

# Reweighted estimators for the common principal components model: Influence functions and Monte Carlo study

G. Boente<sup>1</sup>, A. M. Pires<sup>2</sup> and I. M. Rodrigues<sup>2</sup>

<sup>1</sup> CONICET and Departamento de Matemática and Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Ciudad Universitaria, Pabellón 1. 1428, Buenos Aires, Argentina.

<sup>2</sup> Departamento de Matemática, Instituto Superior Técnico and CEMAT, Av. Rovisco Pais, 1049-001, Lisboa, Portugal

**Keywords:** Asymptotic variances; Common principal components; Outlier detection; Projection–Pursuit; Robust estimation.

## 1 Introduction

Several authors, as Flury (1988), have studied models for common dispersion structure. Those models have been introduced to overcome the problem of an excessive number of parameters, when dealing with several populations, in multivariate analysis. One such basic common structure assumes that the  $k$  covariance matrices have different eigenvalues but identical eigenvectors, i.e.,  $\Sigma_i = \beta \Lambda_i \beta^T$ ,  $1 \leq i \leq k$ , where  $\Lambda_i$  are diagonal matrices,  $\beta$  is the orthogonal matrix of the common eigenvectors and  $\Sigma_i$  is the covariance matrix of the  $i$ th population. This model was proposed in Flury (1984) and became known as the *Common Principal Components* (CPC) model.

It is well known that in practice the classical CPC analysis can be affected by the existence of outliers in a sample. The replacement of the classical covariance matrices  $\Sigma_i$ ,  $i = 1, \dots, k$  by robust affine equivariant estimators is perhaps the most simple and intuitive robust approach. This robust plug-in (PI) technique was studied by Boente and Orellana (2001) and recently by Boente, Pires and Rodrigues (2002a). These authors also considered another robust approach based on projection–pursuit (PP) principles. Moreover, a more general robust approach is to apply a general increasing score function,  $f$ , to the robust scale estimate (Boente, Pires and Rodrigues, 2002b). In this case, the estimates of the common principal axes are obtained by solving iteratively

$$r(\hat{\beta}_1) = \sup_{\|\mathbf{b}\|=1} \sum_{i=1}^k \tau_i f \left\{ s^2(\mathbf{X}_i^T \mathbf{b}) \right\} \quad r(\hat{\beta}_j) = \sup_{\mathbf{b} \in \mathcal{B}_j} \sum_{i=1}^k \tau_i f \left\{ s^2(\mathbf{X}_i^T \mathbf{b}) \right\} \quad 2 \leq j \leq p, \quad (1)$$

where  $\mathcal{B}_j = \{\mathbf{b} : \|\mathbf{b}\| = 1, \mathbf{b}^T \hat{\beta}_m = 0 \text{ for } 1 \leq m \leq j-1\}$  and  $s$  is a univariate robust scale estimate.

## 2 Reweighted estimators

As it is well known, reweighted estimators allow to improve the asymptotic efficiency of the initial estimators. We consider a reweighted estimator of the scatter matrices of each population, where the weights do not depend on the Mahalanobis distance as usually, but on the outlier detection measures defined in Boente, Pires and Rodrigues (2002a). In order to avoid masking, those measures depend on the influences on the classical functionals but with the unknown parameters robustly estimated. To summarize our proposal, given an observation  $\mathbf{x}$  from the  $i$ th population, the expressions for the

diagnostics are

$$\begin{aligned}
 IML(\mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Lambda}}_i) &= IML_i(\mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Lambda}}) = \left[ \sum_{r=1}^p \frac{\left\{ \left( \hat{\boldsymbol{\beta}}_r^T \mathbf{x} \right)^2 - \hat{\lambda}_{ir} \right\}^2}{2\hat{\lambda}_{ir}^2} \right]^{\frac{1}{2}} \\
 IMB(\mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Lambda}}_i) &= IMB_i(\mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Lambda}}) = \left[ \sum_{r=1}^p \sum_{s \neq r} \frac{\left\{ \left( \hat{\boldsymbol{\beta}}_r^T \mathbf{x} \right) \left( \hat{\boldsymbol{\beta}}_s^T \mathbf{x} \right) \right\}^2}{\hat{\lambda}_{ir} \hat{\lambda}_{is}} \right]^{\frac{1}{2}}, \quad (2)
 \end{aligned}$$

where  $\hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{\Lambda}}_i = \text{diag}(\hat{\lambda}_{i1}, \dots, \hat{\lambda}_{ip})$  are the robust estimators.

Assuming that  $\boldsymbol{\mu}_i = \mathbf{0}_p$ , an estimate of  $\boldsymbol{\Sigma}_i$  can be defined as

$$\hat{\boldsymbol{\Sigma}}_i = \frac{\sum_{j=1}^{n_i} w \left( IML^2(\mathbf{x}_{ij}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Lambda}}_i), IMB^2(\mathbf{x}_{ij}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Lambda}}_i) \right) \mathbf{x}_{ij} \mathbf{x}_{ij}^T}{\sum_{j=1}^{n_i} w \left( IML^2(\mathbf{x}_{ij}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Lambda}}_i), IMB^2(\mathbf{x}_{ij}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Lambda}}_i) \right)}. \quad (3)$$

The reweighted estimators of the principal axes and of the eigenvalues are obtained by plugging-in these estimators into the equations defining the maximum likelihood estimators for normal data.

### 3 Influence functions and asymptotic variances

With the aim of evaluating the robustness of this approaches we derive the partial influence functions for the reweighted functionals. With these expressions we obtain heuristically the asymptotic variances of the related estimators for ellipsoidal distributions. This allows us to calibrate the estimates in order to improve the efficiency of the initial estimates.

### 4 Monte Carlo study

We performed a simulation to evaluate the finite sample behaviour of this approach. The weights considered are smooth functions that penalize observations with large values of the influence measures defined in (2).

### 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 (2002a). Influence functions and outlier detection under the common principal components model: A robust approach. *Biometrika*, 89, 861–875.
- G. Boente, A. M. Pires and I. Rodrigues (2002b). General projection-pursuit estimators for the common principal components model: Influence functions and Monte Carlo study. Submitted.
- 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*. New York: John Wiley.