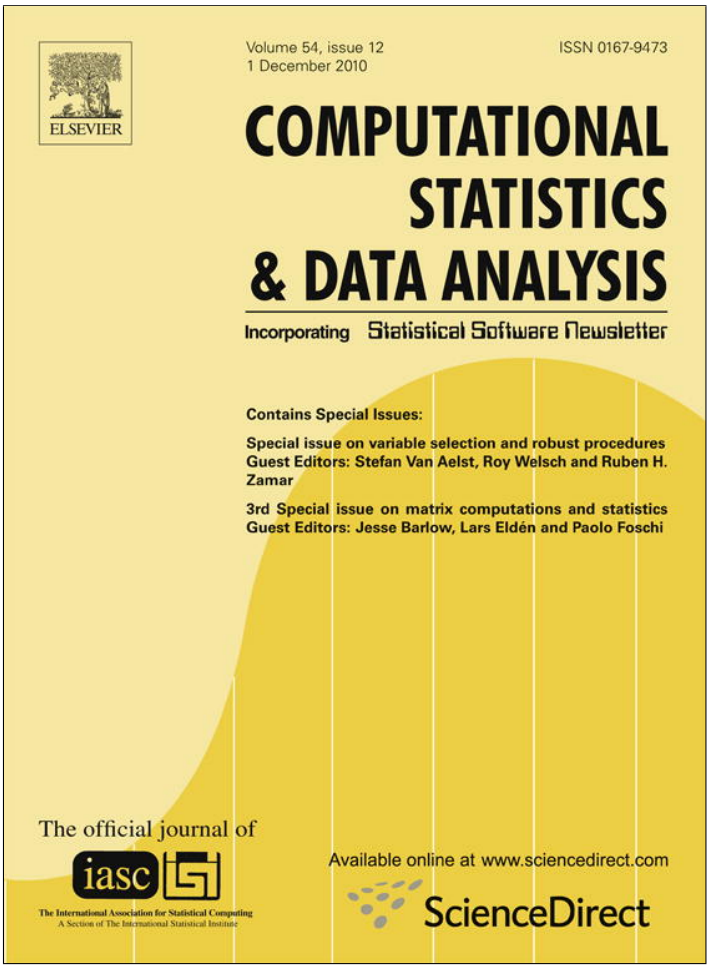

Computational Statistics and Data Analysis 54 (2010) 2967–2975

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Detecting influential observations in principal components and common principal components

Graciela Boente [a,b,*], Ana M. Pires [c], Isabel M. Rodrigues [c]

[a] Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina

[b] CONICET, Argentina

[c] Departamento de Matemática and CEMAT, Instituto Superior Técnico, Technical University of Lisbon (TULisbon), Lisboa, Portugal

ARTICLE INFO

Article history:
Received 30 January 2009
Received in revised form 31 December 2009
Accepted 2 January 2010
Available online 14 January 2010

Keywords:
Common principal components
Detection of outliers
Influence functions
Robust estimation

ABSTRACT

Detecting outlying observations is an important step in any analysis, even when robust estimates are used. In particular, the robustified Mahalanobis distance is a natural measure of outlyingness if one focuses on ellipsoidal distributions. However, it is well known that the asymptotic chi-square approximation for the cutoff value of the Mahalanobis distance based on several robust estimates (like the minimum volume ellipsoid, the minimum covariance determinant and the *S*-estimators) is not adequate for detecting atypical observations in small samples from the normal distribution. In the multi-population setting and under a common principal components model, aggregated measures based on standardized empirical influence functions are used to detect observations with a significant impact on the estimators. As in the one-population setting, the cutoff values obtained from the asymptotic distribution of those aggregated measures are not adequate for small samples. More appropriate cutoff values, adapted to the sample sizes, can be computed by using a cross-validation approach. Cutoff values obtained from a Monte Carlo study using *S*-estimators are provided for illustration. A real data set is also analyzed.

1. Introduction

Detecting outlying observations is an important step in any analysis, even when robust estimates are used, either because there is some specific interest in finding anomalous observations or as a pre-processing task before the application of some multivariate method, in order to preserve the results from possible harmful effects of those observations. If one focuses on ellipsoidal distributions, the robustified Mahalanobis distance (Rousseeuw and van Zomeren, 1990) is a natural measure of outlyingness which is generally used to detect outliers from the central normal distribution. More recently, Filzmoser et al. (2008) introduced a computationally fast procedure for identifying outliers based on a principal component analysis that is particularly effective in high dimensions while Hubert et al. (2009) considered the situation of skewed distributions. We refer the reader to Serneels and Verdonck (2008) and Chen et al. (2009) for some recent proposals on principal component analysis and outlier detection for data sets with missing observations. On the other hand, as is well known, influence functions can be used to detect influential/outlying observations. It is worth noticing that, in general, an outlier may not be an influential observation for the estimation of the parameter of interest but an influential observation is usually an outlier. An influential observation can be described as an observation with high influence on something, usually an estimate of the parameters of interest. For the one-population case, Croux and Haesbroeck (1999) discussed the use of empirical influence

* Corresponding address: Instituto de Cálculo, Ciudad Universitaria, Pabellón 2, Buenos Aires, C1428EHA, Argentina. Tel.: +54 11 45763375; fax: +54 11 45763375.
E-mail address: gboente@dm.uba.ar (G. Boente).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.01.001

functions of the eigenvalues and eigenvectors of the sample covariance matrix (Critchley, 1985; Shi, 1997) and those of the one-step reweighted minimum covariance determinant estimator (Rousseeuw, 1985). As expected, empirical influence functions of the robust estimators hardly change when contaminated data points are included in the sample, because outliers usually have small influence on robust estimators, while, if we consider the empirical influence of the classical estimators, a masking effect may appear, preventing the detection of outlying observations. An alternative approach, in order to avoid masking, is to consider robust empirical influence functions for the classical estimators but with the parameters estimated through a robust procedure (see Pison et al., 2000). This procedure is analogous to the use of the robustified version of the Mahalanobis distance introduced by Rousseeuw and van Zomeren (1990). Usually, the cutoff values are computed assuming implicitly that the researcher is interested in detecting just one independent observation as an outlier, i.e., a new observation independent of the sample at hand. Instead, Becker and Gather (2001) considered multivariate outlier detection rules based on Mahalanobis-type distances such that, for a multivariate normal sample of size n, no observation within the sample is identified as an outlier with probability $1-\alpha$. Considering $\mathbf{x}_j \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\mathbf{t}_n$ and $\mathbf{V}_n$ the robust location and scatter estimators of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the observation $\mathbf{x}$ is declared an outlier relatively to the observed sample if the squared Mahalanobis distance, $MD^2(\mathbf{x}, \mathbf{t}_n, \mathbf{V}_n)$, is greater than a detection limit c_{MD^2}. If the goal is to detect just one new independent observation, $\mathbf{x}^{new}$, as an outlier, c_{MD^2} satisfies $P(MD^2(\mathbf{x}^{new}, \mathbf{t}_n, \mathbf{V}_n) < c_{MD^2}) = 1-\alpha$. On the other hand, for the experiment for which "no observation, in a sample of size n, is identified as an outlier", the cutoff value c_{MD^2} should be calculated as $P(\max_{1\le j\le n} MD^2(\mathbf{x}_j, \mathbf{t}_n, \mathbf{V}_n) < c_{MD^2}) = 1-\alpha$. For large samples the value c_{MD^2} can be approximated by its asymptotic value, i.e., by the chi-square percentile, $\chi^2_{p,\beta}$, where $\beta = 1-\alpha$ in the first approach and $\beta = (1-\alpha)^{\frac{1}{n}}$ in the latter. For not very large samples, and depending on the estimators $\mathbf{t}_n$ and $\mathbf{V}_n$, the exact cutoffs may be quite different from the asymptotic values. Becker and Gather (2001) studied this effect for the minimum volume ellipsoid estimator (Rousseeuw, 1985), the minimum covariance determinant estimator (Rousseeuw, 1985) and the S-estimator (Rousseeuw and Yohai, 1984) based on Tukey's biweight function. Those authors suggest that more reliable cutoff values can easily be determined by simulation, taking into account the dimension and the sample size of the data set. Besides, the rule based on the S-estimator leads to the best results in most situations. Hardin and Rocke (2005) also considered the robustified Mahalanobis distance using the minimum covariance determinant estimator, usually denoted as the *MCD*-estimator.

As mentioned above, an influential observation is an observation with high influence on an estimator of some of the parameters of interest. This has motivated the introduction of detection measures under a *Common Principal Components* (CPC) model. This model (Flury, 1984) deals with several populations with a common scatter structure, i.e. assumes that independent observations $\mathbf{x}_{ij}$, $1 \le j \le n_i$, $1 \le i \le k$, from k independent samples of size n_i in $\mathbb{R}^p$, are identically distributed within each sample, with location parameter $\boldsymbol{\mu}_i$ and scatter matrix $\boldsymbol{\Sigma}_i$ such that

$$\boldsymbol{\Sigma}_i = \boldsymbol{\beta}\boldsymbol{\Lambda}_i\boldsymbol{\beta}^{\mathrm{T}}, \quad 1 \le i \le k, \tag{1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$ is the orthogonal matrix of the common eigenvectors and $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i1}, \ldots, \lambda_{ip})$ are the diagonal matrices containing the eigenvalues for each population. Boente et al. (2002) proposed using aggregated measures based on standardized empirical influence functions in trying to detect observations with a significant impact on the estimators of the parameters of the CPC model for k multivariate normal samples. This is a natural approach as not all the outliers detected by the Mahalanobis distance have a high influence on those estimators and downweighting only the observations which are influential avoids efficiency losses. If, on the other hand, interest is in finding anomalous observations (relatively to the model distribution or the majority of the data) without reference to a specific parameter, this procedure may also be interesting because unlike the usual outlyingness measures it gives additional information on the reasons why the highlighted observations differ from the bulk of the data. The proposed summary diagnostic measures can also be applied when dealing with just one multivariate population.

The cutoff values used in Boente et al. (2002) were asymptotic approximations for detecting a new observation as influential. As for the Mahalanobis distance, more reliable cutoff values can be obtained taking into account the number of populations and the sample sizes. Following these ideas, in this paper we propose an adaptive method for computing the cutoff values for detecting influential/outlying observations in the CPC setting. In Section 2 we describe the method. A procedure for computing the cutoff values is given in Section 3. An example is studied in Section 4. Finally, some conclusions are given in Section 5.

2. The proposal

Let us consider the multi-population setting under a CPC model. As mentioned in Becker and Gather (1999), *"the identification of outliers heavily relies on the assumption of some underlying model for the data. An observation can finally only be considered as an outlier in respect to such a model in mind"*. The same arguments apply when detecting influential observations for the principal axis and their sizes. For this reason, throughout this paper, we will assume that the multivariate normal distribution is the central model (the one we want to detect deviations from). Thus, let $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})$ be k independent samples of independent observations such that $\mathbf{x}_{i1} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ where the scatter matrices $\boldsymbol{\Sigma}_i$ satisfy the model (1). Let $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Lambda}}_i$ be robust equivariant estimators of the common directions and the related eigenvalues of the i-th population based on $\mathbf{X}_1, \ldots, \mathbf{X}_k$. Instead of considering standardized robust scores or separate influence plots for each parameter, Boente et al. (2002) consider just two aggregated influential measures, one for the eigenvalues and one for the eigenvectors, denoted

Table 1
Cutoff values of $G^2_{\boldsymbol{\lambda}}$ and $G^2_{\boldsymbol{\beta}}$, $G^2_{\boldsymbol{\lambda},\gamma}$ and $G^2_{\boldsymbol{\beta},\gamma}$, corresponding to $\gamma = 1-(1-\alpha)^{\frac{1}{n}}$ for different values of n and p.

p	α	$n=1$		$n=20$		$n=30$		$n=40$		$n=50$		$n=100$	
		0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
2	$G^2_{\boldsymbol{\lambda}}$	8.545	24.526	44.496	75.946	52.022	83.374	57.630	90.555	61.463	93.845	75.651	105.692
2	$G^2_{\boldsymbol{\beta}}$	9.551	25.971	46.791	78.395	53.971	88.187	59.376	94.539	63.780	98.461	78.083	112.122
3	$G^2_{\boldsymbol{\lambda}}$	12.384	30.841	53.382	86.052	61.267	96.590	66.387	103.545	71.211	106.273	85.646	119.657
3	$G^2_{\boldsymbol{\beta}}$	25.604	57.154	92.133	144.302	102.886	159.032	113.349	168.621	120.640	174.086	143.618	191.238
5	$G^2_{\boldsymbol{\lambda}}$	18.713	40.629	66.002	102.818	73.414	110.641	80.084	118.871	85.543	123.156	102.463	136.284
5	$G^2_{\boldsymbol{\beta}}$	71.523	135.376	202.789	297.667	222.477	317.571	237.229	328.862	251.843	340.909	295.580	380.688
10	$G^2_{\boldsymbol{\lambda}}$	31.378	58.626	88.323	131.911	97.871	141.301	105.050	147.906	110.452	149.943	131.505	164.678
10	$G^2_{\boldsymbol{\beta}}$	253.343	408.048	557.683	739.712	602.753	791.481	635.780	814.346	664.542	834.136	736.033	909.870

as *IML* and *IMB* (short for *influence measure for* λ and *influence measure for* $\boldsymbol{\beta}$), respectively. To avoid the problem of the different sizes of the eigenvalues, those measures are defined using the standardized robust empirical functions leading to the following simple functions of the standardized robust scores:

$$IML_i^2(\mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i) = \sum_{r=1}^{p} \frac{\left\{\left(\widehat{\boldsymbol{\beta}}_r^{\mathrm{T}}(\mathbf{x}-\widehat{\boldsymbol{\mu}}_i)\right)^2 - \widehat{\lambda}_{ir}\right\}^2}{2\widehat{\lambda}_{ir}^2},$$

$$IMB_i^2(\mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i) = \sum_{r=1}^{p}\sum_{s\neq r} \frac{\left\{\left(\widehat{\boldsymbol{\beta}}_r^{\mathrm{T}}(\mathbf{x}-\widehat{\boldsymbol{\mu}}_i)\right)\left(\widehat{\boldsymbol{\beta}}_s^{\mathrm{T}}(\mathbf{x}-\widehat{\boldsymbol{\mu}}_i)\right)\right\}^2}{\widehat{\lambda}_{ir}\widehat{\lambda}_{is}},$$

where $\widehat{\boldsymbol{\mu}}_i$ is a robust equivariant estimator of the location $\boldsymbol{\mu}_i$ of the i-th population, $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_p)$ and $\widehat{\boldsymbol{\Lambda}}_i = \mathrm{diag}(\widehat{\lambda}_{i1}, \ldots, \widehat{\lambda}_{ip})$, $1 \le i \le k$. In this way, the eigenvector diagnostics, IMB_i^2, turn out to be invariant through orthogonal transformations. As mentioned in the introduction, robust estimators need to be considered in the estimation of the unknown parameters in order to avoid masking and/or swamping effects (see, for instance, the discussion in Hardin and Rocke, 2005). As pointed out by Boente et al. (2002), to detect influential observations with respect to the multivariate normal distribution, one must compare the observed values of IML^2 and IMB^2 with high percentiles of the distribution functions of the random variables $G^2_{\boldsymbol{\lambda}} = \sum_{r=1}^{p}\left(z_r^2-1\right)^2/2$ and $G^2_{\boldsymbol{\beta}} = \sum_{r=1}^{p}\sum_{s\neq r} z_r^2 z_s^2$, respectively, where $z_1, \ldots, z_p$ are independent and identically distributed $N(0,1)$ random variables. These (asymptotic) percentiles, denoted as $G^2_{\boldsymbol{\lambda},\alpha}$ and $G^2_{\boldsymbol{\beta},\alpha}$, were obtained through a simulation study.

The previous approach has fixed cutoff values and does not account for the sample size n_i and/or the number of populations k of the data structure. However, as argued before, it is better to adjust them to the data set at hand. As in Becker and Gather (2001), one possibility is to consider the following measures:

$$AL_i(\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i) = \max_{1\le j\le n_i} IML^2(\mathbf{x}_{ij}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i),$$

$$AB_i(\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i) = \max_{1\le j\le n_i} IMB^2(\mathbf{x}_{ij}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i),$$

where $\mathbf{X}_i = \left(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}\right)$ is the sample of the i-th population. Due to the equivariance of the estimators, the cutoff values $a_{IML^2,\alpha}$ and $a_{IMB^2,\alpha}$ depend only on the sample sizes and not on the estimated eigenvalues of each population. Therefore, they can be chosen such that

$$P(AL_1(\mathbf{X}_1, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_1) < a_{IML^2,\alpha}) = 1-\alpha \qquad P(AB_1(\mathbf{X}_1, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_1) < a_{IMB^2,\alpha}) = 1-\alpha,$$

for different values of n_i and they may be easily derived through a Monte Carlo study with the selected robust location and scatter estimation procedure. However, when considering the Donoho (1982), Stahel (1981) and S-estimators, a simulation study showed that IML^2 and IMB^2 have very heavy tailed distributions, leading to extremely large values of $a_{IML^2,\alpha}$ and $a_{IMB^2,\alpha}$ when $\alpha = 0.01$ or even $\alpha = 0.05$. This fact may be explained by the behavior of the random variables $G_{\boldsymbol{\lambda}}$ and $G_{\boldsymbol{\beta}}$. Effectively, the cutoff values computed taking into account the sample sizes but not the estimation of $\boldsymbol{\Lambda}_i$ and $\boldsymbol{\beta}$, that is the percentiles $\gamma_i = 1-(1-\alpha)^{1/n_i}$ of $G_{\boldsymbol{\lambda}}$ and $G_{\boldsymbol{\beta}}$ (Table 1), increase considerably with the sample size, especially as dimension increases. The cutoff values reported in Boente et al. (2002) correspond to the choice $n = 1$ and are also given, for comparison. It is worth noticing that as dimension increases, with large sample sizes, the observations influential for the common directions are those more difficult to detect if the asymptotic cutoff values are considered.

For the sake of simplicity, from now on, the subscript α in cutoff values will be omitted whenever the meaning is clear.

Another approach for obtaining cutoff values adapted to the sample sizes and the estimation is to consider a procedure related to a leave-one-out cross-validation method. Given the cutoff constants $c_{IML_i^2}$ and $c_{IMB_i^2}$, an observation $\mathbf{x}_{ij}$ is influential

for the i-th population if

$$IML_i^2(\mathbf{x}_{ij}, \widehat{\boldsymbol{\beta}}^{(-j)}, \widehat{\boldsymbol{\Lambda}}_i^{(-j)}) > c_{IML_i^2} \quad \text{or} \quad IMB_i^2(\mathbf{x}_{ij}, \widehat{\boldsymbol{\beta}}^{(-j)}, \widehat{\boldsymbol{\Lambda}}_i^{(-j)}) > c_{IMB_i^2},$$

where $\widehat{\boldsymbol{\beta}}^{(-j)}$ and $\widehat{\boldsymbol{\Lambda}}_i^{(-j)}$ are the estimates computed without the observation $\mathbf{x}_{ij}$. We have to repeat this process for all observations of all populations, like in a leave-one-out cross-validation procedure.

Since the observation to be detected is independent of those used for computing the estimates, the cutoff values can be defined as follows. For $1 \leq i \leq k$, denote by $\mathbf{x}_i^{new}$ a new independent observation from the i-th population, $\mathbf{X}_i$, and compute $IML_i^2(\mathbf{x}_i^{new}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i)$ and $IMB_i^2(\mathbf{x}_i^{new}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i)$. The cutoff values $c_{IML_i^2}$ and $c_{IMB_i^2}$ are defined as the values satisfying

$$P(IML_i^2(\mathbf{x}_i^{new}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i) \leq c_{IML_i^2} | (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i)) = 1 - \alpha \tag{2}$$

and

$$P(IMB_i^2(\mathbf{x}_i^{new}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i) \leq c_{IMB_i^2} | (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i)) = 1 - \alpha, \tag{3}$$

with $\mathbf{x}_i^{new} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

As mentioned in the Introduction, the influential measures, IML_i and IMB_i, may give additional information on the type of *outlyingness* of a detected observation. If $IMB_i(\mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i)$ is large it means that $\mathbf{x}$ is distant from the center of the i-th population on a diagonal direction relatively to some of the principal axes, i.e., those with the largest contributions to $IMB_i(\mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i)$. Also, if $IML_i(\mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i)$ is large, it means that $\mathbf{x}$ has a large score on some of the common principal components, i.e., those with the largest contribution to $IML_i(\mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i)$. In contrast, if the Mahalanobis distance is large, we only know that the point is distant from the center of the data. Therefore, if an observation $\mathbf{x}$ is detected as an outlier and it also has a large value of $IML_i(\mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i)$ but a small value of $IMB_i(\mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i)$, one can suspect that this point might be influential when estimating the size of the components. On the other hand, if $\mathbf{x}$ is an outlier with a large value of $IMB_i(\mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i)$ and a small value of $IML_i(\mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i)$, $\mathbf{x}$ might be distant from the center of the i-th population on a diagonal direction relatively to some of the principal axes. This is especially interesting if the common principal components have a meaningful interpretation.

3. Computation of the adaptive procedure

To compute the cutoff points defined in (2) and (3), we can proceed as follows.

Step 1. Generate independent observations $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}, 1 \leq i \leq k$, such that $\mathbf{x}_{i1} \sim N_p(\mathbf{0}_p, \boldsymbol{\Sigma}_i)$. Compute the robust estimates of the common directions and their sizes $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Lambda}}$.

Step 2. Generate a new independent observation from the i-th population, $\mathbf{x}_i^{new}$.

Step 3. Compute $IML_i^2(\mathbf{x}_i^{new}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i)$ and $IMB_i^2(\mathbf{x}_i^{new}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_i)$. Call the value of interest u.

Step 4. Repeat Step 1 to Step 3 N times, saving the value u at each step.

Step 5. Sort the values obtained, $\{u_j\}_{1 \leq j \leq N}$, as $u^{(1)} \leq \cdots \leq u^{(N)}$. The $(1-\alpha)$-th quantile, $u^{(N\,(1-\alpha))}$, gives an approximation to the cutoff points $c_{IML_i^2}$ or $c_{IMB_i^2}$.

It should be noticed that the diagnostic measures are invariant for translations and orthogonal transformations. Hence, since the observations are centered using an equivariant location estimator, we can assume that $\boldsymbol{\mu}_i = \mathbf{0}$ and $\boldsymbol{\beta} = \mathbf{I}_p$ when generating the data for computing the cutoff values. Therefore, given k independent samples of independent observations $\mathbf{Y}_i = (\mathbf{y}_{i1}, \ldots, \mathbf{y}_{in_i})$, let us denote by $\widehat{\boldsymbol{\beta}}^{(\mathbf{Y})}$ and $\widehat{\boldsymbol{\Lambda}}_i^{(\mathbf{Y})}$ the robust estimators of the common directions and the related eigenvalues of the i-th population based on the samples $\mathbf{Y}_1, \ldots, \mathbf{Y}_k$. In order to detect influential observations in the samples, we perform a parametric bootstrap using Steps 1 to 5 by taking $\boldsymbol{\Sigma}_i$ as the diagonal matrix $\widehat{\boldsymbol{\Lambda}}_i^{(\mathbf{Y})}$. To stabilize the variability one might perform N_R replications.

To illustrate the procedure, we have computed the cutoff values for $k = 2$ populations in dimensions $p = 2$ and 5 as described above. We have also considered a situation in dimension $p = 10$, to be described below. In dimension 2, we generated normal data with covariance matrices $\boldsymbol{\Sigma}_1 = \mathrm{diag}(14, 4)$ and $\boldsymbol{\Sigma}_2 = \mathrm{diag}(12, 2)$, which have well separated eigenvalues, to avoid consistency problems with the projection-pursuit estimators. In dimension 5, we have considered as scatter matrices $\boldsymbol{\Sigma}_1 = \mathrm{diag}(33.849, 7.375, 2.472, 2.061, 0.764)$ and $\boldsymbol{\Sigma}_2 = \mathrm{diag}(85.613, 21.085, 3.820, 1.130, 0.986)$ that correspond to the estimated eigenvalues in the example below. To center the observations, the location estimate, $\mathbf{t}_{n_i}$, obtained using S-estimators was used, while to estimate robustly $\widehat{\boldsymbol{\Lambda}}_i$ and $\widehat{\boldsymbol{\beta}}$, the projection-pursuit procedure defined in Boente et al. (2006) was used with $f(t) = \ln(t)$ and an M-scale estimator. Tables 2–4 give the resulting cutoff values computed using $N = 5000$ and $N_R = 10$ when $n_i \leq 50$, $i = 1, 2$, and $N_R = 3$, otherwise. Since the mean and median values are quite similar, in all tables we report the median over replications. The S-estimates were computed using the MATLAB programs provided on Christophe Croux's personal Web site taking 1000 random p-subsets and using, as function ρ, Tukey's biweight function calibrated to attain a 25% breakdown point. To allow fair comparisons, we also report, in those tables, the cutoff values, c_{MD^2}, obtained for the robust Mahalanobis distance computed considering as location and scatter

Table 2
Cutoff values for $n_1 = 30$ and $n_2 = 22$.

	α	1st population			2nd population		
		0.10	0.05	0.01	0.10	0.05	0.01
$p = 2$	$c_{IML_i^2}$	6.857	14.895	52.972	8.596	19.564	74.855
	$c_{IMB_i^2}$	7.027	14.267	46.691	8.949	18.947	68.729
	c_{MD^2}	5.599	7.525	12.384	6.062	8.237	14.246
$p = 5$	$c_{IML_i^2}$	19.926	35.153	92.469	31.398	60.427	180.390
	$c_{IMB_i^2}$	77.515	118.656	264.237	112.297	196.696	521.590
	c_{MD^2}	12.886	16.021	23.707	15.175	19.367	31.047

Table 3
Cutoff values for $n_1 = 50$ and $n_2 = 50$.

	α	1st population			2nd population		
		0.10	0.05	0.01	0.10	0.05	0.01
$p = 2$	$c_{IML_i^2}$	5.762	12.007	38.780	5.805	12.292	39.236
	$c_{IMB_i^2}$	6.286	12.096	37.266	6.430	12.836	40.562
	c_{MD^2}	5.158	6.839	10.945	5.171	6.841	10.963
$p = 5$	$c_{IML_i^2}$	16.550	28.394	73.150	16.744	28.641	72.954
	$c_{IMB_i^2}$	66.836	102.084	215.880	69.285	112.185	236.395
	c_{MD^2}	11.199	13.877	19.834	11.175	13.719	19.565

Table 4
Cutoff values for $n_1 = 100$ and $n_2 = 20$ for dimension $p = 2, 5$.

	α	1st population			2nd population		
		0.10	0.05	0.01	0.10	0.05	0.01
$p = 2$	$c_{IML_i^2}$	5.041	9.661	29.587	9.578	23.559	79.927
	$c_{IMB_i^2}$	5.912	11.433	32.820	8.870	17.923	59.013
	c_{MD^2}	4.937	6.365	9.969	6.282	8.596	16.038
$p = 5$	$c_{IML_i^2}$	13.184	22.067	54.803	29.482	55.197	175.951
	$c_{IMB_i^2}$	57.427	89.621	188.159	106.899	171.846	419.117
	c_{MD^2}	10.039	12.206	17.067	16.601	21.214	32.903

Table 5
Cutoff values for $n_1 = 100$ and $n_2 = 100$ for dimension $p = 2, 5, 10$.

	α	1st population			2nd population		
		0.10	0.05	0.01	0.10	0.05	0.01
$p = 2$	$c_{IML_i^2}$	5.123	9.680	30.866	5.218	11.154	32.288
	$c_{IMB_i^2}$	5.516	10.228	27.982	6.101	11.304	32.401
	c_{MD^2}	4.876	6.275	9.751	5.013	6.535	10.177
$p = 5$	$c_{IML_i^2}$	12.966	21.920	54.150	13.833	23.038	55.992
	$c_{IMB_i^2}$	57.326	85.757	160.526	60.143	93.785	209.319
	c_{MD^2}	10.035	12.348	16.900	10.095	12.263	17.520
$p = 10$	$c_{IML_i^2}$	26.911	40.664	78.571	26.799	41.020	85.279
	$c_{IMB_i^2}$	233.408	328.315	616.109	237.382	340.598	610.016
	c_{MD^2}	18.632	21.635	28.416	18.746	22.032	28.148

estimators the S-estimators, using a procedure similar to that described in Step 1 to Step 5. In dimension $p = 2$, it is possible to avoid resampling when computing the projection-pursuit eigenvector estimators by maximizing over a fixed number κ of equally spaced directions. A possible choice could be $\kappa = 1000$ which gives quite reliable results.

It is worth noticing that as dimension increases the cutoff values are much larger for a fixed sample size than those corresponding to the asymptotic cutoff values reported in Boente et al. (2002) and also in the first column of Table 1. For dimensions $p = 2$ and 5, we need more than 100 observations in each sample to attain the asymptotic cutoff. Moreover, if the percentiles $G^2_{\boldsymbol{\lambda},\gamma_i}$ and $G^2_{\boldsymbol{\beta},\gamma_i}$, with $\gamma_i = 1 - (1-\alpha)^{1/n_i}$, reported in Table 1, are considered, some influential observations may not be detected, since the exact cutoff values are much smaller. Table 5 also reports the cutoff values computed with $n_1 = 100$ and $n_2 = 100$ for dimension $p = 2, 5$ and 10. In this last situation, we have considered $\boldsymbol{\Sigma}_1 = \mathrm{diag}(1, 5, 10, 20, 30, 50, 65, 80, 95, 115)$ and $\boldsymbol{\Sigma}_2 = 4\boldsymbol{\Sigma}_1$. Note that when $p = 2$ we are still far from the asymptotic cutoff $G^2_{\boldsymbol{\lambda},\alpha}$, and the difference between the asymptotic cutoff and the resampling ones becomes larger as the dimension

Table 6
Cutoff values for different values of n and p.

p	α	$n = 20$		$n = 50$		$n = 100$		$n = 200$		$n = 400$		$n = 1000$	
		0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01		
2	$c_{IML_1^2}$	22.550	96.168	12.576	40.366	10.853	32.937	9.465	27.072	8.608	25.284	8.548	23.024
2	$c_{IMB_1^2}$	21.286	73.951	14.298	46.270	11.261	34.302	10.182	30.908	9.817	27.408	9.387	24.380
3	$c_{IML_1^2}$	42.290	165.709	18.337	56.202	14.777	40.789	14.007	36.219	12.875	33.702	12.456	29.761
3	$c_{IMB_1^2}$	65.924	203.209	38.217	102.677	31.618	81.370	28.371	69.705	26.800	59.164	26.211	59.984
5	$c_{IML_1^2}$	131.359	562.097	30.534	79.223	23.538	59.116	20.637	46.368	19.677	42.681	18.371	39.221
5	$c_{IMB_1^2}$	304.374	854.292	115.940	270.480	89.077	214.390	83.220	177.510	79.674	157.503	75.450	138.475
10	$c_{IML_1^2}$	* * *	* * *	* * *	* * *	46.093	94.686	36.030	71.083	32.858	62.736	31.409	56.689
10	$c_{IMB_1^2}$	* * *	* * *	* * *	* * *	367.635	689.145	310.787	576.205	290.055	515.678	267.336	456.851

increases, especially for $\alpha = 0.01$. It is worth noticing that $c_{IMB_i^2,\alpha}$ seems to approximate faster to $G^2_{\boldsymbol{\beta},\alpha}$, since for sample sizes equal to 100, the relative deviations, i.e., the ratios $|c_{IMB_i^2,\alpha} - G^2_{\boldsymbol{\beta},\alpha}|/G^2_{\boldsymbol{\beta},\alpha}$, are not larger than 18%, while for $c_{IML_i^2,\alpha}$ that ratio is larger than 33% for $\alpha = 0.01$.

From a practical point of view, it is important to have a guideline on the sample sizes for which the asymptotic cutoff values are a reasonable approximation. With that purpose we have computed the cutoff values $c_{IML_1^2}$ and $c_{IMB_1^2}$ for different dimensions and for one population ($k = 1$). The sample sizes were taken as $n = 20, 50, 100, 200, 400$ and 1000. We generated normal data sets in dimension $p = 2, 3, 5$ and 10 with covariance matrices $\boldsymbol{\Sigma} = \text{diag}(5, 1)$, $\boldsymbol{\Sigma} = \text{diag}(10, 5, 1)$, $\boldsymbol{\Sigma} = \text{diag}(50, 20, 10, 5, 1)$ and $\boldsymbol{\Sigma} = \text{diag}(115, 95, 80, 65, 50, 30, 20, 10, 5, 1)$, respectively. These covariance matrices have well separated eigenvalues, to avoid consistency problems with the projection-pursuit estimators. As above, to center the observations, we considered the S-estimators, while to estimate robustly $\widehat{\boldsymbol{\Lambda}}_1$ and $\widehat{\boldsymbol{\beta}}$, projection-pursuit estimators based on an M-scale estimator were computed. Table 6 gives the resulting cutoff values, using $N = 5000$ and $N_R = 3$, computed as the median values over replications. Note that in dimension $p = 10$, we do not give the cutoff values for $n = 20$ and 50 since the number of observations is too low for estimating all the parameters. The results given in Table 6 show that, even for dimension $p = 2$, very large samples, $n \geq 400$, are needed so that the asymptotic cutoff value provides a reliable value. This shows the advantage of the proposed procedure, since, otherwise, for moderate sample sizes some extra points would be detected as influential when using the asymptotic cutoff values.

4. Example

To illustrate the proposed outlier detection procedure and its relevance, we have selected a real data set with $k = 2$ populations and five variables. This data set is part of a larger data set described in Oliveira (1995) where a principal component analysis was performed and it was also studied in Boente et al. (2002). The variables are the following measurements made on two varieties, *Lada*, $n_1 = 100$, and *Longal*, $n_2 = 47$, of chestnut tree leaves of the genus *Castanea*, in 1989:

x_1: petiole length (in mm);
x_2: number of nervures from the right side of the leaf;
x_3: number of nervures from the left side of the leaf;
x_4: number of teeth from the right side of the leaf;
x_5: number of teeth from the left side of the leaf.

As mentioned in Boente et al. (2002), a common principal components model was judged adequate after a robust principal component analysis of each variety showed similar principal axes with different amounts of variability. Therein, the robust common principal components obtained using plug-in estimates and the projection-pursuit estimates with $f(t) = t$ are reported. We have used the projection-pursuit procedure defined in Boente et al. (2006) with $f(t) = \ln(t)$. The estimated eigenvalue matrices are $\widehat{\boldsymbol{\Lambda}}_1 = \text{diag}(33.849, 7.375, 2.472, 2.061, 0.764)$ and $\widehat{\boldsymbol{\Lambda}}_2 = \text{diag}(85.613, 21.085, 3.820, 1.130, 0.986)$ while the common eigenvector matrix is

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 0.797 & 0.594 & 0.104 & 0.007 & -0.042 \\ -0.338 & 0.578 & -0.582 & -0.378 & 0.266 \\ -0.346 & 0.328 & 0.660 & -0.457 & -0.359 \\ -0.291 & 0.372 & -0.152 & 0.686 & -0.533 \\ -0.218 & 0.261 & 0.439 & 0.422 & 0.717 \end{pmatrix}.$$

Detection of outliers becomes an important feature in this setting. In Boente et al. (2002), several influential observations were detected in each group using the asymptotic cutoff values given in the first column of Table 1. The cutoff values corresponding to the procedure described in Section 3 are reported in Table 7.

Table 7
Cutoff values for $n_1 = 100$ and $n_2 = 47$ for dimension $p = 5$.

α	1st population			2nd population		
	0.10	0.05	0.01	0.10	0.05	0.01
$c_{IML_i^2}$	13.742	22.711	51.275	16.783	29.090	79.677
$c_{IMB_i^2}$	58.444	90.181	181.517	72.358	110.900	267.172
c_{MD^2}	10.279	12.455	16.967	11.314	13.756	20.0618

Table 8
Observations detected with the asymptotic cutoff values at level $1-(1-\alpha)^{1/n_i}$, in increasing order of the influential measures.

	Lada			*Longal*			
IML^2	87	88	30	1	24		
IMB^2	30			22	6	24	1
MD^2	30						

The labels of the observations detected as possible outliers when $\alpha = 0.05$, using *IML* and *IMB* for the projection-pursuit method, using the asymptotic cutoff values and those computed using Steps 1 to 5, are indicated in Fig. 1. Besides, observations detected as outliers with the asymptotic values but that are not detected with the cutoff values reported in Table 7 are plotted with a solid circle. The horizontal dashed lines correspond to the asymptotic cutoff values while the solid ones correspond to the proposed cutoff values. As in Becker and Gather (2001), and according to the discussion given in Section 3, the asymptotic values detect more observations than the exact ones.

When comparing the influence measures *IML* and *IMB* with the robust Mahalanobis distance, we see that the highest influential observations are detected by both methods. There are however some discrepancies for the observations near the detection limits, which is to be expected since (*IML*, *IMB*) and the robust Mahalanobis distance are measuring different effects. Note that observation 45 of the *Lada* variety only appears as an influential observation for the eigenvalues with our cutoff values while with the asymptotic ones it is detected by all procedures. Besides, observations 85 and 93 are influential with respect to the eigenvector estimators but are not detected as outliers using the robust Mahalanobis distance. On the other hand, observations 1, 6, 22 and 42 of the *Longal* variety seem to influence the common eigenvectors and the eigenvalues estimation while they are not considered as potential outliers using the Mahalanobis distance. Note also that observations 1, 6 and 42 were detected by means of the squared Mahalanobis distance when using the asymptotic cutoff $\chi^2_{p,0.05}$.

It is worth noticing that using the asymptotic detection measures defined in Boente et al. (2002) with $1-(1-\alpha)^{1/n_i}$ points for $G^2_{\boldsymbol{\lambda}}$ and $G^2_{\boldsymbol{\beta}}$, many observations are not detected. Table 8 reports the labels of the observations detected as influential.

This example shows clearly the advantage of the procedure described in Section 2, since it allows one to detect observations masked by other methods. It is worth noticing that the observations in Table 8 are mainly the observations detected as influential with our cutoff values at the 1% level.

5. Concluding remarks

A new method for identifying influential/outlying observations for the principal axes and their sizes, in principal component analysis and/or the CPC setting, was developed according to adaptive percentiles. This method accounts not only for the sample size n_i but also for the number of populations k of the data structure. The simulation study showed that for small samples the adaptive percentiles could be much larger than the percentiles of $G^2_{\boldsymbol{\lambda}}$ and $G^2_{\boldsymbol{\beta}}$, recommended in Boente et al. (2002). This is analogous to the behavior of the cutoff values corresponding to the robustified Mahalanobis distance described in Becker and Gather (2001).

For the real data set analyzed, this adaptive procedure not only detected more influential/outlying observations than the adaptive method based on the Mahalanobis distances but it also provides a nice interpretation of its *outlyingness*. Besides, it allows one to detect observations that are not detected with the asymptotic $1-(1-\alpha)^{1/n_i}$ cutoff values.

It is worth noting that in many practical applications, principal component analysis is based on correlation matrices rather than on covariance matrices. This is particularly the case in situations in which the units of measurement are arbitrary. As mentioned in Flury (1988), the *FG*-algorithm can be applied to the estimated correlation matrices of the k populations. However, in the classical setting, the estimates obtained may not be the maximum likelihood estimates. In the robust situation, robust plug-in estimators for several populations can be defined as in the one-population case (see Croux and Haesbroeck, 2000) by using robust correlation matrix estimators. Thus, after computation of the influence measures of the new defined estimators, cutoff values based on a plug-in procedure can be easily adapted to this particular setting by considering the algorithm described. Nevertheless, it is not clear how to define a procedure analogous to the projection-pursuit approach considered in Boente et al. (2006) and so the computation of cutoff values based on this approach needs further research.

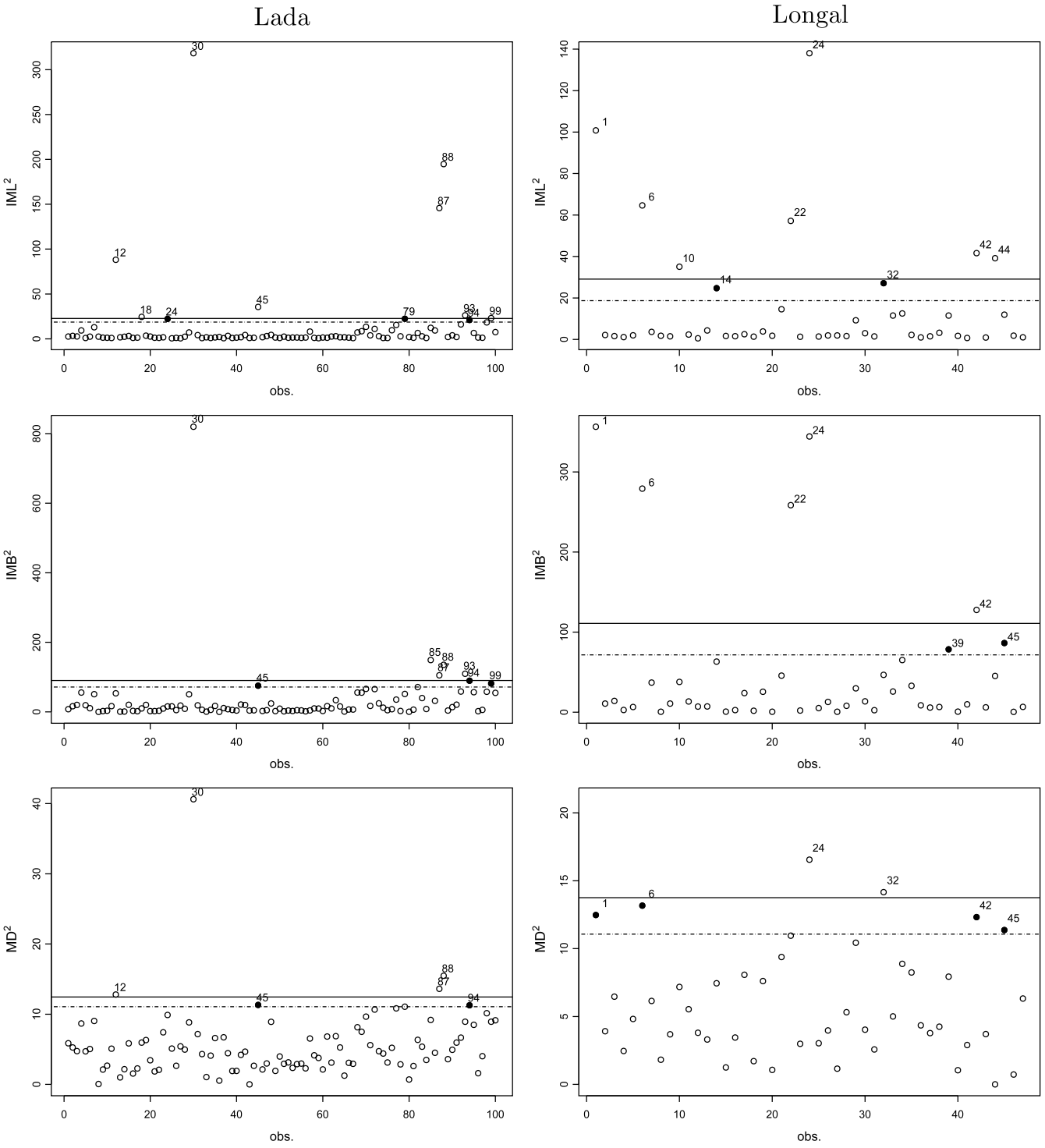

Fig. 1. Chestnut tree data. Varieties Lada $n_1 = 100$, Longal $n_2 = 47$. Observations detected as outliers at the 5% detection level. The solid line corresponds to the computed cutoff values while the dashed line corresponds to the asymptotic ones.

Acknowledgements

This research was partially supported by Grants X-018 from the Universidad de Buenos Aires, PIP 0216 from CONICET and PICT 00821 from ANPCYT, Argentina, and also by the Center for Mathematics and its Applications, Lisbon, Portugal, through *Programa Operacional Ciência, Tecnologia, Inovação* (POCTI) of the Fundação para a Ciência e a Tecnologia (FCT), cofinanced by the European Community fund FEDER. We also wish to thank two anonymous referees for valuable comments which led to an improved version of the original paper.

References

Becker, C., Gather, U., 1999. The masking breakdown point of multivariate outlier identification rules. Journal of the American Statistical Association 94, 947–955.

Becker, C., Gather, U., 2001. The largest nonidentifiable outliers: A comparison of multivariate simultaneous outliers identification rules. Computational Statistics and Data Analysis 36, 119–127.

Boente, G., Pires, A.M., Rodrigues, I.M., 2002. Influence functions and outlier detection under the common principal components model: A robust approach. Biometrika 89, 861–875.

Boente, G., Pires, A.M., Rodrigues, I.M., 2006. General projection-pursuit estimators for the common principal components model: Influence functions and Monte Carlo study. Journal of Multivariate Analysis 97, 124–147.

Chen, T., Martin, E., Montague, G., 2009. Robust probabilistic PCA with missing data and contribution analysis for outlier detection. Computational Statistics and Data Analysis 53, 3706–3716.

Critchley, F., 1985. Influence in principal components analysis. Biometrika 72, 627–636.

Croux, C., Haesbroeck, G., 1999. Empirical influence functions for robust principal component analysis. In: Proceedings of the Statistical Computing Section of the American Statistical Association. Am. Statist. Assoc., Alexandria, VA, pp. 201–206.

Croux, C., Haesbroeck, G., 2000. Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. Biometrika 87, 603–618.

Donoho, D.L., 1982. Breakdown Properties of Multivariate Location Estimators. Ph.D. Thesis. Harvard University (in English).

Filzmoser, P., Maronna, R., Werner, M., 2008. Outlier identification in high dimensions. Computational Statistics and Data Analysis 52, 1694–1711.

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

Flury, B.N., 1988. Common Principal Components and Related Multivariate Models. John Wiley, New York.

Hardin, J., Rocke, D., 2005. The distribution of robust distances. Journal of Computational and Graphical Statistics 14, 928–946.

Hubert, M., Rousseeuw, P., Verdonck, T., 2009. Robust PCA for skewed data and its outlier map. Computational Statistics and Data Analysis 53, 2264–2274.

Oliveira, I., 1995. Variedades de castanheiros em Trás-os-Montes. Uma análise em componentes principais dos caracteres morfológicos da folha. Master Thesis. Universidade de Lisboa (in Portuguese).

Pison, G., Rousseeuw, P.J., Filzmoser, P., Croux, C., 2000. A robust version of principal factor analysis. In: Bethlehem, J., van der Heijden, P. (Eds.), Compstat: Proceedings in Computational Statistics. Physica-Verlag, Heidelberg, pp. 385–390.

Rousseeuw, P.J., 1985. Multivariate estimation with high breakdown point. In: Grossmann, W., et al. (Eds.), Mathematical Statistics and Applications, Vol. B. Akadémiai Kiadó, Budapest, pp. 283–297.

Rousseeuw, P.J., van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association 85, 633–639.

Rousseeuw, P.J., Yohai, V.J., 1984. Robust regression by means of *S*-estimators. In: Franke, J., et al. (Eds.), Robust and Nonlinear Time Series Analysis. In: Lecture Notes in Statistics, vol. 26. Springer-Verlag, New York, pp. 256–272.

Serneels, S., Verdonck, T., 2008. Principal component analysis for data containing outliers and missing elements. Computational Statistics and Data Analysis 52, 1712–1727.

Shi, L., 1997. Local influence in principal components analysis. Biometrika 84, 175–186.

Stahel, W.A., 1981. Robust Estimation: Infinitesimal Optimality and Covariance Matrix Estimators. Ph.D. Thesis. ETH, Zurich (in German).