Complex Moment-Based Supervised Eigenmap for Dimensionality Reduction

Akira Imakura, Momo Matsuda, Xiucai Ye, Tetsuya Sakurai
University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

Abstract

Dimensionality reduction methods that project high-dimensional data to a low-dimensional space by matrix trace optimization are widely used for clustering and classification. The matrix trace optimization problem leads to an eigenvalue problem for a low-dimensional subspace construction, preserving certain properties of the original data. However, most of the existing methods use only a few eigenvectors to construct the low-dimensional space, which may lead to a loss of useful information for achieving successful classification. Herein, to overcome the deficiency of the information loss, we propose a novel complex moment-based supervised eigenmap including multiple eigenvectors for dimensionality reduction. Furthermore, the proposed method provides a general formulation for matrix trace optimization methods to incorporate with ridge regression, which models the linear dependency between covariate variables and univariate labels. To reduce the computational complexity, we also propose an efficient and parallel implementation of the proposed method. Numerical experiments indicate that the proposed method is competitive compared with the existing dimensionality reduction methods for the recognition performance. Additionally, the proposed method exhibits high parallel efficiency.

Introduction

Dimensionality reduction is an efficient technique for data analysis which maps high-dimensional data to a low dimensional space. Dimensionality reduction methods that utilize matrix trace optimization are widely used for clustering and classification.

The matrix trace optimization problem leads to an eigenvalue problem for low-dimensional space construction, preserving certain properties of the original data. Principal component analysis (PCA) (Pearson 1901; Jolliffe 1986) and locality preserving projections (LPP) (He and Niyogi 2004) are two of the typical unsupervised dimensionality reduction methods. PCA aims to maximize the variance of the projected vectors, while LPP devotes to preserve the local similarity of the original data. Discriminant analysis is the typical supervised method which maximizes the between-class scatter and reduce the within-class scatter. A family of discriminant analysis methods are proposed for dimensionality reduction, including Fisher discriminant analysis (FDA) (Fisher 1936; Fukunaga 2013), local FDA (LFDA) (Sugiyama 2007), semi-supervised LFDA (SELF) (Sugiyama et al. 2010) and locality adaptive discriminant analysis (LADA) (Li et al. 2017).

However, most of the existing dimensionality reduction methods use only ℓ eigenvectors to construct the low-dimensional space with ℓ dimensions, which may lead to a loss of useful information for achieving successful classification. Herein, to overcome the deficiency of information loss, we propose a novel complex moment-based supervised eigenmap for dimensionality reduction. The proposed method allows us to achieve better recognition performance by using a complex moment-based subspace that includes $d > \ell$ eigenvectors, where d can be set independently of ℓ.

The proposed method is inspired by complex moment-based parallel eigensolvers (Sakurai and Sugiura 2003; Imakura, Du, and Sakurai 2016) that are one of the hottest parallel eigensolvers for computing multiple eigenpairs of large and sparse eigenvalue problems. Since a subspace including a large number of eigenvectors is expected to have rich information, the usage of complex moment-based subspace can help to achieve a good recognition performance.

Furthermore, we incorporate ridge regression (or called as the Tikhonov regularization) (Saunders, Gammerman, and Vovk 1998) to matrix trace optimization in the objective function of the proposed method. Ridge regression aims to find a linear function that models the dependencies between covariate variables and univariate labels. By the corporation, the proposed method can find the subspace that most compactly expresses the target and rejects other possible but less compact candidates. As far as our knowledge, no existing method combines a matrix trace and ridge regression.

The proposed method is flexible and extendable since there are a lot of matrix trace optimization methods can be incorporated, such as the methods mentioned above. We use a specific parameter to control the trade-off between the matrix trace optimization and ridge regression. In extreme cases, the proposed method simplifies to the matrix trace optimization methods and the ridge regression method. To reduce the computational complexity, we also propose an efficient and parallel implementation of the proposed method based on some techniques of the complex moment-based

parallel eigensolvers (Sakurai and Sugiura 2003; Imakura, Du, and Sakurai 2016).

The main aspects of the proposed method to achieve high recognition performance are

- The usage of a complex moment-based subspace including multiple eigenvectors to preserve more data properties than the existing methods.
- Providing a general formulation for matrix trace optimization methods to incorporate with ridge regression, which models the linear dependency between covariate variables and univariate labels in addition to the existing methods.
- Proposing an efficient and parallel implementation based on techniques used for the complex moment-based parallel eigensolvers to reduce the computational complexity.

Numerical results are demonstrated to evaluate how much effect the aspects of the proposed method make it possible to have competitive advantages over existing dimensionality reduction methods.

Throughout the manuscript, the following notation is used. We define the range space of a matrix $V = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_L]$ by $\mathcal{R}(V) := \mathrm{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ \ldots, \boldsymbol{v}_L\}$. We also use MATLAB notations.

Related methods

Dimensionality reduction methods

Let m and n be the dimension of the features and the number of samples for training, and let $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{m\times n}$ be the training dataset. We consider linear and nonlinear dimensionality reduction methods that construct low-dimensional data $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n] \in \mathbb{R}^{\ell\times n}$, which retain some of the properties of the original data.

Linear dimensionality reduction methods reduce the original data X to the low-dimensional data Y using a linear map $B \in \mathbb{R}^{m\times\ell}$, i.e.

$$Y = B^{\mathrm{T}}X.$$

Let a symmetric matrix $A_1 \in \mathbb{R}^{m\times m}$ and a symmetric positive definite matrix $A_2 \in \mathbb{R}^{m\times m}$ be defined, respectively, in each dimensionality reduction method. Then, the linear map B is formulated by the minimization or maximization of a matrix trace:

$$\begin{aligned}&\min_{B\in\mathbb{R}^{m\times\ell}} \mathrm{Tr}\left(B^{\mathrm{T}}A_1B\right) \quad \text{or} \quad \max_{B\in\mathbb{R}^{m\times\ell}} \mathrm{Tr}\left(B^{\mathrm{T}}A_1B\right)\\ &\text{s.t.} \quad B^{\mathrm{T}}A_2B = I\end{aligned}$$

and is computed as ℓ eigenvectors of the corresponding generalized eigenvalue problem:

$$A_1\boldsymbol{t}_i = \lambda_i A_2\boldsymbol{t}_i. \tag{1}$$

Here, we have $B = [\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_\ell]$.

Nonlinear dimensionality reduction methods, which use a nonlinear map and the kernel trick (Schölkopf, Smola, and Müller 1998), are widely used as improvements over linear dimensionality reduction methods. Nonlinear dimensionality reduction methods transform the original data X to $\phi(X) = [\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2), \ldots, \phi(\boldsymbol{x}_n)]$ with a nonlinear kernel function, and reduce the dimension of $\phi(X)$ using a nonlinear map $\widetilde{B}$ such that $Y = \widetilde{B}^{\mathrm{T}}\phi(X)$. With appropriate nonlinear functions, nonlinear dimensionality reduction methods are expected to improve the recognition performance.

In general, we set $\widetilde{B} = \phi(X)\widehat{B}$ with $\widehat{B} \in \mathbb{R}^{n\times\ell}$ and directly set the Gram matrix $K = \phi(X)^{\mathrm{T}}\phi(X) \in \mathbb{R}^{n\times n}$ without computing $\phi(X)$ to reduce the computational costs. The Gaussian kernel, polynomial kernel and sigmoid kernel are commonly used as the kernel functions.

Ridge regression

Linear ridge regression is a classical statistical algorithm which computes a linear function $B \in \mathbb{R}^{m\times\ell}$ such that minimizes the squared error with the ground truth data $Z \in \mathbb{R}^{\ell\times n}$:

$$\min_{B\in\mathbb{R}^{m\times\ell}} \|Z - B^{\mathrm{T}}X\|_{\mathrm{F}}^2 + \lambda\|B\|_{\mathrm{F}}^2,$$

where λ is a regularization parameter. The kernel version of the ridge regression is also used for improving recognition performance.

Complex moment-based parallel eigensolvers

Complex moment-based eigensolvers were first proposed by (Sakurai and Sugiura 2003) for solving interior generalized eigenvalue problems of the form:

$$\begin{aligned}&A\boldsymbol{x}_i = \lambda_i B\boldsymbol{x}_i,\\ &A, B \in \mathbb{C}^{n\times n}, \quad \boldsymbol{x}_i \in \mathbb{C}^n \setminus \{\boldsymbol{0}\}, \quad \lambda_i \in \Omega \subset \mathbb{C},\end{aligned}$$

where $zB - A$ is non-singular in a boundary Γ of the target region Ω. This method is based on Cauchy's integral formula and constructs certain complex moment matrices using a contour integral.

Let $L, M \in \mathbb{N}_+$ be the input parameters and $V \in \mathbb{C}^{n\times L}$ be an input matrix. We define $S = [S_0, S_1, \ldots, S_{M-1}] \in \mathbb{C}^{n\times LM}$ and $S_k \in \mathbb{C}^{n\times L}$ as

$$S_k := \frac{1}{2\pi\mathrm{i}}\oint_\Gamma z^k(zB - A)^{-1}BV\,\mathrm{d}z. \tag{2}$$

Complex moment-based eigensolvers are mathematically designed on the basis of the properties of the matrices S_k and S. Then, practical algorithms are derived by approximating the contour integral (2) using the numerical integration rule:

$$\widehat{S}_k := \sum_{j=1}^{N} \omega_j z_j^k (z_jB - A)^{-1}BV,$$

where z_j is a quadrature point and ω_j is its corresponding weight.

The most time-consuming part of using complex moment-based eigensolvers involves solving linear systems at each quadrature point. However, as these linear systems can be independently solved, the complex moment-based eigensolvers have good scalability. For this reason, complex moment-based eigensolvers have attracted considerable attention. Currently, there are several methods including direct extensions of Sakurai and Sugiura's approach (Sakurai and Tadano 2007; Ikegami, Sakurai, and Nagashima

2010; Ikegami and Sakurai 2010; Imakura, Du, and Sakurai 2014; 2016; Imakura and Sakurai 2017; Imakura, Futamura, and Sakurai 2017), the FEAST eigensolver developed by (Polizzi 2009) and its improvements (Tang and Polizzi 2014; Güttel et al. 2015; Kestyn et al. 2016).

For details of these methods and the relationship among typical methods, refer to the study by (Imakura, Du, and Sakurai 2016) and the references therein.

A complex moment-based supervised eigenmap for dimensionality reduction

In this section, to achieve high recognition performance, we propose a novel complex moment-based supervised eigenmap for dimensionality reduction. The proposed method minimizes a novel objective function that combines a matrix trace and a squared error with the ground truth data $Z \in \mathbb{R}^{\ell \times n}$. The proposed method also uses a complex moment-based subspace from complex moment-based eigensolvers that includes multiple eigenvectors.

Basic concepts of the proposed method

Let A_1 and A_2 be the matrices used in a given dimensionality reduction method such as LPP or LFDA. We also let $\mathcal{S}_\Omega$ be a complex moment-based subspace with respect to a given real interval $\Omega = [a, b] \subset \mathbb{R}$ defined by

$$\mathcal{S}_\Omega = \mathcal{R}(S), \quad S = [S_0, S_1, \ldots, S_{M-1}],$$
$$S_k := \frac{1}{2\pi \mathrm{i}} \oint_\Gamma z^k (zA_2 - A_1)^{-1} A_2 V \mathrm{d}z, \tag{3}$$

where $L, M \in \mathbb{N}_+, V \in \mathbb{R}^{m \times L}$ and Γ is a positively oriented Jordan curve around Ω. Here, we assume that there are only the target eigenvalues in the Jordan curve Γ. Then, to obtain the linear map $B \in \mathbb{R}^{m \times \ell}$, we introduce the following minimization problem:

$$\min_{B=[\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_\ell], \boldsymbol{b}_i \in \mathcal{S}_\Omega} E(B) \quad \text{s.t.} \quad B^{\mathrm{T}} A_2 B = I,$$
$$E(B) = (1-\mu)\mathrm{Tr}\left(B^{\mathrm{T}} f(A_1) B\right) + \mu \|Z - B^{\mathrm{T}} X\|_{\mathrm{F}}^2, \tag{4}$$

whose objective function $E(B)$ combines a matrix trace derived from dimensionality reduction methods and a squared error straightforwardly using the ground truth data Z like the ridge regression.

The column vectors of the linear map B are constrained by A_2-orthonormal bases of the complex moment-based subspace $\mathcal{S}_\Omega$. Here, $\mu \in [0, 1]$ is a weight parameter for both terms and $f(\cdot)$ is a (meromorphic) weight function of each eigenvector for minimization.

In a trace minimization, solutions contain rich eigenvectors with a small weight. If $f(\lambda) = 1$, then each eigenvector is not scaled, and if $f(\lambda) = \lambda$, then each eigenvector is scaled by the corresponding eigenvalue. If A_1 and A_2 are from a trace minimization-type method like LPP, since eigenvectors associated with small eigenvalues have rich information, we set $f(\lambda)$ such that the eigenvectors associated with small eigenvalues in Ω have small weights. In contrast, if A_1 and A_2 are from a trace maximization-type method like LFDA, since eigenvectors associated with large eigenvalues have rich information, we set $f(\lambda)$ such that the eigenvectors associated with large eigenvalues in Ω have small weights. Note that, for a diagonalizable matrix $A = XDX^{-1}$ with $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$, we have $f(A) = Xf(D)X^{-1}$ with $f(D) = \mathrm{diag}(f(d_1), f(d_2), \ldots, f(d_n))$ (Higham 2008).

For the meaning of the complex moment-based subspace $\mathcal{S}_\Omega$, we have the following theorem; see e.g., (Imakura, Du, and Sakurai 2016).

Theorem 1. *The complex moment-based subspace is equivalent to an invariant subspace with respect to the multiple eigenvectors corresponding to the eigenvalues in a given real interval $\Omega = [a, b] \subset \mathbb{R}$, that is,*

$$\mathcal{S}_\Omega = \mathcal{T}_\Omega := \mathrm{span}\{\boldsymbol{t}_i | \lambda_i \in \Omega\},$$

if and only if $\mathrm{rank}(S) = d$.

The basic concepts of the proposed method for high recognition performance are summarized as follows:

- Use the complex moment-based subspace $\mathcal{S}_\Omega$, which is equivalent to the invariant subspace $\mathcal{T}_\Omega$ with respect to the multiple eigenvectors to solve the novel minimization problem (4).
- Use the novel minimization problem (4) that combines the matrix trace derived from the dimensionality reduction methods and the squared error straightforwardly using the ground truth data.

Derivation of a practical algorithm

Here, we propose a practical algorithm for the proposed method based on some techniques of complex moment-based eigensolvers. Let $U \in \mathbb{R}^{m \times d}$ and $U^\perp \in \mathbb{R}^{m \times (m-d)}$ be A_2-orthogonal matrices whose columns are A_2-orthonormal bases of the complex moment-based subspace $\mathcal{S}_\Omega$ and of its A_2-orthogonal complement, respectively, i.e.,

$$U = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d], \quad \boldsymbol{u}_k \in \mathcal{S}_\Omega,$$
$$U^{\mathrm{T}} A_2 U = I, \quad U^{\mathrm{T}} A_2 U^\perp = O, \tag{5}$$

where $d = |\{\lambda_i | \lambda_i \in \Omega\}|$ is the number of eigenvalues of (1) in $\Omega = [a, b]$. Then, the linear map B is given by

$$B = UC, \quad C^{\mathrm{T}} C = I, \quad C \in \mathbb{R}^{d \times \ell}.$$

Using (5) and the relation $\mathrm{Tr}(A^{\mathrm{T}} A) = \|A\|_{\mathrm{F}}^2$, we have

$$\begin{aligned}
&\mathrm{Tr}\left(B^{\mathrm{T}} f(A_1) B\right) \\
&= \mathrm{Tr}\left((C^{\mathrm{T}} U^{\mathrm{T}})(A_2[U, U^\perp]) f\left([U, U^\perp]^{\mathrm{T}} A_1 [U, U^\perp]\right)\right. \\
&\qquad \left. \cdot ([U, U^\perp]^{\mathrm{T}} A_2)(UC)\right) \\
&= \mathrm{Tr}\left(C^{\mathrm{T}} f(U^{\mathrm{T}} A_1 U) C\right) \\
&= \|f(T)^{1/2} C\|_{\mathrm{F}}^2,
\end{aligned}$$

where $T = U^{\mathrm{T}} A_1 U$. Also, we have

$$\|Z - B^{\mathrm{T}} X\|_{\mathrm{F}}^2 = \|Z - C^{\mathrm{T}} U^{\mathrm{T}} X\|_{\mathrm{F}}^2.$$

From the relation $\|A\|_{\mathrm{F}}^2 + \|B\|_{\mathrm{F}}^2 = \|[A; B]\|_{\mathrm{F}}^2$, the objective function $E(B)$ in (4) can be written as

$$E(B) = (1-\mu)\|f(T)^{1/2}C\|_{\mathrm{F}}^2 + \mu\|Z - C^{\mathrm{T}}U^{\mathrm{T}}X\|_{\mathrm{F}}^2 = \left\| \begin{bmatrix} \mu^{1/2}Z^{\mathrm{T}} \\ O \end{bmatrix} - \begin{bmatrix} \mu^{1/2}X^{\mathrm{T}}U \\ (1-\mu)^{1/2}f(T)^{1/2} \end{bmatrix} C \right\|_{\mathrm{F}}^2.$$

Therefore, the minimization problem (4) becomes

$$\min_{C \in \mathbb{R}^{d \times \ell}} \left\| \begin{bmatrix} \mu^{1/2}Z^{\mathrm{T}} \\ O \end{bmatrix} - \begin{bmatrix} \mu^{1/2}X^{\mathrm{T}}U \\ (1-\mu)^{1/2}f(T)^{1/2} \end{bmatrix} C \right\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad C^{\mathrm{T}}C = I. \tag{6}$$

A minimization problem with an orthogonal constraint (6) is called an unbalanced orthogonal Procrustes (UOP) problem, which is solved using an iterative method (Eldén and Park 1999; Park 1991).

In practice, the contour integral (3) is approximated by a numerical integration rule such as the N-point trapezoidal rule, as follows:

$$\widehat{S}_k := \sum_{j=1}^{N} \omega_j z_j^k (z_j A_2 - A_1)^{-1} A_2 V,$$

where $(z_j, \omega_j), j = 1, 2, \ldots, N$ are the quadrature points and the corresponding weights, respectively. Because of the symmetric property of the matrix pencil (A_1, A_2), if quadrature points and the corresponding weights are symmetric about the real axis, $(z_j, \omega_j) = (\overline{z}_{j+N/2}, \overline{\omega}_{j+N/2}), j = 1, 2, \ldots, N/2$, we can reduce the number of linear systems,

$$\widehat{S}_k = 2\sum_{j=1}^{N/2} \mathrm{Re}(\omega_j z_j^k (z_j A_2 - A_1)^{-1} A_2 V). \tag{7}$$

To improve the numerical stability, we also apply a low-rank approximation of $\widehat{S}$ with a singular value decomposition on an A_2-inner product:

$$\widehat{S} = [\widehat{U}, \widehat{U}'] \begin{bmatrix} \widehat{\Sigma} & \\ & \widehat{\Sigma}' \end{bmatrix} \begin{bmatrix} \widehat{W}^{\mathrm{T}} \\ \widehat{W}'^{\mathrm{T}} \end{bmatrix} \approx \widehat{U}\widehat{\Sigma}\widehat{W}^{\mathrm{T}},$$
$$\widehat{U}^{\mathrm{T}} A_2 \widehat{U} = I, \quad \widehat{W}^{\mathrm{T}}\widehat{W} = I,$$

where $\widehat{\Sigma}$ is a diagonal matrix whose diagonal entries are the larger part of the singular values, i.e., $\sigma_i/\sigma_1 \geq \delta$ ($\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{LM}$), and the columns of $\widehat{U}, \widehat{W}$ are the corresponding singular vectors. Let $\widehat{d}$ be a numerical rank, $\sigma_{\widehat{d}}/\sigma_1 \geq \delta > \sigma_{\widehat{d}+1}/\sigma_1$. Then, the UOP problem (6) is rewritten as

$$\min_{\widehat{C} \in \mathbb{R}^{\widehat{d} \times \ell}} \left\| \begin{bmatrix} \mu^{1/2}Z^{\mathrm{T}} \\ O \end{bmatrix} - \begin{bmatrix} \mu^{1/2}X^{\mathrm{T}}\widehat{U} \\ (1-\mu)^{1/2}f(\widehat{T})^{1/2} \end{bmatrix} \widehat{C} \right\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \widehat{C}^{\mathrm{T}}\widehat{C} = I, \tag{8}$$

where $\widehat{T} = \widehat{U}^{\mathrm{T}} A_1 \widehat{U}$ and the map is obtained from $B = \widehat{U}\widehat{C}$.

The algorithm of the proposed method is summarized in Algorithm 1. One of the most time-consuming parts of the proposed method is computing the solutions of the $N/2$ linear systems with L right-hand sides in (7) and Step 2 of Algorithm 1 as follows:

$$(z_j A_2 - A_1)P_j = A_2 V, \quad j = 1, 2, \ldots, N/2. \tag{9}$$

For solving these linear systems, the proposed method has hierarchical parallelism. By making this hierarchical structure of the algorithms responsive to the hierarchical structure of the recent architecture, the proposed method is expected to achieve a scalability as high as the complex moment-based eigensolvers, which is a significant advantage for parallel computation. For complex moment-based eigensolvers, their parallel efficiency was demonstrated in previous research (Kestyn et al. 2016; Iwase et al. 2017).

Algorithm 1 A complex moment-based supervised eigenmap for dimensionality reduction

Input: Training dataset: $X \in \mathbb{R}^{m \times n}, Z \in \mathbb{R}^{\ell \times n}$ and parameters: $L, M, N \in \mathbb{N}_+, \delta \in \mathbb{R}, V \in \mathbb{R}^{m \times L}, (z_j, \omega_j)$ for $j = 1, 2, \ldots, N, \Omega = [a, b], \mu, f(\cdot)$.
Output: Linear eigenmap $B \in \mathbb{R}^{m \times \ell}$

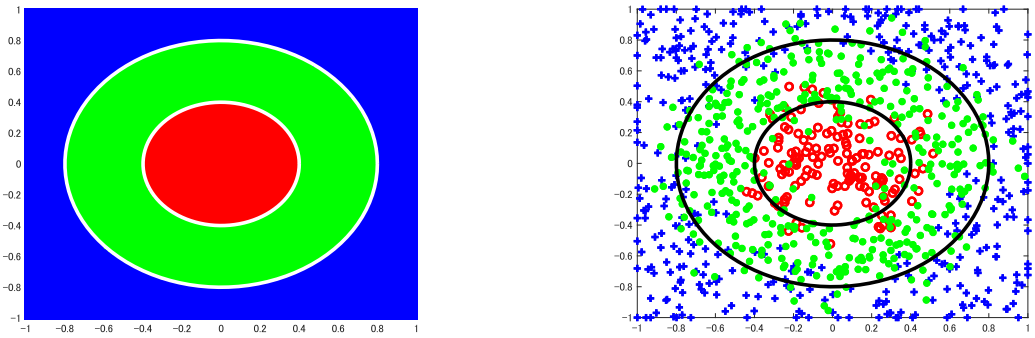
1: Construct a matrix pencil (A_1, A_2) from the training dataset X (and Z if required)
2: Compute $\widehat{S}_k = \sum_{j=1}^{N/2} \mathrm{Re}(\omega_j z_j^k (z_j A_2 - A_1)^{-1} A_2 V)$, and set $\widehat{S} = [\widehat{S}_0, \widehat{S}_1, \ldots, \widehat{S}_{M-1}]$
3: Compute a low-rank approximation of $\widehat{S}$ using the threshold δ: $\widehat{S} = [\widehat{U}, \widehat{U}'][\widehat{\Sigma}, O; O, \widehat{\Sigma}'][\widehat{W}, \widehat{W}'^{\mathrm{T}}] \approx \widehat{U}\widehat{\Sigma}\widehat{W}^{\mathrm{T}}$ such that $\widehat{U}^{\mathrm{T}} A_2 \widehat{U} = I$
4: Solve UOP problem (8) and set $B = \widehat{U}\widehat{C}$

Extensions of the proposed method and relations between existing methods

In the proposed method, the matrices A_1 and A_2 are derived from existing dimensionality reduction methods, that is, we can consider several variants of the proposed method based on each dimensionality reduction method. The concept of the nonlinear dimensionality reduction can also be applied to the proposed method to improve the recognition performance.

There are no direct methods for computing an optimal solution to the UOP problem (8) because of its nonlinear constraint. Therefore, we use an approximate solution obtained by an iterative method (Eldén and Park 1999; Park 1991). As another choice, in future, we may relax or ignore the orthogonal constraint. If we ignore the orthogonal constraint, the UOP problem (8) is reduced to a linear least-squares problem and can be solved efficiently.

The proposed method can be recognized as a combination of the existing dimensionality reduction method based on trace minimization/maximization and ridge regression. If the contour integral (3) is computed exactly, then the proposed method with specific parameters simplifies to dimensionality reduction methods and ridge regression. For example, the matrices A_1 and A_2 are those used in trace minimization or maximization methods like LPP or LFDA. In this case, if



(a) Ground truth (b) Training dataset

Figure 1: Ground truth and training dataset for the artificial problem.

we set Ω to include nonzero ℓ smallest or largest eigenvalues, then the proposed method with $\mu = 0$ and $f(\lambda) = \lambda$ or $f(\lambda) = 1/\lambda$ are mathematically equivalent to the corresponding dimensionality reduction methods. Also, if we set Ω to include all eigenvalues and ignore the orthogonal constraint in (6), then the proposed method with $f(\lambda) = 1$ is mathematically equivalent to ridge regression.

Numerical experiments

The main aspects of the proposed method for achieving high recognition performance are (i) a complex moment-based subspace including multiple eigenvectors, (ii) a novel objective function that combines a matrix trace and a squared error and (iii) an efficient and parallel implementation based on techniques used for complex moment-based parallel eigensolvers. To evaluate the effect of these aspects of the proposed method on recognition performance, here we compare the performance of the kernel version of the proposed method (complex moment-based supervised eigenmap, K–CMSE) with the performance of Kernel LPP (K–LPP), Kernel LFDA (K–LFDA) and Kernel ridge regression (K–RR).

We use the Gaussian kernel as the kernel function. The dimension size ℓ is set as the number of classes for each problem. The similarity matrix is sparsified with the k-nearest neighbor approach ($k = 7$). For each problem, a line search tunes the regularization parameter of K–RR.

For K–CMSE, we use the same matrices A_1 and A_2 as those used for K–LPP. We also use $(M, N, \delta) = (8, 32, 10^{-15})$, which are the default parameters for complex moment-based eigensolvers. The input matrix V is a random matrix generated by the Mersenne Twister. We set $\Omega = [0, b]$ and the quadrature points as on an ellipse with center $\gamma = b/2$, major axis $\rho = b/2$ and aspect ratio $\alpha = 0.1$ as follows:

$$z_j = \gamma + \rho\left(\cos(\theta_j) + \alpha \mathrm{i}\sin(\theta_j)\right), \quad \theta_j = \frac{2\pi}{N}\left(j - \frac{1}{2}\right)$$

for $j = 1, 2, \ldots, N/2$. The corresponding weights are set as

$$\omega_j = \frac{\rho}{N}\left(\alpha\cos(\theta_j) + \mathrm{i}\sin(\theta_j)\right)$$

for $j = 1, 2, \ldots, N/2$. The nonlinear function $f(\cdot)$ is defined as $f(\lambda) = 1/(b - \lambda)^2$. We solve the UOP problem (8) using an iterative method (Zhao, Wang, and Nie 2016).

In the training phase, we use the ground truth Z as a binary matrix whose (i, j) entry is 1 if the training data $\boldsymbol{x}_j$ is in class i. This type of ground truth Z is used for several classification algorithms including the ridge regression and deep neural networks (Bishop 2006). Then, in the prediction phase, we firstly apply the trained dimensionality reduction and apply the k-nearest neighbors (Altman 1992) for classification to the obtained low dimensional data.

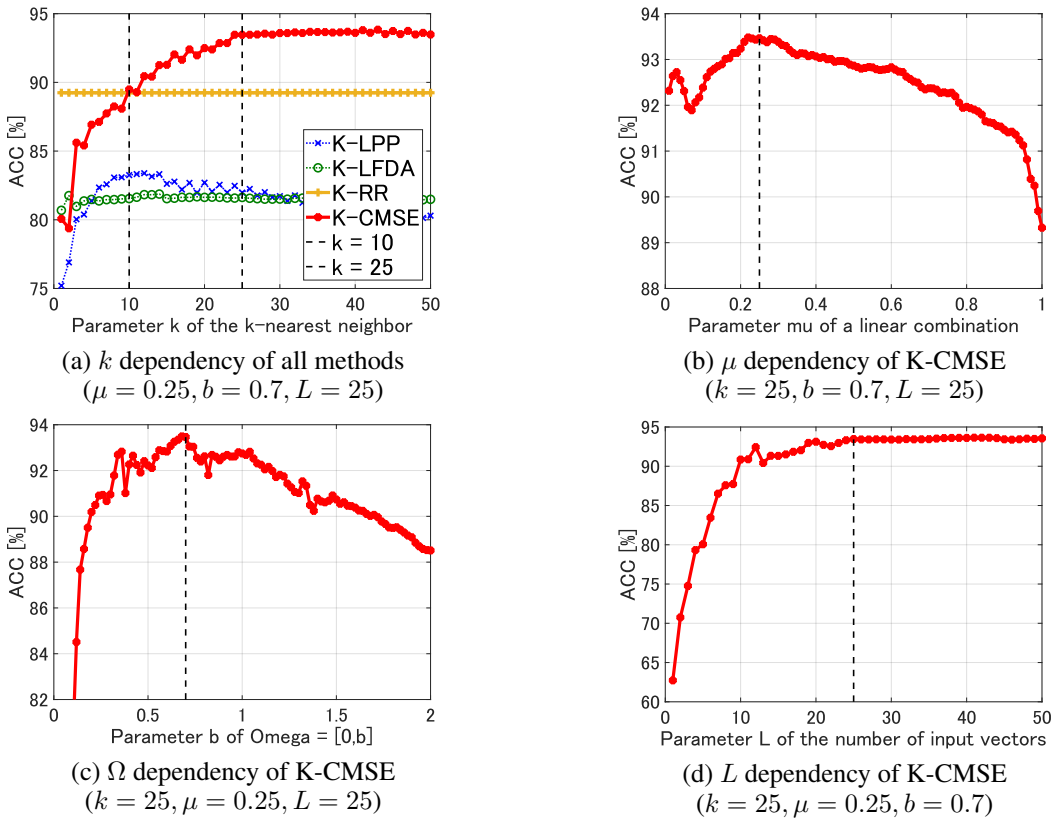
Numerical experiments I and II were performed using MATLAB2017b, and numerical experiment III was performed using Fortran 90 and MPI.

Experiment I: artificial data

In this experiment, we compare the recognition performance of the dimensionality reduction methods for the three-class classification of 10-dimensional artificial data. The first two dimensions of the ground truth are shown in Figure 1(a). In Figure 1(b), we show 1000 training data points of the first two dimensions with the corresponding labels: $\circ$, $\bullet$ and $+$. As shown in Figure 1, the training dataset has noise and deviates from the ground truth to evaluate the overfitting of the methods due to the noise. For the test dataset, we use 201×201 data points whose first two dimensions are square grid points in $[-1, 1]\times[-1, 1]$. The other eight dimensions of the training and test dataset are random values in $[-0.1, 0.1]$ generated by the Mersenne Twister in MATLAB.

Firstly, we evaluate the parameter dependency of the recognition performance, accuracy (ACC) defined by the rate of correct predictions to the number of test data points, for $k = 1, 2, \ldots, 50$ of the k-nearest neighbors, $\mu = 0.01, 0.02, \ldots, 1.00$ of a linear combination, $b = 0.02, 0.04, \ldots, 2.00$ of Ω and $L = 1, 2, \ldots, 50$ of the number of input vectors.

Figure 2 (a) shows the k dependency of the performance of all methods, and Figures 2 (b), (c) and (d) show the μ, Ω and L dependencies, respectively, of the performance of the proposed method, K–CMSE. From Figure 2 (a), we can see that K–CMSE outperforms the existing methods for $k > 10$,



(a) k dependency of all methods ($\mu = 0.25, b = 0.7, L = 25$)

(b) μ dependency of K-CMSE ($k = 25, b = 0.7, L = 25$)

(c) Ω dependency of K-CMSE ($k = 25, \mu = 0.25, L = 25$)

(d) L dependency of K-CMSE ($k = 25, \mu = 0.25, b = 0.7$)

Figure 2: Parameter dependency of accuracy (ACC) for the artificial problem. For each figure, one parameter is changed and the other parameters are fixed at $k = 25, \mu = 0.25, b = 0.7$ and $L = 25$. Vertical dashed lines denote the fixed parameters.

although the performance of K–LPP and K–CMSE are dependent on k. Figure 2 (b) demonstrates that the performance of K–CMSE with $\mu \approx 1$ is relatively low. Note that μ is the weight for a linear combination of two objective functions and K–CMSE with $\mu = 0$ is strongly related to K-LPP. Therefore, the result in Figure 2 (b) indicates that the combination of objective functions gives a better recognition performance than using each objective function alone. Figure 2 (c) and (d) also show that the complex moment-based subspace with wide region Ω and large L, which include a larger number of eigenvectors, improves recognition performance, although an overlarge Ω leads to limited recognition performance.

The classification results of the existing methods and K–CMSE are shown in Figure 3. At first glance, the existing methods give good classification results; however, the predicted boundary of each class is not sharp, which indicates that is overfitting due to the noise in the training dataset. In contrast, in comparison with the existing methods, the proposed method (K–CMSE) gives a better and sharper classification result and less overfitting.

Experiment II: real-world data

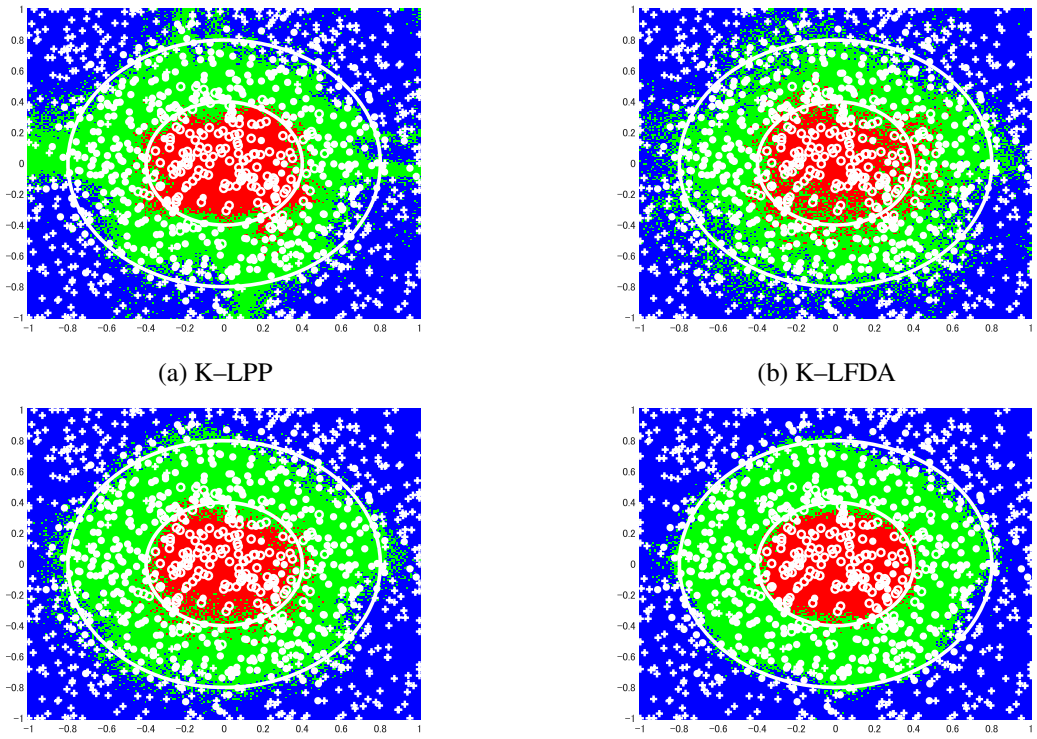
Here, we evaluate the performance of the dimensionality reduction methods on normalized mutual information (NMI) (Strehl and Ghosh 2002), accuracy (ACC) and Rand index (RI) (Rand 1971). As test problems, we treat the binary and multiclass classification problems obtained from (LeCun 1998; Samaria and Harter 1994) and feature selection datasets which is available at http://featureselection.asu.edu/datasets.php.

In these numerical experiments, k of the k-nearest neighbor, the regularization parameter of K–RR and (μ, b, L) of K–CMSE are tuned by applying a line search to each parameter sequentially and by a 10-fold cross-validation until convergence. Then, the performance of each method with the tuned parameters is evaluated by a 10-fold cross-validation using a different validation set from that used for parameter tuning.

The numerical results (average $\pm$ standard error) are summarised in Table 1. We can observe from Table 1 that the proposed method (K–CMSE) has a recognition performance higher than those of existing methods for binary and multiclass classifications.

Experiment III: scalability

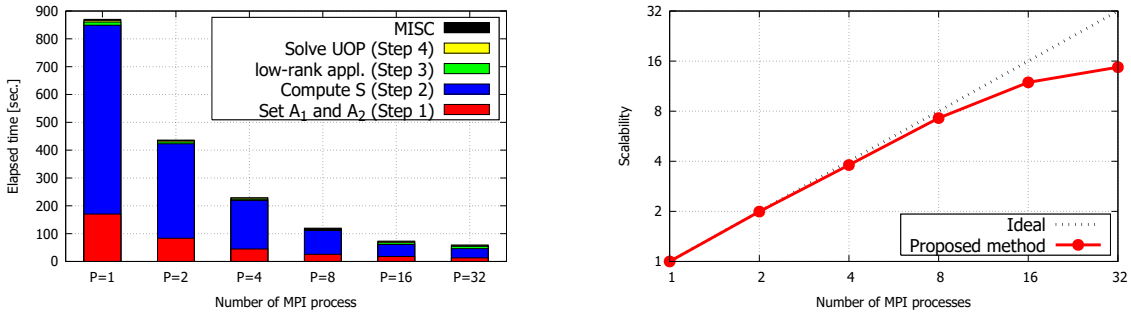
In this experiment, we evaluate the strong scalability of the main parts of K–CMSE which construct the matrices A_1 and A_2, compute the complex moment matrix $\widehat{S}$ (7), low-rank approximation and solve the UOP problem. For the test problem, we use the 10-class classification MNIST with 60,000 training data points (LeCun 1998).



(a) K–LPP (b) K–LFDA

(c) K–RR (d) K–CMSE

Figure 3: Classification results of the existing methods ($k = 10$) and K–CMSE ($k = 25, \mu = 0.25, b = 0.7, L = 25$) for the artificial problem.



(a) elapsed time (b) scalability

Figure 4: Elapsed time and strong scalability of the main part of K–CMSE for MNIST with 60,000 training data points.

The numerical experiments were conducted on COMA at the Center for Computational Sciences, University of Tsukuba, Japan. COMA has two Intel Xeon E5-2670v2 (2.5 GHz) processors and two Intel Xeon Phi 7110P (61 cores) processors per node. In this numerical experiment, we use only the CPU. The algorithms are implemented in Fortran 90 and MPI and executed with 1 to 32 nodes (one MPI process per node). The sparse linear systems (9) are solved using "cluster_sparse_solver" in Intel MKL.

Figure 4 shows the elapsed time and the strong scalability of the main parts of the proposed method. We observe from Figure 4 that the most time-consuming part of the proposed method is computing the complex moment matrix $\widehat{S}$, which is scaled well by parallelization. As a result, the proposed method exhibits good strong scalability.

Remarks on numerical results

The results of experiments I and II indicate that our combined objective function and the complex moment-based subspace achieve better recognition performance over existing methods for artificial and real-world problems. Experiment III demonstrates that the proposed method exhibits a high parallel efficiency based on techniques used for complex moment-based eigensolvers.

Table 1: Recognition performance (average $\pm$ standard error) for real-world problems. The parameters m, n, ℓ denote the number of the features, samples and classes. The bold fonts denote the best method(s) for each evaluation index.

att644 $(m, n, \ell) = (644, 400, 40)$

Method	NMI	ACC	RI
K-LPP	0.92±0.01	69.75±3.31	97.56±0.32
K-LFDA	0.95±0.01	83.75±1.63	98.42±0.21
K-RR	0.95±0.01	86.25±1.24	98.59±0.21
K-CMSE	**0.96±0.01**	**86.50±1.28**	**98.64±0.22**

GLI-85 $(m, n, \ell) = (22283, 85, 2)$

Method	NMI	ACC	RI
K-LPP	0.18±0.10	66.92±5.87	57.53±5.23
K-LFDA	0.41±0.12	85.19±2.90	73.30±4.86
K-RR	**0.48±0.11**	85.48±3.03	73.82±5.10
K-CMSE	0.47±0.11	**86.73±2.83**	**75.60±4.79**

GLIOMA $(m, n, \ell) = (4434, 50, 4)$

Method	NMI	ACC	RI
K-LPP	0.83±0.04	66.00±8.02	83.00±4.48
K-LFDA	0.86±0.04	74.00±8.02	**88.00±3.41**
K-RR	0.84±0.04	74.00±8.02	85.00±4.30
K-CMSE	**0.87±0.04**	**76.00±7.38**	**88.00±3.69**

Isolet1 $(m, n, \ell) = (617, 1560, 26)$

Method	NMI	ACC	RI
K-LPP	0.85±0.00	73.78±1.06	97.16±0.15
K-LFDA	0.97±0.00	96.86±0.38	99.56±0.06
K-RR	**0.98±0.00**	97.12±0.34	99.59±0.05
K-CMSE	**0.98±0.00**	**97.31±0.38**	**99.62±0.06**

MNIST $(m, n, \ell) = (784, 2000, 10)$

Method	NMI	ACC	RI
K-LPP	0.80±0.01	83.85±0.82	94.78±0.21
K-LFDA	**0.93±0.01**	95.50±0.49	98.27±0.18
K-RR	**0.93±0.01**	95.55±0.47	98.30±0.16
K-CMSE	**0.93±0.01**	**95.70±0.33**	**98.35±0.11**

pixraw10P $(m, n, \ell) = (10000, 100, 10)$

Method	NMI	ACC	RI
K-LPP	0.95±0.02	90.00±4.00	96.67±1.55
K-LFDA	**0.98±0.01**	**97.00±2.02**	98.67±0.90
K-RR	**0.98±0.01**	**97.00±2.02**	98.67±0.90
K-CMSE	**0.98±0.01**	**97.00±1.45**	**98.89±0.57**

SMK-CAN-187 $(m, n, \ell) = (19993, 187, 2)$

Method	NMI	ACC	RI
K-LPP	0.12±0.03	65.84±3.09	54.48±1.28
K-LFDA	0.17±0.04	68.71±2.57	55.94±1.95
K-RR	**0.18±0.05**	68.71±2.69	56.07±2.25
K-CMSE	**0.18±0.04**	**71.89±2.48**	**58.56±2.31**

TOX-171 $(m, n, \ell) = (5748, 171, 4)$

Method	NMI	ACC	RI
K-LPP	0.42±0.03	50.92±2.69	69.14±1.37
K-LFDA	0.91±0.03	94.77±1.92	94.34±2.13
K-RR	0.93±0.02	95.92±1.45	95.81±1.70
K-CMSE	**0.94±0.02**	**96.50±1.48**	**96.32±1.74**

Therefore, we conclude that these numerical experiments confirm the efficiency of the main aspects of the proposed method.

Conclusions

Here, we proposed a novel complex moment-based supervised dimensionality reduction method, which achieves high recognition performance by extending the existing dimensionality reduction methods using a complex moment-based subspace. The main aspects of the proposed method are (i) a complex moment-based subspace including multiple eigenvectors, (ii) a novel objective function that combines a matrix trace and a squared error and (iii) an efficient and parallel implementation based on techniques used for the complex moment-based parallel eigensolvers. The numerical results confirm that these aspects help to achieve high recognition performance and high parallel efficiency.

Note that we also tested the proposed method with a matrix pencil (A_1, A_2) derived from K–LFDA and obtained almost the same good results as demonstrated in this paper.

In the future, we will evaluate the performance of the proposed method based on other dimensionality reduction methods and compared it with those of existing methods when solving large real-world problems in parallel environments. We will also investigate tuning strategies for the hyperparameters of the proposed method.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful and constructive comments. The present study is supported in part by the Japan Science and Technology Agency (JST), ACT-I (No. JPMJPR16U6), the New Energy and Industrial Technology Development Organization (NEDO) and the Japan Society for the Promotion of Science (JSPS), Grants-in-Aid for Scientific Research (Nos. 17K12690, 18H03250).

References

Altman, N. S. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46(3):175–185.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag Berlin, Heidelberg.

Eldén, L., and Park, H. 1999. A procrustes problem on the Stiefel manifold. *Numerische Mathematik* 82(4):599–619.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of human genetics* 7(2):179–188.

Fukunaga, K. 2013. *Introduction to statistical pattern recognition*. Academic press.

Güttel, S.; Polizzi, E.; Tang, P. T. P.; and Viaud, G. 2015. Zolotarev quadrature rules and load balancing for the FEAST eigensolver. *SIAM Journal on Scientific Computing* 37(4):A2100–A2122.

He, X., and Niyogi, P. 2004. Locality preserving projections. In *Advances in neural information processing systems*, 153–160.

Higham, N. J. 2008. *Functions of matrices: theory and computation*, volume 104. Siam.

Ikegami, T., and Sakurai, T. 2010. Contour integral eigensolver for non-Hermitian systems: a Rayleigh-Ritz-type approach. *Taiwanese Journal of Mathematics* 825–837.

Ikegami, T.; Sakurai, T.; and Nagashima, U. 2010. A filter diagonalization for generalized eigenvalue problems based on the Sakurai–Sugiura projection method. *Journal of Computational and Applied Mathematics* 233(8):1927–1936.

Imakura, A., and Sakurai, T. 2017. Block Krylov-type complex moment-based eigensolvers for solving generalized eigenvalue problems. *Numerical Algorithms* 75(2):413–433.

Imakura, A.; Du, L.; and Sakurai, T. 2014. A block Arnoldi-type contour integral spectral projection method for solving generalized eigenvalue problems. *Applied Mathematics Letters* 32:22–27.

Imakura, A.; Du, L.; and Sakurai, T. 2016. Relationships among contour integral-based methods for solving generalized eigenvalue problems. *Japan Journal of Industrial and Applied Mathematics* 33(3):721–750.

Imakura, A.; Futamura, Y.; and Sakurai, T. 2017. Structure-preserving technique in the block SS–Hankel method for solving Hermitian generalized eigenvalue problems. In *International Conference on Parallel Processing and Applied Mathematics*, 600–611. Springer.

Iwase, S.; Futamura, Y.; Imakura, A.; Sakurai, T.; and Ono, T. 2017. Efficient and scalable calculation of complex band structure using Sakurai-Sugiura method. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 40. ACM.

Jolliffe, I. T. 1986. Principal component analysis and factor analysis. In *Principal component analysis*. Springer. 115–128.

Kestyn, J.; Kalantzis, V.; Polizzi, E.; and Saad, Y. 2016. PFEAST: a high performance sparse eigenvalue solver using distributed-memory linear solvers. In *High Performance Computing, Networking, Storage and Analysis, SC16: International Conference for*, 178–189. IEEE.

LeCun, Y. 1998. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Li, X.; Chen, M.; Nie, F.; and Wang, Q. 2017. Locality adaptive discriminant analysis. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2201–2207. AAAI Press.

Park, H. 1991. A parallel algorithm for the unbalanced orthogonal Procrustes problem. *Parallel Computing* 17(8):913–923.

Pearson, K. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2(11):559–572.

Polizzi, E. 2009. A density matrix-based algorithm for solving eigenvalue problems. *Phys. Rev. B* 79:115112.

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* 66(336):846–850.

Sakurai, T., and Sugiura, H. 2003. A projection method for generalized eigenvalue problems using numerical integration. *Journal of computational and applied mathematics* 159(1):119–128.

Sakurai, T., and Tadano, H. 2007. CIRR: a Rayleigh-Ritz type method with counter integral for generalized eigenvalue problems. *Hokkaido Math. J.* 36:745–757.

Samaria, F., and Harter, A. 1994. Parameterisation of a stochastic model for human face identification. In *Proceeding of IEEE Workshop on Applications of Computer Vision*.

Saunders, C.; Gammerman, A.; and Vovk, V. 1998. Ridge regression learning algorithm in dual variables.

Schölkopf, B.; Smola, A.; and Müller, K.-R. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation* 10(5):1299–1319.

Strehl, A., and Ghosh, J. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research* 3(Dec):583–617.

Sugiyama, M.; Idé, T.; Nakajima, S.; and Sese, J. 2010. Semi-supervised local Fisher discriminant analysis for dimensionality reduction. *Machine learning* 78(1-2):35.

Sugiyama, M. 2007. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *Journal of machine learning research* 8(May):1027–1061.

Tang, P. T. P., and Polizzi, E. 2014. FEAST as a subspace iteration eigensolver accelerated by approximate spectral projection. *SIAM Journal on Matrix Analysis and Applications* 35(2):354–390.

Zhao, H.; Wang, Z.; and Nie, F. 2016. Orthogonal least squares regression for feature extraction. *Neurocomputing* 216:200–207.