PSYCHOMETRIKA—VOL. 57, NO. 2, 203–215
JUNE 1992

BEYOND PRINCIPAL COMPONENT ANALYSIS: A TRILINEAR DECOMPOSITION MODEL AND LEAST SQUARES ESTIMATION

TUAN DINH PHAM

C. N. R. S. AND UNIVERSITY OF GRENOBLE

JOACHIM MÖCKS

DEPARTMENT BIOMETRIE (FK-BR), BOEHRINGER MANNHEIM GMBH

The paper derives sufficient conditions for the consistency and asymptotic normality of the least squares estimator of a trilinear decomposition model for multiway data analysis.

Key Words: consistency, asymptotic normality, factor analysis, principal component analysis, least squares, multidimensional scaling.

1. Introduction

Principal components and factor analysis are well-known tools in multivariate data analysis. For data represented by X_{it} over subjects $i = 1, \ldots, I$ and variables $t = 1, \ldots, T$, these approaches assume a bilinear decomposition:

$$X_{it} = \sum_{k=1}^{K} A_{ki} C_{kt} + E_{it}, \tag{1}$$

where K is the number of components (factors), $\mathbf{A} = (A_{ki})$ and $\mathbf{C} = (C_{kt})$ are unknown matrices of coefficients, and E_{it} denotes the error. The $K \times T$ matrix $\mathbf{C}$ is usually called the "loadings" matrix while the $K \times I$ matrix $\mathbf{A}$ is termed the "scores" matrix. Generalizing from the bilinear form, a data set may also have an intrinsic three-way (three-mode) structure, arising for instance when the variables are repeatedly observed for each subject. An example is provided by the evoked brain potential data from which the present work was actually motivated. Here, X_{ilt} denotes the voltage recorded for subject i at electrode l at time t, where t is measured relative to the onset of an external event (stimulus) being processed by the subject's brain. In the case of three-mode data, X_{ilt}, a natural generalization of (1) is the trilinear decomposition

$$X_{ilt} = \sum_{k=1}^{K} A_{ki} B_{kl} C_{kt} + E_{ilt}, \quad 1 \le i \le I, \ 1 \le l \le L, \ 1 \le t \le T, \tag{2a}$$

where $\mathbf{B} = (B_{kl})$ is a further $K \times L$ matrix of coefficients, corresponding to the repetition mode. For brain potential data, (2a) can be well-motivated from biophysical considerations (Möcks, 1988a, 1988b). It also has been introduced within the framework of multidimensional scaling (Carroll & Chang, 1970; Harshman, 1970). Harshman first noticed an important algebraic difference between the decompositions (1) and (2a).

Requests for reprint should be sent to Tuan Dinh Pham, Laboratory of Modelling and Computation, C. N. R. S., B. P. 53x, 38041 Grenoble Cedex, FRANCE.

0033-3123/92/0600-89168$00.75/0

The matrix **A** and **C** in (1) are obviously not unique, since one can premultiply **A** by any nonsingular matrix $\mathbf{H}'$ (where a prime denotes the transpose), provided that **C** is premultiplied by $\mathbf{H}^{-1}$ (hence, one usually imposes orthogonality conditions on **A**, but then **A** and **C** are still unique only up to a rotation). By contrast, the decomposition in (2a) can be shown (Harshman, 1970) to be essentially unique under mild conditions, a result subsequently generalized by Kruskal (1976, 1977). (Note that a different model for three-mode data has been introduced by Tucker, 1966, under the name "three-mode factor analysis". This model, however, does not possess the uniqueness property).

Although previous work (e.g., Harshman & Lundy, 1984; Kruskal, 1984) studied algebraic properties and algorithmic questions for obtaining the least squares estimator in the model in (2a), nothing is known about their statistical properties. The present work attempts to partially fill this gap. It is shown that the least squares estimator is consistent and asymptotically normal, and the limiting covariance matrix is computed.

2. Preliminaries and Notations

Equation (2a) can be written in compact form using tensor (or Kronecker) product notation. The data X_{ilt} can be viewed as an element **X** of the tensor space $\mathbb{R}^I \otimes \mathbb{R}^L \otimes \mathbb{R}^T$, and (2a) may be written

$$\mathbf{X} = \sum_{k=1}^{K} \mathbf{a}_k \otimes \mathbf{b}_k \otimes \mathbf{c}_k + \mathbf{E},$$

where $\mathbf{a}_k$, $\mathbf{b}_k$, and $\mathbf{c}_k$ denote the k-th rows of **A**, **B**, and **C**, considered as elements of $\mathbb{R}^I$, $\mathbb{R}^L$, and $\mathbb{R}^T$, respectively, and **E** is the tensor (E_{ilt}). It is also convenient to retain the matrix notation, but care is needed not to confuse the dimension of the matrices involved. For this reason, we will sometimes adopt Tucker's (1966) notation, where the dimensions of a matrix are indicated by a pre- and a post-subscript: ${}_K\mathbf{B}_L$, for example, denotes a $K \times L$ matrix. The data **X** can then be viewed as a $IL \times T$ matrix by writing it as ${}_{IL}\mathbf{X}_T$. Likewise, one may regard the data as a $IT \times L$ matrix ${}_{IT}\mathbf{X}_L$, or as a $LT \times I$ matrix ${}_{LT}\mathbf{X}_I$. To write (2a) in matrix form, we shall use Rao's (1973, p. 30) "new product" (with a different notation): $\mathbf{A}*\mathbf{B}$ denotes the $K \times IL$ matrix the rows of which are the vectors $\mathbf{a}_k \times \mathbf{b}_k$. Likewise for $\mathbf{B} * \mathbf{C}$ and $\mathbf{A} * \mathbf{C}$. Then (2a) can be written in the following equivalent forms:

$$\begin{aligned} {}_{IL}\mathbf{X}_T &= (\mathbf{A} * \mathbf{B})'\mathbf{C} + {}_{IL}E_T, \\ {}_{IT}\mathbf{X}_L &= (\mathbf{A} * \mathbf{C})'\mathbf{B} + {}_{IT}E_L, \\ {}_{LT}\mathbf{X}_I &= (\mathbf{B} * \mathbf{C})'\mathbf{A} + {}_{LT}\mathbf{E}_I. \end{aligned} \tag{2b}$$

Any of the above notations will be used according to convenience; also, for simplicity, we will drop pre- and post-subscripts on the matrices **X** and **E** when there is no risk of confusion.

Suppose that the errors E_{ilt} in (2) are independent identically distributed random variables with zero means and variance σ^2. Then (2) can be viewed as a variant of the factor analysis model in which the errors have the same variance, the factor scores A_{ki} can be regarded as unknown constants rather than random variables, and the loading matrix has a special structure, namely $\mathbf{B} * \mathbf{C}$. The assumption that the A_{ki} are fixed constants is not essential. It only serves to justify the use of the least squares method, and *one could also work with random scores without affecting the asymptotic results* shown below. Note that the above assumption makes it possible to treat **A**, **B**, and **C** in

a symmetric way; for example, one can regard $\mathbf{A} * \mathbf{B}$ as the loading matrix and $\mathbf{C}$ as the scores. Likewise, $\mathbf{A} * \mathbf{C}$ can be conceived of as a loading matrix and $\mathbf{B}$ as the scores.

The simplest method to estimate the parameters of the above model is the least squares which consists of minimizing the sum of squares of the errors,

$$Q = \sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{t=1}^{T} \left(X_{ilt} - \sum_{k=1}^{K} A_{ki} B_{kl} C_{kt} \right)^2 = \left\| \mathbf{X} - \sum_{k=1}^{K} \mathbf{a}_k \otimes \mathbf{b}_k \otimes \mathbf{c}_k \right\|^2, \qquad (3)$$

where $\| \cdot \|$ denotes the Euclidean norm. To minimize Q, one may equate to zero its (partial) derivatives with respect to the parameters A_{ki}, B_{kl}, and C_{kt}. For fixed $\mathbf{B}$ and $\mathbf{C}$, (2) is a linear model with respect to $\mathbf{A}$, hence the derivative of the criterion Q with respect to A_{ki} can be obtained in the same way as in linear models. Thus, we obtain the equation

$$\mathbf{S}_{bc}\mathbf{A} = (\mathbf{B} * \mathbf{C})_{LT}\mathbf{X}_I,$$

where $\mathbf{S}_{bc} = (\mathbf{B} * \mathbf{C})(\mathbf{B} * \mathbf{C})'$. Similarly, by equating to zero the derivatives of Q with respect to $\mathbf{B}$ and $\mathbf{C}$, respectively, one gets the equations

$$\mathbf{S}_{ac}\mathbf{B} = (\mathbf{A} * \mathbf{C})_{IT}\mathbf{X}_L, \qquad \mathbf{S}_{ab}\mathbf{C} = (\mathbf{A} * \mathbf{B})_{IL}\mathbf{X}_T,$$

where $\mathbf{S}_{ab}$ and $\mathbf{S}_{ac}$ are defined in the same way. The above three equations define the least squares estimators $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, and $\hat{\mathbf{C}}$ of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$. It can be seen that these estimators are defined only up to a scale factor and a permutation of their rows. Indeed, the above equations are unchanged when one premultiplies $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ with $\mathbf{PD}_a$, $\mathbf{PD}_b$, and $\mathbf{PD}_c$, respectively, where $\mathbf{P}$ is a permutation matrix and $\mathbf{D}_a$, $\mathbf{D}_b$, and $\mathbf{D}_c$ are diagonal matrices with a product equal to the identity matrix.

In the following, the asymptotic properties of the least squares estimators will be studied when one of the dimensions, say I, goes to infinity while the others remain fixed. This is one possible way to introduce "asymptotics." Clearly, it makes no sense to assume all three dimensions tend to infinity simultaneously, since then the number of parameters would also increase to infinity. We could consider the case when two of three dimensions go to infinity while the remaining stays fixed, but for simplicity, we do not. In the case when I goes to infinity and L and T remain fixed, the parameters A_{ki}, $i = 1, \ldots, I$, cannot be estimated consistently, but we can expect consistency and asymptotic normality of the estimators of the parameters B_{kl}, $l = 1, \ldots, L$ and C_{kt}, $t = 1, \ldots, T$. One may regard our model as specified by $K(L + T)$ parameters $B_{kl}, \ldots, B_{kL}$, and $C_{k1}, \ldots, C_{kT}$, $k = 1, \ldots, K$, and K sequences of numbers A_{k1}, $A_{k2}, \ldots$, $k = 1, \ldots, K$. Only the first $K(L + T)$ parameters will be estimated; the A_{ki} play the role of nuisance parameters. Another possibility is to assume $(A_{1i}, \ldots, A_{Ki})'$, $i = 1, 2, \ldots, I$ be independent identically distributed (i.i.d.) random vectors, in which case the index i may be viewed as a replication index and I as the *sample size*. Our results hold for both situations. Specifically, we shall assume

(M0) $(A_{1i}, \ldots, A_{Ki})'$, $i = 1, 2, \ldots$ are either a deterministic or random sequence of vectors in $\mathbb{R}^K$, such that the matrix with (k, k') element $(\sum_{i=1}^{I} A_{ki}A_{k'i})/I$ converge almost surely to a limit $\mathbf{R}_a$ as $I \to \infty$.

Note that in the case where $(A_{1i}, \ldots, A_{Ki})'$ are random vectors, (M0) holds by the strong law of large numbers if they are iid and admit second moments. Also when $(A_{1i}, \ldots, A_{Ki})'$ are deterministic, they denote the *true* unknown parameters, and not free parameters as in the criterion (3).

For the asymptotic normality of the estimators, we will need a further assumption as follows:

(M1) For all $k = 1, \ldots, K$, $\max_{k=1}^{I} A_{ki}^2/I \to 0$ almost surely as $I \to \infty$.

This is a very weak assumption. For example, it would hold if for some $\delta > 0$, $\sum_{i=1}^{I} |A_{ki}|^{2+\delta}/I$ is bounded almost surely as $I \to \infty$, since $[\max A_{ki}^2/I]^{1+\delta/2} \leq [\sum_{i=1}^{I} |A_{ki}|^{2+\delta}/I]/I^{\delta/2}$. In the case where the vectors $(A_{1i}, \ldots, A_{Ki})'$ are random, the latter condition follows from strong law of large numbers, provided that they are iid. and admit $2 + \delta$ absolute moments.

Although (2a) has the form of a (nonlinear) regression model (in the case where A_{ik} are deterministic), standard asymptotic theory for nonlinear least squares estimation does not apply, since the number of the parameters A_{ki} goes to infinity with I. Our method consists of eliminating these parameters leading to an estimation criterion Q^*, given by (4) below, containing only the parameters B_{kl} and C_{kt}. However, this criterion does not have the form of a sum of squares so that new arguments are needed to obtain the almost sure consistency and the asymptotic normality of the estimate (although for the latter property, the proof follows rather closely the standard arguments).

For fixed $\mathbf{B}$ and $\mathbf{C}$, the least squares criterion Q, defined in (3), is minimized when $\mathbf{A} = \mathbf{S}_{bc}^{-1}(\mathbf{B} * \mathbf{C})\mathbf{X}$. Inserting this into Q, we obtain

$$\text{tr } ({}_I\mathbf{X}'_{LT}{}_{LT}\mathbf{X}_I) - \text{tr } [{}_I\mathbf{X}'_{LT}(\mathbf{B} * \mathbf{C})'\mathbf{S}_{bc}^{-1}(\mathbf{B} * \mathbf{C})_{LT}\mathbf{X}_I].$$

Using the relation tr $(\mathbf{AB})$, = tr $(\mathbf{BA})$, the above expression equals

$$Q^* = \text{tr } \{[\mathbf{I} - (\mathbf{B} * \mathbf{C})'\mathbf{S}_{bc}^{-1}(\mathbf{B} * \mathbf{C})]\mathbf{X}\mathbf{X}'\} = \text{tr } [(\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}})_{LT}\mathbf{X}\mathbf{X}'_{LT}], \tag{4}$$

where $\mathbf{I}$ denotes the identity matrix, and

$$\mathbf{P}_{\mathbf{B}*\mathbf{C}} = (\mathbf{B} * \mathbf{C})'\mathbf{S}_{bc}^{-1}(\mathbf{B} * \mathbf{C})$$

denotes the projector that projects orthogonally onto the linear subspace spanned by the rows of $\mathbf{B} * \mathbf{C}$. Thus, the least squares estimators $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ of $\mathbf{B}$ and $\mathbf{C}$ are those that minimize the criterion Q^*, defined by (4).

3. Consistency of the Least Squares Estimators

Before discussing the consistency of the estimator, one must ensure that the model is identifiable; that is there exists no other set of parameters leading to the same model. The following conditions are sufficient for identifiability, up to a scale factor and a permutation:

(A0) The vectors $\mathbf{a}_1, \ldots, \mathbf{a}_K$ are linearly independent.

(A1) The tensor products $\mathbf{b}_1 \otimes \mathbf{c}_1, \ldots, \mathbf{b}_K \otimes \mathbf{c}_K$ are linearly independent and the linear subspace spanned by them contains no other tensor product (of a vector in $\mathbb{R}^L$ with a vector in $\mathbb{R}^T$) but their multiples.

To see that (A0) and (A1) imply uniqueness of the model (up to a scale factor and a permutation) note that $E({}_{LT}\mathbf{X}_I) = (\mathbf{B} * \mathbf{C})'\mathbf{A}$, and hence, the linear independence of the $\mathbf{a}_k$ implies that the columns of the matrix $E({}_{LT}\mathbf{X}_I)$ span the same linear subspace as the $\mathbf{b}_k \otimes \mathbf{c}_k$. The latter subspace is thus unique and the uniqueness of the $\mathbf{b}_k$ and $\mathbf{c}_k$ then follows from (A1).

Since we are interested in the asymptotic behavior of the estimates, (A0) needs hold only for I large enough and it should continue to hold for $I \to \infty$. Thus, we can replace (A0) by

(A0′) The matrix $\mathbf{R}_a$ in (M0) is nonsingular.

Condition (A1) is not sufficient for proving the consistency of the estimators, and therefore, a stronger condition must be introduced (which was also considered in Harshman, 1970, and Möcks, 1988a):

(A2) The vectors $\mathbf{c}_1, \ldots, \mathbf{c}_K$ are linearly independent and no two of $\mathbf{b}_1, \ldots, \mathbf{b}_K$ are a multiple of each other.

(It is noted that condition (A2) is not symmetric with respect to **B** and **C** since the stronger linear independence requirement on the $\mathbf{c}_k$ makes it possible to weaken the requirement on the $\mathbf{b}_k$. In practice, it may happen that the dimension L is less than K so the $\mathbf{b}_k$ cannot be linearly independent. Of course, one could replace (A2) by an analogous condition in which the matrices **B** and **C** are interchanged.)

To see that (A2) implies (A1), consider the linear relation $\sum_{k=1}^{K} \mu_k \mathbf{b}_k \otimes \mathbf{c}_k = \mathbf{0}$, or equivalently, $\sum_{k=1}^{K} \mu_k B_{kl} \mathbf{c}_k = \mathbf{0}$ for all l. The linear independence of the $\mathbf{c}_k$ implies $\mu_k B_{kl} = 0$ for all k and l, but this is not possible unless $\mu_k = 0$ for all k. In the same way, if $\sum_{k=1}^{K} \mu_k \mathbf{b}_k \otimes \mathbf{c}_k = \mathbf{b} \otimes \mathbf{c}$, for some set of coefficients μ_k and some vectors **b** and **c** of $\mathbb{R}^L$ and $\mathbb{R}^T$, respectively, then the linear independence of the $\mathbf{c}_k$ imply that $\mu_k B_{kl}$ must be of the form $\beta_l \gamma_k$. But then the second condition in (A2) implies that all but one μ_k must be zero.

The vectors $\mathbf{a}_k$, $\mathbf{b}_k$, and $\mathbf{c}_k$ in the above conditions refer to the *true values* of the model parameters while the matrices **B** and **C** in the criterion Q^* represent free parameters. To avoid confusion, we shall reserve in this section, the notation **B** and **C**, and $\mathbf{b}_k$ and $\mathbf{c}_k$ for the true values. We also normalize the $\mathbf{b}_k$ and $\mathbf{c}_k$ to have unit norm to eliminate the scale factor. The free parameters are viewed as an element in the set Θ of all $*$ products of a $K \times L$ and a $K \otimes T$ matrix, with rows having unit norm. Thus, the least squares estimators of $\mathbf{B} * \mathbf{C}$ is the matrix $\hat{\mathbf{H}}$ realizing the minimum of $\text{tr}\,[(\mathbf{I} - \mathbf{P}_\mathbf{H})_{LT}\mathbf{X}\mathbf{X}'_{LT}]$ for among all **H** in Θ. Once $\hat{\mathbf{H}}$ has been found, it may be factored (with respect to the $*$ product) to get the least squares estimators $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ of **B** and **C**.

The consistency study of the $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ presents two technical difficulties. First, the function $\text{tr}\,[(\mathbf{I} - \mathbf{P}_\mathbf{H})_{LT}\mathbf{X}\mathbf{X}'_{LT}]$ is not everywhere continuous in Θ, since $\mathbf{H}\mathbf{H}'$ may be singular (in this case, $\mathbf{P}_\mathbf{H}$ is still defined as the projector to the linear subspace spanned by the rows of **H**, but this space has dimension less than K, while the one corresponding to other points in Θ, however close to **H**, generally has dimension K; hence, the discontinuity). Secondly, the matrix $\mathbf{H} \in \Theta$ may not admit a unique factorization, unless its factors satisfy (A1). Of course, one may exclude such **H** from the parameter set, but then it *will not be closed*, and the standard argument for proving consistency no longer applies. Moreover, in practice, the minimization of the criterion Q^* is usually performed without any restriction on Θ. Therefore, the full set Θ will be retained as parameter set and a specific proof for the consistency of the estimator will be provided. Observe that the criterion $\text{tr}\,[(\mathbf{I} - \mathbf{P}_\mathbf{H})_{LT}\mathbf{X}\mathbf{X}'_{LT}]$, as a function of $\mathbf{P}_\mathbf{H}$, is a well-behaved function. Thus, we introduce the set $\Pi = \{\mathbf{P}_\mathbf{H}, \mathbf{H} \in \Theta\}$ and consider the minimization of $\text{tr}\,[(\mathbf{I} - \mathbf{P})_{LT}\mathbf{X}\mathbf{X}'_{LT}]$ with respect to $\mathbf{P} \in \Pi$. Since Π is not closed (if $\mathbf{H}_n$ is a sequence in Θ converging to a matrix of rank less than K, then $\mathbf{P}_{\mathbf{H}_n}$ could converge to a projector not of the form $\mathbf{P}_\mathbf{H}$ for some $\mathbf{H} \in \Theta$), we must enlarge it to $\bar{\Pi}$, the closure of Π (i.e., the set of points that can be approached arbitrarily close by points in Π). Then we can apply

standard techniques to prove that the minimizer $\hat{\mathbf{P}}$ of tr $[(\mathbf{I} - \mathbf{P})_{LT}\mathbf{X}\mathbf{X}'_{LT}]$ in $\bar{\Pi}$, converges almost surely to $\mathbf{P}_{\mathbf{B}*\mathbf{C}}$ as $\mathbf{I} \to \infty$. The difficult part consists of showing that $\hat{\mathbf{P}}$ must belong to Π (i.e., $\hat{\mathbf{P}} = \mathbf{P}_{\hat{\mathbf{H}}}$ for some $\hat{\mathbf{H}} \in \Theta$ or the existence of the least squares estimator) for $\mathbf{I}$ large enough, and that the convergence of $\hat{\mathbf{P}}$ to $\mathbf{P}_{\mathbf{B}*\mathbf{C}}$ implies that of $\hat{\mathbf{H}}$ to $\mathbf{B} * \mathbf{C}$. However, due to the fact that the model is unchanged when the rows of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are permuted, one can only expect that $\hat{\mathbf{H}}$ converges to $\mathbf{B} * \mathbf{C}$ *up to a permutation*. To make this concept more precise, we introduce the "distance" $d(\mathbf{H}, \mathbf{G}) = d_1(\mathbf{H}, \mathbf{G}) + d_2(\mathbf{H}, \mathbf{G})$, where

$$d_1(\mathbf{H}, \mathbf{G}) = \operatorname*{Max}_{k=1,\ldots,K}\left(\operatorname*{Min}_{r=1,\ldots,K} \|\mathbf{h}_k - \mathbf{g}_r\|\right),$$

$$d_2(\mathbf{H}, \mathbf{G}) = \operatorname*{Max}_{k=1,\ldots,K}\left(\operatorname*{Min}_{r=1,\ldots,K} \|\mathbf{h}_r - \mathbf{g}_k\|\right),$$

and $\mathbf{h}_k$ and $\mathbf{g}_k$ denote the row vectors of $\mathbf{H}$, $\mathbf{G}$. It can be shown that $d(\cdot, \cdot)$ possesses the property of a distance if matrices having the same row vectors *up to a permutation* are identified as the same (clearly, $d_1(\mathbf{H}, \mathbf{G}) = 0$ implies $\mathbf{h}_k$ is one of the $\mathbf{g}_1, \ldots, \mathbf{g}_K$; and $d_1(\mathbf{H}, \mathbf{G}) = 0$ implies $\mathbf{g}_k$ is one of the $\mathbf{h}_1, \ldots, \mathbf{h}_K$; hence, $d(\mathbf{H}, \mathbf{G}) = 0$ implies $\mathbf{h}_1, \ldots, \mathbf{h}_K$ is a permutation of $\mathbf{g}_1, \ldots, \mathbf{g}_K$).

Theorem 1. Under the assumptions (M0) and (A0′) and supposing that $\mathbf{B} * \mathbf{C}$ has rank K, the minimizer $\hat{\mathbf{P}}$ of tr $[(\mathbf{I} - \mathbf{P})\mathbf{X}\mathbf{X}']$ in $\bar{\Pi}$ converges almost surely to $\mathbf{P}_{\mathbf{B}*\mathbf{C}}$ as $I \to \infty$.

The proof of this Theorem is based on the following Lemma.

Lemma 1. Under assumption (M0), ${}_{LT}\mathbf{X}\mathbf{X}'_{LT}/I$ converges almost surely to $(\mathbf{B} * \mathbf{C})'\mathbf{R}_a(\mathbf{B} * \mathbf{C}) + \sigma^2\mathbf{I}$ as $I \to \infty$.

Proof. A simple computation shows that ${}_{LT}\mathbf{X}\mathbf{X}'_{LT}/I$ equals

$$(\mathbf{B} * \mathbf{C})'\left(\frac{\mathbf{A}\mathbf{A}'}{I}\right)(\mathbf{B} * \mathbf{C}) + (\mathbf{B} * \mathbf{C})'\,\frac{\mathbf{B}\mathbf{E}'}{I} + \mathbf{E}\mathbf{B}'\,\frac{(\mathbf{B} * \mathbf{C})}{I} + \frac{\mathbf{E}\mathbf{E}'}{I}.$$

The first term converges to $(\mathbf{B} * \mathbf{C})'\mathbf{R}_a(\mathbf{B} * \mathbf{C})$ by assumption. The last term converges to $\sigma^2\mathbf{I}$ by the strong law of large numbers. The second term is the matrix with the element $((l, t), (l', t'))$ of the form $\sum_{k=1}^{K} \mathbf{B}_{kl}\mathbf{C}_{kt}$ $(\sum_{i=1}^{I} \mathbf{A}_{ki}\mathbf{E}_{il't'}/I)$, which converges to 0 as $I \to \infty$, by a strong law for sums of independent random variables (e.g., see Loève, 1963). The same is true for the third term and the result follows. □

Proof of Theorem 1. We first note that $\hat{\mathbf{P}}$ is the maximizer of tr $(\mathbf{P}_{LT}\mathbf{X}\mathbf{X}'_{LT})$ in $\bar{\Pi}$. By Lemma 1, this function converges almost surely as $I \to \infty$ to tr $[\mathbf{P}(\mathbf{B} * \mathbf{C})'$ $\mathbf{R}_a(\mathbf{B} * \mathbf{C}] + \sigma^2$ tr $(\mathbf{P})$. Now, $\mathbf{P}$ can be written as $\sum_{k=1}^{K'} \mathbf{p}_k\mathbf{p}'_k$, where $\{\mathbf{p}_1, \ldots, \mathbf{p}_{K'}\}$ is a orthonormal basis of the image space of $\mathbf{P}$; hence, tr $(\mathbf{P}) = K' \leq K$, and $\mathbf{P}$ is bounded in $\bar{\Pi}$, implying that the above convergence is uniform in $\bar{\Pi}$. Thus, to show that $\hat{\mathbf{P}} \to \mathbf{P}_{\mathbf{B}*\mathbf{C}}$, one needs only to show, by a simple standard argument, that the function tr $[\mathbf{P}(\mathbf{B} * \mathbf{C})'\mathbf{R}_a(\mathbf{B} * \mathbf{C})] + \sigma^2$ tr $(\mathbf{P})$ is maximized in $\bar{\Pi}$ at $\mathbf{P} = \mathbf{P}_{\mathbf{B}*\mathbf{C}}$. Since $\mathbf{B} * \mathbf{C}$ has rank K, tr $(\mathbf{P}_{\mathbf{B}*\mathbf{C}}) = K \geq$ tr $(\mathbf{P})$, and since tr $[\mathbf{P}_{\mathbf{B}*\mathbf{C}}(\mathbf{B} * \mathbf{C})'\mathbf{R}_a(\mathbf{B} * \mathbf{C})] =$ tr $[(\mathbf{B} * \mathbf{C})'\mathbf{R}_a(\mathbf{B} * \mathbf{C})]$, it suffices to show that

$$\text{tr}\,[(\mathbf{I}-\mathbf{P})(\mathbf{B}*\mathbf{C})'\mathbf{R}_a(\mathbf{B}*\mathbf{C})] > 0, \qquad \text{for all } \mathbf{P} \neq \mathbf{P}_{\mathbf{B}*\mathbf{C}},\ \mathbf{P} \in \bar{\Pi}.$$

The above left-hand-side can be written as tr $[(\mathbf{I}-\mathbf{P})(\mathbf{B}*\mathbf{C})'\mathbf{R}_a(\mathbf{B}*\mathbf{C})(\mathbf{I}-\mathbf{P})]$ since tr $(\mathbf{AB})$ = tr $(\mathbf{BA})$ and the projector $\mathbf{P}$ is idempotent ($\mathbf{P}^2 = \mathbf{P}$). Now,

$$\text{tr}\,[(\mathbf{I}-\mathbf{P})(\mathbf{B}*\mathbf{C})'\mathbf{R}_a(\mathbf{B}*\mathbf{C})(\mathbf{I}-\mathbf{P})] \geq 0,$$

since the trace of a positive semidefinite matrix is nonnegative. Equality is attained if and only if $(\mathbf{B}*\mathbf{C})(\mathbf{I}-\mathbf{P}) = 0$, since $\mathbf{R}_a$ is positive definite. Suppose that $(\mathbf{B}*\mathbf{C})(\mathbf{I}-\mathbf{P}) = 0$, or equivalently, $\mathbf{P}_{\mathbf{B}*\mathbf{C}}(\mathbf{I}-\mathbf{P}) = 0$. Then, again from the identity tr $(\mathbf{AB})$ = tr $(\mathbf{BA})$ and the idempotence of projectors,

$$\begin{aligned}\text{tr}\,(\mathbf{P}_{\mathbf{B}*\mathbf{C}}-\mathbf{P})^2 &= \text{tr}\,(\mathbf{P}_{\mathbf{B}*\mathbf{C}}+\mathbf{P}-2\mathbf{P}_{\mathbf{B}*\mathbf{C}}\mathbf{P}) \\ &= 2\,\text{tr}\,[\mathbf{P}_{\mathbf{B}*\mathbf{C}}(\mathbf{I}-\mathbf{P})] + \text{tr}\,(\mathbf{P}) - \text{tr}\,(\mathbf{P}_{\mathbf{B}*\mathbf{C}}) \leq 0.\end{aligned}$$

But the left-hand-side is the sum of squares of all elements of $\mathbf{P}_{\mathbf{B}*\mathbf{C}} - \mathbf{P}$, and therefore, $\mathbf{P}$ must equal $\mathbf{P}_{\mathbf{B}*\mathbf{C}}$. □

Lemma 2. Under (A1), for any sequence $\mathbf{H}(I)$, $I = 1, 2, \ldots$ in Θ such that $\mathbf{P}_{\mathbf{H}(I)} \to \mathbf{P}_{\mathbf{B}*\mathbf{C}}$, one has $d_1[\mathbf{H}(I), \mathbf{B}*\mathbf{C}] \to 0$. Suppose that additionally the following condition holds:

(C) For any sequence of matrices $\mathbf{H}(I)$ with $\mathbf{P}_{\mathbf{H}(I)} \to \mathbf{P}_{\mathbf{B}*\mathbf{C}}$, and with row $\mathbf{h}_k(I)$ converging as $I \to \infty$ to some row $\mathbf{b}_{j_k} \otimes \mathbf{c}_{j_k}$ of $\mathbf{B} \otimes \mathbf{C}$, the subscripts $j_1, \ldots, j_K$ must be distinct;

then also $d_2[\mathbf{H}(I), \mathbf{B}*\mathbf{C}] \to 0$.

Proof. Denote by $\mathbf{h}_k(I)$ the row vectors of $\mathbf{H}(I)$. Since the set of sequences $\mathbf{h}_k(I)$, $k = 1, \ldots, K$, are bounded, for any set of subsequences $\mathbf{h}_k(I_n)$, one can extract a further set of subsequences, denoted by $\mathbf{h}_k(I'_n)$, that converge to $\mathbf{h}_k$, say. But $\mathbf{P}_{\mathbf{H}(I)}\mathbf{h}_k(I) = \mathbf{h}_k(I)$ and by assumption $\mathbf{P}_{\mathbf{H}(I)} \to \mathbf{P}_{\mathbf{B}*\mathbf{C}}$ as $I \to \infty$. Thus, by taking the limit along the subsequence I'_n, $\mathbf{h}_k$ must be in the linear subspace spanned by the rows of $\mathbf{B}*\mathbf{C}$, namely the tensor products $\mathbf{b}_1 \otimes \mathbf{c}_1, \ldots, \mathbf{b}_K \otimes \mathbf{c}_K$. By (A1), $\mathbf{h}_k$ must be a multiple of one of these tensor products (i.e., $\mathbf{h}_k = \mathbf{b}_{j_k} \otimes \mathbf{c}_{j_k}$ for some j_k in $\{1, \ldots, K\}$), since the $\mathbf{b}_k \otimes \mathbf{c}_k$ and $\mathbf{h}_k$ all have unit norm. The indexes $j_1, \ldots, j_K$ need not be distinct, but under the condition (C) they are so.

Now, suppose that the first conclusion of the lemma does not hold. Then there exists some $k \in \{1, \ldots, K\}$ such that one can extract from the sequence $\mathbf{h}_k(I)$, a subsequence, say $\mathbf{h}_k(I_n)$, for which its distances to $\mathbf{b}_1 \otimes \mathbf{c}_1, \ldots, \mathbf{b}_K \otimes \mathbf{c}_K$ are all greater than a positive constant. From the above result, one can extract a further couple of subsequences, denoted by $\mathbf{h}_k(I'_n)$, that converge to $\mathbf{b}_{j_k} \otimes \mathbf{c}_{j_k}$, for some j_k in $\{1, \ldots, K\}$, leading to a contradiction. This proves the first result of the lemma.

We prove the second conclusion of the lemma again by contradiction. If it does not hold, there exists k^* in $\{1, \ldots, K\}$ and subsequences $\mathbf{h}_1(I_n), \ldots, \mathbf{h}_K(I_n)$ that are at a distance to $\mathbf{b}_{k*} \otimes \mathbf{c}_{k*}$ greater than a positive constant. But by the result at the beginning of the proof, there is a further subsequence $\mathbf{h}_1(I'_n), \ldots, \mathbf{h}_K(I'_n)$, converging to $\mathbf{b}_{j_1} \otimes \mathbf{c}_{j_1}, \ldots, \mathbf{b}_{j_K} \otimes \mathbf{c}_{j_K}$. Clearly, the indexes $j_1, \ldots, j_K$ must differ from k^*, contradicting the condition (C). Hence the desired result. □

Corollary: Under the condition of Lemma 2, any element $\bar{\mathbf{P}}$ of $\bar{\Pi}$ close enough to $\mathbf{P}_{\mathbf{B}*\mathbf{C}}$ is also an element of Π.

Proof. Since the row vectors $\mathbf{b}_k$ and $\mathbf{c}_k$ of $\mathbf{B}$ and $\mathbf{C}$ satisfy the condition (A1), any set of vectors sufficiently close to them also satisfy the same condition. Thus, there exists $\eta > 0$ such that if $\mathbf{H} = \bar{\mathbf{B}} * \bar{\mathbf{C}} \in \Pi$ and $d(\mathbf{H}, \mathbf{B} * \mathbf{C}) \leq \eta$, then the row vectors $\bar{\mathbf{b}}_k$ and $\bar{\mathbf{c}}_k$ of $\mathbf{B}$ and $\mathbf{C}$ also satisfy (A1). It follows that that mapping $\mathbf{H} \mapsto \mathbf{P}_\mathbf{H}$ is continuous and one-to-one from $\Theta_\eta = \{\mathbf{H}, d(\mathbf{H}, \mathbf{B} * \mathbf{C}) \leq \eta\}$ onto a subset Π_η of Π. The latter is compact being the image of a compact set by a continuous mapping. Now, let ε_n, $n = 1, 2, \ldots$ be a sequence of positive numbers tending to 0. We claim that for n sufficiently large, the set $\{\mathbf{P} \in \Pi, \|\mathbf{P} - \mathbf{P}_{\mathbf{B}*\mathbf{C}}\| < \varepsilon_n\}$ is included in Π_η (here, $\|\mathbf{M}\| = \text{tr}\,(\mathbf{M}\mathbf{M}')^{1/2}$ denotes the Euclidean norm of the matrix $\mathbf{M}$). Indeed, if this is not true, there is a sequence $\mathbf{P}_n = \mathbf{P}_{\mathbf{H}_n}$ in Π such that $\|\mathbf{P}_n - \mathbf{P}_{\mathbf{B}*\mathbf{C}}\| < \varepsilon_n$, but $\mathbf{P}_n \notin \Pi_\eta$; that is, $d(\mathbf{H}_n, \mathbf{B} * \mathbf{C}) > \eta$. But this contradicts Lemma 2 since $\mathbf{P}_n \to \mathbf{P}_{\mathbf{B}*\mathbf{C}}$. Thus, we have proved that for some $\varepsilon > 0$, $\mathbf{P} \in \Pi$ and $\|\mathbf{P} - \mathbf{P}_{\mathbf{B}*\mathbf{C}}\| < \varepsilon$ implies $\mathbf{P} \in \Pi_\eta$. Now, for any $\bar{\mathbf{P}} \in \bar{\Pi}$ and satisfying $\|\bar{\mathbf{P}} - \mathbf{P}_{\mathbf{B}*\mathbf{C}}\| < \varepsilon$, there exists $\mathbf{P}$ in Π arbitrarily close to it so that $\|\mathbf{P} - \mathbf{P}_{\mathbf{B}*\mathbf{C}}\| < \varepsilon$, and hence, $\mathbf{P} \in \Pi_\eta$. But since $\mathbf{P}$ can be taken arbitrarily close $\bar{\mathbf{P}}$, the latter belongs to the closure of Π_η which coincides with Π_η because of compactness. □

Condition (C), used in Lemma 2, is rather technical and difficult to check. The following lemma shows that it is implied by the simpler condition (A2).

Lemma 3. Under (A2), condition (C) of Lemma 2 is satisfied.

Proof. Suppose that there exists a sequence $\mathbf{h}_k(I) = \mathbf{b}_k(I) \otimes \mathbf{c}_K(I)$, $k = 1, \ldots, K$, satisfying the properties stated in (C), but the subscripts $j_1, \ldots, j_K$ as defined there are not distinct. We shall show that this leads to a contradiction. By renumbering, we may assume that $j_k = k$ for $k = 1, \ldots, K'$, $j_k \in \{1, \ldots, K'\}$ for $k = K' + 1, \ldots, K$, with K' being the number of distinct j_k. Note that the assumption $\mathbf{h}_k(I) \to \mathbf{b}_{j_k} \otimes \mathbf{c}_{j_k}$ implies $\mathbf{b}_k(I) \to \mathbf{b}_{j_k}$ and $\mathbf{c}_k(I) \to \mathbf{c}_{j_k}$. Indeed, if this is not true, there exists a pair of subsequences $\mathbf{b}_k(I_n)$ and $\mathbf{c}_k(I_n)$ that are at a distance to $\mathbf{b}_{j_k}$ and $\mathbf{c}_{j_k}$ greater than some positive constant. Since these subsequences are bounded, one may extract from them a further pair of subsequences $\mathbf{b}_k(I'_n)$ and $\mathbf{c}_k(I'_n)$ converging to $\bar{\mathbf{b}}_k$ and $\bar{\mathbf{c}}_k$ distinct from $\mathbf{b}_{j_k}$, $\mathbf{c}_{j_k}$. But this contradicts the fact that $\mathbf{h}_k(I) \to \mathbf{b}_{j_k} \otimes \mathbf{c}_{j_k}$.

Let $\mathbf{c}_k(I)^*$ be obtained by orthogonalizing the $\mathbf{c}_k(I)$; that is, the $\mathbf{c}_k(I)^*$ form an orthonormal system of vectors satisfying

$$\mathbf{c}_k(I) = \lambda_k(I)\mathbf{c}_k(I)^* + \mathbf{u}_{k-1}(I), \qquad k = 1, \ldots, K,$$

where $\lambda_k(I)$ is a scalar and $\mathbf{u}_{k-1}(I)$ is an element of the linear subspace $U_{k-1}(I)$ spanned by $\mathbf{c}_1(I), \ldots, \mathbf{c}_{k-1}(I)$ ($\mathbf{u}_0 = \mathbf{0}$) and $U_0(I) = \{\mathbf{0}\}$, by convention). Thus,

$$\mathbf{h}_k(I) = \lambda_k(I)\mathbf{b}_k(I) \otimes \mathbf{c}_k(I)^* + \mathbf{b}_k(I) \otimes \mathbf{u}_{k-1}(I).$$

Clearly, $\mathbf{b}_k(I) \otimes \mathbf{c}_k(I)^*$ is orthogonal to the space $\mathbb{R}^L \otimes U_{k-1}(I)$ consisting of all linear combinations of tensor products of a vector in $\mathbb{R}^L$ and a vector in $U_{k-1}(I)$. This space contains $\mathbf{h}_1(I), \ldots, \mathbf{h}_{k-1}(I)$ and $\mathbf{b}_k(I) \otimes \mathbf{u}_{k-1}(I)$. Let $V_{k-1}(I)$ be spanned by $\mathbf{h}_1(I), \ldots, \mathbf{h}_{k-1}(I)$ ($V_0(I) = \{\mathbf{0}\}$, by convention) and denote by $\mathbf{y}_{k-1}$ the difference between $\mathbf{b}_k(I) \otimes \mathbf{u}_{k-1}(I)$ and its orthogonal projection onto $V_{k-1}(I)$. Then,

$$\mathbf{h}_k(I) = \lambda_k(I)\mathbf{b}_k(I) \otimes \mathbf{c}_k(I)^* + \mathbf{y}_{k-1}(I) + \text{a tensor in } V_{k-1}(I),$$

and the sum of the first two terms of the above right-hand-side is orthogonal to V_{k-1}. Define

$$\mathbf{v}_k(I) = \mu_k(I)[\lambda_k(I)\mathbf{b}_k(I) \otimes \mathbf{c}_k(I)^* + \mathbf{y}_{k-1}(I)],$$

where $\mu_k(I)$ is chosen such that $\mathbf{v}_k(I)$ has unit norm. Then $\mathbf{v}_1(I), \ldots, \mathbf{v}_k(I)$ constitute an orthonormal basis of $V_k(I)$. Since $\mathbf{z}_{k-1}(I) = \mu_k(I)\mathbf{y}_{k-1}(I)$ and $\mu_k(I)\lambda_k(I)\mathbf{b}_k(I) \otimes \mathbf{c}_k(I)^*$ are orthogonal to each other, both have norm less than 1, and since $\|\mathbf{b}_k(I) \otimes \mathbf{c}_k(I)^*\| = 1$, $\mu_k(I)\lambda_k(I)$ is bounded. Therefore, one can extract a subsequence, say I_n, such that $\mu_k(I_n)\lambda_k(I_n)$, $\mathbf{c}_k(I_n)^*$, and $\mathbf{z}_{k-1}(I_n)$ all converge to some limit ν_k, $\mathbf{c}_k$ and $\mathbf{z}_{k-1}$, say. Since $\mathbf{b}_k(I_n)$ converges to $\mathbf{b}_{j_k}$, $\mathbf{v}_k(I_n)$ converges to $\mathbf{v}_k = \nu_k\mathbf{b}_{j_k}(I) \otimes \mathbf{c}_k(I)^* + \mathbf{z}_{k-1}$. Since $\mathbf{v}_k(I) \in V_k(I)$, which is spanned by the row vectors of $\mathbf{H}(I)$, $\mathbf{P}_{\mathbf{H}(I)}\mathbf{v}_k(I) = \mathbf{v}_k(I)$, and by taking the limit along I_n, one gets $\mathbf{P}_{\mathbf{B}*\mathbf{C}}\mathbf{v}_k = \mathbf{v}_k$. The linear subspace spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_K$ thus is contained in the one spanned by the row vectors of $\mathbf{B} * \mathbf{C}$, and since both have dimension K, they must be identical. On the other hand, the linear subspace $U_k(I)$ is spanned by $\mathbf{c}_1(I), \ldots, \mathbf{c}_K(I)$ which converge to $\mathbf{c}_1, \ldots, \mathbf{c}_K$ as $I \to \infty$; hence, any converging sequence in $U_k(I)$ must converge to some vector in U_k, the linear subspace spanned by $\mathbf{c}_1, \ldots, \mathbf{c}_K$. It follows that $\mathbf{v}_k$, being the limit of a sequence in $\mathbb{R}^L \otimes U_k(I)$, must belong to $\mathbb{R}^L \otimes U_k$. Since the $\mathbf{v}_1, \ldots, \mathbf{v}_K$ spanned the same linear subspace as $\mathbf{b}_k \otimes \mathbf{c}_k$, $k = 1, \ldots, K$, the latter must be in $\mathbb{R}^L \otimes U_k$, and hence, $\mathbf{c}_k$, $k = 1, \ldots, K$, must be in U_K. But the above vectors, by assumption, are linearly independent, implying that there is a k_0 in $1, \ldots, K$ for which $\mathbf{c}_{k_0}$ is not in U_{K-1}, since this linear subspace has dimension $K - 1$. Note that k_0 is necessarily greater than K' since we have assumed $\mathbf{c}_k(I) \to \mathbf{c}_k$, for $k \le K'$. Let λ'_k and λ''_k denote the coordinates of $\mathbf{c}_{k_0}$ and $\mathbf{b}_{k_0} \otimes \mathbf{c}_{k_0}$ with respect to the basis $\mathbf{c}_1^*, \ldots, \mathbf{c}_K^*$ of U_K and $\mathbf{v}_1, \ldots, \mathbf{v}_K$ of V_K, respectively. Then, $\mathbf{b}_{k_0} \otimes \mathbf{c}_{k_0}$ can be expressed alternatively as

$$\sum_{k=1}^{K} \lambda'_k \mathbf{b}_{k_0} \otimes \mathbf{c}_k^* = \sum_{k=1}^{K} \lambda''_k(\nu_k \mathbf{b}_{j_k} \otimes \mathbf{c}_k^* + \mathbf{z}_{k-1}).$$

From the above equality and the fact that $\mathbf{z}_k \in \mathbb{R}^L \otimes U_k$, it is seen that for fixed l, the K-th coordinate of the vector $B_{k_0 l}\mathbf{c}_{k_0}$ with respect to the basis $\mathbf{c}_1^*, \ldots, \mathbf{c}_K^*$, is equal to $\lambda'_K B_{k_0 l}$ and also to $\lambda''_K \nu_K B_{j_K l}$. Since $\lambda'_k \neq 0$ by the choice of k_0, this implies that $\mathbf{b}_{k_0}$ must be a multiple of $\mathbf{b}_{j_K}$. This contradicts (A2) because $k_0 > K' \ge j_K$. The proof of the Lemma is completed. □

Theorem 2. Under the assumptions (M0), (A0′) and (A2), for I sufficiently large the least squares estimators $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ of $\mathbf{B}$ and $\mathbf{C}$ exist and converge almost surely to their true values as $I \to \infty$, up to scaling and permutation of their rows.

Proof. From the above lemmas and corollary, for I sufficiently large, there exists $\hat{\mathbf{H}}$ in Θ such that $\mathbf{P}_{\hat{\mathbf{H}}}$ minimizes $\operatorname{tr}(\mathbf{I} - \mathbf{P}_{\mathbf{H}})_{LT}\mathbf{X}\mathbf{X}'_{LT}]$ in Θ and $d(\hat{\mathbf{H}}, \mathbf{B} * \mathbf{c}) \to 0$ as $I \to \infty$, almost surely. By definition, $\hat{\mathbf{H}}$ is of the form $\hat{\mathbf{B}} \otimes \hat{\mathbf{C}}$, and $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ are precisely the least squares estimators of $\mathbf{B}$ and $\mathbf{C}$. Let $\hat{\mathbf{h}}_1, \ldots, \hat{\mathbf{h}}_K$ be the row vectors $\hat{\mathbf{H}}$ and denote by r_k (which depends on I) the index r for which $\mathbf{h}_r$ is closest to $\mathbf{b}_k \otimes \mathbf{c}_k$. Then, from $d_2(\hat{\mathbf{H}}, \mathbf{B} * \mathbf{C}) \to 0$, $\hat{\mathbf{h}}_{r_k} \to \mathbf{b}_k \otimes \mathbf{c}_k$ and from $d_2(\hat{\mathbf{H}}, \mathbf{B} * \mathbf{C}) \to 0$, the r_k must be distinct for I large enough, meaning that $\{r_1, \ldots, r_K\}$ is a permutation of $\{1, \ldots, K\}$. On the other hand, $\hat{\mathbf{h}}_{r_k} \to \mathbf{b}_k \otimes \mathbf{c}_k$ implies $\hat{\mathbf{b}}_{r_k} \to \mathbf{b}_k$ and $\hat{\mathbf{c}}_{r_k} \to \mathbf{c}_k$ by the same argument as at the beginning of the proof of Lemma 3. □

4. Asymptotic Normality of the Least Squares Estimator

To eliminate scale factors, we have normalized the rows of our estimators $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ to have unit norm. The result is that the elements of these matrices are not functionally

independent. The set of all possible values of them is a manifold in $\mathbb{R}^{KL} \times \mathbb{R}^{KT} = \mathbb{R}^{K(L+T)}$. Any sufficiently small open neighborhood of a point in it can be mapped "smoothly" to some open set in $\mathbb{R}^D$, the integer D denoting the number of functionally independent coordinates. Here $D = K(L + T - 2)$ since each row vector of $\hat{\mathbf{B}}$ or $\hat{\mathbf{C}}$ can be specified by $L - 1$ or $T - 1$ of its coordinates. These coordinates may be taken as the first ones, but since they do not contain information on the sign of the last coordinate, it is necessary to restrict oneself to a small neighborhood of the point of interest where the unspecified coordinates keep a constant sign (one may have to choose the unspecified coordinate other than the last one to satisfy the above condition). In this section we shall restrict ourselves to a small open neighborhood of the *true* parameter point, since our estimators, being consistent, will eventually enter it. This neighborhood is then mapped one-to-one to an open subset of $\mathbb{R}^D$, denoted by Θ (note that the definition of Θ has been somewhat changed with respect to that of section 3). Thus, a point $\boldsymbol{\theta}$ in Θ corresponds to matrices $\mathbf{B} = \mathbf{B}(\theta)$ and $\mathbf{C} = \mathbf{C}(\theta)$ and the criterion Q^* becomes a function of θ: $Q^* = \text{tr}\,[(\mathbf{I} - \mathbf{P}_{\mathbf{B}(\theta) * \mathbf{C}(\theta)})_{LT}\mathbf{X}\mathbf{X}'_{LT}]$ (note that at this stage, θ, $\mathbf{B}$ and $\mathbf{C}$ denote free parameters). The maps $\boldsymbol{\theta} \mapsto \mathbf{B}$ and $\boldsymbol{\theta} \mapsto \mathbf{C}$ can be assumed to be twice continuously differentiable, and thus, Q^* is twice continuously differentiable with respect to $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}$ denotes the least squares estimator of θ. It is the solution of the equations $(\partial Q^*/\partial\theta_r)(\hat{\theta}) = 0$, $r = 1, \ldots, D$, where the notation $\partial Q^*/\partial\theta_r$ denotes the partial derivative of Q^* with respect to the component θ_r of $\boldsymbol{\theta}$ (the argument $\hat{\boldsymbol{\theta}}$ means that it is evaluated at this point). Then a Taylor development of $(\partial Q^*/\partial\theta_r)(\hat{\boldsymbol{\theta}})$ around the *true* parameter point $\boldsymbol{\theta}$ yields

$$0 = \frac{\partial Q^*}{\partial\theta_r}(\hat{\theta}) = \frac{\partial Q^*}{\partial\theta_r} + \sum_{s=1}^{D} \frac{\partial^2 Q^*}{\partial\theta_r \partial\theta_s}(\bar{\boldsymbol{\theta}})(\hat{\theta}_s - \theta_s),$$

where $\partial^2 Q^*/\partial\theta_r\partial\theta_s$ denotes the partial second derivative of Q^* with respect to θ_r and θ_s, and $\bar{\boldsymbol{\theta}}$ is a point lying on the segment joining $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ (here, $\boldsymbol{\theta}$ denotes the true value and $\partial Q^*/\partial\theta_r$ denotes the derivative evaluated at the true value; a similar convention is used for the second derivative). Since $\bar{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}$, it can be shown that the difference between $\partial^2 Q^*/\partial\theta_r\partial\theta_s)(\bar{\theta})/I$ and $(\partial^2 Q^*/\partial\theta_r\partial\theta_s)/I$ tends to zeros as $I \to \infty$. Thus, one has

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \left[\frac{1}{I}\frac{\partial^2 Q^*}{\partial\boldsymbol{\theta}^2} + o(1)\right]^{-1}\frac{1}{I}\frac{\partial Q^*}{\partial\boldsymbol{\theta}}, \tag{5}$$

where $o(1)$ denotes a term tending to 0 almost surely as $I \to \infty$, and $\partial Q^*/\partial\boldsymbol{\theta}$ and $\partial^2 Q^*/\partial\boldsymbol{\theta}^2$ denote the vector with components $\partial Q^*/\partial\theta_r$ and the matrix with elements $\partial^2 Q^*/\partial\theta_r\partial\theta_s$, respectively.

Computing the derivative in (5),

$$\frac{\partial}{\partial\theta_r}\mathbf{P}_{\mathbf{B}*\mathbf{C}} = (\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}})\left[\frac{\partial}{\partial\theta_r}(\mathbf{B} * \mathbf{C})'\right]\mathbf{S}_{bc}^{-1}(\mathbf{B} * \mathbf{C}) + (\mathbf{B} * \mathbf{C})'\mathbf{S}_{bc}^{-1}\left[\frac{\partial}{\partial\theta_r}(\mathbf{B} * \mathbf{C})\right](\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}}). \tag{6}$$

Hence, from the identity $\text{tr}\,(\mathbf{AB}) = \text{tr}\,(\mathbf{BA})$,

$$\frac{1}{I}\frac{\partial Q^*}{\partial\theta_r} = -\frac{2}{I}\,\text{tr}\left\{\left[\frac{\partial}{\partial\theta_r}(\mathbf{B} * \mathbf{C})'\right]\mathbf{S}_{bc}^{-1}(\mathbf{B} * \mathbf{C})_{LT}\mathbf{X}\mathbf{X}'_{LT}(\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}})\right\}. \tag{7}$$

The second derivatives of Q^* are more complicated. But they simplify when taking the limit as $I \to \infty$. Indeed, by Lemma 1, ${}_{LT}\mathbf{X}\mathbf{X}'_{LT}/I$ tends to $\mathbf{\Lambda} = (\mathbf{B} * \mathbf{C})'\mathbf{R}_a(\mathbf{B} * \mathbf{C}) + \sigma^2\mathbf{I}$. Thus, the limit of $(\partial^2 Q^*/\partial\theta_r\,\partial\theta_s)/I$ can be obtained by differentiating (7) and replacing ${}_{LT}\mathbf{X}\mathbf{X}'_{LT}/I$ by $\mathbf{\Lambda}$. Noting that $(\mathbf{B} * \mathbf{C})\mathbf{\Lambda}(\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}}) = \mathbf{0}$, and from (6),

$$\left[\frac{\partial}{\partial\theta_s}(\mathbf{B} * \mathbf{C})\right]\mathbf{\Lambda}(\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}}) - (\mathbf{B} * \mathbf{C})\mathbf{\Lambda}\frac{\partial}{\partial\theta_s}\mathbf{P}_{\mathbf{B}*\mathbf{C}}$$

$$= \left\{\left[\frac{\partial}{\partial\theta_s}(\mathbf{B} * \mathbf{C})\right]\sigma^2 - [\mathbf{S}_{bc}\mathbf{R}_a + \sigma^2\mathbf{I}]\left[\frac{\partial}{\partial\theta_s}(\mathbf{B} * \mathbf{C})\right]\right\}(\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}})$$

$$= \mathbf{S}_{bc}\mathbf{R}_a\left[\frac{\partial}{\partial\theta_s}(\mathbf{B} * \mathbf{C})\right](\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}}),$$

one obtains,

$$\frac{1}{I}\frac{\partial^2 Q^*}{\partial\theta_r\,\partial\theta_s} \to 2\,\mathrm{tr}\left\{\left[\frac{\partial}{\partial\theta_r}(\mathbf{B} * \mathbf{C})'\right]\mathbf{R}_a\left[\frac{\partial}{\partial\theta_s}(\mathbf{B} * \mathbf{C})\right](\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}})\right\} = 2W_{rs}, \text{ say.} \tag{8}$$

We now derive the asymptotic distribution of $(\partial Q^*/\partial\theta_r)/\sqrt{I}$.

Lemma 4. Under the assumption assumptions (M0), (M1) and suppose that the E_{ilt} have zero third and fourth cumulant, the random variables $(\partial Q^*/\partial\theta_r)/(2\sqrt{I})$ is asymptotically normal with mean zero and covariance matrix $\mathbf{V}$ with general element

$$V_{rs} = \sigma^2\,\mathrm{tr}\left\{(\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}})\left[\frac{\partial}{\partial\theta_r}(\mathbf{B} * \mathbf{C})'\right](\mathbf{R}_a + \mathbf{S}_{bc}^{-1}\sigma^2)\left[\frac{\partial}{\partial\theta_s}(\mathbf{B} * \mathbf{C})\right]\right\}.$$

It is noted that the above condition of third and fourth cumulants of E_{ilt} being zero is not indispensable, but it allows simpler expression for the asymptotic covariance matrix. As can be seen in the proof below, the asymptotic normality of $(\partial Q^*/\partial\theta_r)/\sqrt{I}$ only requires that E_{jlt} has finite fourth (hence third) cumulant.

Proof. Letting $\mathbf{M}_r = \partial(\mathbf{B} * \mathbf{C})'/\partial\theta_r$, we have from (7) and (2b)

$$\frac{1}{2\sqrt{I}}\frac{\partial Q^*}{\partial\theta_r} = -\frac{1}{\sqrt{I}}\{\mathrm{tr}\,[(\mathbf{M}_r\mathbf{A}_I\mathbf{E}'_{LT})(\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}})]$$

$$+ \mathrm{tr}\,[\mathbf{M}_r\mathbf{S}_{bc}^{-1}(\mathbf{B} * \mathbf{C})_{LT}\mathbf{E}\mathbf{E}'_{LT}(\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}})]\}$$

$$= \frac{1}{\sqrt{I}}\sum_{i=1}^{K}[\mathbf{e}'_i\mathbf{m}_{r,i} + \mathbf{e}'_i(\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}})(\mathbf{B} * \mathbf{C})\mathbf{S}_{bc}^{-1}\mathbf{M}_r\mathbf{e}_i],$$

where $\mathbf{e}_i$ and $\mathbf{m}_{r,i}$ denote the i-th column of ${}_{LT}\mathbf{E}_I$ and of $(\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}})\mathbf{M}_r\mathbf{A}$, respectively. To obtain the joint asymptotic normality of the above random variables, a simple method is to show that any linear combination of them is asymptotically normal. Such a linear combination can be written as $\sum_{i=1}^{K}(Y_i + Z_i)/\sqrt{I}$ where $Y_i = \sum_{r=1}^{D}\alpha_r\mathbf{e}'_i\mathbf{m}_{r,i}$, $Z_i = \sum_{r=1}^{D}\alpha_r\mathbf{e}'_i(\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}})\mathbf{M}_r\mathbf{S}_{bc}^{-1}(\mathbf{B} * \mathbf{C})\mathbf{e}_i$ and α_r are given coefficients. Assume for the moment that the A_{ki}, $i = 1, 2, \ldots$ are deterministic sequences. Then, the random variables $Y_i + Z_i$ are independent having zero mean (since tr [(I −

$P_{\mathbf{B}*\mathbf{C}})\mathbf{M}_r\mathbf{S}_{bc}^{-1}(\mathbf{B}*\mathbf{C})] = 0$) and finite variance, and the Central Limit Theorem applies. A sufficient condition for this Theorem is the Lindeberg's condition (e.g., Loève, 1963, p. 280). It can be checked that this condition is satisfied if $\max_{k=1}^{I} \|\mathbf{m}_{r,i}\|^2/I \to 0$, for all r, which from the definition of $\mathbf{m}_{r,i}$, is implied by (M1). Let us now compute the variance of $\sum_{i=1}^{K}(Y_i + Z_i)$. Since the third cumulant of E_{ilt} is zero, Y_i and Z_i are uncorrelated, and hence, this variance is the sum of the variances of Y_i and Z_i. We have

$$\sum_{i=1}^{K} \text{var}\,(Y_i) = \sigma^2 \sum_{r=1}^{D}\sum_{s=1}^{D}\sum_{i=1}^{K} \alpha_r\alpha_s\mathbf{m}_{r,i}\mathbf{m}'_{s,i}$$

$$= \sigma^2 \sum_{r=1}^{D}\sum_{s=1}^{D} \alpha_r\alpha_s \text{ tr } [(\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}})\mathbf{M}_r\mathbf{A}\mathbf{A}'\mathbf{M}_s],$$

$$\sum_{i=1}^{K} \text{var}\,(Z_i) = K\sigma^4 \sum_{r=1}^{D}\sum_{s=1}^{D} \alpha_r\alpha_s \text{ tr } [(\mathbf{I} - \mathbf{P}_{\mathbf{B}*\mathbf{C}})\mathbf{M}_r\mathbf{S}_{bc}^{-1}\mathbf{M}_s].$$

The last equality follows from the fact that the E_{ilt} are independent with mean zero, variance σ^2 and fourth cumulant zero; hence, cov $(E_{ilt}E_{il't'}, E_{i\lambda\tau}E_{i\lambda'\tau'}) = 0$ unless $(1, t) = (\lambda, \tau)$, $(1', t') = (\lambda', \tau')$ or $(1, t) = (\lambda', \tau')$, $(1', t') = (\lambda, \tau)$, in which case it equals σ^4, or $(1, t) = (\lambda, \tau) = (1', t') = (\lambda', \tau')$, in which case it equals $2\sigma^4$, which yields the formula: cov $\{\text{tr } (\mathbf{e}_i'\mathbf{F}\mathbf{e}_i), \text{ tr } (\mathbf{e}_i'\mathbf{G}\mathbf{e}_i)\} = \sigma^4 \text{ tr } (\mathbf{F}\mathbf{G}' + \mathbf{F}\mathbf{G})$ for any matrices $\mathbf{F}$, $\mathbf{G}$. From (M0) and the above computations, the asymptotic variance of $\sum_{i=1}^{K}(Y_i + Z_i)/\sqrt{I}$ is precisely $\sum_{r=1}^{D}\sum_{s=1}^{D} \alpha_r\,\alpha_s\,V_{rs}$ and the result follows. In the case where the A_{ki} are random, one considers the conditional distribution given these random variables. Since the limiting conditional distribution does not depend on the distribution of the A_{ki} (provided that (M0) holds), it is the same as the limiting unconditional distribution. □

From (5), (8), and the above lemma, we obtain

Theorem 3. Under the assumptions of Theorem 2 and Lemma 4, the least squares estimator $\hat{\theta}$ of θ is asymptotically normal with mean $\boldsymbol{\theta}$ and covariance matrix $\mathbf{W}^{-1}\mathbf{V}\mathbf{W}^{-1}/I$, where $\mathbf{W}$ is defined by its general element given in (8), and $\mathbf{V}$ denotes the matrix of Lemma 4.

In practice, σ^2 is often small and $\mathbf{S}_{bc}^{-1}\sigma^2$ may be negligible with respect to $\mathbf{R}_a$. If this term is neglected, $\mathbf{V}$ reduces to $2\mathbf{W}$, and the asymptotic covariance matrix of the estimator then simply equals $(2/I)\mathbf{W}^{-1}$.

References

Carrol, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika, 35*, 283–319.

Harshman, R. A. (1970). Foundation of the PARAFAC procedure: Models and conditions for an "explanatory" multi-mode factor analysis. *UCLA Working Papers in Phonetics, 16*, 1–84.

Harshman, R. A., & Lundy, M. E. (1984). The PARAFAC model for the three-way factor analysis and multidimensional scaling. In H. G. Law, C. W. Snyder, J. A. Hattie, & R. P. McDonald (Eds.), *Research methods for multimode data analysis* (pp. 123–215). New York: Praeger.

Kruskal, J. B. (1976). More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika, 41*, 281–293.

Kruskal, J. B. (1977). Three-way arrays: Rank and uniqueness of trilinear decomposition with application to arithmetic complexity and statistics. *Linear algebra and its applications*, *18*, 95–138.

Kruskal, J. B. (1984). Multilinear methods. In H. G. Law, C. W. Snyder, J. A. Hattie, & R. P. McDonald (Eds.), *Research methods for multimode data analysis* (pp. 36–62). New York: Praeger.

Loève, M. (1963). *Probability theory*. New York: Van Nostrand.

Möcks, J. (1988a). Topographical components model for event-related potentials and some biophysical considerations. *IEEE Transactions on Biomedical Engineering*, *35*, 482–484.

Möcks, J. (1988b). Decomposing event-related potentials: A new topographic components model. *Biological Psychology*, *26*, 129–215.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

Tucker, L. R. (1966). Some mathematical notes on the three modes factor analysis. *Psychometrika*, *31*, 279–311.

Manuscript received 10/17/89
Final version received 5/13/91