Directionally Paired Principal Component Analysis for Bivariate Estimation Problems

Yifei Fan*, Navdeep Dahiya*, Samuel Bignardi*, Romeil Sandhu†, and Anthony Yezzi*
*School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30308, USA
† Computer Science Department, Stony Brook University, Stony Brook, NY 11794, USA

***Abstract*—We propose Directionally Paired Principal Component Analysis (DP-PCA), a novel linear dimension-reduction model for estimating coupled yet partially observable variable sets. Unlike partial least squares methods (e.g., partial least squares regression and canonical correlation analysis) that maximize correlation/covariance between the two datasets, our DP-PCA directly minimizes, either conditionally or unconditionally, the reconstruction and prediction errors for the observable and unobservable part, respectively. We demonstrate the optimality of the proposed DP-PCA approach, we compare and evaluate relevant linear cross-decomposition methods with data reconstruction and prediction experiments on synthetic Gaussian data, multi-target regression datasets, and a single-channel image dataset. Results show that when only a single pair of bases is allowed, the conditional DP-PCA achieves the lowest reconstruction error on the observable part and the total variable sets as a whole; meanwhile, the unconditional DP-PCA reaches the lowest prediction errors on the unobservable part. When an extra budget is allowed for the observable part's PCA basis, one can reach an optimal solution using a combined method: standard PCA for the observable part and unconditional DP-PCA for the unobservable part.**

I. Introduction

The dimension of useful features in data is generally much smaller than that of the data themselves, which implies that plenty of redundancy or irrelevant information exists in data samples. Such a redundancy or irrelevance in data samples often leads to unnecessary complexity and issues such as the "curse of dimensionality" [1]. To mitigate those issues, various dimension reduction approaches are invented to reduce the number of variables under consideration by obtaining a set of principal variables, among which Principal Component Analysis (PCA) [2] is the most widely used one.

In principle, PCA handles a single set of variables (i.e., measurements) by maximizing the variance along the principal components or equivalently minimizing the reconstruction errors. In certain scenarios, we may have more than one set of correlated samples. For those situations, Canonical Correlation Analysis (CCA) [3], together with its general framework Partial Least Squares (PLS) [4] methods, measures the linear correlation between two multi-dimensional variables by seeking a pair of bases such that the corresponding variables expressed in those bases are maximally correlated. Changing the objective to maximizing correlation, however, leads to sub-optimal bases in terms of maximizing variance and minimizing reconstruction errors for each respective set.

Our research on Directionally Paired Principal Component Analysis (DP-PCA) originates from a use case in which we have access to a pair of correlated datasets at training but can observe only one of them at test time. Henceforth, the two datasets are referred to as the observable and unobservable. The goal of the special use case is to conduct dimension reduction for both the observable and unobservable variables and provide high-quality reconstruction/predictions at test time. In cases of both datasets being observable, we could develop two independent PCA models for each, by which we might ignore the correlation between the two sets and perhaps adopt higher dimensional representation than necessary. One naïve version that considers the correlation between the two sets is called joint PCA, in which we stack both sets of samples and extract a single set of principal components. Such a technique forces the model to learn the correlations between the two while keeping the dimensionality lower than the sum of two independent PCA models. In joint PCA, reconstruction for neither variable set is optimal because the model's capacity is split between the two variable sets. Unfortunately, when one of the variable sets becomes unobservable, the budget and efforts spent on those variables will be in vain.

To overcome the above issue and achieve our goal for the special use case, we derive DP-PCA, which minimizes reconstruction/prediction errors of the coupled datasets by minimizing the least-squares errors. In the unconditional situation, we pursue optimal prediction for the unobservable variables regardless of the reconstruction of the observable ones. In the conditional situation, we aim at the best possible prediction for the unobservable under the premise that the observable is optimally preserved. The major contribution of the paper, therefore, is the proposed DP-PCA, a linear dimension-reduction model for predicting the unobservable part in a coupled variable sets by minimizing the least-squares estimation errors, both unconditionally and conditionally on the optimal reconstruction of the observable. In another concurrent paper, we explore the usage of the proposed DP-PCA framework in the context of inversion problems.

The rest of the paper is organized as follows. Section II presents the proposed DP-PCA approach. Section III reviews related linear models for dimension reduction and estimation

This work was funded in part by National Institutes of Health (NIH) grant R01 HL143350, Army Research Office grant ARO W911NF-18-1-0281, U.S. Air Force Office of Scientific Research (AFOSR) grant FA9550-18-1-0130 and National Science Foundation (NSF) grant ECCS-1749937.

of coupled yet partially observable data, and address the connections to the proposed DP-PCA. In Section IV, we evaluate relevant dimension reduction and estimation methods using data reconstruction and prediction experiments. Finally, Section V concludes with discussions.

II. Proposed Directionally Paired PCA

In this section, we derive the Directionally Paired Principal Component Analysis (DP-PCA) approach that best predicts the unobservable from the observable through the linear dimensional reduction in both unconditional and conditional scenarios. To begin with, we establish a notation system that we will adhere to for the rest of the paper.

Let us assume that an $M_1 \times N$ matrix $X = [\mathbf{x}_1\, \mathbf{x}_2 \cdots \mathbf{x}_N]$ and an $M_2 \times N$ matrix $Y = [\mathbf{y}_1\, \mathbf{y}_2 \cdots \mathbf{y}_N]$ contain the separate collected components of pairs from N data measurements represented as vectors in $\mathbb{R}^{M_1}$ and $\mathbb{R}^{M_2}$. In particular, $\mathbf{x}_i$ and $\mathbf{y}_i$ represent the observable (high-confidence) and unobservable (low-confidence) components of the i^{th} data measurement, respectively. We further assume that the mean values of both sets of measurements are zero. If such a condition is not satisfied, the respective means should be pre-subtracted from each $\mathbf{x}_i$ and $\mathbf{y}_i$ for $i = 1, \cdots, N$. In addition, let $M_1 \times L$ matrix $U = [\mathbf{u}_1 \cdots \mathbf{u}_L]$ and $M_2 \times L$ matrix $V = [\mathbf{u}_1 \cdots \mathbf{u}_L]$ denote the bases of dimension reduction for X and Y, respectively. As a result of the dimension reduction, we obtain $L \times N$ matrix $A = [\mathbf{a}_1 \cdots \mathbf{a}_N]$ and $B = [\mathbf{b}_1 \cdots \mathbf{b}_N]$ as the expansion coefficients (i.e., low-dimensional representation) of X and Y, respectively.

A. Optimal Estimation for the Unobservable Part

Inspired by the strategy in Partial Least Squares (PLS) methods [4] for predicting the unobservable part Y from the observable part X, we can derive the optimal pair of bases U, V that directly minimizes the reconstruction error ε_Y for the unobservable part Y, regardless of any correlation between the two variable sets X and Y. Henceforth, we refer to such estimation as the *optimal-Y* (i.e., unconditional) mode of the proposed DP-PCA. We start with the loss function on U, V:

$$\begin{aligned}\varepsilon_Y(U,V) &= \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{y}_n - VU^T\mathbf{x}_n\|^2 \\ &= \frac{1}{N}\sum_{n=1}^{N} \mathbf{y}_n^T\mathbf{y}_n - 2\mathbf{x}_n^T UV^T\mathbf{y}_n + \mathbf{x}_n^T UV^T VU^T\mathbf{x}_n \end{aligned} \quad (1)$$

Differentiating ε_Y with respect to U and V, we obtain:

$$\begin{aligned}\frac{\partial \varepsilon_Y}{\partial U} &= \frac{1}{N}\sum_{n=1}^{N} -2\mathbf{x}_n\left(V^T\mathbf{y}_n\right)^T + \left(2V^TVU^T\mathbf{x}_n\mathbf{x}_n^T\right)^T \\ &= -\frac{2}{N}\sum_{n=1}^{N} \mathbf{x}_n\mathbf{y}_n^TV - \mathbf{x}_n\mathbf{x}_n^TUV^TV \\ &= -\frac{2}{N}\left(XY^TV - XX^TUV^TV\right) \end{aligned} \quad (2)$$

and

$$\begin{aligned}\frac{\partial \varepsilon_Y}{\partial V} &= \frac{1}{N}\sum_{n=1}^{N} -2\mathbf{y}_n\mathbf{x}_n^TU + 2VU^T\mathbf{x}_n\mathbf{x}_n^TU \\ &= -\frac{2}{N}\left(YX^TU - VU^TXX^TU\right) \end{aligned} \quad (3)$$

Setting both matrix derivatives to zero yields the conditions:

$$\begin{cases} XY^TV = XX^TUV^TV \\ YX^TU = VU^TXX^TU \end{cases} \quad \text{or} \quad \begin{cases} XX^TU = XY^TV\left(V^TV\right)^{-1} \\ V = YX^TU\left(U^TXX^TU\right)^{-1} \end{cases} \quad (4)$$

In certain situations, we would hope to estimate the unobservable part under the premise that the observable part is optimally preserved. Thus, only the condition in equation (3) is taken account because U is pre-determined for ε_Y. We refer to such a mode as conditional DP-PCA. In reality, we first derive the conditional version for inversion problems; thus, we still call it DP-PCA and refer to the unconditional version as the "optimal-Y" mode or unconditional DP-PCA. The usage of DP-PCA in inversion problems is documented in our concurrent paper titled "Dependently Coupled Principal Component Analysis for Bivariate Inversion Problems.

B. Closed-form Solution

The solution to the conditional DP-PCA is straight-forward: one should compute the basis U for the observable part via standard PCA and apply the bottom-right formula in equation (4) to obtain the paired basis V. For the rest part of the section, we derive a closed-form solution for the unconditional DP-PCA.

If we substitute V in the top left equation of (4) with $YX^TU\left(U^TXX^TU\right)^{-1}$ as specified by the bottom right equation (thus, $V^T = (U^TXX^TU)^{-1}U^TXY^T$), we obtain:

$$\begin{aligned}&XY^TYX^TU\left(U^TXX^TU\right)^{-1} = \\ &XX^TU\left(U^TXX^TU\right)^{-1}U^TXY^TYX^TU\left(U^TXX^TU\right)^{-1}\end{aligned} \quad (5)$$

which can be further simplified by right-multiplying U^TXX^TU on both sides:

$$XY^TYX^TU = XX^TU\left(U^TXX^TU\right)^{-1}U^TXY^TYX^TU \quad (6)$$

Therefore, the optimal pair of bases U, V which minimizes the reconstruction error of the unobservable part Y can be computed by solving the above equation (6) for U and plugging the solution of U into the bottom right equation in (4) for V.

Notice that if we allow any solution (not necessarily orthogonal matrices) to equation (6), then it is easy to show by direct substitution that UW also solves the equation as long as W is any invertible $L \times L$ matrix.

$$\begin{aligned}&XY^TYX^TUW = XX^TU\left(U^TXX^TU\right)^{-1}U^TXY^TYX^TUW \\ =&XX^TUWW^{-1}\left(U^TXX^TU\right)^{-1}(W^T)^{-1}W^TU^TXY^TYX^TUW \\ =&XX^TUW\left(W^TU^TXX^TUW\right)^{-1}W^TU^TXY^TYX^TUW\end{aligned} \quad (7)$$

Thus, equation (6) depends only upon the column space of U. We may exploit such a property and seek a solution to the following simpler equation using any convenient choice of W:

$$XY^TYX^TU = XX^TUW \tag{8}$$

which has the exact same set of solutions. One sufficient yet not necessary condition of the above equation (8) is the following:

$$Y^TYX^TU = X^TUW \tag{9}$$

Now, if we perform an eigenvalue decomposition on the $N \times N$ matrix Y^TY and select L of the eigenvectors (with non-zero eigenvalues) to form the columns of an $N \times L$ matrix Z, and store their associated eigenvalues in an $L \times L$ diagonal matrix D, then we may write

$$Y^TYZ = ZD \tag{10}$$

Finally, solving

$$X^TU = Z \tag{11}$$

for U and setting $W = D$, we obtain a solution to equation (9) and hence a solution to equation (6).

Technically, the proposed DP-PCA can be described as a least-squares regression analysis using a low-dimensional (i.e., principal) subspace. The abbreviation of the principal least-squares regression, however, coincides with the existing partial least-squares regression (PLSR). Therefore, we name the method as DP-PCA, which also reflects the directional prediction in the subspace of principal components.

The asymmetry of the proposed method is also reflected in the solutions. For (conditional) DP-PCA, we first solve the eigenvalue problem on XX^T in standard PCA on the observable part X, then minimize the least-squares prediction error by setting the partial derivative of paired basis V to zero. For the unconditional ("optimal-Y" mode) DP-PCA, we begin by solving another eigenvalue problem on Y^TY (not YY^T, but their eigenvalues and eigenvectors are closely related), then solve a linear equation by minimizing the least-squares error.

III. Related Work

In this section, we review relevant linear approaches for the estimation of coupled data and elaborate the connections to the proposed Directionally Paired PCA.

A. Canonical Correlation Analysis (CCA)

A well-known yet symmetric method that also produces a paired set of bases for a correlated pair of variables is Canonical Correlation Analysis (CCA). First introduced in [3], CCA manages to measure the linear relationship between two multi-dimensional variables $\mathbf{x} = (x_1, \cdots, x_{m_1})^T$ and $\mathbf{y} = (y_1, \cdots, y_{m_2})^T$. It seeks a pair of vectors $a \in \mathbb{R}^{m_1}$ and $b \in \mathbb{R}^{m_2}$ such that the linear combinations $u = a^T\mathbf{x} = \sum_{i=0}^{m_1} a_i x_i$ and $v = b^T\mathbf{y} = \sum_{j=0}^{m_2} b_j y_j$ maximize the *correlation* $\rho = \operatorname{corr}\left(a^T\mathbf{x}, b^T\mathbf{y}\right)$. The random variables u and v in $\mathbb{R}$ are called the first pair of canonical variates. The methods then iteratively seeks pairs of canonical variates that maximize the above correlation, subjecting to the constraint(s) that the new canonical variates shall be uncorrelated with previous canonical variates. The entire process can take up to $\min(m_1, m_2)$ iterations, and correlations between the canonical variates u and v indicate correlations among the terms $a_i x_i$ and $b_j y_j$.

In practice, CCA is applied primarily for modeling and correlation analysis, which tends to overfit data when it comes to reconstruction and prediction. Its customized version, namely Canonical Regression (CR) [5] [6], performs additional regression analysis in the low-dimensional subspaces, which enables prediction of $\mathbf{y}$ from $\mathbf{x}$.

B. Partial Least Squares Regression (PLSR)

The CCA approach can be considered the special mode "B" of a more general framework named Partial Least Squares (PLS) methods [4]. The equivalence between CCA and orthonormalized PLS is further addressed in [7]. One slight difference between Partial Least Squares Regression (PLSR, mode "A") and CCA is that instead of maximizing the correlation between $\mathbf{X}$ and $\mathbf{Y}$ as in CCA, the objective function for maximization in PLSR becomes the *covariance* between X and Y [8].

The PLSR approach has an "opposite" motivation compared with the proposed conditional DP-PCA. In conditional DP-PCA, we assume that the predictor X remains observable with high confidence at all times and hope to predict the values of unobservable low-confidence variables Y at test time with the help of the correlation under the premise that X is optimally preserved. On the contrary, the goal of PLSR is to predict as accurately as possible the values of the important predictands Y (which might be expensive to capture) based on the less important predictors X (which are cheaper to capture) by utilizing the covariance between the two. It is, therefore, acceptable that the predictors X may not be optimally reconstructed when necessary.

Mathematically, both CCA and PLSR can be formulated as solving eigenvalue equations with slightly different matrix coefficients [8] [9]:

$$\mathbf{B}^{-1}\mathbf{A}\hat{\mathbf{w}} = \rho\hat{\mathbf{w}} \tag{12}$$

in which for CCA

$$\mathbf{A} = \begin{bmatrix} 0 & S_{xy} \\ S_{yx} & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} S_{xx} & 0 \\ 0 & S_{yy} \end{bmatrix}, \tag{13}$$

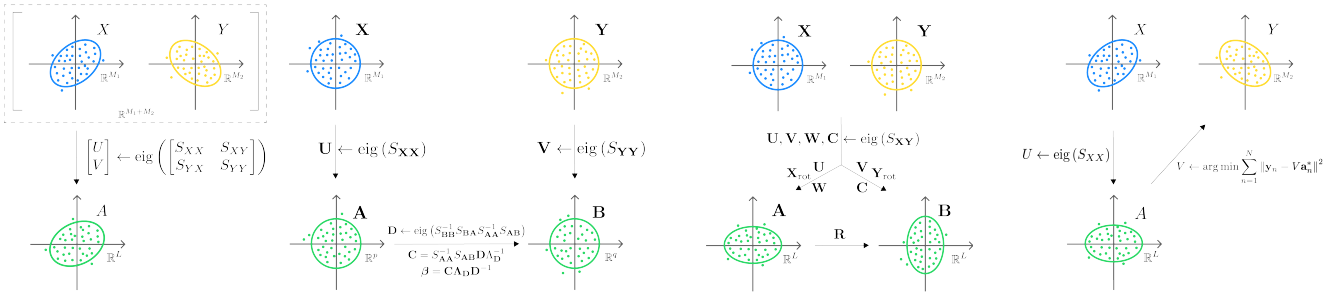
and for PLSR

$$\mathbf{A} = \begin{bmatrix} 0 & S_{xy} \\ S_{yx} & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{I} \end{bmatrix}, \tag{14}$$

In the above equations, $S_{\cdot\cdot}$ are covariance matrices and $\mathbf{I}$ stands for an identity block.

C. Correlation Analysis for Coupled Data

We can now compare the correlation analysis in relevant approaches with the help of Fig. 1. Joint PCA maximizes the full covariance matrix and obtains a concatenated basis, which is split into U, V (Fig. 1a). Different from the original CCA whose goal is to maximize the correlation $\operatorname{corr}(\mathbf{X}, \mathbf{Y})$ between the two sets of variables, the customized Canonical Regression



(a) Joint PCA (b) Canonical Regression (c) PLSR (d) conditional DP-PCA

Fig. 1: Comparison on correlation analysis in related approaches.

(CR) first project both sets of standardized (i.e., subtracted by mean and divided by standard deviation) variables into the lower-dimensional subspace such that their own variance is maximally captured. The regression coefficient is then computed by maximizing the covariance terms constructed by the paired standardized data after dimension reduction (Fig. 1b). As for PLSR, the loadings $\mathbf{U}, \mathbf{V}$ and weights $\mathbf{W}, \mathbf{C}$ for transforms are estimated by maximizing the covariance $\mathrm{cov}(\mathbf{X}, \mathbf{Y})$, after which a regression analysis is performed between the scores (Fig. 1c). Finally, in the proposed conditional DP-PCA, the correlation between X and Y is maximized by sharing the expansion coefficients A in the lower-dimensional subspace (Fig. 1d).

IV. Evaluation of Dimension Reduction via Reconstruction and Prediction

In this section, we evaluate the performance of relevant dimension reduction approaches via data reconstruction/prediction experiments. We assume that better dimension reduction methods are more capable of capturing the principal components of the data and lead to lower reconstruction/prediction errors. The lower bound of the reconstruction error is provided by independent PCA under the premise that both groups of coupled data X and Y are observable at all times (i.e., not practical). In reality, the paired data are available only during training, and we no longer have access to the unobservable variables Y at test time. Thus, the experiment's goal with involved dimension-reduction methods is to both reconstruct the observable variables X and predict the unobservable variables Y using the bases U, V learned during training.

A. Experiment Design and Procedures

It should be highlighted that the purpose of the experiment is to evaluate the relevant algorithms in terms of dimension reduction rather than data reconstruction or prediction. At first glance, one may argue that the goal of the experiment is also achievable via auto-encoders [10] (i.e., for reconstructing the observable variables X) and neural-network regressors (i.e., for predicting the unobservable variables Y as multi-target regression). Those methods, however, fail to provide dimension reduction and correlation analysis in a similar manner as the involved algorithms do. In particular, correlation analysis is missing in auto-encoders, and neural-network regressors are not suitable for dimension reduction. Therefore, we consider the following linear approaches in our experiments: independent (standard) PCA,[1] joint PCA,[1] Partial Least Squares Regression[2] (PLSR) [4], Canonical Regression[3] (CR) [5], [6], and the proposed DP-PCA. Among the approaches, the independent PCA serves as the lower bound of reconstruction error, and both X, Y remain observable at all times. The Canonical Regression (CR) approach is a customized version of the Canonical Correlation Analysis (CCA), which supports predicting the values of Y from X.

Following the notation defined at the beginning of II, M_1 and M_2 denote the dimensions of observable (high-confidence) and unobservable (low-confidence) variables in a data sample, respectively. For each method, we compute the paired bases (i.e., loadings) U, V with the training set, reducing the dimensions of $X_{\text{train}}, Y_{\text{train}}$ from M_1, M_2 to L. We then apply the corresponding basis U or rotations $\mathbf{X}_{\text{rot}}$ for dimension reduction to the observable test data[4] X_{test} and obtain the dimension-reduced data A_{test} with dimension L. The observable part of the reconstructed test data $\hat{X}_{\text{test}}$ is obtained by taking the inverse transform of the dimension reduction with its corresponding basis U or loadings $\mathbf{U}$.

The prediction of the unobservable part $\hat{Y}_{\text{test}}$ is handled differently in the involved approaches. For independent and joint PCA, the basis V characterizes a transformation between the unobservable variables Y and their corresponding dimension-reduced expansion coefficients (i.e., scores) B. The only difference between independent and joint PCA is that we assume Y_{test} remains available for computing B_{test} in independent PCA whereas B_{test} is replaced by A_{test} in joint PCA because only X_{test} is available. To predict the values of the unobservable part $\hat{Y}_{\text{test}}$, both independent PCA and joint PCA take the inverse transform of B_{test}. For PLSR, CR, and the proposed DP-PCA, the basis/loadings V characterizes a transformation from the expansion coefficients A (i.e., scores)

[1] scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA

[2] scikit-learn.org/stable/modules/generated/sklearn.cross_decomposition.PLSRegression.html

[3] in R: https://rdrr.io/github/jmhewitt/telefit/man/cca.predict.html

[4] Applying the basis U to X_{train} leads to expansion coefficient A.

TABLE I: Storage requirement on dimension reduction approaches for coupled data.

Method	Results to be stored after training
JPCA	$\bar{X}, \bar{Y}$: mean values of training data (size M_1 and M_2); U, V: bases for X and Y (size $M_1 \times L$ and $M_2 \times L$).
PLSR	$\bar{\mathbf{X}}, \bar{\mathbf{Y}}$: mean values of training data (size M_1 and M_2); $\sigma_{\mathbf{X}}, \sigma_{\mathbf{Y}}$: std. of training data (size M_1 and M_2); $\mathbf{U}, \mathbf{X}_{\text{rot}}$: loadings and rotations for $\mathbf{X}$ (both size $M_1 \times L$); One of the following: (1) $\mathbf{V}$: loadings for $\mathbf{Y}$ (size $M_2 \times L$) and $\mathbf{R}$: regression matrix between $\mathbf{A}$ and $\mathbf{B}$ (size $L \times L$), (2) $\boldsymbol{\beta} = \mathbf{VR}$: regression coefficients (size $M_2 \times M_1$).
CR	$\bar{\mathbf{X}}, \bar{\mathbf{Y}}$: mean values of training data (size M_1 and M_2); $\sigma_{\mathbf{X}}, \sigma_{\mathbf{Y}}$: std. of training data (size M_1 and M_2); $\mathbf{U}, \mathbf{V}$: bases for $\mathbf{X}$ and $\mathbf{Y}$ (size $M_1 \times L$ and $M_2 \times L$); $\bar{\mathbf{A}}, \bar{\mathbf{B}}$: mean values in the subspace (size L); $\sigma_{\mathbf{A}}, \sigma_{\mathbf{B}}$: standard deviation in the subspace (size L); $\boldsymbol{\beta}$: correlation coefficient between $\mathbf{A}$ and $\mathbf{B}$ (size $L \times L$).
DP-PCA	same as those in the J(oint) PCA

of the observable part to the unobservable data Y. To predict the values of the unobservable test data $\hat{Y}_{\text{test}}$, those three approaches apply a prediction transform to A_{test}. Based on the reconstructed and predicted values of the test data, we finally compute the mean squared error per element between $\{X_{\text{test}}, Y_{\text{test}}\}$ and $\{\hat{X}_{\text{test}}, \hat{Y}_{\text{test}}\}$.

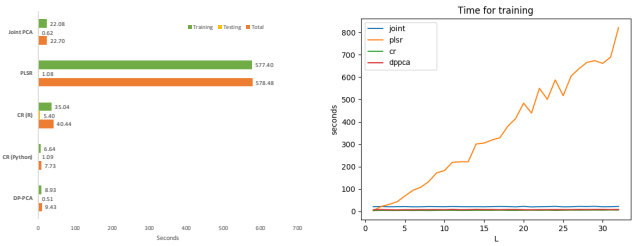
Table I provides a list of the storage requirement for each approach to facilitate the reconstruction and prediction process. The data structures are consistent with the publicly available implementations discussed above. Compared with baseline methods PLSR and CR, the proposed DP-PCA requires minimal storage that remains the same as joint PCA. In the following subsection, we introduce the benchmark datasets for our experiments and report execution time for those methods.

B. Benchmark Datasets and Execution Time

1) Datasets: We conduct experiments on three types of datasets: *synthetic*, *multi-target regression*, and *single-channel image* data. The details of those datasets elaborated as follows.

Synthetic data: We generate a data matrix D of size $(M_1 + M_2) \times N$, containing N data measurements from a multivariate Gaussian distribution with random mean $\mu_{M_1+M_2}$ and random covariance matrix $\Sigma_{M_1+M_2}$. We then split the rows of D into the observable part X with size $M_1 \times N$ and unobservable part Y with size $M_2 \times N$. Thus, the correlation between the two parts are established via the covariance matrix $\Sigma_{M_1+M_2}$. By keeping 70% samples for training and the rest for testing, the N data samples are further divided into training set $\{X_{\text{train}}, Y_{\text{train}}\}$ and test set $\{X_{\text{test}}, Y_{\text{test}}\}$. Following the procedures illustrated in the previous subsection, the training set is used for computing bases U, V and other required results listed in Table I, whereas the test set is reserved for computing reconstruction (for X_{test}) and prediction (for Y_{test}) errors.

Multi-target regression data: As discussed at the beginning of Chapter IV-A, predicting the values of the unobservable variables Y can be formulated as a multi-target regression problem. In multi-target regression datasets, the observable variables X are called "features" while the unobservable variables Y are considered "targets." Among all 18 datasets in [11], we select 4 of them which satisfy the following two conditions. (1) the dimensions of both X and Y are larger than 10 so that there is room for varying the dimension L of the subspace, and (2) no missing values exist in the data.



(a) Execution time for 100 runs ($M_1 = M_2 = 128$, $L = 32$)

(b) Training time for 100 runs ($M_1 = M_2 = 128$, $L = 1$ to 32)

Fig. 2: Comparison on execution time on the synthetic dataset.

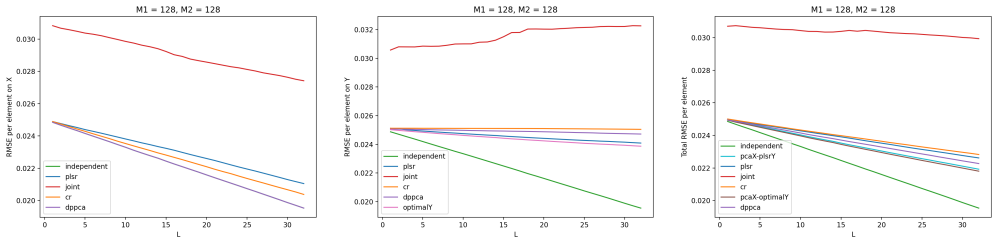
Single-channel image data: We further repeat the experiments on MNIST [12], which are real datasets with larger dimensions than those of the multi-target regression datasets. Pixels of each image are split into two halves as observable and unobservable either (1) according to the sequence of indices (i.e., sequential split) or (2) randomly yet consistently across images (i.e., random split). In addition to measuring the reconstruction/prediction errors, we also evaluate the classification accuracy of the reconstructed data using a pre-trained classifier.

2) Execution time: We report the execution time of each method for 100 runs on the synthetic dataset in Fig. 2. The execution time is benchmarked on a Ubuntu 16.04 Desktop with 8-core Intel Core i7-6700K CPU @ 4.00GHz and 16GB DDR4 RAM @ 2133MHz. In a strict sense, the reported execution time does not necessarily demonstrate the time complexity of the approaches because they are not optimized uniformly. Instead, the chart in Fig. 2a reflects the experience with popular implementations that are publicly available. According to the chart, the required training time for PLSR is substantially longer than others. Besides, as illustrated by Fig. 2b, the training time in PLSR also increases significantly as the budget (i.e., the dimension of the target subspaces) increases. As for the testing time, all approaches have testing time fluctuated within a small range. We also compare the execution time for CR between the original R implementation and our translated version[5] in Python, and find out that the Python version is about 5 times faster. In sum, our python implementation of DP-PCA is relatively faster than its open-source competitors.

C. Experiment Results

Fig. 3 illustrates the reconstruction and prediction errors on the synthetic multi-variate Gaussian data using involved

[5]https://gist.github.com/thelittlekid/89630241f5b90a838a7b583a5836d350



(a) Reconstruction error on X_{test}

(b) Prediction error on unobservable Y_{test}

(c) Total error on $\{X_{\text{test}}, Y_{\text{test}}\}$

Fig. 3: Evaluation on dimension reduction via data reconstruction and prediction of coupled synthetic data. $N = 10^4$, $M_1 = M_2 = 128$. Horizontal axis: dimension L of the target subspace (i.e., budget); vertical axis: reconstruction/prediction error.

approaches. During training, joint PCA shares the budget on the dimension of the subspace L and between the observable part X and unobservable part Y. Consequently, the principal components of X are no longer optimal, and the effort spent on Y is wasted because Y becomes unobservable at test time. Such results motivate us to design DP-PCA that accounts for both the correlation between X and Y, as well as optimally preserving the observable part X. Thus, the reconstruction error on the observable part X remains optimal for DP-PCA and meets the lower bound given by the independent PCA.[6] As the objective of CCA (i.e., CR) is to maximize the correlation between X and Y, neither bases are optimal in terms of preserving information for data reconstruction. By design, PLSR aims at predicting the unobservable part Y by leveraging the covariance between the two variable sets X and Y, leading to more precise predictions on Y yet larger distortion on X. If we disregard the correlation and target at the best prediction with the optimal-Y mode in unconditional DP-PCA, we can further push higher the accuracy on the unobservable Y at the cost of huge reconstruction errors on the observable X, which are orders of magnitude larger and not suitable for plotting with other methods in Fig. 3a.

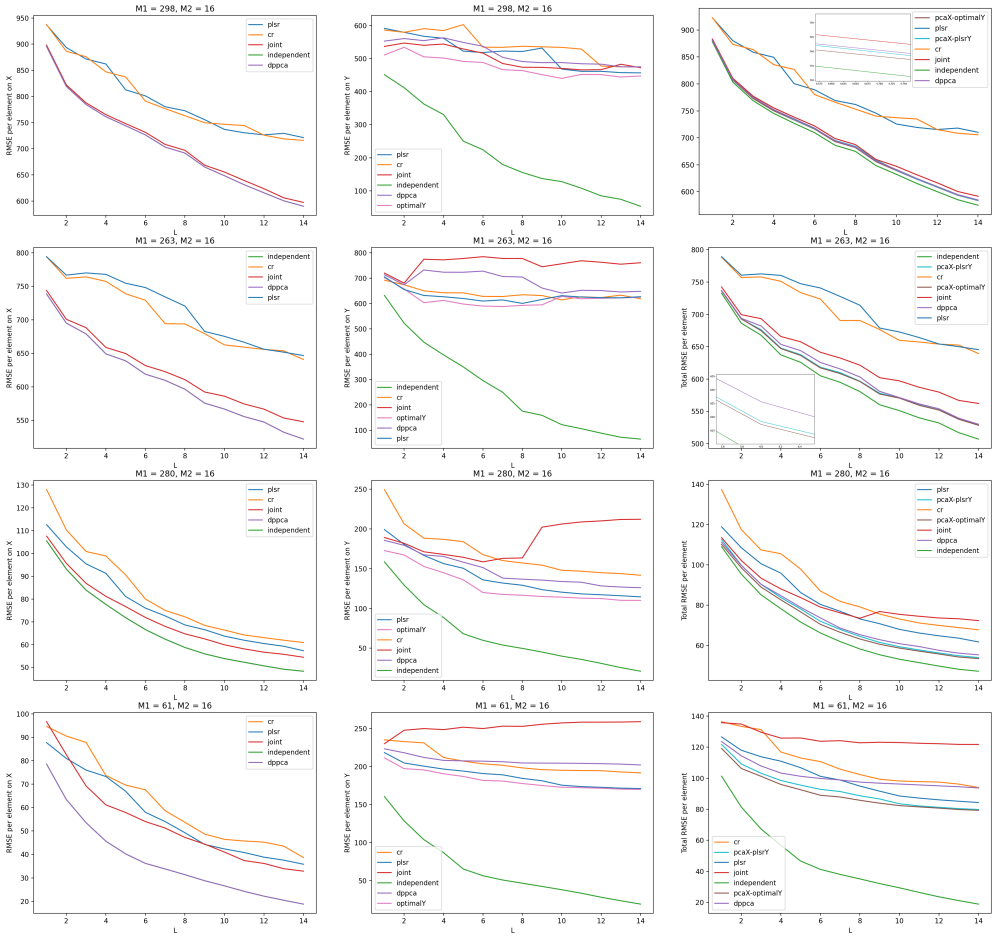
Overall, when combining both observable and unobservable variables (Fig. 3c), the proposed DP-PCA achieves the lowest errors under the condition that only one pair of bases are allowed. With an additional budget for the PCA basis of the observable part X, we can fully enjoy the benefits from the optimal prediction on Y using a combined method: applying standard independent PCA for the observable part X and DP-PCA in mode optimal-Y for the unobservable part Y (i.e., "pcaX-optimalY"). Theoretically, such a combination is the best possible linear model for estimating coupled data.

Fig. 4 shows the reconstruction and prediction errors on the 4 selected multi-target regression datasets. Despite slight differences, results on real data exhibit a similar pattern as those on the synthetic data in Fig. 3. On real data, the simplest joint PCA does not necessarily lead to the largest errors, even though the proposed DP-PCA consistently outperforms it. The pursue of maximum correlation from CR leads to sub-optimal reconstruction on both the observable and unobservable variables. With PLSR and the optimal Y mode of DP-PCA, one may achieve lower prediction errors on the unobservable part Y. When it comes to the total reconstruction error, however, the proposed DP-PCA beats others with single pair of bases most of the time (with a few exceptions on the 4^{th} row scm20d dataset with large budgets L, but we are more interested in lower-budget scenarios for dimension reduction). The combined method of "pcaX-optimalY" remains to be the best linear solution on real data. In Fig. 5, we further repeat the experiments on MNIST [12], which is a real dataset with larger input dimension. Results on MNIST agree with the previous ones; when pixels are randomly split, the proposed DP-PCA even obtains the lowest errors on both the observable and unobservable variables (bottom row of Fig. 5).

Finally, to demonstrate the effectiveness in capturing principal components of the signals, we classify the reconstructed images using a pre-trained classifier, which was trained on clean, complete samples. The experiment corresponds to a use case in which half of the pixels (i.e., Y_{test}) are unobservable at test time. Similar to Fig. 5, the pixels either missing sequentially (i.e., the bottom half of the image) or random yet consistently across images. Moreover, the other observable half (i.e., X_{test}) is interfered by zero-mean Gaussian noise with $\sigma = 0.3$ (in an intensity scale of $[0, 1]$). The input images to the classifier are built with two different modes: mixture or reconstruction. In the mixture mode, we combine the available X_{test} with the predicted $\hat{Y}_{\text{test}}$ whereas in the reconstruction mode, we integrate the reconstructed signal for the observable part $\hat{X}_{\text{test}}$ and the predicted $\hat{Y}_{\text{test}}$. Models for linear reconstruction and prediction are trained on 10,000 randomly selected samples from the original training set. In contrast to retraining the classifier, the ground-truth labels are no longer required for training those linear models.

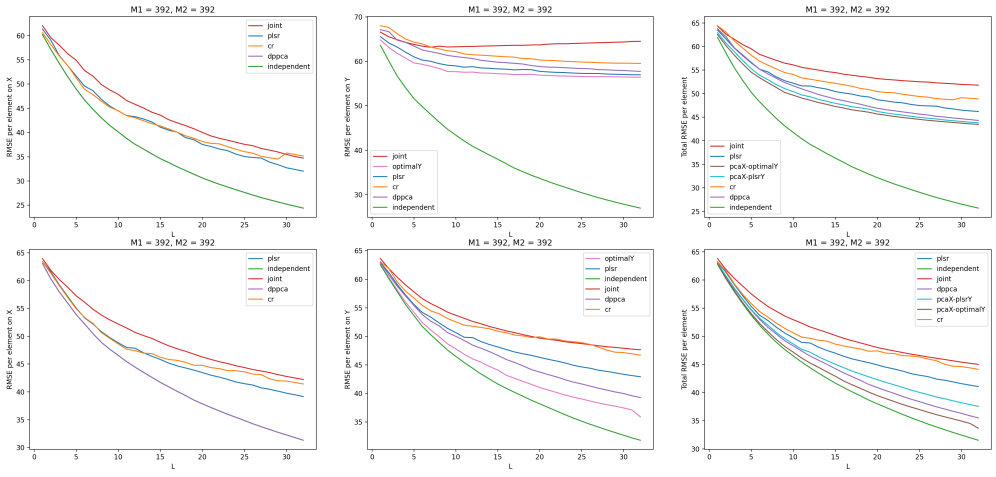
Fig. 6 illustrate the classification accuracy on reconstructed signals of MNIST images. Under low budgets (i.e., small L), mixing the predicted unobservable $\hat{Y}_{\text{test}}$ from unconditional (i.e., optimalY) DP-PCA with observable X_{test} leads to highest accuracy. As the budget becomes sufficiently large, replacing the observable X_{test} with the reconstructed version $\hat{X}_{\text{test}}$ results in higher accuracy. It is worth mentioning that even though

[6]The proposed (unconditional) DP-PCA adopts standard PCA for the observable part X; thus, the two curves (i.e., independent and dppca) in Figs 3a, 4a, and 5a overlap.



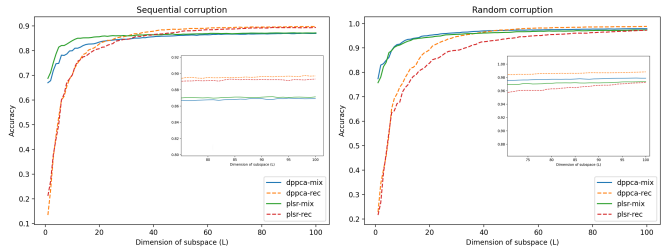
(a) Reconstruction error on X_{test} (b) Prediction error on unobservable Y_{test} (c) Total error on $\{X_{\text{test}}, Y_{\text{test}}\}$

Fig. 4: Evaluation on dimension reduction via data reconstruction and prediction of real multi-target regression datasets [13], [14]: (top to bottom) oes10, oes97, scm1d, and scm20d. Horizontal axis: dimension L of the target subspace (i.e., budget); vertical axis: reconstruction/prediction error.



(a) Reconstruction error on X_{test} (b) Prediction error on unobservable Y_{test} (c) Total error on $\{X_{\text{test}}, Y_{\text{test}}\}$

Fig. 5: Evaluation on dimension reduction via data reconstruction and prediction of MNIST. $L = 1$ to 32, $M_1 = M_2$ (equal split of variables). Top row: sequential split; bottom row: random split.



(a) sequential missing (b) random-consistent missing

Fig. 6: Classification accuracy on partially observable and noisy MNIST after reconstruction and prediction: half of the pixels are missing and the other half noisy at test time.

that conditional DP-PCA obtain higher prediction errors on $\hat{Y}_{\text{test}}$ than PLSR, it achieves lower errors both on $\hat{X}_{\text{test}}$ and in total, thus leading to the highest classification accuracy under a larger budget.

D. Result Analysis

We now analyze the results in terms of degrees of freedom (i.e., budgets) in the optimization process. The degrees of freedom characterizes the number of free variables to be optimized in each method. For independent, joint, and the proposed DP- PCA, the budget equals the total number of variables in U and V, that is, $(M_1 + M_2) \times L$. When it comes to PLSR and CR, an additional budget of L^2 is introduced to learn the mapping between the paired data after dimension reduction. In an ideal case such as two independent PCAs, the degrees of freedom are proportionally split between the observable part X and unobservable part Y at a ratio of $M_1 : M_2$, and each part of the budget is optimally spent to minimize the reconstruction errors, respectively. When one set of variables Y becomes unobservable, the corresponding part of the budget is dissipated while the other part for X is utilized in a sub-optimal manner to also take into account the correlation between the two sets. The proposed DP-PCA ensures that the budget spent on the observable part is utilized optimally such that it maximally captures the variance and minimizes the reconstruction error. The other part of the budget on unobservable Y, moreover, is consumed in the best possible fashion for minimizing the reconstruction error given the shared expansion coefficients. As far as the optimal Y mode, all budgets are allocated to predict the unobservable part Y. In practice, we may assume that M_1 and M_2 are much larger than L. Thus, the majority of the degrees of freedom (i.e., $(M_1 + M_2) \times L$) in PLSR is allocated for maximizing the covariance between the two sets. The method does not explicitly capture variance or minimize reconstruction errors for the observable part X, sometimes leading to higher reconstruction errors on the observable part. On the contrary, with a better correlation and extra budget of L^2 on regression, it tends to better predict the values of the unobservable part Y. In Canonical Regression, the most critical L^2 degrees of freedom are reserved for correlation analysis in the sub-spaces instead of starting from the original high-dimensional data, leading to worse prediction than PLSR. In addition, although dividing the inputs by their standard deviation appears to be a valid strategy for data visualization and regression analysis, it is less desirable for minimizing reconstruction errors.

V. Conclusion

In conclusion, we make the following statements: When estimating coupled yet partially observable data using linear models, one can achieve the lowest overall reconstruction errors by applying standard PCA for the observable part X and the optimal Y mode of the proposed unconditional DP-PCA for the unobservable part Y. Such a combined approach, however, requires two separate sets of bases, resulting in longer computation time and larger storage. When the unobservable part Y is no more critical than the observable part X, the proposed conditional DP-PCA approach can achieve the lowest total error in estimation with a single pair of bases at a fast speed.

References

[1] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

[2] K. Pearson, "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*, vol. 2, no. 11, pp. 559–572, 1901.

[3] H. HOTELLING, "Relations between two sets of variates*." *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[4] J. A. Wegelin *et al.*, "A survey of partial least squares (pls) methods, with emphasis on the two-block case," *University of Washington, Tech. Rep*, 2000.

[5] E. R. Cook, K. R. Briffa, and P. D. Jones, "Spatial regression methods in dendroclimatology: a review and comparison of two techniques," *International Journal of Climatology*, vol. 14, no. 4, pp. 379–402, 1994.

[6] H. R. Glahn, "Canonical correlation and its relationship to discriminant analysis and multiple regression," *Journal of the atmospheric sciences*, vol. 25, no. 1, pp. 23–31, 1968.

[7] L. Sun, S. Ji, S. Yu, and J. Ye, "On the equivalence between canonical correlation analysis and orthonormalized partial least squares," in *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[8] T. De Bie, N. Cristianini, and R. Rosipal, "Eigenproblems in pattern recognition," in *Handbook of Geometric Computing*. Springer, 2005, pp. 129–167.

[9] M. Borga, "Canonical correlation: a tutorial," *On line tutorial http://people. imt. liu. se/magnus/cca*, vol. 4, no. 5, 2001.

[10] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE journal*, vol. 37, no. 2, pp. 233–243, 1991.

[11] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas, "Mulan: A java library for multi-label learning," *Journal of Machine Learning Research*, vol. 12, pp. 2411–2414, 2011.

[12] Y. LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[13] E. Spyromitros-Xioufis, G. Tsoumakas, W. Groves, and I. Vlahavas, "Multi-target regression via input space expansion: treating targets as inputs," *Machine Learning*, vol. 104, no. 1, pp. 55–98, 2016.

[14] S. Džeroski, D. Demšar, and J. Grbović, "Predicting chemical parameters of river water quality from bioindicator data," *Applied Intelligence*, vol. 13, no. 1, pp. 7–17, 2000.