IPSJ Transactions on Mathematical Modeling and Its Applications Vol.6 No.1 136–145 (Mar. 2013)

Regular Paper

Designing Various Multivariate Analysis at Will via Generalized Pairwise Expression*

Akisato Kimura[1,a)] Masashi Sugiyama[2] Hitoshi Sakano[1] Hirokazu Kameoka[3,1]

Received: August 22, 2012, Revised: October 12, 2012,
Accepted: November 26, 2012

Abstract: It is well known that dimensionality reduction based on multivariate analysis methods and their kernelized extensions can be formulated as generalized eigenvalue problems of scatter matrices, Gram matrices or their augmented matrices. This paper provides a generic and theoretical framework of multivariate analysis introducing a new expression for scatter matrices and Gram matrices, called Generalized Pairwise Expression (GPE). This expression is quite compact but highly powerful. The framework includes not only (1) the traditional multivariate analysis methods but also (2) several regularization techniques, (3) localization techniques, (4) clustering methods based on generalized eigenvalue problems, and (5) their semi-supervised extensions. This paper also presents a methodology for designing a desired multivariate analysis method from the proposed framework. The methodology is quite simple: adopting the above mentioned special cases as templates, and generating a new method by combining these templates appropriately. Through this methodology, we can freely design various tailor-made methods for specific purposes or domains.

Keywords: multivariate analysis, dimensionality reduction, generalized eigenvalue problem, pairwise expression, kernel method, clustering, semi-supervised learning, regularization

1. Introduction

We can easily obtain a massive collection of texts (long articles [*1], microblogs [*2]), images [1], [2], [3], [4], videos [5], [6] and musics [7] [*3] nowadays. However, we are now facing a difficulty in finding an intrinsic trend and nature of such a massive collection of data. Multivariate analysis [8] is traditional, quite simple but might be one of the powerful tools to obtain a hidden structure embedded in the data, via classification, regression and clustering [9], [10]. Actually, multivariate analysis has been still an important tool, and recent reports showed its effectiveness for several applications, e.g., human detection [11], image annotation [12], [13], and sensor data mining [14], [15], [16], [17].

Principal component analysis (PCA) [18], Fisher discriminant analysis (FDA) [19], multivariate linear regression (MLR), canonical correlation analysis (CCA) [18], and partial least squares (PLS) [20] are well known as standard multivariate analysis methods. These methods can be formulated as a generalized eigenvalue problem of a scatter matrix or an augmented matrix composed of several scatter matrices. Several extended researches tried to tackle the small sample size problem [21], i.e., the situation where the number of training samples is small compared with their dimensionality (e.g., robust PCA [22], [23], [24], [25] and robust FDA [26], [27], [28]). Kernelized extensions of those standard methods have been also developed to deal with non-vector samples and non-linear analysis (e.g., kernel PCA [29], kernel FDA [30], [31], [32], kernel MLR [33], and kernel CCA [34], [35]). They can be formulated as a generalized eigenvalue problem of an augmented matrix composed of Gram matrices, instead of scatter matrices. Kernel multivariate analysis often needs some regularization techniques such as ℓ_2-norm regularization [36], [37], [38] to inhibit overfitting and the graph Laplacian method [39] to fit underlying data manifolds smoothly. In addition, improvements of robustness against outliers and non-Gaussianity (i.e., multi-dimensional scaling (MDS) [40], locality preserving projection (LPP) [41] and local Fisher discriminant analysis (LFDA) [42]) and their extensions to semi-supervised dimensionality reduction [39], [43], [44] have been considered.

In addition, a lot of multivariate analysis methods and several trials to unify these methods have been presented so far. Borge et al. [45] and De Bie et al. [46] showed that several major linear multivariate analysis method can be formulated by a unified form of generalized eigenvalue problems by introducing the augmented matrix expression. Sun et al. [47], [48] showed the equivalence between a certain class of generalized eigenvalue problems and least squares problems under a mild assumption. De la Torre [49], [50] further extended their work to a various kind of component analysis methods by introducing the formulation of least-squares weighted kernel reduced rank regression (LS-WKRRR). However, freely designing a tailor-made multi-

[1] NTT Communication Science Laboratories, NTT Corporation, Soraku, Kyoto 619–0237, Japan
[2] Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Meguro, Tokyo 152–8552, Japan
[3] Graduate School of Information Science and Technologies, the University of Tokyo, Bunkyo, Tokyo 113–8656, Japan
[a)] akisato@ieee.org

* A preliminary version of this paper was previously presented in Conference on International Association for Pattern Recognition (ICPR2012).
[*1] New York Times Article Archive: http://www.nytimes.com/ref/membercenter/nytarchive.html
[*2] Tweets2011 corpus for TREC2011 microblog track: http://trec.nist.gov/data/tweets/
[*3] Last.fm: http://www.lastfm.jp, Freesound: http://www.freesound.org

variate analysis for a specific purpose or domain still remains an open problem. Until now, researchers and engineers have had to choose one of the existing methods that seems best to address the problem of interest, or had to laboriously develop a new analysis method tailored specifically for that purpose.

In view of the above discussions, this paper provides a new expression of second-order statistics including covariance matrices and Gram matrices, which we call the *Generalized Pairwise Expression* (*GPE*). GPE is originated in Pairwise Expression (PE) [42], [43], [51] that tries to describe the relation between pairs of samples regarding whether pairs are close together or far apart. Through the PE framework, we can obtain an interpretation of a multivariate analysis method as to how it works. Our GPE framework extends the PE framework to larger classes of multivariate analysis methods, and it can be also diverted to designing a new multivariate analysis method. Our main contributions of this paper can be summarized as follows:

(1) GPE makes it easy to design a new multivariate analysis method with desired properties without any special knowledge of multivariate analysis.
(2) The methodology is quite simple: Exploiting the above mentioned existing methods as templates, and constructing a new method by combining these templates appropriately. This property has not been discussed yet in any previous researches to our best knowledge.
(3) It is also possible to individually select and arrange samples for calculating the second-order statistics of the methods to be combined, which enables us to extend multivariate analysis methods to semi-supervised ones and multi-modal ones, where some parts are calculated from only labeled samples, and the other parts are obtained from both labeled and unlabeled samples.

The rest of this paper is organized as follows: Section 2 defines a class of multivariate analysis methods we are concerned with in this paper. Next, Section 3 describes our proposed framework, GPE, and its fundamental properties. These properties provide a methodology to design multivariate analysis methods with desired characteristics. Then, Section 4 reviews major multivariate analysis methods from the viewpoint of GPE. This review will give us templates of the GPEs for designing desired methods. After the above preparations, Section 5 demonstrates how to design a new multivariate analysis method. By replicating the methodology shown in the preceding sections, we can easily design various multivariate analysis methods at will. Additionally, Section 6 describes a non-linear and/or non-vector extension of GPE with the help of the kernel trick, which is not trivial. With this extension, non-linear dimensionality reduction, and several clustering methods are all in the class of multivariate analysis methods we are concerned with.

2. Multivariate Analysis for Vector Data

2.1 Preliminaries

Consider two sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ of samples [*4], where each set contains N_x and N_y samples, and each sample can be expressed as a vector with d_x and d_y dimensions, respectively, as follows:

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_x}\},$$
$$\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N, \boldsymbol{y}_{N_x+1}, \ldots, \boldsymbol{y}_{N_x+N_y-N}\} \quad (N \le N_x).$$

For brevity, both of the sample sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ are supposed to be centered on the origin by subtracting the mean from each component. Suppose that samples $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$ with the same suffix are co-occurring. Each set $\boldsymbol{X}$ and $\boldsymbol{Y}$ of samples is separated into the following two types: *Complete sample sets* $\boldsymbol{X}^{(\mathrm{C})}$ and $\boldsymbol{Y}^{(\mathrm{C})}$ so that every sample $\boldsymbol{x}_n$ (resp. $\boldsymbol{y}_n$) has co-occurring sample $\boldsymbol{y}_n$ (resp. $\boldsymbol{x}_n$), and *incomplete sample sets* $\boldsymbol{X}^{(\mathrm{I})}$ and $\boldsymbol{Y}^{(\mathrm{I})}$ so that every sample $\boldsymbol{x}_n$ (resp. $\boldsymbol{y}_n$) cannot find the co-occurring sample.

$$\begin{aligned}
\boldsymbol{X}^{(\mathrm{C})} &= \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\},\\
&= \{\boldsymbol{x}_1^{(\mathrm{C})}, \boldsymbol{x}_2^{(\mathrm{C})}, \ldots, \boldsymbol{x}_N^{(\mathrm{C})}\},\\
\boldsymbol{Y}^{(\mathrm{C})} &= \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\},\\
&= \{\boldsymbol{y}_1^{(\mathrm{C})}, \boldsymbol{y}_2^{(\mathrm{C})}, \ldots, \boldsymbol{y}_N^{(\mathrm{C})}\},\\
\boldsymbol{X}^{(\mathrm{I})} &= \{\boldsymbol{x}_{N+1}, \boldsymbol{x}_{N+2}, \ldots, \boldsymbol{x}_{N_x}\},\\
&= \{\boldsymbol{x}_1^{(\mathrm{I})}, \boldsymbol{x}_2^{(\mathrm{I})}, \ldots, \boldsymbol{x}_{N_x-N}^{(\mathrm{I})}\},\\
\boldsymbol{Y}^{(\mathrm{I})} &= \{\boldsymbol{y}_{N_x+1}, \boldsymbol{y}_{N_x+2}, \ldots, \boldsymbol{y}_{N_x+N_y-N}\},\\
&= \{\boldsymbol{y}_1^{(\mathrm{I})}, \boldsymbol{y}_2^{(\mathrm{I})}, \ldots, \boldsymbol{y}_{N_y-N}^{(\mathrm{I})}\}
\end{aligned}$$

First, we concentrate on the case that $N_x = N_y = N$, namely all the samples are paired, unless otherwise stated.

2.2 Formulation

Many linear multivariate analysis methods developed so far involve an optimization problem of the following form for a d-dimensional vector $\boldsymbol{w}$:

$$\boldsymbol{w}^{(\mathrm{opt})} = \arg\max_{\boldsymbol{w}\in\mathcal{R}^d} R(\boldsymbol{w}), \tag{1}$$
$$R(\boldsymbol{w}) = \boldsymbol{w}^\top \overline{\boldsymbol{C}}\boldsymbol{w}(\boldsymbol{w}^\top \underline{\boldsymbol{C}}\boldsymbol{w})^{-1},$$

where $\overline{\boldsymbol{C}}$ and $\underline{\boldsymbol{C}}$ are square matrices with certain statistical nature. For example, $\overline{\boldsymbol{C}}$ is a scatter matrix of $\boldsymbol{X}$ and $\underline{\boldsymbol{C}}$ is an identity matrix in PCA, and $\overline{\boldsymbol{C}}$ is a between-class scatter matrix and $\underline{\boldsymbol{C}}$ is a within-class scatter matrix in FDA. Roughly speaking, $\overline{\boldsymbol{C}}$ encodes the quantity that we want to increase, and $\underline{\boldsymbol{C}}$ corresponds to the quantity that we want to decrease. The denominator of the function $R(\boldsymbol{w})$ is often normalized to remove scale ambiguity, resulting in the following form:

$$\boldsymbol{w}^{(\mathrm{opt})} = \arg\max_{\boldsymbol{w}\in\mathcal{R}^d} R_1(\boldsymbol{w}) \text{ s.t. } R_2(\boldsymbol{w}) = 1, \tag{2}$$
$$R_1(\boldsymbol{w}) = \boldsymbol{w}^\top \overline{\boldsymbol{C}}\boldsymbol{w}, \quad R_2(\boldsymbol{w}) = \boldsymbol{w}^\top \underline{\boldsymbol{C}}\boldsymbol{w}.$$

The above optimization problem can be converted to the following generalized eigenvalue problem via the Lagrange multiplier method:

$$\overline{\boldsymbol{C}}\boldsymbol{w} = \lambda \underline{\boldsymbol{C}}\boldsymbol{w}. \tag{3}$$

The solution $\boldsymbol{w}_k$ $(k = 1, 2, \ldots, r)$ of the above generalized eigenvalue problem gives a solution of the original multivariate analysis formulated in Eq. (1).

It can be confirmed that Eq. (1) is invariant against any kinds of linear transformations, i.e., a vector $\boldsymbol{U}\boldsymbol{w}^{(\mathrm{opt})}$ transformed by any

[*4] The following discussion can be easily extended to more than 2 sets of samples sets [52].

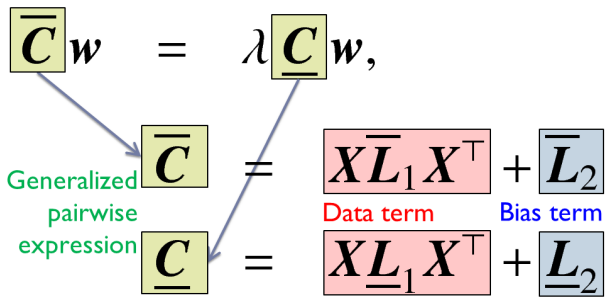

Fig. 1 Various multivariate analysis methods can be described via generalized pairwise expression (GPE).

r-dimensional unitary matrix $\boldsymbol{U}$ is also a global solution. This implies that the range of the embedding space can be uniquely determined by Eq. (1), but the metric in the embedding space is arbitrary. A practically useful heuristic is to set

$$\boldsymbol{U} = \mathrm{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_d}), \tag{4}$$

where $\mathrm{diag}(a, b, \cdots, c)$ denotes the diagonal matrix with the diagonal elements $a, b, \ldots, c$, and $\{\lambda_k\}_{k=1}^{r}$ denotes the generalized eigenvalues. Finally, we obtain the solution as

$$\boldsymbol{W}^{(\mathrm{opt})} = \{\sqrt{\lambda_1}\boldsymbol{w}_1, \sqrt{\lambda_2}\boldsymbol{w}_2, \ldots, \sqrt{\lambda_r}\boldsymbol{w}_r\}. \tag{5}$$

Thus, the minor eigenvectors are de-emphasized according to the square root of the eigenvalues.

3. Generalized Pairwise Expression

3.1 Definition

When addressing linear multivariate analysis methods, we often deal with the following type of second-order statistics [42], [51] as an extension of scatter matrices, since it is convenient to describe the relation between two features regarding whether they are close together or far apart:

$$\boldsymbol{S}_{\boldsymbol{Q},xx} = \sum_{n=1}^{N}\sum_{m=1}^{N} Q_{n,m}(\boldsymbol{x}_n - \boldsymbol{x}_m)(\boldsymbol{x}_n - \boldsymbol{x}_m)^\top,$$

where $\boldsymbol{Q}$ is an $N \times N$ non-negative, positive semi-definite and symmetric matrix [*5]. A typical example is the scatter matrix:

$$\boldsymbol{S}_{xx} = N^{-1}\textstyle\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n{}^\top.$$

Let $\boldsymbol{D}_{\boldsymbol{Q}}$ be the $N \times N$ diagonal matrix with

$$D_{\boldsymbol{Q},n,n} = \sum_{n_2=1}^{N} Q_{n,n_2},$$

and let $\boldsymbol{L}_{\boldsymbol{Q}}$ be $\boldsymbol{L}_{\boldsymbol{Q}} = \boldsymbol{D}_{\boldsymbol{Q}} - \boldsymbol{Q}$. Then, the matrix $\boldsymbol{S}_{\boldsymbol{Q},xx}$ can be expressed in terms of $\boldsymbol{L}_{\boldsymbol{Q}}$ as follows:

$$\boldsymbol{S}_{\boldsymbol{Q},xx} = \boldsymbol{X}\boldsymbol{L}_{\boldsymbol{Q}}\boldsymbol{X}^\top.$$

The above expression is called the *pairwise expression* (*PE*) of the second-order statistics $\boldsymbol{S}_{\boldsymbol{Q},xx}$. If $\boldsymbol{Q}$ is a weight matrix for a graph with N nodes, $\boldsymbol{L}_{\boldsymbol{Q}}$ can be regarded as a graph Laplacian matrix in the spectral graph theory. If $\boldsymbol{Q}$ is symmetric and its elements are all non-negative, $\boldsymbol{L}_{\boldsymbol{Q}}$ is known to be positive semi-definite.

[*5] When dealing with 2 sample sets in this framework, it is sufficient to introduce a concatenated sample set $\boldsymbol{Z} = (\boldsymbol{X}^\top, \boldsymbol{Y}^\top)^\top$.

Here, we extend PE to the following expression introducing an additional matrix independent of $\boldsymbol{Q}$:

$$\hat{\boldsymbol{S}}_{\boldsymbol{Q},xx} = \boldsymbol{X}\boldsymbol{L}_{\boldsymbol{Q},1}\boldsymbol{X}^\top + \boldsymbol{L}_2,$$

where $\boldsymbol{L}_{\boldsymbol{Q},1}$ is an $N \times N$ positive semi-definite matrix, and $\boldsymbol{L}_2$ is a $d_x \times d_x$ non-negative symmetric matrix. We do not have to explicitly consider the matrix $\boldsymbol{Q}$ for the following discussions:

$$\hat{\boldsymbol{S}}_{xx} = \boldsymbol{X}\boldsymbol{L}_1\boldsymbol{X}^\top + \boldsymbol{L}_2. \tag{6}$$

After all, we call this expression as the *generalized pairwise expression* (*GPE*) (See **Fig. 1**). The first term of Eq. (6) is called the *data term* since it depends on the sample data, and the second term is called the *bias term*.

3.2 Properties

We can derive the following fundamental properties of GPE from the definition, if the number of samples, N, is sufficiently large:

(1) If $\boldsymbol{A}$ is GPE and $\beta > 0$ is a constant, then $\beta\boldsymbol{A}$ is also GPE.
(2) If both $\boldsymbol{A}$ and $\boldsymbol{B}$ are GPE, then $\boldsymbol{A} + \boldsymbol{B}$ is also GPE.
(3) If both $\boldsymbol{A}$ and $\boldsymbol{B}$ are GPE, then $\boldsymbol{A}\boldsymbol{B}$ is also GPE.

Proof. The first and second claims can be easily proved, so we concentrate on proving the third one.

First, let us denote $\boldsymbol{A}$ and $\boldsymbol{B}$ as follows:

$$\boldsymbol{A} = \boldsymbol{X}\boldsymbol{L}_{A1}\boldsymbol{X}^\top + \boldsymbol{L}_{A2},$$
$$\boldsymbol{B} = \boldsymbol{Y}\boldsymbol{L}_{B1}\boldsymbol{X}^\top + \boldsymbol{L}_{B2},$$

where $\boldsymbol{L}_{A1}$ and $\boldsymbol{L}_{B1}$ are positive semi-definite $N \times N$ matrices, and $\boldsymbol{L}_{A2}$ and $\boldsymbol{L}_{B2}$ are $d_x \times d_x$ non-negative symmetric matrices. Then, we obtain

$$\begin{aligned}\boldsymbol{A}\boldsymbol{B} &= (\boldsymbol{X}\boldsymbol{L}_{A1}\boldsymbol{X}^\top + \boldsymbol{L}_{A2})(\boldsymbol{X}\boldsymbol{L}_{B1}\boldsymbol{X}^\top + \boldsymbol{L}_{B2}),\\ &= \boldsymbol{X}(\boldsymbol{L}_{A1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{L}_{B1})\boldsymbol{X}^\top\\ &\quad + \boldsymbol{L}_{A2}\boldsymbol{X}\boldsymbol{L}_{B1}\boldsymbol{X}^\top + \boldsymbol{X}\boldsymbol{L}_{A1}\boldsymbol{X}^\top\boldsymbol{L}_{B2} + \boldsymbol{L}_{A2}\boldsymbol{L}_{B2}.\end{aligned}$$

We can find some matrices $\boldsymbol{L}_{Ci}$ $(i = 1, 2, 3)$ satisfying the following relationships, if $N \geq d_x$:

$$\boldsymbol{L}_{C1} = \boldsymbol{L}_{A1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{L}_{B1},$$
$$\boldsymbol{L}_{C2}\boldsymbol{X}^\top = \boldsymbol{L}_{A1}\boldsymbol{X}^\top\boldsymbol{L}_{B2},$$
$$\boldsymbol{X}\boldsymbol{L}_{C3} = \boldsymbol{L}_{A2}\boldsymbol{X}\boldsymbol{L}_{B1}.$$

Here, we will show that those matrices $\boldsymbol{L}_{Ci}$ $(i = 1, 2, 3)$ are all positive semi-definite. For any matrix $\boldsymbol{X} \in \mathcal{R}^{d_x \times N}$, we have

$$\boldsymbol{X}\boldsymbol{L}_{C1}\boldsymbol{X}^\top = (\boldsymbol{X}\boldsymbol{L}_{A1}\boldsymbol{X}^\top)(\boldsymbol{X}\boldsymbol{L}_{B1}\boldsymbol{X}^\top),$$
$$\boldsymbol{X}\boldsymbol{L}_{C2}\boldsymbol{X}^\top = (\boldsymbol{X}\boldsymbol{L}_{A1}\boldsymbol{X}^\top)\boldsymbol{L}_{B2},$$
$$\boldsymbol{X}\boldsymbol{L}_{C3}\boldsymbol{X}^\top = \boldsymbol{L}_{A2}(\boldsymbol{X}\boldsymbol{L}_{B1}\boldsymbol{X}^\top).$$

Recalling that $\boldsymbol{L}_{A1}$ and $\boldsymbol{L}_{B1}$ are positive semi-definite and $\boldsymbol{L}_{A2}$ and $\boldsymbol{L}_{B2}$ are non-negative symmetric, we can see that all the above matrix product $\boldsymbol{X}\boldsymbol{L}_{Ci}\boldsymbol{X}^\top$ $(i = 1, 2, 3)$ are non-negative, which means that the matrices $\boldsymbol{L}_{Ci}$ $(i = 1, 2, 3)$ are all positive semi-definite. This implies that

$$\begin{aligned}&\boldsymbol{A}\boldsymbol{B}\\ &= \boldsymbol{X}(\boldsymbol{L}_{C1} + \boldsymbol{L}_{C2} + \boldsymbol{L}_{C3})\boldsymbol{X}^\top + \boldsymbol{L}_{A2}\boldsymbol{L}_{B2}\end{aligned}$$

$$= \boldsymbol{X}\boldsymbol{L}_{D1}\boldsymbol{X}^\top + \boldsymbol{L}_{D2},$$

for a positive semi-definite matrix $\boldsymbol{L}_{D1}$ and a non-negative symmetric matrix $\boldsymbol{L}_{D2}$, which means $\boldsymbol{A}\boldsymbol{B}$ is also GPE. □

These fundamental properties of GPE provide us a promising way to design various multivariate analysis methods very easily, namely with addition, weighting and multiplication of GPEs of existing methods with desired characteristics. The rest of the problem is to reveal GPE of existing methods and the function of every type of combinations (addition and/or multiplication), which will be described in the next section.

4. Reviewing Multivariate Analysis

4.1 Preliminaries

This section reviews major multivariate analysis methods from the viewpoint of GPE. As shown in Sections 2 and 3, the GPEs of PCA and FDA respectively are given by

$$\overline{\boldsymbol{C}}^{(\mathrm{PCA})} = \boldsymbol{S}_{xx}, \quad \underline{\boldsymbol{C}}^{(\mathrm{PCA})} = \boldsymbol{I}_{d_x},$$
$$\overline{\boldsymbol{C}}^{(\mathrm{FDA})} = \boldsymbol{S}_{xx}^{(\mathrm{b})}, \quad \underline{\boldsymbol{C}}^{(\mathrm{FDA})} = \boldsymbol{S}_{xx}^{(\mathrm{w})},$$

where $\boldsymbol{S}_{xx}^{(\mathrm{b})}$ and $\boldsymbol{S}_{xx}^{(\mathrm{w})}$ are respectively between-class and within-class scatter matrices of $\boldsymbol{X}$. From these examples, a scatter matrix $\boldsymbol{S}_{xx}$ is a typical example of the data term in GPE, and an identity matrix $\boldsymbol{I}_d$ is a typical example of the bias term. Note that unlike FDA, $\underline{\boldsymbol{C}}^{(\mathrm{PCA})}$ does not have a PE since $\boldsymbol{I}_{d_x}$ cannot be expressed in a pairwise form. This indicates the significance of introducing GPE when reviewing various multivariate analysis methods within a unified framwork.

4.2 Canonical Correlation Analysis (CCA)

Canonical correlation analysis (CCA) [18] is a method of correlating linear relationships between two sample sets. Formally, CCA finds a new coordinate $(\boldsymbol{w}_x, \boldsymbol{w}_y)$ to maximize the correlation between the two vectors in the new coordinates. In other words, the function $\rho(\boldsymbol{w}_x, \boldsymbol{w}_y|\boldsymbol{X}, \boldsymbol{Y})$ to be maximized is

$$\begin{aligned}
&\rho^{(\mathrm{CCA})}(\boldsymbol{w}_x, \boldsymbol{w}_y|\boldsymbol{X}, \boldsymbol{Y}) \\
&= \frac{\langle \boldsymbol{X}^\top\boldsymbol{w}_x, \boldsymbol{Y}^\top\boldsymbol{w}_y\rangle}{\|\boldsymbol{X}^\top\boldsymbol{w}_x\| \cdot \|\boldsymbol{Y}^\top\boldsymbol{w}_y\|} \\
&= \max_{(\boldsymbol{w}_x,\boldsymbol{w}_y)} \frac{\widehat{E}[\langle \boldsymbol{w}_x, \boldsymbol{x}\rangle\langle \boldsymbol{w}_y, \boldsymbol{y}\rangle]}{\sqrt{\widehat{E}[\langle \boldsymbol{w}_x, \boldsymbol{x}\rangle^2] \cdot \widehat{E}[\langle \boldsymbol{w}_y, \boldsymbol{y}\rangle^2]}} \\
&= \max_{(\boldsymbol{w}_x,\boldsymbol{w}_y)} \frac{\boldsymbol{w}_x^\top \widehat{E}[\boldsymbol{x}\boldsymbol{y}^\top]\boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^\top \widehat{E}[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{w}_x \boldsymbol{w}_y^\top \widehat{E}[\boldsymbol{y}\boldsymbol{y}^\top]\boldsymbol{w}_y}} \\
&= \frac{\boldsymbol{w}_x^\top \boldsymbol{S}_{xy}\boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^\top \boldsymbol{S}_{xx}\boldsymbol{w}_x \boldsymbol{w}_y^\top \boldsymbol{S}_{yy}\boldsymbol{w}_y}}, \qquad (7)
\end{aligned}$$

where $\hat{E}[\cdot]$ denotes an empirical expectation. The maximum of the function $\rho(\boldsymbol{X}^{(\mathrm{C})}, \boldsymbol{Y}^{(\mathrm{C})})$ is not affected by re-scaling $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ either together or independently. Therefore, the maximization of $\rho(\boldsymbol{X}^{(\mathrm{C})}, \boldsymbol{Y}^{(\mathrm{C})})$ is equivalent to maximizing the numerator of $\rho(\boldsymbol{X}^{(\mathrm{C})}, \boldsymbol{Y}^{(\mathrm{C})})$ subject to

$$\boldsymbol{w}_x^\top \boldsymbol{S}_{xx}\boldsymbol{w}_x = \boldsymbol{w}_y^\top \boldsymbol{S}_{yy}\boldsymbol{w}_y = 1.$$

Taking derivatives of the corresponding Lagrangian with respect to $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$, we obtain

$$\boldsymbol{S}_{xy}\boldsymbol{w}_y - \lambda \boldsymbol{S}_{xx}\boldsymbol{w}_x = \boldsymbol{0},$$
$$\boldsymbol{S}_{yx}\boldsymbol{w}_x - \lambda \boldsymbol{S}_{yy}\boldsymbol{w}_y = \boldsymbol{0},$$

where λ is a Lagrange multiplier.

From the above discussion, the GPE of CCA can be obtained as follows:

$$\overline{\boldsymbol{C}}^{(\mathrm{CCA})} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{S}_{xy} \\ \boldsymbol{S}_{yx} & \boldsymbol{0} \end{pmatrix}, \quad \underline{\boldsymbol{C}}^{(\mathrm{CCA})} = \begin{pmatrix} \boldsymbol{S}_{xx} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_{yy} \end{pmatrix},$$
$$\boldsymbol{w} = (\boldsymbol{w}_x^\top, \boldsymbol{w}_y^\top)^\top.$$

We additionally note that when every sample $\boldsymbol{y}_n$ in $\boldsymbol{Y}$ represents a class indicator vectors, namely $\boldsymbol{y}_n \in \{0, 1\}^M$, $\sum_{m=1}^M y_{n,m} = 1$ and M is the number of classes, CCA is reduced to FDA [53] [*6]. Thus, CCA can be regarded as a generalized variant of FDA so that each sample can belong to multiple classes.

4.3 Multiple Linear Regression (MLR)

Multiple linear regression (MLR) is a method of finding a projection matrix $\boldsymbol{W}$ with the minimum squared error between $\boldsymbol{y}$ and its linear approximation $\boldsymbol{W}\boldsymbol{x}$. For simplicity, we first consider the case that the projection matrix $\boldsymbol{W}$ is with rank 1, which can be written as a direct product of two bases $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$. This assumption is useful to understand MLR from the viewpoint of GPE. Then, the objective function to be minimized is the following squared error:

$$\begin{aligned}
&\epsilon^{(\mathrm{MLR})}(\boldsymbol{w}_x, \boldsymbol{w}_y|\boldsymbol{X}, \boldsymbol{Y}) \\
&= \hat{E}\left[\|\boldsymbol{y} - \alpha\boldsymbol{w}_y\boldsymbol{w}_x^\top\boldsymbol{x}\|^2\right] \\
&= \hat{E}\left[\boldsymbol{y}^\top\boldsymbol{y}\right] - 2\alpha\boldsymbol{w}_y^\top\hat{E}\left[\boldsymbol{y}^\top\boldsymbol{x}\right]\boldsymbol{w}_x + \alpha^2\boldsymbol{w}_x^\top\hat{E}\left[\boldsymbol{x}^\top\boldsymbol{x}\right]\boldsymbol{w}_x \\
&= \hat{E}\left[\boldsymbol{y}^\top\boldsymbol{y}\right] - 2\alpha\boldsymbol{w}_y^\top\boldsymbol{S}_{xy}^\top\boldsymbol{w}_x + \alpha^2\boldsymbol{w}_x^\top\boldsymbol{S}_{xx}\boldsymbol{w}_x.
\end{aligned}$$

To get an expression for α, we calculate the derivative

$$\begin{aligned}
&\frac{\partial}{\partial\alpha}\epsilon^{(\mathrm{MLR})}(\boldsymbol{w}_x, \boldsymbol{w}_y|\boldsymbol{X}, \boldsymbol{Y}) \\
&= 2(\alpha\boldsymbol{w}_x^\top\boldsymbol{S}_{xx}\boldsymbol{w}_x - \boldsymbol{w}_y^\top\boldsymbol{S}_{xy}^\top\boldsymbol{w}_x) = 0,
\end{aligned}$$

which gives

$$\alpha = (\boldsymbol{w}_y^\top\boldsymbol{S}_{xy}^\top\boldsymbol{w}_x)(\boldsymbol{w}_x^\top\boldsymbol{S}_{xx}\boldsymbol{w}_x)^{-1}.$$

Then, we obtain

$$\begin{aligned}
&\epsilon^{(\mathrm{MLR})}(\boldsymbol{w}_x, \boldsymbol{w}_y|\boldsymbol{X}, \boldsymbol{Y}) \\
&= E\left[\boldsymbol{y}^\top\boldsymbol{y}\right] - \frac{(\boldsymbol{w}_y^\top\boldsymbol{S}_{xy}^\top\boldsymbol{w}_x)^2}{\boldsymbol{w}_x^\top\boldsymbol{S}_{xx}\boldsymbol{w}_x}. \qquad (8)
\end{aligned}$$

Since the squared error cannot be negative and the first term of the objective function is independent of the two directions $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$, we can minimize it by maximizing the following generalized Rayleigh quotient:

$$\rho^{(\mathrm{MLR})}(\boldsymbol{w}_x, \boldsymbol{w}_y|\boldsymbol{X}, \boldsymbol{Y}) = \frac{\boldsymbol{w}_x^\top\boldsymbol{S}_{xy}\boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^\top\boldsymbol{S}_{xx}\boldsymbol{w}_x\boldsymbol{w}_y^\top\boldsymbol{w}_y}},$$

[*6] Note that the technical report [53] includes several mistakes in the discussion as to the equivalence between CCA and FDA.

where $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ are supposed to be normalized as $\boldsymbol{w}_x^\top \boldsymbol{S}_{xx}\boldsymbol{w}_x = 1$ and $\boldsymbol{w}_y^\top \boldsymbol{w}_y = 1$. By comparing the above equation and Eq. (7) and the objective function for CCA, we can see that MLR is a special case of CCA, and

$$\overline{\boldsymbol{C}}^{(\mathrm{MLR})} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{S}_{xy} \\ \boldsymbol{S}_{yx} & \boldsymbol{0} \end{pmatrix}, \quad \underline{\boldsymbol{C}}^{(\mathrm{MLR})} = \begin{pmatrix} \boldsymbol{S}_{xx} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{d_y} \end{pmatrix},$$
$$\boldsymbol{w} = (\boldsymbol{w}_x^\top, \boldsymbol{w}_y^\top)^\top.$$

The above derivation shows a part of the equivalence between the generalized eigenproblem and the least squares, which have been already revealed by Sun et al. [47], [48]. This equivalence property will be often exploited in the following discussions.

4.4 Principal Component Regression (PCR)

Principal component regression (PCR) [54] is a variant of MLR that uses PCA when estimating regression coefficients $\boldsymbol{W}$. It is a procedure used to overcome problems which arise when the exploratory variables are nearly co-linear. In PCR, instead of regressing the dependent variable $\boldsymbol{y}$ on the independent variables $\boldsymbol{x}$ directly, the principal components $\boldsymbol{Vx}$ of the independent variables are used. One typically only uses a subset of the principal components in the regression, making a kind of regularized estimation. Often the principal components with the highest variance are selected. A larger class of multivariate analysis methods that introduces a latent model into the standard linear regression is called latent variable regression (LVR) [55].

In the same way as MLR, we assume that the projection matrix $\boldsymbol{W}$ is with rank 1, namely $\boldsymbol{W} = \boldsymbol{w}_y\boldsymbol{w}_x^\top$. A rank-$K$ approximation $\hat{\boldsymbol{X}}$ of the data matrix $\boldsymbol{X}$ can be obtained by singular value decomposition as

$$\hat{\boldsymbol{X}} = \boldsymbol{U}_K \Sigma_K \boldsymbol{V}_K^\top, \tag{9}$$

where Σ_K is a $K \times K$ diagonal matrix whose diagonal components are top-K eigenvalues obtained by PCA of $\boldsymbol{X}$, and $\boldsymbol{V}_K$ is a $K \times d_x$ matrix whose columns are the top-k eigenvectors. Then, the objective function of PCR to be minimized can be obtained by substituting $\hat{\boldsymbol{X}}$ into $\boldsymbol{X}$ in the objective function of MLR, as follows:

$$\begin{aligned}
&\epsilon^{(\mathrm{PCR})}(\boldsymbol{w}_x, \boldsymbol{w}_y | \boldsymbol{X}, \boldsymbol{Y}) \\
&= \epsilon^{(\mathrm{MLR})}(\boldsymbol{w}_x, \boldsymbol{w}_y | \hat{\boldsymbol{X}}, \boldsymbol{Y}) \\
&= \hat{E}\left[\|\boldsymbol{y} - \alpha \boldsymbol{w}_y \boldsymbol{w}_x^\top \hat{\boldsymbol{x}}\|^2\right] \\
&= \hat{E}\left[\boldsymbol{y}^\top \boldsymbol{y}\right] - 2\alpha \boldsymbol{w}_y^\top \hat{E}\left[\boldsymbol{y}^\top \hat{\boldsymbol{x}}\right] \boldsymbol{w}_x + \alpha^2 \boldsymbol{w}_x^\top \hat{E}\left[\boldsymbol{y}^\top \hat{\boldsymbol{x}}\right] \boldsymbol{w}_x \\
&= \hat{E}\left[\boldsymbol{y}^\top \boldsymbol{y}\right] - 2\alpha \boldsymbol{w}_y^\top \boldsymbol{S}_{\hat{x}y}^\top \boldsymbol{w}_x + \alpha^2 \boldsymbol{w}_x^\top \boldsymbol{S}_{\hat{x}\hat{x}} \boldsymbol{w}_x,
\end{aligned}$$

where $\boldsymbol{S}_{\hat{x}y}$ and $\boldsymbol{S}_{\hat{x}\hat{x}}$ can be obtained as follows:

$$\begin{aligned}
\boldsymbol{S}_{\hat{x}y} &= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{u}_{K,n} \Sigma_K \boldsymbol{V}_K^\top \boldsymbol{y}_n^\top, \\
\boldsymbol{S}_{\hat{x}\hat{x}} &= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{u}_{K,n} \Sigma_K \boldsymbol{V}_K^\top \boldsymbol{V}_K \Sigma_K \boldsymbol{u}_{K,n}, \\
&= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{u}_{K,n} (\Sigma_K)^2 \boldsymbol{u}_n^\top.
\end{aligned}$$

From the description of the previous subsection, we can obtain

$$\overline{\boldsymbol{C}}^{(\mathrm{PCR})} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{S}_{\hat{x}y} \\ \boldsymbol{S}_{\hat{x}y}^\top & \boldsymbol{0} \end{pmatrix}, \quad \underline{\boldsymbol{C}}^{(\mathrm{PCR})} = \begin{pmatrix} \boldsymbol{S}_{\hat{x}\hat{x}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{d_y} \end{pmatrix},$$
$$\boldsymbol{w} = (\boldsymbol{w}_x^\top, \boldsymbol{w}_y^\top)^\top.$$

4.5 Partial Least Squares (PLS)

Partial Least Squares (PLS) [20] (or sometimes called PLS regression) belongs to a family of latent variable regression (LVR), and tries to finds a direction for the observable sample set $\boldsymbol{X}$ that explains the maximum variance direction for the predicted sample set $\boldsymbol{Y}$. The contribution of PLS against the standard MLR and PCR is to simultaneously estimate the latent model and regression from the latent space to the predicted space, which leads to robust regression against noisy observations.

Although PLS cannot be formulated as a generalized eigenproblem in general, orthogonal PLS (OPLS) [56], [57] as a variant of the original PLS has a form of generalized eigenproblem. This improves the interpretability (but not the predictivity) of the original PLS. OPLS can be formulated as follows:

$$\boldsymbol{X}\boldsymbol{Y}^\top \boldsymbol{Y}\boldsymbol{X}^\top \boldsymbol{w} = \lambda \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{w},$$
$$\boldsymbol{w} = (\boldsymbol{w}_x^\top, \boldsymbol{w}_y^\top)^\top$$

meaning,

$$\begin{aligned}
\overline{\boldsymbol{C}}^{(\mathrm{OPLS})} &= \boldsymbol{X}\boldsymbol{Y}^\top \boldsymbol{Y}\boldsymbol{X}^\top &&\propto \boldsymbol{S}_{xy}\boldsymbol{S}_{xy}^\top, \\
\underline{\boldsymbol{C}}^{(\mathrm{OPLS})} &= \boldsymbol{X}\boldsymbol{X}^\top &&\propto \boldsymbol{S}_{xx}.
\end{aligned}$$

When every sample $\boldsymbol{y}_n$ in $\boldsymbol{Y}$ represents a class indicator vectors (cf. Section 4.2), OPLS is called OPLS-discriminant analysis (OPLS-DA) [57], which has been often used for the task of bio-marker identification [58].

4.6 ℓ_2-norm Regularization

ℓ_2-norm regularization is a popular regularization technique for various optimization problems including multivariate analysis. In the area of statistics or machine learning, this is sometimes called Tikhonov regularization [9], [37]. The most popular method that utilizes ℓ_2-norm regularization is ridge regression [36], which combines MLR and ℓ_2-norm regularization. The objective function to be minimized is the following squared error:

$$\begin{aligned}
&\epsilon^{(\mathrm{Ridge})}(\boldsymbol{w}_x, \boldsymbol{w}_y | \boldsymbol{X}, \boldsymbol{Y}) \\
&= \hat{E}\left[\|\boldsymbol{y} - \alpha \boldsymbol{w}_y \boldsymbol{w}_x^\top \boldsymbol{x}\|^2\right] + \delta \|\boldsymbol{w}_x\|^2 \\
&= \hat{E}\left[\boldsymbol{y}^\top \boldsymbol{y}\right] - 2\alpha \boldsymbol{w}_y^\top \boldsymbol{S}_{xy}^\top \boldsymbol{w}_x + \alpha^2 \boldsymbol{w}_x^\top \boldsymbol{S}_{xx} \boldsymbol{w}_x + \delta \|\boldsymbol{w}_x\|^2 \\
&= \hat{E}\left[\boldsymbol{y}^\top \boldsymbol{y}\right] - 2\alpha \boldsymbol{w}_y^\top \boldsymbol{S}_{xy}^\top \boldsymbol{w}_x + \alpha^2 \boldsymbol{w}_x^\top (\boldsymbol{S}_{xx} + \hat{\delta}\boldsymbol{I}_{d_x}) \boldsymbol{w}_x,
\end{aligned}$$

where $\hat{\delta} = \delta/\alpha^2$. From the above equation and the objective function of MLR, the GPE of ridge regression can be derived as

$$\overline{\boldsymbol{C}}^{(\mathrm{Ridge})} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{S}_{xy} \\ \boldsymbol{S}_{yx} & \boldsymbol{0} \end{pmatrix},$$
$$\underline{\boldsymbol{C}}^{(\mathrm{Ridge})} = \begin{pmatrix} \boldsymbol{S}_{xx} + \hat{\delta}\boldsymbol{I}_{d_x} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{d_y} \end{pmatrix},$$
$$\boldsymbol{w} = (\boldsymbol{w}_x^\top, \boldsymbol{w}_y^\top)^\top.$$

In a similar way to ridge regression, we can derive the GPE of

CCA with ℓ_2-norm regularization [38], [59] as

$$\overline{\boldsymbol{C}}^{(\mathrm{CCA}-\ell_2)} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{S}_{xy} \\ \boldsymbol{S}_{yx} & \boldsymbol{0} \end{pmatrix},$$
$$\underline{\boldsymbol{C}}^{(\mathrm{CCA}-\ell_2)} = \begin{pmatrix} \boldsymbol{S}_{xx} + \hat{\delta}\boldsymbol{I}_{d_x} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_{yy} + \hat{\delta}\boldsymbol{I}_{d_y} \end{pmatrix}.$$

In addition, we can incorporate ℓ_1-norm regularization into the GPE framework only if the objective generalized eigenproblem has the following form:

$$\boldsymbol{X}\boldsymbol{L}_{\boldsymbol{Q}}\boldsymbol{X}^\top\boldsymbol{w} = \lambda\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{w},$$

meaning

$$\boldsymbol{S}_{\boldsymbol{Q},xx}\boldsymbol{w} = \lambda\boldsymbol{S}_{xx}\boldsymbol{w}.$$

PCA, FDA, MLR, CCA, OPLS and several variants can be included in this form. The details can be found in the previous work [47].

As shown in the above discussion, one of the major motivations that introduce the bias term of GPE is to integrate some regularization techniques within the framework of GPE.

4.7 Locality Preserving Projection (LPP)

Locality preserving projections (LPP) [41] seeks for an embedding transformation such that nearby data pairs in the original space close in the embedding space. Thus, LPP can reduce the dimensionality without losing the local structure.

Let $\boldsymbol{A}$ be an affinity matrix, that is, the N-dimensional matrix with the (n,m)-th element $A_{n,m}$ being the affinity between $\boldsymbol{x}_n$ and $\boldsymbol{x}_m$. We assume that $A_{n,m} \in [0,1]$; $A_{n,m}$ is large if $\boldsymbol{x}_n$ and $\boldsymbol{x}_m$ are close and $A_{n,m}$ is small if $\boldsymbol{x}_n$ and $\boldsymbol{x}_m$ are far apart. There are several different manners of defining $\boldsymbol{A}$, such as using the local scaling heuristics [60], i.e.,

$$A_{n,m} = \exp\left\{-\frac{\|\boldsymbol{x}_n - \boldsymbol{x}_m\|^2}{\sigma_n\sigma_m}\right\},$$
$$\sigma_n = \|\boldsymbol{x}_n - \boldsymbol{x}_n^{(k)}\|,$$

where $\boldsymbol{x}_n^{(k)}$ is the k-th nearest neighbor of $\boldsymbol{x}_n$. A heuristic choice of $k = 7$ was shown to be useful through experiments [60]. The objective function to be minimized is the following weighted squared error:

$$\epsilon^{(\mathrm{LPP})}(\boldsymbol{w}|\boldsymbol{X}) = \sum_{n=1}^{N}\sum_{m=1}^{N} A_{n,m}\|\boldsymbol{w}^\top\boldsymbol{x}_n - \boldsymbol{w}^\top\boldsymbol{x}_m\|^2 \quad \text{s.t. } \boldsymbol{w}^\top\boldsymbol{X}\boldsymbol{D}_{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{w} = 1.$$

In the same way as the derivation of GPE (see Section 3), the above minimization can be converted to the following generalized eigenvalue problem:

$$\boldsymbol{X}\boldsymbol{L}_{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{w} = \lambda\boldsymbol{X}\boldsymbol{D}_{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{w}.$$

Thus, the GPE of LPP can be obtained as

$$\overline{\boldsymbol{C}}^{(\mathrm{LPP})} = \boldsymbol{X}\boldsymbol{L}_{\boldsymbol{A}}\boldsymbol{X}^\top, \qquad \underline{\boldsymbol{C}}^{(\mathrm{LPP})} = \boldsymbol{X}\boldsymbol{D}_{\boldsymbol{A}}\boldsymbol{X}^\top.$$

4.8 Local Fisher Discriminant Analysis (LFDA)

Local Fisher discriminant analysis (LFDA) [42] is a method for supervised dimensionality reduction, and an extension of Fisher discriminant analysis (FDA). LFDA can overcome the weakness of the original FDA against outliers. The point is the introduction of between-sample similarity matrix Q obtained from the affinity matrix, for calculating the between-class scatter matrix $\boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{lb})}$ and the within-class scatter matrix $\boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{lw})}$.

$$\boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{lb})} = \sum_{n=1}^{N}\sum_{m=1}^{N} Q_{n,m}^{(\mathrm{lb})}(\boldsymbol{x}_n - \boldsymbol{x}_m)(\boldsymbol{x}_n - \boldsymbol{x}_m)^\top,$$
$$\boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{lw})} = \sum_{n=1}^{N}\sum_{m=1}^{N} Q_{n,m}^{(\mathrm{lw})}(\boldsymbol{x}_n - \boldsymbol{x}_m)(\boldsymbol{x}_n - \boldsymbol{x}_m)^\top,$$

where $\boldsymbol{Q}^{(\mathrm{lb})}$ and $\boldsymbol{Q}^{(\mathrm{lw})}$ are the $N \times N$ matrices with

$$Q_{n,m}^{(\mathrm{lb})} = \begin{cases} A_{n,m}(1/N - 1/N_c) & \text{if } y_n = y_m = c, \\ 1/N & \text{if } y_n \neq y_m, \end{cases}$$
$$Q_{n,m}^{(\mathrm{lw})} = \begin{cases} A_{n,m}/N_c & \text{if } y_n = y_m = c, \\ 1/N & \text{if } y_n \neq y_m, \end{cases}$$

and N_c is the number of samples in class c. Note that the local scaling is computed in a class-wise manner in LFDA, since we want to preserve the within-class local structure. This also contributes to reducing the computational cost for nearest neighbor search when computing the local scaling.

From the above discussion, the GPE of LFDA can be obtained as follows:

$$\overline{\boldsymbol{C}}_{\boldsymbol{Q}}^{(\mathrm{LFDA})} = \boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{lb})}, \qquad \underline{\boldsymbol{C}}_{\boldsymbol{Q}}^{(\mathrm{LFDA})} = \boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{lw})}.$$

4.9 Semi-supervised LFDA (SELF)

Semi-supervised local Fisher discriminant analysis, called SELF [43], integrates LFDA as a supervised dimensionality reduction and PCA as an unsupervised dimensionality reduction. SELF brings us one example for designing multivariate analysis methods via the GPE framework from the following two viewpoints:

(1) combining several multivariate analysis methods via GPE by following the properties shown in Section 3.2,
(2) changing sample sets to calculate the data term in GPE, which provides us to extend the method to a semi-supervised one.

Assume that there are two samples sets $\boldsymbol{X}$ and $\boldsymbol{Y}$, each sample in $\boldsymbol{Y}$ represents a class indicator vector, and an incomplete sample set $\boldsymbol{X}^{(\mathrm{I})}$ only exists, namely there are at least one unlabeled samples in the sample set $\boldsymbol{X}$. In such cases, we can search for solutions that lie in the span of the larger sample set $\boldsymbol{X}$, and regularize the solution using the additional data. SELF looks for solutions that lie along an empirical estimate of the subspace spanned by all the samples. This gives increased robustness to the algorithm, and increases class separability in the absence of label information. In detail, SELF integrates the GPE ($\boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{C,lb})}$ and $\boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{C,lb})}$) of LFDA calculated only from the labeled samples (in other words, complete sample sets) and the GPE $\boldsymbol{S}_{xx}$ of PCA calculated from all the samples, as follows:

$$\overline{\boldsymbol{C}}_{\boldsymbol{Q}}^{(\mathrm{SELF})} = \beta \boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{C,lb})} + (1-\beta)\boldsymbol{S}_{xx},$$
$$\underline{\boldsymbol{C}}_{\boldsymbol{Q}}^{(\mathrm{SELF})} = \beta \boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{C,lw})} + (1-\beta)\boldsymbol{I}_{d_x},$$

where β is a hyper parameter satisfying $0 \leq \beta \leq 1$. When $\beta = 1$, SELF is equivalent to LFDA with only the labeled samples $(\boldsymbol{X}^{(\mathrm{C})}, \boldsymbol{Y}^{(\mathrm{C})})$. Meanwhile, when $\beta = 0$, SELF is equivalent to PCA with all samples in $\boldsymbol{X}$. Generally speaking, SELF inherits the properties of both LFDA and PCA, and their influences can be controlled by the parameter β.

4.10 Semi-supervised CCA

In a similar way to that of SELF, a semi-supervised extension of CCA can be derived, which is called SemiCCA [44].

Assume that there are two samples sets $\boldsymbol{X}$ and $\boldsymbol{Y}$, and each includes incomplete sample set $\boldsymbol{X}^{(\mathrm{I})}$ and $\boldsymbol{Y}^{(\mathrm{I})}$, namely there are at least one unpaired samples in both $\boldsymbol{X}$ and $\boldsymbol{Y}$. SemiCCA integrates the GPE of CCA calculated only from the complete sample sets and the GPE of PCA calculated from the complete and incomplete sample sets, as follows:

$$\overline{\boldsymbol{C}}^{(\mathrm{SemiCCA})} = \beta \begin{pmatrix} \boldsymbol{0} & \boldsymbol{S}_{xy}^{(\mathrm{C})} \\ \boldsymbol{S}_{yx}^{(\mathrm{C})} & \boldsymbol{0} \end{pmatrix} + (1-\beta)\begin{pmatrix} \boldsymbol{S}_{xx} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_{yy} \end{pmatrix},$$
$$\underline{\boldsymbol{C}}^{(\mathrm{SemiCCA})} = \beta \begin{pmatrix} \boldsymbol{S}_{xx}^{(\mathrm{I})} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_{yy}^{(\mathrm{I})} \end{pmatrix} + (1-\beta)\begin{pmatrix} \boldsymbol{I}_{d_x} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{d_y} \end{pmatrix},$$

When $\beta = 1$, SemiCCA is equivalent to CCA with only the complete samples $(\boldsymbol{X}^{(\mathrm{C})}, \boldsymbol{Y}^{(\mathrm{C})})$. Meanwhile, when $\beta = 0$, SemiCCA is equivalent to PCA with all samples in $\boldsymbol{X}$ and $\boldsymbol{Y}$ under the assumption that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are uncorrelated with each other.

Another type of semi-supervised extension of CCA has been developed by Blaschko et al. [39]. Please see the detail in Section 6.

5. How to Design New Methods

To summarize the discussions so far, we describe (1) GPEs of major existing methods, (2) the way for integrating several GPEs and (3) some semi-supervised extensions by changing the sample sets for calculating GPEs. This section shows that we can easily design new multivariate analysis methods at will by replicating those steps. Note that another way to generate new methods would be possible, and the following one is only one example.

One of the simple extensions is to integrate FDA as supervised dimensionality reduction and CCA as unsupervised dimensionality reduction with a latent model. Consider a problem of video categorization, where its training data includes image features $\boldsymbol{X}$, audio features $\boldsymbol{Y}$ and class indexes. Finding appropriate correlations of such three different modals would be still challenging. Several approaches might be possible: (1) FDA for concatenated features $(\boldsymbol{X}^\top, \boldsymbol{Y}^\top)^\top$, which cannot obtain appropriate correlations between two different types of feature vectors, (2) CCA for two features $(\boldsymbol{X}, \boldsymbol{Y})$ followed by FDA on the compressed domain, which cannot find class-wise differences of correlations.

Here, we newly introduce an integration of CCA and FDA, which enables us to extract class-wise differences of feature correlations as well as to achieve discriminative embedding simultaneously. In the following, we call this method CFDA for the simplicity. CFDA can be formulated by the following equations:

$$\overline{\boldsymbol{C}}_{\boldsymbol{Q}}^{(\mathrm{CFDA})} = \beta \begin{pmatrix} \boldsymbol{0} & \boldsymbol{S}_{xy} \\ \boldsymbol{S}_{yx} & \boldsymbol{0} \end{pmatrix} + (1-\beta)\boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{lb})}, \tag{10}$$

$$\underline{\boldsymbol{C}}_{\boldsymbol{Q}}^{(\mathrm{CFDA})} = \beta \begin{pmatrix} \boldsymbol{S}_{xx} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_{yy} \end{pmatrix} + (1-\beta)\boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{lw})}. \tag{11}$$

When $\beta = 1$ CFDA is equivalent to CCA, while when $\beta = 0$ CFDA is equivalent to FDA for concatenated features $(\boldsymbol{X}^\top, \boldsymbol{Y}^\top)^\top$.

We note that we do not have to explicitly consider GPEs of FDA and CCA when constructing CFDA. All what we need to design a new multivariate analysis method are that existing methods to be combined can be described by GPE and operations for the combination are shown in Section 3.2.

6. Kernelized Extensions

6.1 Kernelization of Standard Methods

Almost all the methods in the GPE framework can be kernelized in a similar manner to the existing ones. First, we describe kernel CCA [34], [35] and related regularization techniques.

The original CCA can be extended to, e.g., non-vectorial domains by defining kernels over $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$k_x(\boldsymbol{x}_n, \boldsymbol{x}_m) = \langle \phi_x(\boldsymbol{x}_n), \phi_x(\boldsymbol{x}_m) \rangle,$$
$$k_y(\boldsymbol{y}_n, \boldsymbol{y}_m) = \langle \phi_y(\boldsymbol{y}_n), \phi_y(\boldsymbol{y}_m) \rangle,$$

and searching for solutions that lie in the span of $\phi_x(\boldsymbol{x})$ and $\phi_y(\boldsymbol{y})$

$$\boldsymbol{w}_x = \sum_{n=1}^{N} \alpha_n \phi_x(\boldsymbol{x}_n), \quad \boldsymbol{w}_y = \sum_{n=1}^{N} \beta_n \phi_y(\boldsymbol{y}_n).$$

In this setting, we use the following empirical scatter matrix,

$$\hat{\boldsymbol{S}}_{xy} = \sum_{n=1}^{N} \phi_x(\boldsymbol{x}_i)\phi_y(\boldsymbol{y}_i)^\top.$$

Denoting the Gram matrices defined by the samples as $\boldsymbol{K}_x$ and $\boldsymbol{K}_y$, we can obtain the solution from the following optimization problem with respect to coefficient vectors, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$,

$$\rho^{(kCCA)}(\boldsymbol{w}_x, \boldsymbol{w}_y | \boldsymbol{X}, \boldsymbol{Y}) = \frac{\boldsymbol{\alpha}^\top \boldsymbol{K}_x \boldsymbol{K}_y \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^\top \boldsymbol{K}_x^2 \boldsymbol{\alpha} \boldsymbol{\beta}^\top \boldsymbol{K}_y^2 \boldsymbol{\beta}}}.$$

In the same way as CCA, the optimization can be achieved by solving the following generalized eigenvalue problem:

$$\overline{\boldsymbol{C}}^{(\mathrm{kCCA})} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \lambda \underline{\boldsymbol{C}}^{(\mathrm{kCCA})} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix},$$
$$\overline{\boldsymbol{C}}^{(\mathrm{kCCA})} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{K}_x \boldsymbol{K}_y \\ \boldsymbol{K}_y \boldsymbol{K}_x & \boldsymbol{0} \end{pmatrix},$$
$$\underline{\boldsymbol{C}}^{(\mathrm{kCCA})} = \begin{pmatrix} \boldsymbol{K}_x^2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{K}_y^2 \end{pmatrix}.$$

Although the bases $(\boldsymbol{w}_x, \boldsymbol{w}_y)$ cannot be explicitly obtained, the projection to those bases can be calculated with the help of the kernel trick:

$$\boldsymbol{w}_x^\top \phi_x(\boldsymbol{x}) = \sum_{n=1}^{N} \alpha_n k_x(\boldsymbol{x}, \boldsymbol{x}_n),$$

$$\boldsymbol{w}_y^\top \phi_y(\boldsymbol{y}) = \sum_{n=1}^{N} \beta_n k_y(\boldsymbol{y}, \boldsymbol{y}_n).$$

As discussed in Ref. [38], this optimization leads to degenerate solutions in the case that either $\boldsymbol{K}_x$ or $\boldsymbol{K}_y$ is not invertible. Therefore, the following ℓ_2-regularized formulation should be necessary in general:

$$\overline{\boldsymbol{C}}^{(\mathrm{kCCA}-\ell_2)} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{K}_x\boldsymbol{K}_y \\ \boldsymbol{K}_y\boldsymbol{K}_x & \boldsymbol{0} \end{pmatrix},$$
$$\underline{\boldsymbol{C}}^{(\mathrm{kCCA}-\ell_2)} = \begin{pmatrix} \boldsymbol{K}_x^2 + \delta_x\boldsymbol{K}_x & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{K}_y^2 + \delta_y\boldsymbol{K}_y \end{pmatrix}.$$

Another popular regularization technique is the graph Laplacian method [39], [61]. By using Laplacian regularization, we are able to learn directions that tend to lie along the data manifold estimated from a collection of data. Denoting the empirical graph Laplacian $\hat{\boldsymbol{L}}_x$ and $\hat{\boldsymbol{L}}_y$ obtained from $\boldsymbol{K}_x$ and $\boldsymbol{K}_y$, the formulation is replaced by the following equations:

$$\overline{\boldsymbol{C}}^{(\mathrm{kCCA-Lap})} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{K}_x\boldsymbol{K}_y \\ \boldsymbol{K}_y\boldsymbol{K}_x & \boldsymbol{0} \end{pmatrix},$$
$$\underline{\boldsymbol{C}}^{(\mathrm{kCCA-Lap})} = \begin{pmatrix} \boldsymbol{K}_x^2 + \gamma_x\boldsymbol{R}_x & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{K}_y^2 + \gamma_y\boldsymbol{R}_y \end{pmatrix},$$
$$\boldsymbol{R}_x = \boldsymbol{K}_x\hat{\boldsymbol{L}}_x\boldsymbol{K}_x, \quad \boldsymbol{R}_y = \boldsymbol{K}_y\hat{\boldsymbol{L}}_y\boldsymbol{K}_y.$$

6.2 Non-linear Embedding Methods

With the kernelized extension, non-linear dimensionality reduction such as locally linear embedding [62] and Laplacian eigenmaps [51] are also in the GPE framework.

6.2.1 Laplacian Eigenmaps

Laplacian eigenmaps [51] is one of the popular methods for non-linear embedding. The goal of Laplacian eigenmaps is to find an embedding that preserves the local structure of nearby high-dimensional samples. Laplacian eigenmaps exploits graph Laplacian of a neighborhood graph on the samples $\boldsymbol{X}$, where each edge measures the affinity between two samples. Since a set of edge weights can be expressed by a Gram matrix $\boldsymbol{K}_x$, the objective function of Laplacian eigenmaps to be minimized is

$$\rho^{(\mathrm{LE})}(\boldsymbol{w}_x|\boldsymbol{X}) = (\boldsymbol{\alpha}^\top\hat{\boldsymbol{L}}_x\boldsymbol{\alpha})(\boldsymbol{\alpha}^\top\hat{\boldsymbol{D}}_x\boldsymbol{\alpha})^{-1},$$

where $\hat{\boldsymbol{D}}_x$ is a diagonal matrix satisfying $\hat{\boldsymbol{D}}_x = \boldsymbol{K}_x + \hat{\boldsymbol{L}}_x$. Therefore, the GPE of Laplacian eigenmaps can be obtained as

$$\overline{\boldsymbol{C}}^{(\mathrm{LE})} = \hat{\boldsymbol{L}}_x, \qquad \underline{\boldsymbol{C}}^{(\mathrm{LE})} = \hat{\boldsymbol{D}}_x.$$

6.2.2 Locally Linear Embedding (LLE)

Locally linear embedding (LLE) [62] finds an embedding of the samples $\boldsymbol{X}$ that preserves the local structure of nearby samples in the high-dimensional space. LLE builds the embedding by preserving the geometry of pairwise relations between samples in the high-dimensional manifold. LLE first computes a Gram matrix $\boldsymbol{K}_x$ containing the structural information of the embedding by minimizing the following function:

$$\rho^{(\mathrm{LLE1})}(\boldsymbol{K}_x|\boldsymbol{X}) = \|\boldsymbol{X}(\boldsymbol{I}_N - \boldsymbol{K}_x)\|_F^2 \quad \text{s.t. } \boldsymbol{K}_x\boldsymbol{1}_N = \boldsymbol{1}_N,$$

where each column of the Gram matrix $\boldsymbol{K}_x$ has k non-zero values. This minimization can be solved via a linear system of equations. Once $\boldsymbol{K}_x$ is calculated, LLE next finds a base that minimizes

$$\rho^{(\mathrm{LLE2})}(\boldsymbol{w}_x|\boldsymbol{X}) = \boldsymbol{\alpha}^\top(\boldsymbol{I}_N - \boldsymbol{K}_x)^2\boldsymbol{\alpha} \quad \text{s.t. } \boldsymbol{\alpha}^\top\boldsymbol{\alpha} = 1.$$

Therefore, the GPE of LLE can be obtained as

$$\overline{\boldsymbol{C}}^{(\mathrm{LLE})} = (\boldsymbol{I}_N - \boldsymbol{K}_x)^2, \qquad \underline{\boldsymbol{C}}^{(\mathrm{LLE})} = \boldsymbol{I}_N.$$

6.3 Clustering Methods

With the benefit of the kernelized extension of the GPE framework, several clustering methods such as spectral clustering (SC) [13], [63], [64] and normalized cuts (NC) [65] can also be included in the class of multivariate analysis we are concerned with.

$$\overline{\boldsymbol{C}}^{(\mathrm{SC})} = \boldsymbol{L}_x, \qquad \underline{\boldsymbol{C}}^{(\mathrm{SC})} = \boldsymbol{D}_x,$$
$$\overline{\boldsymbol{C}}^{(\mathrm{NC})} = \boldsymbol{D}_x^{-1/2}\boldsymbol{L}_x\boldsymbol{D}_x^{-1/2}, \qquad \underline{\boldsymbol{C}}^{(\mathrm{NC})} = \boldsymbol{I}_N.$$

Kernel k-means [66] is also known to belong to this family if we admit ntroducing a certain iterative procedure [67]. The details can be seen in the preceding work by De la Torre [49], [50].

6.4 How to Design New Kernelized Methods

Integrating two methods within the kernelized GPE framework is not straightforwad, since a simple addition of Gram matrices is not GPE.

For example, let us consider the following generalized eigenproblem characterized by two matrices of GPE-style second-order statistics:

$$\overline{\boldsymbol{C}}\boldsymbol{w} = \lambda\underline{\boldsymbol{C}}\boldsymbol{w}, \qquad (12)$$
$$\overline{\boldsymbol{C}} = \boldsymbol{X}\overline{\boldsymbol{L}}_1\boldsymbol{X}^\top + \overline{\boldsymbol{L}}_2, \quad \underline{\boldsymbol{C}} = \boldsymbol{X}\underline{\boldsymbol{L}}_1\boldsymbol{X}^\top + \underline{\boldsymbol{L}}_2,$$

where $\overline{\boldsymbol{L}}_1$ and $\underline{\boldsymbol{L}}_1$ are both positive semi-definite matrices and $\overline{\boldsymbol{L}}_2$ and $\underline{\boldsymbol{L}}_2$ are both non-negative symmetric matrices, from the definition of GPE. The above generalized eigenproblem can be replaced with the following one without essentially changing the solution [43]

$$\overline{\boldsymbol{C}}_2\boldsymbol{w} = \lambda\underline{\boldsymbol{C}}_2\boldsymbol{w},$$
$$\overline{\boldsymbol{C}}_2 = \boldsymbol{X}\{\overline{\boldsymbol{L}}_1 + (\boldsymbol{X}^\top\overline{\boldsymbol{L}}_2\boldsymbol{X})^\dagger\}\boldsymbol{X}^\top = \boldsymbol{X}\overline{\boldsymbol{L}}_3\boldsymbol{X}^\top,$$
$$\underline{\boldsymbol{C}}_2 = \boldsymbol{X}\{\underline{\boldsymbol{L}}_1 + (\boldsymbol{X}^\top\underline{\boldsymbol{L}}_2\boldsymbol{X})^\dagger\}\boldsymbol{X}^\top = \boldsymbol{X}\underline{\boldsymbol{L}}_3\boldsymbol{X}^\top,$$

where † denotes the Moore-Penrose pseudo-inverse [68]. In this derivation, we used the fact that the matrices $\overline{\boldsymbol{C}}$ and $\overline{\boldsymbol{C}}_2$ (resp. $\underline{\boldsymbol{C}}$ and $\underline{\boldsymbol{C}}_2$) that characterize the generalized eigenproblems share the same range. Following the procedure shown in Sugiyama et al. [43], we can obtain a kernelized variant of the multivariate analysis method formulated by Eq. (12).

From the above discussion, we can see that when dealing with kernelized multivariate analysis, we have to explicitly derive GPEs of existing methods, and replace the data matrix with its Gram matrix.

7. Concluding Remarks

This paper provided a new expression of covariance matrices

and Gram matrices, which we call generalized pairwise expression (GPE). This provided a unified insight into various multivariate analysis methods and their extensions. GPE made it easy to design desired multivariate analysis methods by simple combinations of GPEs of existing methods as templates. According to this methodology, we designed several new multivariate analysis methods.

The GPE framework covers a wide variety of multivariate analysis methods, and thus the way we have presented in this paper for designing new methods is still one of the examples. Developing more general guidelines would be promising future work.

References

[1] Russell, B.C., Torralba, A., Murphy, K.P. and Freeman, W.T.: LabelMe: A database and web-based tool for image annotation, *International Journal on Computer Vision*, Vol.77, No.5, pp.157–173 (2008).

[2] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L.: ImageNet: A large-scale hierarchical image database, *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.248–255 (2009).

[3] Torralba, A., Fergus, R. and Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.30, No.11, pp.1958–1970 (2008).

[4] Wang, X.J., Zhang, L., Liu, M., Li, Y. and Ma, W.Y.: ARISTA - Image search to annotation on billions of web photos, *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.2987–2994 (2010).

[5] Smeaton, A.F., Over, P. and Kraaij, W.: Evaluation campaigns and trecvid, *Proc. ACM International Workshop on Multimedia Information Retrieval* (*MIR*), pp.321–330 (2006).

[6] Yuen, J., Russell, B.C., Liu, C. and Torralba, A.: LabelMe video: Building a video database with human annotations, *Proc. International Conference on Computer Vision* (*ICCV*), pp.1451–1458 (2009).

[7] Bertin-Mahieux, T., Ellis, D.P., Whitman, B. and Lamere, P.: The million song dataset, *Proc. International Conference on Music Information Retrieval* (*ISMIR*), pp.591–596 (2011).

[8] Anderson, T.W.: *An Introduction to Multivariate Statistical Analysis* (*Wiley Series in Probability and Statistics*), 3rd ed., Wiley-Interscience (2003).

[9] Bishop, C.M.: *Pattern Recognition and Machine Learning*, Springer (2006).

[10] Hastie, T., Tibshirani, R. and Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (*Springer Series in Statistics*), 2nd ed., Springer-Verlag (2009).

[11] Schwartz, W.R., Kembhavi, A., Harwood, D. and Davis, L.: Human detection using partial least squares analysis, *Proc. International Conference on Computer Vision* (*ICCV*), pp.24–31 (2009).

[12] Nakayama, H., Harada, T. and Kuniyoshi, Y.: Evaluation of dimensionality reduction methods for image auto-annotation, *Proc. British Machine Vision Conference* (*BMVC*), pp.94.1–94.12 (2010).

[13] Blaschko, M.B. and Lampert, C.H.: Correlational spectral clustering, *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.1–8 (2008).

[14] Wang, M., Perera, A. and Gutierrez-Osuna, R.: Principal discriminants analysis for small-sample-size problems: Application to chemical sensing, *Proc. IEEE Sensors*, pp.591–594 (2004).

[15] Pezeshki, A., Azimi-Sadjadi, M.R. and Scharf, L.L.: Undersea target classification using canonical correlation analysis, *IEEE Journal of Oceanic Engineering*, Vol.32, No.4, pp.948–955 (2007).

[16] Nielsen, A.A.: Multiset canonical correlations analysis and multispectral, truly multi-temporal remote sensing data, *IEEE Trans. Image Processing*, Vol.11, No.3, pp.293–305 (2002).

[17] Schizas, I., Giannakis, G. and Luo, Z.Q.: Distributed estimation using reduced-dimensionality sensor observations, *IEEE Trans. Signal Processing*, Vol.55, No.8, pp.4284–4299 (2007).

[18] Hotelling, H.: Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, Vol.24 (1933).

[19] Fisher, R.A.: The use of multiple measurements in taxonomic problems, *Annals Eugen.*, Vol.7, pp.179–188 (1936).

[20] Wold, H.: Estimation of principal components and related models by iterative least squares, *Multivariate Analysis*, pp.391–420, Academic Press (1966).

[21] Kanal, L. and Chandrasekaran, B.: On dimensionality and sample size in statistical pattern classification, *Pattern Recogn.*, Vol.3, No.3, pp.225–234 (1971).

[22] Campbell, N.A.: Robust procedures in multivariate analysis I: Robust covariance estimation, *Applied Statistics*, Vol.29, No.3, pp.231–237 (1980).

[23] De la Torre, F. and Black, M.: Robust principal component analysis for computer vision, *Proc. IEEE International Conference on Computer Vision* (*ICCV*), pp.362–369 (2001).

[24] De La Torre, F. and Black, M.J.: A framework for robust subspace learning, *International Journal of Computer Vision*, Vol.54, p.2003 (2003).

[25] Inoue, K., Hara, K. and Urahama, K.: Robust multilinear principal component analysis, *Proc. IEEE International Conference on Computer Vision* (*ICCV*), pp.591–597 (2009).

[26] Lu, J., Plataniotis, K. and Venetsanopoulos, A.: Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition, *Pattern Recogn. Lett.*, Vol.26, No.2, pp.181–191 (2005).

[27] Zhu, M. and Martinez, A.: Subclass discriminant analysis, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.28, No.8, pp.1274–1286 (2006).

[28] Gkalelis, N. Mezaris, V. and Kompatsiaris, I.: Mixture subclass discriminant analysis, *IEEE Signal Processing Letters*, Vol.18, No.5, pp.319–322 (2011).

[29] Schölkopf, B., Smola, A. and Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, Vol.10, No.5, pp.1299–1319 (1998).

[30] Mika, S., Ratsch, G., Weston, J., Scholkopf, B. and Mullers, K.R.: Fisher discriminant analysis with kernels, *Proc. IEEE Workshop on Neural Networks for Signal Processing* (*NNSP*), No.8, pp.41–48 (1999).

[31] Baudat, G. and Anouar, F.: Generalized discriminant analysis using a kernel approach, *Neural Computation*, Vol.12, No.10, pp.2385–2404 (2000).

[32] Dai, G., yan Yeung, D. and Chang, H.: Extending kernel Fisher discriminant analysis with the weighted pairwise Chernoff criterion, *Proc. European Conference on Computer Vision* (*ECCV*), pp.308–320 (2006).

[33] Bishop, C.: Linear models for regression, *Pattern Recognition and Machine Learning*, ch. 3, Springer (2006).

[34] Akaho, S.: A kernel method for canonical correlation analysis, *Proc. International Meeting of the Psychometric Society* (*IMPS*), pp.1–5, Springer-Verlag (2001).

[35] Lai, P.L. and Fyfe, C.: Kernel and nonlinear canonical correlation analysis, *Proc. International Joint Conference on Neural Networks* (*IJCNN*), p.4614 (2000).

[36] Hoerl, A.E.: Application of ridge analysis to regression problem, *Chemical Engineering Progress*, Vol.58, pp.54–59 (1962).

[37] Tikhonov, A.: On the stability of inverse problems, *Dokl. Akad. Nauk SSSR*, Vol.39, No.5, pp.195–198 (1943).

[38] Hardoon, D.R., Szedmak, S. and Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods, *Neural Computation*, Vol.16, No.12, pp.2639–2664 (2004).

[39] Blaschko, M., Lampert, C. and Gretton, A.: Semi-supervised Laplacian regularization of kernel canonical correlation analysis, *Proc. European Conference on Machine Learning and Knowledge Discovery in Databases* (*ECML-PKDD*), pp.133–145 (2008).

[40] Cox, T. and Cox, M.: *Multidimensional Scaling, Monographs on Statistics and Applied Probability*, No.1, Chapman & Hall (1994).

[41] He, X. and Niyogi, P.: Locality preserving projections, *Advances in Neural Information Processing Systems* (*NIPS*), pp.1–8 (2003).

[42] Sugiyama, M.: Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis, *Journal of Machine Learning Research*, Vol.8, No.5, pp.1027–1061 (2007).

[43] Sugiyama, M., Idé, T., Nakajima, S. and Sese, J.: Semi-supervised local Fisher discriminant analysis for dimensionality reduction, *Machine Learning*, Vol.78, No.1–2, pp.35–61 (2010).

[44] Kimura, A., Kameoka, H., Sugiyama, M., Nakano, T., Maeda, E., Sakano, H. and Ishiguro, K.: SemiCCA: Efficient semi-supervised learning of canonical correlations, *Proc. IAPR International Conference on Pattern Recognition* (*ICPR*), pp.2933–2936 (2010).

[45] Borga, M., Landelius, T. and Knutsson, H.: A unified approach to PCA, PLS, MLR and CCA, Report LiTH-ISY-R-1992, ISY, SE-581 83 Linköping, Sweden (Nov. 1997).

[46] De Bie, T., Cristianini, N. and Rosipal, R.: Eigenproblems in pattern recognition, *Handbook of Geometric Computing: Applications in Pattern Recognition, Computer Vision, Neuralcomputing, and Robotics*, pp.129–170, Springer (2005).

[47] Sun, L., Ji, S. and Ye, J.: A least squares formulation for a class of generalized eigenvalue problems in machine learning, *Proc. International Conference on Machine Learning* (*ICML*), pp.977–984 (2009).

[48] Sun, L., Ji, S. and Ye, J.: Canonical correlation analysis for multil-

abel classification: A least-squares formulation, extensions, and analysis, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.33, pp.194–200 (2011).
[49] De la Torre, F.: A unification of component analysis methods, *Handbook of Pattern Recognition and Computer Vision*, 4th ed., Chen, C. (Ed.), ch. 1, pp.3–22, World Scientific Pub Co Inc (2010).
[50] De la Torre, F.: A least-squares framework for component analysis, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.34, No.6, pp.1041–1055 (2012).
[51] Belkin, M. and Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Computation*, Vol.15, pp.1373–1396 (2002).
[52] Yanai, H. and Puntanen, S.: Partial canonical correlation associated with the inverse and some generalized inverse of a partitioned dispersion matrix, *Proc. Pacific Area Statistical Conference on Statistical Sciences and Data Analysis*, pp.253–264 (1993).
[53] Bach, F.R. and Jordan, M.I.: A probabilistic interpretation of canonical correlation analysis, Tech. Rep. 688, Department of Statistics, University of California, Berkeley (2005).
[54] Jolliffe, I.T.: A note on the use of principal components in regression, *Journal of the Royal Statistical Society*, Vol.31, No.3 (1982).
[55] Burnham, A.J., MacGregor, J.F. and Viveros, R.: Latent variable multivariate regression modeling, *Chemometrics and Intelligent Laboratory Systems*, pp.167–180 (Aug. 1999).
[56] Worsley, K.J., Poline, J.b., Friston, K.J. and Evans, A.C.: Characterizing the response of PET and fMRI data using multivariate linear models, *NeuroImage*, Vol.6, pp.305–319 (1997).
[57] Trygg, J. and Wold, S.: Orthogonal projections to latent structures (o-pls), *Journal of Chemometrics*, Vol.16, No.3, pp.119–128 (2002).
[58] Wang, H., Gottfries, J., Barrenäs, F. and Benson, M.: Identification of novel biomarkers in seasonal allergic rhinitis by combining proteomic, multivariate and pathway analysis, *PLoS ONE*, Vol.6, No.8, p.e23563 (2011).
[59] Bach, F.: Kernel independent component analysis, *Journal of Machine Learning Research*, Vol.3, pp.1–48 (2002).
[60] Zelnik-manor L. and Perona, P.: Self-tuning spectral clustering, *Advances in Neural Information Processing Systems* (*NIPS*), pp.1601–1608 (2004).
[61] Belkin, M. Niyogi, P. and Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples, *Journal of Machine Learning Research*, Vol.7, pp.2399–2434 (Dec. 2006).
[62] Roweis, S.T. and Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding, *Science*, Vol.290, pp.2323–2326 (Dec. 2000).
[63] Weiss, Y.: Segmentation using eigenvectors: A unifying view, *Proc. IEEE International Conference on Computer Vision* (*ICCV*), pp.975–982 (1999).
[64] Yu, S. and Shi, J.: Multiclass spectral clustering, *Proc. IEEE International Conference on Computer Vision* (*ICCV*), No.10, pp.313–319 (2003).
[65] Shi, J. and Malik, J.: Normalized cuts and image segmentation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.22, No.8, pp.888–905 (2000).
[66] Dhillon, I.S., Guan, Y. and Kulis, B.: Kernel k-means: spectral clustering and normalized cuts, *Proc. ACM International Conference on Knowledge Discovery and Data Mining* (*KDD*), pp.551–556, ACM (2004).
[67] Zass, R. and Shashua, A.: A unifying approach to hard and probabilistic clustering, *Proc. IEEE International Conference on Computer Vision* (*ICCV*), pp.294–301 (2005).
[68] Albert, A.: *Regression and the Moore-Penrose pseudoinverse, Mathematics in Science and Engineering*, Elsevier Science (1972).

Akisato Kimura received his B.E., M.E. and D.E. degrees in communications and integrated systems from Tokyo Institute of Technology, Japan in 1998, 2000 and 2007, respectively. Since 2000, he has been with NTT Communication Science Laboratories, NTT Corporation, where he is currently a senior research scientist in Innovative Communication Laboratory. He has been engaged in work on multimedia content identification, automatic multimedia annotation, human visual attention modeling and social media mining. His research interests include pattern recognition, computer vision, image processing, human visual perception, statistical signal processing, data mining and social media. He is a member of IEICE, IEEE and ACM SIGMM/SIGKDD.

Masashi Sugiyama received his B.E., M.E., and Ph.D. degrees from Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan, in 1997, 1999, and 2001, respectively. In 2001, he was appointed as a Research Associate in the same institute, and from 2003, he is an Associate Professor. His research interests include theory and application of machine learning.

Hitoshi Sakano received his B.S. degree in physics from Chuo University Tokyo and the M.S. degrees in physics from Saitama University, and the Ph.D. in applied physics from Waseda University, Tokyo in 1988, 1990 and 2008, respectively. He joined NTT Communication Science Laboratories in 2008 and studied pattern recognition technology. He is a member of IEEE, IEICE and Physical Society of Japan (PSJ).

Hirokazu Kameoka received his B.E., M.E. and Ph.D. degrees all from the University of Tokyo, Japan, in 2002, 2004 and 2007, respectively. He is currently a research scientist at NTT Communication Science Laboratories and an Adjunct Associate Professor at the University of Tokyo. His research interests include computational auditory scene analysis, statistical signal processing, speech and music processing, and machine learning. He is a member of IEEE, IEICE, IPSJ and ASJ. He received 13 awards over the past 9 years, including the IEEE Signal Processing Society 2008 SPS Young Author Best Paper Award.