *Multivariate Descriptive Statistical Analysis*. Chichester, UK: Wiley, 1984.

[166] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, Nov. 1998, pp. 2278–2324.

[167] Y. LeCun, C. Cortes, and C. J. C. Burges, *MNIST handwritten digit database*. URL: http://yann.lecun.com/exdb/mnist.

[168] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivariate Anal.*, vol. 88, 2004, pp. 365–411.

[169] J. K.-W. Lee, Y. Bu, P. Sattigeri, R. Panda, G. W. Wornell, L. Karlinsky, and R. Feris, "A maximal correlation framework for fair machine learning," *Entropy*, vol. 24, no. 4, Mar. 2022.

[170] J. K.-W. Lee, Y. Bu, P. Sattigeri, R. Panda, G. W. Wornell, L. Karlinsky, and R. Feris, "A maximal correlation framework to imposing fairness in machine learning," in *Proc. Int. Conf. Acoust. Speech, Signal Processing (ICASSP)*, Singapore, May 2022.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[171] J. K.-W. Lee, P. Sattigeri, and G. W. Wornell, “Learning new tricks from old dogs: Multi-source transfer learning from pretrained networks,” in *Advances Neural Inform. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2019.

[172] T.-W. Lee, M. Girolami, A. J. Bell, and T. J. Sejnowski, “A unifying information-theoretic framework for independent component analysis,” *Comput., Math., Appl.*, vol. 39, no. 11, Jun. 2000, pp. 1–21.

[173] O. Levy and Y. Goldberg, “Neural word embedding as implicit matrix factorization,” in *Advances Neural Inform. Process. Syst. (NIPS)*, pp. 2177–2185, Montréal, Canada, Dec. 2014.

[174] A. S. Lewis, “The convex analysis of unitarily invariant matrix functions,” *J. Convex Anal.*, vol. 2, no. 1/2, 1995, pp. 173–183.

[175] L. Li, Y. Li, S.-L. Huang, and L. Zhang, “Maximal correlation embedding network for multilabel learning with missing labels,” in *Proc. Int. Conf. Multimedia, Expo (ICME)*, Shanghai, China, Jul. 2019.

[176] M. Li, Y. Li, S.-L. Huang, and L. Zhang, “Semantically supervised maximal correlation for cross-modal retrieval,” in *Proc. Int. Conf. Image Processing (ICIP)*, Abu Dhabi, UAE, Oct. 2020.

[177] M. Li, X. Xu, S.-L. Huang, and L. Zhang, “Dual feature distributional regularization for defending against adversarial attacks,” in *Proc. Int. Conf. Neural Inform. Process. (ICONIP)*, Bali, Indonesia, Dec. 2021.

[178] J. Lian, Y. Li, W. Gu, S.-L. Huang, and L. Zhang, “Joint mobility pattern mining with urban region partitions,” in *Proc. Int. Conf. Mobile Ubiquit. Syst. (MobiQuitous)*, pp. 362–371, Nov. 2018.

[179] J. Lian, Y. Li, W. Gu, S.-L. Huang, and L. Zhang, “Mining regional mobility patterns for urban dynamic analytics,” *Mobile Netw. Appl.*, vol. 25, Apr. 2019, pp. 459–473.

[180] J. Lian, Y. Li, S.-L. Huang, and L. Zhang, “Mining mobility patterns with trip-based traffic analysis zones: A deep feature embedding approach,” in *Proc. Intell. Transport. Syst. Conf. (ITSC)*, Auckland, New Zealand, Oct. 2019.

[181] Y. Liang, F. Ma, Y. Li, and S.-L. Huang, "Person recognition with HGR maximal correlation on multimodal data," in *Int. Conf. Patt. Recogn. (ICPR)*, Milan, Italy, Jan. 2021.

[182] G. Licciardi, P. R. Marpu, J. Chanussot, and J. A. Benediktsson, "Linear versus nonlinear PCA for the classification of hyperspectral data based on the extended morphological profiles," *IEEE Geosci., Remote Sensing Lett.*, vol. 9, no. 3, May 2012, pp. 447–451.

[183] G. J. Lidstone, "Note on the general case of the Bayes-Laplace formula for inductive or *a Posteriori* probabilities," *Trans. Fac. Actuaries*, vol. 8, 1920, pp. 182–192.

[184] A. Lubotzky, R. Phillips, and P. Sarnak, "Ramanujan graphs," *Combinatorica*, vol. 8, 1988, pp. 261–278.

[185] F. Ma, S.-L. Huang, and L. Zhang, "An efficient approach for audio-visual emotion recognition with missing labels and missing modalities," in *Proc. Int. Conf. Multimedia, Expo (ICME)*, Shenzhen, China, Jul. 2021.

[186] F. Ma, Y. Li, S. Ni, S.-L. Huang, and L. Zhang, "Data augmentation for audio-visual emotion recognition with an efficient multimodal conditional GAN," *Appl. Sci.*, vol. 12, no. 1, Jan. 2022, p. 527.

[187] F. Ma, W. Zhang, Y. Li, S.-L. Huanbg, and L. Zhang, "Learning better representations for audio-visual emotion recognition with common information," *Appl. Sci.*, vol. 10, no. 20, Oct. 2020, p. 7239.

[188] F. Ma, W. Zhang, Y. Li, S.-L. Huang, and L. Zhang, "An end-to-end learning approach for multimodal emotion recognition: Extracting common and private information," in *Proc. Int. Conf. Multimedia, Expo (ICME)*, Shanghai, China, Jul. 2019.

[189] A. Makhdoumi, F. P. Calmon, and M. Médard, "Forgot your password: Correlation dilution," in *Proc. Int. Symp. Inform. Theory (ISIT)*, Hong Kong, China, Jun. 2015.

[190] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *Proc. Inform. Theory Workshop (ITW)*, pp. 501–505, Hobart, TAS, Australia, Nov. 2014.

[191] A. Makur, “Information contraction and decomposition,” Ph.D. dissertation, Massachusetts Insitute of Technology, Cambridge, MA, 2019.

[192] A. Makur, F. Kozynski, S.-L. Huang, and L. Zheng, “An efficient algorithm for information decomposition and extraction,” in *Proc. Allerton Conf. Commun., Contr., Computing*, Monticello, IL, Sep. 2015.

[193] A. Makur, G. W. Wornell, and L. Zheng, “On estimation of modal decompositions,” in *Proc. Int. Symp. Inform. Theory (ISIT)*, Los Angeles, CA, Jun. 2020.

[194] A. Makur and L. Zheng, “Polynomial spectral decomposition of conditional expectation operators,” in *Proc. Allerton Conf. Commun., Contr., Computing*, Monticello, IL, Sep. 2016.

[195] A. Makur and L. Zheng, “Polynomial singular value decompositions of a family of source-channel models,” *IEEE Trans. Inform. Theory*, vol. 63, no. 12, Dec. 2017, pp. 7716–7728.

[196] A. Makur and L. Zheng, “Comparison of contraction coefficients for f-divergences,” *Probl. Inf. Transm.*, vol. 56, Apr. 2020, pp. 103–156.

[197] J. Mary, C. Calauzènes, and N. E. Karoui, “Fairness-aware learning for continuous attributes and treatments,” in *Proc. Int. Conf. Machine Learning (ICML)*, ser. Proc. Mach. Learn. Res. (PMLR), vol. 97, pp. 4382–4391, Long Beach, CA, Jun. 2019.

[198] D. A. McAllester and R. E. Schapire, “On the convergence rate of Good-Turing estimators,” in *Proc. Conf. Comput. Learning Theory (COLT)*, pp. 1–6, Palo Alto, CA, Jul. 2000.

[199] L. R. Mead and N. Papanicolaou, “Maximum entropy in the problem of moments,” *J. Math. Phys.*, vol. 25, 1984, pp. 2404–2417.

[200] F. G. Mehler, “Ueber die entwicklung einer function von beliebig vielen variabeln nach Laplaceschen functionen höherer ordnung,” *J. Reine, Angewandte Math.*, vol. 66, 1866, pp. 161–176.

[201] P. Melville and V. Sindhwani, “Recommender systems,” in *Encyclopedia of Machine Learning and Data Mining*, C. Sammut and G. I. Webb, Eds., Boston, MA: Springer, 2017, pp. 1056–1066.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[202] T. Melzer, M. Reiter, and H. Bischof, “Nonlinear feature extraction using generalized canonical correlation analysis,” in *Proc. Int. Conf. Artif. Neural Netw. (ICANN)*, ser. Lecture Notes in Computer Science (LNCS), vol. 2130, Vienna, Austria, Aug. 2001.

[203] T. Michaeli, W. Wang, and K. Livescu, “Nonparametric canonical correlation analysis,” in *Proc. Int. Conf. Machine Learning (ICML)*, ser. Proc. Mach. Learn. Res. (PMLR), vol. 48, pp. 1967–1976, New York, NY, Jun. 2016.

[204] G. Michailidis and J. de Leeuw, “The Gifi system of descriptive multivariate analysis,” *Stat. Sci.*, vol. 13, no. 4, 1998, pp. 307–336.

[205] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *CoRR*, vol. abs/-1301.3781, 2013. arXiv: 1301.3781. URL: http://arxiv.org/abs/1301.3781.

[206] D. J. Miller, A. Rao, K. Rose, and A. Gersho, “An information-theoretic learning algorithm for neural network classification,” in *Advances Neural Inform. Process. Syst. (NIPS)*, pp. 591–597, Denver, CO, Dec. 1996.

[207] M. Minsky and S. Papert, *Perceptrons.* Cambridge, MA: MIT Press, 1969.

[208] S. Negahban and M. J. Wainwright, “Estimation of (near) low-rank matrices with noise and high-dimensional scaling,” *Ann. Stat.*, vol. 39, no. 2, Apr. 2011, pp. 1069–1097.

[209] H. Ney, U. Essen, and R. Kneser, “On structuring probabilistic dependences in stochastic language modeling,” *Comput., Speech, Lang.*, vol. 8, no. 1, Jan. 1994, pp. 1–38.

[210] J. Neyman, “‘Smooth’ test for goodness of fit,” *Scand. Actuar. J.*, vol. 20, no. 3–4, 1937, pp. 149–199.

[211] J. Neyman, “Contribution to the theory of the χ^2 test,” in *Proc. Berkeley Symp. Math. Stat. Prob.*, pp. 239–273, Berkeley, CA, 1949.

[212] S. Nishisato, *Multivariate Nonlinear Descriptive Analysis.* London, UK: Chapman & Hall/CRC, 2006.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[213] E. Oja, “A simplified neuron model as a principal component analyzer,” *J. Math. Biology*, vol. 15, no. 3, Nov. 1982, pp. 267–273.

[214] E. Oja, “Principal components, minor components, and linear neural networks,” *Neural Netw.*, vol. 5, no. 6, Nov. 1992, pp. 927–935.

[215] E. Oja, “The nonlinear PCA learning rule in independent component analysis,” *Neurocomputing*, vol. 17, no. 1, Sep. 1997, pp. 25–45.

[216] A. Orlitsky and A. T. Suresh, “Competitive distribution estimation: Why is Good-Turing good,” in *Advances Neural Inform. Process. Syst. (NeurIPS)*, Montréal, Canada, Dec. 2015.

[217] A. Painsky, “Generalized independent components analysis over finite alphabets,” Ph.D. dissertation, Tel Aviv University, Tel Aviv, Israel, Sep. 2016.

[218] A. Painsky, M. Feder, and N. Tishby, “Nonlinear canonical correlation analysis: A compressed representation approach,” *Entropy*, vol. 22, no. 2, Feb. 2020.

[219] A. Painsky, S. Rosset, and M. Feder, “Generalized independent component analysis over finite alphabets,” *IEEE Trans. Inform. Theory*, vol. 62, no. 2, Feb. 2016, pp. 1038–1053.

[220] L. Paninski, “Variational minimax estimation of discrete distributions under KL loss,” in *Advances Neural Inform. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2004.

[221] E. Parzen, “On estimation of a probability density function and mode,” *Ann. Math. Stat.*, vol. 33, no. 3, Sep. 1962, pp. 1065–1076.

[222] K. Pearson, “Contributions to the mathematical theory of evolution,” *Phil. Trans. Roy. Soc. London, A*, vol. 185, 1894, pp. 71–110.

[223] K. Pearson, “On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling,” *Philos. Mag., Series 5*, vol. 50, no. 302, 1900, pp. 157–175.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

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

[225] K. Pearson, *On the Theory of Contingency and Its Relation to Association and Normal Correlation*, ser. Drapers' Company Research Memoirs, Biometric Series, I. Mathematical Contributions to the Theory of Evolution. London, England: Dulau and Co., 1904.

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

[227] D. Qiu, A. Makur, and L. Zheng, "Probabilistic clustering using maximal matrix norm couplings," in *Proc. Allerton Conf. Commun., Contr., Computing*, Monticello, IL, Oct. 2018.

[228] M. Raginsky, "Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels," *IEEE Trans. Inform. Theory*, vol. 62, no. 6, Jun. 2016, pp. 3355–3389.

[229] C. R. Rao, *Linear Statistical Inference and its Applications.* New York, NY: Wiley, 1965.

[230] B. Recht, M. Fazel, and P. A. Parillo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, 2010, pp. 471–501.

[231] A. Rényi, "A new version of the probabilistic generalization of the large sieve," *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 10, no. 1–2, Mar. 1959, pp. 217–226.

[232] A. Rényi, "On measures of dependence," *Acta Math. Acad. Sci. Hung.*, vol. 10, no. 3–4, Sep. 1959, pp. 441–451.

[233] A. Rohde and A. B. Tsybakov, "Estimation of high-dimensional low-rank matrices," *Ann. Stat.*, vol. 39, no. 2, Apr. 2011, pp. 887–930.

[234] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Ann. Math. Stat.*, vol. 27, no. 3, Sep. 1956, pp. 832–837.

[235] W. Rudin, *Principles of Mathematical Analysis*, 3rd. New York, NY: McGraw-Hill, 1976.

[236] S.Watanabe, “Information theoretical analysis of multivariate correlation,” *IBM J. Res. Develop.*, vol. 4, no. 1, Jan. 1960, pp. 66–82.

[237] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, 2nd ed. Philadelphia, PA: SIAM, 2011.

[238] O. V. Sarmanov, “The maximum correlation coefficient (nonsymmetric case),” *Dockl. Akad. Nauk SSSR*, vol. 121, no. 4, 1958, pp. 52–55.

[239] O. V. Sarmanov, “The maximum correlation coefficient (symmetric case),” *Dockl. Akad. Nauk SSSR*, vol. 120, no. 4, 1958, pp. 715–718.

[240] O. V. Sarmanov and V. K. Zaharov, “Maximum coefficients of multiple correlation,” *Dokl. Akad. Nauk SSSR*, vol. 121, 1960, pp. 269–271.

[241] S. Satpathy and P. Cuff, “Gaussian secure source coding and Wyner’s common information,” in *Proc. Int. Symp. Inform. Theory (ISIT)*, Hong Kong, China, Jun. 2015.

[242] P. Sattigeri, S. Ghosh, and S. C. Hoffman, “Chi-square information for invariant learning,” in *Proc. ICML Workshop Uncert., Robust. Deep Learn. (ICML-UDL*, (virtual), Jul. 2020.

[243] L. K. Saul, K. Q. Weinberger, F. Sha, J. Ham, and D. D. Lee, “Spectral methods for dimensionality reduction,” in *Semi-Supervised Learning*, O. Chapelle, B. Schölkopf, and A. Zien, Eds., Cambridge, MA: MIT Press, 2006, ch. 16, pp. 293–308.

[244] E. Schmidt, “Zur theorie der linearen und nichtlinearen integralgleichungen. I. Teil: Entwicklung willkürlicher funktionen nach systemen vorgeschriebener,” *Math. Ann.*, vol. 63, 1907, pp. 433–476.

[245] B. Schölkopf, A. Smola, and K.-R. Müller, “Kernel principal component analysis,” in *Proc. Int. Conf. Artif. Neural Netw. (ICANN)*, ser. Lecture Notes in Computer Science (LNCS), vol. 1327, pp. 583–588, Lausanne, Switzerland, Jun. 1997.

[246] B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Comput.*, vol. 10, no. 5, 1998, pp. 1299–1319.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[247] H. Scudder, “Probability of error of some adaptive pattern-recognition machines,” *IEEE Trans. Inform. Theory*, vol. 11, no. 3, Jul. 1965, pp. 363–371.

[248] A. Shah, Y. Bu, J. K.-W. Lee, S. Das, R. Panda, P. Sattigeri, and G. W. Wornell, “Selective regression under fairness criteria,” in *Proc. Int. Conf. Machine Learning (ICML)*, ser. Proc. Mach. Learn. Res. (PMLR), vol. 162, pp. 19 598–19 615, Baltimore, MD, Jul. 2022.

[249] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, Aug. 2000, pp. 888–905.

[250] R. Shwartz-Ziv and N. Tishby, “Opening the black box of deep neural networks via information,” *CoRR*, vol. abs/1703.00810, 2017. arXiv: 1703.00810. URL: http://arxiv.org/abs/1703.00810.

[251] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, UK: Chapman & Hall/CRC, 1986.

[252] D. Slepian, “On the symmetrized Kronecker power of a matrix and extensions of Mehler’s formula for Hermite polynomials,” *SIAM J. Math. Anal.*, vol. 3, no. 4, 1972, pp. 606–616.

[253] N. Slonim and N. Tishby, “Document clustering using word clusters via the information bottleneck method,” in *Proc. Int. Conf. Res., Dev. Inform. Retrieval (ACM SIGIR)*, ACM, pp. 208–215, Athens, Greece, Jul. 2000.

[254] C. Spearman, “‘General intelligence,’ objectively determined and measured,” *Amer. J. Psychol.*, vol. 15, no. 2, Apr. 1904, pp. 201–292.

[255] N. Srebro, “Learning with matrix factorizations,” Ph.D. dissertation, MIT, Cambridge, MA, Aug. 2004.

[256] G. W. Stewart, “On the early history of the singular value decomposition,” *SIAM Rev.*, vol. 35, no. 4, Dec. 1993, pp. 551–566.

[257] G. W. Stewart, “Perturbation theory for the singular value decomposition,” in *SVD and Signal Processing, II: Algorithms, Analysis and Applications*, R. J. Vaccaro, Ed., Elsevier, pp. 99–109, 1991.

[258] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” *Proc. Allerton Conf. Commun., Contr., Computing*, Oct. 1999, pp. 368–377.

[259] N. Tishby and N. Slonim, “Data clustering by Markovian relaxation and the information bottleneck method,” in *Advances Neural Inform. Process. Syst. (NIPS)*, pp. 619–625, Denver, CO, Dec. 2000.

[260] N. Tishby and N. Zaslavsky, “Deep learning and the information bottleneck principle,” in *Proc. Inform. Theory Workshop (ITW)*, Jerusalem, Israel, Apr. 2015.

[261] X. Tong, J. Xu, and S.-L. Huang, “An information-theoretic method for collaborative distributed learning with limited communication,” *CoRR*, vol. abs/2205.06515, 2022. arXiv: 2205.06515. URL: https://arxiv.org/abs/2205.06515.

[262] X. Tong, X. Xu, and S.-L. Huang, “On sample complexity of learning shared representations: The asymptotic regime,” in *Proc. Allerton Conf. Commun., Contr., Computing*, Monticello, IL, Sep. 2022.

[263] X. Tong, X. Xu, S.-L. Huang, and L. Zheng, “A mathematical framework for quantifying transferability in multi-source transfer learning,” in *Advances Neural Inform. Process. Syst. (NeurIPS)*, (virtual), Dec. 2021.

[264] L. N. Trefethen and I. D. Bau, *Numerical Linear Algebra.* Philadelphia, PA: SIAM, 1997.

[265] J. A. Tropp, “User-friendly tail bounds for sums of random matrices,” *Found. Comp. Math.*, vol. 12, no. 4, Aug. 2012, pp. 389–434.

[266] V. Uurtio, S. Bhadra, and J. Rousu, “Large-scale sparse kernel canonical correlation analysis,” in *Proc. Conf. Comput. Learning Theory (COLT)*, pp. 6383–6391, Long Beach, CA, Jun. 2019.

[267] N. Vaswani, T. Bouwmans, S. Javed, and P. Narayanamurthy, “Robust subspace learning: Robust PCA, robust subspace tracking, and robust subspace recovery,” *IEEE Signal Processing Mag.*, vol. 35, no. 4, Jul. 2018, pp. 32–55.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[268] J. M. Vegas and P. J. Zufiria, “Generalized neural networks for spectral analysis: Dynamics and Liapunov functions,” *Neural Netw.*, vol. 17, no. 2, Mar. 2004, pp. 233–245.

[269] C. Wang, “Variational Bayesian approach to canonical correlation analysis,” *IEEE Trans. Neural Netw.*, vol. 18, no. 3, May 2007, pp. 905–910.

[270] D. Wang and M. Murphy, “Estimating optimal transformations for multiple regression using the ACE algorithm,” *J. Data Science*, vol. 2, 2004, pp. 329–346.

[271] H. Wang, L. Vo, F. P. Calmon, M. Médard, K. R. Duffy, and M. Varia, “Privacy with estimation guarantees,” *IEEE Trans. Inform. Theory*, vol. 65, no. 12, Dec. 2019, pp. 8025–8042.

[272] H. Wang, L. Vo, F. P. Calmon, M. Médard, K. R. Duffy, and M. Varia, “Privacy with estimation guarantees,” *IEEE Trans. Inform. Theory*, vol. 65, no. 12, Dec. 2019, pp. 8025–8042.

[273] L. Wang, J. Wu, S.-L. Huang, L. Zheng, X. Xu, L. Zhang, and J. Huang, “An efficient approach to informative feature extraction from multimodal data,” in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 33, pp. 5281–5288, Honolulu, HI, Jan. 2019.

[274] W. Wang, R. Arora, K. Livescu, and J. Bilmes, “On deep multi-view representation learning,” in *Proc. Int. Conf. Machine Learning (ICML)*, ser. Proc. Mach. Learn. Res. (PMLR), vol. 37, pp. 1083–1092, Lille, France, Jul. 2015.

[275] R. W. M. Wedderburn, “Generalized linear models specified in terms of constraints,” *J. Roy. Stat. Soc., Ser. B*, vol. 36, no. 3, 1974, pp. 449–454.

[276] P. Whittle, “On the smoothing of probability density functions,” *J. Roy. Stat. Soc., Ser. B*, vol. 20, no. 2, Jul. 1958, pp. 334–343.

[277] H. S. Witsenhausen, “On sequences of pairs of dependent random variables,” *SIAM J. Appl. Math.*, vol. 28, no. 1, Jan. 1975, pp. 100–113.

[278] J. Wolf and J. Ziv, “Transmission of noisy information to a noisy receiver with minimum distortion,” *IEEE Trans. Inform. Theory*, vol. 16, no. 4, Jul. 1970, pp. 406–411.

[279] A. D. Wyner, “The common information of two dependent random variables,” *IEEE Trans. Inform. Theory*, vol. 21, no. 2, Mar. 1975, pp. 163–179.

[280] X. Xu and S.-L. Huang, “On the asymptotic sample complexity of HGR maximal correlation functions in semi-supervised learning,” in *Proc. Allerton Conf. Commun., Contr., Computing*, Monticello, IL, Sep. 2019.

[281] X. Xu and S.-L. Huang, “Maximal correlation regression,” *IEEE Access*, vol. 8, 2020, pp. 26 591–26 601.

[282] X. Xu and S.-L. Huang, “On the optimal tradeoff between computational efficiency and generalizability of Oja’s algorithm,” *IEEE Access*, vol. 8, 2020, pp. 102 616–102 628.

[283] X. Xu and S.-L. Huang, “An information theoretic framework for distributed learning algorithms,” in *Proc. Int. Symp. Inform. Theory (ISIT)*, Melbourne, Australia, Jul. 2021.

[284] X. Xu, W. Wang, and S.-L. Huang, “On the sample complexity of estimating small singular modes,” in *Proc. Int. Symp. Inform. Theory (ISIT)*, Los Angeles, CA, Jun. 2020.

[285] X. Xu and L. Zheng, “Multivariate feature extraction,” in *Proc. Allerton Conf. Commun., Contr., Computing*, Monticello, IL, Sep. 2022.

[286] S. Yin, F. Ma, and S.-L. Huang, “A semi-supervised learning approach for visual question answering based on maximal correlation,” in *Proc. Int. Conf. Syst., Man, Cybern. (SMC)*, Melbourne, Australia, Oct. 2021.

[287] G. Young, “Maximum likelihood estimation and factor analysis,” *Psychometrika*, vol. 6, no. 1, Feb. 1940, pp. 49–53.

[288] J. Yu, K. Wang, L. Ye, and Z. Song, “Accelerated kernel canonical correlation analysis with fault relevance for nonlinear process fault isolation,” *Int. Eng. Chem. Res.*, vol. 58, no. 39, Oct. 2019, pp. 18 280–18 291.

[289] T. Zhang, R. Ramakrishnan, and M. Livny, “BIRCH: An efficient data clustering method for very large databases,” in *Proc. ACM Conf. Management Data (SIGMOD)*, pp. 103–114, Montréal, Canada, Jun. 1996.

The version of record is available at: http://dx.doi.org/10.1561/0100000107

[290] W. Zhang, W. Gu, F. Ma, S. Ni, L. Zhang, and S.-L. Huang, “Multimodal emotion recognition by extracting common and modality-specific information,” in *Proc. Conf. Embed. Netw. Sensor Syst. (SENSYS)*, pp. 396–397, Shenzhen, China, Nov. 2018.

The version of record is available at: http://dx.doi.org/10.1561/0100000107