Designing various multivariate analysis at will via generalized pairwise expression

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

Abstract: 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). 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 images [1], [2], [3], [4] and videos [5], [6] nowadays. However, we are now facing a difficulty in finding an intrinsic trend and nature of such a massive collection of data. Multivariate analysis is traditional, quite simple but might be one of the powerful tools to obtain a hidden structure embedded in the data. Actually, multivariate analysis has been still an important tool, and recent reports showed its effectiveness for several tasks, e.g. human detection [7], image annotation [8], [9], sensor data mining [10], [11], [12].

Principal component analysis (PCA) [13], Fisher discriminant analysis (FDA) [14], multivariate linear regression (MLR), canonical correlation analysis (CCA) [13], and partial least squares (PLS) [15] 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 so-called small sample size problem, i.e., the situation where the number of training samples is small compared with their dimensionality [16], [17], [18], [19], [20], [21]). Kernel multivariate analysis methods as kernelized extensions of those standard methods have been also developed to deal with non-vector samples and non-linear analysis [22], [23], [24], [25], [26], [27]). 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 [28], [29], [30] and graph Laplacian method [31]. In addition, improvements of robustness against outliers and non-Gaussiaity (i.e. multi-dimensional scaling (MDS) [32], locality preserving projection (LPP) [33] and local Fisher discriminant analysis (LFDA) [34]) and their extensions to semi-supervised dimensionality reduction [31], [35], [36] have been considered.

A lot of multivariate analysis methods and several trials to unify these methods have been presented so far. Borge et al [37] and De Bie et al [38] presented that several major linear multivariate analysis method can be formulated by a unified form of generalized eigenvalue problems. Sun et al [39], [40] showed the equivalence between a certain class of generalized eigenvalue problems and least squares ones under a mild assumption. De la Torre [41], [42] extended the work by Sun et al 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 multivariate analysis for a specific purpose or domain still remains an open problem. Until now, researchers have had to choose one of the existing methods that seems best to address the problem of interest, or

[1] NTT Communication Science Laboratories, NTT Corporation, 2-4 Hikaridai, Seika, Soraku, Kyoto, 619–0237 Japan.
[2] Graduate School of Information Science and Engineering, Tokyo Institute of Technology, 2-12-1 Oookayama, Meguro, Tokyo, 152–8552 Japan.
[3] Graduate School of Information Science and Technologies, the University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo, 113–8656 Japan.
[a)] akisato@ieee.org

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 covariance matrices and Gram matrices, which we call *Generalized pairwise expression (GPE)* to make it easy to design a new multivariate analysis method with desired property. The methodology is quite simple: adopting the above mentioned special cases as templates, and generating a new method by combining these templates appropriately. This characteristics has not been discussed yet in any previous researches to our best knowledge. It is also possible to individually select and arrange samples for calculating the scatter matrices of the methods to be combined, which enables us to extend CA methods to semi-supervised ones and multi-modal ones.

Our contributions can be summarized as follows:

(1) Providing a unified formulation of various multivariate analysis methods via GPE.
(2) Making it easy to implement a multivariate analysis method with desired property by simply combining the GPEs of existing methods.
(3) Designing new multivariate analysis methods based on the methodology derived from the GPE.

2. Multivariate analysis for vector data

Consider two sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ of samples[*1], 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:

$$\begin{aligned}\boldsymbol{X} &= \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_x}\},\\ \boldsymbol{Y} &= \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N, \boldsymbol{y}_{N_x+1}, \ldots, \boldsymbol{y}_{N_x+N_y-N}\}.\end{aligned}$$

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 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.

[*1] The following discussion can be easily extended to more than 2 sets of samples sets [43].

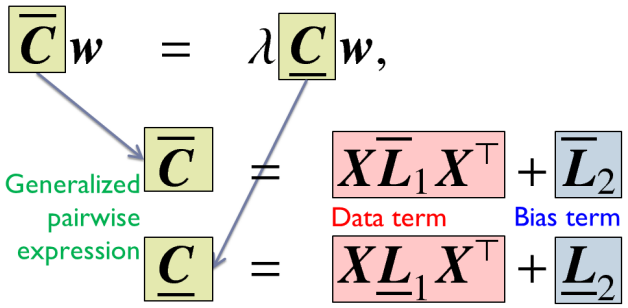

Fig. 1 Various multivariate analysis methods can be described via generalized pairwise expression (GPE)

Many linear multivariate analysis methods developed so far involve an optimization problem of the following form:

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

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:

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

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

$$\overline{\boldsymbol{C}}\boldsymbol{w} = \lambda \underline{\boldsymbol{C}}\boldsymbol{w}. \qquad (3)$$

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 Equation (1).

3. Generalized pairwise expression

When addressing linear multivariate analysis methods, we often deal with the following type of second-order statistics 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

$$\boldsymbol{S}_{\boldsymbol{Q},xy} = \sum_{n=1}^{N}\sum_{m=1}^{N} Q_{n,m}(\boldsymbol{x}_n - \boldsymbol{x}_m)(\boldsymbol{y}_n - \boldsymbol{y}_m)^\top, \qquad (4)$$

where $\boldsymbol{Q}$ is an $N \times N$ non-negative, semi-definite and symmetric matrix. A typical example is the scatter matrix[*2]:

[*2] Due to the limited space, we describe only the scatter matrix $\boldsymbol{S}_{xy}$ and its extensions with the pairwise form. The scatter matrices $\boldsymbol{S}_{xx}$ and $\boldsymbol{S}_{yy}$, and their extensions can be easily derived in the same way.

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

Let $\boldsymbol{D_Q}$ be the $N \times N$ diagonal matrix with

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

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

$$\boldsymbol{S}_{\boldsymbol{Q},xy} = \boldsymbol{X}\boldsymbol{L_Q}\boldsymbol{Y}^\top.$$

The above expression is called the *pairwise expression (PE)* of the second-order statistics $\boldsymbol{S}_{\boldsymbol{Q},xx}$[35]. If $\boldsymbol{Q}$ is a weight matrix for a graph with n nodes, $\boldsymbol{L_Q}$ 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, $\boldsymbol{L_Q}$ is known to be positive semi-definite.

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

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

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

$$\hat{\boldsymbol{S}}_{xy} = \boldsymbol{X}\boldsymbol{L}_1\boldsymbol{Y}^\top + \boldsymbol{L}_2. \tag{5}$$

After all, we call this expression as the *generalized pairwise expression (GPE)*. The first term of Equation (5) is called the *data term* since it depends on the sample data, and the second term is called the *bias term*.

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 with d_x rows and d_y columns, then $\boldsymbol{A}+\boldsymbol{B}$ is also GPE with d_x rows and d_y columns.
(3) If $\boldsymbol{A}$ is GPE with d_x rows and d_y columns, and $\boldsymbol{B}$ is GPE with d_y rows and d_z columns, then $\boldsymbol{AB}$ is also GPE with d_x rows and d_z columns.

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:

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

where $\boldsymbol{L}_{A1}$ (resp. $\boldsymbol{L}_{B1}$) is a positive semi-definite matrix with d_x (resp. d_y) rows and d_y (resp. d_z) columns, and $\boldsymbol{L}_{A2}$ (resp. $\boldsymbol{L}_{B2}$) is a $d_x \times d_y$ (resp. $d_y \times d_z$) non-negative matrix. Then, we obtain

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

Table 1 GPEs of standard methods

Method	$\overline{\boldsymbol{C}}$	$\underline{\boldsymbol{C}}$
PCA	$\boldsymbol{S}_{xx}$	$\boldsymbol{I}_{d_X}$
FDA	$\boldsymbol{S}_{xx}^{(b)}$	$\boldsymbol{S}_{xx}^{(w)}$
CCA	$\begin{bmatrix}\boldsymbol{0} & \boldsymbol{S}_{xy}\\ \boldsymbol{S}_{yx} & \boldsymbol{0}\end{bmatrix}$	$\begin{bmatrix}\boldsymbol{S}_{xx} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{S}_{yy}\end{bmatrix}$
MLR	$\begin{bmatrix}\boldsymbol{0} & \boldsymbol{S}_{xy}\\ \boldsymbol{S}_{yx} & \boldsymbol{0}\end{bmatrix}$	$\begin{bmatrix}\boldsymbol{S}_{xx} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_{d_y}\end{bmatrix}$
PCR[44]	$\begin{bmatrix}\boldsymbol{0} & \boldsymbol{S}_{\hat{x}y}\\ \boldsymbol{S}_{y\hat{x}} & \boldsymbol{0}\end{bmatrix}$	$\begin{bmatrix}\boldsymbol{S}_{\hat{x}\hat{x}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_{d_y}\end{bmatrix}$
OPLS	$\boldsymbol{S}_{xy}\boldsymbol{S}_{xy}^\top$	$\boldsymbol{S}_{xx}$
Ridge regression	$\begin{bmatrix}\boldsymbol{0} & \boldsymbol{S}_{xy}\\ \boldsymbol{S}_{yx} & \boldsymbol{0}\end{bmatrix}$	$\begin{bmatrix}\boldsymbol{S}_{xx}+\delta\boldsymbol{I}_{d_x} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_{d_y}\end{bmatrix}$
LPP[33]	$\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^\top$	$\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^\top$
LFDA[45]	$\boldsymbol{S}_{\boldsymbol{Q},xx}^{(lb)}$	$\boldsymbol{S}_{\boldsymbol{Q},xx}^{(lw)}$

PCR: Principal component regression, OPLS: Orthogonal partial least-squares.
$\boldsymbol{S}_{xx}^{(b)}$ and $\boldsymbol{S}_{xx}^{(w)}$: Between-class and within-class scatter matrices of $\boldsymbol{X}$, $\hat{\boldsymbol{X}} = \boldsymbol{U}_K\Sigma_K\boldsymbol{V}_K^\top$: K-rank approximation of $\boldsymbol{X}$ by SVD, $\boldsymbol{I}_d$: $d \times d$ identity matrix, $\delta > 0$: constant, $\boldsymbol{S}_{\boldsymbol{Q}}^{(b)}$ and $\boldsymbol{S}_{\boldsymbol{Q}}^{(w)}$: between-class and within-class scatter matrices of $\boldsymbol{X}$ weighted by an $N \times N$ non-negative symmetric matrix.

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

$$\begin{aligned}\boldsymbol{L}_{C1} &= \boldsymbol{L}_{A1}\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{L}_{B1},\\ \boldsymbol{X}\boldsymbol{L}_{C2} &= \boldsymbol{L}_{A2}\boldsymbol{Y},\\ \boldsymbol{L}_{C3}\boldsymbol{Z}^\top &= \boldsymbol{Y}^\top\boldsymbol{L}_{B2}.\end{aligned}$$

This implies that

$$\begin{aligned}&\boldsymbol{AB}\\ &= \boldsymbol{X}\boldsymbol{L}_{C1}\boldsymbol{Z}^\top + \boldsymbol{X}\boldsymbol{L}_{C2}\boldsymbol{L}_{B1}\boldsymbol{Z}^\top\\ &\qquad + \boldsymbol{X}\boldsymbol{L}_{A1}\boldsymbol{L}_{C3}\boldsymbol{Z}^\top + \boldsymbol{L}_{A2}\boldsymbol{L}_{B2}\\ &= \boldsymbol{X}(\boldsymbol{L}_{C1} + \boldsymbol{L}_{C2}\boldsymbol{L}_{B1} + \boldsymbol{L}_{A1}\boldsymbol{L}_{C3})\boldsymbol{Z}^\top + \boldsymbol{L}_{A2}\boldsymbol{L}_{B2}\\ &= \boldsymbol{X}\boldsymbol{L}_{D1}\boldsymbol{Z}^\top + \boldsymbol{L}_{D2},\end{aligned}$$

for some matrices $\boldsymbol{L}_{D1}$ and $\boldsymbol{L}_{D2}$, which means $\boldsymbol{AB}$ is also GPE. □

Recall that the class of multivariate analysis we are dealing with can be expressed as $\overline{\boldsymbol{C}}\boldsymbol{w} = \lambda\underline{\boldsymbol{C}}\boldsymbol{w}$, and both $\overline{\boldsymbol{C}}$ and $\underline{\boldsymbol{C}}$ can be expressed by GPEs or their augmented matrices. The notable point is that various multivariate analysis methods can be easily designed with the help of these GPE properties, namely by combining GPEs of existing methods with desired properties. 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

The GPEs of the standard CA methods are listed in Table 1. Several detailed derivations can be seen in [37], [38]. Instead, this paper provides several significant examples that would be quite an important hint when constructing new CA methods.

4.2 Locality preserving projection (LPP)

Locality preserving projections (LPP) [33] 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 [46], i.e.

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

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 [46]. The objective function to be minimized is the following weighted squared error:

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

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

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

4.3 Local Fisher discriminant analysis (LFDA)

Local Fisher discriminant analysis (LFDA) [34] 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 $\boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{lw})}$.

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

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

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

where 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:

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

4.4 Semi-supervised LFDA (SELF)

Semi-supervised local fisher discriminant analysis, called SELF [35], integrates LFDA as a supervised dimensionality reduction and PCA as a 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,
(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 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:

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

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.

4.5 Semi-supervised CCA

In a similar way to that of SELF, a semi-supervised extension of CCA can be derived, which is called SemiCCA [36].

Assume that there are two samples sets $\boldsymbol{X}$ and $\boldsymbol{Y}$, and each includes incomplete sample set $\boldsymbol{X}^{(\mathrm{I})}$ and $\boldsymbol{Y}^{(\mathrm{I})}$ only exists, 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 $(\boldsymbol{X}^{(\mathrm{C})}, \boldsymbol{Y}^{(\mathrm{C})})$.

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 equation:

$$\overline{\boldsymbol{C}}_{\boldsymbol{Q}}^{(\mathrm{CFDA})} = \beta \begin{pmatrix} \boldsymbol{0} & \boldsymbol{S}_{xy} \\ \boldsymbol{S}_{yx} & \boldsymbol{0} \end{pmatrix} + (1-\beta)\boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{lb})}, \tag{6}$$

$$\underline{\boldsymbol{C}}_{\boldsymbol{Q}}^{(\mathrm{CFDA})} = \beta \begin{pmatrix} \boldsymbol{S}_{xx} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_{yy} \end{pmatrix} + (1-\beta)\boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{lw})}. \tag{7}$$

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

6. Kernelized extensions

6.1 Kernelization of standard methods

A lot of methods in the GPE framework can be kernelized in a similar manner to the existing ones. The GPEs of major kernelized CA methods are listed in Table 1. By introducing kernelized expression, several methods for clustering and local embedding can be included in this framework, e.g.

Table 2 GPEs of kernelized CA methods

Method	$\overline{\boldsymbol{C}}$	$\underline{\boldsymbol{C}}$
kPCA	$\boldsymbol{K}_x$	$\boldsymbol{I}_N$
kFDA	$\boldsymbol{K}_x^{(b)}$	$\boldsymbol{K}_x^{(w)}$
kCCA	$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{K}_x\boldsymbol{K}_y \\ \boldsymbol{K}_y\boldsymbol{K}_x & \boldsymbol{0} \end{bmatrix}$	$\begin{bmatrix} \boldsymbol{K}_x^2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{K}_y^2 \end{bmatrix}$
kMLR	$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{K}_x\boldsymbol{K}_y \\ \boldsymbol{K}_y\boldsymbol{K}_x & \boldsymbol{0} \end{bmatrix}$	$\begin{bmatrix} \boldsymbol{K}_x^2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_N \end{bmatrix}$
kCCA+ℓ_2[30]	$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{K}_x\boldsymbol{K}_y \\ \boldsymbol{K}_y\boldsymbol{K}_x & \boldsymbol{0} \end{bmatrix}$	$\begin{bmatrix} \boldsymbol{K}_x^{(\ell_2)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{K}_y^{(\ell_2)} \end{bmatrix}$
L-kCCA[31]	$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{K}_x\boldsymbol{K}_y \\ \boldsymbol{K}_y\boldsymbol{K}_x & \boldsymbol{0} \end{bmatrix}$	$\begin{bmatrix} \boldsymbol{K}_x^{(L)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{K}_x^{(L)} \end{bmatrix}$
LE, SC	$\boldsymbol{L}_x$	$\boldsymbol{D}_x$
LLE	$\boldsymbol{K}_x^{(LL)}\boldsymbol{K}_x^{(LL)\top}$	$\boldsymbol{I}_N$
NC, nSC	$\boldsymbol{D}_x^{-1/2}\boldsymbol{L}_x\boldsymbol{D}_x^{-1/2}$	$\boldsymbol{I}_N$

L-kCCA: Laplacian-regularized kernel CCA, $\boldsymbol{K}_x$, $\boldsymbol{K}_x$: Gram matrices, $\boldsymbol{K}_x^{(\ell_2)} = \boldsymbol{K}_x^2 + \delta_x\boldsymbol{K}_x$, $\boldsymbol{K}_x^{(L)} = \boldsymbol{K}_x^2 + \gamma_x\boldsymbol{R}_x$, $\boldsymbol{R}_x = \boldsymbol{K}_x\boldsymbol{L}_x\boldsymbol{K}_x$m, LE: Laplacian eigenmap, SC: Spectral clustering, NC: Normalized cuts, nSC: normalized SC, $\boldsymbol{K}_x^{(LL)} = \boldsymbol{I}_N - \boldsymbol{K}_x$

Laplacian eigenmap (LE), locally linear embedding (LLE), spectral clustering (SC) and normalized cuts (NC).

6.2 How to design new kernelized methods

Integrating two methods within the kernelized GPE framework is not obvious, since a simple addition of Gram matrices is not GPE. One example can be seen in a kernelized extension of SELF, called kernel SELF [35]. Remember that the original SELF integrates LFDA with labeled samples and PCA with all the samples (see Section 4.4), and it can be formulated by a localized between-class scatter matrix $\boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{C,lb})}$, localized within-class matrix $\boldsymbol{S}_{\boldsymbol{Q}}^{(\mathrm{C,lw})}$ and the ordinary scatter matrix $\boldsymbol{S}_{xx}$. Kernel SELF can be formulated via their Laplacian matrices $\boldsymbol{L}_{\boldsymbol{Q}}^{(\mathrm{C,lb})}$, $\boldsymbol{L}_{\boldsymbol{Q}}^{(\mathrm{C,lw})}$, $\boldsymbol{L}_{xx}$, as follows:

$$\overline{\boldsymbol{C}}^{(\mathrm{kSELF})} = \boldsymbol{K}_x\{\beta\boldsymbol{L}_{\boldsymbol{Q}}^{(\mathrm{C,lb})} + (1-\beta)\boldsymbol{L}_{xx}\}\boldsymbol{K}_x,$$

$$\underline{\boldsymbol{C}}^{(\mathrm{kSELF})} = \beta\boldsymbol{K}_x\boldsymbol{L}_{\boldsymbol{Q}}^{(\mathrm{C,lb})}\boldsymbol{K}_x + (1-\beta)\boldsymbol{K}_x.$$

From this formulation, we can see that a weighted sum of GPEs in original multivariate analysis corresponds to a weighted sum of Laplacian matrices in kernelized multivariate analysis. Namely, when dealing with kernelized multivariate analysis, we have to explicitly derive GPEs of existing methods, and replace the data matrix into its Gram matrix.

7. Concluding remarks

This paper provided a new theoretical 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 significant future work.

References

[1] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman, "Labelme: A database and web-based tool for image annotation," IJCV, vol.77, pp.157–173, 2008.
[2] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," Proc. CVPR, 2009.
[3] A. Torralba, R. Fergus, and W.T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," IEEE Trans. PAMI, vol.30, no.11, pp.1958–1970, 2008.
[4] X.J. Wang, L. Zhang, M. Liu, Y. Li, and W.Y. Ma, "ARISTA - image search to annotation on billions of web photos," Proc. CVPR, pp.2987–2994, 2010.
[5] A.F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," Proc, MIR, pp.321–330, 2006.
[6] J. Yuen, B.C. Russell, C. Liu, and A. Torralba, "Labelme video: Building a video database with human annotations," Proc. ICCV, pp.1451–1458, 2009.
[7] W.R. Schwartz, A. Kembhavi, D. Harwood, and L.S. Davis, "Human detection using partial least squares analysis," Proc. ICCV, 2009.
[8] H. Nakayama, T. Harada, and Y. Kuniyoshi, "Evaluation of dimensionality reduction methods for image auto-annotation," Proc. BMVC, pp.1–12, 2010.
[9] M.B. Blaschko and C.H. Lampert, "Correlational spectral clustering," Proc. CVPR, pp.1–8, 2008.
[10] A. Pezeshki, M.R. Azimi-Sadjadi, and L.L. Scharf, "Undersea target classification using canonical correlation analysis," IEEE Journal of Oceanic Engineering, vol.32, no.4, pp.948–955, 2007.
[11] A.A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multi-temporal remote sensing data," IEEE Trans IP, vol.11, no.3, pp.293–305, 2002.
[12] I. Schizas, G. Giannakis, and Z.Q. Luo, "Distributed estimation using reduced-dimensionality sensor observations," IEEE Trans. SP, vol.55, no.8, pp.4284 –4299, aug. 2007.
[13] H. Hotelling, "Analysis of a complex of statistical variables into principal components," Journal of Educational Psychology, vol.24, 1933.
[14] R.A. Fisher, "The use of multiple measurements in taxonomic problems," Annals Eugen., vol.7, pp.179–188, 1936.
[15] H. Wold, Estimation of Principal Components and Related Models by Iterative Least squares, pp.391–420, Academic Press, 1966.
[16] F. De la Torre and M. Black, "Robust principal component analysis for computer vision," Proc. ICCV, pp.362–369, 2001.
[17] D. Fernando and M.J. Black, "A framework for robust subspace learning," IJCV, vol.54, p.2003, 2003.
[18] K. Inoue, K. Hara, and K. Urahama, "Robust multilinear principal component analysis," Proc. ICCV, pp.591–597, 2009.
[19] J. Lu, K. Plataniotis, and A. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," Pattern Recognition Letters, vol.26, no.2, pp.181 – 191, 2005.
[20] M. Zhu and A. Martinez, "Subclass discriminant analysis," IEEE Trans PAMI, vol.28, no.8, pp.1274–1286, 2006.
[21] N. Gkalelis, V. Mezaris, and I. Kompatsiaris, "Mixture subclass discriminant analysis," IEEE Signal Processing Letters, vol.18, no.5, pp.319–322, 2011.
[22] B. Schölkopf, A. Smola, and K.R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Computation, vol.10, no.5, pp.1299–1319, 1998.
[23] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.R. Mullers, "Fisher discriminant analysis with kernels," Proc. NNSP, pp.41–48, 1999.
[24] G. Dai, D. yan Yeung, and H. Chang, "Extending kernel Fisher discriminant analysis with the weighted pairwise Chernoff criterion," Proc. ECCV, pp.308–320, 2006.
[25] C. Bishop, "Linear models for regression," in Pattern Recognition and Machine Learning, ch. 3, Springer, 2006.
[26] S. Akaho, "A kernel method for canonical correlation analysis," Proc. IMPS2001, 2001.
[27] S. Harmeling, A. Ziehe, M. Kawanabe, and K.R. Müller, "Kernel-based nonlinear blind source separation," Neural Computation, vol.15, pp.1089–1124, 2003.
[28] A.E. Hoerl, "Application of ridge analysis to regression problem," Chemical Engineering Progress, vol.58, pp.54–59, 1962.
[29] A. Tikhonov, "On the stability of inverse problems," Dokl. Akad. Nauk SSSR, vol.39, no.5, pp.195–198, 1943.
[30] D.R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," Neural Computation, vol.16, no.12, pp.2639–2664, 2004.
[31] M. Blaschko, C. Lampert, and A. Gretton, "Semi-supervised Laplacian regularization of kernel canonical correlation analysis," Proc. ECML-PKDD, pp.133–145, Springer-Verlag, 2008.
[32] T. Cox and M. Cox, Multidimensional scaling, Monographs on statistics and applied probability, no.1, Chapman & Hall, 1994.
[33] X. He and P. Niyogi, "Locality preserving projections," Proc. NIPS, 2003.
[34] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," JMLR, vol.8, pp.1027–1061, 2007.
[35] M. Sugiyama, T. Idé, S. Nakajima, and J. Sese, "Semi-supervised local Fisher discriminant analysis for dimensionality reduction," Machine Learning, vol.78, no.1–2, pp.35–61, 2010.
[36] A. Kimura, H. Kameoka, M. Sugiyama, T. Nakano, E. Maeda, H. Sakano, and K. Ishiguro, "SemiCCA: Efficient semi-supervisedl learning of canonical correlations," Proc. ICPR, pp.2933–2936, 2010.
[37] M. Borga, T. Landelius, and H. Knutsson, "A unified approach to PCA, PLS, MLR and CCA," Report LiTH-ISY-R-1992, ISY, SE-581 83 Linköping, Sweden, November 1997.
[38] T. De Bie, N. Cristianini, and R. Rosipal, "Eigenproblems in pattern recognition," in Handbook of Geometric Computing: Applications in Pattern Recognition, Computer Vision, Neuralcomputing, and Robotics, pp.129–170, Springer, 2005.
[39] L. Sun, S. Ji, and J. Ye, "A least squares formulation for a class of generalized eigenvalue problems in machine learning," Proc. ICML, pp.977–984, 2009.
[40] L. Sun, S. Ji, and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," IEEE Trans PAMI, vol.33, pp.194–200, 2011.
[41] F. De la Torre, "A unification of component analysis methods," in Handbook of Pattern Recognition and Computer Vision, ed. C. Chen, ch. 1, pp.3–22, World Scientific Pub Co Inc, 4 ed., 2010.
[42] F. De la Torre, "A least-squares framework for component analysis," IEEE Trans PAMI, vol.34, no.6, pp.1041–1055, 2012.
[43] H. Yanai and S. Puntanen, "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.
[44] I.T. Jolliffe, "A note on the use of principal components in regression," Journal of the Royal Statistical Society, vol.31, no.3, 1982.
[45] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," JMLR, vol.8, pp.1027–1061, 2007.
[46] L. Zelnik-manor and P. Perona, "Self-tuning spectral clustering," Proc. NIPS, pp.1601–1608, 2004.