Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2011, Article ID 748430, 7 pages
doi:10.1155/2011/748430

Research Article

Matrix-Variate Probabilistic Model for Canonical Correlation Analysis

Mehran Safayani and Mohammad Taghi Manzuri Shalmani

Department of Computer Engineering, Sharif University of Technology P.O. Box 11155-8639, Tehran 1458889694, Iran

Correspondence should be addressed to Mehran Safayani, safayani@ce.sharif.edu

Received 29 October 2010; Revised 29 January 2011; Accepted 31 January 2011

Academic Editor: Enrico Capobianco

Motivated by the fact that in computer vision data samples are matrices, in this paper, we propose a matrix-variate probabilistic model for canonical correlation analysis (CCA). Unlike probabilistic CCA which converts the image samples into the vectors, our method uses the original image matrices for data representation. We show that the maximum likelihood parameter estimation of the model leads to the two-dimensional canonical correlation directions. This model helps for better understanding of two-dimensional Canonical Correlation Analysis (2DCCA), and for further extending the method into more complex probabilistic model. In addition, we show that two-dimensional Linear Discriminant Analysis (2DLDA) can be obtained as a special case of 2DCCA.

1. Introduction

Recently, a probabilistic interpretation of statistical dimension reduction algorithms has been proposed by several authors. Tipping and Bishop have derived a latent variable model for principal component analysis (PPCA) and have shown that how the principal subspace of the set of data vectors can be obtained within a maximum likelihood framework [1]. Lawrence has proposed another probabilistic model for Principal Component Analysis (PCA); he integrated out the weights and optimized the positions of the latent variables in the q dimensional latent space [2]. Roweis has presented an expectation-maximization (EM) algorithm for PCA. The algorithm allows a few eigenvectors and eigenvalues to be extracted from large collections of high dimensional data [3]. Unlike PCA, which works with a single random vector and maximizes the variance in the projected space, Canonical Correlation Analysis (CCA) works with a pair of random vectors (or in general with a set of m random vectors) and maximizes correlation between sets of projections. In [4], a latent variable model for CCA has been proposed by Bach and Jordan. Other probabilistic models are also known [5–8]. In general, the probabilistic models have many advantages including the following:

(i) the potential to extending the scope of the methods into the mixture models [9],

(ii) extending the methods so as to handle the missing data values [1],

(iii) automatic model selection can be applied by combining the likelihood with a prior [10],

(iv) Extending the model into supervised or semi supervised cases [5].

One major drawback of aforementioned methods is that they only work for data vectors while in computer vision research, samples are often multidimensional arrays such as matrix or tensor. Hence, in the preprocessing step, the image matrices should be converted into the long vectors. This results in losing the spatial structure of the image and consequently the huge covariance matrices, high computational cost, and small sample size problem. Recently, some statistical methods that directly perform on the image matrices without the image to vector conversion procedure have been proposed. These methods make use of

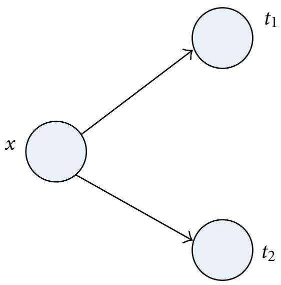

Figure 1: Probabilistic graphical model for CCA.

the spatial information in the image structure and reduce the computational cost to a great extent. General Low Rank Approximation of Matrix (GLRAM) [11], Two-Dimensional Canonical Correlation Analysis (2DCCA) [12, 13], and Two-Dimensional Linear Discriminant Analsis (2DLDA) [14] are some well-known matrix-based algorithms constructed based on this idea. Some other researchers have applied multilinear algebra and have extended this concept to higher-order tensor data [15–20].

Because of the success of the matrix-based methods, recently some researchers have developed probabilistic model for matrix and tensor extensions of PCA [21–24]. However, they do not show the maximum likelihood relationship between their models and corresponding PCA.

Armed with probabilistic principal component analysis [1] and probabilistic canonical correlation analysis [4], in this paper, we propose a matrix-variate factor analysis model that has the property that its maximum likelihood solution extracts the canonical correlation directions of two random matrices. In addition, we show that 2DCCA can be converted to 2DLDA by considering special kind of random matrices. This means that 2DLDA can be interpreted as 2DCCA between appropriately defined random matrices.

The remaining part of the paper is organized as follows: in Section 2, we review CCA and its probabilistic interpretation. Two-Dimensional CCA is described in Sections 3, and 4 introduces our probabilistic model and derivation of canonical directions using maximum likelihood estimation. The relationship between 2DCCA and 2DLDA is discussed in Section 5. Finally, conclusions are presented in Section 6.

2. Probabilistic CCA

Canonical Correlation Analysis determines the linear relationship between two multidimensional variables $t_1 \in \mathfrak{R}^m$ and $t_2 \in \mathfrak{R}^n$. It finds a pair of linear transforms u_1 and u_2 such that correlations between transformed variables $u_1^T t_1$ an $u_2^T t_2$ are maximized. The objective function of CCA can be written as

$$\underset{u_1,u_2}{\arg\max} \frac{\operatorname{cov}\left\langle u_1^T t_1, u_2^T t_2\right\rangle}{\sqrt{\operatorname{var}\left\langle u_1^T t_1\right\rangle}\sqrt{\operatorname{var}\left\langle u_2^T t_2\right\rangle}}. \tag{1}$$

The solutions to this problem can be obtained as $u_1 = \tilde{\Sigma}_{11}^{1/2} q_1$ and $u_2 = \tilde{\Sigma}_{22}^{1/2} q_2$, where q_1 and q_2 contain left-right singular vectors of $(\tilde{\Sigma}_{11})^{-1/2}\tilde{\Sigma}_{12}(\tilde{\Sigma}_{22})^{-1/2}$ and $\tilde{\Sigma}_{ij}$ denotes the sample covariance matrix of t_i and t_j random vectors. A latent variable model for CCA has been proposed in [4] whose graphical model is depicted in Figure 1. The model is defined as follows:

$$\begin{aligned} x &\sim N(0, I_d), \quad \min\{m,n\} \geq d \geq 1, \\ t_1 \mid x &\sim N(W_1 x, \Psi_1), \quad W_1 \in \mathfrak{R}^{m\times d},\ \Psi_1 \succeq 0, \\ t_2 \mid x &\sim N(W_2 x, \Psi_2), \quad W_2 \in \mathfrak{R}^{n\times d},\ \Psi_2 \succeq 0, \end{aligned} \tag{2}$$

where we assume that the data is centered. The negative log likelihood of the data is equal to

$$\mathrm{L} = \frac{(m+n)N}{2}\log(2\pi) + \frac{N}{2}\log|\Sigma| + \frac{N}{2}\operatorname{tr}\left(\Sigma^{-1}\tilde{\Sigma}\right), \tag{3}$$

where N is the number of the samples, $\Sigma = \begin{pmatrix} W_1W_1^T+\Psi_1 & W_1W_2^T \\ W_2W_1^T & W_2W_2^T+\Psi_2 \end{pmatrix}$, and $\tilde{\Sigma} = \begin{pmatrix} \tilde{\Sigma}_{11} & \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{21} & \tilde{\Sigma}_{22} \end{pmatrix}$ denotes the $(m+n) \times (m+n)$ sample covariance matrix obtained from data $t_1^j, t_2^j, j = 1, \ldots, N$.

The maximum likelihood solution is given by

$$\begin{aligned} \widehat{W}_1 &= \tilde{\Sigma}_{11}^{1/2} Q_{1d} P_d^{1/2}, \\ \widehat{W}_2 &= \tilde{\Sigma}_{22}^{1/2} Q_{2d} P_d^{1/2}, \end{aligned} \tag{4}$$

where the columns of Q_{1d} and Q_{2d} are the first d left-right singular vectors of the matrix $(\tilde{\Sigma}_{11})^{-1/2}\tilde{\Sigma}_{12}(\tilde{\Sigma}_{22})^{-1/2}$, and P_d is the diagonal matrix containing the singular values of $(\tilde{\Sigma}_{11})^{-1/2}\tilde{\Sigma}_{12}(\tilde{\Sigma}_{22})^{-1/2}$.

3. 2-Dimensional CCA

The main difference between classical CCA and 2DCCA lies in the way the data are represented. Unlike classical CCA which uses the vectorized representation, 2DCCA works with the data in matrix representation. Therefore, 2DCCA preserves some implicit structural information among elements of the original images. It also overcomes the singularity problem of scatter matrices resulting from the high dimensionality of vectors [12, 13].

2DCCA considers two random matrices $T_1 \in \mathfrak{R}^{m_1\times n_1}$ and $T_2 \in \mathfrak{R}^{m_2\times n_2}$ and seeks left transforms $L_1 \in \mathfrak{R}^{m_1\times m'}$, $L_2 \in \mathfrak{R}^{m_2\times m'}$ and right transforms $R_1 \in \mathfrak{R}^{n_1\times n'}$, $R_2 \in \mathfrak{R}^{n_2\times n'}$ such that the following criteria is maximized:

$$\underset{L_1,L_2,R_1,R_2}{\arg\max} \frac{\operatorname{tr}\left(\operatorname{cov}\left\langle L_1^T T_1 R_1, L_2^T T_2 R_2\right\rangle\right)}{\sqrt{\operatorname{tr}\left(\operatorname{var}\left\langle L_1^T T_1 R_1\right\rangle\right)}\sqrt{\operatorname{tr}\left(\operatorname{var}\left\langle L_2^T T_2 R_2\right\rangle\right)}}. \tag{5}$$

There is no closed form solution for maximizing all projection matrices simultaneously. Hence, 2DCCA adopts an iterative algorithm for finding the local optimal projections.

At first left transforms L_1, L_2 are assumed known and the following covariance matrices are defined:

$$\begin{aligned}
\operatorname{cov}\left\langle T_1^l, T_2^l \right\rangle &= \tilde{\Sigma}_{12}^l = \frac{1}{Nm'} \Sigma_{n=1}^N T_{1,n}^l \left(T_{2,n}^l\right)^T, \\
\operatorname{var}\left\langle T_1^l \right\rangle &= \tilde{\Sigma}_{11}^l = \frac{1}{Nm'} \Sigma_{n=1}^N T_{1,n}^l \left(T_{1,n}^l\right)^T, \\
\operatorname{var}\left\langle T_2^l \right\rangle &= \tilde{\Sigma}_{22}^l = \frac{1}{Nm'} \Sigma_{n=1}^N T_{2,n}^l \left(T_{2,n}^l\right)^T,
\end{aligned} \tag{6}$$

where $T_1^l = T_1^T L_1$ and $T_2^l = T_2^T L_2$ are left projected sample matrices. Then, the formula (1) becomes

$$\underset{R_1,R_2}{\arg\max} \frac{\operatorname{tr}\left(R_1^T \tilde{\Sigma}_{12}^l R_1\right)}{\sqrt{\operatorname{tr}\left(R_1^T \tilde{\Sigma}_{11}^l R_1\right)}\sqrt{\operatorname{tr}\left(R_2^T \tilde{\Sigma}_{22}^l R_2\right)}}. \tag{7}$$

The optimal projection can be obtained as follows:

$$\begin{aligned}
R_1 &= \left(\tilde{\Sigma}_{11}^l\right)^{-1/2} Q_1^l, \\
R_2 &= \left(\tilde{\Sigma}_{22}^l\right)^{-1/2} Q_2^l,
\end{aligned} \tag{8}$$

where Q_1^l and Q_2^l contain n' first left-right singular vector of $(\tilde{\Sigma}_{11}^l)^{-1/2} \tilde{\Sigma}_{12}^l (\tilde{\Sigma}_{22}^l)^{-1/2}$.

Alternatively, we can rewrite (5) as

$$\underset{L_1,L_2}{\arg\max} \frac{\operatorname{tr}\left(L_1^T \tilde{\Sigma}_{12}^r L_1\right)}{\sqrt{\operatorname{tr}\left(L_1^T \tilde{\Sigma}_{11}^r L_1\right)}\sqrt{\operatorname{tr}\left(L_2^T \tilde{\Sigma}_{22}^r L_2\right)}}, \tag{9}$$

where,

$$\begin{aligned}
\operatorname{cov}\langle T_1^r, T_2^r \rangle &= \tilde{\Sigma}_{12}^r = \frac{1}{Nn'} \Sigma_{n=1}^N T_{1,n}^r \left(T_{2,n}^r\right)^T, \\
\operatorname{var}\langle T_1^r \rangle &= \tilde{\Sigma}_{11}^r = \frac{1}{Nn'} \Sigma_{n=1}^N T_{1,n}^r \left(T_{1,n}^r\right)^T, \\
\operatorname{var}\langle T_2^r \rangle &= \tilde{\Sigma}_{22}^r = \frac{1}{Nn'} \Sigma_{n=1}^N T_{2,n}^r \left(T_{2,n}^r\right)^T,
\end{aligned} \tag{10}$$

where $T_1^r = T_1 R_1$ and $T_2^r = T_2 R_2$ are right-projected sample matrices. The optimal solution can be obtained as

$$\begin{aligned}
L_1 &= \left(\tilde{\Sigma}_{11}^r\right)^{-1/2} Q_1^r, \\
L_2 &= \left(\tilde{\Sigma}_{22}^r\right)^{-1/2} Q_2^r,
\end{aligned} \tag{11}$$

where Q_1^r and Q_2^r contain m' first left-right singular vector of $(\tilde{\Sigma}_{11}^r)^{-1/2} \tilde{\Sigma}_{12}^r (\tilde{\Sigma}_{22}^r)^{-1/2}$. left projections ($L_1$ and L_2) and right projections (R_1 and R_2) are determined by iteratively solving (7) and (9) until convergence.

4. Matrix-Variate Probabilistic Model for CCA

In this section, we propose an extension of probabilistic canonical correlation model to deal with 2D data. One limitation of probabilistic canonical correlation model is that, in this method, samples are represented by vectors while in computer vision research, data (images) are often matrices, and structural information can be used for improving the conventional model. We show that the estimating the parameters of the proposed model leads to the two-dimensional canonical correlation analysis directions.

We relate the random matrices $T_1 \in \Re^{m_1 \times n_1}$ and $T_2 \in \Re^{m_2 \times n_2}$ with the latent matrix $X \in \Re^{m' \times n'}$ as follows:

$$\begin{aligned}
T_1 &= U_1 X V_1^T + \Xi_1, \\
T_2 &= U_2 X V_2^T + \Xi_2,
\end{aligned} \tag{12}$$

where $U_1 \in \Re^{m_1 \times m'}$, $V_1 \in \Re^{n_1 \times n'}$, $U_2 \in \Re^{m_2 \times m'}$, and $V_2 \in \Re^{n_2 \times n'}$ are the factor loading matrices. Ξ_1 and Ξ_2 are the noise sources, and every entry of them follows from $N(0, \psi_1)$ and $N(0, \psi_2)$, respectively. Let $\theta = \{U_1, U_2, V_1, V_2, \psi_1, \psi_2\}$ be the parameter of the model. The observations T_1 and T_2 are conditionally independent given the value of latent matrix X; so, we have

$$P(T_1, T_2 \mid X, \theta) = P(T_1 \mid X, \theta) P(T_2 \mid X, \theta). \tag{13}$$

Marginal distribution of observed variables is then given by the integrating out the latent variable as

$$P(T_1, T_2 \mid \theta) = \int P(T_1 \mid X, \theta) P(T_2 \mid X, \theta) P(X) dX. \tag{14}$$

Maximum likelihood is one method for setting the values of these parameters which involves consideration of the log probability of the observed data set given the parameters, that is,

$$L(D_1, D_2 \mid \theta) = \ln\, p(D_1, D_2 \mid \theta) = \sum_{n=1}^N \ln\, P\left(T_{1;n}, T_{2;n} \mid \theta\right), \tag{15}$$

where $D_i = \{T_{i;n}\}_{n=1}^N$ and $i \in \{1, 2\}$ consist of N data matrix. One difficulty here is that all the projection matrices $\{U_i, V_i\}_{i=1}^2$ should be obtained simultaneously and there is no closed-form solution for it. Therefore, two probabilistic models are proposed so as to obtain each projection direction separately and from an alternating optimization procedure.

We assume that the value of U_i, $i = 1, 2$ is known and proceed to project the observations over these matrices. The left probabilistic model is defined as

$$T_i^l = V_i X^l + \Xi_i^l, \quad i = 1, 2, \tag{16}$$

where $T_i^l = T_i^T U_i$, $X^l = X^T$, and Ξ_i^l is the noise in this model. We define the left probabilistic function $P(T_1^l, T_2^l)$ as the marginal distribution over the latent variables, that is,

$$P\left(T_1^l, T_2^l \mid \theta^l\right) = \int P\left(T_1^l, T_2^l \mid X^l, \theta^l\right) P\left(X^l\right) dX, \tag{17}$$

where $\theta^l = \{V_i, \Psi_i^l\}|_{i=1}^{2}$. The projected observations T_1^l and T_2^l are conditionally independent given the value of latent matrix X^l; so, we have

$$P\left(T_1^l, T_2^l \mid X^l, \theta^l\right) = \prod_{i=1}^{2} P\left(T_i^l \mid X^l, \theta_i^l\right). \tag{18}$$

One major problem here is that the probabilistic distributions are defined over vectors but in this case observed data are matrices.

Suppose that $t_{i,j}^l \in \mathfrak{R}^{n_i}$ be the jth column of the projected matrices $T_i^l \in \mathfrak{R}^{n_i \times m'}$, then the probabilistic function $P(T_i^l)$ is defined as

$$P\left(T_i^l\right) = \Pi_{j=1}^{m'} p\left(t_{i,j}^l\right), \quad i = 1, 2. \tag{19}$$

x-conditional probability distribution over $t_{i,j}^l$ space is given by

$$p\left(t_{i,j}^l \mid x_j^l\right) \sim N\left(V_i x_j^l, \Psi_i^l\right), \quad \Psi_i^l \geq 0, \ i = 1, 2, \tag{20}$$

where $x_j^l \in \mathfrak{R}^{n'}$ is defined as the jth column vector of X^l and marginal distribution of x_j is $N(0, I)$. Therefore, the marginal distribution for the observed data $t_{i,j}^l$ is readily obtained by integrating out the latent variables, giving

$$p\left(t_{i,j}^l\right) \sim N\left(0, V_i (V_i)^T + \Psi_i^l\right), \quad i = 1, 2. \tag{21}$$

Suppose that $\tau_j^l = [(t_{1,j}^l)^T (t_{2,j}^l)^T]^T \in \mathfrak{R}^{(n_1+n_2)}$, $V = [(V_1)^T (V_2)^T]^T \in \mathfrak{R}^{(n_1+n_2)\times n'}$, $\Psi^l = \begin{pmatrix} \Psi_1^l & 0 \\ 0 & \Psi_2^l \end{pmatrix}$, and $\Sigma^l = VV^T + \Psi^l$. Therefore, $P(\tau_j^l)$ can be obtained as follows:

$$P\left(\tau_j^l\right) = N\left(0, \Sigma^l\right), \tag{22}$$

where $\Sigma^l = VV^T + \Psi^l$. It can be shown that the negative log likelihood of the left projected data is equal to

$$L = \Sigma_{n=1}^{N} \Sigma_{j=1}^{m'} \log P\left(\tau_{n,j}^l\right). \tag{23}$$

After some manipulations, equation (23) becomes

$$\begin{aligned} L &= \frac{(n_1+n_2)Nm'}{2}\log(2\pi) + \frac{Nm'}{2}\log\left|\Sigma^l\right| \\ &\quad + \frac{1}{2}\Sigma_{n=1}^{N}\Sigma_{j=1}^{m'} \operatorname{tr}\left(\left(\Sigma^l\right)^{-1}\tau_{n,j}^l\left(\tau_{n,j}^l\right)^T\right) \\ &= \frac{(n_1+n_2)Nm'}{2}\log(2\pi) + \frac{Nm'}{2}\log\left|\Sigma^l\right| \\ &\quad + \frac{Nm'}{2}\operatorname{tr}\left(\left(\Sigma^l\right)^{-1}\widetilde{\Sigma}^l\right), \end{aligned} \tag{24}$$

where $\widetilde{\Sigma}^l = (1/Nm')\Sigma_{n=1}^{N}\Sigma_{j=1}^{m'}\tau_{n,j}^l(\tau_{n,j}^l)^T$ is the sample covariance matrix of left projected data, and $|A|$ denotes the determinant of matrix A. For log likelihood not become infinite, we assume $\Sigma^l \succ 0$. Figure 2(a) depicts the left probabilistic graphical model.

In this stage, we should maximize L with differentiating with respect to V, Ψ_1^l, and Ψ_2^l, where the solutions is straightforward. As shown in [4], the solutions can be obtained as

$$\begin{aligned} \widetilde{V}_1 &= \left(\widetilde{\Sigma}_{11}^l\right)^{1/2} Q_1^l \left(P^l\right)^{1/2} = \widetilde{\Sigma}_{11}^l R_1 \left(P^l\right)^{1/2}, \\ \widetilde{V}_2 &= \left(\widetilde{\Sigma}_{22}^l\right)^{1/2} Q_2^l \left(P^l\right)^{1/2} = \widetilde{\Sigma}_{22}^l R_2 \left(P^l\right)^{1/2}, \end{aligned} \tag{25}$$

where Q_1^l and Q_2^l are composed of n' first left-right singular vectors of the matrix, $(\widetilde{\Sigma}_{11}^l)^{-1/2}\widetilde{\Sigma}_{12}^l(\widetilde{\Sigma}_{22}^l)^{-1/2}$ with corresponding singular values on the diagonal of the matrix, $P^l \in \mathfrak{R}^{n' \times n'}$, and the matrices R_1 and R_2 are composed of first n' canonical directions. Note that the size of matrix $(\widetilde{\Sigma}_{11}^l)^{-1/2}\widetilde{\Sigma}_{12}^l(\widetilde{\Sigma}_{22}^l)^{-1/2}$ is $n_1 \times n_2$ which is much smaller than the size of the matrix $(\widetilde{\Sigma}_{11})^{-1/2}\widetilde{\Sigma}_{12}(\widetilde{\Sigma}_{22})^{-1/2} \in \mathfrak{R}^{m_1 n_1 \times m_2 n_2}$ in classical probabilistic CCA.

After computing V_1 and V_2, the observations are projected onto these matrices. The right probabilistic model is defined as

$$T_i^r = X^r + \Xi_i^r, \quad i = 1, 2, \tag{26}$$

where $T_i^r = T_i V_i$. $X^r = X$ is the latent matrix and Ξ_i^r represents the noise source in this model. Similar to left probabilistic model, we define $t_{i,j}^r \in \mathfrak{R}^{m_i}$, and $x_j^r \in \mathfrak{R}^{m'}$ as the jth column vector of T_i^r and X^r, respectively, where the marginal distribution of x_j^r is $N(0, I)$. Let $\tau_j^r = [(t_{1,j}^r)^T (t_{2,j}^r)^T]^T \in \mathfrak{R}^{(m_1+m_2)}$, $U = [(U_1)^T (U_2)^T]^T \in \mathfrak{R}^{(m_1+m_2)\times m'}$, $\Psi^r = \begin{pmatrix} \Psi_1^r & 0 \\ 0 & \Psi_2^r \end{pmatrix}$, and $\Sigma^r = UU^T + \Psi^r$. Therefore, $P(\tau_j^r) = N(0, \Sigma^r)$. The negative log likelihood of the right projected data is equal to

$$\begin{aligned} L &= \frac{(m_1+m_2)Nn'}{2}\log(2\pi) + \frac{Nn'}{2}\log|\Sigma^r| \\ &\quad + \frac{1}{2}\Sigma_{n=1}^{N}\Sigma_{j=1}^{n'} \operatorname{tr}\left((\Sigma^r)^{-1}\tau_{n,j}^r\left(\tau_{n,j}^r\right)^T\right) \\ &= \frac{(m_1+m_2)Nn'}{2}\log(2\pi) + \frac{Nn'}{2}\log|\Sigma^r| \\ &\quad + \frac{Nn'}{2}\operatorname{tr}\left((\Sigma^r)^{-1}\widetilde{\Sigma}^r\right), \end{aligned} \tag{27}$$

where $\widetilde{\Sigma}^r = (1/Nn')\Sigma_{n=1}^{N}\Sigma_{j=1}^{n'}\tau_{n,j}^r(\tau_{n,j}^r)^T$ is the sample covariance matrix of right projected data samples, and assume $\Sigma^r \succ 0$. The solution to this optimization can be obtained as

$$\begin{aligned} \widetilde{U}_1 &= \left(\widetilde{\Sigma}_{11}^r\right)^{1/2} Q_1^r (P^r)^{1/2} = \widetilde{\Sigma}_{11}^r L_1 (P^r)^{1/2}, \\ \widetilde{U}_2 &= \left(\widetilde{\Sigma}_{22}^r\right)^{1/2} Q_2^r (P^r)^{1/2} = \widetilde{\Sigma}_{11}^r L_2 (P^r)^{1/2}, \end{aligned} \tag{28}$$

where in this case L_1 and L_2 contain the first m' canonical directions, Q_1^r and Q_2^r are composed of m' first left-right singular vectors of $(\widetilde{\Sigma}_{11}^r)^{-1/2}\widetilde{\Sigma}_{12}^r(\widetilde{\Sigma}_{22}^r)^{-1/2}$, and $P^r \in \mathfrak{R}^{m' \times m'}$ contains the corresponding singular values on the diagonal. The right graphical probabilistic model is shown in Figure 2(b).

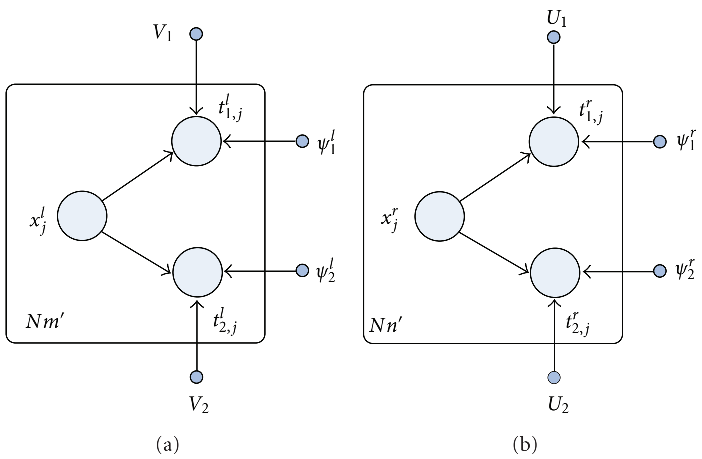

FIGURE 2: Probabilistic graphical model for 2DCCA, (a) left model and (b) right model.

It can be seen that the left and right canonical directions of 2DCCA can be obtained by maximizing the likelihood function. Posterior expectations can be obtained as follows:

$$E\left(x^l_j \mid \tau^l_j\right) = \left(\left(P^l\right)^{1/2}\right)^T L^T_1 \tau^l_j,$$
$$E\left(x^r_j \mid \tau^r_j\right) = \left((P^r)^{1/2}\right)^T R^T_1 \tau^r_j. \tag{29}$$

5. Relationship of 2DCCA and 2DLDA

In this section, we show that 2DCCA and 2DLDA [11] are closely related. 2DLDA uses original sample matrices for constructing between-class and within-class covariance matrices. It adapts an iterative algorithm where, in each iteration one projection direction is assumed known, and other projection is obtained by solving generalized eigenvalue problem. Let $X^l_j \in \mathfrak{R}^{r' \times c}$, $j = 1, \dots, N$ the N image samples which are projected onto the left projection matrix, and $L \in \mathfrak{R}^{r \times r'}$. These samples are clustered into C classes with $y_j \in \{1 \cdots C\}$ class label where ith class has n_i data samples. Define $\overline{X}^l_i \in \mathfrak{R}^{r' \times c}$ as the mean of ith class, π as a vector where its ith element is $\pi_i = n_i/N$, and $N = \Sigma^C_{j=1} n_j$.

Lemma 1. *In 2DLDA, between class scatter matrix is obtained as* $SB^l = MPM^T$, *where* $P = (\text{diag}(\pi) \otimes I_{r'} - (\pi \otimes I_{r'})(\pi \otimes I_{r'})^T)$, $M^l = ((\overline{X}^l_1)^T, \dots, (\overline{X}^l_C)^T) \in \mathfrak{R}^{c \times (r'C)}$, $\otimes$ *is the kronecker product,* diag(π) *is a diagonal matrix with* π_i*'s on its diagonal, and I is the identity matrix.*

The proof of lemma is shown in Appendix A. Consider two sets of multivariate data, $\{T^j_1 = (X^l_j)^T \in R^{c \times r'}, j = 1, \dots, N\}$ and $\{T^j_2 = [Q_1, \dots, Q_C]^T \in R^{Cr' \times r'},\ j = 1, \dots, N\}$ which are realizations of random matrices T_1 and T_2, respectively. Where $Q_i = I_{r'}$ if $y_j = i$, and otherwise $Q_i = 0_{r'}$.

For example, for image matrix X^l_1 with class label $y_1 = 2$, T^1_1 and T^1_2 are defined as follows:

$$T^1_1 = \left(X^l_1\right)^T, \qquad T^1_2 = \begin{pmatrix} 0_{r'} \\ I_{r'} \\ 0_{r'} \\ \vdots \\ 0_{r'} \end{pmatrix}_{(r'C) \times r'}, \tag{30}$$

where $0_{r'}$ is a $r' \times r'$square matrix of all zeros. The following lemma shows the relationship between two methods.

Lemma 2. *2DCCA finds the optimal correlation directions of* $T_1 = (X^l)^T$ *and* $T_2 = [Q_1, \dots, Q_C]^T$ *random matrices by solving the generalized eigenvalue problem* $SB^l u = (\lambda/1 - \lambda) SW^l u$, *where* SB^l *and* SW^l *are between-class and within-class covariance matrices, respectively.*

The proof is shown in Appendix B. As we know, The right projection vector of 2DLDA is computed using generalized eigenvalue problem $SB^l w = \lambda SW^l w$, while in Lemma 2, we proved that the canonical correlation direction of 2DCCA for (T_1, T_2) is obtained by solving the generalized eigenvalue problem $SB^l u = (\lambda/1 - \lambda) SW^l u$. These show the relationship between two methods. Therefore, the proposed probabilistic model can also be used for modeling 2DLDA technique.

6. Conclusion

Conventional probabilistic model only works for vectors data while the data samples in computer vision applications are matrices. In this paper, we presented a probabilistic interpretation of matrix-based canonical correlation analysis. We introduced a model and expressed that two-dimensional canonical correlation directions could be archived using maximum likelihood parameter estimation. This model can be applied for extending the matrix based CCA. We also showed that matrix-based Linear Discriminant Analysis can be obtained by setting the input random matrices of CCA.

Appendices

A. Proof of Lemma 1

$$\begin{aligned} SB^l &= \sum_{i=1}^{C} \pi_i \left(\overline{X}_i - \overline{X}\right)\left(\overline{X}_i - \overline{X}\right)^T \\ &= \sum_{i=1}^{C} \pi_i \left(\overline{X}_i\right)\left(\overline{X}_i\right)^T - \overline{X}\sum_{i=1}^{C} \pi_i \left(\overline{X}_i\right)^T - \sum_{i=1}^{C} \pi_i \left(\overline{X}_i\right)\left(\overline{X}\right)^T \\ &\quad + \sum_{i=1}^{C} \pi_i \left(\overline{X}\right)\left(\overline{X}\right)^T. \end{aligned} \tag{A.1}$$

By substituting $\overline{X} = \sum_{i=1}^{C} \pi_i(\overline{X}_i)$ and $\sum_{i=1}^{C} \pi_i = 1$ in to above equation, we have

$$SB^l = \sum_{i=1}^{C} \pi_i \left(\overline{X}_i\right)\left(\overline{X}_i\right)^T - \left(\overline{X}\right)\left(\overline{X}\right)^T. \tag{A.2}$$

The following equations can be easily obtained:

$$\begin{aligned} \overline{X} &= M(\pi \otimes I_{r'}), \\ \sum_{i=1}^{C} \pi_i \left(\overline{X}_i\right)\left(\overline{X}_i\right)^T &= M(\operatorname{diag}(\pi) \otimes I_{r'})M^T. \end{aligned} \tag{A.3}$$

So, we have

$$\begin{aligned} SB^l &= M(\operatorname{diag}(\pi) \otimes I_{r'})M^T - M(\pi \otimes I_{r'})(\pi \otimes I_{r'})^T M^T \\ &= M\Big((\operatorname{diag}(\pi) \otimes I_{r'}) - (\pi \otimes I_{r'})(\pi \otimes I_{r'})^T\Big)M^T \\ &= MPM^T. \end{aligned} \tag{A.4}$$

B. Proof of Lemma 2

Joint sample covariance matrix of T_1 and T_2 is computed as

$$\Sigma = \begin{pmatrix} SB^l + SW^l & MP \\ PM^T & P \end{pmatrix}. \tag{B.1}$$

2DCCA obtains the optimal canonical directions by finding the eigenvectors of $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$ which by some computations, we have

$$\begin{aligned} \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2} &= \left(SB^l + SW^l\right)^{-1/2} \mathrm{MPM}^T \\ &\quad \times \left(SB^l + SW^l\right)^{-1/2} \\ &= \left(SB^l + SW^l\right)^{-1/2} SB^l \left(SB^l + SW^l\right)^{-1/2} \end{aligned} \tag{B.2}$$

which is equivalent to solving the generalized eigenvalue problem $SB^l u = \lambda(SB^l + SW^l)u$ which is equal to $SB^l u = (\lambda/1 - \lambda)SW^l u$.

Acknowledgment

This paper is partially supported by Iran Telecommunication Research Center (ITRC).

References

[1] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 61, no. 3, pp. 611–622, 1999.

[2] N. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, 2005.

[3] S. Roweis, "Em algorithms for pca and spca," in *Advances in Neural Information Processing Systems*, pp. 626–632, MIT Press, Cambridge, Mass, USA, 1998.

[4] F. Bach and M. Jordan, "A probabilistic interpretation of canonical correlation analysis," Tech. Rep. 688, University of California, Berkeley, Calif, USA, 2005.

[5] S. Yu, K. Yu, V. Tresp, H. P. Kriegel, and M. Wu, "Supervised probabilistic principal component analysis," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*, pp. 464–473, ACM Press, August 2006.

[6] Y. Zhang and D.-Y. Yeung, "Heteroscedastic probabilistic linear discriminant analysis withsemi-supervised extension," in *Learning and Knowledge Discovery in Databases*, vol. 5782 of *Lecture Notes in Computer Science*, pp. 602–616, 2009.

[7] Z. Zhao, L. Sun, S. Yu, H. Liu, and J. Ye, "Multiclass probabilistic kernel discriminant analysis," in *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI '09)*, pp. 1363–1368, 2009.

[8] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proceedings of the European Conference on Computer Vision (ECCV '06)*, vol. 3954 of *Lecture Notes in Computer Science*, pp. 531–542, 2006.

[9] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[10] C. Bishop, M. Mozer, M. Jordan, and T. Petche, "Bayesian pca," *Advances in Neural Information Processing Systems*, vol. 9, pp. 382–388, 1996.

[11] J. Ye, "Generalized low rank approximations of matrices," *Machine Learning*, vol. 61, no. 1–3, pp. 167–191, 2005.

[12] S. H. Lee and S. Choi, "Two-dimensional canonical correlation analysis," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 735–738, 2007.

[13] N. Sun, Z. H. Ji, C. R. Zou, and LI. Zhao, "Two-dimensional canonical correlation analysis and its application in small sample size face recognition," *Neural Computing and Applications*, vol. 19, no. 3, pp. 377–382, 2010.

[14] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS '05)*, pp. 1569–1576, 2005.

[15] D. Cai, X. He, and J. Han, "Subspace learning based on tensor analysis," Tech. Rep. UIUCDCS-R-2005-2572, University of Illinois, Champaign, Ill, USA, 2005.

[16] X. He, D. Cai, and P. Niyogi, "Tensor subspace analysis," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS '05)*, pp. 499–506, 2005.

[17] S. Yan, D. Xu, B. Zhang, and H. J. Zhang, "Graph embedding: a general framework for dimensionality reduction," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 830–837, June 2005.

[18] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H. J. Zhang, "Multilinear discriminant analysis for face recognition," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 212–220, 2007.

[19] T. K. Kim and R. Cipolla, "Canonical correlation analysis of video volume tensors for action categorization and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1415–1428, 2009.

[20] M. Safayani and M. M. Shalmani, "Heteroscedastic multilinear discriminant analysis for face recognition," in *Proceeding of the International Conference on Pattern Recognition (ICPR '10)*, pp. 4287–4290, 2010.

[21] X. Xie, S. Yan, J. T. Kwok, and T. S. Huang, "Matrix-variate factor analysis and its applications," *IEEE Transactions on Neural Networks*, vol. 19, no. 10, pp. 1821–1826, 2008.

[22] D. Tao, M. Song, X. Li et al., "Bayesian tensor approach for 3-D face modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 10, pp. 1397–1410, 2008.

[23] D. Tao, J. Sun, X. Wu et al., "Probabilistic tensor analysis with akaike and bayesian information criteria," in *Proceedings of the International Conference on Neural Information Processing (ICONIP '07)*, pp. 791–801, 2007.

[24] D. Tao, J. Sun, J. Shen et al., "Bayesian tensor analysis," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '08)*, pp. 1402–1409, June 2008.