General projection–pursuit estimates for the common principal components model: Influence functions and Monte Carlo study

G. Boente[1], A. M. Pires[2] and I. M. Rodrigues[2]

[1] Departamento de Matemática and Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Ciudad Universitaria, Pabellón 1. 1428, Buenos Aires, Argentina.

[2] Departamento de Matemática, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001, Lisboa, Portugal

Keywords: Common principal components, Influence functions, Projection–pursuit, Robust estimation.

1 Introduction

In multivariate analysis, we often deal with situations involving several populations, such as discriminant analysis, where the assumption of equality of covariance matrices is usually assumed. Yet sometimes, this assumption is not adequate but problems related to an excessive number of parameters will arise if we estimate the covariance matrices separately for each population. In many practical situations, this problem can be avoided if the covariance matrices of the different populations exhibit some common structure. Several authors, as for instance Flury (1988), have studied models for common structure dispersion. One such basic common structure assumes that the k covariance matrices have different eigenvalues but identical eigenvectors, *i.e.*, there is an orthogonal matrix $\boldsymbol{\beta} \in I\!R^{p\times p}$ such that $\boldsymbol{\Lambda}_i = \boldsymbol{\beta}'\boldsymbol{\Sigma}_i\boldsymbol{\beta}$, $i = 1,\dots,k$ where $\boldsymbol{\Sigma}_i$ is the covariance matrix of the *i-th* population and $\boldsymbol{\Lambda}_i = diag\,(\lambda_{i1},...,\lambda_{ip})$. This model, proposed by Flury (1984), became known as the Common Principal Components (CPC) model. He derived the maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}_i$ assuming multivariate normality of the original variables $\mathbf{X}_i$, $i = 1,...,k$. It is well known that in practice the classical CPC analysis can be affected by the existence of outliers in a sample. In order to obtain robust estimates, one approach is to consider robust affine equivariant estimators of the covariance matrices $\boldsymbol{\Sigma}_i$, $i = 1,...,k$, as done by Boente and Orellana (2001) and Boente, Pires and Rodrigues (2001). These authors also studied an approach based on the projection–pursuit principles. In this last case, the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}_i$ are the solution of

$$r(\widehat{\boldsymbol{\beta}}_1) = \sup_{\|\mathbf{b}\|=1} \sum_{i=1}^{k} \tau_i\, s^2(\mathbf{X}_i'\mathbf{b}) \qquad r(\widehat{\boldsymbol{\beta}}_j) = \sup_{\mathbf{b}\in\mathcal{B}_j} \sum_{i=1}^{k} \tau_i s^2(\mathbf{X}_i'\mathbf{b}) \quad 2 \le j \le p\,, \tag{1}$$

where $\mathcal{B}_j = \{\mathbf{b} : \|\mathbf{b}\| = 1, \mathbf{b}'\widehat{\boldsymbol{\beta}}_m = 0 \text{ for } 1 \le m \le j-1\}$ and s is a univariate scale estimate. A more general approach is to consider a score function applied to the scale estimate. In this case, the estimates of the common principal axes are obtained by solving iteratively

$$r(\widehat{\boldsymbol{\beta}}_1) = \sup_{\|\mathbf{b}\|=1} \sum_{i=1}^{k} \tau_i\, f\left\{s^2(\mathbf{X}_i'\mathbf{b})\right\} \qquad r(\widehat{\boldsymbol{\beta}}_j) = \sup_{\mathbf{b}\in\mathcal{B}_j} \sum_{i=1}^{k} \tau_i f\left\{s^2(\mathbf{X}_i'\mathbf{b})\right\} \quad 2 \le j \le p\,, \tag{2}$$

where f is a strictly increasing score function.

In both cases, the estimates of the eigenvalues and the covariance matrix of the i-th population are computed as $\widehat{\lambda}_{ij} = s^2(\mathbf{X}_i'\widehat{\boldsymbol{\beta}}_j)$, for $1 \le j \le p$, $\mathbf{V}_i = \sum_{j=1}^{p} \widehat{\lambda}_{ij}\widehat{\boldsymbol{\beta}}_j\widehat{\boldsymbol{\beta}}_j'$, for $1 \le i \le k$.

2 Influence functions and Asymptotic variances

With the aim of evaluating the robustness of our procedure we derive the partial influence functions of the estimates defined by (2). Let $\sigma(\cdot)$ be the scale functional related to the scale estimate s.

Assume that $\mathbf{\Lambda}_i^{-\frac{1}{2}}\mathbf{x}_{i1} = \mathbf{z}_i$ has the same spherical distribution G for all $1 \leq i \leq k$ and that $\sigma(G_0) = 1$ where G_0 is the distribution of z_{11}. Moreover, assume that f is twice continuously differentiable and that the function $(\epsilon, y) \to \sigma\left((1-\epsilon)G_0 + \epsilon\delta_y\right)$ is twice continuously differentiable at $(0, y)$. Then, we have that for any $\mathbf{x}$

$$\begin{aligned} \mathrm{PIF}_i(\mathbf{x}, \lambda_{\sigma,\ell j}, F) &= 2\ \delta_{\ell i}\ \lambda_{ij}\ \mathrm{IF}\left(\frac{\mathbf{x}'\boldsymbol{\beta}_j}{\sqrt{\lambda_{ij}}}, \sigma, G_0\right) & (3) \\ \mathrm{PIF}_i(\mathbf{x}, \boldsymbol{\beta}_{\sigma,j}, F) &= \tau_i\ \boldsymbol{\beta}_j'\mathbf{x}\sum_{s=1}^{j-1}\frac{1}{\nu_{sj}-\nu_s}\sqrt{\lambda_{is}}f'\left(\lambda_{is}\right)\ \mathrm{DIF}\left(\frac{\boldsymbol{\beta}_s'\mathbf{x}}{\sqrt{\lambda_{is}}}, \sigma, G_0\right)\ \boldsymbol{\beta}_s + \\ &\quad + \tau_i\sqrt{\lambda_{ij}}f'\left(\lambda_{ij}\right)\mathrm{DIF}\left(\frac{\boldsymbol{\beta}_j'\mathbf{x}}{\sqrt{\lambda_{ij}}}, \sigma, G_0\right)\sum_{s=j+1}^{p}\frac{1}{\nu_j-\nu_{js}}\boldsymbol{\beta}_s'\mathbf{x}\,\boldsymbol{\beta}_s\ , & (4) \end{aligned}$$

where $\mathrm{DIF}(y, \sigma, G)$ denotes the derivative of $\mathrm{IF}(y, \sigma, G)$ with respect to y, $\nu_{js} = \sum_{i=1}^{k}\tau_i f'\left(\lambda_{ij}\right)\lambda_{is}$, $\nu_j = \nu_{jj}$ and $\nu_{js} \neq \nu_{jj}$ for $s \neq j$.

These expressions allow us to derive heuristically the asymptotic variance of the estimates defined by (2), as

$$\begin{aligned} \mathrm{ASVAR}\left(\widehat{\lambda}_{\ell j}\right) &= 4\lambda_{\ell j}^2\frac{1}{\tau_\ell}\mathrm{ASVAR}\left(\sigma, G_0\right) & (5) \\ \mathrm{ASVAR}\left(\widehat{\boldsymbol{\beta}}_{jm}\right) &= \sum_{i=1}^{k}\tau_i\lambda_{ij}\lambda_{im}\left\{\frac{\delta_{m>j}\left\{f'\left(\lambda_{ij}\right)\right\}^2}{\left(\nu_j-\nu_{jm}\right)^2} + \right. \\ &\quad + \left.\frac{\delta_{m<j}\left\{f'\left(\lambda_{im}\right)\right\}^2}{\left(\nu_{mj}-\nu_m\right)^2}\right\}E_G\left\{\mathrm{DIF}\left(z_{1j}, \sigma, G_0\right)z_{1m}\right\}^2 . & (6) \end{aligned}$$

For the particular case of a proportional model, the optimal estimates defined through (2) in the sense of minimizing the asymptotic variance given by (6), for any strictly increasing score function f twice continuously differentiable, are those related to $f(t) = \ln(t)$.

Through a simulation study these estimates are compared with those related to $f(t) = t$ for small samples.

References

G. Boente and L. Orellana (2001). A Robust Approach to Common Principal Components. In *Statistics in Genetics and in the Environmental Sciences*, eds. L. T. Fernholz, S. Morgenthaler, and W. Stahel, pp. 117-147. Basel: Birkhauser.

G. Boente, A.M. Pires and I. Rodrigues (2001). Influence functions and outlier detection under the common principal components model. Submitted to *Biometrika.*

B. Flury (1984). Common principal components in k groups. *Journal of the American Statistical Association*, 79, 892–898.

B. Flury (1988). *Common Principal Components and Related Multivariate Models.* Wiley & Sons, New York.

Please fill in this form and mail it together with your abstract.

My abstract fits best to topic number 14

List of Topics:

1. Algorithms
2. Applications
3. Biostatistics
4. Computing and graphics
5. Data analysis
6. Data mining
7. Economics, finance
8. Efficiency and robustness
9. Functionals and bias
10. Fuzzy statistics
11. Geostatistics
12. Inference for robust methods, model testing
13. Location depth and regression depth
14. Multivariate methods
15. Neural networks
16. Rank-based methods
17. Regression quantiles, trimming
18. Robust covariance
19. Robust designs
20. Robust regression
21. Time series analysis
22. Wavelets