Gaussian Universal Features, Canonical Correlations, and Common Information

Shao-Lun Huang
DSIT Research Center, TBSI, SZ, China 518055
Email: shaolun.huang@sz.tsinghua.edu.cn

Gregory W. Wornell, Lizhong Zheng
Dept. EECS, MIT, Cambridge, MA 02139 USA
Email: {gww, lizhong}@mit.edu

***Abstract*—We address the problem of optimal feature selection for a Gaussian vector pair in the weak dependence regime, when the inference task is not known in advance. In particular, we show that multiple formulations all yield the same solution, and correspond to the singular value decomposition (SVD) of the canonical correlation matrix. Our results reveal key connections between canonical correlation analysis (CCA), principal component analysis (PCA), the Gaussian information bottleneck, Wyner's common information, and the Ky Fan (nuclear) norms.**

I. Introduction

Typical applications of machine learning involve data whose dimension is high relative to the amount of training data that is available. As a consequence, it is necessary to perform dimensionality reduction before the regression or other inference task is carried out. This reduction corresponds to extracting a set of comparatively low-dimensional features from the data. When the inference task is fully specified, classical statistics establishes that the appropriate features take the form of a (minimal) sufficient statistic. However, in most contemporary settings, the task is not known in advance—or equivalently there are multiple tasks—and we require a set of *universal* features that are, in an appropriate sense, uniformly good.

With this motivation, the Gaussian universal feature selection problem can be expressed as: given a pair of high-dimensional jointly distributed Gaussian data vectors $X \in \mathbb{R}^{K_X}$ and, $Y \in \mathbb{R}^{K_Y}$, how should we choose low-dimensional features $\mathbf{f}(X)$ and $\mathbf{g}(Y)$ before knowing the desired inference task so to ensure that after the task is revealed, inference based on the features performs as well as possible?

Mathematically, we express this problem as one of making inference about latent variables $U, V \in \mathbb{R}^k$, for $1 \leq k \leq K \triangleq \min\{K_X, K_Y\}$, in the Gauss-Markov chain

$$U \leftrightarrow X \leftrightarrow Y \leftrightarrow V, \tag{1}$$

where the (Gaussian) distributions for these variables, i.e., P_U, $P_{X|U}$, P_V, and $P_{Y|V}$ are not known at the time of feature extraction. Our results can be viewed as an extension of the framework for discrete variables described in [1]. We note in advance that to simplify the exposition, we treat $P_{X,Y}$ as known, though in practice we must estimate the relevant aspects of this distribution from training samples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$.

Our contribution of this paper is summarized as follows. To deal with the inference for unknown attributes, in section III we define a rotation-invariant ensemble (RIE) that assigns a uniform prior for the unknown attributes, and formulate a universal feature selection problem that aims to select optimal features minimizing the averaged MSE over RIE. We show that the optimal features can be obtained from the SVD of a canonical dependence matrix (CDM). In addition, we demonstrate that in a weak dependence regime, this SVD also provides the optimal features and solutions for several problems, such as CCA, information bottleneck, and Wyner's common information, for jointly Gaussian variables. This reveals important connections between information theory and machine learning problems.

II. Gaussian Local Analysis Framework

In the sequel, we restrict our attention to zero-mean variables, for simplicity of exposition. In the model of interest, $X \in \mathbb{R}^{K_X}$ and $Y \in \mathbb{R}^{K_Y}$. Moreover,

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathtt{N}(\mathbf{0}, \mathbf{\Lambda}_Z), \quad \mathbf{\Lambda}_Z = \mathbb{E}\left[ZZ^{\mathrm{T}}\right] = \begin{bmatrix} \mathbf{\Lambda}_X & \mathbf{\Lambda}_{XY} \\ \mathbf{\Lambda}_{YX} & \mathbf{\Lambda}_Y \end{bmatrix},$$

so $X \sim \mathtt{N}(\mathbf{0}, \mathbf{\Lambda}_X)$, $Y \sim \mathtt{N}(\mathbf{0}, \mathbf{\Lambda}_Y)$, $\mathbf{\Lambda}_{YX} = \mathbb{E}\big[YX^{\mathrm{T}}\big]$, and $\mathbf{\Lambda}_{XY} = \mathbf{\Lambda}_{YX}^{\mathrm{T}}$. We assume without loss of generality that $\mathbf{\Lambda}_X$ and $\mathbf{\Lambda}_Y$ are (strictly) positive definite. The joint distribution takes the form

$$P_{X,Y}(x,y) = P_Z(z) = \frac{|\mathbf{\Lambda}_Z|^{-1/2}}{(2\pi)^{K_Z/2}} \exp\left\{-\frac{1}{2}\, z^{\mathrm{T}}\, \mathbf{\Lambda}_Z^{-1}\, z\right\} \tag{2}$$

where $K_Z = K_X + K_Y$, and with $|\cdot|$ denoting the determinant of its argument. It will be convenient to normalize X and Y according to

$$\tilde{X} = \mathbf{\Lambda}_X^{-1/2} X \quad \text{and} \quad \tilde{Y} = \mathbf{\Lambda}_Y^{-1/2} Y$$

so that

$$\tilde{Z} = \begin{bmatrix} \tilde{X} \\ \tilde{Y} \end{bmatrix}, \qquad \mathbf{\Lambda}_{\tilde{Z}} = \begin{bmatrix} \mathbf{I} & \mathbf{B}^{\mathrm{T}} \\ \mathbf{B} & \mathbf{I} \end{bmatrix}, \tag{3}$$

where

$$\mathbf{B} \triangleq \mathbf{\Lambda}_Y^{-1/2}\, \mathbf{\Lambda}_{YX}\, \mathbf{\Lambda}_X^{-1/2} = \mathbf{\Lambda}_Y^{-1/2}\, \mathbf{\Gamma}_{Y|X}\, \mathbf{\Lambda}_X^{1/2} \tag{4}$$

is called the canonical dependence matrix (CDM). CDM plays the key role in Gaussian local analysis as the divergence transfer matrix (DTM) does in the discrete case [1].

We note that the MMSE estimate of $\tilde{Y}$ based on $\tilde{X}$ is $\hat{\tilde{Y}}(\tilde{X}) = \mathbf{B}\,\tilde{X}$, and the associated error $\tilde{\nu} \triangleq \tilde{Y} - \hat{\tilde{Y}}(\tilde{X})$ has covariance $\mathbb{E}\big[\tilde{\nu}\tilde{\nu}^{\mathrm{T}}\big] = \mathbf{I} - \mathbf{B}\mathbf{B}^{\mathrm{T}}$, so the resulting MSE is $\tilde{\sigma}_{\mathrm{e}}^2 = \mathrm{tr}\big(\mathbf{I} - \mathbf{B}\mathbf{B}^{\mathrm{T}}\big) = K_Y - \|\mathbf{B}\|_{\mathrm{F}}^2$. with $\|\cdot\|_{\mathrm{F}}$ denoting the Frobenius norm.

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

$$\mathbf{B} = \mathbf{\Psi}^Y \, \mathbf{\Sigma} \, \big(\mathbf{\Psi}^X\big)^{\mathrm{T}} = \sum_{i=1}^{K} \sigma_i \, \boldsymbol{\psi}_i^Y \, \big(\boldsymbol{\psi}_i^X\big)^{\mathrm{T}}, \tag{5}$$

where $K = \min\{K_X, K_Y\}$ and where we order the singular values according to $\sigma_1 \geq \cdots \geq \sigma_K$. Note that since (3) is positive semidefinite, it follows that $\sigma_i \leq 1$ for $i = 1, \dots, K$.

As in the discrete case [1], it is useful to define a local analysis regime for such variables. In particular, we make use of the following notion of neighborhood.

Definition 1 (Gaussian ϵ-Neighborhood)**.** For a given $\epsilon > 0$, the ϵ-neighborhood of a K_0-dimensional Gaussian distribution $P_0 = \mathtt{N}(\mathbf{0}, \mathbf{\Lambda}_0)$ is the set of Gaussian distributions in a covariance-divergence ball of radius ϵ^2 about P_0, i.e.,

$$\mathcal{G}_\epsilon^{K_0}(P_0) \triangleq \Big\{P = \mathtt{N}(\mathbf{0}, \mathbf{\Lambda})\colon \big\|\mathbf{\Lambda}_0^{-1/2}\big(\mathbf{\Lambda} - \mathbf{\Lambda}_0\big)\mathbf{\Lambda}_0^{-1/2}\big\|_{\mathrm{F}}^2 \leq \epsilon^2 K_0\Big\}$$

Note that $P_{X,Y}$ lies in an ϵ-neighborhood of $P_X P_Y$ if and only if $P_{\tilde{X},\tilde{Y}}$ lies in an ϵ-neighborhood of $P_{\tilde{X}} P_{\tilde{Y}}$. Hence, $P_{X,Y} \in \mathcal{G}_\epsilon^{K_X+K_Y}(P_X P_Y)$ when $\|\mathbf{\Lambda}_{\tilde{Z}} - \mathbf{I}\|_{\mathrm{F}}^2 \leq \epsilon^2(K_X + K_Y)$. We conclude that the neighborhood constraint limits how much the mean-square error (MSE) in the estimate of $\tilde{Y}$ based on observing $\tilde{X}$ can be reduced relative to the MSE in the estimate of $\tilde{Y}$ based on no data. In the rest of this paper, we focus on the regime that ϵ is small. The K-L divergence and mutual information in this regime admits the following useful asymptotic expressions.

Lemma 1. *In the weak dependence regime,*

$$D(P_{Y|X}(\cdot|x) \| P_Y) = \frac{1}{2}\big\|\mathbf{B}\tilde{x}\big\|^2 + o(\epsilon^2),$$

and

$$I(X;Y) = \frac{1}{2}\sum_{i=1}^{K} \sigma_i^2 + o(\epsilon^2). \tag{6}$$

Proof. This is straightforward from the fact that, for an arbitrary matrix $\mathbf{A}$, $\ln|\mathbf{I} - \epsilon^2 \mathbf{A}\,\mathbf{A}^{\mathrm{T}}| = -\epsilon^2\|\mathbf{A}\|_{\mathrm{F}}^2 + o(\epsilon^2)$. □

To interpret (6), consider the modal decomposition of $P_{X,Y}$. In particular, observe that as $\epsilon \to 0$,

$$\mathbf{\Lambda}_{\tilde{Z}}^{-1} = \begin{bmatrix} \mathbf{I} & -\mathbf{B}^{\mathrm{T}} \\ -\mathbf{B} & \mathbf{I} \end{bmatrix} + o(\epsilon). \tag{7}$$

Hence,

$$P_{X,Y}(x,y) = P_X(x)\,P_Y(y)\,\left(\prod_{i=1}^{K} e^{\sigma_i\, f_i^*(x)\, g_i^*(y)}\right)\big(1 + o(\epsilon)\big), \tag{8a}$$

where f_i^* and g_i^* are (linear) functions given by

$$f_i^*(x) = \underbrace{(\boldsymbol{\psi}_i^X)^{\mathrm{T}}\mathbf{\Lambda}_X^{-1/2}}_{\triangleq \mathbf{f}_i^{*\mathrm{T}}}\, x \quad \text{and} \quad g_i^*(y) = \underbrace{(\boldsymbol{\psi}_i^Y)^{\mathrm{T}}\mathbf{\Lambda}_Y^{-1/2}}_{\triangleq \mathbf{g}_i^{*\mathrm{T}}}\, y. \tag{8b}$$

Moreover, using (8b) with (4) and (5) we obtain the covariance expansion

$$\mathbf{\Lambda}_Y^{-1}\,\mathbf{\Lambda}_{YX}\,\mathbf{\Lambda}_X^{-1} = \mathbf{\Lambda}_Y^{-1/2}\,\mathbf{B}\,\mathbf{\Lambda}_X^{-1/2} = \sum_{i=1}^{K}\sigma_i\,\mathbf{g}_i^*\,\mathbf{f}_i^{*\mathrm{T}}.$$

These linear functions can be computed from $P_{X,Y}$ (or estimated from training data) using a linearly-constrained version of the alternating conditional expectation (ACE) algorithm [2], which interprets the power iteration method for SVD computation.

From this perspective, we see that approximations to $P_{X,Y}$ can be obtained by truncating the representation (8) to the first $k < K$ of the terms in the product, yielding

$$\begin{aligned} &P_{X^{(k)},Y^{(k)}}(x,y) \\ &\qquad = P_X(x)\,P_Y(y)\,\left(\prod_{i=1}^{k} e^{\sigma_i\, f_i^*(x)\, g_i^*(y)}\right)\big(1 + o(\epsilon)\big), \end{aligned}$$

This corresponds to jointly Gaussian $X^{(k)}$ and $Y^{(k)}$ with the same marginals as X and Y, respectively, but

$$\mathbf{\Lambda}_{Y^{(k)}X^{(k)}} = \mathbf{\Lambda}_Y^{1/2}\underbrace{\left(\sum_{i=1}^{k}\sigma_i\,\boldsymbol{\psi}_i^Y\big(\boldsymbol{\psi}_i^X\big)^{\mathrm{T}}\right)}_{\triangleq \mathbf{B}^{(k)}}\mathbf{\Lambda}_X^{1/2}, \tag{9}$$

so

$$I(X^{(k)};Y^{(k)}) = \frac{1}{2}\sum_{i=1}^{k}\sigma_i^2 + o(\epsilon^2).$$

III. Universal Linear Feature Selection

We use our framework to address the problem of Gaussian universal feature selection. In our analysis, $\tilde{U}, \tilde{X}, \tilde{Y}, \tilde{V}$ denote normalized versions of the variables in (1), so are $\mathtt{N}(\mathbf{0}, \mathbf{I})$ random vectors of appropriate dimension. In the sequel, we consider several different formulations, all of which yield the same linear features, and coincide with those defined by the modal expansion (8). In our development, the following lemma will be useful (see, e.g., [3, Corollary 4.3.39, p. 248]).

Lemma 2. *Given an arbitrary $k_1 \times k_2$ matrix $\mathbf{A}$ and any $k \in \big\{1, \dots, \min\{k_1, k_2\}\big\}$, we have*

$$\max_{\left\{\mathbf{M}\in\mathbb{R}^{k_2\times k}\colon\, \mathbf{M}^{\mathrm{T}}\mathbf{M}=\mathbf{I}\right\}} \big\|\mathbf{A}\mathbf{M}\big\|_{\mathrm{F}}^2 = \sum_{i=1}^{k}\sigma_i(\mathbf{A})^2, \tag{10}$$

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 (10) is achieved by $\mathbf{M} = \big[\boldsymbol{\psi}_1(\mathbf{A}) \;\; \cdots \;\; \boldsymbol{\psi}_k(\mathbf{A})\big]$, with $\boldsymbol{\psi}_i(\mathbf{A})$ denoting the right singular vector of $\mathbf{A}$ corresponding to $\sigma_i(\mathbf{A})$, for $i = 1, \dots, \min\{k_1, k_2\}$.

A. Optimum Features, Rotation-Invariant Ensembles

In this formulation, we seek to determine optimum features for estimating an unknown k-dimensional U from Y, in the case where U and X are weakly dependent; specifically, $P_{X,U} \in \mathcal{G}_\epsilon^{K_X+k}(P_X P_U)$. Accordingly, from the innovations form $\tilde{X} = \epsilon\,\mathbf{\Phi}^{X|U}\,\tilde{U} + \nu_{\tilde{U}\to\tilde{X}}$, where $\tilde{U}$ and $\nu_{\tilde{U}\to\tilde{X}}$ are independent, it follows that weak dependence means, using Definition 1, that $\mathbf{\Phi}^{X|U}$ satisfies

$$\big\|\mathbf{\Phi}^{X|U}\big\|_{\mathrm{F}}^2 \leq \frac{1}{2}\,(K_X + k), \tag{11}$$

but is otherwise unknown.

We observe $\tilde{Y} = \mathbf{B}\,\tilde{X} + \nu_{\tilde{X}\to\tilde{Y}} = \epsilon\,\mathbf{B}\,\mathbf{\Phi}^{X|U}\,\tilde{U} + \nu_{\tilde{U}\to\tilde{Y}}$. From this data—and before knowing $\mathbf{\Phi}^{X|U}$—we construct a linear feature (statistic) of the form

$$T = \big(\mathbf{\Xi}^Y\big)^{\mathrm{T}}\,\tilde{Y}, \tag{12}$$

which we normalize for convenience (and without loss of generality) according to

$$\big(\mathbf{\Xi}^Y\big)^{\mathrm{T}}\mathbf{\Xi}^Y = \mathbf{I}, \tag{13}$$

so $T = \tilde{T}$ since $\mathbf{\Lambda}_T = \mathbf{I}$. We refer to $\mathbf{\Xi}^Y$ as the *feature weights* associated with the linear feature T. When $\mathbf{\Phi}^{X|U}$ is determined, we can generate the MMSE estimate of U based on T, and express the resulting MSE in the form

$$\mathrm{tr}(\mathbf{\Lambda}_{U|T}) = \mathrm{tr}\,\mathbf{\Lambda}_U - \epsilon^2 \big\| \big(\mathbf{\Xi}^Y\big)^{\mathrm{T}}\, \mathbf{B}\, \mathbf{\Phi}^{X|U}\, \mathbf{\Lambda}_U^{1/2} \big\|_{\mathrm{F}}^2. \tag{14}$$

Since $\mathbf{\Phi}^{X|U}$ is unknown, to determine the optimum choice of $\mathbf{\Xi}^Y$, we assume that (X, U) configuration is randomly drawn from a rotation-invariant ensemble (RIE), defined as follows.

Definition 2. RIE is a collection of configurations such that given any $\mathbf{\Phi}^{X|U}$, $\mathbf{\Phi}^{X|U} \overset{\mathrm{d}}{=} \mathbf{Q}\mathbf{\Phi}^{X|U}$ for any unitary matrix $\mathbf{Q}$, where $\overset{\mathrm{d}}{=}$ denotes the two configurations have equal probability, i.e., the prior distribution of $\mathbf{\Phi}^{X|U}$ is spherically symmetric.

To derive the optimal features, the following lemma from [4] is useful.

Lemma 3. *Let* $\mathbf{Z}$ *be a* $k_1 \times k_2$ *spherically symmetric [4] random matrix, i.e., 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* $\mathbf{Z} \overset{\mathrm{d}}{=} \mathbf{Q}_1^{\mathrm{T}}\, \mathbf{Z}\, \mathbf{Q}_2$*. Then if* $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are any fixed matrices of appropriate dimension, then*

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

Theorem 1. *The optimal features* T_* *to minimize the averaged MSE* (14) *over the RIE prior is given by* $T_* = \mathbf{g}^*(Y)$*, where* $\mathbf{g}^*(Y)$ *is from the Gaussian modal decomposition from* (8).

Proof. Taking the expectation of (14) over the RIE and applying Lemma 3, with some computations, we obtain that the averaged MSE is

$$\begin{aligned} &\mathbb{E}\left[\mathrm{tr}(\mathbf{\Lambda}_{U|T})\right] \\ &\quad = \mathrm{tr}(\mathbf{\Lambda}_U)\left(1 - \frac{\epsilon^2}{2}\left(\frac{1}{k} + \frac{1}{K_X}\right) \big\|\big(\mathbf{\Xi}^Y\big)^{\mathrm{T}}\mathbf{B}\big\|_{\mathrm{F}}^2\right). \end{aligned} \tag{15}$$

Then via Lemma 2 it follows that (15) is minimized subject to (13) by choosing the columns of $\mathbf{\Xi}^Y$ as the left singular vectors of $\mathbf{B}$ corresponding to the k largest singular values, i.e., $\mathbf{\Xi}^Y = \mathbf{\Psi}_{(k)}^Y \triangleq \begin{bmatrix}\boldsymbol{\psi}_1^Y & \cdots & \boldsymbol{\psi}_k^Y\end{bmatrix}$. Hence, the optimum feature vector for inferences about the resulting variable U from Y is, via (12),

$$T_* = \big(\mathbf{\Psi}_{(k)}^Y\big)^{\mathrm{T}} \mathbf{\Lambda}_Y^{-1/2}\, Y = \underbrace{\begin{bmatrix} g_1^*(Y) & \cdots & g_k^*(Y)\end{bmatrix}^{\mathrm{T}}}_{\triangleq \mathbf{g}^*(Y)}, \tag{16}$$

which we note coincides with the features of Y that arise in the Gaussian modal decomposition (8).

□

By symmetry, an analogous derivation yields the corresponding results for constructing feature weights $\mathbf{\Xi}^X$ and linear feature (statistic)

$$S = \big(\mathbf{\Xi}^X\big)^{\mathrm{T}} \tilde{X} \tag{17}$$

with [cf. (13)]

$$\big(\mathbf{\Xi}^X\big)^{\mathrm{T}} \mathbf{\Xi}^X = \mathbf{I} \tag{18}$$

for estimating an unknown V from X. Similarly, we obtain that the optimal $\mathbf{\Xi}^X$ has as its columns the right singular vectors $\mathbf{\Psi}_{(k)}^X \triangleq \begin{bmatrix}\boldsymbol{\psi}_1^X & \cdots & \boldsymbol{\psi}_k^X\end{bmatrix}$ of $\mathbf{B}$ corresponding to the k largest singular values, and the optimum feature vector

$$S_* = \big(\mathbf{\Psi}_{(k)}^X\big)^{\mathrm{T}} \mathbf{\Lambda}_X^{-1/2}\, X = \underbrace{\begin{bmatrix} f_1^*(X) & \cdots & f_k^*(X)\end{bmatrix}^{\mathrm{T}}}_{\triangleq \mathbf{f}^*(X)}, \tag{19}$$

is from the Gaussian modal decomposition (8).

B. Cooperative Optimization and CCA

In this subsection, we consider a variant optimization problem, where the system designer and nature share the goal of constructing and estimating U from Y in the chain (1) so as to minimize the relative MSE subject to the weak dependence constraint $P_{X,U} \in \mathcal{G}_\epsilon^{K_X+k}(P_X P_U)$. In this game, nature chooses the variable U via $\mathbf{\Phi}^{X|U}$ and attribute covariance $\mathbf{\Lambda}_U$ and the system designer chooses the feature weights $\mathbf{\Xi}^Y$ for the statistic (12). Nature is constrained in that $\mathbf{\Lambda}_U$ cannot have eigenvalues larger than 1, i.e., its spectral norm $\|\cdot\|_{\mathrm{s}}$, denoted the largest eigenvalue, satisfies $\big\|\mathbf{\Lambda}_U\big\|_{\mathrm{s}} \le 1$, and $\mathbf{\Phi}^{X|U}$ is constrained according to [cf. (11)]

$$\big\|\mathbf{\Phi}^{X|U}\big\|_{\mathrm{F}}^2 \le \frac{1}{2}\,\frac{K_X + k}{k}. \tag{20}$$

The system designer is constrained in that the columns of $\mathbf{\Xi}^Y$ must be orthonormal, i.e., (13) must be satisfied.

Theorem 2. *To minimize the MSE* (14)*, in the cooperative game the nature should choose* $\mathbf{\Lambda}_U = \mathbf{I}$ *and* $\mathbf{\Phi}^{X|U} = \mathbf{\Psi}_{(k)}^X$*, and the system designer should choose* $\mathbf{\Xi}^Y = \mathbf{\Psi}_{(k)}^Y$.

Proof. Although we omit the steps due to space constraints, it is straightforward to show, again using Lemma 2, that

$$\mathrm{tr}(\mathbf{\Lambda}_U) - \mathrm{tr}(\mathbf{\Lambda}_{U|T}) \le \frac{\epsilon^2}{2}\,\frac{K_X + k}{k} \sum_{i=1}^{k} \sigma_i^2, \tag{21}$$

and the equality holds when nature chooses the variable U according to $\mathbf{\Lambda}_U = \mathbf{I}$ and $\mathbf{\Phi}^{X|U} = \mathbf{\Psi}_{(k)}^X$, and the system designer chooses the feature weights according to $\mathbf{\Xi}^Y = \mathbf{\Psi}_{(k)}^Y$.

□

Note that the analysis of this section is not asymptotic; it holds for all ϵ sufficiently small. We further emphasize that this analysis establishes that (16) is a sufficient statistic for inferences about the resulting variable U from Y.

An identical analysis yields the solution to the cooperative game for estimating V from X. In particular, nature chooses the variable V according to $\mathbf{\Lambda}_V = \mathbf{I}$ and $\mathbf{\Phi}^{Y|V} = \mathbf{\Psi}_{(k)}^Y$, and the system designer chooses the feature weights according to $\mathbf{\Xi}^X = \mathbf{\Psi}_{(k)}^X$. Moreover, analogously, (19) is a sufficient statistic for inferences about the resulting variable V from X.

CCA Interpretation: The cooperative game can be viewed as selecting the most detectable attribute of X and detecting this attribute by the most correlated feature of Y. These optimizations can be equivalently and directly interpreted via CCA [5]. To demonstrate the connection, the following (von Neumann) lemma (see, e.g., [6]), is useful.

Lemma 4. *Given an arbitrary $k_1 \times k_2$ matrix $\mathbf{A}$ and any $k \in \{1, \dots, \min\{k_1, k_2\}\}$, we have*

$$\max_{\substack{\{\mathbf{M}_1\in\mathbb{R}^{k_1\times k},\ \mathbf{M}_2\in\mathbb{R}^{k_2\times k}:\\ \mathbf{M}_1^{\mathrm{T}}\mathbf{M}_1=\mathbf{M}_2^{\mathrm{T}}\mathbf{M}_2=\mathbf{I}\}}} \operatorname{tr}\big(\mathbf{M}_1^{\mathrm{T}}\mathbf{A}\mathbf{M}_2\big)=\sum_{i=1}^{k} \sigma_i(\mathbf{A}), \tag{22}$$

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 (22) is achieved by $\mathbf{M}_j = \big[\boldsymbol{\psi}_1^{(j)}(\mathbf{A}) \ \cdots \ \boldsymbol{\psi}_k^{(j)}(\mathbf{A})\big]$, $j = 1, 2$, 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 $i = 1, \dots, \min\{k_1, k_2\}$.

To develop this perspective, first note that for zero-mean random vectors $X \in \mathbb{R}^{K_X}$ an $Y \in \mathbb{R}^{K_Y}$ with given covariance structure $\mathbf{\Lambda}_{X,Y}$, CCA seeks to find k-dimensional linear feature vectors $\mathbf{f}(X) = \big(\mathbf{\Xi}^X\big)^{\mathrm{T}}\tilde{X}$ and $\mathbf{g}(Y) = \big(\mathbf{\Xi}^Y\big)^{\mathrm{T}}\tilde{Y}$, for $k \leq K$, normalized according to (18) and (13) so that

$$\mathbb{E}\left[\mathbf{f}(X)\right] = \mathbb{E}\left[\mathbf{g}(X)\right] = \mathbf{0} \tag{23a}$$

$$\mathbb{E}\left[\mathbf{f}(X)\,\mathbf{f}(X)^{\mathrm{T}}\right] = \mathbb{E}\left[\mathbf{g}(Y)\,\mathbf{g}(Y)^{\mathrm{T}}\right] = \mathbf{I}, \tag{23b}$$

so as to maximize the vector correlation (generalized Pearson correlation coefficient)

$$\sigma(\mathbf{f},\mathbf{g}) \triangleq \mathbb{E}\left[\mathbf{f}(X)^{\mathrm{T}}\mathbf{g}(Y)\right] = \big(\mathbf{\Xi}^Y\big)^{\mathrm{T}}\mathbf{B}\,\mathbf{\Xi}^X, \tag{24}$$

which via Lemma 4 immediately yields that the optimizing $\mathbf{f}$ an $\mathbf{g}$ are as given by (19) and (16), and the maximal correlation is $\sigma(\mathbf{f}^*, \mathbf{g}^*) = \sum_{i=1}^{k} \sigma_i$, which corresponds to the Ky Fan k-norm of $\mathbf{B}$. Note that the special case $k = 1$ corresponds to standard CCA [5].

C. The Local Gaussian Information Bottleneck

For the Gauss-Markov chain (1), we want to determine the Gaussian vector $U = (U_1, \dots, U_k)$ that maximizes $I(Y;U)$ subject to the constraints: 1) U_i are i.i.d. and unit-variance; 2) $I(X;U_i) \leq \epsilon^2/2$ for all i; 3) U_i and U_j are conditionally independent given X. This can be viewed as a variation of the Gaussian information bottleneck problem [7] in the weak dependence regime. In this subsection, we illustrate the connection between our optimal features and this information bottleneck problem.

Theorem 3. *Denote the innovations form of U as $\tilde{X} = \epsilon\,\mathbf{\Phi}^{X|U}\,\tilde{U} + \nu_{\tilde{U}\to\tilde{X}}$, and U_i as $U_i = \epsilon\big(\boldsymbol{\phi}^{X|U_i}\big)^{\mathrm{T}}\tilde{X} + \nu_{\tilde{X}\to U_i}$, then the $\mathbf{\Phi}^{X|U}$ maximizing $I(Y;U)$ is given by $\mathbf{\Psi}^X_{(k)}$.*

Proof. Note that the constraint 1 expresses $\mathbf{\Lambda}_U = \mathbf{I}$, and via the innovations form $U_i = \epsilon\big(\boldsymbol{\phi}^{X|U_i}\big)^{\mathrm{T}}\tilde{X} + \nu_{\tilde{X}\to U_i}$, Constraint 2 imposes that $\big\|\boldsymbol{\phi}^{X|U_i}\big\| \leq 1 + o(1)$ for all i. Finally, Constraint 3 imposes that the $\nu_{\tilde{X}\to U_i}$ are independent for different i, thus $\mathbf{I} - \epsilon^2\big(\mathbf{\Phi}^{X|U}\big)^{\mathrm{T}}\mathbf{\Phi}^{X|U}$ is diagonal, where $\mathbf{\Phi}^{X|U} = \big[\boldsymbol{\phi}^{X|U_1} \ \cdots \ \boldsymbol{\phi}^{X|U_k}\big]$, which, in turn, means $\big(\boldsymbol{\phi}^{X|U_i}\big)^{\mathrm{T}}\boldsymbol{\phi}^{X|U_j} = 0$ for $i \neq j$.

Hence, it follows from —omitting the computations due to space limitations—using Lemma 1 and Lemma 2 that

$$I(Y;U) \leq \frac{\epsilon^2}{2}\sum_{i=1}^{k}\sigma_i^2 + o(\epsilon^2), \tag{25}$$

with equality when $\mathbf{\Phi}^{X|U}$ defining U is given by $\mathbf{\Psi}^X_{(k)}$. □

An analogous development determines the Gaussian vector V that maximizes $I(X;V)$ subject to the constraints: 1) V_i are i.i.d. and unit-variance; 2) $I(Y;V_i) \leq \epsilon^2/2$ for all i; 3) V_i and V_j are conditionally independent given Y. In particular, via the innovations form $\tilde{V}_i = \epsilon\big(\boldsymbol{\phi}^{Y|V_i}\big)^{\mathrm{T}}\tilde{Y} + \nu_{\tilde{Y}\to\tilde{V}_i}$ with $\mathbf{\Phi}^{Y|V} = \big[\boldsymbol{\phi}^{Y|V_1} \ \cdots \ \boldsymbol{\phi}^{Y|V_k}\big]$, we obtain $I(X;V) \leq \frac{\epsilon^2}{2}\sum_{i=1}^{k}\sigma_i^2 + o(\epsilon^2)$, with equality when $\mathbf{\Phi}^{Y|V}$ defining V is given by $\mathbf{\Psi}^Y_{(k)}$.

D. Most Strongly Revealed Joint Dependency

In this formulation, for the Gauss-Markov chain (1) we determine Gaussian $U, V \in \mathbb{R}^k$ that maximize $I(U;V)$ subject to the constraints: 1) the elements of U and V are each i.i.d. with unit-variance; 2) $I(X;U_i) \leq \epsilon^2/2$ and $I(Y;V_i) \leq \epsilon^2/2$ for all i; 3) U_i and U_j are conditionally independent given X, and V_i and V_j are conditionally independent given Y.

Theorem 4. *The optimal $\mathbf{\Phi}^{X|U}$ and $\mathbf{\Phi}^{Y|V}$ in the innovations form of U and V that maximizes $I(U;V)$ are given by $\mathbf{\Psi}^X_{(k)}$ and $\mathbf{\Psi}^Y_{(k)}$, respectively.*

Proof. Using the fact that

$$\mathbf{\Lambda}_{UV} = \epsilon^2\left(\mathbf{\Phi}^{X|U}\right)^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{\Phi}^{Y|V}. \tag{26}$$

we obtain, using Lemma 1 and Lemma 2 and the analysis of Section III-C (again, omitted the computations),

$$I(U;V) \leq \frac{\epsilon^4}{2}\sum_{i=1}^{k}\sigma_i^2 + o(\epsilon^4), \tag{27}$$

which is achieved with equality when $\mathbf{\Phi}^{Y|V}$ and $\mathbf{\Phi}^{X|U}$ are given by the singular vectors $\mathbf{\Psi}^Y_{(k)}$ and $\mathbf{\Psi}^X_{(k)}$, respectively. □

The resulting optimizing (jointly Gaussian) U, V then has covariance $\mathbf{\Lambda}_{UV} = \epsilon^2\,\mathbf{\Sigma}_{(k)}$, where $\mathbf{\Sigma}_{(k)}$ is diagonal with diagonal elements $\sigma_1, \dots, \sigma_k$, so $\mathbb{E}\left[U_i\,V_j\right] = \epsilon^2\,\sigma_i\,\mathbb{1}_{i=j}$ and

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

Note, finally, that (S,T) given by (19) and (16) are sufficient statistics for inferences about the resulting (U,V), which we emphasize are obtained by *separate* processing of X and Y.

IV. Gaussian Common Information

We interpret the dominant structure in terms of the common information associated with the jointly Gaussian pair (X,Y) as defined by Wyner [8]. In our analysis, the following (variational) characterization of the Ky Fan (nuclear) norm (see, e.g., [9], [10]) is useful.

Lemma 5. *Given an arbitrary $k_1 \times k_2$ matrix $\mathbf{A}$ and any positive integer k, we have*

$$\min_{\substack{\{\mathbf{M}_1\in\mathbb{R}^{k_1\times k},\ \mathbf{M}_2\in\mathbb{R}^{k\times k_2}:\\ \mathbf{M}_1\mathbf{M}_2=\mathbf{A}\}}} \left(\frac{1}{2}\|\mathbf{M}_1\|_{\mathrm{F}}^2 + \frac{1}{2}\|\mathbf{M}_2\|_{\mathrm{F}}^2\right) = \|\mathbf{A}\|_* \tag{28}$$

where $\|\cdot\|_*$ denotes the Ky Fan norm, i.e.,

$$\|\mathbf{A}\|_* \triangleq \operatorname{tr}\left(\sqrt{\mathbf{A}^{\mathrm{T}}\mathbf{A}}\right) = \sum_{i=1}^{k_* \triangleq \min\{k_1,k_2\}} \sigma_i(\mathbf{A}), \tag{29}$$

with $\sigma_1(\mathbf{A}), \ldots, \sigma_{k_*}(\mathbf{A})$ denoting the singular values of $\mathbf{A}$.

To begin, we express the common information $C(X, Y)$ as

$$C(X,Y) = \min I(\tilde{W}; \tilde{X}, \tilde{Y}), \tag{30}$$

where the minimum is over all (zero-mean) random vectors $\tilde{W} \in \mathbb{R}^k$ for some k such that $(\tilde{W}, \tilde{X}, \tilde{Y})$ are jointly Gaussian and have the Markov structure $\tilde{X} \leftrightarrow \tilde{W} \leftrightarrow \tilde{Y}$, and where we have normalized $\tilde{W}$ such that $\mathbf{\Lambda}_{\tilde{W}} = \mathbf{I}$. This corresponds to the Wyner's common information for Gaussian vectors [13]. Then, we express this Gaussian structure via innovations form

$$\tilde{Z} = \begin{bmatrix} \mathbf{A}_X \\ \mathbf{A}_Y \end{bmatrix} \tilde{W} + \nu_{W \to Z}, \tag{31}$$

where $\mathbf{A}_X$ and $\mathbf{A}_Y$ are $K_X \times k$ and $K_Y \times k$ matrices, respectively.

We focus on the case where $P_{X|W}(\cdot|w) \in \mathcal{G}_\epsilon^{K_X}(P_X)$ and $P_{Y|W}(\cdot|w) \in \mathcal{G}_\epsilon^{K_Y}(P_Y)$, so $\|\mathbf{A}_X\|_\mathrm{F} \leq \epsilon$ and $\|\mathbf{A}_Y\|_\mathrm{F} \leq \epsilon$. In addition, note that $\mathbf{A}_X$ and $\mathbf{A}_Y$ are related according to

$$\mathbb{E}\left[\tilde{Y}\tilde{X}^\mathrm{T}\right] = \mathbf{A}_Y \mathbf{A}_X^\mathrm{T} = \mathbf{B}. \tag{32}$$

Theorem 5. *The optimal $\mathbf{A}_X$ and $\mathbf{A}_Y$ that maximizes* (30) *are given by $\mathbf{\Psi}^X_{(K)} \mathbf{\Sigma}^{1/2}_{(K)}$ and $\mathbf{\Psi}^Y_{(K)} \mathbf{\Sigma}^{1/2}_{(K)}$, where $\mathbf{\Sigma}_{(k)}$ is diagonal with diagonal elements $\sigma_1, \ldots, \sigma_k$.*

Proof. Note that via Lemma 1—but omitting the computations due to space constraints—we obtain

$$I(\tilde{W}; \tilde{X}, \tilde{Y}) = \frac{1}{2}\,\mathrm{tr}\big(\mathbf{A}_X^\mathrm{T}\mathbf{A}_X\big) + \frac{1}{2}\,\mathrm{tr}\big(\mathbf{A}_Y^\mathrm{T}\mathbf{A}_Y\big) + o(\epsilon^2) \tag{33}$$

Minimizing (33) subject to the constraint (32) yields, via Lemma 5,

$$C(X,Y) = \min_{\substack{\{\mathbf{A}_X, \mathbf{A}_Y : \\ \mathbf{A}_Y\mathbf{A}_X^\mathrm{T} = \mathbf{B}\}}} I(\tilde{W}; \tilde{X}, \tilde{Y}) = \sum_{i=1}^{K} \sigma_i + o(\epsilon^2),$$

for which an optimizing W is defined via

$$\begin{bmatrix} \mathbf{A}_X \\ \mathbf{A}_Y \end{bmatrix} = \begin{bmatrix} \mathbf{\Psi}^X_{(K)} \\ \mathbf{\Psi}^Y_{(K)} \end{bmatrix} \mathbf{\Sigma}^{1/2}_{(K)}. \tag{34}$$

□

In addition, we note that $R \triangleq S + T$—with S and T from (19) and (16)—is a sufficient statistic for inferences about W.

We can interpret the common information variable W_i as capturing the common information between U_i and V_i. Indeed, it can be verified that $C(U_i, V_i) = \sigma_i + o(\epsilon^2) = I(W_i; X, Y)$.

V. Connection to PCA

PCA [11], [12] can be interpreted as a special case of the preceding results. Specifically, in some instances, the dimensionality reduction realized by PCA corresponds to the optimum statistics defined in (19) and (16), respectively, for the universal estimation of the unknown attributes U and Vunder any of our formulations.

Example 1. *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 $\mathbf{\Upsilon}$ are orthonormal, and $\mathbf{\Lambda}$ is diagonal with entries $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K$. Then it is immediate that $\mathbf{\Lambda}_Y = \mathbf{\Upsilon}\left(\mathbf{\Lambda} + \sigma_\nu^2\,\mathbf{I}\right)\mathbf{\Upsilon}^\mathrm{T}$, so*

$$\mathbf{B} = \mathbf{\Lambda}_Y^{-1/2}\mathbf{\Lambda}_X^{1/2} = \mathbf{\Upsilon}\left(\mathbf{I} + \sigma_\nu^2\,\mathbf{\Lambda}^{-1}\right)^{-1/2}\mathbf{\Upsilon}^\mathrm{T}. \tag{35}$$

As a result, we have that (19) *specializes to*

$$\mathbf{f}^*(X) = \mathbf{\Lambda}_{(k)}^{-1/2}\,\mathbf{\Upsilon}_{(k)}^\mathrm{T}\,X, \tag{36a}$$

where $\mathbf{\Upsilon}_{(k)}$ denotes the $K \times k$ matrix consisting of the first k columns of $\mathbf{\Upsilon}$, and where $\mathbf{\Lambda}_{(k)}$ denotes the $k \times k$ upper left submatrix of $\mathbf{\Lambda}$. In turn, it follows that the k-dimensional PCA vector

$$S^\mathrm{PCA} = \mathbf{f}^\mathrm{PCA}(X) \triangleq \mathbf{\Upsilon}_{(k)}^\mathrm{T}\,X \tag{36b}$$

is a sufficient statistic for inferences about the unknown V. Via a similar analysis

$$T^\mathrm{PCA} = \mathbf{g}^\mathrm{PCA}(Y) \triangleq \mathbf{\Upsilon}_{(k)}^\mathrm{T}\,Y \tag{36c}$$

is a sufficient statistic for inferences about the unknown U in this case. More generally $(S^\mathrm{PCA}, T^\mathrm{PCA}) = (\mathbf{f}^\mathrm{PCA}(X), \mathbf{g}^\mathrm{PCA}(Y))$ is a sufficient statistic pair for inferences about (U, V).

Beyond this example, for a general jointly Gaussian pair (X, Y), our statistics $S = \mathbf{f}(X)$ and $T = \mathbf{g}(Y)$ specialize to the PCA statistics (36) 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, which is equivalent to the condition that $\mathbf{\Lambda}_X$ and $\mathbf{\Lambda}_Y$ commute (see, e.g., [3, 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., [3, Theorem 2.4.3.2 and Problem 1.3.P4]).

Acknowledgment

The research of Shao-Lun Huang was funded by the Shenzhen Municipal Scientific Program JCYJ20170818094022586.

References

[1] 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.

[2] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *J. Am. Stat. Assoc.*, vol. 80, no. 391, pp. 580–598, Sep. 1985.

[3] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. Cambridge, UK: Cambridge University Press, 2012.

[4] M. A. Chmielewski, "Elliptically symmetric distributions: A review and bibliography," *Int. Stat. Review*, vol. 49, no. 1, pp. 67–74, Apr. 1981.

[5] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

[6] A. S. Lewis, "The convex analysis of unitarily invariant matrix functions," *J. Convex Anal.*, vol. 2, no. 1/2, pp. 173–183, 1995.

[7] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for Gaussian variables," *J. Machine Learning Res.*, vol. 6, pp. 165–188, May 2005.

[8] A. D. Wyner, "The common information of two dependent random variables," *IEEE Trans. Inform. Theory*, vol. 21, no. 2, pp. 163–179, Mar. 1975.

[9] F. Bach, J. Mairal, and J. Ponce, "Convex sparse matrix factorizations," CNRS, Paris, France, Tech. Rep. HAL-00345747, 2008.

[10] 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, pp. 471–501, 2010.

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

[12] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Ed. Psych.*, vol. 24, pp. 417–441, 498–520, 1933.

[13] S. Satpathy and P. Cuff, "Gaussian secure source coding and Wyner's common information," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2015, pp. 116–120.