Probabilistic Principal Component Analysis for 2D data

Zhao, Jianhua
Yunnan University of Finance and Economics, School of Statistics and Mathematics
Kunming, 650221, China
E-mail: jhzhao.ynu@gmail.com

Yu, Philip L.H.
The University of Hong Kong, Department of Statistics and Actuarial Science
Pokfulam Road, Hong Kong
E-mail: plhyu@hku.hk

Kwok, James T.
Hong Kong University of Science and Technology, Department of Computer Science and Engineering
Clear Water Bay, N.T., Hong Kong
E-mail: jamesk@cse.ust.hk

1 Introduction

Probabilistic modeling for dimension reduction is a central research area in statistics, data mining, pattern recognition and machine learning. Compared with non-probabilistic counterparts, probabilistic models enables different sources of uncertainty inherent in the data to be well studied by means of probability theory. Principal component analysis (PCA) (Jolliffe, 2002) is one of most popular techniques for dimension reduction. Due to the non-probabilistic nature of PCA, Moghaddam and Pentland (1997) formulates PCA in a probabilistic framework and Tipping and Bishop (1999) derives the probabilistic PCA (PPCA) from the classical linear latent variable model. PPCA is an important development of PCA since it inherits all the advantages as a probabilistic model and includes PCA as a special case.

However, PPCA is simply formulated for 1D data (in which observations are in vector form). To apply PPCA for 2D data (in which observations are in matrix form), one possible solution is applying PPCA to vectorized data. However, this might not obtain the result as expected because the vectorization breaks the natural matrix structure, which may incur loss of the potentially more compact or useful representation (Ye et al., 2004). Moreover, for 2D data such as images, the resulting 1D data (typically, over tens of thousands) by vectorization is easily trapped into the so-called *curse of dimensionality* which could degenerate the performance of PPCA (Xie et al., 2008).

In recent years, several novel tools have been proposed to perform dimension reduction using 2D data directly. For example, 2DPCA (Yang et al., 2004), generalized low rank approximation of matrices (GLRAM) (Ye, 2005), etc. Later, to attain the similar advantages as PPCA enjoys over PCA, a probabilistic formulation for GLRAM called probabilistic second-order PCA (PSOPCA) has been proposed (Yu et al., 2008; Xie et al., 2008). However, unlike the relationship between PPCA and PCA (Tipping and Bishop, 1999), the theoretical relationship between PSOPCA and GLRAM has not been full established yet.

In this paper, we propose a novel probabilistic PCA model for 2D data (2DPPCA) to address these problems. Different from PSOPCA, which is based on minimum-error formulation, 2DPPCA is a development of PPCA for 2D data based on maximum-variance formulation. Due to their different

formulations, it is expected that the strength of PSOPCA depends on Euclidean distance while the strength of 2DPPCA relies on Mahalanobis distance. The remainder of the paper is organized as follows. Sec. 2 briefly reviews some related works. Sec. 3 proposes 2DPPCA model. Sec. 4 gives some empirical studies to compare 2DPPCA and some related competitors. Sec. 5 closes the paper with a conclusion.

2 Related works

2.1 PPCA

Let $\mathbf{x}$ be a d-dimensional data vector. PPCA model is defined as a *restricted factor analysis* model (Tipping and Bishop, 1999)

$$(1) \quad \begin{cases} \mathbf{x} = \mathbf{C}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \\ \mathbf{z} \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}_d(\mathbf{0}, \sigma^2\mathbf{I}), \end{cases}$$

where $\mathbf{z}$ is a q-dimensional whitened latent representations and assumed to be independent of $\boldsymbol{\epsilon}$, $\boldsymbol{\mu}$ is a d-dimensional mean vector, $\mathbf{C}$ is a $d \times q$ factor loadings matrix, and isotropic noise variance $\sigma^2 > 0$, $\mathbf{I}$ is an identity matrix whose dimension should be apparent from the context.

Under model (1), the probability distribution of $\mathbf{x}$ and the conditional probability distribution of $\mathbf{z}$ given $\mathbf{x}$ follow multivariate normal: $\mathbf{x} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{z}|\mathbf{x} \sim \mathcal{N}_q\left(\mathbf{M}^{-1}\mathbf{C}'(\mathbf{x} - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}\right)$, where $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}' + \sigma^2\mathbf{I}, \quad \mathbf{M} = \mathbf{C}'\mathbf{C} + \sigma^2\mathbf{I}$.

2.2 Minimum-error formulation for bilinear dimension reduction

For 2D data compression, e.g. image or image patches, it is often expected that an observation $\mathbf{X} \in \mathbb{R}^{d_c \times d_r}$ is projected onto a smaller one $\mathbf{T} \in \mathbb{R}^{q_c \times q_r}$ with $q_c < d_c$ and $q_r < d_r$ while reserving the interesting information as much as possible. To this end, Ye (2005) proposed a method called generalized low rank approximation of matrices (GLRAM). Let $\{\mathbf{X}_n\}_{n=1}^N$ be a set of 2D observations and $||\cdot||_F$ denote the Frobenius norm. Assume the data has been centered w.r.t. $\bar{\mathbf{X}}$ given by $\bar{\mathbf{X}} = \frac{1}{N}\sum_{n=1}^N \mathbf{X}_n$, i.e., $\mathbf{X}_n = \mathbf{X}_n - \bar{\mathbf{X}}$. GLRAM finds optimal transformation matrices $\mathbf{U}_c \in \mathbb{R}^{d_c \times q_c}(q_c < d_c), \mathbf{U}_r \in \mathbb{R}^{d_r \times q_r}(q_r < d_r)$, and projected low-dimensional representations $\mathbf{T}_n \in \mathbb{R}^{q_c \times q_r}, n = 1, \ldots, N$ to minimize MSE of all reconstructed observations given by $\frac{1}{N}\sum_{n=1}^N ||\mathbf{X}_n - \mathbf{U}_c\mathbf{T}_n\mathbf{U}_r'||_F^2$, under the condition that $\mathbf{U}_c'\mathbf{U}_c = \mathbf{I}, \mathbf{U}_r'\mathbf{U}_r = \mathbf{I}$. Note that 2DPCA can be viewed as a special case of GLRAM.

2.3 Probabilistic extensions of GLRAM

Recently, several works attempt to formulate a probabilistic model for GLRAM to attain the similar advantages PPCA enjoys over PCA, e.g. Yu et al. (2008); Xie et al. (2008). These works formulate the same model (called PSOPCA in Yu et al. (2008)) but use different algorithms:

$$(2) \quad \begin{cases} \mathbf{X} = \mathbf{C}\mathbf{Z}\mathbf{R} + \mathbf{W} + \boldsymbol{\epsilon}, \\ \mathbf{Z} \sim \mathcal{N}_{q_c,q_r}(\mathbf{0}, \mathbf{I}, \mathbf{I}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}_{d_c,d_r}(\mathbf{0}, \sigma^2\mathbf{I}, \sigma^2\mathbf{I}), \end{cases}$$

where $\mathcal{N}_{q_c,q_r}$ and $\mathcal{N}_{d_c,d_r}$ denote matrix-variate normals (Gupta and Nagar, 1999). This formulation looks like PPCA (1), following from the classical linear latent variable model.

3 2DPPCA model

In this section, we propose 2DPPCA model, which is formulated as a bilinear latent variable model.

We assume that the covariance matrix of 2D data $\mathbf{X}$ is separable, as

$$\text{(3)} \quad \text{cov}(\text{vec}(\mathbf{X})) = \mathbf{\Sigma}_r \otimes \mathbf{\Sigma}_c,$$

where $\mathbf{\Sigma}_r$ and $\mathbf{\Sigma}_c$ are row and column covariance matrices respectively. Actually, separable covariance models have been successfully used in many applications where the structure of the problem suggests such an assumption. Examples include spatial-temporal modeling for environmental data (Mardia and Goodall, 1993), channel modeling for multiple-input multiple-out communications (Werner and Jansson, 2009), signal modeling of MEG/EEG data (de Munck et al., 2002), etc.

For the aim of dimension reduction, we follow the idea of classical factor analysis model to assign factor structures on $\mathbf{\Sigma}_c$ and $\mathbf{\Sigma}_r$, respectively. In this paper, factor structures with isotropic noises are studied, i.e.,

$$\text{(4)} \quad \mathbf{\Sigma}_c = \mathbf{C}\mathbf{C}' + \sigma_c^2\mathbf{I}, \quad \mathbf{\Sigma}_r = \mathbf{R}\mathbf{R}' + \sigma_r^2\mathbf{I},$$

where $\mathbf{C} : d_c \times q_c (q_c < d_c)$ and $\mathbf{R} : d_r \times q_r (q_r < d_r)$. For simplicity, normal distribution is assumed. All assumptions we make for 2DPPCA are summarized as follows.

(A1) Separable covariance matrix (3).

(A2) Normal distribution.

(A3) Factor structures with isotropic noises (4).

The resulting model under (A1-A3) is called 2DPPCA, which can be formulated as a bilinear latent variable model as follows.

$$\text{(5)} \quad \begin{cases} \mathbf{X} = \mathbf{C}\mathbf{Z}\mathbf{R}' + \mathbf{W} + \mathbf{C}\boldsymbol{\epsilon}_r + \boldsymbol{\epsilon}_c\mathbf{R}' + \boldsymbol{\epsilon}, \\ \mathbf{Z} \sim \mathcal{N}_{q_c,q_r}(\mathbf{0}, \mathbf{I}, \mathbf{I}), \quad \boldsymbol{\epsilon}_r \sim \mathcal{N}_{q_c,d_r}(\mathbf{0}, \mathbf{I}, \sigma_r^2\mathbf{I}), \\ \boldsymbol{\epsilon}_c \sim \mathcal{N}_{d_c,q_r}(\mathbf{0}, \sigma_c^2\mathbf{I}, \mathbf{I}), \boldsymbol{\epsilon} \sim \mathcal{N}_{d_c,d_r}(\mathbf{0}, \sigma_c^2\mathbf{I}, \sigma_r^2\mathbf{I}) \end{cases}$$

where latent matrix $\mathbf{Z}$, column noise $\boldsymbol{\epsilon}_c (d_c \times q_r)$, row noise $\boldsymbol{\epsilon}_r (q_c \times d_r)$, and common noise $\boldsymbol{\epsilon} (d_c \times d_r)$ are assumed to be independent of each other. $\mathbf{C}(d_c \times q_c)$ and $\mathbf{R}(d_r \times q_r)$ are column and row factor loadings matrices, respectively. Noise variances $\sigma_c^2 > 0$ and $\sigma_r^2 > 0$.

2DPPCA (5) is different from PPCA (1) and PSOPCA (2). It implies a breakthrough from conventional 1D probabilistic model to the 2D one. To understand model (5) better, it is helpful to further introduce two latent matrices $\mathbf{Y}^r (q_c \times d_r), \mathbf{Y}_\epsilon^r (d_c \times d_r)$ to write model (5) in the form

$$\text{(6)} \quad \begin{cases} \mathbf{X} = \mathbf{C}\mathbf{Y}^r + \mathbf{W} + \mathbf{Y}_\epsilon^r, \\ \mathbf{Y}^r = \mathbf{Z}\mathbf{R}' + \boldsymbol{\epsilon}_r, \\ \mathbf{Y}_\epsilon^r = \boldsymbol{\epsilon}_c\mathbf{R}' + \boldsymbol{\epsilon}. \end{cases}$$

(6) is a two-stage representation of 2DPPCA. In stage 1, $\mathbf{X}$ is projected onto $\mathbf{Y}^r$ in column direction. In stage 2, $\mathbf{Y}^r$ and the residual $\mathbf{Y}_\epsilon^r$ are further projected onto $\mathbf{Z}$ and $\boldsymbol{\epsilon}_c$ in row direction. Equivalently, model (6) can be rewritten as a first projection in the row direction followed by one in the column direction.

It can be verified from (5) and (6) that all $\mathbf{X}, \mathbf{Y}^r$, $\mathbf{Y}^r_\epsilon$, $\mathbf{Z}|\mathbf{Y}^r$, $\mathbf{Y}|\mathbf{X}$ follow matrix-variate normal, e.g.

$$\mathbf{X} \sim \mathcal{N}(\mathbf{W}, \mathbf{\Sigma}_c, \mathbf{\Sigma}_r). \quad (7)$$

where $\mathbf{\Sigma}_c$ and $\mathbf{\Sigma}_r$ are given by (4). Details about this can be found in our technical report (Zhao et al., 2011).

3.1 Maximum Likelihood Estimation of 2DPPCA

Given a set of observations $\mathcal{X} = \{\mathbf{X}_n\}_{n=1}^N$, using the p.d.f. of $\mathbf{X}$ (7), we obtain that the global MLE of $\mathbf{W}$ is obviously the sample mean of $\mathcal{X}$ given by $\bar{\mathbf{X}} = \frac{1}{N}\sum_{n=1}^N \mathbf{X}_n$. Assume that the data has been centered, i.e., $\mathbf{X}_n = \mathbf{X}_n - \bar{\mathbf{X}}$. The MLE of $\boldsymbol{\theta} = (\mathbf{C}, \sigma_c^2, \mathbf{R}, \sigma_r^2)$ can be obtained by maximizing the log likelihood of 2DPPCA model, apart from a constant, given by

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{X}) = -\frac{1}{2}\sum_{n=1}^N \left\{ d_r \ln|\mathbf{\Sigma}_c| + d_c \ln|\mathbf{\Sigma}_r| + \mathrm{tr}\left(\mathbf{\Sigma}_c^{-1}\mathbf{X}_n\mathbf{\Sigma}_r^{-1}\mathbf{X}_n'\right)\right\}. \quad (8)$$

Due to the bilinear nature of 2DPPCA: given $(\mathbf{R}, \sigma_r^2)$, the model is linear w.r.t. $(\mathbf{C}, \sigma_c^2)$ and vice versa, it is natural to develop iterative procedures to maximize $\mathcal{L}$ in (8). A condition maximization (CM) algorithm to maximize $\mathcal{L}$ in (8) (Meng and Rubin, 1993) could be developed. Interested readers are referred to our technical report (Zhao et al., 2011) for details.

4 Experiments

In this section, we use real data to investigate the performance of 2DPPCA, GLRAM and PPCA in dimension reduction for classification. For comparison, the performance of PCA is also included. We use the following benchmark datasets:

- YALE[1] contains 165 face images of 15 individuals. Each person has 11 images in different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink. The image size is 160×121. Some images of two randomly chosen people are show in Fig. 1.
- PIX[2] consists of the images in the folder 'test-easy', containing 300 face images of 30 individuals. Each person has 10 images. We subsample the images to the size 100×100.

Figure 1: Images of two people in YALE.

For simplicity, the Nearest-Neighbors (NN) classifier based on the reduced features is employed for classification. To measure the misclassification error rate, we randomly split each data set into two

[1] http://cvc.yale.edu/projects/yalefaces/yalefaces.html.
[2] http://peipa.essex.ac.uk/ipa/pix/faces/manchester/.

parts: one part for training and the other for test. The training part consists of randomly chosen $r = 6$ or 8 images per individual with labels. We report the results from 10 replications.

Given an observed image $\mathbf{X}$, the reduced features for PCA and GLRAM are $\mathbf{U}'\text{vec}(\mathbf{X})(\mathbf{U}'\mathbf{U} = \mathbf{I})$ and $\mathbf{U}_c'\mathbf{X}\mathbf{U}_r$, respectively. For PPCA and 2DPPCA, the conditional expectations of latent representations are taken as the reduced features. Since the covariance matrices of latent representations are $\mathbf{I}$, classification in PPCA and 2DPPCA is actually based on Mahalanobis distance while classification in PCA and GLRAM is based on Euclidean distance. For all these methods, all possible dimensionalities of the reduced representation are tried and the best results are reported.

Tab. 1 shows the optimal averaged misclassification rates. The main observations include:

1. *Mahalanobis distance vs. Euclidean distance.* Mahalanobis distance-based methods 2DPPCA and PPCA substantially perform better than Euclidean distance-based methods GLRAM and PCA, respectively. The reveals the advantage of Mahalanobis distance over Euclidean distance for classification.

2. *2DPPCA vs. PPCA.* 2DPPCA is better than PPCA. This superiority of 2DPPCA should be attributed to the utilization of underlying 2D data structure.

Table 1: The averaged lowest error rates shown as mean±std. by different methods. Bold face indicates the best one.

Data	Method	$r = 6$	$r = 8$
YALE	2DPPCA	**10.8**±6.2	**7.2**±5.9
	PPCA	12.6±8.1	9.8±8.4
	GLRAM	17.0±9.7	15.2±11.9
	PCA	17.5±10.0	16.4±11.8
PIX	2DPPCA	**16.3**±4.0	**13.4**±3.2
	PPCA	19.8±4.6	15.4±4.2
	GLRAM	17.1±4.3	13.7±3.7
	PCA	19.6±4.3	15.9±3.7

5 Conclusion

To perform probabilistic dimension reduction for 2D data, we have proposed in this paper a bilinear probabilistic model called 2DPPCA. The novelty is that 2DPPCA signals a breakthrough from classical 1D latent variable model to the 2D one. The model parameters of 2DPPCA could be estimated by maximum likelihood method. Empirical studies with face recognition are investigated and the result reveals the advantages of 2DPPCA.

Acknowledgements

The work of J.H. Zhao is supported by a NSF of Yunnan (2010CD070) and partly by two small SF (YC10D028,YCT1013) from YNUFE. The work of P.L.H. Yu is supported by a small project funding from HKU and partly supported by GRF (HKU 706710P).

REFERENCES

de Munck, J., Huizenga, H., Waldorp, L., and Heethaar, R. (2002). Estimating stationary dipoles from MEG/EEG data contaminated with spatially and temporally correlated background noise. *IEEE Transactions on Signal Processing*, 50(7):1565–1572.

Gupta, A. K. and Nagar, D. K. (1999). *Matrix Variate Distributions*. Chapman and Hall-CRC.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, New York, 2 edition.

Mardia, K. and Goodall, C. (1993). Spatial-temporal analysis of multivariate environmental monitoring data. In *Multivariate Environmental Statistics*, pages 347–386. Amsterdam, The Netherlands:Elsevier.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278.

Moghaddam, B. and Pentland, A. (1997). Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61:611–622.

Werner, K. and Jansson, M. (2009). Estimating MIMO channel covariances from training data under the kronecker model. *Signal Processing*, 89(1):1–13.

Xie, X., Yan, S., Kwok, J., and Huang, T. (2008). Matrix-variate factor analysis and its applications. *IEEE Transactions on Neural Networks*, 19(10):1821–1826.

Yang, J., Zhang, D., Frangi, A. F., and Yang, J. (2004). Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):131–137.

Ye, J. (2005). Generalized low rank approximations of matrices. *Machine Learning*, 61:167–191.

Ye, J., Janardan, R., and Li, Q. (2004). GPCA: An efficient dimension reduction scheme for image compression and retrieval. In *Proc. 10th ACM SIGKDD Int. Cof. Knowl. Discovery Data Mining*.

Yu, S., Bi, J., and Ye, J. (2008). Probabilistic interpretations and extensions for a family of 2D PCA-style algorithms. In *the KDD'2008 Workshop on Data Mining using Matrices and Tensors*.

Zhao, J., Yu, P. L. H., and Kwok, J. T. (2011). Bilinear probabilistic principal component analysis. Technical report, School of Statistics and Mathematics, Yunnan University of Finance and Economics.