

PCA based on Multivariate MM-estimators with Fast and Robust Bootstrap

Matías Salibián-Barrera ¹, Stefan Van Aelst ², and Gert Willems ³

Abstract

We consider robust principal component analysis based on multivariate MM-estimators. We first study the robustness and efficiency of these estimators, in particular regarding eigenvalues and eigenvectors. Then we focus on inference procedures based on a fast and robust bootstrap for MM-estimators. This method is an alternative to the approach based on the asymptotic distribution of the estimators, and can also be used to assess the stability of the principal components. A formal consistency proof for the bootstrap method is given and its finite-sample performance is investigated through simulations. We illustrate the use of the robust principal components method and the bootstrap inference on a real dataset.

1 Introduction

Principal component analysis (PCA) is a very common technique in multivariate statistics. It aims to explain the covariance structure of the data by projecting the observations

¹Department of Statistics, University of British Columbia, 6356 Agriculture Road, Vancouver, B.C., Canada V6T 1Z2. E-mail: matias@stat.ubc.ca. The research of Matías Salibián-Barrera has been partially funded by the National Science and Engineering Research Council of Canada.

²Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, B-9000 Ghent, Belgium. E-mail: stefan.vanaelst@ugent.be. The research of Stefan Van Aelst was supported by a grant of the Fund for Scientific Research-Flanders (FWO-Vlaanderen).

³Department of Statistics, University of British Columbia, 6356 Agriculture Road, Vancouver, B.C., Canada V6T 1Z2. E-mail: gert.willems@ua.ac.be. Part of this research was done while Gert Willems was a researcher at the University of Antwerp funded by the GOA grant “Statistical and numerical techniques for the modelling and optimisation of computer- and communication networks.”

onto a small number of principal components, which are linear combinations of the original variables. Classical PCA is based on the eigenvectors and eigenvalues of the sample covariance or correlation matrix. The eigenvector associated with the largest eigenvalue gives the projection of the data which has the largest variance, while the eigenvalue itself measures the amount of information that can be explained by that projection. The sample covariance matrix however is notorious for being sensitive to outliers. Therefore, several robustifications of classical PCA have been proposed in the literature, such as projection-pursuit approaches (Li and Chen 1985, Croux and Ruiz-Gazen 2005, Hubert et al. 2002) and the use of robust estimators of scatter instead of the sample covariance. The latter approach has been proposed and investigated by many authors starting with Maronna (1976), Campbell (1980) and Devlin et al. (1981). More recently Croux and Haesbroeck (2000) compared several robust estimators in a more formal manner. In this paper we consider a related type of PCA based on robust estimates of shape. In particular we will use the eigenvectors and eigenvalues of multivariate MM-estimators of shape as introduced by Tatsuoka and Tyler (2000). MM-estimators are designed to be both highly robust against outliers and highly efficient for normal data. Essentially, an S-estimator is used to obtain a robust scale estimate, after which the location vector and shape matrix are estimated with a more efficient M-estimator.

We are primarily interested in the inference part of the PCA method based on the MM-estimator. As in classical PCA, results based on asymptotic normality can be used to construct confidence intervals or to estimate standard errors (see e.g. Croux and Haesbroeck, 2000). However, these results only hold under the assumption of some underlying elliptical distribution. Such assumptions are often not appropriate in those cases where robust estimation is most recommended. Inference based on the asymptotic variances derived at the central model may still yield reasonable results for small amounts of contamination when the sample size is large. The bootstrap (Efron 1979) provides a computer-intensive

alternative that can work better for smaller sample sizes and for larger deviations from the central model. Moreover, since the bootstrap can estimate the sampling distribution of the estimator of interest, it offers a wider range of inference applications and allows to assess the stability of the PCA results. However, applying bootstrap on robust estimators such as the MM-estimator raises some difficulties. One serious problem is the high computational cost of these estimators. Indeed, computing the MM-estimator, particularly the initial S-estimator, is a time-consuming task. Recalculating the estimates many times, as the bootstrap requires, can therefore lead to an extreme computational cost. This is especially true for large datasets in high dimensions. Another typical problem that arises is the instability of the classical bootstrap procedure. If harmful outlying observations are present in the original sample, then these will occur in various numbers in the bootstrap samples as well. Even if the MM-estimator on the original data yields a robust PCA solution, it might fail in bootstrap samples with many outliers. In other words, inference based on the bootstrap may be less robust than the MM-estimator itself. The robustness of the bootstrap in general has been investigated by Singh (1998) and Stromberg (1997).

Recently, Salibián-Barrera and Zamar (2002) proposed a fast and robust bootstrap method for univariate regression MM-estimators. Here we will investigate an adaptation of their method to the multivariate location and shape setting. This extension can be used to obtain many recalculations of the MM-shape or scatter matrix. As with the classical bootstrap, we will base our inference on the eigenvalues and eigenvectors of these recomputed matrices.

The rest of the paper is organized as follows. In Section 2 we derive asymptotic and robustness properties of the multivariate MM-estimators. In Section 3 some theoretical aspects of PCA based on the MM-estimates are discussed. Section 4 is devoted to the fast and robust bootstrap. Here we present formal consistency results and investigate the finite-sample performance through a simulation study that compares the method with the

approach based on the asymptotic variances at the central model. In Section 5 a real data example is given. Section 6 contains some concluding remarks, and all the proofs can be found in the Appendix.

2 Multivariate MM-estimators

2.1 Definition

Analogously to MM-estimators of regression (Yohai 1987), MM-estimators of multivariate location and shape are based on two loss functions. In what follows we will require the following regularity conditions for a loss function ρ :

(R1) ρ is real, symmetric, twice continuously differentiable and $\rho(0) = 0$.

(R2) ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty)$ for some finite constant c .

Multivariate MM-estimators of location, shape and covariance are now defined as follows:

Definition 1 Let $\mathcal{X}_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^p$ with $n \geq p + 1$. Let ρ_0 and ρ_1 satisfy (R1) and (R2) and let $(\tilde{\boldsymbol{\mu}}_n, \tilde{\boldsymbol{\Sigma}}_n)$ be multivariate S-estimators, that is $(\tilde{\boldsymbol{\mu}}_n, \tilde{\boldsymbol{\Sigma}}_n)$ minimize $|C|$ subject to

$$\frac{1}{n} \sum_{i=1}^n \rho_0 \left([(\mathbf{x}_i - T)^t C^{-1} (\mathbf{x}_i - T)]^{\frac{1}{2}} \right) = b$$

among all $(T, C) \in \mathbb{R}^p \times PDS(p)$. Here, $PDS(p)$ denotes the set of positive definite symmetric $p \times p$ matrices. Denote $\hat{\sigma}_n := |\tilde{\boldsymbol{\Sigma}}_n|^{1/(2p)}$. Then the multivariate MM-estimators for location and shape $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Gamma}}_n)$ minimize

$$\frac{1}{n} \sum_{i=1}^n \rho_1 \left([(\mathbf{x}_i - T)^t G^{-1} (\mathbf{x}_i - T)]^{\frac{1}{2}} / \hat{\sigma}_n \right)$$

among all $(T, G) \in \mathbb{R}^p \times PDS(p)$ for which $|G|=1$. The MM-estimator for the covariance matrix is $\hat{\boldsymbol{\Sigma}}_n = \hat{\sigma}_n^2 \hat{\boldsymbol{\Gamma}}_n$.

The multivariate MM-estimators were introduced by Tatsuoka and Tyler (2000) as belonging to a broad class of estimators which they call ‘multivariate M-estimators with auxiliary scale’. The idea is to estimate the scale by means of a very robust S-estimator, and then estimate the location and shape using a different ρ -function that yields better efficiency at the central model. The location and shape estimates inherit the breakdown point of the auxiliary scale and can be seen as a generalization of the regression MM-estimators of Yohai (1987). Another multivariate version of MM-estimators was proposed by Lopuhaä (1992). He uses the entire initial covariance matrix ($\tilde{\Sigma}_n$) as an auxiliary statistic, instead of just the scale ($\hat{\sigma}_n$).

In this paper we will consider loss functions in the well-known family of Tukey’s biweight functions, given by

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4}, & |x| \leq c \\ \frac{c^2}{6}, & |x| \geq c, \end{cases} \quad (1)$$

where $c > 0$ is a user-chosen tuning constant. Note that S-estimators can be seen as a special case of MM-estimators, since choosing ρ_1 equal to ρ_0 yields the initial S-estimator.

Remark : Throughout this paper, when referring to a $p \times p$ matrix G as a ‘shape matrix’ or ‘shape estimator’ we mean that G corresponds to a covariance matrix (or estimator) C through $G = |C|^{-1/p}C$. That is, G is a symmetric positive definite matrix with $|G| = 1$.

2.2 Breakdown point

Concerning the robustness properties of MM-estimators, Tatsuoka and Tyler (2000) indicate that MM-estimators for multivariate location and shape inherit the breakdown point of the initial S-estimator. The asymptotic breakdown point has been investigated by Tyler (2002). Theorem 1 below considers the finite-sample breakdown point of MM-estimators.

Following Donoho and Huber (1983), the finite-sample breakdown point of a location

estimator T_n is defined as the smallest fraction of observations of the sample \mathcal{X}_n that needs to be replaced to carry T_n beyond all bounds. Formally,

$$\epsilon_n^*(T_n, \mathcal{X}_n) = \min\left\{\frac{m}{n} : \sup_{\mathcal{X}'_n} \|T_n(\mathcal{X}_n) - T_n(\mathcal{X}'_n)\| = \infty\right\},$$

where the supremum is over all possible collections \mathcal{X}'_n that differ from \mathcal{X}_n in at most m points. The breakdown point of a covariance or shape estimator is usually defined as the smallest proportion of outliers that can carry its largest eigenvalue λ_1 over all bounds or make its smallest eigenvalue λ_p arbitrarily small. However, for a shape estimator it should be noted that $\lambda_1 \rightarrow \infty$ and $\lambda_p \rightarrow 0$ can only occur simultaneously, since its determinant is constant.

For a dataset $\mathcal{X}_n \subset \mathbb{R}^p$, let $k(\mathcal{X}_n)$ be the maximum number of observations lying on the same hyperplane of \mathbb{R}^p .

Theorem 1 *Let $\mathcal{X}_n \subset \mathbb{R}^p$. Assume $\rho_1(s) \leq \rho_0(s)$ for all $s \in \mathbb{R}$ and $\rho_1(\infty) = \rho_0(\infty)$. Denote $r := b/\rho_0(\infty)$. If $k(\mathcal{X}_n) < \lceil n - nr \rceil$ then, for MM-estimators defined in Definition 1, we have*

$$\epsilon_n^*(\hat{\boldsymbol{\mu}}_n, \hat{\Gamma}_n, \hat{\Sigma}_n; \mathcal{X}_n) \geq \epsilon_n^*(\tilde{\boldsymbol{\mu}}_n, \tilde{\Sigma}_n; \mathcal{X}_n) = \frac{1}{n} \min(\lceil nr \rceil, \lceil n - nr \rceil - k(\mathcal{X}_n)).$$

Note that the assumption $\rho_1(\infty) = \rho_0(\infty)$ in Theorem 1 is not a restriction since rescaling ρ_1 by multiplying the function with some constant has no effect on the M-estimates. Furthermore, the assumption that $\rho_1(s) \leq \rho_0(s)$ for all $s \in \mathbb{R}$ is quite natural when the second ρ -function is designed to improve the efficiency. For example, let us consider loss functions from Tukey's biweight family, then ρ_0 and ρ_1 can be chosen similarly as outlined in Remark 4.1 of Yohai (1987) for the regression setting. For instance, suppose that $p = 5$. In order to have consistency at the normal model and 50% asymptotic breakdown point, we need to set $b = 1.803$ and $c_0 = 4.652$ in ρ_0 . To obtain 95% normal shape-efficiency we should set $c_1 = 6.596$ in ρ_1 , as follows from (9) in Section 2.4. It can now easily be seen that

ρ_0 and a properly rescaled version of ρ_1 satisfy the assumptions in Theorem 1. Because in PCA the parameter space for the eigenvectors is bounded, it is not straightforward to define a breakdown point in this context (see e.g. Davies and Gather 2005), so we will not consider breakdown of PCA methods here.

2.3 Influence function

The MM-functionals for location and shape are defined analogously to the MM-estimators. That is, $(\boldsymbol{\mu}_{\text{MM}}(F), \Gamma_{\text{MM}}(F))$ minimize

$$\int \rho_1 \left([(\mathbf{x} - T)^t G^{-1} (\mathbf{x} - T)]^{\frac{1}{2}} / \sigma_S(F) \right) dF(\mathbf{x})$$

over all $T \in \mathbb{R}^p$ and $G \in \text{PDS}(p)$ with $|G| = 1$, and $\sigma_S(F)$ is the scale of the S-functional corresponding to some function ρ_0 . The MM-covariance functional is defined by $\Sigma_{\text{MM}}(F) = \sigma_S(F)^2 \Gamma_{\text{MM}}(F)$.

In the following we will focus on unimodal elliptical distributions $F_{\boldsymbol{\mu}, \Sigma}$ defined by a density of the form

$$f_{\boldsymbol{\mu}, \Sigma}(\mathbf{x}) = |\Sigma|^{-1/2} g((\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})),$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$, $\Sigma \in \text{PDS}(p)$ and g has a strictly negative derivative. When the covariance matrix of $F_{\boldsymbol{\mu}, \Sigma}$ exists, it is proportional to Σ . It equals Σ in case of the multivariate normal distribution corresponding to $g(t) = (2\pi)^{-\frac{p}{2}} e^{-\frac{t}{2}}$. Denote by Γ the shape matrix of the distribution, i.e. $\Gamma = |\Sigma|^{-1/p} \Sigma$. It is shown in Tatsuoka and Tyler (2000) that the MM-estimators for location and shape are Fisher consistent for a broad class of distributions including elliptical distributions. That is, $(\boldsymbol{\mu}_{\text{MM}}(F_{\boldsymbol{\mu}, \Sigma}), \Gamma_{\text{MM}}(F_{\boldsymbol{\mu}, \Sigma})) = (\boldsymbol{\mu}, \Gamma)$. The same is true for the covariance MM-estimator, provided that the initial covariance S-estimator is consistent through the choice $b = E_{\mathbf{0}, I}[\rho_0(\|\mathbf{x}\|)]$. Here, $E_{\mathbf{0}, I}$ denotes expectation with respect to the distribution $F_{\mathbf{0}, I}$, and $\|\cdot\|$ denotes the Euclidian norm.

The influence function of a functional measures the local robustness of the functional and the corresponding estimator. Following Hampel et al. (1986), the influence function

of a functional T at a distribution H is defined as:

$$IF(\mathbf{x}; T, H) = \lim_{\epsilon \rightarrow 0} \frac{T(H_{\epsilon, \mathbf{x}}) - T(H)}{\epsilon} = \frac{\partial}{\partial \epsilon} T(H_{\epsilon, \mathbf{x}})|_{\epsilon=0}$$

where $H_{\epsilon, \mathbf{x}} = (1 - \epsilon)H + \epsilon\Delta_{\mathbf{x}}$ and $\Delta_{\mathbf{x}}$ denotes the point mass at \mathbf{x} .

First note that the influence functions of affine equivariant covariance functionals C at elliptical distributions can be characterized by two functions $\alpha_C, \gamma_C : [0, \infty] \rightarrow \mathbb{R}$ as follows (see e.g. Lemma 1 of Croux and Haesbroeck (2000)):

$$IF(\mathbf{x}; C, F_{\boldsymbol{\mu}, \Sigma}) = \alpha_C(d(\mathbf{x}))(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t - \beta_C(d(\mathbf{x}))\Sigma \quad (2)$$

where $d^2(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$ and

$$\beta_C(d(\mathbf{x})) = \alpha_C(d(\mathbf{x})) \frac{d^2(\mathbf{x})}{p} - \gamma_C(d(\mathbf{x})).$$

For the corresponding shape functional $G = |C|^{-1/p}C$ we have that

$$IF(\mathbf{x}; G, F_{\boldsymbol{\mu}, \Sigma}) = |\Sigma|^{-1/p} \left[IF(\mathbf{x}; C, F_{\boldsymbol{\mu}, \Sigma}) - \frac{1}{p} \text{tr}(\Sigma^{-1} IF(\mathbf{x}; C, F_{\boldsymbol{\mu}, \Sigma})) \Sigma \right].$$

In this way we obtain the following general form for the influence function of a shape functional G corresponding to an affine equivariant covariance functional C :

$$IF(\mathbf{x}; G, F_{\boldsymbol{\mu}, \Sigma}) = \alpha_C(d(\mathbf{x})) |\Sigma|^{-1/p} \left((\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t - \frac{d^2(\mathbf{x})}{p} \Sigma \right). \quad (3)$$

Hence, the function α_C can be seen as describing the shape part in (2), while the function γ_C is associated with the scale part of the estimator. Kent and Tyler (1996) show that in fact the influence function of any shape component of C does not depend on γ_C . Here, a shape component is any function H such that $H(C) = H(\nu C)$ for all $\nu > 0$.

It is not difficult to see that if the initial scale estimator is consistent, the influence function of the shape MM-estimator does not depend on that initial estimator. The same holds for the location MM-estimator. Moreover, the influence functions of the MM-estimators for location and shape equal those of S-estimators of location and shape that use the function

ρ_1 . The influence function of the covariance MM-estimator turns out to be a mixture of the influence functions of S-estimators with ρ_1 and ρ_0 . Formally, it can be shown that

$$\alpha_{\text{MM}} = \alpha_{S_1} \quad \text{and} \quad \gamma_{\text{MM}} = \gamma_{S_0}. \quad (4)$$

For S-estimators S_l ($l = 0, 1$) with function ρ_l , Lopuhaä (1989) shows that

$$\begin{aligned} \alpha_{S_l}(t) &= \rho_l'(t)p/(\gamma_1 t) \\ \gamma_{S_l}(t) &= 2(\rho_l(t) - b)/\gamma_3 \end{aligned}$$

where $b = E_{\mathbf{0}, I}[\rho_l(\|\mathbf{x}\|)]$ and

$$\begin{aligned} \gamma_1 &= (p+2)^{-1} E_{\mathbf{0}, I}[\rho_l''(\|\mathbf{x}\|)\|\mathbf{x}\|^2 + (p+1)\rho_l'(\|\mathbf{x}\|)\|\mathbf{x}\|] \\ \gamma_3 &= E_{\mathbf{0}, I}[\rho_l'(\|\mathbf{x}\|)\|\mathbf{x}\|]. \end{aligned}$$

2.4 Asymptotic variance

The asymptotic variances of the MM-estimators for location and shape again equal those of the S-estimators with function ρ_1 . In many cases the asymptotic variance of an estimator T_n (with associated functional T) at the distribution F can be computed through its influence function as follows:

$$ASV(T, F) = E_F[\text{vec}(IF(\mathbf{x}, T, F))\text{vec}(IF(\mathbf{x}, T, F))^t] \quad (5)$$

For each affine equivariant covariance functional C , there exist scalars $\sigma_1 = \sigma_1(C)$ and $\sigma_3 = \sigma_3(C)$ such that:

$$ASV(C, F_{\boldsymbol{\mu}, \Sigma}) = \sigma_1(I + K_{pp})(\Sigma \otimes \Sigma) + \sigma_2 \text{vec}(\Sigma)\text{vec}(\Sigma)^t, \quad (6)$$

where we use the notation of Tyler (1983), and

$$\sigma_2 = -2\sigma_1/p + \sigma_3.$$

From expressions (2) and (5) it follows that

$$\sigma_1 = \frac{1}{p(p+2)} E_{\mathbf{0},I}[\alpha_C^2(\|\mathbf{x}\|)\|\mathbf{x}\|^4] \quad (7)$$

$$\sigma_3 = E_{\mathbf{0},I}[\gamma_C^2(\|\mathbf{x}\|)]. \quad (8)$$

For the corresponding shape functional G , we have

$$ASV(G, F_{\boldsymbol{\mu},\Sigma}) = \sigma_1 |\Sigma|^{-2/p} \left((I + K_{pp})(\Sigma \otimes \Sigma) - \frac{2}{p} \text{vec}(\Sigma) \text{vec}(\Sigma)^t \right), \quad (9)$$

which solely depends on the scalar σ_1 . From (4) we immediately obtain that the asymptotic variances of the MM-estimators for shape and covariance are given by

$$\sigma_1(MM) = \sigma_1(S_1) \quad \text{and} \quad \sigma_3(MM) = \sigma_3(S_0). \quad (10)$$

Hence, the asymptotic efficiency of MM-estimators of shape does not depend on the initial S -estimator. On the other hand, the efficiency of the whole covariance matrix obviously is related to both ρ_0 and ρ_1 . See the next section for some numerical efficiency results.

3 PCA based on the MM-estimator

Now suppose that Γ has distinct eigenvalues $\lambda_1 > \lambda_2 > \dots > \lambda_p > 0$ with corresponding eigenvectors v_1, v_2, \dots, v_p . The robust PCA method based on the MM-estimator essentially consists of estimating these eigenvalues and eigenvectors by the eigenvalues and eigenvectors of the MM-estimator of shape $\hat{\Gamma}_n$. In principle, performing PCA based on $\hat{\Sigma}_n$ instead of $\hat{\Gamma}_n$ would yield the same method, since $\hat{\Sigma}_n$ and $\hat{\Gamma}_n$ have the same eigenvectors and the same eigenvalue ratios. However, we prefer to work mainly with the shape matrix. This is natural when the interest lies in the eigenvalues, for example, since in this way we can avoid the potentially damaging effect of the bias of the scale component of $\hat{\Sigma}_n$.

Throughout this paper, $\lambda_j(A)$ and $v_j(A)$ will denote the j -th eigenvalue and eigenvector of the matrix A , and we will also use the notation $\hat{\lambda}_j$ and \hat{v}_j for the eigenvalues and

eigenvectors of the MM-estimator of shape. We next present results on influence functions and asymptotic efficiencies for the PCA method based on multivariate MM-estimators.

For a distribution F denote by $\lambda_{\Gamma_{\text{MM}},j}(F)$ and $v_{\Gamma_{\text{MM}},j}(F)$ the j -th eigenvalue and corresponding eigenvector of the MM-functional of shape and by $\lambda_{\Sigma_{\text{MM}},j}(F)$ the j -th eigenvalue of the MM-functional of covariance ($j = 1, \dots, p$). Note that $\lambda_{\Sigma_{\text{MM}},j}(F) = |\Sigma_{\text{MM}}(F)|^{1/p} \lambda_{\Gamma_{\text{MM}},j}(F)$. These eigenvector and eigenvalue functionals inherit the Fisher consistency of the MM-functionals which implies that

$$\lambda_{\Gamma_{\text{MM}},j}(F_{\boldsymbol{\mu},\Sigma}) = \lambda_j \quad \text{and} \quad v_{\Gamma_{\text{MM}},j}(F_{\boldsymbol{\mu},\Sigma}) = v_j.$$

For the influence function of the eigenvalues and eigenvectors the following holds:

$$\begin{aligned} IF(\mathbf{x}; \lambda_{\Gamma_{\text{MM}},j}, F_{\boldsymbol{\mu},\Sigma}) &= v_j^t IF(\mathbf{x}; \Gamma_{\text{MM}}, F_{\boldsymbol{\mu},\Sigma}) v_j \\ IF(\mathbf{x}; v_{\Gamma_{\text{MM}},j}, F_{\boldsymbol{\mu},\Sigma}) &= \sum_{k=1; k \neq j}^p \frac{1}{\lambda_j - \lambda_k} (v_k^t IF(\mathbf{x}; \Gamma_{\text{MM}}, F_{\boldsymbol{\mu},\Sigma}) v_j) v_k \end{aligned}$$

for $j = 1, \dots, p$. The expression for the eigenvalues of the covariance matrix is analogous (see e.g. Croux and Haesbroeck, 2000). Through (2)-(3) we obtain

$$IF(\mathbf{x}; \lambda_{\Gamma_{\text{MM}},j}, F_{\boldsymbol{\mu},\Sigma}) = \alpha_{\text{MM}}(d(\mathbf{x})) \left(z_j^2 |\Sigma|^{-\frac{1}{p}} - \frac{d^2(\mathbf{x})}{p} \lambda_j \right) \quad (11)$$

$$IF(\mathbf{x}; v_{\Gamma_{\text{MM}},j}, F_{\boldsymbol{\mu},\Sigma}) = \alpha_{\text{MM}}(d(\mathbf{x})) |\Sigma|^{-\frac{1}{p}} \sum_{k=1; k \neq j}^p \frac{z_k z_j}{\lambda_j - \lambda_k} v_k \quad (12)$$

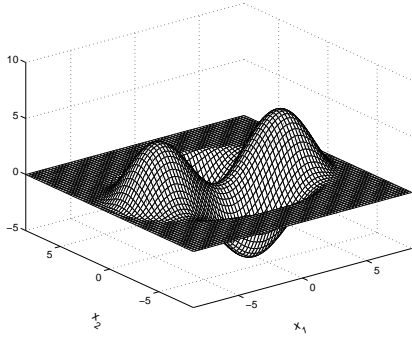
$$IF(\mathbf{x}; \lambda_{\Sigma_{\text{MM}},j}, F_{\boldsymbol{\mu},\Sigma}) = \alpha_{\text{MM}}(d(\mathbf{x})) z_j^2 - \beta_{\text{MM}}(d(\mathbf{x})) \lambda_j |\Sigma|^{\frac{1}{p}} \quad (13)$$

where $z_j = v_j^t(\mathbf{x} - \boldsymbol{\mu})$ for $j = 1, \dots, p$, and

$$\beta_{\text{MM}}(d(\mathbf{x})) = \frac{d^2(\mathbf{x})}{p} \alpha_{\text{MM}}(d(\mathbf{x})) - \gamma_{\text{MM}}(d(\mathbf{x})).$$

Figure 1a shows the influence function at the bivariate normal distribution $N(\mathbf{0}, \text{diag}(2, 1))$ for the largest eigenvalue of the shape MM-estimator. Here ρ_1 is Tukey's

(a)



(b)

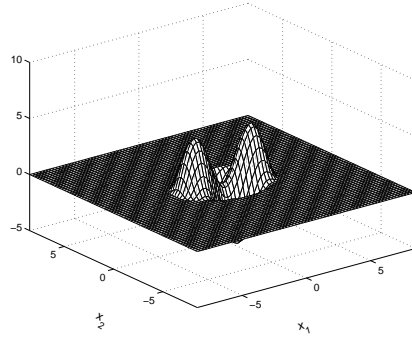


Figure 1: Influence functions for the largest eigenvalue of the MM-shape estimator at $N(\mathbf{0}, \text{diag}(2, 1))$; (a) MM-estimator (95% eff.); (b) initial S-estimator

biweight with tuning constant c_1 chosen so that the shape estimator attains 95% shape-efficiency at the normal model. For comparison, in Figure 1b we plotted the influence function corresponding to the initial S-estimator of shape, where ρ_0 was again the biweight function but now chosen to have maximal breakdown point. We see that the influence functions are smooth and vanish outside some ellipse, indicating that a single extreme outlier will not influence the eigenvalue estimators. Note that the difference between the two plots is solely due to the choice of the constant c in the ρ -functions. This choice is directly responsible for the size of the ellipse and in this way determines the efficiency of the estimator.

As in Croux and Haesbroeck (2000, Corollary 1), the asymptotic variances of the eigenvalues and eigenvectors can now be derived from (5) and (11)-(13). We obtain for $j = 1, \dots, p$

$$ASV(\lambda_{\Gamma_{\text{MM},j}}, F_{\boldsymbol{\mu}, \Sigma}) = \lambda_j^2 ASV(\Gamma_{\text{MM},11}, F_{\mathbf{0}, I}) \quad (14)$$

and

$$ASV(v_{\Gamma_{\text{MM},j}}, F_{\boldsymbol{\mu}, \Sigma}) = ASV(\Gamma_{\text{MM},12}, F_{\mathbf{0}, I}) \sum_{k=1; k \neq j}^p \frac{\lambda_j \lambda_k}{(\lambda_j - \lambda_k)^2} v_k v_k^t. \quad (15)$$

The ASV for the covariance eigenvalues is analogously related to the ASV of the MM-estimator for the covariance matrix.

Asymptotic relative efficiencies w.r.t. the classical estimators are defined as

$$ARE(\lambda_{\Gamma_{MM},j}, F_{\boldsymbol{\mu},\Sigma}) = \frac{ASV(\lambda_{\Gamma_{Cov},j}, F_{\boldsymbol{\mu},\Sigma})}{ASV(\lambda_{\Gamma_{MM},j}, F_{\boldsymbol{\mu},\Sigma})} = \frac{2 - 2/p}{ASV(\Gamma_{MM,11}, F_{\mathbf{0},I})},$$

$$ARE(v_{\Gamma_{MM},j}, F_{\boldsymbol{\mu},\Sigma}) = \left(\frac{|ASV(v_{\Gamma_{Cov},j}, F_{\boldsymbol{\mu},\Sigma})|}{|ASV(v_{\Gamma_{MM},j}, F_{\boldsymbol{\mu},\Sigma})|} \right)^{\frac{1}{p}} = \frac{1}{ASV(\Gamma_{MM,12}, F_{\mathbf{0},I})}$$

and

$$ARE(\lambda_{\Sigma_{MM},j}, F_{\boldsymbol{\mu},\Sigma}) = \frac{ASV(\lambda_{\Sigma_{Cov},j}, F_{\boldsymbol{\mu},\Sigma})}{ASV(\lambda_{\Sigma_{MM},j}, F_{\boldsymbol{\mu},\Sigma})} = \frac{2}{ASV(\Sigma_{MM,11}, F_{\mathbf{0},I})},$$

where Σ_{Cov} and Γ_{Cov} denote the classical covariance and corresponding shape matrix respectively. It can immediately be seen from (9) that the efficiency of the eigenvalues of the shape matrix equals that of the eigenvectors. In particular we have

$$ARE(\lambda_{\Gamma_{MM},j}, F_{\boldsymbol{\mu},\Sigma}) = \sigma_1^{-1}$$

$$ARE(v_{\Gamma_{MM},j}, F_{\boldsymbol{\mu},\Sigma}) = \sigma_1^{-1}$$

$$ARE(\lambda_{\Sigma_{MM},j}, F_{\boldsymbol{\mu},\Sigma}) = ((1 - 1/p)\sigma_1 + \sigma_3/2)^{-1}$$

where $\sigma_1 = \sigma_1(MM)$ and $\sigma_3 = \sigma_3(MM)$ as given in (10).

In order to obtain a highly robust and highly efficient method, one needs to choose the (Tukey biweight) ρ -functions accordingly. An obvious choice for ρ_0 is the one that corresponds to maximal breakdown point (and yields consistency at the normal model). The choice for ρ_1 is somewhat less obvious. For MM-estimators in a univariate setting, a quite natural choice would be the ρ_1 -function that yields 95% relative efficiency at the normal model for the location (or regression) estimates. In this way one obtains a considerable improvement over the S-estimator. In the multivariate situation S-estimators become more efficient when the dimension increases, see e.g. Croux and Haesbroeck (1999) and Lopuhaä (1989). Therefore, choosing ρ_1 such that the estimators attain 95% efficiency is only a good choice up to a certain dimension. Furthermore, there are three different efficiency values

Table 1: Asymptotic relative efficiencies for the 50% breakdown MM- and S-estimators at the normal model; MM designed to have 95% shape-efficiency

		$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 10$	$p = 15$	$p = 20$
$\text{Eff}_{\boldsymbol{\mu}}$	MM	0.970	0.967	0.966	0.964	0.960	0.958	0.957
	S	0.580	0.722	0.800	0.846	0.933	0.958	0.970
Eff_{Γ}	MM	0.950	0.950	0.950	0.950	0.950	0.950	0.950
	S	0.377	0.579	0.702	0.778	0.915	0.950	0.965
$\text{Eff}_{\Sigma_{11}}$	MM	0.878	0.921	0.937	0.944	0.952	0.952	0.952
	S	0.515	0.652	0.743	0.803	0.920	0.952	0.966

upon which to base the choice, that can be called location-efficiency, shape-efficiency and variance-efficiency respectively. The location-efficiency, which for MM equals that of the S-estimator using the ρ_1 -function, will be denoted by $\text{Eff}_{\boldsymbol{\mu}}$. The shape-efficiency, denoted by Eff_{Γ} , represents the efficiency of the shape matrix and its eigenvalues and eigenvectors, as well as that of the off-diagonal elements of the covariance matrix. Finally, the variance-efficiency $\text{Eff}_{\Sigma_{11}}$ corresponds to the efficiency of the diagonal elements and eigenvalues of the covariance matrix.

Table 1 lists the efficiencies at the normal model for various dimensions, for the maximal breakdown S-estimator and for the MM-estimator where ρ_1 is chosen in order to achieve 95% shape-efficiency. It can be seen that for dimensions $p \geq 15$ the MM-estimator does not yield an improvement over the initial S-estimator anymore but would even have an adverse effect. Instead of settling for 95% efficiency, one could design the MM so that it always attains, say, 99% efficiency. However, it has to be noted that, although the maximal breakdown value is assured by ρ_0 , the choice of the constant c in ρ_1 does have some effect on the robustness of the estimator. Larger values of c lead to more sensitivity to outliers and thus to larger bias. Therefore, the choice of 99% efficiency may have a cost in robustness

that is not negligible for relatively low dimensions. Another possibility is to let the choice for the efficiency depend on the dimension.

In our opinion the class of MM-estimators should be used to improve the efficiency of S-estimators when improvement is desirable, that is, in relatively low dimensions ($p < 15$). For higher dimensions, the S-estimator is highly efficient itself and there is no need to work with an additional M-estimator. In our experience, also the finite-sample performance of MM-estimates improves the performance of the initial S-estimates up to dimension 15. Finally, the difference in computational complexity is not a major factor since the extra M-step for the MM-estimator, computed by iteratively reweighted least squares, is almost negligible in comparison with the computation of the initial S-estimator. It should be noted in this matter that the results concerning the bootstrap method introduced in the next section apply to MM-estimators, but of course are also valid for S-estimators. Indeed, by choosing $\rho_1 = \rho_0$ the resulting MM-estimator would be just the initial S-estimator.

In the simulations and the example in this paper we will use the 95% shape-efficiency design, since the dimensions do not exceed $p = 15$. Note that if one would like to use the MM-estimator in other settings than PCA, it may be desirable to choose ρ_1 to achieve a certain location-efficiency instead of shape-efficiency. However, Table 1 shows that the difference between the location and shape strategy is rather small.

We finish this section with a short simulation study in which we compare four high-breakdown point estimators with the sample covariance matrix estimator (Cov). In particular, we are interested in their precision for estimating the eigenvectors of the shape matrix Γ . The robust estimators included in this experiment were the MM-estimator (95% shape-efficiency), the S-estimator with 50% breakdown, the minimum covariance determinant (MCD) estimator (Rousseeuw 1984) with 50% breakdown and its re-weighted version (RMCD), see e.g. Croux and Haesbroeck (1999).

We first generated 5000 random samples of size $n = 50$ and dimension $p = 5$ following

Table 2: Average angle (with standard error) between v_j and \hat{v}_j ; $N_p(\mathbf{0}, \Sigma_1)$; $n = 50$

	v_1	v_2	v_3	v_4	v_5
Cov	0.115 (.001)	0.161 (.001)	0.345 (.003)	0.548 (.005)	0.456 (.005)
MM	0.119 (.001)	0.166 (.001)	0.353 (.003)	0.559 (.005)	0.466 (.005)
S	0.136 (.001)	0.191 (.001)	0.400 (.004)	0.613 (.005)	0.514 (.005)
MCD	0.275 (.003)	0.422 (.003)	0.732 (.005)	0.917 (.006)	0.826 (.006)
RMCD	0.186 (.002)	0.271 (.002)	0.554 (.005)	0.791 (.006)	0.688 (.006)

a $N_p(\mathbf{0}, \Sigma_1)$ distribution, where Σ_1 is as in (27) (see Section 4.3). Table 2 shows for each estimator and for each eigenvector v_j ; $j = 1, \dots, 5$ the average angle (with standard error) in radians between v_j and its estimates (see Subsection 4.3.2). We see that the MM-estimator attains a finite-sample accuracy that is close to that of the classical estimator. As was to be expected, it yields an improvement over the initial S-estimator. The MM-estimator also outperforms both the MCD and its reweighted version.

Table 3 shows the simulation results for samples generated with 20% outliers as in Section 4.3, i.e. outliers lying in the subspace spanned by the last two eigenvectors. The classical estimator is severely misled by the outliers, leading to angles close to the maximum of $\pi/2$ while the performance order of the high breakdown estimators is preserved.

4 Bootstrap for PCA based on MM

Assuming (near) normality of the data, inference methods for the PCA model based on MM-estimators can be derived from the asymptotic results of (14) and (15). As discussed in Section 1, this may not always be appropriate and the nonparametric bootstrap may give better results in this case. Examples of bootstrap applied to classical PCA can be found

Table 3: Average angle (with standard error) between v_j and \hat{v}_j ; $N_p(\mathbf{0}, \Sigma_1)$ w/ 20% outliers; $n = 50$

	v_1	v_2	v_3	v_4	v_5
Cov	1.086 (.005)	1.477 (.001)	1.525 (.001)	1.447 (.002)	1.531 (.001)
MM	0.131 (.001)	0.184 (.001)	0.434 (.004)	0.643 (.005)	0.520 (.005)
S	0.140 (.001)	0.195 (.001)	0.435 (.004)	0.644 (.005)	0.529 (.005)
MCD	0.247 (.003)	0.366 (.003)	0.683 (.005)	0.885 (.005)	0.771 (.006)
RMCD	0.166 (.002)	0.237 (.002)	0.498 (.004)	0.731 (.005)	0.618 (.005)

in Diaconis and Efron (1983), Daudin et al. (1988), Beran and Srivastava (1985,1987) and Eaton and Tyler (1991). To overcome the problems associated with applying the classical bootstrap to robust estimators on potentially contaminated data (as explained in Section 1), we investigate an extension of the fast and robust bootstrap of Salibián-Barrera and Zamar (2002) to multivariate MM-estimators.

4.1 Fast and robust bootstrap for multivariate MM

The multivariate MM-estimators as defined in Definition 1 can be written as a system of fixed-point equations as follows:

$$\hat{\boldsymbol{\mu}}_n = \left(\sum_{i=1}^n \frac{\rho'_1(d_i/|\tilde{\Sigma}_n|^{1/(2p)})}{d_i} \right)^{-1} \left(\sum_{i=1}^n \frac{\rho'_1(d_i/|\tilde{\Sigma}_n|^{1/(2p)})}{d_i} \mathbf{x}_i \right) \quad (16)$$

$$\hat{\Gamma}_n = G \left(\sum_{i=1}^n \frac{\rho'_1(d_i/|\tilde{\Sigma}_n|^{1/(2p)})}{d_i} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_n)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_n)^t \right) \quad (17)$$

$$\tilde{\Sigma}_n = \frac{1}{nb} \left(\sum_{i=1}^n p \frac{\rho'_0(\tilde{d}_i)}{\tilde{d}_i} (\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_n)(\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_n)^t + \left(\sum_{i=1}^n \tilde{w}_i \right) \tilde{\Sigma}_n \right) \quad (18)$$

$$\tilde{\boldsymbol{\mu}}_n = \left(\sum_{i=1}^n \frac{\rho'_0(\tilde{d}_i)}{\tilde{d}_i} \right)^{-1} \left(\sum_{i=1}^n \frac{\rho'_0(\tilde{d}_i)}{\tilde{d}_i} \mathbf{x}_i \right) \quad (19)$$

where we denote $G(A) = |A|^{-1/p}A$ for $p \times p$ matrices A , and where $d_i = [(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_n)^t \hat{\Gamma}_n^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_n)]^{1/2}$, $\tilde{d}_i = [(\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_n)^t \tilde{\Sigma}_n^{-1} (\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_n)]^{1/2}$ and $\tilde{w}_i = \rho_0(\tilde{d}_i) - \rho'_0(\tilde{d}_i)\tilde{d}_i$. Such a system of equations allows us to apply the bootstrap procedure of Salibian-Barrera and Zamar (2002). The idea is to make use of the equations to compute fast approximations to the MM-estimates in each bootstrap sample. In particular, given a bootstrap sample $\mathcal{X}_n^* = \{\mathbf{x}_1^*, \dots, \mathbf{x}_n^*\}$, an intuitive way to obtain fast approximated re-calculations would be as follows:

$$\hat{\boldsymbol{\mu}}_n^* = \left(\sum_{i=1}^n \frac{\rho'_1(d_i^*/|\tilde{\Sigma}_n|^{1/(2p)})}{d_i^*} \right)^{-1} \left(\sum_{i=1}^n \frac{\rho'_1(d_i^*/|\tilde{\Sigma}_n|^{1/(2p)})}{d_i^*} \mathbf{x}_i^* \right) \quad (20)$$

$$\hat{\Gamma}_n^* = G \left(\sum_{i=1}^n \frac{\rho'_1(d_i^*/|\tilde{\Sigma}_n|^{1/(2p)})}{d_i^*} (\mathbf{x}_i^* - \hat{\boldsymbol{\mu}}_n) (\mathbf{x}_i^* - \hat{\boldsymbol{\mu}}_n)^t \right) \quad (21)$$

$$\tilde{\Sigma}_n^* = \frac{1}{nb} \left(\sum_{i=1}^n p \frac{\rho'_0(\tilde{d}_i^*)}{\tilde{d}_i^*} (\mathbf{x}_i^* - \tilde{\boldsymbol{\mu}}_n) (\mathbf{x}_i^* - \tilde{\boldsymbol{\mu}}_n)^t + \left(\sum_{i=1}^n \tilde{w}_i^* \right) \tilde{\Sigma}_n \right) \quad (22)$$

$$\tilde{\boldsymbol{\mu}}_n^* = \left(\sum_{i=1}^n \frac{\rho'_0(\tilde{d}_i^*)}{\tilde{d}_i^*} \right)^{-1} \left(\sum_{i=1}^n \frac{\rho'_0(\tilde{d}_i^*)}{\tilde{d}_i^*} \mathbf{x}_i^* \right) \quad (23)$$

where $d_i^* = [(\mathbf{x}_i^* - \hat{\boldsymbol{\mu}}_n)^t \hat{\Gamma}_n^{-1} (\mathbf{x}_i^* - \hat{\boldsymbol{\mu}}_n)]^{1/2}$, $\tilde{d}_i^* = [(\mathbf{x}_i^* - \tilde{\boldsymbol{\mu}}_n)^t \tilde{\Sigma}_n^{-1} (\mathbf{x}_i^* - \tilde{\boldsymbol{\mu}}_n)]^{1/2}$ and $\tilde{w}_i^* = \rho_0(\tilde{d}_i^*) - \rho'_0(\tilde{d}_i^*)\tilde{d}_i^*$. Note that since we are keeping the estimators $\hat{\boldsymbol{\mu}}_n$, $\hat{\Gamma}_n$, $\tilde{\Sigma}_n$ and $\tilde{\boldsymbol{\mu}}_n$ fixed on the right-hand side of (20)-(23), these approximations will likely underestimate the variability of the MM-estimator. To remedy this a linear correction can be applied as follows. Denote the equations (16)-(19) by means of a function $\mathbf{f} : \mathbb{R}^{2(p+p^2)} \rightarrow \mathbb{R}^{2(p+p^2)}$ such that

$$\mathbf{f}(\hat{\Theta}_n) = \hat{\Theta}_n$$

where $\hat{\Theta}_n := ((\hat{\boldsymbol{\mu}}_n)^t, \text{vec}(\hat{\Gamma}_n)^t, \text{vec}(\tilde{\Sigma}_n)^t, (\tilde{\boldsymbol{\mu}}_n)^t)^t$. Given the smoothness of \mathbf{f} we can calculate a Taylor expansion about the limiting value of $\hat{\Theta}_n$

$$\hat{\Theta}_n = \mathbf{f}(\Theta) + \nabla \mathbf{f}(\Theta)(\hat{\Theta}_n - \Theta) + R_n \quad (24)$$

where $\Theta = (\boldsymbol{\mu}^t, \text{vec}(\Gamma)^t, \text{vec}(\Sigma)^t, \boldsymbol{\mu}^t)^t$, R_n is the remainder term and $\nabla \mathbf{f}(\cdot)$ is the matrix of partial derivatives. If the remainder term is sufficiently small, we can rewrite (24) as

$$\sqrt{n}(\widehat{\Theta}_n - \Theta) \approx [I - \nabla \mathbf{f}(\Theta)]^{-1} \sqrt{n}(\mathbf{f}(\Theta) - \Theta). \quad (25)$$

Since both sides of the above equation are asymptotically equivalent, the distribution of the bootstrapped statistics will also converge to the same limit. Moreover, we can estimate the matrix $[I - \nabla \mathbf{f}(\Theta)]^{-1}$ by the sample version $[I - \nabla \mathbf{f}(\widehat{\Theta}_n)]^{-1}$ and obtain

$$\sqrt{n}(\widehat{\Theta}_n^* - \widehat{\Theta}_n) \approx [I - \nabla \mathbf{f}(\widehat{\Theta}_n)]^{-1} \sqrt{n}(\mathbf{f}^*(\widehat{\Theta}_n) - \widehat{\Theta}_n). \quad (26)$$

The fast and robust bootstrap is then obtained by computing the right-hand side instead of the left-hand side for each bootstrap sample. Note that $\mathbf{f}^*(\widehat{\Theta}_n)$ corresponds to (20)-(23) and that the matrix of partial derivatives provides the linear correction. Let us denote the approximations to $\widehat{\Theta}_n^*$, obtained by (26), as $\widehat{\Theta}_n^{R*} = ((\widehat{\boldsymbol{\mu}}_n^{R*})^t \text{vec}(\widehat{\Gamma}_n^{R*})^t \text{vec}(\widetilde{\Sigma}_n^{R*})^t (\widetilde{\boldsymbol{\mu}}_n^{R*})^t)^t$.

Now for bootstrapping the eigenvalues and eigenvectors of $\widehat{\Gamma}_n$, we propose to recalculate shape estimates $\widehat{\Gamma}_n^{R*}$ using the fast and robust bootstrap, and to take $\lambda_j(\widehat{\Gamma}_n^{R*})$ and $v_j(\widehat{\Gamma}_n^{R*})$ as recalculated versions of the j -th eigenvalue and eigenvector estimates.

Remark: Because of the linear correction in (26), the recalculated shape estimates $\widehat{\Gamma}_n^{R*}$ may not be positive definite and hence the bootstrapped eigenvalue estimates can be negative. In practice this seems to occur very rarely, but care must be taken, especially for small sample sizes. A simple solution is to discard those bootstrap samples where this happens. Alternatively, one can consider transformations of $\widehat{\Gamma}_n^{R*}$ such as those described in Rousseeuw and Molenberghs (1993). In our simulations we used the first approach.

To see why the fast bootstrap is more robust than the classical bootstrap method, suppose that there were outliers present in the data and that the MM-estimator was not severely affected by them. Since bootstrap samples are obtained by drawing observations at random with replacement, some of these samples may contain more outliers than the

original dataset. Moreover, the number of outliers in the bootstrap samples may exceed the breakdown point of the MM-estimator. Thus the classical bootstrap might yield *affected* re-computed estimates, such that the resulting inference is distorted by the outliers.

On the other hand, the fast bootstrap method is effectively as stable as the estimator itself. If the MM-estimator was able to resist the outlying observations, then the latter are associated with large robust distances (d_i and \tilde{d}_i) and thus receive a small or zero weight in the fast bootstrap calculations according to (20)-(23). If ρ_0 and ρ_1 satisfy (R1) and (R2), we have

$$\frac{\rho_1'(d)}{d} \xrightarrow{d \rightarrow \infty} 0 \quad \text{and} \quad \frac{\rho_0'(d)}{d} \xrightarrow{d \rightarrow \infty} 0.$$

Furthermore, these weights vanish for d outside some bounded interval. As for the weights \tilde{w}_i , it can be seen that

$$\rho_0(d) - \rho_0'(d)d \xrightarrow{d \rightarrow \infty} \rho_0(c).$$

Hence, the influence of harmful outlying observations is limited, regardless of the number in which they appear in the bootstrap sample. An illustration of the gain in robustness of the fast bootstrap method over the classical bootstrap can be found in the example in Section 5 (see Figure 4).

4.2 Consistency of the fast and robust bootstrap

In this section we show that, given the consistency of the estimators for some underlying distribution F , the fast and robust bootstrap distribution converges to the same limiting distribution as the distribution of the MM-estimator does. This consistency will first be shown for the estimator $\hat{\Gamma}_n$, after which the property for the eigenvalues and eigenvectors follows fairly easily.

Both ρ_0 and ρ_1 will need to satisfy the following regularity conditions:

(A.1) The following functions are bounded and almost everywhere continuous:

$$\frac{\rho'(x)}{x}, \frac{\rho''(x)}{x^2} - \frac{\rho'(x)}{x^3}, \frac{\rho'''(x)}{x^3} - 3\frac{\rho''(x)}{x^4} + 3\frac{\rho'(x)}{x^5}, \rho''(x) \text{ and } \frac{\rho'''(x)}{x}$$

(A.2) $E_F[\frac{\rho'(d)}{d}] \neq 0$.

Some additional conditions are needed for the function ρ_1 :

(A.1a) The function $\rho_1'''(x)x$ is bounded and almost everywhere continuous.

(A.2a) $E_F[\frac{\rho_1'(d)}{d}(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^t]^{-1}$ exists.

Remark : Tukey's biweight satisfies (A.1) and (A.1a). Assumptions (A.2) and (A.2a) depend on the central distribution F and are satisfied for elliptical distributions.

The following theorem proves the consistency of the fast and robust bootstrap for multivariate MM-estimators.

Theorem 2 *Let ρ_0 and ρ_1 be real functions defined as before and assume that (A.1) and (A.1a) are satisfied. Let $(\hat{\boldsymbol{\mu}}_n, \hat{\Gamma}_n)$ be the corresponding multivariate MM-estimators and $(\tilde{\boldsymbol{\mu}}_n, \tilde{\Sigma}_n)$ the initial S-estimators. Assume that $\tilde{\boldsymbol{\mu}}_n \xrightarrow{P} \boldsymbol{\mu}$, $\tilde{\Sigma}_n \xrightarrow{P} \Sigma$, $\hat{\boldsymbol{\mu}}_n \xrightarrow{P} \boldsymbol{\mu}$ and $\hat{\Gamma}_n \xrightarrow{P} \Gamma$. Then, given assumptions (A.2) and (A.2a) are satisfied, conditional on the first n observations, along all sample sequences the distributions of $\sqrt{n}(\hat{\boldsymbol{\mu}}_n^{R*} - \hat{\boldsymbol{\mu}}_n)$ and $\sqrt{n}(\hat{\Gamma}_n^{R*} - \hat{\Gamma}_n)$ converge weakly to the same limit distributions as those of $\sqrt{n}(\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu})$ and $\sqrt{n}(\hat{\Gamma}_n - \Gamma)$ respectively.*

Bickel and Freedman (1981) show that the bootstrap commutes with smooth functions. We can use this to prove the consistency of bootstrapping the eigenvalues of $\hat{\Gamma}_n$. Here we need the restriction to simple eigenvalues since otherwise we do not have the necessary smoothness conditions.

Theorem 3 *Let λ_Γ be a simple eigenvalue of Γ , with normalized eigenvector v_Γ . Then in some neighborhood $N(\Gamma)$ of Γ there exists a real-valued function λ and vector function v such that $\lambda(\Gamma) = \lambda_\Gamma$ and $v(\Gamma) = v_\Gamma$, as well as $Cv = \lambda v, v^t v = 1$ for all $C \in N(\Gamma)$. With the assumptions from the previous theorem we have that the distributions of $\sqrt{n}(\lambda(\widehat{\Gamma}_n^{R*}) - \lambda(\widehat{\Gamma}_n))$ and $\sqrt{n}(v(\widehat{\Gamma}_n^{R*}) - v(\widehat{\Gamma}_n))$ converge weakly to the same limit distributions as those of $\sqrt{n}(\lambda(\widehat{\Gamma}_n) - \lambda(\Gamma))$ and $\sqrt{n}(v(\widehat{\Gamma}_n) - v(\Gamma))$ respectively.*

4.3 Applications for PCA: a simulation study

In this section we consider three applications of the bootstrap as a tool for inference and to assess the stability of PCA: constructing confidence intervals for the eigenvalues of the shape matrix (Subsection 4.3.1); estimating the distribution of the angles between the eigenvector estimators and their population versions (Subsection 4.3.2); and constructing confidence intervals for the percentage of variance explained by the first k principal components ($k = 1, \dots, p - 1$) (Subsection 4.3.3).

For each application we present the results of a simulation study where we investigated the finite-sample performance of the fast and robust bootstrap. A comparison to the classical bootstrap is not included in the simulations due to the high computational cost of the latter.

The study involved samples of sizes $n = 50, 100$ and 200 in $p = 5$ dimensions. The samples were constructed by randomly drawing observations from a multivariate normal distribution $N_p(\mathbf{0}, \Sigma_1)$ with

$$\Sigma_1 = \begin{pmatrix} 1 & .8 & .6 & .4 & .2 \\ .8 & 1 & .8 & .6 & .4 \\ .6 & .8 & 1 & .8 & .6 \\ .4 & .6 & .8 & 1 & .8 \\ .2 & .4 & .6 & .8 & 1 \end{pmatrix}. \quad (27)$$

The eigenvalues of the corresponding shape matrix $\Gamma_1 = |\Sigma_1|^{-1/5}\Sigma_1$ are [7.92, 2.41, 0.59, 0.35, 0.25]. Next we introduced contamination by replacing a specified percentage of the observations by outliers which were randomly scattered within the subspace spanned by the last two eigenvectors of Σ_1 . The direction of the outliers within the subspace was uniform and the distance to the bulk of the data was chosen such that they severely affect classical PCA (as seen in Table 3) but can not be identified by univariate methods. We considered four degrees of contamination: 0%, 10%, 20% and 30% outliers. In this paper we only present results for the cases of 0% and 20% of outliers. Simulation results for 10% and 30% outliers were very similar. For each combination of n and proportion of outliers, we generated $m = 1000$ samples. For each of these samples we computed the MM-estimator with 50% breakdown and 95% shape-efficiency, and subsequently performed the robust bootstrap with $B = 1000$ recalculations. As a competitor for the robust bootstrap, we also computed an empirical version of the asymptotic variance of the estimators, which can be used to construct confidence intervals based on the corresponding normal approximation. Note that this empirical asymptotic variance (EASV) approach cannot be used to assess the stability of the eigenvector estimates.

4.3.1 Eigenvalues

A first straightforward application consists of assessing the variability of the MM-estimates of the eigenvalues. In particular the bootstrap can estimate the variance of the eigenvalues of the shape MM-estimator or can be used to construct confidence intervals for the eigenvalues of the shape matrix Γ . In our simulation study we investigated bootstrap confidence intervals constructed using the bias corrected and accelerated (BCA) method (see e.g. Davison and Hinkley, 1997, page 202). The nominal confidence level was 95%. The intervals based on the EASV and the asymptotic normal approximation are of the form

$$\left[\frac{\hat{\lambda}_j}{1 + \Phi^{-1}(0.975)\sqrt{\hat{V}_\Gamma/(n-1)}}, \frac{\hat{\lambda}_j}{1 - \Phi^{-1}(0.975)\sqrt{\hat{V}_\Gamma/(n-1)}} \right]$$

Table 4: Coverage and length of 95% confidence intervals for the eigenvalues of Γ ; normal data

n	λ_1	λ_2	λ_3	λ_4	λ_5
	Bootstrap				
50	96.2 (6.584)	95.4 (2.021)	91.9 (0.431)	89.5 (0.242)	86.5 (0.130)
100	94.7 (4.360)	96.0 (1.324)	92.1 (0.304)	89.9 (0.170)	91.3 (0.114)
200	95.6 (2.996)	96.1 (0.904)	95.9 (0.215)	92.4 (0.123)	93.5 (0.089)
	EASV				
50	91.5 (7.286)	93.3 (2.152)	94.5 (0.526)	97.5 (0.302)	94.9 (0.186)
100	92.4 (4.533)	96.0 (1.355)	94.3 (0.337)	97.9 (0.196)	96.8 (0.132)
200	94.3 (3.048)	95.7 (0.915)	95.8 (0.224)	96.8 (0.133)	97.1 (0.093)

where \widehat{V}_Γ denotes the empirical version of $ASV(\Gamma_{MM,11}, F_{\mathbf{0},I}) = (2 - 2/p)\sigma_1$.

Table 4 contains the observed coverage levels (among the $m = 1000$ intervals) for each eigenvalue (from largest to smallest) when the data do not contain outliers. Similarly, Table 5 displays the results obtained with the contaminated data. Between brackets we also report the average lengths of the intervals.

First note that the bootstrap coverage levels for the first two eigenvalues are better than for the other eigenvalues, particularly the last one. This is due to the difference in magnitude of the eigenvalues of Γ_1 . Indeed, reasonably large ratios between successive eigenvalues are needed in order to have good estimates and accurate bootstrap inference results. In our case we have that λ_1/λ_2 and λ_2/λ_3 are quite large while the other ratios are comparatively small. Also note that the intervals based on EASV have better coverage than the bootstrap intervals for the last eigenvalues, but perform worse for the first eigenvalues. This seems to be related to the bias of the eigenvalue estimates. Indeed, the asymmetric form of the EASV intervals is favorable for the smallest eigenvalues since they have a

Table 5: Coverage and length of 95% confidence intervals for the eigenvalues of Γ ; 20% outliers

n	λ_1	λ_2	λ_3	λ_4	λ_5
Bootstrap					
50	94.4 (7.100)	92.8 (2.172)	90.3 (0.458)	88.5 (0.271)	82.8 (0.135)
100	93.8 (4.751)	93.8 (1.432)	90.9 (0.320)	86.8 (0.189)	88.2 (0.120)
200	93.3 (3.232)	94.2 (0.987)	93.6 (0.231)	89.5 (0.139)	87.2 (0.097)
EASV					
50	92.5 (8.090)	92.7 (2.386)	92.9 (0.601)	97.0 (0.349)	95.0 (0.211)
100	91.9 (5.053)	93.9 (1.499)	95.7 (0.371)	96.6 (0.223)	95.4 (0.147)
200	94.4 (3.316)	96.0 (1.004)	96.2 (0.247)	95.0 (0.150)	94.6 (0.104)

negative bias, but works adversely for the positively biased largest eigenvalues.

Finally, note that there is little difference between the contaminated and ‘clean’ data sets (Tables 5 and 4 respectively). The performance of the bootstrap is again good for the larger eigenvalues and somewhat less satisfactory for the other eigenvalues. Also note that the empirical asymptotic variance is not severely affected by the outliers. This is probably due to the weights in (7) which are small for outlying observations. The lengths of the intervals are larger than those for the data without outliers, however not dramatically.

Overall the fast bootstrap intervals for the eigenvalues of the shape matrix have good coverage and are reasonably short, yet do not always outperform the EASV intervals in the situations considered here. The performance depends on the ratios of the population eigenvalues.

4.3.2 Eigenvector angles

Often the interest lies in the eigenvectors or principal components, rather than in the eigenvalues. A performance measure for an estimator in this case is the angle of the estimated eigenvector with respect to the population one. For example, estimated eigenvectors relatively aligned with their population counterparts provide valuable information regarding the principal directions of the underlying distribution. Eigenvector estimators that can be almost orthogonal to the true eigenvector are less reliable. We can assess the variability of the principal component estimates by looking at the bootstrap distribution of the angles that the recalculated eigenvectors have with the originally estimated eigenvector. The angle between the normalized eigenvectors \hat{v}_j^* and \hat{v}_j is given by $\text{acos}(|\hat{v}_j^t \hat{v}_j^*|) \in [0, \pi/2]$. The bootstrap distribution of these angles is then an estimate of the distribution of the angles $\text{acos}(|v_j^t \hat{v}_j|)$ between the eigenvector estimator and the population eigenvector.

For each simulated sample and for each eigenvector v_j , we computed the mean angle between the B bootstrap estimates $v_j(\hat{\Gamma}_n^{R*})$ and the original MM-estimate $v_j(\hat{\Gamma}_n)$. The average and standard deviation of the $m = 1000$ values are displayed in Table 6 for the normal data and in Table 7 for the data with 20% outliers. The average is compared to the Monte Carlo estimate of the mean angle between $v_j(\hat{\Gamma}_n)$ and $v_j(\Gamma_1)$, based on the same m simulated samples.

Concerning the bootstrap performance, we mostly have a slight overestimation of the variability. Nevertheless, the bootstrap estimates of the mean angles seem to be quite efficient and undeterred by the outliers. They also appear to become more accurate as the sample size grows, as expected.

4.3.3 Percentage of variance explained by first k components

Consider the following statistics which are of considerable importance in PCA:

$$\hat{p}_k = \frac{\sum_{j=1}^k \hat{\lambda}_j}{\sum_{j=1}^p \hat{\lambda}_j}, \quad \text{for } k = 1, \dots, p-1.$$

They estimate the percentage of variance explained by the first k principal components. Often the statistics $\hat{p}_k; k = 1, \dots, p-1$ are used to determine how many principal components should be used in further analyses. Therefore, it is of interest to measure the variability and stability of \hat{p}_k . Again, the bootstrap can provide an estimate of the variance or standard error of these statistics, or can be used to construct confidence intervals for the ‘true’ percentages of variance explained. Furthermore, instead of using the point estimates \hat{p}_k for choosing the number of components, one might use the lower limit of a confidence interval.

Table 6: Average bootstrap estimates (with standard deviations) of the mean angle between MM-eigenvectors and distribution eigenvectors; normal data

n		v_1	v_2	v_3	v_4	v_5
50	Monte Carlo	0.121	0.169	0.369	0.567	0.462
	Bootstrap	0.134	0.193	0.395	0.566	0.434
	(s.d.)	(0.041)	(0.035)	(0.124)	(0.133)	(0.148)
100	Monte Carlo	0.084	0.115	0.235	0.382	0.327
	Bootstrap	0.087	0.124	0.270	0.428	0.341
	(s.d.)	(0.016)	(0.013)	(0.100)	(0.125)	(0.139)
200	Monte Carlo	0.057	0.080	0.166	0.258	0.221
	Bootstrap	0.059	0.083	0.176	0.292	0.245
	(s.d.)	(0.007)	(0.005)	(0.051)	(0.091)	(0.102)

Table 7: Average bootstrap estimates (with standard deviations) of the mean angle between MM-eigenvectors and distribution eigenvectors; 20% outliers

n		v_1	v_2	v_3	v_4	v_5
50	Monte Carlo	0.135	0.188	0.431	0.651	0.529
	Bootstrap	0.149	0.220	0.443	0.618	0.471
	(s.d.)	(0.054)	(0.051)	(0.135)	(0.144)	(0.157)
100	Monte Carlo	0.090	0.126	0.282	0.466	0.394
	Bootstrap	0.095	0.136	0.313	0.480	0.373
	(s.d.)	(0.020)	(0.017)	(0.120)	(0.135)	(0.141)
200	Monte Carlo	0.063	0.088	0.179	0.310	0.274
	Bootstrap	0.065	0.092	0.204	0.343	0.282
	(s.d.)	(0.010)	(0.007)	(0.070)	(0.113)	(0.125)

Here we compare the fast and robust bootstrap confidence intervals to intervals based on the EASV. For the EASV approach, it can be shown that the asymptotic variance of the statistic \widehat{p}_k at the distribution $F_{\mu, \Sigma}$ is given by

$$\frac{ASV(\Gamma_{MM,11}, F_{\mathbf{0}, I})}{(\sum_{j=1}^p \lambda_j)^2} \left((1 - p_k)^2 \sum_{j=1}^k \lambda_j^2 + p_k^2 \sum_{j=k+1}^p \lambda_j^2 \right) \quad (28)$$

where $p_k = \sum_{j=1}^k \lambda_j / \sum_{j=1}^p \lambda_j$. The EASV intervals are then constructed using the empirical version of (28) and the usual normal approximation.

Table 8 shows the coverage and the average length of the 95% confidence intervals for the normal data. Note that the true values, corresponding to Γ_1 , are given by [.687 .897 .947 .978]. The coverage percentages for the bootstrap are very close to the nominal value, even for small samples, except for the case $n = 50$ and $k = 4$. The asymptotic variance intervals are unduly short, resulting in poor coverage. Apparently larger samples are required for

Table 8: Coverage and length of 95% confidence intervals for percentage of variance explained by the first k principal components; normal data

n	$k = 1$	$k = 2$	$k = 3$	$k = 4$
	Bootstrap			
50	94.7 (0.220)	96.8 (0.085)	95.0 (0.047)	90.6 (0.020)
100	94.1 (0.155)	95.4 (0.058)	95.3 (0.032)	93.4 (0.016)
200	96.0 (0.108)	95.4 (0.039)	95.3 (0.022)	94.8 (0.011)
	EASV			
50	90.5 (0.189)	90.2 (0.067)	83.9 (0.035)	73.9 (0.016)
100	90.9 (0.134)	91.3 (0.048)	88.5 (0.026)	84.6 (0.013)
200	91.8 (0.095)	90.8 (0.034)	90.1 (0.019)	90.0 (0.009)

the asymptotic variance to produce accurate inference results for \hat{p}_k . Table 9 displays the results for the contaminated data (20% of outliers). The performance for the bootstrap is almost as good as in the case of normal data. The intervals are slightly longer now, but the robustness of the fast bootstrap method can be clearly seen. In this case the EASV intervals have also a smaller coverage level than the bootstrap ones.

Finally, note that we considered just one specific outlier configuration in our simulation study. However, it is clear that the performance of the robust bootstrap is likely to be similar for other configurations, provided that the MM-estimator itself is able to identify the outliers.

Table 9: Coverage and length of 95% confidence intervals for percentage of variance explained by the first k principal components; 20% outliers

n	$k = 1$	$k = 2$	$k = 3$	$k = 4$
	Bootstrap			
50	94.2 (0.237)	94.4 (0.124)	93.1 (0.079)	89.7 (0.045)
100	94.2 (0.167)	94.7 (0.065)	91.2 (0.037)	90.3 (0.017)
200	93.8 (0.119)	93.4 (0.045)	91.4 (0.026)	90.3 (0.013)
	EASV			
50	90.4 (0.206)	89.7 (0.075)	86.5 (0.040)	75.7 (0.018)
100	89.5 (0.146)	89.3 (0.053)	87.8 (0.029)	85.2 (0.014)
200	90.2 (0.104)	91.6 (0.038)	90.6 (0.021)	89.0 (0.011)

5 Example

For an illustration on a real data example we consider the measurements on $n = 100$ forged old Swiss 1000 franc bills, which are part of the ‘Swiss bank notes data’ from Flury and Riedwyl (1988). The data consist of $p = 6$ variables corresponding to length, height and other distance measurements on the bills. We applied PCA based on the 50% breakdown MM-estimator with 95% shape-efficiency. Let us first take a look at the diagnostic plot in Figure 2 resulting from the PCA. Here we plotted for each observation its overall empirical influence for the MM-eigenvectors versus its robust distance based on the MM-location and covariance estimates as proposed by Pison and Van Aelst (2004). It can be seen that the MM-estimator detected a group of 15 outlying observations (with large robust distances) and that most of these appear to be highly influential points for the eigenvectors.

The MM-estimates for the eigenvalues of the shape matrix are given by $\hat{\lambda}=[10.25, 1.94,$

1.05, 0.51, 0.39, 0.24]. The weights in the first PC are $\hat{v}_1 = [-0.070 \ 0.028 \ -0.019 \ 0.813 \ -0.569 \ -0.094]$. Hence, the first principal component can be interpreted as the difference between the 4th and the 5th variable. These correspond to the distance from the inner frame on the bill to respectively the lower border and the upper border.

Figure 3 shows histograms for the weights in the first PC, obtained by performing fast and robust bootstrap with $B = 1000$. This is another application of bootstrap for PCA, not yet considered in the previous section. Some care is needed here, since the coefficients of the principal components are not uniquely defined. In order to obtain meaningful inference results, we imposed that, for each $\hat{v}_j; j = 1, \dots, 6$, the coefficient with the largest absolute value should be positive in every bootstrap recalculation, as well as in the MM-estimate \hat{v}_j itself. The bootstrap result in Figure 3 indicates that the coefficients of \hat{v}_1 , i.e. the weights associated with the first PC, are quite stable. Note that the bootstrap can also be used to construct confidence intervals for the weights and to determine which original components contribute significantly to a PC.

As an alternative way of assessing the stability of the first PC, we look at the bootstrap

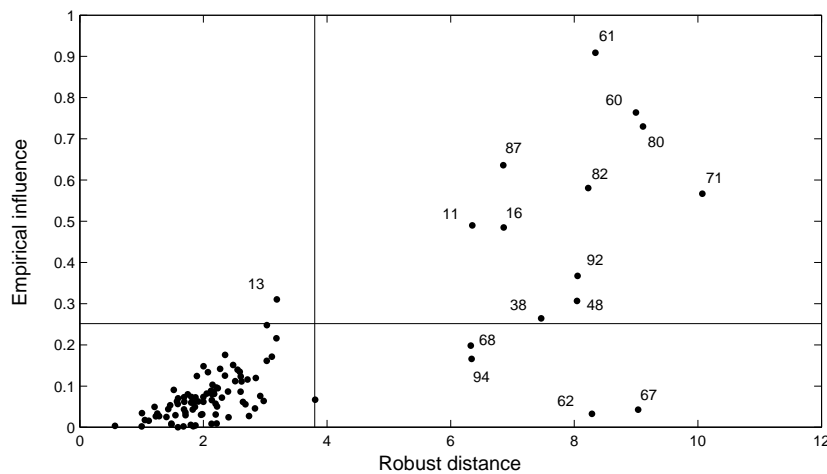


Figure 2: Swiss bank notes data: diagnostic plot: overall empirical influence for the eigenvectors versus the robust distance based on MM

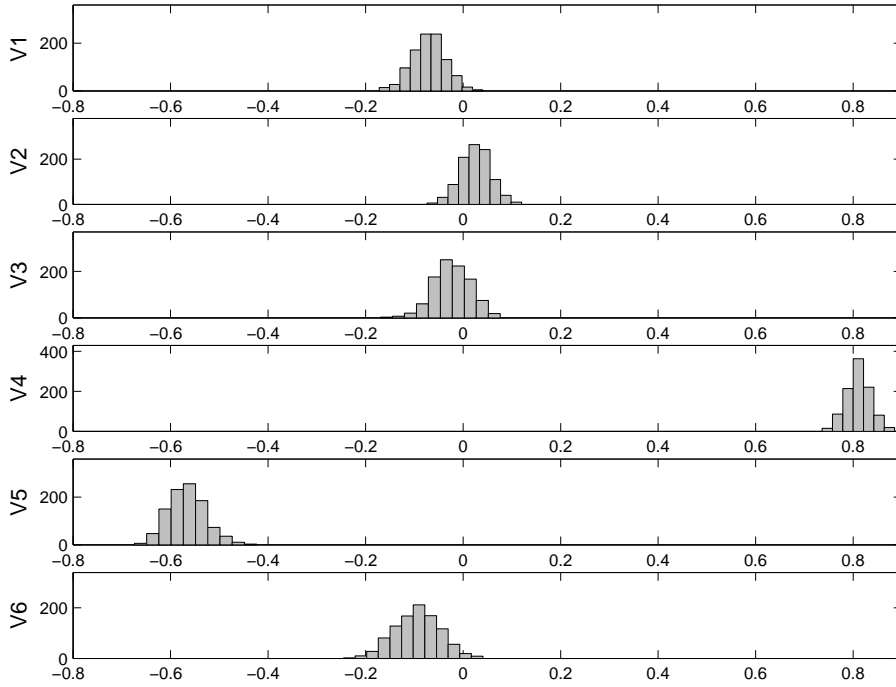


Figure 3: Swiss bank notes data: fast bootstrap histograms for the weights in the first principal component

distribution of the angles $\arccos(|\hat{v}_1^t \hat{v}_1^*|)$. The upper panels in Figure 4 compare the result from the classical bootstrap (left) with that from the fast bootstrap (right).

While the fast bootstrap recalculations yield angles no larger than about 0.2, the classical bootstrap suggests a somewhat higher variability. This is in fact an illustration of the instability of the classical bootstrap. Indeed, it turns out that the bootstrap samples corresponding to the larger angles all contained more than 15 replications from the original group of 15 outliers. This is shown in the lower panels in Figure 4, where we plotted for each bootstrap sample the angle between \hat{v}_1 and \hat{v}_1^* versus the total number of replications of the 15 outliers in that bootstrap sample. The left panel again corresponds to the classical bootstrap, the right to the fast and robust bootstrap. Clearly, in the classical procedure higher proportions of outlying observations give rise to a larger variability concerning the

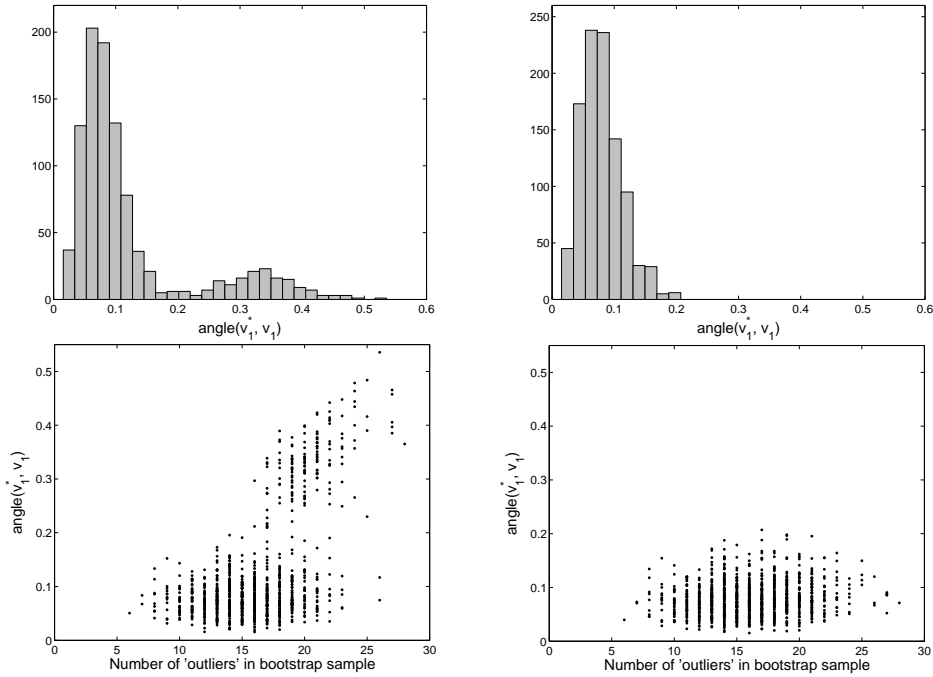


Figure 4: Swiss bank notes data: Upper: bootstrap histograms for the angles between \hat{v}_1^* and \hat{v}_1 ; Lower: angles between \hat{v}_1^* and \hat{v}_1 versus number of observations from the outlier group present in the bootstrap sample; Left: classical bootstrap; Right: fast bootstrap

eigenvector angles. On the other hand, in the fast bootstrap procedure the group of 15 outliers is severely downweighted, due to the large robust distances, and hence has very little influence on the recalculated eigenvectors.

We now turn to the question of how many components should be retained. The MM-estimates yield $\hat{p}_1 = 71.3\%$, $\hat{p}_2 = 84.8\%$, $\hat{p}_3 = 92.1\%$, $\hat{p}_4 = 95.6\%$ and $\hat{p}_5 = 98.3\%$. We used the fast bootstrap to construct 95% confidence intervals for these percentages. The left panel of Figure 5 shows the bootstrap intervals (BCA) and as a comparison also gives the intervals based on the empirical version of the asymptotic variance. The classical bootstrap intervals, which are not shown on the plot, were very similar to the fast bootstrap intervals. The dotted vertical lines indicate the point estimates. The difference between both methods here is rather small, although the intervals based on asymptotic normality are slightly more

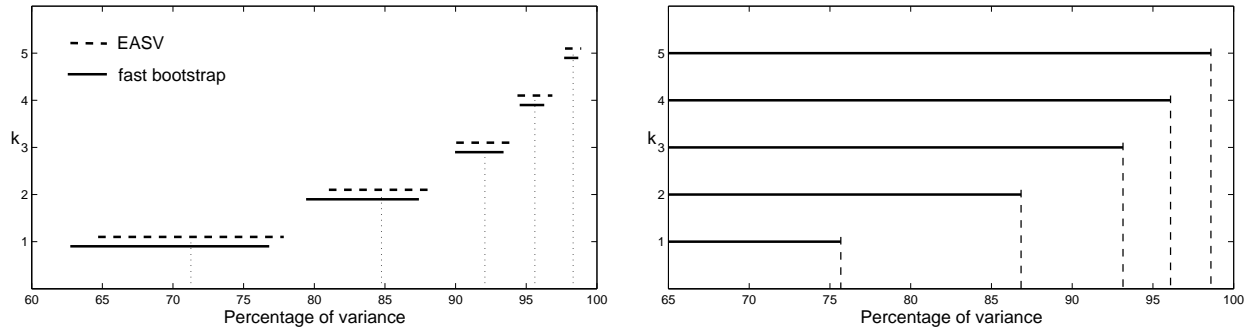


Figure 5: Swiss bank notes data. Left: 95% confidence intervals for the percentage of variance explained by the first k principal components; fast bootstrap (solid) compared to intervals based on asymptotic normality (dashed). Right: 95% one-sided confidence intervals for the percentage of variance explained by the first k principal components

optimistic. Note that the simulation study demonstrated that bootstrap intervals usually have better coverage. The intervals in general are somewhat long, although there is no overlap present. Concerning the number of PCs to retain, suppose we adopt the criterion of choosing the smallest k for which the percentages p_k exceeds some cutoff value. Then, if the cutoff is set at 70%, we might decide to play it safe and take two components into account, instead of just the first. Indeed, the point estimate \hat{p}_1 equals 71.3%, but the lower limit of the confidence interval is as low as 62.5%. If the cutoff is set at 80%, the choice for $k = 2$ would be obvious since both the point estimate and almost the whole interval exceeds 80%. In case of a 90% cutoff, the choice of $k = 3$ based on \hat{p}_3 is confirmed by the corresponding interval.

Finally, note that we could apply the duality between hypothesis tests and confidence intervals to test null hypotheses of the form $H_0 : p_k \geq \pi$. Here, π would then equal e.g. 80% or 90%. We can decide to accept H_0 at the level α whenever π is contained in a $(1 - \alpha)100\%$ one-sided confidence interval for p_k . As an example we plotted in the right panel of Figure 5 the 95% one-sided intervals for p_k based on the fast bootstrap, where we

again used the BCA method. On this plot we can immediately see which hypotheses would be accepted on the 5% significance level. E.g. $H_0 : p_k \geq 80\%$ is accepted for $k \geq 2$, while $H_0 : p_k \geq 90\%$ is accepted for $k \geq 3$.

Remark: The robustness of the fast bootstrap has been illustrated in this section, and was demonstrated in the simulation study as well. We would like to stress that its speed is also a very important feature. As an example, the computation of the classical bootstrap on the Swiss bank notes data took about 20 minutes (on a Pentium IV, 1.9 GHz), while the fast bootstrap only took a few seconds. Software code in MATLAB and R/SPLUS to compute robust PCA based on multivariate MM-estimates and the fast, robust bootstrap together with a detailed description of its use is available at <http://users.ugent.be/~svaelst/software/MMPCAboot.html>.

6 Conclusion

In this paper we considered robust PCA based on multivariate MM-estimators of shape. We first studied the breakdown point, influence function and asymptotic variance of MM-estimators and their eigenvalues and eigenvectors and found that MM-estimators have very good efficiency and robustness properties. We also investigated an adaptation of the fast and robust bootstrap method of Salibián-Barrera and Zamar (2002). We showed the consistency of this bootstrap procedure and performed a simulation study to investigate its finite-sample properties. The results of the simulations indicate that the method works well and generally slightly outperforms the approach based on asymptotic normality. However, one should be aware of the intrinsic difficulties in bootstrapping a principal component analysis. That is, care must be taken when interpreting bootstrap results in case eigenvalue ratios are small (see e.g. Jolliffe, 2002, page 49). On the other hand, the bootstrap seems to offer more possibilities than the asymptotic approach. The method was illustrated with application to a real data set.

Appendix

Proof of Theorem 1: As can be found in (Van Aelst and Willems, 2005), for the S-estimators we have that

$$\epsilon_n^*(\tilde{\boldsymbol{\mu}}_n, \tilde{\Sigma}_n; \mathcal{X}_n) = \min(\lceil nr \rceil, \lceil n - nr \rceil - k(\mathcal{X}_n))/n. \quad (29)$$

Suppose \mathcal{X}'_n is a dataset obtained from \mathcal{X}_n by replacing m observations, where $m < \min(\lceil nr \rceil, \lceil n - nr \rceil - k(\mathcal{X}_n))$. Denote $(\hat{\boldsymbol{\mu}}'_n, \hat{\Gamma}'_n) := (\hat{\boldsymbol{\mu}}_n(\mathcal{X}'_n), \hat{\Gamma}_n(\mathcal{X}'_n))$ and analogously for the S-estimators. Further abbreviate $[(\mathbf{x} - T)^t C^{-1}(\mathbf{x} - T)]^{1/2}$ by $d(\mathbf{x}; T, C)$. We have the following inequalities:

$$\begin{aligned} \frac{1}{n} \sum_{\mathbf{x}_i \in \mathcal{X}'_n} \rho_1 \left(d(\mathbf{x}_i; \hat{\boldsymbol{\mu}}'_n, \hat{\Gamma}'_n) / \hat{\sigma}'_n \right) &\leq \frac{1}{n} \sum_{\mathbf{x}_i \in \mathcal{X}'_n} \rho_1 \left(d(\mathbf{x}_i; \tilde{\boldsymbol{\mu}}'_n, \tilde{\Sigma}'_n) \right) \\ &\leq \frac{1}{n} \sum_{\mathbf{x}_i \in \mathcal{X}'_n} \rho_0 \left(d(\mathbf{x}_i; \tilde{\boldsymbol{\mu}}'_n, \tilde{\Sigma}'_n) \right) = b = r\rho_0(\infty) \end{aligned}$$

The first of these inequalities holds since $(\hat{\boldsymbol{\mu}}'_n, \hat{\Gamma}'_n)$ form the solution to the M-minimization. The second is due to the assumptions on the ρ -functions. Since $\rho_1(\infty) = \rho_0(\infty)$ we furthermore obtain

$$\frac{1}{n} \sum_{\mathbf{x}_i \in \mathcal{X}'_n} \rho_1 \left([(\mathbf{x}_i - \hat{\boldsymbol{\mu}}'_n)^t \hat{\Gamma}'_n^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}'_n)]^{1/2} / \hat{\sigma}'_n \right) \leq r\rho_1(\infty).$$

It follows that there exists some $d_1 < \infty$, not depending on \mathcal{X}'_n , such that $(\mathbf{x}_i - \hat{\boldsymbol{\mu}}'_n)^t \hat{\Gamma}'_n^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}'_n) / \hat{\sigma}'_n{}^2 < d_1$ for at least $\lceil n - nr \rceil$ points in \mathcal{X}'_n . We know that the S-scale $\hat{\sigma}'_n$ is bounded for $m < \min(\lceil nr \rceil, \lceil n - nr \rceil - k(\mathcal{X}_n))$, hence there also exists some $d_2 < \infty$ such that $(\mathbf{x}_i - \hat{\boldsymbol{\mu}}'_n)^t \hat{\Gamma}'_n^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}'_n) < d_2$ for the same $\lceil n - nr \rceil$ points. Now since $m < \lceil n - nr \rceil - k(\mathcal{X}_n)$, at least $k(\mathcal{X}_n) + 1$ of these points are points in \mathcal{X}_n , and they are not lying on a hyperplane. It follows that the smallest eigenvalue of $\hat{\Gamma}'_n$ must be bounded away from zero, and since its determinant is constant, its largest eigenvalue is bounded from above. It immediately follows that $\|\hat{\boldsymbol{\mu}}'_n\|$ must be bounded as well. Furthermore, the eigenvalues of $\hat{\Sigma}'_n$ are bounded

since $\widehat{\sigma}'_n$ and the eigenvalues of $\widehat{\Gamma}'_n$ are bounded. \square

We need the following lemma to prove Theorem 2.

Lemma 1 *Let $Y_1, \dots, Y_n \sim F$ be a sequence of i.i.d. random vectors. Let (\mathbf{t}_n, S_n) be consistent estimators for $(\boldsymbol{\mu}, \Sigma)$. Let $\kappa : \mathbb{R} \rightarrow \mathbb{R}$ be a function that is bounded and almost everywhere continuous. If $\tilde{\kappa}(Y, \mathbf{t}, S) = \kappa((Y - \mathbf{t})^t S^{-1} (Y - \mathbf{t}))$, then*

$$\frac{1}{n} \sum_{i=1}^n \tilde{\kappa}(Y_i, \mathbf{t}_n, S_n) \xrightarrow{P} E_F[\tilde{\kappa}(Y, \boldsymbol{\mu}, \Sigma)].$$

Proof: The proof is based on an argument used in Davies (1987, proof of Theorem 3).

Denote $\tilde{\kappa}_n(y) := \tilde{\kappa}(y, \mathbf{t}_n, S_n)$ and $\tilde{\kappa}(y) := \tilde{\kappa}(y, \boldsymbol{\mu}, \Sigma)$. For any y such that κ is continuous at $(y - \boldsymbol{\mu})^t \Sigma^{-1} (y - \boldsymbol{\mu})$, and for any sequence $(y_n)_n$ such that $y_n \rightarrow y$, we have that

$$\tilde{\kappa}_n(y_n) \xrightarrow[n \rightarrow \infty]{} \tilde{\kappa}(y). \quad (30)$$

Since κ is almost everywhere continuous, the convergence in (30) holds for almost all y .

Hence we can apply Theorem 5.5 of Billingsley (1968). Define $\gamma : \mathbb{R} \rightarrow \mathbb{R}$ by $\gamma(y) = y$ if $\inf \kappa \leq y \leq \sup \kappa$, $\gamma(y) = \sup \kappa$ if $y \geq \sup \kappa$ and $\gamma(y) = \inf \kappa$ if $y \leq \inf \kappa$. Let F_n be the empirical distribution of Y_1, \dots, Y_n , we then obtain from the theorem that

$$\int \gamma(\tilde{\kappa}_n(y)) dF_n \longrightarrow \int \gamma(\tilde{\kappa}(y)) dF$$

since γ is bounded and uniformly continuous. \square

Proof of Theorem 2: We mostly follow the lines of Salibian-Barrera and Zamar (2002).

We can write the estimating equations (16)-(19) as follows:

$$\widehat{\boldsymbol{\mu}}_n = a_n(\widehat{\boldsymbol{\mu}}_n, \widehat{\Gamma}_n, \widetilde{\Sigma}_n)^{-1} \mathbf{B}_n(\widehat{\boldsymbol{\mu}}_n, \widehat{\Gamma}_n, \widetilde{\Sigma}_n) \quad (31)$$

$$\widehat{\Gamma}_n = |\mathbf{V}_n(\widehat{\boldsymbol{\mu}}_n, \widehat{\Gamma}_n, \widetilde{\Sigma}_n)|^{-1/p} \mathbf{V}_n(\widehat{\boldsymbol{\mu}}_n, \widehat{\Gamma}_n, \widetilde{\Sigma}_n) \quad (32)$$

$$\widetilde{\Sigma}_n = \widetilde{\mathbf{V}}_n(\widetilde{\boldsymbol{\mu}}_n, \widetilde{\Sigma}_n) + \widetilde{w}_n(\widetilde{\boldsymbol{\mu}}_n, \widetilde{\Sigma}_n) \widetilde{\Sigma}_n \quad (33)$$

$$\widetilde{\boldsymbol{\mu}}_n = \widetilde{a}_n(\widetilde{\boldsymbol{\mu}}_n, \widetilde{\Sigma}_n)^{-1} \widetilde{\mathbf{B}}_n(\widetilde{\boldsymbol{\mu}}_n, \widetilde{\Sigma}_n) \quad (34)$$

with properly defined functions $a_n, \mathbf{B}_n, \mathbf{V}_n, \tilde{\mathbf{V}}_n, \tilde{w}_n, \tilde{a}_n$ and $\tilde{\mathbf{B}}_n$.

Consider the function $\mathbf{f} : \mathbb{R}^{2(p+p^2)} \rightarrow \mathbb{R}^{2(p+p^2)}$, for $M, M_0 \in \mathbb{R}^p$ and $G, C \in \mathbb{R}^{p \times p}$:

$$\mathbf{f} \begin{pmatrix} M \\ \text{vec}(G) \\ \text{vec}(C) \\ M_0 \end{pmatrix} := \begin{pmatrix} a_n(M, G, C)^{-1} \mathbf{B}_n(M, G, C) \\ \text{vec}(|\mathbf{V}_n(M, G, C)|^{-1/p} \mathbf{V}_n(M, G, C)) \\ \text{vec}(\tilde{\mathbf{V}}_n(M_0, C) + \tilde{w}_n(M_0, C)C) \\ \tilde{a}_n(M_0, C)^{-1} \tilde{\mathbf{B}}_n(M_0, C) \end{pmatrix}$$

Let $\hat{\Theta}_n := ((\hat{\boldsymbol{\mu}}_n)^t \text{vec}(\hat{\Gamma}_n)^t \text{vec}(\hat{\Sigma}_n)^t (\hat{\boldsymbol{\mu}}_n)^t)^t$. We have that $\mathbf{f}(\hat{\Theta}_n) = \hat{\Theta}_n$. Since ρ_0 and ρ_1 are sufficiently smooth, the function \mathbf{f} allows a Taylor expansion around $\Theta := (\boldsymbol{\mu}^t \text{vec}(\Gamma)^t \text{vec}(\Sigma)^t \boldsymbol{\mu}^t)^t$:

$$\hat{\Theta}_n = \mathbf{f}(\Theta) + \nabla \mathbf{f}(\Theta)(\hat{\Theta}_n - \Theta) + \frac{1}{2}(I \otimes (\hat{\Theta}_n - \Theta)^t) \mathbf{Hf}(\tilde{\Theta}_n)(\hat{\Theta}_n - \Theta) \quad (35)$$

Here $\nabla \mathbf{f}(\cdot) \in \mathbb{R}^{2(p+p^2) \times 2(p+p^2)}$ is the Jacobian and $\mathbf{Hf}(\cdot) \in \mathbb{R}^{4(p+p^2)^2 \times 2(p+p^2)}$ is the Hessian matrix of \mathbf{f} . The value of $\tilde{\Theta}_n$ in the remainder term lies between $\hat{\Theta}_n$ and Θ . The Hessian is obtained by taking the partial derivatives of the entries of the Jacobian, the matrix of the partial derivatives of \mathbf{f} . Straightforward calculations then yield that each entry in the Hessian is a combination of products of means. Taking into account that the derivatives of ρ_0 and ρ_1 vanish outside some interval (due to (R1) and (R2)), assumptions (A.1) and (A.1a) ensure the existence of the population analogues of these means. Furthermore, Lemma 1 then guarantees that $\|\mathbf{Hf}(\tilde{\Theta}_n)\| = O_p(1)$.

From the consistency of the estimators we also have that $\|\hat{\Theta}_n - \Theta\| = O_p(n^{-1/2})$. It follows that the remainder term is $o_p(n^{-1/2})$.

We can now rewrite (35) as follows:

$$\sqrt{n}(\hat{\Theta}_n - \Theta) = [I - \nabla \mathbf{f}(\Theta)]^{-1} \sqrt{n}(\mathbf{f}(\Theta) - \Theta) + o_p(1). \quad (36)$$

It needs to be shown that the bootstrap distribution of the right-hand side of this equation converges to the asymptotic distribution of $\sqrt{n}(\hat{\Theta}_n - \Theta)$. For any X_n, Y_n , by $X_n \sim Y_n$ we

denote that X_n and Y_n have the same limiting distribution. We have

$$\sqrt{n}(\widehat{\Theta}_n - \Theta) \sim [I - \nabla \mathbf{f}(\Theta)]^{-1} \sqrt{n}(\mathbf{f}(\Theta) - \Theta). \quad (37)$$

Define the function $\mathbf{g} : \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^{p \times p} \times \mathbb{R}^{p \times p} \times \mathbb{R}^{p \times p} \times \mathbb{R} \times \mathbb{R}^p \rightarrow \mathbb{R}^{2(p+p^2)}$ by

$$\mathbf{g}(a, B, V, \widetilde{V}, \widetilde{W}, \widetilde{a}, \widetilde{B}) = (a^{-1}B^t, \text{vec}(|V|^{-1/p}V)^t, \text{vec}(\widetilde{V} + \widetilde{W})^t, \widetilde{a}^{-1}\widetilde{B}^t)^t.$$

Denote $T = (M^t \text{vec}(G)^t \text{vec}(C)^t M_0^t)^t$ and

$$\bar{Y}_n(T) := (a_n(\cdot), \mathbf{B}_n(\cdot), \mathbf{V}_n(\cdot), \widetilde{\mathbf{V}}_n(\cdot), \widetilde{\mathbf{W}}_n(\cdot), \widetilde{a}_n(\cdot), \widetilde{\mathbf{B}}_n(\cdot))$$

where $\widetilde{\mathbf{W}}_n(\cdot) := \widetilde{w}_n(\cdot)C$. Note that the components are actually means. Furthermore, denote by $\boldsymbol{\mu}_{Y(T)}$ the limiting values of these means. We then have $\mathbf{g}(\bar{Y}_n(T)) = \mathbf{f}(T)$ for any T and also $\mathbf{g}(\boldsymbol{\mu}_{Y(\Theta)}) = \Theta$.

The regularity conditions on ρ_0 and ρ_1 ensure that \mathbf{g} is smooth, so that we can proceed as follows (see Bickel and Freedman, 1981):

$$\sqrt{n}(\mathbf{g}(\bar{Y}_n(\Theta)) - \mathbf{g}(\boldsymbol{\mu}_{Y(\Theta)})) \sim \nabla \mathbf{g}(\boldsymbol{\mu}_{Y(\Theta)}) \sqrt{n}(\bar{Y}_n(\Theta) - \boldsymbol{\mu}_{Y(\Theta)}) \quad (38)$$

as well as

$$\sqrt{n}(\mathbf{g}(\bar{Y}_n^*(\widehat{\Theta}_n)) - \mathbf{g}(\bar{Y}_n(\widehat{\Theta}_n))) \sim \nabla \mathbf{g}(\boldsymbol{\mu}_{Y(\widehat{\Theta}_n)}) \sqrt{n}(\bar{Y}_n^*(\widehat{\Theta}_n) - \bar{Y}_n(\widehat{\Theta}_n)) \quad (39)$$

(We omit the writing of the vec-operator here for convenience.)

Theorem 2.2 of Bickel and Freedman (1981) yields

$$\sqrt{n}(\bar{Y}_n(\Theta) - \boldsymbol{\mu}_{Y(\Theta)}) \sim \sqrt{n}(\bar{Y}_n^*(\Theta) - \bar{Y}_n(\Theta)).$$

Using the consistency of $\widehat{\Theta}_n$ it can be shown, as in (Salibian-Barrera, 2000), that

$$\sqrt{n}(\bar{Y}_n^*(\Theta) - \bar{Y}_n(\Theta)) \sim \sqrt{n}(\bar{Y}_n^*(\widehat{\Theta}_n) - \bar{Y}_n(\widehat{\Theta}_n)).$$

Once more using the consistency of $\widehat{\Theta}_n$, it then follows from (38) and (39) that

$$\sqrt{n}(\mathbf{g}(\bar{Y}_n(\Theta)) - \mathbf{g}(\boldsymbol{\mu}_{Y(\Theta)})) \sim \sqrt{n}(\mathbf{g}(\bar{Y}_n^*(\widehat{\Theta}_n)) - \mathbf{g}(\bar{Y}_n(\widehat{\Theta}_n)))$$

and thus, noting that $\mathbf{f}(\widehat{\Theta}_n) = \widehat{\Theta}_n$,

$$\sqrt{n}(\mathbf{f}(\Theta) - \Theta) \sim \sqrt{n}(\mathbf{f}^*(\widehat{\Theta}_n) - \widehat{\Theta}_n). \quad (40)$$

Finally, since $[I - \nabla \mathbf{f}(\widehat{\Theta}_n)]^{-1}$ is a consistent estimate of $[I - \nabla \mathbf{f}(\Theta)]^{-1}$, we get from (37) and (40):

$$\sqrt{n}(\widehat{\Theta}_n - \Theta) \sim [I - \nabla \mathbf{f}(\widehat{\Theta}_n)]^{-1} \sqrt{n}(\mathbf{f}^*(\widehat{\Theta}_n) - \widehat{\Theta}_n)$$

The right-hand side is actually $(\sqrt{n}(\widehat{\boldsymbol{\mu}}_n^{R*} - \widehat{\boldsymbol{\mu}}_n)^t \text{vec}(\sqrt{n}(\widehat{\Gamma}_n^{R*} - \widehat{\Gamma}_n))^t \text{vec}(\sqrt{n}(\widetilde{\Sigma}_n^{R*} - \widetilde{\Sigma}_n))^t \sqrt{n}(\widetilde{\boldsymbol{\mu}}_n^{R*} - \widetilde{\boldsymbol{\mu}}_n)^t)^t$, and the proof is complete. \square

Proof of Theorem 3: The proof of the first part can be found in Magnus and Neudecker (1999, Theorem 7, Chapter 8). They furthermore show that in the neighborhood $N(\Gamma) \subset \mathbb{R}^{p \times p}$ the functions λ and v are ∞ times differentiable. Hence we have the necessary smoothness conditions for applying Lemma 8.10 of Bickel and Freedman (1981) to the functions λ and v . The convergence of the bootstrap distributions then follows from this lemma, together with Theorem 2. \square

References

- Beran, R., and Srivastava, M.S. (1985), "Bootstrap Tests and Confidence Regions for Functions of a Covariance Matrix," *The Annals of Statistics*, 13, 95-115.
- Beran, R., and Srivastava, M.S. (1987), "Correction: Bootstrap Tests and Confidence Regions for Functions of a Covariance Matrix," *The Annals of Statistics*, 15, 470-471.
- Bickel, P.J., and Freedman, D.A. (1981), "Some Asymptotic Theory for the Bootstrap," *The Annals of Statistics*, 9, 1196-1217.
- Billingsley, P. (1968), *Convergence of Probability Measures*, New York: Wiley.
- Campbell, N.A. (1980), "Robust Procedures in Multivariate Analysis: Robust Covariance Estimation," *Applied Statistics*, 29, 231-237.

- Croux, C., and Haesbroeck, G. (1999), “Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator,” *Journal of Multivariate Analysis*, 71, 161-190.
- Croux, C., and Haesbroeck, G. (2000), “Principal Components Analysis Based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies,” *Biometrika*, 87, 603-618.
- Croux, C., and Ruiz-Gazen, A. (2005), “High Breakdown Estimators for Principal Components: the Projection-Pursuit Approach Revisited,” *Journal of Multivariate Analysis*, 95, 206-226.
- Davies, P.L. (1987), “Asymptotic Behavior of S-Estimates of Multivariate Location Parameters and Dispersion Matrices,” *The Annals of Statistics*, 15, 1269-1292.
- Davies, P.L., and Gather, U. (2005), “Breakdown and Groups” (with discussion), *The Annals of Statistics*, 33, 977-1016.
- Davison, A.C., and Hinkley, D.V. (1997), *Bootstrap Methods and their Application*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.
- Daudin, J.J., Duby, C., and Trecourt, P. (1988), “Stability of Principal Component Analysis Studied by the Bootstrap Method,” *Statistics*, 19, 241-258.
- Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1981), “Robust Estimation of Dispersion Matrices and Principal Components,” *Journal of the American Statistical Association*, 76, 354-362.
- Diaconis, P., and Efron, B. (1983), “Computer-Intensive Methods in Statistics,” *Scientific American*, 248, 96-108.
- Donoho, D.L., and Huber, P.J. (1983), “The Notion of Breakdown Point,” in *A Festschrift for Erich Lehmann*, eds. P. Bickel, K. Doksum, and J.L. Hodges, Jr., Belmont, CA: Wadsworth, pp. 157-184.
- Eaton, M., and Tyler, D.E. (1991), “On Wielandt’s Inequality and Its Application to

- the Asymptotic Distribution of the Eigenvalues of a Random Symmetric Matrix,” *The Annals of Statistics*, 19, 260-271.
- Efron, B. (1979), “Bootstrap Methods: Another Look at the Jackknife,” *The Annals of Statistics*, 7, 1-26.
- Flury, B., and Riedwyl, H. (1988), *Multivariate Statistics: A Practical Approach*, Cambridge University Press.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: Wiley.
- Hubert, M., Rousseeuw, P.J., and Verboven, S. (2002), “A Fast Method for Robust Principal Components with Applications to Chemometrics,” *Chemometrics and Intelligent Laboratory Systems*, 60, 101-111.
- Jolliffe, I.T. (2002), *Principal Component Analysis* (2nd ed.), New York: Springer-Verlag.
- Kent, J.T., and Tyler, D.E. (1996), “Constrained M-Estimation for Multivariate Location and Scatter,” *The Annals of Statistics*, 24, 1346-1370.
- Li, G., and Chen, Z. (1985), “Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo,” *Journal of the American Statistical Association*, 80, 759-766.
- Lopuhaä, H. (1989), “On the Relation Between S-Estimators and M-Estimators of Multivariate Location and Covariance,” *The Annals of Statistics*, 17, 1662-1683.
- Lopuhaä, H. (1992), “Highly Efficient Estimators of Multivariate Location with High Breakdown Point,” *The Annals of Statistics*, 20, 398-413.
- Magnus, J.R., and Neudecker, H. (1999), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, New York: Wiley.
- Maronna, R.A. (1976), “Robust M-Estimators of Multivariate Location and Scatter,” *The Annals of Statistics*, 4, 51-67.
- Pison, G., and Van Aelst, S. (2004), “Diagnostic Plots for Robust Multivariate Methods,”

- Journal of Computational and Graphical Statistics*, 13, 310-329.
- Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871-880.
- Rousseeuw, P.J., and Molenberghs, G. (1993), "Transformation of Non Positive Semidefinite Correlation Matrices," *Communications in Statistics, Part A - Theory and Methods*, 22, 965-984.
- Salibian-Barrera, M. (2000), "Contributions to the Theory of Robust Inference," Ph.D. thesis, Dept. Statist., Univ. British Columbia, Vancouver.
- Salibian-Barrera, M., and Zamar, R. (2002), "Bootstrapping Robust Estimates of Regression," *The Annals of Statistics*, 30, 556-582.
- Singh, K. (1998), "Breakdown Theory for Bootstrap Quantiles," *The Annals of Statistics*, 26, 1719-1732.
- Stromberg, A.J. (1997), "Robust Covariance Estimates Based on Resampling," *Journal of Statistical Planning and Inference*, 57, 321-334.
- Tatsuoka, K.S., and Tyler, D.E. (2000), "The Uniqueness of S and M-Functionals Under Non-Elliptical Distributions," *The Annals of Statistics*, 28, 1219-1243.
- Tyler, D.E. (1983), "Robustness and Efficiency Properties of Scatter Matrices," *Biometrika*, 70, 411-420.
- Tyler, D.E. (2002), "High-Breakdown Point Multivariate M-Estimation," *Estadística*, 54, 213-247.
- Van Aelst, S., and Willems, G. (2005), "Multivariate Regression S-Estimators for Robust Estimation and Inference," *Statistica Sinica*, 15, 981-1001.
- Yohai, V.J. (1987), "High Breakdown-Point and High Efficiency Robust Estimates for Regression," *The Annals of Statistics*, 15, 642-656.