Principal component analysis from multivariate familial correlation matrix[1]

Martin BILODEAU
Université de Montréal
AND
Pierre DUCHESNE
École des Hautes Études Commerciales

Key words and phrases: familial model, principal components, correlation matrix, elliptical distributions.

ABSTRACT

This paper considers principal component analysis (PCA) in familial models, where the number of siblings can differ among families. Konishi and Rao (1992) used the unified estimator of Konishi and Khatri (1990) to develop a PCA derived from the covariance matrix. However, because of the lack of invariance to componentwise change of scale, an analysis based on the correlation matrix is often preferred. The asymptotic distribution of the estimated eigenvalues and eigenvectors of the correlation matrix are derived under elliptical sampling. A Monte Carlo simulation shows the usefulness of the asymptotic expressions for samples as small as $N = 25$ families.

1. INTRODUCTION

Principal component analysis (PCA) in familial models is complicated by the lack of independence between offspring of the same family. Besides the obvious correlations among variables observed on an individual, correlations between variables observed on two members of the same family may also exist. Such data arises, for example, in a genetics study on the degree of resemblance between members of the same family. Konishi and Rao (1992), using the unified estimator of Konishi and Khatri (1990), developed a PCA derived from the covariance matrix. Since PCA on the covariance matrix is not invariant to a componentwise change of scale, many users prefer doing PCA on the correlation matrix. This paper proposes a PCA derived from the correlation matrix.

The familial model considered is as follows. Suppose we have a random sample of N families on a p-variate random vector $\mathbf{x} = (x_1, \ldots, x_p)'$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let

$$\mathbf{z}_\alpha = (\mathbf{x}'_{1\alpha}, \ldots, \mathbf{x}'_{k_\alpha,\alpha})', \ \alpha = 1, \ldots, N \tag{1}$$

denote the measurements on the αth family with $k_\alpha \geq 1$ members, where $\mathbf{x}_{j\alpha} = (x_{1j,\alpha}, \ldots, x_{pj,\alpha})'$ is the vector on p variables for the jth member of the αth family. Independence holds among families so that $\mathbf{z}_1, \ldots, \mathbf{z}_N$ are assumed mutually

[1]This research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

independent, each having a pk_α-variate distribution with mean vector $(\boldsymbol{\mu}', \ldots, \boldsymbol{\mu}')'$ and covariance matrix

$$\mathbf{I}_{k_\alpha} \otimes \boldsymbol{\Sigma} + \left(\mathbf{1}_{k_\alpha}\mathbf{1}_{k_\alpha}' - \mathbf{I}_{k_\alpha}\right) \otimes \boldsymbol{\Sigma}_s,$$

where $\mathbf{I}_{k_\alpha}$ is the identity matrix of order k_α, $\mathbf{1}_{k_\alpha}$ is the k_α-vector of unit elements, and $\otimes$ denotes Kronecker product. The matrix $\boldsymbol{\Sigma} = (\sigma_{ij}) : p \times p$ comprises the covariances among p variables on a given member, whereas $\boldsymbol{\Sigma}_s : p \times p$ reflects the covariances of the p variables among two members of the same family.

Let $\boldsymbol{\Sigma}_0 = \mathrm{diag}(\boldsymbol{\Sigma}) = \mathrm{diag}(\sigma_{11}, \ldots, \sigma_{pp})$. The subindex 0 will play this diagonal role on various matrices throughout the paper. The correlation matrix between the p variables is $\mathbf{R} = (\rho_{ij}) = \boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1/2}$. Let $\lambda_1 \geq \ldots \geq \lambda_p$ be the ordered eigenvalues of $\mathbf{R}$, with corresponding normalized eigenvectors $\boldsymbol{\gamma}_i = (\gamma_{1i}, \ldots, \gamma_{pi})'$ satisfying $|\boldsymbol{\gamma}_i| = 1$ and $\gamma_{ii} \geq 0$. The ith principal component based on the correlation matrix $\mathbf{R}$ is $y_i = \boldsymbol{\gamma}_i'\boldsymbol{\Sigma}_0^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ and its variance is λ_i. This paper gives the large sample distribution of the estimators, $\hat{\lambda}_i$ and $\hat{\boldsymbol{\gamma}}_i$, of eigenvalues and eigenvectors based on the unified estimator. Konishi and Rao (1992) developed the PCA from the unified estimator of the covariance matrix $\boldsymbol{\Sigma}$ rather than the correlation matrix $\mathbf{R}$.

2. PCA BASED ON THE UNIFIED ESTIMATOR

2.1 Asymptotic distribution of the eigenvalues, eigenvectors and of a statistic for reduction of dimensionality

The unified estimator of Konishi and Khatri (1990) is as follows. Let

$$\bar{\mathbf{X}} = (\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_N) : p \times N, \ \mathbf{S}_\alpha = \sum_{j=1}^{k_\alpha} (\mathbf{x}_{j\alpha} - \bar{\mathbf{x}}_\alpha)(\mathbf{x}_{j\alpha} - \bar{\mathbf{x}}_\alpha)' : p \times p,$$

where $\bar{\mathbf{x}}_\alpha = \sum_{j=1}^{k_\alpha} \mathbf{x}_{j\alpha}/k_\alpha$, $\alpha = 1, \ldots, N$. Let $\mathbf{B} : N \times N$ be a positive semidefinite matrix such that $\mathbf{B}\mathbf{1}_N = \mathbf{0}$. The unified estimator of $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}} = \{\mathrm{tr}(\mathbf{B})\}^{-1}(\bar{\mathbf{X}}\mathbf{B}\bar{\mathbf{X}}' + \sum_\alpha \omega_\alpha \mathbf{S}_\alpha), \tag{2}$$

where $\omega_1, \ldots, \omega_N$ are non-negative weights. If $\mathbf{D}_N = \mathrm{diag}(k_1, \ldots, k_N)$ then the expectation of $\hat{\boldsymbol{\Sigma}}$ is

$$E(\hat{\boldsymbol{\Sigma}}) = \boldsymbol{\Sigma} + \{\mathrm{tr}(\mathbf{B})\}^{-1}\left[\sum_\alpha \omega_\alpha(k_\alpha - 1) - \mathrm{tr}\{\mathbf{B}(\mathbf{I}_N - \mathbf{D}_N^{-1})\}\right](\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_s).$$

Asymptotic derivations require the following hypotheses on the weights $\mathbf{B}$ and ω_α:

H1: The αth diagonal element of $\mathbf{B}$ converges to b_α as $N \to \infty$, and the off-diagonal elements are $O(1/N)$.
H2: If the weights ω_α depend on N, then the ω_α's are regarded as their finite limiting values.
H3: The bias term $\{\mathrm{tr}(\mathbf{B})\}^{-1}\left[\sum_\alpha \omega_\alpha(k_\alpha - 1) - \mathrm{tr}\{\mathbf{B}(\mathbf{I}_N - \mathbf{D}_N^{-1})\}\right](\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_s)$ is $O(1/N)$.

Under conditions H1 to H3, Konishi and Rao (1992) established that $\hat{\mathbf{\Sigma}}$ converges in probability to $\mathbf{\Sigma}$ and that $\mathbf{V} = (v_{ij}) = N^{1/2}\mathbf{\Sigma}_0^{-1/2}(\hat{\mathbf{\Sigma}}-\mathbf{\Sigma})\mathbf{\Sigma}_0^{-1/2}$ converges to a normal distribution as N goes to infinity.

Several estimators in the literature fall within this class. For example, the estimator of Srivastava, Keen and Katapa (1988) (called method S) is obtained in setting $\mathbf{B} = \mathbf{I}_N - \mathbf{1}_N\mathbf{1}_N'/N$ and $\omega_\alpha = (N-1)N^{-1}\sum k_\alpha^{-1}(k_\alpha-1)/\sum(k_\alpha-1)$ and the components of variance model estimator of Konishi and Rao (1992) (called method R) is obtained with $\mathbf{B} = \mathbf{D}_N - \mathbf{K}_N\mathbf{K}_N'/\sum k_\alpha$ and $\omega_\alpha = (N-1)(N_0-1)/(\sum k_\alpha - N)$, where $\mathbf{K}_N = (k_1, \ldots, k_N)'$ and $N_0 = (\sum k_\alpha - \sum k_\alpha^2/\sum k_\alpha)/(N-1)$.

If we want to do a PCA from the correlation matrix, the unified estimator of $\mathbf{R}$ is then $\hat{\mathbf{R}} = \hat{\mathbf{\Sigma}}_0^{-1/2}\hat{\mathbf{\Sigma}}\hat{\mathbf{\Sigma}}_0^{-1/2}$. From Konishi (1979), using the expansion $x^{-1/2} = a^{-1/2} - (1/2)a^{-3/2}(x-a) + O[(x-a)^2]$, we obtain

$$\hat{\mathbf{R}} = \hat{\mathbf{\Sigma}}_0^{-1/2}\hat{\mathbf{\Sigma}}\hat{\mathbf{\Sigma}}_0^{-1/2} = \mathbf{R} + N^{-1/2}\left(\mathbf{V} - \frac{1}{2}\mathbf{V}_0\mathbf{R} - \frac{1}{2}\mathbf{R}\mathbf{V}_0\right) + O_p(N^{-1}).$$

Now, let $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_p) : p \times p$ be an orthogonal matrix such that $\mathbf{H}'\mathbf{R}\mathbf{H} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p) = \mathbf{\Lambda}$. Then, we have

$$N^{1/2}(\mathbf{H}'\hat{\mathbf{R}}\mathbf{H} - \mathbf{\Lambda}) = \mathbf{V}^{(1)} + O_p(N^{-1/2}), \tag{3}$$

where

$$\mathbf{V}^{(1)} = (v_{ij}^{(1)}) = \mathbf{H}'\left(\mathbf{V} - \frac{1}{2}\mathbf{V}_0\mathbf{R} - \frac{1}{2}\mathbf{R}\mathbf{V}_0\right)\mathbf{H}.$$

From the perturbation method [Bellman (1960), p. 61, see also Bilodeau (1999), Section 8.8], the jth eigenvalue $\hat{\lambda}_j$ of $\hat{\mathbf{R}}$ will have the same asymptotic distribution as the jth diagonal element of $\mathbf{V}^{(1)}$, where

$$\begin{aligned} v_{ij}^{(1)} &= \sum_k\sum_l h_{ki}h_{lj}v_{kl} - \frac{1}{2}(\lambda_i+\lambda_j)\sum_k h_{ki}h_{kj}v_{kk} \\ &= \sum_k\sum_l h_{ki}h_{lj}\left[1 - \frac{1}{2}(\lambda_i+\lambda_j)\delta_{kl}\right]v_{kl} \\ &= \sum_k\sum_l a_{ij}(k,l)v_{kl}, \end{aligned}$$

with

$$a_{ij}(k,l) = h_{ki}h_{lj}\left[1 - \frac{1}{2}(\lambda_i+\lambda_j)\delta_{kl}\right].$$

The derivation of the joint asymptotic distribution of $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ will require the evaluation of $\mathrm{cov}(v_{kl}, v_{k'l'})$.

Lemma: If the p-vector $\mathbf{y}$ has an elliptical distribution with mean vector $\mathbf{0}$, covariance matrix $\mathbf{\Sigma}$, and finite kurtosis 3κ, then for any $p \times p$ matrices $\mathbf{A}$ and $\mathbf{B}$

$$\mathrm{cov}(\mathbf{y}'\mathbf{A}\mathbf{y}, \mathbf{y}'\mathbf{B}\mathbf{y}) = 2(1+\kappa)\ \mathrm{tr}(\mathbf{A}\mathbf{\Sigma}\mathbf{B}\mathbf{\Sigma}) + \kappa\ \mathrm{tr}(\mathbf{A}\mathbf{\Sigma})\ \mathrm{tr}(\mathbf{B}\mathbf{\Sigma}).$$

Proof: We have that $E(\mathbf{y}'\mathbf{A}\mathbf{y}) = \mathrm{tr}(\mathbf{A}\mathbf{\Sigma})$ and using product-moments we also have $E\{(\mathbf{y}'\mathbf{A}\mathbf{y})(\mathbf{y}'\mathbf{B}\mathbf{y})\} = \sum_{ijkl} a_{ij}b_{kl}\mu_{1111}^{ijkl}$. From Muirhead (1982), all fourth-order multivariate cumulants of an elliptical distribution are of the form $k_{1111}^{ijkl} = \kappa(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk})$. Finally, using the general relation between product-moments and cumulants $\mu_{1111}^{ijkl} = k_{1111}^{ijkl} + k_{11}^{ij}k_{11}^{kl} + k_{11}^{ik}k_{11}^{jl} + k_{11}^{il}k_{11}^{jk}$ one gets, after

some algebra, the final result.

Since $\hat{\mathbf{R}}$ is invariant to componentwise relocation and rescaling, we can assume without loss of generality that $\mathbf{z}_\alpha$ has a pk_α-variate distribution with mean vector $\mathbf{0}$ and covariance matrix

$$\mathbf{I}_{k_\alpha} \otimes \mathbf{R} + \left(\mathbf{1}_{k_\alpha}\mathbf{1}'_{k_\alpha} - \mathbf{I}_{k_\alpha}\right) \otimes (\mathbf{\Sigma}_0^{-1/2}\mathbf{\Sigma}_s\mathbf{\Sigma}_0^{-1/2}).$$

Let $\mathbf{\Phi} = (\Phi_{ij}) = \mathbf{\Sigma}_0^{-1/2}\mathbf{\Sigma}_s\mathbf{\Sigma}_0^{-1/2}$ be the among offspring correlation matrix. Thus, we require the asymptotic covariance matrix of $N^{1/2}\hat{\mathbf{\Sigma}}$. However,

$$(\sum_\alpha b_\alpha)\hat{\mathbf{\Sigma}} = \sum_\alpha b_\alpha \bar{\mathbf{x}}_\alpha \bar{\mathbf{x}}'_\alpha + \sum_{\alpha\neq\beta} b_{\alpha\beta}\bar{\mathbf{x}}_\alpha\bar{\mathbf{x}}'_\beta + \sum_\alpha \omega_\alpha \mathbf{S}_\alpha,$$

where the term $(N^{1/2}/\sum_\alpha b_\alpha)\sum_{\alpha\neq\beta} b_{\alpha\beta}\bar{\mathbf{x}}_\alpha\bar{\mathbf{x}}'_\beta \to \mathbf{0}$ in probability because $E(\bar{\mathbf{x}}_\alpha) = \mathbf{0}$ and $\mathrm{cov}\left((N^{1/2}/\sum_\alpha b_\alpha)\sum_{\alpha\neq\beta} b_{\alpha\beta}\bar{\mathbf{x}}_\alpha\bar{\mathbf{x}}'_\beta\right) \to \mathbf{0}$. Thus, we consider only

$$\tilde{v}_{kl} = \frac{N^{1/2}}{(\sum_\alpha b_\alpha)}\sum_\alpha (b_\alpha \bar{\mathbf{x}}_\alpha^{(k)}\bar{\mathbf{x}}_\alpha^{(l)} + \omega_\alpha \mathbf{S}_\alpha^{(k,l)}).$$

In (1), $\mathbf{z}_\alpha$ was partitioned along offspring, however, it could be partitioned as well along variables so that we could define $\tilde{\mathbf{z}}_\alpha$, after a suitable permutation, with mean vector $\mathbf{0}$ and covariance matrix

$$\mathbf{R} \otimes \mathbf{I}_{k_\alpha} + \mathbf{\Phi} \otimes \left(\mathbf{1}_{k_\alpha}\mathbf{1}'_{k_\alpha} - \mathbf{I}_{k_\alpha}\right) \equiv \mathbf{\Omega}.$$

With this notation, then

$$\tilde{v}_{kl} = \frac{N^{1/2}}{(\sum_\alpha b_\alpha)}\sum_\alpha \tilde{\mathbf{z}}'_\alpha(\mathbf{e}_k\mathbf{e}'_l \otimes \mathbf{T}_\alpha)\tilde{\mathbf{z}}_\alpha = \frac{N^{1/2}}{(\sum_\alpha b_\alpha)}\sum_\alpha \tilde{\mathbf{z}}'_\alpha \mathbf{A}_{kl}\tilde{\mathbf{z}}_\alpha,$$

where $\mathbf{e}_k$ is a p-vector with a unit element in position k and 0 elsewhere, $\mathbf{M} = \mathbf{1}_{k_\alpha}\mathbf{1}'_{k_\alpha}/k_\alpha$ is idempotent of rank one, $\mathbf{T}_\alpha = (b_\alpha/k_\alpha)\mathbf{M} + \omega_\alpha(\mathbf{I}_{k_\alpha} - \mathbf{M})$, and $\mathbf{A}_{kl} = \mathbf{e}_k\mathbf{e}'_l \otimes \mathbf{T}_\alpha$. Using the lemma, we thus find

$$\mathrm{cov}(\tilde{v}_{kl}, \tilde{v}_{k'l'}) = \frac{N}{(\sum_\alpha b_\alpha)^2}\sum_\alpha [2(1+\kappa)\mathrm{tr}(\mathbf{A}_{kl}\mathbf{\Omega}\mathbf{A}_{k'l'}\mathbf{\Omega}) + \kappa\ \mathrm{tr}(\mathbf{A}_{kl}\mathbf{\Omega})\mathrm{tr}(\mathbf{A}_{k'l'}\mathbf{\Omega})].$$

Algebraic manipulations then yield

$$\begin{aligned}
\mathrm{tr}(\mathbf{A}_{kl}\mathbf{\Omega}) &= \rho_{kl}\left[\frac{b_\alpha}{k_\alpha} + \omega_\alpha(k_\alpha - 1)\right] + \Phi_{kl}\left(\frac{b_\alpha}{k_\alpha} - \omega_\alpha\right)(k_\alpha - 1) \\
&\equiv c_\alpha(\rho_{kl}, \Phi_{kl}), \\
\mathrm{tr}(\mathbf{A}_{kl}\mathbf{\Omega}\mathbf{A}_{k'l'}\mathbf{\Omega}) &= \rho_{k'l}\rho_{kl'}\left[\left(\frac{b_\alpha}{k_\alpha}\right)^2 + \omega_\alpha^2(k_\alpha - 1)\right] \\
&\quad + (\Phi_{k'l}\rho_{kl'} + \rho_{k'l}\Phi_{kl'})\left[\left(\frac{b_\alpha}{k_\alpha}\right)^2 - \omega_\alpha^2\right](k_\alpha - 1) \\
&\quad + \Phi_{k'l}\Phi_{kl'}\left[\left(\frac{b_\alpha}{k_\alpha}\right)^2(k_\alpha - 1) + \omega_\alpha^2\right](k_\alpha - 1) \\
&\equiv d_\alpha(\rho_{k'l}, \rho_{kl'}, \Phi_{k'l}, \Phi_{kl'}).
\end{aligned}$$

Thus, letting

$$\begin{aligned} C(k,l,k',l') &= \sum_\alpha c_\alpha(\rho_{kl},\Phi_{kl})\; c_\alpha(\rho_{k'l'},\Phi_{k'l'}) \\ D(k,l,k',l') &= \sum_\alpha d_\alpha(\rho_{k'l},\rho_{kl'},\Phi_{k'l},\Phi_{kl'}), \end{aligned}$$

$$\mathrm{cov}(\tilde{v}_{kl},\tilde{v}_{k'l'}) = \frac{N}{(\sum_\alpha b_\alpha)^2}\left[2(1+\kappa)D(k,l,k',l') + \kappa\; C(k,l,k',l')\right].$$

For the asymptotic distribution of eigenvalues and eigenvectors the required covariance becomes

$$\begin{aligned} \mathrm{cov}(\tilde{v}^{(1)}_{ij},\tilde{v}^{(1)}_{i'j'}) &= \frac{N}{(\sum_\alpha b_\alpha)^2}\sum_{k,l,k',l'} a_{ij}(k,l)a_{i'j'}(k',l')[2(1+\kappa) \\ & \quad D(k,l,k',l') + \kappa\; C(k,l,k',l')]. \end{aligned}$$

We can now state the two theorems.

Theorem 1: Let $\hat{\lambda}_1 > \ldots > \hat{\lambda}_p$ be the ordered eigenvalues of $\hat{\mathbf{R}}$ in (2) constructed from a sample of N independent families drawn from an elliptical distribution with finite kurtosis 3κ. Let $\mathbf{H} : p \times p$ be orthogonal such that $\mathbf{H}'\mathbf{R}\mathbf{H} = \mathrm{diag}(\lambda_1,\ldots,\lambda_p)$, where $\lambda_1 \geq \ldots \geq \lambda_p$. If λ_i and λ_j, $i \neq j$, are both eigenvalues of multiplicity one, then $N^{1/2}(\hat{\lambda}_i - \lambda_i, \hat{\lambda}_j - \lambda_j)$ is asymptotically normally distributed with mean vector $\mathbf{0}$ and covariance matrix

$$\begin{aligned} \mathrm{var}\left\{N^{1/2}(\hat{\lambda}_j - \lambda_j)\right\} &= \frac{N}{(\sum_\alpha b_\alpha)^2}\sum_{k,l,k',l'} a_{jj}(k,l)a_{jj}(k',l')[2(1+\kappa) \\ & \quad D(k,l,k',l') + \kappa\; C(k,l,k',l')] \\ \mathrm{cov}\left\{N^{1/2}(\hat{\lambda}_i - \lambda_i, \hat{\lambda}_j - \lambda_j)\right\} &= \frac{N}{(\sum_\alpha b_\alpha)^2}\sum_{k,l,k',l'} a_{ii}(k,l)a_{jj}(k',l')[2(1+\kappa) \\ & \quad D(k,l,k',l') + \kappa\; C(k,l,k',l')]. \end{aligned}$$

Examples: Suppose that we want to perform a PCA based on $\hat{\mathbf{R}}$ on the basis of a random sample of N families from a multinormal distribution. For the R method, taking $b_\alpha = k_\alpha$ and $\omega_\alpha = 1$ we have

$$D(k,l,k',l') = \rho_{kl'}\rho_{k'l}\sum_\alpha k_\alpha + \Phi_{k'l}\Phi_{kl'}(\sum_\alpha k_\alpha^2 - \sum_\alpha k_\alpha). \tag{4}$$

For the S method, taking $b_\alpha = 1$ and $\omega_\alpha \equiv \omega = \sum k_\alpha^{-1}(k_\alpha - 1)/\sum(k_\alpha - 1)$, we obtain

$$\begin{aligned} D(k,l,k',l') &= \sum_\alpha k_\alpha^{-2}[\rho_{kl'} + (k_\alpha - 1)\Phi_{kl'}][\rho_{k'l} + (k_\alpha - 1)\Phi_{k'l}] + \\ & \quad \omega^2(\rho_{kl'} - \Phi_{kl'})(\rho_{k'l} - \Phi_{k'l})(\sum_\alpha k_\alpha - N). \end{aligned} \tag{5}$$

These formulas provide useful simplifications for the computation of the asymptotic variances and covariances of the estimated eigenvalues using R and S methods.

The eigenvalues of $\mathbf{H}'\mathbf{R}\mathbf{H}$ are the same as those of $\mathbf{R}$. For the eigenvectors, however, we have the following relation: if $\boldsymbol{\gamma}_j$ is an eigenvector of $\mathbf{R}$ corresponding to λ_j then $\mathbf{H}'\boldsymbol{\gamma}_j$ is an eigenvector of $\mathbf{H}'\mathbf{R}\mathbf{H}$ corresponding to the same eigenvalue λ_j.

From (3) and the perturbation theory, we have the following expansion for the eigenvectors $\mathbf{f}_j \equiv \mathbf{H}'\hat{\boldsymbol{\gamma}}_j$ of $\mathbf{H}'\hat{\mathbf{R}}\mathbf{H}$,

$$\begin{aligned} f_{ij} &= -\lambda_{ij}\left[N^{-1/2}v_{ij}^{(1)} + O_p(N^{-1})\right],\ i \neq j, \\ f_{jj} &= 1 + O_p(N^{-1}), \end{aligned}$$

where $\lambda_{ij} = 1/(\lambda_i - \lambda_j)$. Hence, the vector $N^{1/2}(\mathbf{h}_1'\boldsymbol{\gamma}_j, \ldots, \mathbf{h}_j'\boldsymbol{\gamma}_j - 1, \ldots, \mathbf{h}_p'\boldsymbol{\gamma}_j)$ has the same asymptotic distribution as the vector $\mathbf{v}_j' \equiv (-\lambda_{1j}v_{1j}^{(1)}, \ldots, 0, \ldots, -\lambda_{pj}v_{pj}^{(1)})$. Premultiplying the latter by $\mathbf{H}$ we finally get that the asymptotic distribution of $N^{1/2}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j)$ is the same as the asymptotic distribution of $\mathbf{H}\mathbf{v}_j$.

Theorem 2: Let $\hat{\boldsymbol{\gamma}}_j = (\hat{\gamma}_{1j}, \ldots, \hat{\gamma}_{pj})$ be the normalized eigenvector corresponding to the eigenvalue $\hat{\lambda}_j$ of the unified estimator $\hat{\mathbf{R}}$ from an elliptical sample. Assume $\hat{\gamma}_{jj}$ and γ_{jj} are non-negative. If the eigenvalue λ_j is distinct from all other eigenvalues, then the asymptotic distribution of $N^{1/2}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j)$ is multivariate normal with mean vector $\mathbf{0}$ and singular covariance matrix $\mathbf{H}_j\boldsymbol{\Sigma}_j\mathbf{H}_j'$, where $\mathbf{H}_j = (\mathbf{h}_1, \ldots, \mathbf{h}_{j-1}, \mathbf{h}_{j+1}, \ldots, \mathbf{h}_p)$ and $\boldsymbol{\Sigma}_j = (\sigma_{rt,j}) : (p-1) \times (p-1)$ $(r, t \neq j)$ is a symmetric matrix given by

$$\begin{aligned} \sigma_{rr,j} &= \frac{1}{(\lambda_r - \lambda_j)^2}\frac{N}{(\sum_\alpha b_\alpha)^2}\sum_{k,l,k',l'} a_{rj}(k,l)a_{rj}(k',l')[2(1+\kappa) \\ &\quad D(k,l,k',l') + \kappa\, C(k,l,k',l')], \\ \sigma_{rt,j} &= \frac{1}{(\lambda_r - \lambda_j)}\frac{1}{(\lambda_t - \lambda_j)}\frac{N}{(\sum_\alpha b_\alpha)^2}\sum_{k,l,k',l'} a_{rj}(k,l)a_{tj}(k',l')[2(1+\kappa) \\ &\quad D(k,l,k',l') + \kappa\, C(k,l,k',l')],\ (r \neq t). \end{aligned}$$

The classical sampling of N observation vectors from a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be obtained by specifying $\kappa = 0$, and $b_\alpha = k_\alpha = 1$. The constants then reduce to $C(k,l,k',l') = D(k,l,k',l') = N\rho_{k'l}\rho_{kl'}$. Straightforward algebra (see Bilodeau (1999), Section 10.5) then establishes that

$$\mathrm{var}\left\{N^{1/2}(\hat{\lambda}_j - \lambda_j)\right\} = 2\lambda_j^2\left[1 - 2\lambda_j\sum_k h_{kj}^4 + \sum_{k,k'}\rho_{kk'}^2 h_{kj}^2 h_{k'j}^2\right].$$

This result was first derived by Konishi (1979). Theorem 2 on eigenvectors also generalizes a result of Konishi (1979).

A statistic often considered for reduction of dimensionality is $(\lambda_1 + \cdots + \lambda_q)/p$, where $q < p$. It represents the proportion of total variance of the standardized variables, $\boldsymbol{\Sigma}_0^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$, explained by the first q principal components. The asymptotic distribution of the statistic associated with this reduction of dimensionality parameter is a direct consequence of Theorem 1 coupled with the so-called delta method.

It suffices to define the differentiable function $g(\lambda_1, \ldots, \lambda_p) = (\lambda_1 + \cdots + \lambda_q)/p$ with partial derivatives $g_j = 1/p$, $j = 1, \ldots, q$, and $g_j = 0$, $j = q+1, \ldots, p$.

Corollary: Under the same hypotheses as in Theorem 1, if λ_j, $j = 1, \ldots, q$, are all eigenvalues of multiplicity one, then $N^{1/2}[(\hat{\lambda}_1 + \cdots + \hat{\lambda}_q)/p - (\lambda_1 + \cdots + \lambda_q)/p]$ is asymptotically normally distributed with mean 0 and variance

$$\frac{N}{(p\sum_\alpha b_\alpha)^2} \sum_{k,l,k',l'} \left\{ \sum_{i=1}^{q} a_{ii}(k,l) \right\} \left\{ \sum_{j=1}^{q} a_{jj}(k',l') \right\} \cdot$$
$$[2(1+\kappa)D(k,l,k',l') + \kappa\, C(k,l,k',l')].$$

2.2 Applications in finite sample

In practical applications, it is highly desirable to have the useful formulae for the construction of approximate confidence intervals and hypotheses testing based on the results of Section 2.1. Our asymptotic results suggest consistent estimators for the variance of the eigenvalues, the eigenvectors or the reduction of dimensionality statistics. However, since our asymptotic analysis assume that the number of families N goes go infinity, it is of interest to investigate what size of N will give reasonable coverage rates at the nominal level α. That issue will be discussed in more details in the simulation study of the next section.

Suppose that a confidence interval has to be constructed for the eigenvalue λ_i. Let $\hat{\mathbf{\Sigma}}$ be the unified estimator of $\mathbf{\Sigma}$ calculated with the R or S method. Let $\hat{\mathbf{H}}$ such that $\hat{\mathbf{H}}'\hat{\mathbf{R}}\hat{\mathbf{H}} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_p) = \hat{\mathbf{\Lambda}}$, where $\hat{\mathbf{R}}$ is the corresponding unified estimator for the correlation matrix. We can estimate consistently $a_{ij}(k,l)$ by

$$\hat{a}_{ij}(k,l) = \hat{h}_{ki}\hat{h}_{lj}[1 - \frac{1}{2}(\hat{\lambda}_i - \hat{\lambda}_j)\delta_{kl}].$$

To contruct a consistent estimator for the among offspring correlation matrix $\mathbf{\Phi}$, we need an estimator for $\mathbf{\Sigma}_s$. Such estimators are discussed in Konishi and Rao (1992). For example, we may consider for the R method

$$\hat{\mathbf{\Sigma}}_{s,R} = N_0^{-1}\{\mathbf{A}/(N-1) - \mathbf{W}/(\sum k_\alpha - N)\},$$

where in the components of variance model $\mathbf{A} = \sum k_\alpha(\bar{\mathbf{x}}_\alpha - \bar{\mathbf{x}})(\bar{\mathbf{x}}_\alpha - \bar{\mathbf{x}})'$ is the among families matrix and $\mathbf{W} = \sum \mathbf{S}_\alpha$ is the error matrix, with $\bar{\mathbf{x}} = \sum k_\alpha \bar{\mathbf{x}}_\alpha / \sum k_\alpha$. Another possibility is to use the S method. An estimator for $\mathbf{\Sigma}_s$ is

$$\hat{\mathbf{\Sigma}}_{s,S} = (N-1)^{-1}\{\sum(\bar{\mathbf{x}}_\alpha - \bar{\mathbf{x}}_{(\cdot)})(\bar{\mathbf{x}}_\alpha - \bar{\mathbf{x}}_{(\cdot)})' + \nu \sum \mathbf{S}_\alpha\},$$

where $\bar{\mathbf{x}}_{(\cdot)} = \sum \bar{\mathbf{x}}_\alpha / N$ and $\nu = -(N-1)N^{-1}\sum k_\alpha^{-1} / \sum(k_\alpha - 1)$. See Konishi and Rao (1992) for more details. Then, we can consider $\hat{\mathbf{\Phi}} = \hat{\mathbf{\Sigma}}_0^{-1/2}\hat{\mathbf{\Sigma}}_s\hat{\mathbf{\Sigma}}_0^{-1/2}$ and obtain a consistent estimator $\hat{D}(k,l,k',l')$ of $D(k,l,k',l')$. An estimator of variance of $\mathrm{var}(\hat{\lambda}_i)$ can then be obtained and an approximate confidence interval can be constructed.

3. EMPIRICAL STUDY

We performed a Monte Carlo simulation to examine the finite sample properties of

the proposed methodology. In the first set of experiments, we examined the bias and the accuracy of the asymptotic variance of the estimators of the eigenvalues, whereas in the second set of experiments we considered the computation of confidence intervals for the eigenvalues, at nominal levels 80%, 90%, 95% and 99%. We considered unified estimators based on the covariance and correlation matrices, using R and S methods. As in Konishi and Rao (1992), 100 000 random samples were generated from a multinormal population using

$$\mathbf{\Sigma} = \begin{pmatrix} 1.0 & 0.6 & 0.5 \\ 0.6 & 1.0 & 0.4 \\ 0.5 & 0.4 & 1.0 \end{pmatrix},$$

giving $\lambda_1 = 2.004$, $\lambda_2 = 0.613$ and $\lambda_3 = 0.382$. We used $\mathbf{\Sigma}_s = (\rho_s - \rho_0)\mathbf{I}_3 + \rho_0 \mathbf{1}_3 \mathbf{1}_3'$, with the values

(i) $(\rho_s, \rho_0) = (0.3, 0.1)$,

(ii) $(\rho_s, \rho_0) = (0.5, 0.3)$,

(iii) $(\rho_s, \rho_0) = (0.7, 0.4)$.

We generated family size according to a negative binomial distribution $NB(m, 0.5)$ truncated such that the family size is between 1 and 15. We generated $N = 25$ family sizes once and kept these values fixed in the simulation. Note that although the distributions are identical to those used by Konishi and Rao (1992), the values k_α generated most probably differ. The frequencies of the number of members for $m = 3$ and $m = 7$ are as follows.

$BN(3, 0.5)$

k_α	3	4	5	6	7	8	9
frequencies	4	3	8	4	2	3	1

$BN(7, 0.5)$

k_α	8	9	10	11	12	13	14	15
frequencies	2	1	5	4	3	4	2	4

The empirical means and variances of the different estimators of the eigenvalues were computed. Table 1 contains the results of the unified estimator based on the covariance matrix, whereas Table 2 gives the results based on the estimated correlation matrix. The asymptotic variances are given in parentheses. They were computed using the simplified formulas given in (4) and (5).

Table 1 and Table 2 show that the R and S methods give estimators with generally small biases compared to the variances. The results based on the covariance matrix illustrate that the largest eigenvalue is overestimated, whereas the smallest eigenvalue is underestimated. This is a well-known phenomenon studied for Wishart distributions by Takemura (1984) and observed by Konishi and Rao (1992) in the context of familial models. The behavior is different, however, when the correlation matrix is used. The largest eigenvalue is not consistently overestimated. Moreover, the smallest eigenvalue of the correlation matrix is less severely underestimated. Biases and variances of the largest estimated eigenvalue are markedly smaller for the correlation matrix.

The empirical and asymptotic variances are in close agreement, particularly for λ_2. The two tables show that the variances increase with ρ_s. The R and S methods

Table 1: Simulation Results, covariance matrix

		$\lambda_1 = 2.004$		$\lambda_2 = 0.613$		$\lambda_3 = 0.382$	
				$N = 25,\ NB(3, 0.5)$			
(ρ_s, ρ_0)		Mean	var $\times 10^4$	Mean	var $\times 10^4$	Mean	var $\times 10^4$
(0.3,0.1)	R	2.020	802(777)	0.619	83(85)	0.361	42(51)
	S	2.020	816(790)	0.619	83(85)	0.361	41(49)
(0.5,0.3)	R	2.026	1536(1473)	0.616	82(85)	0.358	41(51)
	S	2.026	1477(1415)	0.616	82(85)	0.359	41(49)
(0.7,0.4)	R	2.038	2321(2227)	0.618	117(122)	0.344	63(88)
	S	2.037	2170(2080)	0.618	113(118)	0.346	61(82)
				$N = 25,\ NB(7, 0.5)$			
(0.3,0.1)	R	2.014	477(461)	0.618	57(56)	0.367	35(40)
	S	2.014	475(459)	0.618	56(55)	0.368	34(39)
(0.5,0.3)	R	2.021	1237(1185)	0.615	56(56)	0.365	34(41)
	S	2.021	1211(1159)	0.615	56(56)	0.365	34(40)
(0.7,0.4)	R	2.033	2054(1970)	0.617	93(95)	0.350	59(79)
	S	2.032	1999(1916)	0.617	92(93)	0.350	58(77)

Table 2: Simulation Results, correlation matrix

		$\lambda_1 = 2.004$		$\lambda_2 = 0.613$		$\lambda_3 = 0.382$	
(ρ_s, ρ_0)		Mean	var $\times 10^4$	Mean	var $\times 10^4$	Mean	var $\times 10^4$
				$N = 25,\ NB(3, 0.5)$			
(0.3,0.1)	R	2.005	142(142)	0.626	76(77)	0.368	46(52)
	S	2.005	143(140)	0.626	76(76)	0.369	46(50)
(0.5,0.3)	R	1.998	218(218)	0.632	108(109)	0.370	59(65)
	S	1.999	213(209)	0.631	106(105)	0.370	57(63)
(0.7,0.4)	R	1.996	332(332)	0.644	161(163)	0.361	88(106)
	S	1.997	316(311)	0.642	154(154)	0.362	85(98)
				$N = 25,\ NB(7, 0.5)$			
(0.3,0.1)	R	2.006	92(92)	0.622	49(49)	0.372	34(38)
	S	2.006	91(89)	0.622	49(48)	0.372	34(37)
(0.5,0.3)	R	1.999	171(170)	0.628	82(82)	0.373	48(52)
	S	1.999	168(166)	0.628	81(80)	0.373	47(50)
(0.7,0.4)	R	1.996	292(290)	0.640	139(138)	0.364	80(94)
	S	1.996	286(282)	0.639	136(135)	0.364	79(91)

Table 3: Coverage rates, covariance matrix

		$\lambda_1 = 2.004$				$\lambda_2 = 0.613$				$\lambda_3 = 0.382$			
(ρ_s, ρ_0)		80 %	90%	95%	99 %	80 %	90%	95%	99 %	80 %	90%	95%	99 %
						$N = 25$, $NB(3, 0.5)$							
(0.3,0.1)	R	80.8	89.9	94.2	97.9	82.7	91.4	95.2	98.3	72.1	81.1	86.3	92.7
	S	80.8	89.9	94.2	97.9	82.5	91.4	95.2	98.3	72.4	81.2	86.5	92.7
(0.5,0.3)	R	80.1	88.8	92.5	96.3	82.2	91.0	94.8	98.1	71.0	80.0	85.4	92.0
	S	80.0	88.7	92.6	96.4	82.2	91.0	94.8	98.1	71.4	80.3	85.6	92.2
(0.7,0.4)	R	80.2	88.6	92.3	96.0	83.1	91.2	94.6	97.7	68.0	76.9	82.3	89.6
	S	80.2	88.6	92.3	96.1	83.1	91.2	94.7	97.8	68.6	77.2	82.7	89.8
						$N = 25$, $NB(7, 0.5)$							
(0.3,0.1)	R	80.5	89.9	94.3	97.9	81.9	90.8	94.7	97.9	73.1	81.9	86.9	92.9
	S	80.5	89.9	94.3	97.9	81.9	90.7	94.6	97.9	73.2	82.0	86.9	92.9
(0.5,0.3)	R	79.7	88.3	92.2	96.1	81.3	90.1	94.0	97.6	71.9	80.7	85.8	92.2
	S	79.7	88.4	92.2	96.1	81.2	90.1	94.0	97.6	72.0	80.7	85.9	92.3
(0.7,0.4)	R	80.0	88.4	92.1	96.0	82.5	90.6	94.0	97.4	69.0	77.7	83.1	90.2
	S	80.1	88.4	92.2	96.0	82.5	90.6	94.1	97.4	69.2	77.9	83.2	90.1

Table 4: Coverage rates, correlation matrix

		$\lambda_1 = 2.004$				$\lambda_2 = 0.613$				$\lambda_3 = 0.382$			
(ρ_s, ρ_0)		80 %	90%	95%	99 %	80 %	90%	95%	99 %	80 %	90%	95%	99 %
						$N = 25$, $NB(3, 0.5)$							
(0.3,0.1)	R	78.7	88.8	94.1	98.6	80.4	90.0	94.9	98.8	75.8	85.5	90.7	96.2
	S	78.3	88.6	93.9	98.5	80.0	89.7	94.6	98.7	75.5	85.3	90.5	96.1
(0.5,0.3)	R	77.7	87.8	93.3	98.0	78.7	88.6	93.6	98.1	76.0	86.1	91.6	97.0
	S	78.0	88.1	93.4	98.0	78.8	88.5	93.6	98.2	75.9	85.9	91.4	96.9
(0.7,0.4)	R	76.5	86.7	92.0	97.1	77.8	87.4	92.4	97.1	74.0	83.9	89.6	95.6
	S	77.3	87.2	92.4	97.2	78.4	87.8	92.7	97.3	74.3	84.1	89.5	95.5
						$N = 25$, $NB(7, 0.5)$							
(0.3,0.1)	R	78.9	89.0	94.2	98.6	80.0	90.0	94.9	98.9	75.8	85.6	90.7	96.1
	S	78.7	88.8	94.1	98.7	79.9	89.8	94.8	98.9	75.7	85.4	90.6	96.0
(0.5,0.3)	R	77.7	87.8	93.1	97.7	78.1	88.0	93.3	98.0	76.1	86.4	92.0	97.3
	S	77.8	87.8	93.0	97.6	78.1	88.2	93.3	98.0	76.1	86.3	91.9	97.3
(0.7,0.4)	R	76.4	86.5	91.9	96.9	77.2	86.8	92.0	96.8	74.0	84.0	89.7	95.8
	S	76.6	86.7	92.0	96.9	77.6	87.1	92.1	96.9	74.1	84.0	89.7	95.7

give very similar variances, although the variance of $\hat{\lambda}_1$ appears smaller with the S method when ρ_s is large. In each case, as it should be expected, the variances decrease when the number of siblings gets large.

Table 3 contains the coverage rates of confidence intervals for the eigenvalues of the unified estimator based on the covariance matrix, whereas Table 4 are the coverage rates based on the correlation matrix. In general, it seems that the R and S methods give very similar results. In general, when ρ_s and ρ_0 become larger, the coverage rates deteriorate slightly.

All the empirical coverage rates are very close to the nominal significance level for λ_2. For the smallest eigenvalue λ_3, the confidence intervals based on the unified estimator of the covariance matrix are underestimated. This is in accordance with Konishi and Rao (1992, p.639), who found that the maximum errors of the normal approximations of the distributions for the smallest eigenvalues were large and erratic. Better results are obtained using the unified estimator of the correlation matrix, particularly at the 99% nominal level. This is probably related to the bias problem when doing a PCA based on the covariance matrix, since λ_3 was then markedly underestimated, and the bias was a more important component of the MSE, compared to the other eigenvalues. Both methods seem to give comparable results for the largest eigenvalue λ_1.

It should be noted that since our asymptotic results assume that the number of families N goes to infinity, we obtain in general very reasonable coverage rates for confidence intervals for N as small as 25 families.

The two sets of experiments were performed using FORTRAN subroutines of the NAG library. Some S-PLUS functions have been written for the computation of variance estimators and asymptotic variances. All the computer code is available from the second author.

4. CONCLUDING REMARKS

The asymptotic distribution of eigenvalues of the correlation matrix was derived in familial models with unequal number of siblings under elliptical sampling. The Monte Carlo simulation shows the usefulness of the asymptotic variances and for approximate confidence interval for the number of families as small as $N = 25$.

The perturbation theory assumes that the eigenvalues of interest are singular. This hypothesis can not be drawn easily even in the simplest case of a Wishart matrix with the bootstrap. In the case of a Wishart matrix with multiple population eigenvalues of multiplicity p_i, $\sum_i p_i = p$, it has been shown that bootstrapping averages of p_i eigenvalues with bootstrap sample size N is consistent. However, bootstrap of a single eigenvalue is inconsistent when $p_i > 1$ unless a bootstrap sample of size $o(N)$ is used; see Beran and Srivastava (1985, 1987), Eaton and Tyler (1991), and Bilodeau (2001). The unequal family sizes would be another difficulty in the application of the bootstrap. The asymptotic theory in this paper assumes that the family sizes k_α are constant quantities. Bootstrap sample of N families would yield different family sizes k_α from one bootstrap sample to another. The bootstrap would treat the k_α as random. It is our guess that boostrap estimate of the variance of the eigenvalue estimates would overestimate the asymptotic variance in Theorem 1.

REFERENCES

Bellman, R. (1960). *Introduction to Matrix Analysis.* McGraw-Hill, New York.

Beran, R., Srivastava, M. S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *Ann. Statist.* **13**, 95-115.

Beran, R., Srivastava, M. S. (1987). Correction: Bootstrap tests and confidence regions for functions of a covariance matrix. *Ann. Statist.* **15**, 470-471.

Bilodeau, M. (2001). *Asymptotic distribution of the largest eigenvalue*, technical report, Université de Montréal.

Bilodeau, M. (1999). *Theory of Multivariate Statistics*, Springer Texts in Statistics, New York.

Eaton, M. L, Tyler, D. E. (1991). On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *Ann. Statist.* **19**, 260-271.

Konishi, S. (1979). Asymptotic espansions for the distributions of statistics based on the sample correlation matrix in principal component analysis. *Hiroshima Math. J.* **9**, 647-700.

Konishi, S., Khatri, C. G. (1990). Inferences on interclass and intraclass correlations in multivariate familial data. *Ann. Inst. Statist. Math.* **42**, 561-580.

Konishi, S., Rao, C. R. (1992). Principal component analysis for multivariate data. *Biometrika* **79**, 631-641.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory.* John Wiley & Sons, New york.

Srivastava, M. S., Keen, K. J., Katapa, R. S. (1988). Estimation of interclass and intraclass correlations in multivariate familial data. *Biometrics* **44**, 141-150.

Takemura, A. (1984). An orthogonally invariant minimax estimator of the covariance matrix of a multivariate normal population. *Tsukuba J. Math.* **8**, 367-376.

Département de mathématiques et de statistique
Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal (Québec)
Canada H3C 3J7
e-mail: bilodeau@dms.umontreal.ca

École des Hautes Études Commerciales
Service d'enseignement des Méthodes Quantitatives de Gestion
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7
e-mail: pierre.duchesne@hec.ca