2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Dependently Coupled Principal Component Analysis for Bivariate Inversion Problems

Navdeep Dahiya*, Yifei Fan*, 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*—Principal Component Analysis (PCA) is a widely used technique for dimensionality reduction in various problem domains, including data compression, image processing, visualization, exploratory data analysis, pattern recognition, time-series prediction, and machine learning. Often, data is presented in a correlated paired manner such that there exist observable and correlated unobservable measurements. Unfortunately, traditional PCA techniques generally fail to optimally capture the leverageable correlations between such paired data as it does not yield a maximally correlated basis between the observable and unobservable counterparts. This instead is the objective of Canonical Correlation Analysis (and the more general Partial Least Squares methods); however, such techniques are still symmetric in maximizing correlation (covariance for PLSR) over all choices of the basis for both datasets without differentiating between observable and unobservable variables (except for the regression phase of PLSR). Further, these methods deviate from PCA's formulation objective to minimize approximation error, seeking instead to maximize correlation or covariance. While these are sensible optimization objectives, they are not equivalent to error minimization. We therefore introduce a new method of leveraging PCA between paired datasets in a dependently coupled manner, which is optimal with respect to approximation error during training. We generate a dependently coupled paired basis for which we relax orthogonality constraints in decomposing unreliable unobservable measurements. In doing so, this allows us to optimally capture the variations of the observable data while conditionally minimizing the expected prediction error for the unobservable component. We show preliminary results that demonstrate improved learning of our proposed method compared to that of traditional techniques.**

***Index Terms*—Principal Component Analysis; Canonical Correlation Analysis; Partial Least Squares; Segmentation; Shape Analysis;**

I. Introduction

Principal Component Analysis (PCA) [1] is an unsupervised statistical technique primarily used for dimensionality reduction [2] in various problem domains, including data compression, image processing, visualization, exploratory data analysis, pattern recognition, and time series prediction. It is also very commonly used for small to medium-sized datasets in machine learning for both dimensionality reduction and developing predictive models.

The dimensionality of a dataset is related to the number of features in each sample of a dataset. High-dimensional datasets with a large number of features, especially in the field of Machine Learning, suffer from the problem of "Curse of Dimensionality [3]". In addition, overfitting often occurs in high-dimensional datasets, leading to poor generalization to examples beyond the training set. This motivates the use of PCA for dimensionality reduction by compressing a set of high-dimensional vectors into a set of lower-dimensional vectors and then reconstructing the original set in machine learning and related fields.

PCA summarizes the variation in potentially correlated multivariate attributes or features to a set of linearly uncorrelated components, each of which is a particular linear combination of the original variables. The extracted non-correlated components are called Principal Components (PC) and are estimated from the eigenvectors of the covariance matrix of the original variables. The PCA procedure uses an orthogonal transformation, and the set of Principal Components are called an orthogonal basis.

PCA can be used in several problem settings but broadly for *visualization*, *prediction*, and *inversion* problems. Our newly formulated model, which we call Dependently Coupled Principal Component Analysis (DC-PCA), can be used in two different ways: either for inversion or prediction. In this paper, we focus on the use case for inversion problems.

In the inversion problem setting, PCA is used to extract a set of orthogonal basis vectors that form our model, and a linear combination of these basis vectors can be used to fit a new unseen data sample. Depending on the problem at hand, some form of fitting function or cost function can be used to estimate the weights of the linear combination based on observable features of the data. In [4]–[8], the authors used this strategy for image segmentation in both 2D and 3D in various domains. Over the years, several extensions of PCA have been proposed and used for various tasks in computer vision. For example, in [9], [10], the authors used a non-linear extension called kernel PCA for incorporating non-linear shape priors for segmentation tasks. In [11], the authors exploited Riemannian geometry to develop a non-linear extension referred to as Principal Geodesic Analysis.

A technique related to PCA is called Canonical Correlation Analysis (CCA). In certain scenarios, we could have more than one set of correlated samples. First introduced in [12], canonical correlation analysis, together with its general framework Partial Least Squares (PLS) [13], is a method that measures the linear relationship between two multi-dimensional variables.

This work was funded in part by National Institutes of Health (NIH) grant R01 HL143350, Army Research Office (ARO) grant 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.

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9413305

The method seeks a pair of basis vectors such that the corresponding variables in these bases are maximally correlated. It can be used for dimension reduction similar to PCA, but unlike PCA, which seeks a basis that explains maximum within-set variation, the goals of CCA and PLSR is to seek a paired basis such that the correlation and covariance of the corresponding variables are maximized, respectively. As for calculation, all those methods can be formulated as solving eigenvalue equations with slightly different matrix coefficients [14]. Similar to PCA, many extensions, enhancements and applications, [15]–[22] have been proposed and developed to augment the basic CCA technique. A comprehensive review of all these extensions is beyond the scope of this article. Interested readers are referred to a recent tutorial [14] and the references therein for a complete treatment of CCA.

A. Motivation for Dependently Coupled Principal Component Analysis

We consider the scenario where we have a pair of correlated datasets (X and Y) at training time, *neither of which* will be directly represented by test time measurements. However, we expect that raw test time measurements will contain features which allow us to invert a low-dimensional representation of X by leveraging PCA style dimensionality reduction (i.e., estimate its expansion coefficients), but that either no features (or only poor, unreliable features) will be available to invert a low-dimensional representation of Y. We propose a novel model named Dependently Coupled Principal Component Analysis (DC-PCA) to address this class of bivariate inversion problems where X is inverted from test time data, and where Y is estimated from the low-dimensional inversion of X. This proposed method is a special case, mathematically[1], of our more general formulation called Directionally Paired PCA (DP-PCA), which is the subject of our concurrent paper titled "Directionally Paired Principal Component Analysis for Bivariate Estimation Problems". It is important to emphasize that we are considering the challenging class of inversion problems from which dimensionality reduction is a crucial ingredient, meaning that the inverted result from X will only contain components within the low dimensional space learned during training time, with no additional component to be leveraged in the subsequent estimation of Y.

We call one set of samples observable (X) and the other partially observable or completely unobservable (Y), meaning that one set has features which could be used for estimating the weights of corresponding principal components during model fitting, and the other set may or may not have any features that can be used to estimate model weights. However, we assume that the sets are correlated, and reconstructing one tells us something about the other. In the case of both datasets being observable, we could develop two independent PCA models for each; but in that case, we ignore the correlation between the two sets and perhaps adopt higher dimensional representations

[1]while the two papers share common mathematical formulation, the resulting algorithms and applications are completely different and hence treated in two separate papers.

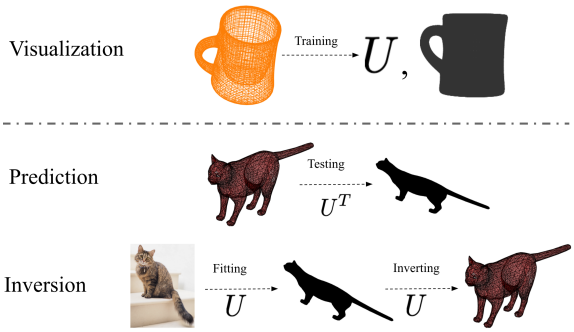


Fig. 1: Use cases of dimension reduction with PCA

than necessary. In such cases, there is a well-known version of PCA called Joint (symmetrically paired) PCA, in which we stack both sets of samples together and extract a single set of Principal Components. This technique would force the model to learn the correlations between the two sets while keeping the dimensionality lower than the case of two independent PCA models. In this case, the capacity of our model is split between optimizing for the two sample sets. In case one of the sample sets is observable and the other unobservable, we lose the ability to reconstruct the observable data well while not gaining much advantage in terms of reconstructing the other unobservable set.

In the following sections, we present a formal description of the existing PCA techniques, their failings in certain cases and describe our new Dependently Coupled Principal Component Analysis technique in more detail. We present a set of synthetic experiments to illustrate the concepts and also show a practical application on a challenging problem of Myocardial Segmentation in medical imagery.

B. Use Case of PCA in Inversion Problems

We may roughly categorize the usage of dimensional reduction with PCA into three use cases: *visualization*, *prediction*, and *inversion*, and summarize the major difference in Fig. 1. The visualization process takes a single step that reduces the dimension of high-dimensional measurements X of data to low-dimensional representations A which are feasible for plotting. In the prediction problem, a basis U is learned during training, which later supports transforming at test time measurements of data X_{test} to a low-dimensional subspace. When it comes to the inversion problem, however, the test phase *starts from the low-dimensional representation*, and the inverse transform is performed to map it to the corresponding high-dimensional measurements, which usually contains statistics of interest (e.g., segmentation maps). In this research, we focus on the *inversion problem*, which differs from the previous two use cases in the following aspects.

One critical concept in the inversion problem is *raw* data D, which are entities that can be quantified by various measurements and statistics (e.g., the input high-dimensional measurement of PCA). Although both the high-dimensional measurement X and its low-dimensional representation A

have fixed dimensions, the dimension of the raw data D might not be fixed or even finite. The goal of the inversion problem is to compute the high-dimensional measurement X of the raw data D by leveraging the low-dimensional representation A, because computing X directly from D is extremely difficult and expensive. Unlike prediction problems in statistics, the dimension M of measurement X in inversion problems is much larger than the number of data samples[2] N. Consequently, direct regression analysis between the raw data D and high-dimensional measurements X is infeasible and subject to overfitting. Alternatively, it is possible to compute[3] the low-dimensional representation A from the raw data D, thus obtaining the high-dimensional measurements X via the inverse PCA transform of A.

II. Traditional PCA for paired datasets

In this section, we establish some notation and review how traditional PCA would be applied in estimating coupled variables in cases where both are observable from a set of measurements (independent PCA) and where only one is observable from the measurements (joint symmetrically coupled PCA). In the subsequent section, we will develop the Dependently Coupled Principal Component Analysis (DC-PCA) methodology, optimized specifically for the latter case, using the same notation presented in this background section.

A. Notation

Let us assume that an $M \times N$ matrix $X = [\mathbf{x}_1\ \mathbf{x}_2 \cdots\ \mathbf{x}_N]$ contains N data measurements represented as column vectors in $\mathbb{R}^M$ and that a $K \times N$ matrix $Y = [\mathbf{y}_1\ \mathbf{y}_2 \cdots\ \mathbf{y}_N]$ contains a different set of N individually paired data measurements (representing a different entity as column vectors in $\mathbb{R}^K$. We further assume that the mean of both sets of measurements is zero (if not, the respective means should be pre-subtracted from each $\mathbf{x}_n$ and $\mathbf{y}_n$ for $n = 1, \ldots, N$.

B. Independent (Unpaired) Principal Component Analysis

Standard PCA, applied independently to each of these paired sets (X and Y), yields independent L-dimensionals subspaces in $\mathbb{R}^M$ and in $\mathbb{R}^K$ that minimize the following mean squared error (MSE):

$$\varepsilon(A,U,B,V) = \frac{1}{N}\sum_{n=1}^{N}\left\|\mathbf{x}_n - \underbrace{\sum_{l=1}^{L}\mathbf{u}_l a_{ln}}_{U\mathbf{a}_n}\right\|^2 + \left\|\mathbf{y}_n - \underbrace{\sum_{l=1}^{L}\mathbf{v}_l b_{ln}}_{V\mathbf{b}_n}\right\|^2 \tag{1}$$

where the columns of $M \times L$ matrices $U = [\mathbf{u}_1 \cdots \mathbf{u}_L]$ and $V = [\mathbf{v}_1 \cdots \mathbf{v}_L]$ denote orthonormal bases of the optimal r-dimensional subspaces and where the coefficients $A = [\mathbf{a}_1 \cdots \mathbf{a}_N]$, $B = [\mathbf{b}_1 \cdots \mathbf{b}_N]$ with $\mathbf{a}_n = (a_{1n}, \ldots, a_{Ln})$, $\mathbf{b}_n = (b_{1n}, \ldots, b_{Ln})$ denote the r-dimensional vectors of coefficients for the linear combinations in these bases of the closest approximations to the measurements $\mathbf{x}_n$ and $\mathbf{y}_n$ in each collected pair. Noting that the orthogonal projections of $\mathbf{x}_n$ and $\mathbf{y}_n$ yield the best approximations for a given choice of subspaces U and V, we may eliminate the parameters a_{ln} and b_{ln} from the optimization problem by substituting $a_{ln} = \mathbf{u}_l^T\mathbf{x}_n$, $b_{ln} = \mathbf{v}_l^T\mathbf{y}_n$ (more compactly $\mathbf{a}_n = U^T\mathbf{x}_n$, $\mathbf{b}_n = V^T\mathbf{y}_n$ or even more compactly $A = U^TX$, $B = V^TY$).

$$\varepsilon^*(U,V) = \frac{1}{N}\sum_{n=1}^{N}\left\|\mathbf{x}_n - \sum_{l=1}^{L}\mathbf{u}_l\mathbf{u}_l^T\mathbf{x}_n\right\|^2 + \left\|\mathbf{y}_n - \sum_{l=1}^{L}\mathbf{v}_l\mathbf{v}_l^T\mathbf{y}_n\right\|^2 \tag{2}$$

Expanding the squared norms into their constituent inner product terms (most of which vanish due to the orthonormality of $\mathbf{u}_1, \ldots, \mathbf{u}_L$ and $\mathbf{v}_1, \ldots, \mathbf{v}_L$) yields

$$\begin{aligned}\varepsilon^* =& \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n^T\mathbf{x}_n - \sum_{l=1}^{L}\mathbf{x}_n^T\mathbf{u}_l\mathbf{u}_l^T\mathbf{x}_n\right) + \\ & \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{y}_n^T\mathbf{y}_n - \sum_{l=1}^{L}\mathbf{y}_n^T\mathbf{v}_l\mathbf{v}_l^T\mathbf{y}_n\right)\end{aligned} \tag{3}$$

From this, it is apparent that an equivalent optimization problem for $\mathbf{u}_1, \ldots, \mathbf{u}_L$ and $\mathbf{v}_1, \ldots, \mathbf{v}_L$ is to maximize

$$\sum_{n=1}^{N}\sum_{l=1}^{L}\mathbf{x}_n^T\mathbf{u}_l\mathbf{u}_l^T\mathbf{x}_n + \mathbf{y}_n^T\mathbf{v}_l\mathbf{v}_l^T\mathbf{y}_n \tag{4}$$

$$= \sum_{l=1}^{L}\mathbf{u}_l^T\left(\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T\right)\mathbf{u}_l + \mathbf{v}_l^T\left(\sum_{n=1}^{N}\mathbf{y}_n\mathbf{y}_n^T\right)\mathbf{v}_l \tag{5}$$

$$= \sum_{l=1}^{L}\mathbf{u}_l^TXX^T\mathbf{u}_l + \mathbf{v}_l^TYY^T\mathbf{v}_l \tag{6}$$

$$= \sum_{l=1}^{L}\|\mathbf{u}_l\|_{XX^T}^2 + \|\mathbf{v}_l\|_{YY^T}^2 \tag{7}$$

where $\|\cdot\|_{XX^T}$ and $\|\cdot\|_{YY^T}$ denote the weighted L^2 norms via the positive definite[4] matrices XX^T and YY^T. Since $\mathbf{u}_1, \ldots, \mathbf{u}_L$ and $\mathbf{v}_1, \ldots, \mathbf{v}_L$ must each be orthonormal, it is clear that the way to maximize this expression under this constraint is to choose the eigenvectors of XX^T and YY^T which, for each matrix, correspond to the L largest eigenvalues $\lambda_1, \ldots, \lambda_L$ (thereby yielding $\lambda_1 + \cdots + \lambda_L$ for each of the two pieces of the sum to be maximized). These eigenvectors as the choice of optimal basis vectors $\mathbf{u}_1, \ldots, \mathbf{u}_L$ and $\mathbf{v}_1, \ldots, \mathbf{v}_L$ are often called the first L *principal components* of each of the data sets X and Y respectively.

In this case, the two orthonormal bases can fit data independently without capturing any correlations between the two datasets X and Y. In case Y is completely unobservable such that we do not have access to measurements during fitting, then we simply cannot use the second set of principal components. In case we had some way of capturing the correlations between the two datasets during training, then we could have been able to use the reconstruction of X to gain some estimate of the unobservable Y. As it turns out, there is one modified way to use PCA to capture such correlations called Joint PCA, which we describe in the next section.

[2]We may have only 10 dataset cube of size 128^3 as training data.

[3]e.g., via minimizing an energy function which depends on the principal components weighted by A

[4]positive semi-definite if X is not full rank.

C. Joint (Symmetric-Paired) Principal Component Analysis

If we only allow a single set of linear combination coefficients $A = [\mathbf{a}_1 \cdots \mathbf{a}_N]$, so that the approximations of $\mathbf{x}_n$ and $\mathbf{y}_n$ in the respective bases U and V must always utilize the same set of expansion coefficients $\mathbf{a}_n$, then we may rewrite the energy by concatenating each $\mathbf{x}_n$ and $\mathbf{y}_n$ into a single measurement vector as well as concatenating each $\mathbf{u}_l$ and $\mathbf{v}_l$ into a single basis vector to yield an equivalent energy as follows.

$$\varepsilon(A,U,V) = \frac{1}{N}\sum_{n=1}^{N}\left\|\mathbf{x}_n - \sum_{l=1}^{L}\mathbf{u}_l a_{ln}\right\|^2 + \left\|\mathbf{y}_n - \sum_{l=1}^{L}\mathbf{v}_l a_{ln}\right\|^2 \tag{8}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left\|\begin{bmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{bmatrix} - \sum_{l=1}^{L}\begin{bmatrix}\mathbf{u}_l\\ \mathbf{v}_l\end{bmatrix} a_{ln}\right\|^2 \tag{9}$$

Assuming that these concatenated basis vectors $(\mathbf{u}_l, \mathbf{v}_l)$ are orthonormal, we may then write the orthogonal projections of the concatenated measurements $(\mathbf{x}_n, \mathbf{y}_n)$ as their closest approximations to obtain

$$\varepsilon^*(U,V) = \frac{1}{N}\sum_{n=1}^{N}\left\|\begin{bmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{bmatrix} - \sum_{l=1}^{L}\begin{bmatrix}\mathbf{u}_l\\ \mathbf{v}_l\end{bmatrix}\begin{bmatrix}\mathbf{u}_l\\ \mathbf{v}_l\end{bmatrix}^T\begin{bmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{bmatrix}\right\|^2 \tag{10}$$

Again, expanding the squared norms into their constituent inner product terms (most of which vanish due to the orthonormality) yields

$$\varepsilon^* = \frac{1}{N}\sum_{n=1}^{N}\left(\begin{bmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{bmatrix}^T\begin{bmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{bmatrix} - \sum_{l=1}^{L}\begin{bmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{bmatrix}^T\begin{bmatrix}\mathbf{u}_l\\ \mathbf{v}_l\end{bmatrix}\begin{bmatrix}\mathbf{u}_l\\ \mathbf{v}_l\end{bmatrix}^T\begin{bmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{bmatrix}\right) \tag{11}$$

Just as before, it is apparent that an equivalent optimization problem for $\mathbf{u}_1, \ldots, \mathbf{u}_L$ and $\mathbf{v}_1, \ldots, \mathbf{v}_L$ is to maximize

$$\sum_{n=1}^{N}\sum_{l=1}^{L}\begin{bmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{bmatrix}^T\begin{bmatrix}\mathbf{u}_l\\ \mathbf{v}_l\end{bmatrix}\begin{bmatrix}\mathbf{u}_l\\ \mathbf{v}_l\end{bmatrix}^T\begin{bmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{bmatrix} \tag{12}$$

$$= \sum_{l=1}^{L}\begin{bmatrix}\mathbf{u}_l\\ \mathbf{v}_l\end{bmatrix}^T\begin{bmatrix}X\\ Y\end{bmatrix}\begin{bmatrix}X\\ Y\end{bmatrix}^T\begin{bmatrix}\mathbf{u}_l\\ \mathbf{v}_l\end{bmatrix} \tag{13}$$

The optimization problem happens when we choose $(\mathbf{u}_1, \mathbf{v}_1), \ldots, (\mathbf{u}_L, \mathbf{v}_L)$ to be the eigenvectors corresponding to the L largest eigenvalues of $\begin{bmatrix}X\\ Y\end{bmatrix}\begin{bmatrix}X\\ Y\end{bmatrix}^T$.

Using this technique, we force the correlations between two datasets to be captured by the Joint PCA model. This is useful only in the case the data can influence both the components during fitting. However, since we impose a single set of linear coefficients, the model capacity is split between learning the variations of both the datasets X and Y. In case Y is completely unobservable during fitting, we get the benefit of the correlation, but we lose the ability to fit the observable data X optimally.

III. Dependently Coupled Principal Component Analysis (DC-PCA)

If we again allow only a single set of linear combination coefficients $A = [\mathbf{a}_1 \cdots \mathbf{a}_N]$, so that the approximations of $\mathbf{x}_n$ and $\mathbf{y}_n$ in the respective bases U and V must always utilize the same set of expansion coefficients $\mathbf{a}_n$, but impose the orthonomal basis U obtained using standard independent PCA on the data set X, together with the coefficients A that yield the best approximation of X within this basis (thereby minimizing the first term below over all choices of U and A, as in standard PCA), then we may seek the "paired basis" V (not necessarily orthonormal) that minimizes the second term below given this choice of U and A.

$$\varepsilon_X(A,U) = \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{x}_n - U\mathbf{a}_n\|^2 \tag{14}$$

$$\varepsilon_Y(A,V) = \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{y}_n - V\mathbf{a}_n^*\|^2 \tag{15}$$

Differentiating ε_X in $\mathbf{a}_n$, and setting the result to zero (to define the optimal $\mathbf{a}_n^*$(U), yields

$$0 = U^T(\mathbf{x}_n - U\mathbf{a}_n^*) = U^T\mathbf{x}_n - \underbrace{U^TU}_{\mathcal{I}}\mathbf{a}_n^* \tag{16}$$

$$\implies \quad \mathbf{a}_n^*(U) = U^T\mathbf{x}_n \tag{17}$$

which, when substituted into ε_Y gives

$$\varepsilon_Y(V) = \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{y}_n - V\mathbf{a}_n^*\|^2$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n^T\mathbf{y}_n - 2\mathbf{x}_n^TUV^T\mathbf{y}_n + \mathbf{x}_n^TUV^TVU^T\mathbf{x}_n \tag{18}$$

Now differentiating with respect to the matrix V we obtain

$$\frac{\partial\varepsilon_Y}{\partial V} = -\frac{1}{N}\sum_{n=1}^{N}-2\mathbf{y}_n(\mathbf{a}_n^*)^T + 2V\mathbf{a}_n^*(\mathbf{a}_n^*)^T \tag{19}$$

$$= -\frac{2}{N}\left(YA^T - VAA^T\right) \tag{20}$$

where

$$A = \begin{bmatrix}\mathbf{a}_1^* & \cdots & \mathbf{a}_N^*\end{bmatrix} = \begin{bmatrix}U^T\mathbf{x}_1 & \cdots & U^T\mathbf{x}_N\end{bmatrix} = U^TX$$

Setting this matrix derivative to zero yields

$$V = YA^T(AA^T)^{-1} = YX^TU\left(U^TXX^TU\right)^{-1} \tag{21}$$

If we use the PCA basis U computed for X, then $U^TU = \mathcal{I}$, and $XX^TU = U\Lambda_X$, in which Λ_X represents the $L \times L$ diagonal matrix with the L largest eigenvalues of XX^T along the diagonal. Plugging those into the above equation yields

$$V = YX^TU\left(\underbrace{U^TU}_{\mathcal{I}}\Lambda_X\right)^{-1} = YX^TU\Lambda_X \tag{22}$$

As such, given a set of expansion coefficients which estimate an observable variable $\mathbf{x}$, we may obtain an optimal prediction (according to our training) for an unobservable variable $\mathbf{y}$ by applying the same weighted linear combination to the matching basis elements in V. We refer to this combination of traditional PCA for X and unidirectional correlation analysis for Y as Dependently Coupled Principal Component Analysis (DC-PCA) for the paired data sets.

A. Relationship to CCA

A well known (but symmetric) method that also produces a paired set of bases for a correlated pair of variables is Canonical Correlation Analysis (CCA). The goal of CCA is to determine bases which maximally correlate linear combinations of two sets of variables. It can also be used for dimensionality reduction when the goal within the low dimensional subspaces is not to optimally approximate coupled sets of measurements X and Y by orthogonal projection (as in PCA) but rather to maximize correlation. These are two different objectives that generally yield two different matched pairs of bases.

Since our *asymmetrically* paired basis V maximizes the correlation between linear expansions of its components with equally weighted expansions of the basis vectors in U, it seems clear there must be a connection with CCA methods in the pairing of V with U. This is indeed the case, with the critical difference that in CCA, correlation is maximized symmetrically by optimizing over all choices for *both* bases U and V. This necessarily results in orthogonal bases for both U and V (just as in joint PCA which also yields different but still orthogonal, paired bases U and V). Here, however, correlation is maximized only over the choice of V whereas U is optimized independently of V with the different goal of maximizing its own variance across the training set X. A consequence of this asymmetric directionality is that while the PCA calculated basis U will be orthogonal, its paired correlated basis V will typically not be. In the exceptional case that PCA applied to the observable measurements X delivers the same basis U as CCA does on the paired measurements X and Y, then the resulting asymmetrically paired basis V would match that given by CCA as well, and would therefore be orthogonal.

B. Relationship to conditional PCA

On the other hand and similar to traditional PCA, we note that the proposed approach can be viewed in a conditional Bayesian context with caveats. Namely, the basis set of V is conditioned on the computation of U computed independently on measurement X (whereas basis V is computed dependently on Y and U). This additional consideration and dependency is a key discerning element between our proposed method and classical conditional PCA. For example, previous work adopting the moniker of conditional PCA [23] focuses on exploiting a PCA basis on measurement Y from measurement X. This, of course, leads to an orthogonal basis V. Here the computation of the basis V remains unchanged provided measurement X and such basis U and V are, in a sense, decoupled. Compared to our proposed method, this is not the case. As a result of coupling, the basis V is not necessarily orthogonal, and this relaxation provides the key advantage argued in this note. This said, a key part of the future work is to provide an alternative characterization in the Bayesian context, and the above is noted to highlight a tacit assumption for which further study is warranted.

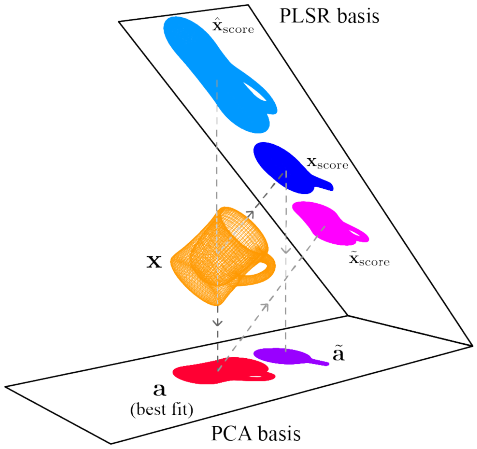


Fig. 2: The "projection of a projection" when applying PLS methods for inversion problems.

C. Relationship to PLS/PLSR

The strength and uniqueness of the proposed DC-PCA model lie in its special customization for inversion problems. On the contrary, those partial least-squares methods (i.e., PLSR and CCA) apply to only the predication scenario, which assumes full access to the measurement X of the observable part. In practice, however, the low-dimensional representation of the test data A_{test} is often obtained by optimizing a loss function such that the representation maximally captures the variance of the raw data D_{test}. In other words, the representation always matches the PCA basis. Therefore, applying the PLS basis for solving the low-dimensional representation (i.e., expansion coefficient or score) would result in a situation of the "projection of a projection," which can be illustrated by Fig. 2.

Let us consider the full high-dimensional representation (which cannot be inverted) of the observable part $\mathbf{x}$ as a volume in the 3D space, which can be projected onto the PCA-basis plane or the PLSR-basis plane (where inversion will occur). Under the PCA basis, $\mathbf{a}$ (illustrated by the red projection at the bottom of Fig. 2) is the vector of expansion coefficients that maximally capture the variance (i.e., best fit) of the high-dimensional signal $\mathbf{x}$. Under the PLSR basis, $\mathbf{x}_{\text{score}}$ (illustrated by the dark blue projection at the right of Fig. 2) is the best low-dimensional representation of the full signal $\mathbf{x}$ that leads to the optimal prediction of the unobservable part $\mathbf{y}$. At the test time of an inversion problem, the evolution of the low-dimensional representation (i.e., expansion coefficient) is driven by the effective projection on the PCA plane.

There are two potential methods to exploit PLS-bases in this class of low dimensional inversion problems, both of which lead to mismatch. On the one hand, if we use the PCA basis for X to directly fit the coefficients $\mathbf{a}$ and then map the result to the corresponding PLSR basis prior to estimating Y, then only the projection $\tilde{\mathbf{x}}_{\text{score}}$ (illustrated by the magenta projection at the right of Fig. 2) from $\mathbf{a}$ to the PLSR basis plane contributes to the estimate of Y. On the other hand, if we were

to directly use the PLSR basis for X to compute the PLSR scores, we would obtain a result $\hat{\mathbf{x}}_{\text{score}}$ (illustrated by the light blue projection at the right of Fig. 2) whose projection onto the PCA basis (the subspace which best captures the signal itself) most resembles the optimal low-dimensional estimate $\mathbf{a}$. While tempting to assume, we would not capture the desired PLSR estimate for X. While this estimate would minimize the projection error for X itself, we do not have access to X but rather must invert it by minimizing the residual of some forward model applied to the raw data. Generally, inversion through the raw data residual will seek to capture the largest variations of the unknown signal, subjected to the low dimensional constraint. The inversion process will be guided by the projection onto the PCA basis, even if the inversion itself is constrained to the PLSR basis. Therefore, the desired PLSR result would effectively be seen during the inversion process as its projection $\tilde{\mathbf{a}}$ (the purple projection at the bottom of Fig. 2) onto the PCA basis, which would not optimize the residual error.

In both cases, therefore, the low-dimensional scores $\tilde{\mathbf{x}}_{\text{score}}$ and $\hat{\mathbf{x}}_{\text{score}}$ do not match the PLSR loadings because they are expected to map the true score $\mathbf{x}_{\text{score}}$ back to the high-dimensional space of observable measurements $\mathbf{x}$ and $\mathbf{y}$, respectively. Unfortunately, the required score $\mathbf{x}_{\text{score}}$ for PLSR is not reachable at test time. Should it be reachable, its projection on the PCA basis would become $\tilde{\mathbf{a}}$, which appears to be different from the actual $\mathbf{a}$. Unless we keep another pair of PCA bases with which we can reconstruct the high-dimensional measurements $\mathbf{x}$, PLS methods are not suitable for inversion problems. While this is conceptually explained by our prior discussion of Fig. 2, we demonstrate this directly by a test example in Fig. 3. Here X represents a 2D cross-section shape and Y a paired 3D teacup shape. A low dimensional shape inversion is applied to an ideal noiseless silhouette image using both a PCA and a PLSR basis for X. As shown on the left, both methods extract a similar shape from the raw image data. However, the estimated 3D surface (Y estimate) from PLSR exhibits a higher mismatch against the ground truth compared to DC-PCA, as shown on the right. This is precisely because the X estimate that would have produced a superior PLSR estimate of Y is not captured during the inversion process. In short, the strategic components of the low-dimensional estimate of the 2D curve that PLSR is optimized to leverage (i.e., those of maximum covariance with Y) were not properly inverted since the residual image segmentation error driving the inversion process responded instead to the independent variance within X itself.

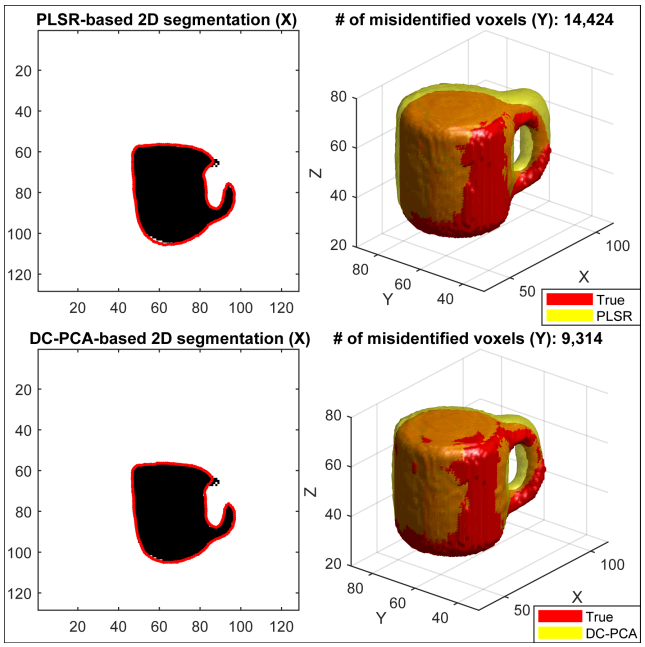
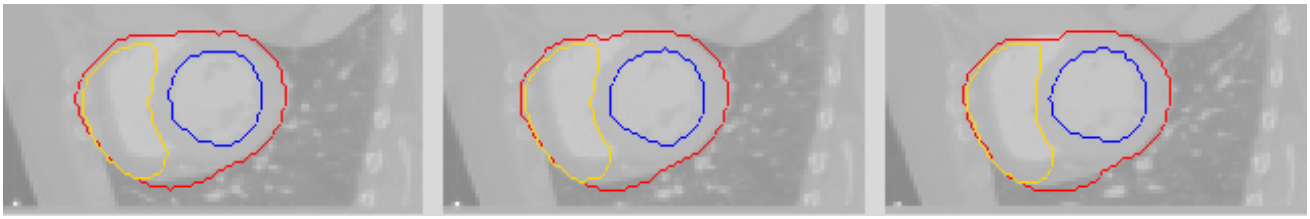


Fig. 3: Demonstration of the "projection of a projection" using 2D/3D teacup segmentation example.

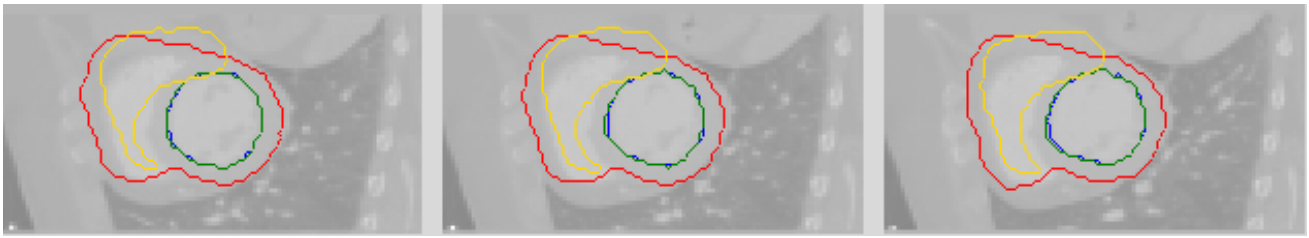
IV. Experimental Results

Although we have presented a general framework applicable to inversion problems, we illustrate the use and demonstrate its superiority over PLSR/CCA based methods using a 3D Cardiac Segmentation problem. There are several uses of PCA in conjunction with 3D deformable models, particularly for image segmentation in medical imagery [5]–[8]. Several such methods use a set of learned shape models to segment medical images. These methods rely on PCA coefficients for matching the models to the features observed in grayscale medical images. In many cases, we are interested in segmenting multiple structures or organs from medical images. In some cases, one organ may be easier to segment than others. In Cardiac Segmentation, the objective is to segment the Left Ventricle (LV), the Right Ventricle (RV), and Epicardium (EPI) structures in cardiac CT or MRI images for disease diagnostic purposes. In Contrast Enhanced CT Angiography (CCTA) images, the LV generally has good contrast and thus is easier to segment. Comparatively, the RV has very little contrast and has wildly changing shapes from patient to patient, and the Epicardium is also notoriously hard to segment. However, the shapes are still anatomically coupled: given a particular LV shape, we can expect the RV and EPI shapes to be correlated with LV. In such a case, we can use a PCA based shape model for the LV, which we can fit the grayscale image to segment the LV and use the DC-PCA framework to estimate the correlated RV and EPI shapes.

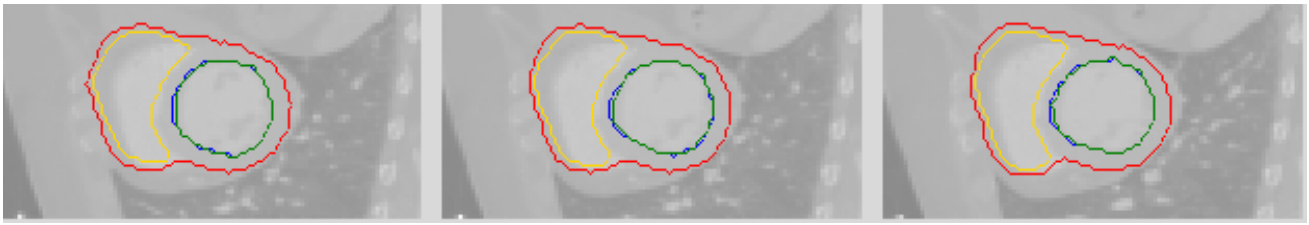
Using the framework developed by the authors in [5], we learn a shape prior model for heart anatomy. We use a set of binary masks obtained from the manual tracing of heart boundaries done by clinical practitioners to develop models of LV, RV, and EPI. Using a set of these binary masks, we use PCA to obtain a set of mean 3D shapes and principal components or principal modes of variations of the anatomical shapes. Since we have manual segmentations of all three shapes during training, we can develop independent models for all three shapes. We can then use a region-based image segmentation model to fit these models to grayscale data. However, as mentioned earlier, RV and EPI are difficult



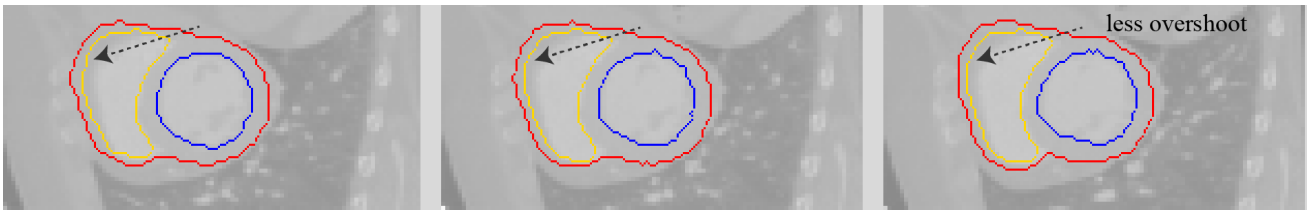
(a) Joint PCA (symmetrically paired) DICE scores: RV = 0.61, EPI = 0.87



(b) CCA DICE scores: RV = 0.40 EPI = 0.89



(c) PLSR DICE scores: RV = 0.65 EPI = 0.91



(d) DC-PCA DICE scores: RV = 0.75 EPI = 0.93

Fig. 4: Several slices of a 3D Cardiac CT Angiography (CCTA) Image. The blue/green curve is the Left Ventricle (LV), the yellow curve is the Right Ventricle (RV), and the red curve is the Epicardium (EPI). In (a), LV is segmented using a joint-PCA-based shape model, and the RV/EPI curves are estimated by applying LV's coefficients to their respective bases obtained by using Joint PCA. In (b) and (c), there are two sets of paired bases, LV/RV and LV/EPI. We estimate two sets of coefficients for LV and apply the same to the respective paired RV and EPI. In (d), DC-PCA uses the independent PCA basis for LV and applies the same set of coefficients to the respective paired bases to estimate RV and EPI. As indicated by the DICE scores, DC-PCA does a clearly better job of segmenting RV/EPI based on LV and needs only half the computation and storage as compared to CCA/PLSR.

to segment due to lack of contrast, hence instead of using three independent models, we use the DC-PCA technique (equation 22) to pair LV (observable part) to RV and EPI (unobservable part) to learn an independent model for LV and asymmetrically paired models for RV and EPI respectively. We then estimate a set of weights for the LV principal components and apply the same weights to the asymmetrically paired RV and EPI bases (obtained using DC-PCA).

Fig. 4(d) shows the results of applying DC-PCA to the task of segmenting cardiac images. The Right Ventricle (yellow curve) and Epicardium (red curve) are simply the asymmetric estimates of corresponding structures based on the precise segmentation of the Left Ventricle, which represents the observable part in this case. Using DC-PCA, we capture the correlation between the observable (high-confidence) LV and the other unobservable (low-confidence, low contrast) structures that are harder to estimate from image data directly. Using DC-PCA, the results show very plausible segmentations of the Epicardium and RV just based on LV. We use the well-known DICE coefficient (also known as F1 score) to quantitatively measure the segmentation accuracy. DC-PCA leads to the highest DICE score for both RV and EPI. In contrast, Fig. 4(a-c) shows the result of using Joint PCA, CCA, and PLSR for EPI and RV based on LV on the same image. Clearly, the DC-PCA-based technique captures the correlations

between LV and RV shapes as well as LV and EPI shapes much better than other techniques. In the case of CCA and PLSR, we have to make two estimates of LV using each of the two separate paired bases (i.e., LV/EPI and LV/RV), leading to double computational load.

V. Conclusions

We have presented a novel method of leveraging PCA between paired datasets in a dependently-coupled manner, which is optimal with respect to approximation error during training. This method, which we have coined Dependently Coupled Principal Component Analysis (DC-PCA), is optimized to capture both variation and correlation between two sets of variables when one of the sets is observable, and the other is not during the model fitting stage. In this paper, we have presented its special customization for inversion problems. While we have presented this methodology in the simple linear framework, kernel-based and other manifold based extensions naturally follow and will be the subject of future work.

References

[1] 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.

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

[3] R. E. Bellman, *Adaptive control processes: a guided tour*. Princeton university press, 2015, vol. 2045.

[4] A. Vikram, B. Ganapathy, A. Abufadel, A. Yezzi, and T. Faber, "A regions of confidence based approach to enhance segmentation with shape priors," in *Proc. of SPIE-IS&T Electronic Imaging, SPIE*, 2010, pp. 7533–12.

[5] M. Leventon, E. Grimson, and O. Faugeras, "Statistical shape influence in geodesic active contours," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, 2000, pp. 316–323.

[6] A. Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, W. E. Grimson, and A. Willsky, "A shape-based approach to the segmentation of medical imagery using level sets," *IEEE Transactions on Medical Imaging*, vol. 22, pp. 137–154, 2003.

[7] N. Dahiya, A. Yezzi, M. Piccinelli, and E. Garcia, "Integrated 3d anatomical model for automatic myocardial segmentation in cardiac ct imagery," in *Lecture Notes in Computational Vision and Biomechanics*, 2017, pp. 1115–1123.

[8] ——, "Integrated 3d anatomical model for automatic myocardial segmentation in cardiac ct imagery," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 7, no. 5-6, pp. 690–706, 2019.

[9] S. Dambreville, Y. Rathi, and A. Tannenbaum, "Shape-based approach to robust image segmentation using kernel pca," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, pp. 977–984.

[10] ——, "A framework for image segmentation using shape models and kernel space shape priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1385–1399, 2008.

[11] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi, "Principal geodesic analysis for the study of nonlinear statistics of shape," *IEEE transactions on medical imaging*, vol. 23, no. 8, pp. 995–1005, 2004.

[12] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*. Springer, 1992, pp. 162–190.

[13] 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.

[14] V. Uurtio, J. a. M. Monteiro, J. Kandola, J. Shawe-Taylor, D. Fernandez-Reyes, and J. Rousu, "A tutorial on canonical correlation methods," *ACM Comput. Surv.*, vol. 50, no. 6, Nov. 2017. [Online]. Available: https://doi.org/10.1145/3136624

[15] M. B. Blaschko, C. H. Lampert, and A. Gretton, "Semi-supervised laplacian regularization of kernel canonical correlation analysis," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2008, pp. 133–145.

[16] A. Kimura, M. Sugiyama, T. Nakano, H. Kameoka, H. Sakano, E. Maeda, and K. Ishiguro, "Semicca: Efficient semi-supervised learning of canonical correlations," *Information and Media Technologies*, vol. 8, no. 2, pp. 311–318, 2013.

[17] X. Chen, L. Han, and J. Carbonell, "Structured sparse canonical correlation analysis," in *Artificial Intelligence and Statistics*, 2012, pp. 199–207.

[18] B. Zhang, J. Hao, G. Ma, J. Yue, and Z. Shi, "Semi-paired probabilistic canonical correlation analysis," in *International Conference on Intelligent Information Processing*. Springer, 2014, pp. 1–10.

[19] P. Horst, "Relations amongm sets of measures," *Psychometrika*, vol. 26, no. 2, pp. 129–149, 1961.

[20] J. D. Carroll, "Generalization of canonical correlation analysis to three or more sets of variables," in *Proceedings of the 76th annual convention of the American Psychological Association*, vol. 3. Washington, DC, 1968, pp. 227–228.

[21] A. Tenenhaus and M. Tenenhaus, "Regularized generalized canonical correlation analysis," *Psychometrika*, vol. 76, no. 2, p. 257, 2011.

[22] A. Tenenhaus, C. Philippe, and V. Frouin, "Kernel generalized canonical correlation analysis," *Computational Statistics & Data Analysis*, vol. 90, pp. 114–131, 2015.

[23] H. Cardot, "Conditional functional principal components analysis," *Scandinavian journal of statistics*, vol. 34, no. 2, pp. 317–335, 2007.