Biometrika (2012), **99**, 3, *pp.* 569–583

Printed in Great Britain

doi: 10.1093/biomet/ass019
Advance Access publication 22 May 2012

On multilinear principal component analysis of order-two tensors

BY HUNG HUNG

Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei 10055, Taiwan

hhung@ntu.edu.tw

PEISHIEN WU, IPING TU AND SUYUN HUANG

Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan

peishien1987@gmail.com iping@stat.sinica.edu.tw syhuang@stat.sinica.edu.tw

SUMMARY

Principal component analysis is commonly used for dimension reduction in analysing high-dimensional data. Multilinear principal component analysis aims to serve a similar function for analysing tensor structure data, and has empirically been shown effective in reducing dimensionality. In this paper, we investigate its statistical properties and demonstrate its advantages. Conventional principal component analysis, which vectorizes the tensor data, may lead to inefficient and unstable prediction due to the often extremely large dimensionality involved. Multilinear principal component analysis, in trying to preserve the data structure, searches for low-dimensional projections and, thereby, decreases dimensionality more efficiently. The asymptotic theory of order-two multilinear principal component analysis, including asymptotic efficiency and distributions of principal components, associated projections, and the explained variance, is developed. A test of dimensionality is also proposed. Finally, multilinear principal component analysis is shown to improve conventional principal component analysis in analysing the Olivetti faces dataset, which is achieved by extracting a more modularly oriented basis set in reconstructing the test faces.

Some key words: Asymptotic theory; Dimension reduction; Image reconstruction; Principal component analysis; Tensor.

1. INTRODUCTION

Principal component analysis is often used to achieve dimension reduction in the analysis of high-dimensional data. It searches for a transformation of the data onto a smaller set of a new coordinate system that retains maximal data variation. It has been applied in many scientific fields for dimension reduction and compact data representation (Jolliffe, 2002). When the data are tensor objects, traditional analysis generates a design matrix by vectorizing each of the tensor objects into a long vector. This usually produces a large number of variables, even where the available sample size is relatively small, and many existing statistical methods fail to work satisfactorily. A real dataset, the Olivetti faces, used in § 4, is typical. There are 400 images in this set and each contains 64×64 pixels. Vectorizing the images would contribute to a design matrix of size 400×4096.

One strategy for overcoming this difficulty is to take advantage of the natural tensor structure of the data, as is done in singular value decomposition. Given a matrix, i.e., an order-two tensor, singular value decomposition can decompose two directional spaces simultaneously. De Lathauwer et al. (2000a) generalized the singular value decomposition to an Nth-order tensor object $A \in \Re^{I_1 \times \cdots \times I_N}$. Further, they formulated the problem of best rank-$(R_1, \ldots, R_N)$ approximation of higher order tensors in a least squares sense (De Lathauwer et al., 2000b). Yang et al. (2004) proposed two-dimensional principal component analysis for analysing image data. An improved two-directional two-dimensional principal component analysis was developed in Zhang & Zhou (2005), which was shown through simulation studies to perform better than two-dimensional principal component analysis. Ye (2005) formulated the problem of generalized low rank approximation of matrices, which can be treated as a sampling extension of the best rank-(R_1, R_2) approximation for order-two tensors in De Lathauwer et al. (2000b). Lu et al. (2008) further generalized the work of Ye (2005) and proposed a multilinear principal component analysis for tensor objects of arbitrary order. Other tensor decomposition methods used for dimension reduction exist. Kolda & Bader (2009) provided a general overview of the current development of tensor decomposition methods used for unsupervised learning, including examples of their application and the software available for implementing them. Li et al. (2010) considered tensor decomposition methods for supervised learning such as regression and classification.

Like conventional principal component analysis, multilinear principal component analysis seeks low-dimensional multilinear projections of tensor objects that capture the maximal data variation. For the Olivetti faces, one eigenvector in conventional principal component analysis creates an image basis element having 4095 free parameters. By contrast, one image basis element in multilinear principal component analysis involves the Kronecker product of a column vector and a row vector, which contains only 126 free parameters. Due to the number of parameters required to specify one basis element, multilinear principal component analysis is expected to perform better than conventional principal component analysis when the sample size is small to moderate, as in this example. Compared to two-directional two-dimensional principal component analysis, multilinear principal component analysis can capture more signal variation and has less noise contamination in the chosen image basis, because of its estimation criterion. Multilinear principal component analysis has been successfully applied in real data analysis and its performance has been verified by simulations (Ye, 2005; Lu et al., 2008). Asymptotic properties regarding principal component analysis have been rigorously studied in Anderson (1963). To the best of our knowledge, however, no statistical justification has been given for multilinear principal component analysis nor has any asymptotic study demonstrated its advantages. In this paper, we establish some of its statistical properties including the asymptotic distributions of the principal components, the associated projections and the explained variance. It is also shown that multilinear principal component analysis is asymptotically more efficient than two-directional two-dimensional principal component analysis in estimating the target dimension reduction subspace. Furthermore, an adequacy test of dimensionality is developed.

Some notation is defined here for ease of reference. Let $\mathrm{vec}(\cdot)$ be the operator that stacks the columns of a matrix into a long vector and $\otimes$ be the Kronecker product. For any matrix M, $\mathrm{span}(M)$ is the linear space spanned by the columns of M, P_M is the orthogonal projection matrix onto $\mathrm{span}(M)$, $Q_M = I - P_M$, and M^+ is its Moore–Penrose generalized inverse.

2. MULTILINEAR PRINCIPAL COMPONENT ANALYSIS

2·1. *Statistical model*

Let $X \in \Re^{p \times q}$ be a tensor object of interest. We start from the following model for conventional principal component analysis:

$$\mathrm{vec}(X - \mu) = \Gamma \nu + \mathrm{vec}(\varepsilon), \tag{1}$$

where μ is the mean parameter of X, $\Gamma \in \Re^{m \times r_0}$ is an orthonormal basis with $m = pq$, and $\nu \in \Re^{r_0}$ is a random coordinate vector with $E(\nu) = 0$ and with full-rank diagonal covariance matrix. The error term $\varepsilon \in \Re^{p \times q}$ is independent of ν with $E(\varepsilon) = 0$ and $\mathrm{cov}\{\mathrm{vec}(\varepsilon)\} = \sigma^2 I_m$. Without considering the error term ε, $\Gamma\nu$ belongs to the r_0-dimensional subspace $\mathrm{span}(\Gamma)$. Principal component analysis aims to estimate a basis of $\mathrm{span}(\Gamma)$ for data compression.

Conventional principal component analysis for $\mathrm{vec}(X)$ can require a large number of parameters. Fortunately, we can apply the notion of Kronecker envelope (Li et al., 2010) to achieve parsimony. Following their Theorem 1, there must exist orthonormal matrices $A_0 \in \Re^{p \times p_0}$ and $B_0 \in \Re^{q \times q_0}$, with $\dim(A_0) = p_0 \leqslant p$ and $\dim(B_0) = q_0 \leqslant q$, such that $\mathrm{span}(B_0 \otimes A_0)$ is the unique minimal subspace that contains $\mathrm{span}(\Gamma)$. That is, there exists a full rank matrix $G \in \Re^{m_0 \times r_0}$ with $m_0 = p_0 q_0 \geqslant r_0$ such that $\Gamma = (B_0 \otimes A_0)G$. This subspace $\mathrm{span}(B_0 \otimes A_0)$ is called the Kronecker envelope of Γ. Let $U \in \Re^{p_0 \times q_0}$ be such that $\mathrm{vec}(U) = G\nu$. Then, since $\mathrm{vec}(A_0 U B_0^{\mathrm{T}}) = (B_0 \otimes A_0)\mathrm{vec}(U)$, model (1) becomes

$$\mathrm{vec}(X - \mu) = (B_0 \otimes A_0)\mathrm{vec}(U) + \mathrm{vec}(\varepsilon), \quad \text{or equivalently,} \quad X = \mu + A_0 U B_0^{\mathrm{T}} + \varepsilon. \tag{2}$$

The latter expression is more suitable for reflecting the matrix structure of X. Instead of using Γ, model (2) adopts $(B_0 \otimes A_0)$ as the basis for $\mathrm{vec}(X)$, or equivalently, adopts A_0 as the column basis and B_0 as the row basis for data compression of X. In the following study, we will assume the validity of model (2) and Condition 1 below:

Condition 1. The matrices $E(UU^{\mathrm{T}})$ and $E(U^{\mathrm{T}}U)$ are nonsingular and have distinct characteristic roots.

Under model (2), the Kronecker envelope $\mathrm{span}(B_0 \otimes A_0)$ is the target subspace for dimension reduction, and we will show that multilinear principal component analysis aims to estimate $\mathrm{span}(B_0 \otimes A_0)$. Note that $T = \mathrm{cov}\{\mathrm{vec}(U)\}$ need not be of full rank, as $\dim(T) = r_0 \leqslant m_0$; see § 2·3. To distinguish different dimensions under consideration, we call r_0 the effective dimension, and (p_0, q_0) the Kronecker dimension.

2·2. *Statistical properties*

Let $X_1, \ldots, X_n$ be independent copies of X. At the sample level, multilinear principal component analysis aims to extract the basis pairs that best approximate the X_i while preserving their tensor structure. In particular, for a prespecified dimensionality $(\tilde{p}, \tilde{q})$, Ye (2005) proposed to find $A \in \mathcal{O}_{p,\tilde{p}}$, $B \in \mathcal{O}_{q,\tilde{q}}$, and $\{U_i\}_{i=1}^n$ that minimize

$$\frac{1}{n}\sum_{i=1}^{n} \|(X_i - \bar{X}) - AU_iB^{\mathrm{T}}\|_F^2 = \frac{1}{n}\sum_{i=1}^{n} \|\mathrm{vec}(X_i - \bar{X}) - (B \otimes A)\mathrm{vec}(U_i)\|^2, \tag{3}$$

where $\bar{X} = n^{-1}\sum_{i=1}^n X_i$, $\|\cdot\|_F$ is the Frobenius norm of a matrix, $\|\cdot\|$ is the Euclidean norm of a vector, and $\mathcal{O}_{\ell,\tilde{\ell}}$ is the collection of all orthonormal matrices of size $\ell \times \tilde{\ell}$. By replacing $(B \otimes A)$ with $\Gamma \in \mathcal{O}_{m,\tilde{m}}$, (3) becomes a conventional principal component analysis. Thus, multilinear principal component analysis can be treated as constrained principal component analysis with the tensor constraint $\Gamma = (B \otimes A)$. The following theorem gives some useful properties of the minimizer of (3).

THEOREM 1 (Ye, 2005). *Let* $(\hat{A}, \hat{B})$ *and* $\{\hat{U}_i\}_{i=1}^n$ *be a minimizer of* (3) *under* $(\tilde{p}, \tilde{q})$. *Then,* (a) $\hat{U}_i = \hat{A}^{\mathrm{T}}(X_i - \bar{X})\hat{B}$; (b) $(\hat{A}, \hat{B})$ *is the maximizer of* $n^{-1}\sum_{i=1}^n \|A^{\mathrm{T}}(X_i - \bar{X})B\|_F^2$; (c) $\hat{A}$ *consists of the* $\tilde{p}$ *leading eigenvectors of* $n^{-1}\sum_{i=1}^n (X_i - \bar{X})P_{\hat{B}}(X_i - \bar{X})^{\mathrm{T}}$, *and* $\hat{B}$ *consists of the* $\tilde{q}$ *leading eigenvectors of* $n^{-1}\sum_{i=1}^n (X_i - \bar{X})^{\mathrm{T}}P_{\hat{A}}(X_i - \bar{X})$.

Similarly, we can define the population version of (3): $E\{\|(X-\mu)-AUB^{\mathrm{T}}\|_F^2\}$. Following Theorem 1(b), the corresponding minimizers over $A\in\mathcal{O}_{p,\tilde{p}}$ and $B\in\mathcal{O}_{q,\tilde{q}}$ will be equal to the maximizers of the maximization problem

$$\operatorname*{argmax}_{A\in\mathcal{O}_{p,\tilde{p}},B\in\mathcal{O}_{q,\tilde{q}}} E\{\|A^{\mathrm{T}}(X-\mu)B\|_F^2\}=\operatorname*{argmax}_{A\in\mathcal{O}_{p,\tilde{p}},B\in\mathcal{O}_{q,\tilde{q}}} \operatorname{tr}\{(B\otimes A)^{\mathrm{T}}\Sigma(B\otimes A)\},\tag{4}$$

where $\Sigma=\operatorname{cov}\{\operatorname{vec}(X)\}$. The following proposition gives the existence of the solution.

PROPOSITION 1. *For a fixed but arbitrary positive semidefinite $m\times m$ matrix Σ, a solution to the maximization problem,* $\operatorname{argmax}_{A\in\mathcal{O}_{p,\tilde{p}},B\in\mathcal{O}_{q,\tilde{q}}}\operatorname{tr}\{(B\otimes A)^{\mathrm{T}}\Sigma(B\otimes A)\}$, *exists.*

We do not need assumption (2) for Proposition 1, which also applies to problem (3) by replacing Σ with its sample estimate S_n, and by rephrasing the minimization problem into an equivalent maximization problem. With the existence of the maximizer in (4) we can formally define the tensor principal components and the multilinear principal component analysis subspace.

DEFINITION 1. *For a prespecified dimension $(\tilde{p},\tilde{q})$, assume (A,B) to be a solution to (4), where A and B are expressed in their columns as $A=[a_1,\ldots,a_{\tilde{p}}]$ and $B=[b_1,\ldots,b_{\tilde{q}}]$. We call $\{b_j\otimes a_i : i=1,\ldots,\tilde{p},\ j=1,\ldots,\tilde{q}\}$ the tensor principal components, and span$(B\otimes A)$ the multilinear principal component analysis subspace of dimensionality $(\tilde{p},\tilde{q})$.*

Similar to Theorem 1(c), A and B will consist of the leading $\tilde{p}$ and $\tilde{q}$ eigenvectors of $E\{(X-\mu)P_B(X-\mu)^{\mathrm{T}}\}$ and $E\{(X-\mu)^{\mathrm{T}}P_A(X-\mu)\}$. Since $E\{(X-\mu)P_B(X-\mu)^{\mathrm{T}}\}=\sum_{j=1}^{\tilde{q}}(b_j\otimes I_p)^{\mathrm{T}}\Sigma(b_j\otimes I_p)$ and $E\{(X-\mu)^{\mathrm{T}}P_A(X-\mu)\}=\sum_{i=1}^{\tilde{p}}(I_q\otimes a_i)^{\mathrm{T}}\Sigma(I_q\otimes a_i)$, we have, equivalently, that (A,B) consists of the leading solutions to the stationary equations

$$\left\{\sum_{j=1}^{\tilde{q}}(b_j\otimes I_p)^{\mathrm{T}}\Sigma(b_j\otimes I_p)\right\}a_i=\lambda_i a_i,\quad (i=1,\ldots,\tilde{p}),\tag{5}$$

$$\left\{\sum_{i=1}^{\tilde{p}}(I_q\otimes a_i)^{\mathrm{T}}\Sigma(I_q\otimes a_i)\right\}b_j=\xi_j b_j,\quad (j=1,\ldots,\tilde{q}),\tag{6}$$

over $A\in\mathcal{O}_{p\times\tilde{p}}$ and $B\in\mathcal{O}_{q\times\tilde{q}}$, where the ordering is by the eigenvalues $\lambda_1>\cdots>\lambda_{\tilde{p}}\geqslant 0$ and $\xi_1>\cdots>\xi_{\tilde{q}}\geqslant 0$ which under Condition 1 are simple roots for $\tilde{p}\leqslant p_0$ and $\tilde{q}\leqslant q_0$. The indices (i,j) in the above system of equations can go beyond $\tilde{p}$ and $\tilde{q}$ up to p and q, but those a_i and b_j with $i>\tilde{p}$ and $j>\tilde{q}$ will not be included in the solution pair (A,B). Obviously λ_i, a_i, ξ_j and b_j depend on Σ and the dimensionality $(\tilde{p},\tilde{q})$. For simplicity, we use λ_i, a_i, ξ_j and b_j unless we want to emphasize their dependence on $(\Sigma,\tilde{p},\tilde{q})$. The sample analogues $(\hat{A},\hat{B})$, $\hat{\lambda}_i$, and $\hat{\xi}_j$ are similarly defined by replacing Σ with S_n. Hereafter, with prespecified dimensionality $(\tilde{p},\tilde{q})$, we denote the maximizer of (4) by (A,B) and its sample analogue by $(\hat{A},\hat{B})$.

Finding conventional principal components is simply an eigenvalue problem, but such is not the case for $(\hat{A},\hat{B})$. We therefore adopt the following iterative alternating least squares approach (Ye, 2005) to obtain $(\hat{A},\hat{B})$.

Algorithm 1. Given a random initial $A^{(0)}\in\mathcal{O}_{p\times\tilde{p}}$, for $k=1,2,\ldots,$

Step 1. Obtain the maximizer $B^{(k+1)}=\operatorname{argmax}_{B\in\mathcal{O}_{q\times\tilde{q}}} n^{-1}\sum_{i=1}^n\|A^{(k)\mathrm{T}}(X_i-\bar{X})B\|_F^2$.

Step 2. Obtain the maximizer $A^{(k+1)} = \text{argmax}_{A \in \mathcal{O}_{p \times \tilde{p}}} n^{-1} \sum_{i=1}^{n} \| A^{\mathrm{T}}(X_i - \bar{X}) B^{(k+1)} \|_F^2$.

Step 3. Repeat Steps 1–2 until convergence between $n^{-1} \sum_{i=1}^{n} \| A^{(k)\mathrm{T}}(X_i - \bar{X}) B^{(k)} \|_F^2$ and $n^{-1} \sum_{i=1}^{n} \| A^{(k+1)\mathrm{T}}(X_i - \bar{X}) B^{(k+1)} \|_F^2$ is reached. Output $(\hat{A}, \hat{B}) = (A^{(k+1)}, B^{(k+1)})$.

For any fixed $A^{(k)}$ or $B^{(k+1)}$, the optimization problems in Steps 1 and 2 are standard eigenvalue problems. Hence, $B^{(k+1)}$ and $A^{(k+1)}$ can be easily obtained. Moreover, Algorithm 1 ensures that $n^{-1} \sum_{i=1}^{n} \| A^{(k)\mathrm{T}}(X_i - \bar{X}) B^{(k)} \|_F^2$ is monotonically increasing as k increases and, hence, it must converge since $n^{-1} \sum_{i=1}^{n} \| A^{(k)\mathrm{T}}(X_i - \bar{X}) B^{(k)} \|_F^2$ is bounded above by $n^{-1} \sum_{i=1}^{n} \| X_i - \bar{X} \|_F^2$. Because Algorithm 1 may find only a local maximum, multiple random initial values are suggested by Ye (2005) to ensure that the global maximum is found.

Unlike conventional principal component analysis, a hierarchical nesting structure may not exist for multilinear principal component analysis. More precisely, for any two pairs $(\tilde{p}', \tilde{q}')$ and $(\tilde{p}, \tilde{q})$ such that $\tilde{p}' \leqslant \tilde{p}$ and $\tilde{q}' \leqslant \tilde{q}$ with the corresponding solution pairs $(\hat{A}', \hat{B}')$ and $(\hat{A}, \hat{B})$, there is no guarantee that $\text{span}(\hat{A}') \subseteq \text{span}(\hat{A})$ or that $\text{span}(\hat{B}') \subseteq \text{span}(\hat{B})$. At the population level, however, there exist certain relationships between the target subspaces and the multilinear principal component analysis subspaces.

PROPOSITION 2. *Assume model* (2) *and Condition* 1. *Then,* (a) $\text{span}(A) \supseteq \text{span}(A_0)$ *and* $\text{span}(B) \supseteq \text{span}(B_0)$ *for* $\tilde{p} \geqslant p_0$ *and* $\tilde{q} \geqslant q_0$*;* (b) $\text{span}(A) \subset \text{span}(A_0)$ *and* $\text{span}(B) \supseteq \text{span}(B_0)$ *for* $\tilde{p} < p_0$ *and* $\tilde{q} \geqslant q_0$*;* (c) $\text{span}(A) \supseteq \text{span}(A_0)$ *and* $\text{span}(B) \subset \text{span}(B_0)$ *for* $\tilde{p} \geqslant p_0$ *and* $\tilde{q} < q_0$*;* (d) $\text{span}(A) \subset \text{span}(A_0)$ *and* $\text{span}(B) \subset \text{span}(B_0)$ *for* $\tilde{p} < p_0$ *and* $\tilde{q} < q_0$.

Proposition 2 ensures the existence of a nesting structure in which the extracted multilinear principal component analysis subspace is a proper subspace of the target subspace if the dimensionality is underspecified, and contains the target subspace if the dimensionality is overspecified. It also implies that multilinear principal component analysis does indeed search the Kronecker envelope $\text{span}(B_0 \otimes A_0)$ when $(\tilde{p}, \tilde{q}) = (p_0, q_0)$. As a result, it provides justification for using $\text{span}(\hat{B} \otimes \hat{A})$ at the sample level for dimension reduction.

2·3. *Connections with other dimension reduction methods*

Two-directional two-dimensional principal component analysis is another method of extracting bases for tensor objects. For a given dimensionality $(\tilde{p}, \tilde{q})$, let $A^* = [a_1^*, \ldots, a_{\tilde{p}}^*]$ and $B^* = [b_1^*, \ldots, b_{\tilde{q}}^*]$ be the leading $\tilde{p}$ and $\tilde{q}$ eigenvectors of $E\{(X - \mu)(X - \mu)^{\mathrm{T}}\}$ and $E\{(X - \mu)^{\mathrm{T}}(X - \mu)\}$, respectively, with the corresponding eigenvalues $\lambda_1^* > \cdots > \lambda_{\tilde{p}}^* \geqslant 0$ and $\xi_1^* > \cdots > \xi_{\tilde{q}}^* \geqslant 0$ which, following Condition 1, are simple roots for $\tilde{p} \leqslant p_0$ and $\tilde{q} \leqslant q_0$. Then, $\{b_j^* \otimes a_i^* : i = 1, \ldots, \tilde{p}, j = 1, \ldots, \tilde{q}\}$ are the population two-directional two-dimensional principal components at dimensionality $(\tilde{p}, \tilde{q})$. The sample analogues, denoted by $\hat{A}^*$, $\hat{B}^*$, $\hat{\lambda}_i^*$, and $\hat{\xi}_j^*$, are similarly defined by using $n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}}$ and $n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^{\mathrm{T}}(X_i - \bar{X})$ instead.

PROPOSITION 3. *Assume model* (2) *and Condition* 1. (a) *If* $\tilde{q} \geqslant q_0$*, then* $a_i = a_i^*$ *for* $i = 1, \ldots, \min\{p_0, \tilde{p}\}$*, and* $\lambda_i^* - \lambda_i = (q - \tilde{q})\sigma^2$ *for* $i = 1, \ldots, \tilde{p}$. (b) *If* $\tilde{p} \geqslant p_0$*, then* $b_j = b_j^*$ *for* $j = 1, \ldots, \min\{q_0, \tilde{q}\}$*, and* $\xi_j^* - \xi_j = (p - \tilde{p})\sigma^2$ *for* $j = 1, \ldots, \tilde{q}$.

When the dimensionality $(\tilde{p}, \tilde{q})$ is adequate, Proposition 3 implies that two-directional two-dimensional principal component analysis and multilinear principal component analysis, at the population level, target the same subspace $\text{span}(B_0 \otimes A_0)$. However, there is no guarantee that the extracted bases $(\hat{A}^*, \hat{B}^*)$ of two-directional two-dimensional principal component analysis

also maximize the sample version of (4). Although it is the case, from Proposition 3, that $E\{(X-\mu)P_B(X-\mu)^{\mathrm{T}}\}$ and $E\{(X-\mu)(X-\mu)^{\mathrm{T}}\}$ have the same leading eigenvectors, we still expect an efficiency gain in using $E\{(X-\mu)P_B(X-\mu)^{\mathrm{T}}\}$, since it is less noise-contaminated than $E\{(X-\mu)(X-\mu)^{\mathrm{T}}\}$. We will prove this in § 3.

There is also a connection between multilinear principal component analysis and conventional principal component analysis. First note that $\Sigma = \Gamma \mathrm{cov}(\nu)\Gamma^{\mathrm{T}} + \sigma^2 I_m$ under model (1). Without loss of generality, we may assume that $\mathrm{cov}(\nu)$ is diagonal, and it is obvious that conventional principal component analysis uses Γ as a new coordinate system for dimension reduction. On the other hand, under model (2), we have $\Sigma = (B_0 \otimes A_0)T(B_0 \otimes A_0)^{\mathrm{T}} + \sigma^2 I_m$, where $T = \mathrm{cov}\{\mathrm{vec}(U)\}$, and multilinear principal component analysis is shown to target the Kronecker envelope $\mathrm{span}(B_0 \otimes A_0)$ for dimension reduction. However, as T is neither necessarily a diagonal matrix nor of full rank, $(B_0 \otimes A_0)$ does not equal Γ in general. As mentioned in § 2·1, Γ takes the form $\Gamma = (B_0 \otimes A_0)G$ for rank-r_0 matrix $G \in \Re^{m_0 \times r_0}$. Thus, $\mathrm{span}(\Gamma)$ must be a subspace of $\mathrm{span}(B_0 \otimes A_0)$. Depending on the relationships between the effective dimension r_0 and the Kronecker dimension (p_0, q_0), we have different inclusion properties for these two subspaces. When $r_0 < m_0$, $\mathrm{span}(\Gamma)$ is a proper subspace of $\mathrm{span}(B_0 \otimes A_0)$, and a further dimension reduction for $\mathrm{vec}(X)$ under model (2) is possible. When $r_0 = m_0$, $\mathrm{span}(B_0 \otimes A_0)$ coincides with the minimum dimension reduction subspace $\mathrm{span}(\Gamma)$. In general, therefore, multilinear principal component analysis uses a larger subspace for dimension reduction. However, via imposing tensor structure, it requires fewer parameters to specify its low-dimensional subspace than the conventional approach does. See Remark 1 below for details.

We remind the reader that even if $\Gamma = (B_0 \otimes A_0)$, there is no obvious ordering relationship between tensor principal components and conventional principal components. This can be seen in a simple example with uncorrelated U_{ij}s, that is, with $T = \mathrm{diag}\{\mathrm{vec}(C)\}$, where C is a matrix with $C_{ij} = \mathrm{var}(U_{ij})$. In this situation, conventional principal component analysis and multilinear principal component analysis share the same eigenvectors. The leading m_0 eigenvalues of conventional principal component analysis are $\{C_{ij} + \sigma^2 : i = 1, \ldots, p_0, j = 1, \ldots, q_0\}$, which have a natural ordering depending on the C_{ij}s. On the other hand, the leading eigenvalues of multilinear principal component analysis at (p_0, q_0) are derived to be $\{\lambda_i = \sum_{j=1}^{q_0} C_{ij} + q_0\sigma^2 : i = 1, \ldots, p_0\}$ and $\{\xi_j = \sum_{i=1}^{p_0} C_{ij} + p_0\sigma^2 : j = 1, \ldots, q_0\}$, where the ordering depends on the column sums and row sums of C. Hence, even if a_i and b_j are leading eigenvectors of A_0 and B_0, there is no guarantee that $(b_j \otimes a_i)$ is the leading eigenvector in Γ.

Remark 1. The number of free parameters required for multilinear principal component analysis is $\{pp_0 - p_0(p_0+1)/2\} + \{qq_0 - q_0(q_0+1)/2\} = O(pp_0 + qq_0)$, while it is $\{pqr_0 - r_0(r_0+1)/2\} = O(pqr_0)$ for conventional principal component analysis.

Multilinear principal component analysis is an unsupervised dimension reduction method with tensor structure. Recently, Li et al. (2010) proposed dimension folding, which is a supervised dimension reduction method with tensor structure imposed on the target subspace. One common point is that both methods use Kronecker structure to achieve a parsimonious usage of parameters. Moreover, the objective function (3) has a similar form as that of dimension folding, and the minimization of the objective function is solved in both cases through iterative alternating least squares algorithms.

2·4. *Selection of dimensionality*

The Kronecker dimension of multilinear principal component analysis can be determined by the explained proportion of total variance.

DEFINITION 2. *Assume that* (4) *has a unique global solution* (A, B). *We call the quantity* $\Phi(\tilde{p}, \tilde{q}) = E\{\|A^{\mathrm{T}}(X-\mu)B\|_F^2\}$ *the cumulative variance and* $\rho(\tilde{p}, \tilde{q}) = \Phi(\tilde{p}, \tilde{q})/\Phi(p, q)$ *the explained proportion of total variance of* X *at dimension* $(\tilde{p}, \tilde{q})$. *Note that* $\Phi(p, q) = E\{\|X-\mu\|_F^2\}$. *The corresponding sample analogues are defined to be* $\hat{\Phi}(\tilde{p}, \tilde{q}) = n^{-1}\sum_{i=1}^n \|\hat{A}^{\mathrm{T}}(X_i - \bar{X})\hat{B}\|_F^2$, $\hat{\Phi}(p, q) = n^{-1}\sum_{i=1}^n \|X_i - \bar{X}\|_F^2$ *and* $\hat{\rho}(\tilde{p}, \tilde{q}) = \hat{\Phi}(\tilde{p}, \tilde{q})/\hat{\Phi}(p, q)$.

Note that $\Phi(\tilde{p}, \tilde{q}) = \sum_{i=1}^{\tilde{p}} \lambda_i = \sum_{j=1}^{\tilde{q}} \xi_j$ and $\hat{\Phi}(\tilde{p}, \tilde{q}) = \sum_{i=1}^{\tilde{p}} \hat{\lambda}_i = \sum_{j=1}^{\tilde{q}} \hat{\xi}_j$ from the stationary equations (5)–(6). Also note that $\Phi(\tilde{p}, \tilde{q}) \leqslant \Phi(p, q)$ always holds. Thus, $\rho(\tilde{p}, \tilde{q}) \leqslant 1$ and is used as a measure of adequacy for multilinear principal component analysis at dimensionality $(\tilde{p}, \tilde{q})$. For a given $\rho_0 \in (0, 1)$, consider the hypothesis test:

$$H_0 : \rho(\tilde{p}, \tilde{q}) \leqslant \rho_0, \quad H_1 : \rho(\tilde{p}, \tilde{q}) > \rho_0. \tag{7}$$

A rejection of H_0 indicates that $\rho(\tilde{p}, \tilde{q})$ reaches the required level of explained variance at a certain confidence level. The asymptotic distribution used to determine the approximate critical value for $\hat{\rho}(\tilde{p}, \tilde{q})$ is derived in § 3.

3. ASYMPTOTIC PROPERTIES FOR MULTILINEAR PRINCIPAL COMPONENT ANALYSIS

3·1. *Preliminary*

Without loss of generality, we assume $\mu = 0$ in the discussion of asymptotic theory. This implies $\Sigma = E\{\mathrm{vec}(X)\mathrm{vec}(X)^{\mathrm{T}}\}$, and the solution (A, B) of multilinear principal component analysis in Definition 1 equals the system eigenvector solution of $E(XP_BX^{\mathrm{T}})$ and $E(X^{\mathrm{T}}P_AX)$. The solution (A^*, B^*) of two-directional two-dimensional principal component analysis in § 2·3 is to solve the eigenvectors of $E(XX^{\mathrm{T}})$ and $E(X^{\mathrm{T}}X)$. Let S_n be the sample covariance matrix of $\{\mathrm{vec}(X_i)\}_{i=1}^n$, where the X_is are independent copies of X with finite second moments following model (2). By the central limit theorem, $n^{1/2}(S_n - \Sigma)$ converges weakly to a random matrix N, where $\mathrm{vec}(N)$ is an m^2-variate normal with zero-mean and covariance matrix $\Sigma_N = \mathrm{cov}\{\mathrm{vec}(X) \otimes \mathrm{vec}(X)\}$. If $\mathrm{vec}(X)$ is further assumed to be normally distributed, then S_n follows a Wishart distribution and $\Sigma_N = (I_{m^2} + K_{m,m})(\Sigma \otimes \Sigma)$, where $K_{\ell,k} = \sum_{i=1}^{\ell}\sum_{j=1}^{k} H_{ij} \otimes H_{ij}^{\mathrm{T}}$ is the commutation matrix, and H_{ij} is an $\ell \times k$ matrix with a one in the (i, j)th entry and zeros elsewhere (Magnus & Neudecker, 1979). Unless explicitly specified, the asymptotic properties derived in this section do not rely on the normality of $\mathrm{vec}(X)$.

3·2. *Asymptotic normality*

In the following theorem we derive the asymptotic distributions of tensor principal components and cumulative variances for $\tilde{p} \leqslant p_0$ and $\tilde{q} \leqslant q_0$ only, since for $\tilde{p} > p_0$ or $\tilde{q} > q_0$ the eigenvalues are multiple roots and the tensor principal components are not uniquely determined.

THEOREM 2. *Assume model* (2), *Condition* 1, *and let* $\tilde{p} \leqslant p_0$ *and* $\tilde{q} \leqslant q_0$. *Then, as* $n \to \infty$, (a) $n^{1/2}[\{\hat{\Phi}(\tilde{p}, \tilde{q}), \hat{\Phi}(p, q)\} - \{\Phi(\tilde{p}, \tilde{q}), \Phi(p, q)\}]^{\mathrm{T}}$ *converges weakly to* $\{D_{\Phi(\tilde{p},\tilde{q})}^{\mathrm{T}}, \mathrm{vec}(I_m)\}^{\mathrm{T}}\mathrm{vec}(N)$, *where* $D_{\Phi(\tilde{p},\tilde{q})} = \partial\Phi(\tilde{p}, \tilde{q})/\partial\mathrm{vec}(\Sigma)$ *is calculated in Lemma* 1; (b) $n^{1/2}[\{\mathrm{vec}^{\mathrm{T}}(\hat{A}), \mathrm{vec}^{\mathrm{T}}(\hat{B})\}^{\mathrm{T}} - \{\mathrm{vec}^{\mathrm{T}}(A), \mathrm{vec}^{\mathrm{T}}(B)\}^{\mathrm{T}}]$ *converges weakly to* $D_{H_{\tilde{p},\tilde{q}}}\mathrm{vec}(N)$ *with* $D_{H_{\tilde{p},\tilde{q}}} = [\{\partial a_1/\partial\mathrm{vec}(\Sigma)\}^{\mathrm{T}}, \ldots, \{\partial a_{\tilde{p}}/\partial\mathrm{vec}(\Sigma)\}^{\mathrm{T}}, \{\partial b_1/\partial\mathrm{vec}(\Sigma)\}^{\mathrm{T}}, \ldots, \{\partial b_{\tilde{q}}/\partial\mathrm{vec}(\Sigma)\}^{\mathrm{T}}]^{\mathrm{T}}$. *When* $(\tilde{p}, \tilde{q}) = (p_0, q_0)$, $D_{H_{p_0,q_0}}$ *has an explicit expression, which is given in Lemma* 1.

LEMMA 1. *Assume model* (2) *and Condition* 1. (a) *For* $\tilde{p} \leqslant p_0$ *and* $\tilde{q} \leqslant q_0$, *we have* $D_{\Phi(\tilde{p},\tilde{q})} = \mathrm{vec}^{\mathrm{T}}(P_{B\otimes A})$. (b) *When* $(\tilde{p},\tilde{q}) = (p_0, q_0)$, *for* $i = 1, \ldots, p_0$ *and* $j = 1, \ldots, q_0$,

$$\frac{\partial a_i}{\partial \mathrm{vec}(\Sigma)} = [a_i \otimes \mathrm{vec}(P_{B_0}) \otimes \{\lambda_i I_p - E(XP_{B_0}X^{\mathrm{T}})\}^{+}]^{\mathrm{T}}(K_{p,q} \otimes I_m),$$
$$\frac{\partial b_j}{\partial \mathrm{vec}(\Sigma)} = [b_j \otimes \mathrm{vec}(P_{A_0}) \otimes \{\xi_j I_q - E(X^{\mathrm{T}}P_{A_0}X)\}^{+}]^{\mathrm{T}}(I_m \otimes K_{p,q}).$$

We are now in a position to establish the asymptotic normality of the explained variance $\hat{\rho}(\tilde{p},\tilde{q})$ and of the projection matrix onto the multilinear principal component subspace $P_{\hat{B}\otimes\hat{A}}$.

COROLLARY 1. *Assume model* (2), *Condition* 1, *and let* $\tilde{p} \leqslant p_0$ *and* $\tilde{q} \leqslant q_0$. *Then, as* $n \to \infty$, $n^{1/2}\{\hat{\rho}(\tilde{p},\tilde{q}) - \rho(\tilde{p},\tilde{q})\}$ *converges weakly to* $N(0, \sigma^2_{\rho(\tilde{p},\tilde{q})})$, *where*

$$\sigma^2_{\rho(\tilde{p},\tilde{q})} = \left\{\frac{D_{\Phi(\tilde{p},\tilde{q})}}{\Phi(p,q)} - \frac{\Phi(\tilde{p},\tilde{q})\mathrm{vec}^{\mathrm{T}}(I_m)}{\Phi^2(p,q)}\right\}\Sigma_N\left\{\frac{D_{\Phi(\tilde{p},\tilde{q})}}{\Phi(p,q)} - \frac{\Phi(\tilde{p},\tilde{q})\mathrm{vec}^{\mathrm{T}}(I_m)}{\Phi^2(p,q)}\right\}^{\mathrm{T}}.$$

Moreover, $n^{1/2}\{\mathrm{vec}(P_{\hat{B}\otimes\hat{A}}) - \mathrm{vec}(P_{B\otimes A})\}$ *converges weakly to* $D_{P_{B\otimes A}}\mathrm{vec}(N)$, *where* $D_{P_{B\otimes A}} = \partial\mathrm{vec}(P_{B\otimes A})/\partial\mathrm{vec}(\Sigma)$. *When* $(\tilde{p},\tilde{q}) = (p_0, q_0)$, $D_{P_{B_0\otimes A_0}}$ *has the explicit expression*

$$(I_{m^2} + K_{m,m}) \times \left\{\sum_{i=1}^{p_0}(K_{q,p}\otimes I_m)[P_{a_i}\otimes\{\mathrm{vec}(P_{B_0})\mathrm{vec}^{\mathrm{T}}(P_{B_0})\}\otimes\{\lambda_i I_p - E(XP_{B_0}X^{\mathrm{T}})\}^{+}](K_{p,q}\otimes I_m) + \sum_{j=1}^{q_0}(I_m\otimes K_{q,p})[P_{b_j}\otimes\{\mathrm{vec}(P_{A_0})\mathrm{vec}^{\mathrm{T}}(P_{A_0})\}\otimes\{\xi_j I_q - E(X^{\mathrm{T}}P_{A_0}X)\}^{+}](I_m\otimes K_{p,q})\right\}.$$

Corollary 1 is the cornerstone of our asymptotic test for hypothesis (7). Before practical implementation of the test, however, we need a consistent estimator of $\sigma^2_{\rho(\tilde{p},\tilde{q})}$. The asymptotic covariance Σ_N can be empirically estimated by

$$\hat{\Sigma}^{(1)}_N = \frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}\{\mathrm{vec}(X_i - \bar{X})\mathrm{vec}^{\mathrm{T}}(X_i - \bar{X}) - S_n\}\mathrm{vec}^{\mathrm{T}}\{\mathrm{vec}(X_i - \bar{X})\mathrm{vec}^{\mathrm{T}}(X_i - \bar{X}) - S_n\}.$$

Moreover, if $\mathrm{vec}(X)$ is normally distributed, $\hat{\Sigma}^{(2)}_N = (I_{m^2} + K_{m,m})(S_n \otimes S_n)$ can also be used to estimate Σ_N. Thus, the asymptotic variance $\sigma^2_{\rho(\tilde{p},\tilde{q})}$ is estimated by

$$\hat{\sigma}^2_{\rho(\tilde{p},\tilde{q})} = \left\{\frac{\hat{D}_{\Phi(\tilde{p},\tilde{q})}}{\hat{\Phi}(p,q)} - \frac{\hat{\Phi}(\tilde{p},\tilde{q})\mathrm{vec}^{\mathrm{T}}(I_m)}{\hat{\Phi}^2(p,q)}\right\}\hat{\Sigma}^{(i)}_N\left\{\frac{\hat{D}_{\Phi(\tilde{p},\tilde{q})}}{\hat{\Phi}(p,q)} - \frac{\hat{\Phi}(\tilde{p},\tilde{q})\mathrm{vec}^{\mathrm{T}}(I_m)}{\hat{\Phi}^2(p,q)}\right\}^{\mathrm{T}}$$

for $i = 1$ without the normality assumption or $i = 2$ with it, where $\hat{D}_{\Phi(\tilde{p},\tilde{q})} = \mathrm{vec}^{\mathrm{T}}(P_{\hat{B}\otimes\hat{A}})$. Direct calculations based on $\hat{\Sigma}^{(1)}_N$ lead to

$$\hat{\sigma}^2_{\rho(\tilde{p},\tilde{q})} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\|\hat{U}_i\|^2_F - \frac{1}{n}\sum_{j=1}^{n}\|\hat{U}_j\|^2_F}{\hat{\Phi}(p,q)} - \frac{\hat{\Phi}(\tilde{p},\tilde{q})}{\hat{\Phi}^2(p,q)}\left(\|X_i - \bar{X}\|^2_F - \frac{1}{n}\sum_{j=1}^{n}\|X_j - \bar{X}\|^2_F\right)\right\}^2$$

and based on $\hat{\Sigma}_N^{(2)}$, we have

$$\hat{\sigma}^2_{\rho(\tilde{p},\tilde{q})} = \frac{2\,\mathrm{tr}[\{(B\otimes A)^{\mathrm{T}} S_n (B\otimes A)\}^2 - 2\frac{\hat{\Phi}(\tilde{p},\tilde{q})}{\hat{\Phi}(p,q)}(B\otimes A)^{\mathrm{T}} S_n^2 (B\otimes A)]}{\hat{\Phi}^2(p,q)} + \frac{2\hat{\Phi}^2(\tilde{p},\tilde{q})\,\mathrm{tr}(S_n^2)}{\hat{\Phi}^4(p,q)}.$$

The consistency of $\hat{\sigma}^2_{\rho(\tilde{p},\tilde{q})}$ follows as a direct consequence by standard arguments, which also enables us to construct an approximate level α test to determine the adequacy at dimensionality $(\tilde{p},\tilde{q})$ as stated in the following theorem. The proof is straightforward.

THEOREM 3. *Assume model* (2), *Condition* 1, *and let* $\tilde{p}\leqslant p_0$ *and* $\tilde{q}\leqslant q_0$. *For the hypothesis* (7), *an approximate level* α *test is to reject* H_0 *if* $\hat{\rho}(\tilde{p},\tilde{q}) > \ell_\alpha(\rho_0;\tilde{p},\tilde{q})$, *where* $\ell_\alpha(\rho_0;\tilde{p},\tilde{q}) = \rho_0 + n^{-1/2}\hat{\sigma}_{\rho(\tilde{p},\tilde{q})}z_\alpha$ *and* z_α *is the upper* α *quantile of the standard normal distribution.*

3·3. *Asymptotic efficiency*

Both $(\hat{B}\otimes\hat{A})$ of multilinear principal component analysis and $(\hat{B}^*\otimes\hat{A}^*)$ of two-directional two-dimensional principal component analysis target $\mathrm{span}(B_0\otimes A_0)$ when $(\tilde{p},\tilde{q}) = (p_0,q_0)$. The following theorem shows that multilinear principal component analysis is asymptotically more efficient than two-directional two-dimensional principal component analysis in estimating the Kronecker envelope $\mathrm{span}(B_0\otimes A_0)$.

THEOREM 4. *Assume model* (2), *Condition* 1 *and the normality of* $\mathrm{vec}(X)$. *When* $(\tilde{p},\tilde{q}) = (p_0,q_0)$, *we have* $\mathrm{acov}\{\mathrm{vec}(P_{\hat{B}^*\otimes\hat{A}^*})\} - \mathrm{acov}\{\mathrm{vec}(P_{\hat{B}\otimes\hat{A}})\} \geqslant 0$, *where* acov *denotes the asymptotic covariance. The equality holds if and only if* $(p_0,q_0) = (p,q)$.

The only case for which we will gain nothing from multilinear principal component analysis over the two-directional two-dimensional principal component analysis is when $(p_0,q_0) = (p,q)$, when there is no room for dimension reduction.

4. OLIVETTI FACES DATASET

We test and compare the performance of multilinear principal component analysis and conventional principal component analysis on the Olivetti faces dataset, which is available at http://www.cs.nyu.edu/~roweis/data.html. This dataset consists of 400 greyscale, i.e., 8-bit, face images of 64×64 pixels. Different facial expressions and/or views exist for each individual in this dataset. A simulation experiment is designed as follows: 400 face images are randomly partitioned into a training set of size 100 and a test set of size 300. This 100/300 partition, where the training set is smaller than the test set, is chosen to reflect a scenario in which a small portion of data is used to train a basis set which will in turn be used to represent the complete data contained in a large data archive. Each principal component analysis scheme is applied on the 100 training images to produce an image basis, which is then used to reconstruct the remaining 300 test images. The average of the 100 training images, which is called the mean face, has been subtracted from each of the 400 images, and is added back to the reconstructions to arrive at the final reconstructed images. To determine the Kronecker dimension (p_0,q_0) of multilinear principal component analysis, we set $\alpha = 0{\cdot}05$ and $\rho_0 = 0{\cdot}95$, and then apply Theorem 3 to implement a sequence of hypotheses (7). Here, due to its flexibility, the empirical estimator of $\hat{\sigma}^2_{\rho(\tilde{p},\tilde{q})}$ based on $\hat{\Sigma}_N^{(1)}$ is adopted. The test results are summarized in Table 1. We then determine $(p_0,q_0) = (24,24)$ since it is the first pair that satisfies $\hat{\rho}(\tilde{p},\tilde{q}) > \ell_\alpha(\rho_0;\tilde{p},\tilde{q})$ and thus leads to

Table 1. *The first pair* $(\tilde{p}, \tilde{q})$ *that meets the criterion, where the sample explained proportion of total variance* $\hat{\rho}(\tilde{p}, \tilde{q})$ *exceeds its* α *critical value* $\ell_\alpha(\rho_0; \tilde{p}, \tilde{q})$*, is selected for the Olivetti faces dataset. Here, we use* $\alpha = 0{\cdot}05$ *and* $\rho_0 = 0{\cdot}95$

$(\tilde{p}, \tilde{q})$	(20, 20)	(21, 21)	(22, 22)	(23, 23)	(24, 24)*
$\hat{\rho}(\tilde{p}, \tilde{q})$	0·9350	0·9410	0·9460	0·9507	0·9551
$\ell_\alpha(\rho_0; \tilde{p}, \tilde{q})$	0·9532	0·9529	0·9527	0·9525	0·9523

Fig. 1. Twenty randomly drawn test faces from the Olivetti faces dataset. Rows 1–2 are the original faces. Rows 3–4 are reconstructed faces using 24 × 24 trained tensor principal components. Rows 5–6 are reconstructed faces using 576 trained conventional principal components.

rejection of the null hypothesis. For comparison, the dimensionality of the conventional principal component analysis is therefore set to be $576 = 24 \times 24$.

In Fig. 1, 20 test images were randomly chosen from the test set to show the visual performance of the image reconstructions. Obviously, multilinear principal component analysis produced superior reconstructed images. In multilinear principal component analysis, 24 row eigenvectors and 24 column eigenvectors, both of size 64, were used to generate 576 basis images, of which the 100 leading ones are shown in Fig. 2(a). In conventional principal component analysis, 576 eigenvectors each of size 4096 were used, of which the 100 leading ones are shown in Fig. 2(b). Because 100 training images were used with average subtraction, there are at most 99 meaningful eigenvectors in the conventional principal component analysis. The rest are random orthogonal eigenvectors with zero eigenvalues from the remaining subspace. In Fig. 2(b), looking from left to right and top to bottom, we can see the images begin as clear facial shapes and gradually change to vague ones with what looks like an image of random noise appearing at the 100th. On the other hand, multilinear principal component analysis tends to distribute the image

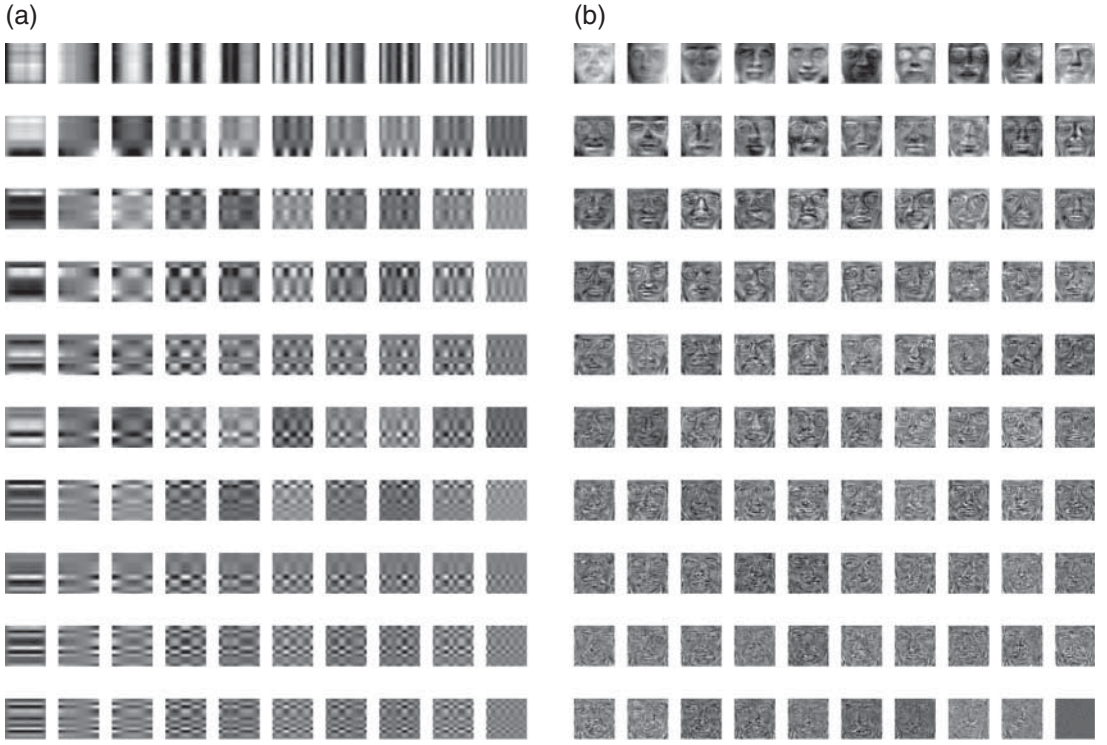

Fig. 2. The leading 100 tensor principal component images (a) and conventional principal component images (b).

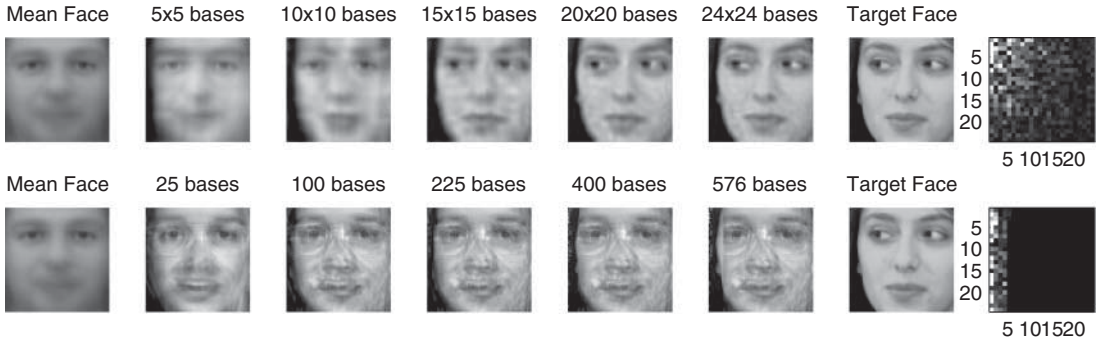

Fig. 3. The reconstructed images of a test face by adding more basis elements using multilinear principal component analysis (top row) and conventional principal component analysis (bottom row). The mean and target faces are put in this figure as references. The right-most panels are absolute projection coefficients, where the lighter color refers to the larger value.

characteristics to more basis elements, which may allow improved localized modifications of the images.

In Fig. 3, one particular image among the 300 test images was picked out to demonstrate the performance of these two methods, with various numbers of basis elements. The mean face and the target image were also included for reference. The right-most column shows the absolute values of the projection scores of the leading 576 basis elements. The conventional principal component analysis concentrates on no more than 99 basis elements while the multilinear principal component analysis expands to many more basis elements. For multilinear principal component

analysis, the image turns its view when 10×10 basis elements are used; the pupil turns to the left when 15×15 basis elements are used; the nostrils and folds of the eyelids show up when 20×20 basis elements are used; the facial curves become clear when 24×24 basis elements are used. While we can observe the reconstruction progressively improve by adding more basis elements for multilinear principal component analysis, we do not see much difference after 100 basis elements for conventional principal component analysis.

We repeated the random training-test partition 500 times with each type of analysis and then compared their mean test errors, defined as the average over 500 replicate runs of the mean Frobenius norm between the original and the reconstructed images over 300 test images. This gave a mean test error for conventional principal component analysis of 2870 ± 43, which was more than six times the error of 452 ± 4 obtained for multilinear principal component analysis. The standard error of multilinear principal component analysis was also smaller. Multilinear principal component analysis has outperformed the conventional approach, which suffered seriously from the large m and the small n problem, wherein there can be at most $n - 1$ meaningful eigenvectors. Because of this, data noise will inevitably still be carried by the chosen principal components. Information will tend to be overconcentrated in a few components, which may not be good for generalization to test images. Multilinear principal component analysis distributes the information to more components, which may allow local modification in the process of image reconstruction. It is our belief that the key factor responsible for the good performance of multilinear principal component analysis is the adoption of the Kronecker envelope and, hence, a parsimonious usage of parameters.

ACKNOWLEDGEMENT

The authors thank the editor and two referees for comments that substantially improved the paper. The research was supported by the National Science Council of Taiwan.

APPENDIX

Technical details

The following properties of commutation matrix $K_{\ell,k}$ are needed in proving propositions and theorems. Given matrices $M_i \in \Re^{a_i \times b_i}$, $i = 1, 2$, then $(M_2 \otimes M_1) = K_{a_2,a_1}(M_1 \otimes M_2)K_{b_1,b_2}$ and $\mathrm{vec}(M_1^{\mathrm{T}}) = K_{a_1,b_1}\mathrm{vec}(M_1)$. Moreover, $K_{a_1,b_1} = K_{b_1,a_1}^{\mathrm{T}}$, $K_{a_1,b_1}K_{b_1,a_1} = I_{a_1b_1}$, and $K_{a_1,b_1} = I_{a_1}$ if $b_1 = 1$.

Proof of Proposition 1. The objective function (4) is continuous and the feasible region $\mathcal{O}_{q,\tilde{q}} \otimes \mathcal{O}_{p,\tilde{p}}$ is compact with respect to the topology induced by the Frobenius norm. Thus, a solution exists. □

Proof of Proposition 2. Let $[B, B_\perp]$ be a $q \times q$ orthogonal matrix. Since $B_0 \in \mathrm{span}([B, B_\perp])$, there exist $\eta_1 \in \Re^{\tilde{q} \times q_0}$ and $\eta_2 \in \Re^{(q-\tilde{q}) \times q_0}$ such that $B_0 = B\eta_1 + B_\perp\eta_2$. As $B_0^{\mathrm{T}}B_0 = I_{q_0}$, we have $B_0^{\mathrm{T}}BB^{\mathrm{T}}B_0 = \eta_1^{\mathrm{T}}\eta_1 = I_{q_0} - \eta_2^{\mathrm{T}}\eta_2$. Hence, direct calculation gives

$$\begin{aligned} E\{\|A^{\mathrm{T}}(X-\mu)B\|_F^2\} &= E\{\mathrm{tr}(A^{\mathrm{T}}A_0UU^{\mathrm{T}}A_0^{\mathrm{T}}A)\} - E\{\mathrm{tr}(A^{\mathrm{T}}A_0U\eta_2^{\mathrm{T}}\eta_2U^{\mathrm{T}}A_0^{\mathrm{T}}A)\} + \tilde{p}\tilde{q}\sigma^2 \\ &\leqslant E\{\mathrm{tr}(A^{\mathrm{T}}A_0UU^{\mathrm{T}}A_0^{\mathrm{T}}A)\} + \tilde{p}\tilde{q}\sigma^2, \end{aligned} \tag{A1}$$

where the equality in (A1) holds if and only if $\eta_2 = 0$, and hence, if and only if $\eta_1^{\mathrm{T}}\eta_1 = I_{q_0}$. Thus, if $\tilde{q} \geqslant q_0$, such an η_1 exists to ensure the equality in (A1). This implies that $B_0 = B\eta_1$ and, hence, $B_0 \in \mathrm{span}(B)$. Similarly, $A_0 \in \mathrm{span}(A)$ which establishes (a). To show (b), we have from (a) that $B_0 \in \mathrm{span}(B)$ when $\tilde{q} \geqslant q_0$. Thus, $\max_{A \in \mathcal{O}_{p,\tilde{p}}, B \in \mathcal{O}_{q,\tilde{q}}} E\{\|A^{\mathrm{T}}(X-\mu)B\|_F^2\} = \max_{A \in \mathcal{O}_{p,\tilde{p}}} \mathrm{tr}\{A^{\mathrm{T}}E(A_0UU^{\mathrm{T}}A_0^{\mathrm{T}})A\} + \tilde{p}\tilde{q}\sigma^2$, and finding A is an eigenvalue problem of $E(A_0UU^{\mathrm{T}}A_0^{\mathrm{T}})$. By Condition 1 and diagonalizing $E(UU^{\mathrm{T}}) = \Gamma_U\Lambda_U\Gamma_U^{\mathrm{T}}$, $E(A_0UU^{\mathrm{T}}A_0^{\mathrm{T}})$ has p_0 distinct nonzero eigenvalues $\mathrm{diag}(\Lambda_U)$ with eigenvectors $A_0\Gamma_U$. When $\tilde{p} < p_0$, the

maximizer A must consist of the first $\tilde{p}$ columns of $A_0\Gamma_U$ and, hence, $\mathrm{span}(A) \subset \mathrm{span}(A_0)$. (c) is similar to (b). As to (d), to maximize $E\{\|A^{\mathrm{T}}(X-\mu)B\|_F^2\} = E\{\mathrm{tr}(A^{\mathrm{T}}A_0UB_0^{\mathrm{T}}BB^{\mathrm{T}}B_0U^{\mathrm{T}}A_0^{\mathrm{T}}A)\} + \tilde{p}\tilde{q}\sigma^2$ over $A \in \mathcal{O}_{p,\tilde{p}}$, $B \in \mathcal{O}_{q,\tilde{q}}$ with $\tilde{p} < p_0$ and $\tilde{q} < q_0$, the rank of $A^{\mathrm{T}}A_0$ and $B^{\mathrm{T}}B_0$ must be $\tilde{p}$ and $\tilde{q}$, respectively, in order to attain the maximal value. This can happen only if $\mathrm{span}(A) \subset \mathrm{span}(A_0)$ and $\mathrm{span}(B) \subset \mathrm{span}(B_0)$. □

Proof of Proposition 3. We will prove only (a), as (b) is similar. If $\tilde{q} \geqslant q_0$, from Proposition 2 we have $\mathrm{span}(B_0) \subseteq \mathrm{span}(B)$. This implies that $E\{(X-\mu)P_B(X-\mu)^{\mathrm{T}}\} = E(A_0UU^{\mathrm{T}}A_0^{\mathrm{T}}) + \tilde{q}\sigma^2 I_p$ and $E\{(X-\mu)(X-\mu)^{\mathrm{T}}\} = E(A_0UU^{\mathrm{T}}A_0^{\mathrm{T}}) + q\sigma^2 I_p$, so these two matrices share the same leading $\min\{p_0, \tilde{p}\}$ eigenvectors of $E(A_0UU^{\mathrm{T}}A_0^{\mathrm{T}})$. Moreover, $\lambda_i = d_i + \tilde{q}\sigma^2$ and $\lambda_i^* = d_i + q\sigma^2$, where d_i is the ith eigenvalue of $E(A_0UU^{\mathrm{T}}A_0^{\mathrm{T}})$. Thus, $\lambda_i^* - \lambda_i = (q-\tilde{q})\sigma^2$ for $i = 1, \ldots, \tilde{p}$. □

Proof of Theorem 2. Let $H_{\tilde{p},\tilde{q}}(S_n) = \{\mathrm{vec}^{\mathrm{T}}(\hat{A}), \mathrm{vec}^{\mathrm{T}}(\hat{B})\}^{\mathrm{T}}$ be the function that maps S_n to its tensor principal components under $(\tilde{p}, \tilde{q})$, which gives $H_{\tilde{p},\tilde{q}}(\Sigma) = \{\mathrm{vec}^{\mathrm{T}}(A), \mathrm{vec}^{\mathrm{T}}(B)\}^{\mathrm{T}}$ and $D_{H_{\tilde{p},\tilde{q}}} = \partial H_{\tilde{p},\tilde{q}}(\Sigma)/\partial\mathrm{vec}(\Sigma)$. Note that $\hat{\Phi}(p,q) = \mathrm{vec}^{\mathrm{T}}(I_m)\mathrm{vec}(S_n)$ and $\Phi(p,q) = \mathrm{vec}^{\mathrm{T}}(I_m)\mathrm{vec}(\Sigma)$. From the weak convergence of $n^{1/2}(S_n - \Sigma)$ to N and the delta method, we have, for $(\tilde{p}, \tilde{q}) \leqslant (p_0, q_0)$, the weak convergence of $n^{1/2}\{\hat{\Phi}(p,q) - \Phi(p,q)\}$, $n^{1/2}\{\hat{\Phi}(\tilde{p},\tilde{q}) - \Phi(\tilde{p},\tilde{q})\}$, and $n^{1/2}\{H_{\tilde{p},\tilde{q}}(S_n) - H_{\tilde{p},\tilde{q}}(\Sigma)\}$ to $\mathrm{vec}(I_m)^{\mathrm{T}}\mathrm{vec}(N)$, $D_{\Phi(\tilde{p},\tilde{q})}\mathrm{vec}(N)$, and $D_{H_{\tilde{p},\tilde{q}}}\mathrm{vec}(N)$, respectively. □

Proof of Lemma 1. For any given pair $(\tilde{p}, \tilde{q})$, A and B will satisfy the system of stationary equations (5)–(6). Also note that $\Phi(\tilde{p},\tilde{q}) = \sum_{i=1}^{\tilde{p}} \lambda_i(\Sigma, \tilde{p}, \tilde{q}) = \sum_{j=1}^{\tilde{q}} \xi_j(\Sigma, \tilde{p}, \tilde{q})$ from (5)–(6). We will apply the perturbation method (Sibson, 1979, Lemma 2.1; Fine, 1987) to derive the expressions for the derivatives $D_{\Phi(\tilde{p},\tilde{q})}$, $\partial a_i/\partial\mathrm{vec}(\Sigma)$ and $\partial b_j/\partial\mathrm{vec}(\Sigma)$. Suppose that Σ is perturbed to $\Sigma_\epsilon = \Sigma + \epsilon\dot{\Sigma}$. Let the corresponding system of stationary equations with Σ_ϵ be

$$\left\{\sum_{j=1}^{\tilde{q}}(b_{j,\epsilon}\otimes I_p)^{\mathrm{T}}\Sigma_\epsilon(b_{j,\epsilon}\otimes I_p)\right\}a_{i,\epsilon} = \lambda_{i,\epsilon}\,a_{i,\epsilon}, \quad (i=1,\ldots,\tilde{p}), \tag{A2}$$

$$\left\{\sum_{i=1}^{\tilde{p}}(I_q\otimes a_{i,\epsilon})^{\mathrm{T}}\Sigma_\epsilon(I_q\otimes a_{i,\epsilon})\right\}b_{j,\epsilon} = \xi_{j,\epsilon}\,b_{j,\epsilon}, \quad (j=1,\ldots,\tilde{q}).$$

Denote the first-order expansions by $\lambda_{i,\epsilon} = \lambda_i + \epsilon\dot{\lambda}_i + o(\epsilon)$, $a_{i,\epsilon} = a_i + \epsilon\dot{a}_i + o(\epsilon)$, $\xi_{j,\epsilon} = \xi_j + \epsilon\dot{\xi}_j + o(\epsilon)$, and $b_{j,\epsilon} = b_j + \epsilon\dot{b}_j + o(\epsilon)$. Following the same arguments as in Lemma 2.1 of Sibson (1979) and by equating the terms involving ϵ in (A2) we have

$$\dot{\lambda}_i = a_i^{\mathrm{T}}\dot{\Sigma}_B a_i, \quad \dot{a}_i = \left\{\lambda_i I_p - \sum_{j=1}^{\tilde{q}}(b_j\otimes I_p)^{\mathrm{T}}\Sigma(b_j\otimes I_p)\right\}^{+}\dot{\Sigma}_B a_i, \quad (i=1,\ldots,\tilde{p}), \tag{A3}$$

where $\dot{\Sigma}_B = E\{X(\dot{B}B^{\mathrm{T}} + B\dot{B}^{\mathrm{T}})X^{\mathrm{T}}\} + \sum_{j=1}^{\tilde{q}}(b_j\otimes I_p)^{\mathrm{T}}\dot{\Sigma}(b_j\otimes I_p)$ with $\dot{B} = [\dot{b}_1, \ldots, \dot{b}_{\tilde{q}}]$. Since $b_{i,\epsilon}^{\mathrm{T}}b_{i,\epsilon} = 1$ and $b_{i,\epsilon}^{\mathrm{T}}b_{j,\epsilon} = 0$ for $i \neq j$, $\dot{B}$ must satisfy $\dot{B}^{\mathrm{T}}B + B^{\mathrm{T}}\dot{B} = 0$.

We first prove (a) for $\tilde{p} \leqslant p_0$ and $\tilde{q} \leqslant q_0$. For the first term of $\dot{\Sigma}_B$, we have

$$\sum_{i=1}^{\tilde{p}} a_i^{\mathrm{T}}E\{X(\dot{B}B^{\mathrm{T}} + B\dot{B}^{\mathrm{T}})X^{\mathrm{T}}\}a_i = \sum_{j=1}^{\tilde{q}}\{\dot{b}_j^{\mathrm{T}}E(X^{\mathrm{T}}P_AX)b_j + b_j^{\mathrm{T}}E(X^{\mathrm{T}}P_AX)\dot{b}_j\}$$

which vanishes because b_j is an eigenvector of $E(X^{\mathrm{T}}P_AX)$ and $b_j^{\mathrm{T}}\dot{b}_j = \dot{b}_j^{\mathrm{T}}b_j = 0$. Thus, $\sum_{i=1}^{\tilde{p}}\dot{\lambda}_i = \sum_{i=1}^{\tilde{p}} a_i^{\mathrm{T}}\dot{\Sigma}_B a_i = \sum_{i=1}^{\tilde{p}} a_i^{\mathrm{T}}\{\sum_{j=1}^{\tilde{q}}(b_j\otimes I_p)^{\mathrm{T}}\dot{\Sigma}(b_j\otimes I_p)\}a_i$. Since $\Phi(\tilde{p},\tilde{q}) = \sum_{i=1}^{\tilde{p}}\lambda_i$, we deduce that $D_{\Phi(\tilde{p},\tilde{q})} = \sum_{i=1}^{\tilde{p}}\sum_{j=1}^{\tilde{q}}(b_j\otimes a_i\otimes b_j\otimes a_i)^{\mathrm{T}} = \mathrm{vec}^{\mathrm{T}}(P_{B\otimes A})$, which proves (a).

To show (b), assume $(\tilde{p},\tilde{q}) = (p_0, q_0)$. We will show the first term of $\dot{\Sigma}_B$ is zero and conclude $\dot{\Sigma}_B = \sum_{j=1}^{q_0}(b_j\otimes I_p)^{\mathrm{T}}\dot{\Sigma}(b_j\otimes I_p)$. This fact together with the expression of $\dot{a}_i$ in (A3) gives $\partial a_i/\partial\mathrm{vec}(\Sigma) = [a_i\otimes\mathrm{vec}(P_{B_0})\otimes\{\lambda_i I_p - E(XP_{B_0}X^{\mathrm{T}})\}^{+}]^{\mathrm{T}}(K_{p,q}\otimes I_m)$, where the equality follows from

Proposition 2 that $\text{span}(B) = \text{span}(B_0)$ when $\tilde{q} = q_0$. To show the first term of $\dot{\Sigma}_B$ vanishes, first note that Proposition 2 ensures the existence of a nonsingular matrix η such that $B_0 = B\eta$. From $X = A_0 U B_0^{\mathrm{T}} + \varepsilon$ and the independence of U and ε, we calculate the first term of $\dot{\Sigma}_B$ to be $E\{X(\dot{B}B^{\mathrm{T}} + B\dot{B}^{\mathrm{T}})X^{\mathrm{T}}\} = E\{A_0 U \eta^{\mathrm{T}}(B^{\mathrm{T}}\dot{B} + \dot{B}^{\mathrm{T}}B)\eta U^{\mathrm{T}} A_0^{\mathrm{T}}\} + \{\sigma^2 \text{tr}(B^{\mathrm{T}}\dot{B} + \dot{B}^{\mathrm{T}}B)\}I_p$. The proof is now completed by using $B^{\mathrm{T}}\dot{B} + \dot{B}^{\mathrm{T}}B = 0$. The case of $\partial b_j / \partial \text{vec}(\Sigma)$ is similar. □

Proof of Corollary 1. Define $F_1(x, y) = xy^{-1}$ and $F_2(A, B) = \text{vec}(P_{B\otimes A})$. Their differentials are calculated to be $D_{F_2(A,B)} = (I_{m^2} + K_{m,m})(I_q \otimes K_{p,q} \otimes I_p)[\text{vec}(P_B) \otimes A \otimes I_p, B \otimes I_q \otimes \text{vec}(P_A)]$ and $D_{F_1(x,y)} = (y^{-1}, -xy^{-2})$. Note that $\rho(\tilde{p}, \tilde{q}) = F_1\{\Phi(\tilde{p}, \tilde{q}), \Phi(p, q)\}$. The proof is now completed by Theorem 2, Lemma 1 and the delta method with $D_{P_{B\otimes A}} = D_{F_2(A,B)} D_{H_{\tilde{p},\tilde{q}}}$. When $(\tilde{p}, \tilde{q}) = (p_0, q_0)$, the expression of $D_{P_{B\otimes A}}$ is obtained by a direct calculation together with Lemma 1(b) and Proposition 2. □

Proof of Theorem 4. Since $(\tilde{p}, \tilde{q}) = (p_0, q_0)$, we have $A = A^*$ and $B = B^*$ from Proposition 3, and $\text{span}(A) = \text{span}(A_0)$ and $\text{span}(B) = \text{span}(B_0)$ from Proposition 2. Let $\{a_i : i = p_0 + 1, \ldots, p\}$ and $\{b_j : j = q_0 + 1, \ldots, q\}$ be orthogonal bases of $\text{span}(Q_{A_0})$ and $\text{span}(Q_{B_0})$, respectively. Let $W_{B,q'} = \sum_{j=1}^{q'} (b_j \otimes I_p \otimes b_j \otimes I_p)^{\mathrm{T}}, q' = 1, \ldots, q$ and $W_{A,p'} = \sum_{i=1}^{p'} (I_q \otimes a_i \otimes I_q \otimes a_i)^{\mathrm{T}}, p' = 1, \ldots, p$. Also define $M_A = [a_1 \otimes M_{A1}, \ldots, a_{p_0} \otimes M_{Ap_0}]^{\mathrm{T}}$ and $M_B = [b_1 \otimes M_{B1}, \ldots, b_{q_0} \otimes M_{Bq_0}]^{\mathrm{T}}$, where $M_{Ai} = \{\lambda_i I_p - E(XP_{B_0}X^{\mathrm{T}})\}^+$ and $M_{Bj} = \{\xi_j I_q - E(X^{\mathrm{T}}P_{A_0}X)\}^+$. From Theorem 2(b), $n^{1/2}\{\text{vec}(\hat{A}, \hat{B}) - \text{vec}(A, B)\}$ converges weakly to

$$\begin{bmatrix} M_A & 0 \\ 0 & M_B \end{bmatrix} \begin{bmatrix} W_{B,q_0} \\ W_{A,p_0} \end{bmatrix} \text{vec}(N) = M_0 W_0 \text{vec}(N). \tag{A4}$$

By Lemma A1 below and the delta method, $n^{1/2}\{\text{vec}(\hat{A}^*, \hat{B}^*) - \text{vec}(A^*, B^*)\}$ converges weakly to

$$M_0(W_0 + W_{0+})\text{vec}(N), \tag{A5}$$

where $W_{0+} = [W_{B,q_0+}^{\mathrm{T}}, W_{A,p_0+}^{\mathrm{T}}]^{\mathrm{T}}$, $W_{A,p_0+} = W_{A,p} - W_{A,p_0}$, and $W_{B,q_0+} = W_{B,q} - W_{B,q_0}$.

To complete the proof, by the delta method and the fact that $A = A^*$ and $B = B^*$, it suffices to show that

$$\text{acov}\{\text{vec}(\hat{A}^*, \hat{B}^*)\} - \text{acov}\{\text{vec}(\hat{A}, \hat{B})\} \geqslant 0. \tag{A6}$$

From (A4)–(A5) we must show $M_0(W_0 \Sigma_N W_{0+}^{\mathrm{T}} + W_{0+}\Sigma_N W_0^{\mathrm{T}} + W_{0+}\Sigma_N W_{0+}^{\mathrm{T}})M_0^{\mathrm{T}} \geqslant 0$, where $\Sigma_N = (I_{m^2} + K_{m,m})(\Sigma \otimes \Sigma)$ under normality of $\text{vec}(X)$. We will prove $M_0 W_0 \Sigma_N W_{0+}^{\mathrm{T}} M_0^{\mathrm{T}} = 0$. This together with $M_0 W_{0+} \Sigma_N W_{0+}^{\mathrm{T}} M_0^{\mathrm{T}} \geqslant 0$ establishes the desired result. Observe that

$$M_0 W_0 \Sigma_N W_{0+}^{\mathrm{T}} M_0^{\mathrm{T}} = \begin{bmatrix} M_A W_{B,q_0} \Sigma_N W_{B,q_0+}^{\mathrm{T}} M_A^{\mathrm{T}} & H_A W_{B,q_0} \Sigma_N W_{A,p_0+}^{\mathrm{T}} M_B^{\mathrm{T}} \\ H_B W_{A,p_0} \Sigma_N W_{B,q_0+}^{\mathrm{T}} M_A^{\mathrm{T}} & H_B W_{A,p_0} \Sigma_N W_{A,p_0+}^{\mathrm{T}} M_B^{\mathrm{T}} \end{bmatrix}.$$

By model (2), $\Sigma = (B_0 \otimes A_0)(T + \sigma^2 I_{m_0})(B_0 \otimes A_0)^{\mathrm{T}} + \sigma^2 Q_{B_0 \otimes A_0}$. Thus, $W_{B,q_0}\Sigma_N W_{B,q_0+}^{\mathrm{T}} = 0$ and $W_{A,p_0}\Sigma_N W_{A,p_0+}^{\mathrm{T}} = 0$ and, hence, the diagonal elements of the above matrix vanish. For the off-diagonal elements, the same reasoning can be used to deduce that $(M_A W_{B,q_0})\Sigma_N W_{A,p_0+}^{\mathrm{T}} = 0$ and $(M_B W_{A,p_0})\Sigma_N W_{B,q_0+}^{\mathrm{T}} = 0$. Therefore, (A6) can be established. A direct calculation further gives

$$M_0 W_{0+} \Sigma_N W_{0+}^{\mathrm{T}} M_0^{\mathrm{T}} = \sigma^4 \begin{bmatrix} (q - q_0) M_A (I_{p^2} + K_{p,p}) M_A^{\mathrm{T}} & 0 \\ 0 & (p - p_0) M_B (I_{q^2} + K_{q,q}) M_B^{\mathrm{T}} \end{bmatrix},$$

which equals a zero matrix if and only if $(p_0, q_0) = (p, q)$. □

LEMMA A1. *Assume model* (2) *and* $(\tilde{p}, \tilde{q}) = (p_0, q_0)$. *Then,* $\partial \text{vec}(A^*)/\partial \text{vec}(\Sigma) = M_A W_{B,q}$ *and* $\partial \text{vec}(B^*)/\partial \text{vec}(\Sigma) = M_B W_{A,p}$.

Proof. We show only the case of A^*, as the case of B^* is similar. Under model (2), A^* are leading p_0 eigenvectors of $E(XX^{\mathrm{T}}) = E(XP_{B_0}X^{\mathrm{T}}) + (q - q_0)\sigma^2 I_p$ with eigenvalues λ_i^*, $i = 1, \ldots, p_0$, and from Proposition 3, $A^* = A$ and $\lambda_i^* = \lambda_i + (q - q_0)\sigma^2$. Thus, a standard argument (Sibson, 1979) gives $\partial \text{vec}(a_i^*)/\partial \text{vec}\{E(XX^{\mathrm{T}})\} = a_i^{*\mathrm{T}} \otimes \{\lambda_i^* I_p - E(XX^{\mathrm{T}})\}^+ = a_i^{\mathrm{T}} \otimes M_{Ai}$, where M_{Ai} is defined in Theorem 4.

We next derive the differential of $\mathrm{vec}\{E(XX^{\mathrm{T}})\}$ with respect to $\mathrm{vec}(\Sigma)$. Since $E(XX^{\mathrm{T}}) = \sum_{j=1}^{q}(b_j \otimes I_p)^{\mathrm{T}}\Sigma(b_j \otimes I_p)$, where $\{b_j : j = q_0 + 1, \ldots, q\}$ are defined in the proof of Theorem 4, we have $\partial\mathrm{vec}\{E(XX^{\mathrm{T}})\}/\partial\mathrm{vec}(\Sigma) = W_{B,q}$. The proof is completed by the chain rule. □

References

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Statist.* **34**, 122–48.

De Lathauwer, L., De Moor, B. & Vandewalle, J. (2000a). A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* **21**, 1253–78.

De Lathauwer, L., De Moor, B. & Vandewalle, J. (2000b). On the best rank-1 and rank-(R_1, R_2, …, R_N) approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.* **21**, 1324–42.

Fine, J. (1987). On the validity of the perturbation method in asymptotic theory. *Statistics* **18**, 401–14.

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

Kolda, T. G. & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51**, 455–500.

Li, B., Kim, M. K. & Altman, N. (2010). On dimension folding of matrix- or array-valued statistical objects. *Ann. Statist.* **38**, 1094–121.

Lu, H., Plataniotis, K. N. & Venetsanopoulos, A. N. (2008). MPCA: Multilinear principal component analysis of tensor objects. *IEEE Trans. Neural Networks* **19**, 18–39.

Magnus, J. R. & Neudecker, H. (1979). The commutation matrix: Some properties and applications. *Ann. Statist.* **7**, 381–94.

Sibson, R. (1979). Studies in the robustness of multidimensional scaling: Perturbational analysis of classical scaling. *J. R. Statist. Soc.* B **41**, 217–29.

Yang, J., Zhang, D., Frangi, A. F. & Yang, J. Y. (2004). Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Trans. Pat. Anal. Mach. Intel.* **26**, 131–7.

Ye, J. (2005). Generalized low rank approximations of matrices. *Mach. Learn.* **61**, 167–91.

Zhang, D. & Zhou, Z. H. (2005). $(2D)^2$PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomputing* **69**, 224–31.

[*Received April* 2011. *Revised February* 2012]