DPCA: DIMENSIONALITY REDUCTION FOR DISCRIMINATIVE ANALYTICS OF MULTIPLE LARGE-SCALE DATASETS

Gang Wang, Jia Chen, and Georgios B. Giannakis

ECE Dept. and Digital Tech. Center, Univ. of Minnesota, Mpls., MN 55455, USA
Emails: {gangwang, chen5625, georgios}@umn.edu

ABSTRACT

Principal component analysis (PCA) has well-documented merits for data extraction and dimensionality reduction. PCA deals with a single dataset at a time, and it is challenged when it comes to analyzing multiple datasets. Yet in certain setups, one wishes to extract the most significant information of one dataset relative to other datasets. Specifically, the interest may be on identifying or extracting features that are specific to a single target dataset but not the others. This paper presents a novel approach for such so-termed *discriminative data analysis*, and establishes its *optimality* in the least-squares sense under suitable assumptions. The criterion reveals linear combinations of variables by maximizing the ratio of the variance of the target data to that of the remainders. The novel approach solves a generalized eigenvalue problem by performing SVD just once. Numerical tests using synthetic and real datasets showcase the merits of the proposed approach relative to its competing alternatives.

Index Terms— Dimensionality reduction, robust principal component analysis, discriminative analytics

1. INTRODUCTION

Principal component analysis (PCA) is arguably the most widely used method for data visualization and dimensionality reduction [1]. PCA originated in statistics [2], but its modern instantiation as well as the term *principal component* (PC) vector was formalized in [3]. The goal of PCA is to extract the most important information from a data table representing observations, and depict it as a few PCs. PCs are uncorrelated linear transformations of the original set of variables, along which the maximum variation in the data is captured [1].

Yet, several application domains involve *multiple* datasets, and the task is to extract trends or features depicted by *component vectors* that are present in one dataset *but not* the other(s) [4]. For example, consider two gene expression datasets of individuals from across different countries and genders: the first includes gene expression levels of cancer patients, which constitutes the *target data* that we want to analyze, while the second is formed by healthy volunteers, and is called *control* or *background data*. Applying PCA on either the target data or the target augmented with the control data is likely to obtain PCs that represent the background information common to both datasets (e.g., the demographic patterns, genders) [5], instead of the component vectors depicting the subtypes of cancer within patients. Despite its practical relevance, such discriminative data analysis has not been widely studied.

Generalizations to PCA include multi-dimensional scaling [6], local linear embedding [7], robust or kernel PCA [8, 9, 10], and canonical component analysis [11, 12]. Given multiple datasets, analysts have to perform these procedures on each individual dataset, and subsequently evaluate manually the obtained projections to identify whether significant patterns representing similarities or differences across datasets are present. A recent approach pursued what is termed contrastive (c) PCA for extracting the most distinct features of one dataset relative to the other [4]. cPCA is able to reveal the dataset-specific information often missed by PCA [4]. This becomes possible through a hyper-parameter that takes into account both target and control data, and critically affects the performance of cPCA. The cPCA solution is often found with SVD. Although possible to automatically select the best from a list of given values, computing SVD multiple times can be computationally cumbersome or even prohibitive in large-scale data extraction settings. Another method related to PCA is linear discriminant analysis (LDA) [13], that is "supervised," and seeks linear combinations of variables to maximize the separation between classes. This is achieved by maximizing the ratio of the variance between classes to the variance within the classes.

Inspired by LDA and cPCA, this paper puts forth a new method for discriminative analytics, which is shown to be optimal in the least-squares (LS) sense provided that the background component vectors are also present in the target data. Our method seeks linear combinations of variables by maximizing the ratio of the variance of target data to the variance of control data, which justifies our chosen description as *discriminative (d) PCA*. dPCA solves a generalized eigenvalue problem. Relative to cPCA, dPCA is parameter-free, and requires only one SVD. As such, dPCA is well-suited for large-scale either discriminative or contrasting data exploration.

Work in this paper was supported in part by NIH 1R01GM104975-01 and NSF 1500713.

2. PRELIMINARIES AND PRIOR ART

Consider two datasets, the target data $\{\boldsymbol{x}_i \in \mathbb{R}^D\}_{1\leq i\leq m}$ that we are interested in analyzing, and data $\{\boldsymbol{y}_j \in \mathbb{R}^D\}_{1\leq j\leq n}$ containing latent background component vectors in the target data. Assume without loss of generality (wlog) that the sample mean of each dataset has been removed, and let $\boldsymbol{C}_{xx} := (1/m)\sum_{i=1}^m \boldsymbol{x}_i\boldsymbol{x}_i^\top$ and $\boldsymbol{C}_{yy} := (1/n)\sum_{i=1}^n \boldsymbol{y}_i\boldsymbol{y}_i^\top$ denote the corresponding sample covariance matrices. To motivate the novel approach in Sec. 3, the basics of PCA and cPCA are briefly reviewed in this section.

One formulation of PCA seeks vectors $\{\boldsymbol{\chi}_i \in \mathbb{R}^d\}_{1\leq i\leq m}$ as linear combinations of $\{\boldsymbol{x}_i \in \mathbb{R}^D\}_{1\leq i\leq m}$ with $d < D$ via maximizing their variances in the low-dimensional subspaces [1]. Specifically for $d = 1$, (linear) PCA obtains $\boldsymbol{\chi}_i := \boldsymbol{u}^\top\boldsymbol{x}_i$, where the direction $\boldsymbol{u} \in \mathbb{R}^D$ is found by

$$\max_{\boldsymbol{u}\in\mathbb{R}^D} \quad \boldsymbol{u}^\top\boldsymbol{C}_{xx}\boldsymbol{u} \tag{1a}$$

$$\text{s. to} \quad \boldsymbol{u}^\top\boldsymbol{u} = 1. \tag{1b}$$

Solving (1) yields $\boldsymbol{u}$ as the principal eigenvector of matrix $\boldsymbol{C}_{xx}$, also known as the first PC. Instead of having constraint (1b) explicitly, we assume wlog that the solution $\boldsymbol{u}$ will always be normalized to have unity norm. For $d > 1$, PCA amounts to computing the first d principal eigenvectors of $\boldsymbol{C}_{xx}$ instead. As alluded to in Sec. 1, when two datasets $\{\boldsymbol{x}_i\}$ and $\{\boldsymbol{y}_j\}$ are available, PCA performed either on $\{\boldsymbol{x}_i\}$, or on $\{\{\boldsymbol{x}_i\},\{\boldsymbol{y}_j\}\}$, can generally not unveil the patterns or trends that are specific to the target relative to the control data.

Contrastive (c) PCA [4], on the other hand, aims to identify directions $\boldsymbol{u}$ along which the target data possesses large variations while the control data has small variations. Concretely, cPCA pursues problem [4]

$$\max_{\|\boldsymbol{u}\|_2=1} \quad \boldsymbol{u}^\top\boldsymbol{C}_{xx}\boldsymbol{u} - \alpha\boldsymbol{u}^\top\boldsymbol{C}_{yy}\boldsymbol{u}. \tag{2}$$

whose solution is given by the eigenvector of $\boldsymbol{C}_{xx} - \alpha\boldsymbol{C}_{yy}$ associated with the largest eigenvalue, and constitutes the first contrastive (c) PC. Here, $\alpha > 0$ is a hyper-parameter that trades off maximizing the target data variance (the first term in (2)) for minimizing the control data variance (second term). However, there is no rule of thumb for choosing α. Although a spectral clustering based algorithm has been developed to automatically select the value of α, its brute-force search discourages its use in large-scale datasets.

3. THE NOVEL APPROACH

Unlike PCA, LDA is "supervised," and seeks directions that yield the largest possible separation between classes via maximizing the ratio of the variance across classes to the variance within classes. In the same vein, when both the target and the background data are available, and one is interested in extracting features, namely component vectors that are *only* present in the target data but *not* in the background data, a meaningful approach would be to maximize the ratio of the variance of the target data over that of the background data

$$\max_{\|\boldsymbol{u}\|_2=1} \quad \frac{\boldsymbol{u}^\top\boldsymbol{C}_{xx}\boldsymbol{u}}{\boldsymbol{u}^\top\boldsymbol{C}_{yy}\boldsymbol{u}} \tag{3}$$

which, with slight abuse of the term "discriminant," we call *discriminative (d) PCA*. Likewise, the solution of (3) will be termed first discriminative PC, or dPC for short.

3.1. dPCA Algorithm

Suppose that $\boldsymbol{C}_{yy}$ is full rank with eigen-decomposition $\boldsymbol{C}_{yy} := \boldsymbol{U}_y^\top\boldsymbol{\Sigma}_y\boldsymbol{U}_y$. Upon defining $\boldsymbol{C}_{yy}^{1/2} := \boldsymbol{\Sigma}_y^{1/2}\boldsymbol{U}_y$, and changing variables $\boldsymbol{v} := \boldsymbol{C}_{yy}^{1/2}\boldsymbol{u}$, (3) admits the same solution as

$$\boldsymbol{v}^* := \arg\max_{\|\boldsymbol{v}\|_2=1} \quad \boldsymbol{v}^\top\boldsymbol{C}_{yy}^{-\top/2}\boldsymbol{C}_{xx}\boldsymbol{C}_{yy}^{-1/2}\boldsymbol{v} \tag{4}$$

which is the principal eigenvector of $\boldsymbol{C}_{yy}^{-\top/2}\boldsymbol{C}_{xx}\boldsymbol{C}_{yy}^{-1/2}$. Finally, the solution to (3) is recovered as $\boldsymbol{u}^* := \boldsymbol{C}_{yy}^{-1/2}\boldsymbol{v}^*$, followed by normalization to obtain a unit norm.

On the other hand, leveraging Lagrangian duality, the solution of (3) can also be obtained as the right eigenvector of $\boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{xx}$. To see this, note that (3) can be rewritten as

$$\max_{\boldsymbol{u}\in\mathbb{R}^D} \quad \boldsymbol{u}^\top\boldsymbol{C}_{xx}\boldsymbol{u} \tag{5a}$$

$$\text{s. to} \quad \boldsymbol{u}^\top\boldsymbol{C}_{yy}\boldsymbol{u} = b \tag{5b}$$

for some constant $b > 0$ such that the solution $\|\boldsymbol{u}^*\|_2 = 1$. One can set $b = 1$ and subsequently normalize the solution of (5). Letting $\lambda \in \mathbb{R}$ be the dual variable corresponding to constraint (5b), the Lagrangian of (5) is

$$\mathcal{L}(\boldsymbol{u};\lambda) = \boldsymbol{u}^\top\boldsymbol{C}_{xx}\boldsymbol{u} + \lambda\big(1 - \boldsymbol{u}^\top\boldsymbol{C}_{yy}\boldsymbol{u}\big). \tag{6}$$

The KKT conditions assert that for the optimal $(\boldsymbol{u}^*;\lambda^*)$, it holds that $\boldsymbol{C}_{xx}\boldsymbol{u}^* = \lambda^*\boldsymbol{C}_{yy}\boldsymbol{u}^*$, which is a generalized eigenvalue problem. Equivalently, one can rewrite

$$\boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{xx}\boldsymbol{u}^* = \lambda^*\boldsymbol{u}^* \tag{7}$$

suggesting that $\boldsymbol{u}^*$ is an eigenvector of $\boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{xx}$ associated with eigenvalue λ^*. Respecting the constraint $(\boldsymbol{u}^*)^\top\boldsymbol{C}_{yy}\boldsymbol{u}^* = 1$, the objective (5a) reduces to

$$(\boldsymbol{u}^*)^\top\boldsymbol{C}_{xx}\boldsymbol{u}^* = \lambda^*(\boldsymbol{u}^*)^\top\boldsymbol{C}_{yy}\boldsymbol{u}^* = \lambda^*. \tag{8}$$

It is clear now that the optimal objective value of problem (5) is equal to the largest eigenvalue of $\boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{xx}$, and the optimal solution $\boldsymbol{u}^*$ is the corresponding eigenvector.

For $d > 1$, one finds the d (right) eigenvectors of $\boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{xx}$ that correspond to the d largest eigenvalues as the first d dPCs. For ease of implementation, the proposed dPCA approach for contrastive data exploration is summarized in Alg. 1. Concerning dPCA, a couple of remarks are in order.

Algorithm 1 Discriminative principal component analysis.

1: **Input:** Nonzero-mean target and background data $\{\mathring{\boldsymbol{x}}_i\}_{1\le i\le m}$, $\{\mathring{\boldsymbol{y}}_j\}_{1\le j\le n}$; number of dPCs d.
2: **Remove** the mean from $\{\mathring{\boldsymbol{x}}_i\}$ and $\{\mathring{\boldsymbol{y}}_j\}$ to yield centered data $\{\boldsymbol{x}_i\}$, and $\{\boldsymbol{y}_j\}$.
3: **Construct** the sample covariance matrices:

$$\boldsymbol{C}_{xx} := \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^\top, \quad \text{and} \quad \boldsymbol{C}_{yy} := \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{y}_j\boldsymbol{y}_j^\top$$

4: **Perform** SVD on matrix $\boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{xx}$.
5: **Output** the d (right) singular vectors corresponding to the d largest singular values.

Remark 1. When there is no background data, with $\boldsymbol{C}_{yy} = \boldsymbol{I}_D$, dPCA boils down to PCA. On the other hand, when there are multiple background datasets, one can first combine them into a single one, and then apply dPCA. Other twists will be discussed in the journal version of this paper.

Remark 2. Performing dPCA on $\{\boldsymbol{x}_i\}$ and $\{\boldsymbol{y}_j\}$ can be seen as performing PCA on the transformed data $\{\boldsymbol{C}_{yy}^{-\top/2}\boldsymbol{x}_i\}$ to yield $\boldsymbol{v}^*$, followed by re-transformation $\boldsymbol{u}^* = \boldsymbol{C}_{yy}^{-1/2}\boldsymbol{v}^*$. The new data can be understood as the data obtained after removing the "background" component vectors from the target data.

Remark 3. Inexpensive power or Lanczos iterations [14] can be employed to compute the principal eigenvectors in (4).

3.2. dPCA vis-à-vis cPCA

Consider again the constrained form of dPCA (5) and its Lagrangian (6). Using Lagrange duality, when choosing $\alpha = \lambda^*$ in (2), cPCA maximizing $\boldsymbol{u}^*(\boldsymbol{C}_{xx} - \lambda^*\boldsymbol{C}_{yy})\boldsymbol{u}$ is equivalent to $\max_{\boldsymbol{u}\in\mathbb{R}^D}\mathcal{L}(\boldsymbol{u};\lambda^*) = \boldsymbol{u}^\top(\boldsymbol{C}_{xx} - \lambda^*\boldsymbol{C}_{yy})\boldsymbol{u} + \lambda^*$, which is exactly dPCA. In other words, cPCA and dPCA are equivalent when α in cPCA is carefully set as the optimal dual variable λ^* for the constrained form (5) of dPCA, namely the largest eigenvalue of $\boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{xx}$.

It will be useful for further analysis to focus on simultaneously diagonalizable matrices $\boldsymbol{C}_{xx}$ and $\boldsymbol{C}_{yy}$, that is

$$\boldsymbol{C}_{xx} := \boldsymbol{U}^\top\boldsymbol{\Sigma}_x\boldsymbol{U}, \quad \text{and} \quad \boldsymbol{C}_{yy} := \boldsymbol{U}^\top\boldsymbol{\Sigma}_y\boldsymbol{U} \tag{9}$$

where $\boldsymbol{U}\in\mathbb{R}^{D\times D}$ is unitary and simultaneously decomposes $\boldsymbol{C}_{xx}$ and $\boldsymbol{C}_{yy}$, while diagonal matrices $\boldsymbol{\Sigma}_x$, $\boldsymbol{\Sigma}_y \succ \boldsymbol{0}$ hold accordingly eigenvalues $\{\lambda_x^i\}_{1\le i\le D}$ of $\boldsymbol{C}_{xx}$ and $\{\lambda_y^i\}_{1\le i\le D}$ of $\boldsymbol{C}_{yy}$ on their main diagonals. It clearly holds that $\boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{xx} = \boldsymbol{U}^\top\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_x\boldsymbol{U} = \boldsymbol{U}^\top\mathrm{diag}\big(\{\frac{\lambda_x^i}{\lambda_y^i}\}_{1\le i\le D}\big)\boldsymbol{U}$. Looking for the first d dPCs boils down to taking the d columns of $\boldsymbol{U}$ associated with the d largest eigenvalue ratios among $\{\frac{\lambda_x^i}{\lambda_y^i}\}_{1\le i\le D}$.

On the other hand, the solution of cPCA under a given α, or the first d cPCs of $\boldsymbol{C}_{xx} - \alpha\boldsymbol{C}_{yy} = \boldsymbol{U}^\top(\boldsymbol{\Sigma}_x - \alpha\boldsymbol{\Sigma}_{\boldsymbol{y}})\boldsymbol{U} = \boldsymbol{U}^\top\mathrm{diag}\big(\{\lambda_x^i - \alpha\lambda_y^i\}_{1\le i\le D}\big)\boldsymbol{U}$, are found as the d columns of $\boldsymbol{U}$ that correspond to the d largest numbers in $\{\lambda_x^i - \alpha\lambda_y^i\}_{1\le i\le D}$. In the ensuing section, we show that when given data obey certain models, dPCA is LS optimal.

4. OPTIMALITY OF dPCA

Adopting a bilinear (factor analysis) model, PCA describes the (non-centered) data $\{\mathring{\boldsymbol{y}}_j\}_{1\le j\le n}$ as

$$\mathring{\boldsymbol{y}}_j = \boldsymbol{m}_y + \boldsymbol{U}_y\boldsymbol{\psi}_j + \boldsymbol{e}_{y,j}, \quad 1\le j\le n \tag{10}$$

where $\boldsymbol{m}_y$ is a location vector, $\boldsymbol{U}_y\in\mathbb{R}^{D\times D}$ has orthonormal columns; $\{\boldsymbol{\psi}_j\}_{1\le j\le n}$ are the coefficients, and $\{\boldsymbol{e}_{y,j}\}_{1\le j\le n}$ zero-mean random variables. The unknowns $\boldsymbol{m}_y$, $\boldsymbol{U}_y$, and $\{\boldsymbol{\psi}_i\}_{1\le j\le n}$ can be estimated using the LS criterion as [15]

$$\min_{\substack{\boldsymbol{m}_y,\ \{\boldsymbol{\psi}_j\}\\ \boldsymbol{U}_y^\top\boldsymbol{U}_y=\boldsymbol{I}}} \ \sum_{j=1}^{n}\|\mathring{\boldsymbol{y}}_j - \boldsymbol{m}_y - \boldsymbol{U}_y\boldsymbol{\psi}_j\|_2^2 . \tag{11}$$

whose solution is given as [15, 10]: $\boldsymbol{m}_y^* := (1/n)\sum_{j=1}^n\mathring{\boldsymbol{y}}_j$, $\boldsymbol{\psi}_j^* := (\boldsymbol{U}_y^*)^\top(\mathring{\boldsymbol{y}}_j - \hat{\boldsymbol{m}}_y)$, $\forall 1\le j\le n$, and $\boldsymbol{U}_y^*$ stacks up as its columns the eigenvectors of $\boldsymbol{C}_{yy} := (1/n)\sum_{j=1}^n\boldsymbol{y}_j\boldsymbol{y}_j^\top$, to form $\boldsymbol{C}_{yy} := \boldsymbol{U}_y^*\boldsymbol{\Sigma}_y(\boldsymbol{U}_y^*)^\top$, where $\boldsymbol{y}_j := \mathring{\boldsymbol{y}}_j - \boldsymbol{m}_y^*$ is the centered data. For notational brevity, the superscript * shall be dropped when clear from the context. Wlog, let $\boldsymbol{U}_y := [\boldsymbol{U}_b\ \boldsymbol{U}_n]$ be partitioned such that $\boldsymbol{U}_b\in\mathbb{R}^{D\times k}$ corresponds to the first k PCs of $\boldsymbol{C}_{yy}$, which capture most background component vectors.

In the context of discriminative data analysis, we *assume* that the target data share some PCs with the background data (say $\boldsymbol{U}_b$ of (10)), and has additionally a few (say d) PCs that capture patterns specific to the target data but are less significant than the PCs in $\boldsymbol{U}_b$. For simplicity, consider $d = 1$, and model $\{\mathring{\boldsymbol{x}}_i\}$ as

$$\mathring{\boldsymbol{x}}_i = \boldsymbol{m}_x + [\boldsymbol{U}_b\ \ \boldsymbol{u}_s]\begin{bmatrix}\boldsymbol{\chi}_{b,i}\\ \chi_{s,i}\end{bmatrix} + \boldsymbol{e}_{x,i}, \quad 1\le i\le m \tag{12}$$

where $\boldsymbol{m}_x$ is the mean of $\{\mathring{\boldsymbol{x}}_i\}_{1\le i\le m}$; and assuming $k + d \le D$, $\boldsymbol{U}_x := [\boldsymbol{U}_b\ \ \boldsymbol{u}_s]\in\mathbb{R}^{D\times(k+1)}$ has orthonormal columns, where $\boldsymbol{U}_b$ describes the component vectors present both in the background as well as target data, while $\boldsymbol{u}_s\in\mathbb{R}^{D\times 1}$ captures the patterns of interest that are present *only* in the target data. Our goal is to obtain this discriminative subspace $\boldsymbol{U}_s$ given solely the two datasets. By modeling this distinctly informative component $\chi_{s,i}\boldsymbol{u}_s$ in (12) explicitly as an outlier vector, it is also possible to employ robust PCA which boils down to solving a nonconvex optimization problem [10].

Likewise, remove the mean $\boldsymbol{m}_x := (1/n)\sum_{i=1}^n\mathring{\boldsymbol{x}}_i$ from the target data yielding $\boldsymbol{x}_i := \mathring{\boldsymbol{x}}_i - \boldsymbol{m}_x$. Recalling $\boldsymbol{C}_{yy}^{1/2} := \boldsymbol{\Sigma}_y^{1/2}\boldsymbol{U}_y$, consider the transformed data model for $1\le i\le m$:

$$\begin{aligned}\boldsymbol{C}_{yy}^{\top/2}\boldsymbol{x}_i &= \boldsymbol{C}_{yy}^{\top/2}[\boldsymbol{U}_b\ \ \boldsymbol{u}_s]\begin{bmatrix}\boldsymbol{\chi}_{b,i}\\ \chi_{s,i}\end{bmatrix} + \boldsymbol{C}_{yy}^{\top/2}\boldsymbol{e}_{x,i}\\ &= \chi_{s,i}\boldsymbol{C}_{yy}^{\top/2}\boldsymbol{u}_s + \boldsymbol{C}_{yy}^{\top/2}\boldsymbol{e}_{x,i} := \chi_{s,i}\tilde{\boldsymbol{u}}_s + \tilde{\boldsymbol{e}}_{x,i}\end{aligned} \tag{13}$$

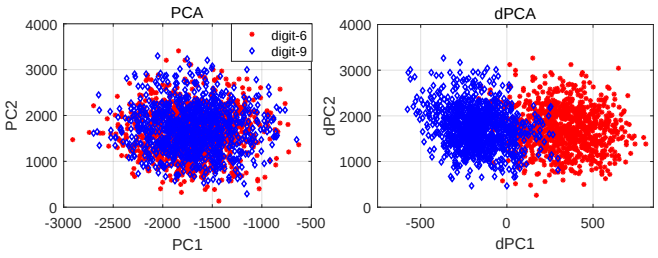

Fig. 1. dPCA versus PCA on semi-synthetic data.

where $\boldsymbol{C}_{yy}^{\top/2}\boldsymbol{U}_b$ vanishes due to the orthogonality of columns of $\boldsymbol{U}_y^* = [\boldsymbol{U}_b\ \boldsymbol{U}_n]$, $\tilde{\boldsymbol{u}}_s := (\boldsymbol{U}_y^*)^\top \boldsymbol{u}_s$, and $\tilde{\boldsymbol{e}}_{x,i}$ is a zero-mean random variable. Similar to (11), the LS optimal estimate $\tilde{\boldsymbol{u}}_s^*$ is given by the first principal eigenvector of

$$\tilde{\boldsymbol{C}}_{xx} := (1/m)\sum_{i=1}^m \boldsymbol{C}_{yy}^{\top/2}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{C}_{yy}^{\top/2} = \boldsymbol{C}_{yy}^{\top/2}\boldsymbol{C}_{xx}\boldsymbol{C}_{yy}^{1/2}.$$

Hence, the discriminative PCs can be recovered from $\tilde{\boldsymbol{u}}_s$ as $\boldsymbol{u}_s^* := \boldsymbol{C}_{yy}^{1/2}\tilde{\boldsymbol{u}}_s^*$, which coincides with solutions of problem (3) or (4), and establishes the LS optimality of dPCA.

5. NUMERICAL TESTS

In this section, the performance of dPCA is assessed relative to PCA and cPCA [4] on a synthetic and a real dataset. In the first experiment, (semi-)synthetic target and background data were generated from real images. Specifically, the target data were constructed using $2,000$ handwritten digits 6 and 9 ($1,000$ for each) of size 28×28 from the MNIST dataset [16] superimposed with $2,000$ frog images from the CIFAR-10 dataset [17]. The raw 32×32 frog images were converted to grayscale and randomly cropped to 28×28. The background data were built with $3,000$ resized images only, which were sampled randomly from the remaining frog images.

We performed PCA on the target data only. The results of the target images embedded onto the first two PCs and two dPCs are depicted in the left and right panels of Fig. 1, respectively. Clearly, PCA is unable to discover the two digit subgroups. This is because the obtained two PCs are likely associated with the background component vectors within the target images, namely features depicting frog images rather than handwritten digits. On the contrary, two clusters emerged in the plot of dPCA, demonstrating its efficacy over PCA in discriminating unique features of one dataset from the other.

The capability of dPCA in discovering subgroups is further tested on real protein expression data. In the second experiment, the target data consist of 267 points, each recording 77 protein expression measurements for a mouse suffering Down Syndrome [18]. There were 135 mice with drug-memantine treatment as well as 134 without treatment. The control data comprise such measurements from 135 healthy mice. The 135 control mice are likely to exhibit similar natural variations (due to e.g., sex and age) as the target mice, but without the differences that result from Down Syndrome. For

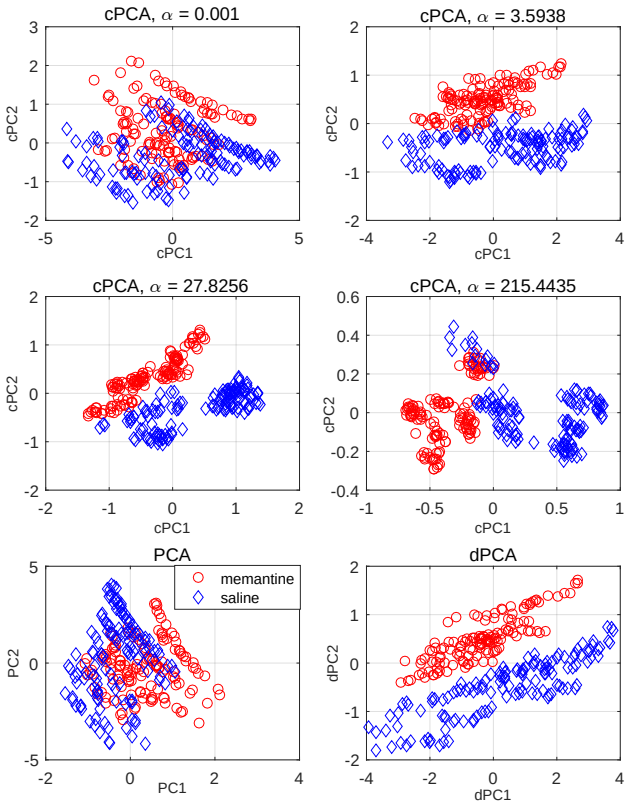

Fig. 2. Discovering subgroups in mice protein expression data.

cPCA, the designed spectral-clustering algorithm was implemented for selecting 4 from a list of 15 logarithmically spaced values between 0.001 and $1,000$ [4]. The simulated results are presented in Fig. 2, with red circles denoting mice with treatment and blue diamonds the other mice. PCA reveals that the two types of mice follow a similar distribution in the space spanned by the first two PCs; see the left bottom plot in Fig. 2. The separation between the two groups of mice becomes clear when dPCA is applied. At the price of runtime (15 times more than dPCA), cPCA with properly *learnt* parameters ($\alpha = 3.5938$ and 27.8256) can work well too.

6. CONCLUSIONS

This paper put forth a novel approach termed dPCA for discriminative analytics, namely for discovering the most informative features that are specific to one dataset but are also distinct from some other correlated datasets. The resultant algorithm amounts to solving a generalized eigenvalue problem. Comparing to existing alternatives, dPCA bypasses parameter tuning, and it incurs complexity required to perform only one SVD. It is provably optimal in the LS sense provided that the background component vectors are present in the target data. Simulated tests using (semi)-synthetic images and real protein expression data corroborated the merits of the developed approach. Investigating dPCA using kernels and over graphs constitutes meaningful future research directions.

7. REFERENCES

[1] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, Jun. 2010.

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

[3] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417–441, Oct. 1933.

[4] A. Abid, V. K. Bagaria, M. J. Zhang, and J. Zou, "Contrastive principal component analysis," *arXiv:1709.06716*, 2017.

[5] S. Garte, "The role of ethnicity in cancer susceptibility gene polymorphisms: The example of CYP1A1," *Carcinogenesis*, vol. 19, no. 8, pp. 1329–1332, Aug. 1998.

[6] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, Mar. 1964.

[7] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[8] B. Scholkopf, A. Smola, and K. B. Muller, *Kernel principal component analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 583–588.

[9] I. D. Schizas and G. B. Giannakis, "Covariance eigenvector sparsity for compression and denoising," *IEEE Trans. Signal Process.*, vol. 60, pp. 2408–2421, May 2012.

[10] G. Mateos and G. B. Giannakis, "Robust PCA as bilinear decomposition with outlier-sparsity regularization," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5176–5190, Oct. 2012.

[11] J. Chen and I. D. Schizas, "Online distributed sparsity-aware canonical correlation analysis," *IEEE Trans. Signal Process.*, vol. 64, no. 3, pp. 688–703, Feb. 2016.

[12] ——, "Distributed information-based clustering of heterogeneous sensor data," *Signal Process.*, vol. 126, pp. 35–51, Sep. 2016.

[13] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936.

[14] Y. Saad, *Iterative Methods for Sparse Linear Systems*. SIAM, 2003.

[15] B. Yang, "Projection approximation subspace tracking," *IEEE Trans. Signal Process.*, vol. 43, no. 1, pp. 95–107, Jan. 1995.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[17] A. Krizhevsky, "Learning multiple layers of features from tiny images," in *Master's thesis*, Department of Computer Science, University of Toronto, 2009.

[18] C. Higuera, K. J. Gardiner, and K. J. Cios, "Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome," *PloS ONE*, vol. 10, no. 6, p. e0129126, Jun. 2015.