Principal Components Analysis Based on Multivariate MM Estimators With Fast and Robust Bootstrap

Matías SALIBIÁN-BARRERA, Stefan VAN AELST, and Gert WILLEMS

We consider robust principal components analysis (PCA) based on multivariate MM estimators. We first study the robustness and efficiency of these estimators, particularly in terms of eigenvalues and eigenvectors. We then 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 PCA and the bootstrap inference on a real dataset.

KEY WORDS: Bootstrap; Inference; MM-estimators; Principal components; Robustness.

1. INTRODUCTION

Principal components analysis (PCA), a very common technique in multivariate statistics, aims to explain the covariance structure of the data by projecting the observations 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 that has the largest variance, whereas the eigenvalue itself measures the amount of information that can be explained by that projection. The sample covariance matrix is notorious for being sensitive to outliers, however. Consequently, several robustifications of classical PCA have been proposed in the literature, including projection-pursuit approaches (Li and Chen 1985; Croux and Ruiz-Gazen 2005; Hubert, Rousseeuw, and Verboven 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, Gnanadesikan, and Kettenring (1981). More recently, Croux and Haesbroeck (2000) compared several robust estimators in a more formal manner. In this article we consider a related type of PCA based on robust estimates of shape. In particular, we 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 interested primarily 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 estimate standard errors (see, e.g., Croux and Haesbroeck 2000). But these results hold only under the assumption of some underlying elliptical distribution, and 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, because the bootstrap can estimate the sampling distribution of the estimator of interest, it offers a wider range of inference applications and enables assessment of the stability of the PCA results. However, applying the 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, thus can entail high computational costs. 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, Salibian-Barrera and Zamar (2002) proposed a fast and robust bootstrap method for univariate regression MM estimators. Here we investigate an adaptation of their method to the multivariate location and shape setting. This extension can be used to obtain many recalculation of the MM shape or scatter matrix. As with the classical bootstrap, we base our inference on the eigenvalues and eigenvectors of these recomputed matrices.

The rest of the article is organized as follows. In Section 2 we derive asymptotic and robustness properties of the multivariate MM estimators. In Section 3 we discuss some theoretical aspects of PCA based on the MM estimates. In Section 4, devoted to the fast and robust bootstrap, 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 Sec-

© 2006 American Statistical Association
Journal of the American Statistical Association
DOI 10.1198/016214506000000096
tion 5 we give a real data example. We provide some concluding remarks in Section 6, and give all of the proofs in an 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 require the following regularity conditions for a loss function $\rho$:

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

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

Multivariate MM estimators of location, shape, and covariance can now be defined as follows.

Definition 1. Let $X_n = \{x_1, \ldots, x_n\} \subset \mathbb{R}^p$ with $n \geq p + 1$. Let $\rho_0$ and $\rho_1$ satisfy (R1) and (R2) and let $(\widehat{\mu}_n, \widehat{\Sigma}_n)$ be multivariate S estimators; that is, $(\widehat{\mu}_n, \widehat{\Sigma}_n)$ minimize $|C|$ subject to

$$\frac{1}{n} \sum_{i=1}^{n} \rho_0\left(\left((x_i - T)^tC^{-1}(x_i - T)\right)^{1/2}/\sigma_S(F)\right) = b$$

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 $\widehat{\sigma}_n := |\Sigma_n|^{1/2}/\rho_0$. Then the multivariate MM estimators for location and shape $(\widehat{\mu}_n, \widehat{\Gamma}_n)$ minimize

$$\frac{1}{n} \sum_{i=1}^{n} \rho_1\left(\left((x_i - T)^tC^{-1}(x_i - T)\right)^{1/2}/\sigma_n\right)$$

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

The multivariate MM estimators were introduced by Tatsuoka and Tyler (2000) as belonging to a broad class of estimators that they call “multivariate M estimators with auxiliary scale.” The idea is to estimate the scale by means of a very robust S estimator, then estimate the location and shape using a different $\rho$-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), who used the entire initial covariance matrix $(\Sigma_n)$ as an auxiliary statistic, instead of just the scale $(\sigma_n)$.

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

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

where $c > 0$ is a user-chosen tuning constant. Note that S estimators can be seen as a special case of MM estimators, because choosing $\rho_1$ equal to $\rho_0$ yields the initial S estimator.

Remark 1. Throughout this article, 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 $G = |C|^{-1/2}C$. 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) indicated 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). Our Theorem 1 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 $X_n$ that needs to be replaced to carry $T_n$ beyond all bounds. Formally, this can be written as

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

where the supremum is over all possible collections $X_n'$ that differ from $X_n$ 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 $\lambda_1$ over all bounds or make its smallest eigenvalue $\lambda_p$ arbitrarily small. However, for a shape estimator, $\lambda_1 \to \infty$ and $\lambda_p \to 0$ can occur only simultaneously, because the determinant is constant.

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

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

$$\epsilon_n^*(\widehat{\mu}_n, \widehat{\Gamma}_n, X_n) \geq \epsilon_n^*(\widehat{\mu}_n, \widehat{\Sigma}_n, X_n) \geq \frac{1}{n} \min\left\{ \lfloor nr \rfloor, [n - nr] - k(X_n) \right\}.$$  

Note that the assumption $\rho_1(\infty) = \rho_0(\infty)$ in Theorem 1 is not a restriction, because rescaling $\rho_1$ 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 $\rho$ function is designed to improve the efficiency. Consider, for example, loss functions from Tukey’s biweight family; then $\rho_0$ and $\rho_1$ can be chosen as outlined in remark 4.1 of Yohai (1987) for the regression setting. For instance, suppose that $p = 5$. 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 $\rho_0$. To obtain 95% normal shape efficiency, we should set $c_1 = 6.596$ in $\rho_1$, as follows from (9) in Section 2.4. It can now easily be seen that $\rho_0$ and a properly rescaled version of $\rho_1$ satisfy the assumptions in Theorem 1. Because in PCA the parameter space for the eigenvectors is bounded, defining a breakdown point in this context is not straightforward (see, e.g., Davies and Gather 2005), so we do 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, $(\mu_{MM}(F), \Gamma_{MM}(F))$ minimize

$$\int \rho_1\left(\left((x - T)^tC^{-1}(x - T)\right)^{1/2}/\sigma_S(F)\right) dF(x)$$
over all \( T \in \mathbb{R}^p \) and \( G \in \text{PDS}(p) \) with \(|G| = 1\), and \( \sigma_\gamma(F) \) is the scale of the \( S \) functional corresponding to some function \( \rho_0 \). The MM covariance functional is defined by \( \Sigma_{\text{MM}}(F) = \sigma_\gamma(F)^2 \Gamma_{\text{MM}}(F) \).

In what follows we focus on unimodal elliptical distributions, \( \mu, \Sigma \), defined by a density of the form

\[
 f_{\mu, \Sigma}(x) = |\Sigma|^{-1/2} g((x - \mu)^T \Sigma^{-1}(x - \mu)),
\]

where \( \mu \in \mathbb{R}^p \), \( \Sigma \in \text{PDS}(p) \), and \( g \) has a strictly negative derivative. When the covariance matrix of \( F_{\mu, \Sigma} \) exists, it is proportional to \( \Sigma \). It equals \( \Sigma \) in the case of the multivariate normal distribution that corresponds to \( g(t) = (2\pi)^{-p/2} e^{-t/2} \).

Let \( \Gamma \) denote the shape matrix of the distribution, that is, \( \Gamma = |\Sigma|^{-1/p} \Sigma \). Tatsuoka and Tyler (2000) showed that the MM estimators for location and shape are Fisher consistent for a broad class of distributions including elliptical distributions, that is, \( (\mu_{\text{MM}}(F_{\mu, \Sigma}), \Gamma_{\text{MM}}(F_{\mu, \Sigma})) = (\mu, \Gamma) \). The same is true for the covariance MM estimator, provided that the initial covariance estimator. Kent and Tyler (1996) showed that in fact the influence function of the covariance MM estimator turns out to be a mixture of the influence functions of S estimators with \( \rho_1 \) and \( \rho_0 \). Formally, it can be shown that

\[
 \alpha_{\text{MM}} = \alpha_{S_l} \quad \text{and} \quad \gamma_{\text{MM}} = \gamma_{S_l}. \quad (4)
\]

For S estimators \( S_l \) \((l = 0, 1)\) with function \( \rho_l \), Lopuhaä (1989) showed that

\[
 \alpha_{S_l}(t) = \rho_l(t)p/(\gamma_l t)
\]

and

\[
 \gamma_{S_l}(t) = 2(\rho_l(t) - b)/\gamma_3,
\]

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

\[
 \gamma_1 = (p + 2)^{-1}E_{0,l}[\rho_l'(||x||)||x||^2 + (p + 1)\rho_l'(||x||)||x||]
\]

and

\[
 \gamma_3 = E_{0,l}[\rho_l'(||x||)||x||].
\]

2.4 Asymptotic Variance

The asymptotic variances of the MM estimators for location and shape again equal those of the \( S \) estimators with function \( \rho_1 \). 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

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

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

\[
 \text{ASV}(C, F_{\mu, \Sigma}) = \sigma_1(I + K_{pp})(\Sigma \otimes \Sigma) + \sigma_2 \text{vec}(\Sigma) \text{vec}(\Sigma)^T + \sigma_3 \text{vec}(\Sigma) \text{vec}(\Sigma)^T.
\]

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

\[
 \sigma_2 = -2\sigma_1 + \sigma_3.
\]

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

\[
 \sigma_1 = \frac{1}{p(p + 2)}E_{0,l}[\alpha_l^2(||x||)||x||^4] \quad (7)
\]

and

\[
 \sigma_3 = E_{0,l}[\gamma_l^2(||x||)]. \quad (8)
\]

For the corresponding shape functional \( G \), we have

\[
 \text{ASV}(G, F_{\mu, \Sigma}) = \sigma_1|\Sigma|^{-2/p}(I + K_{pp})(\Sigma \otimes \Sigma) - \frac{2}{p} \text{vec}(\Sigma) \text{vec}(\Sigma)^T,
\]

(9)
which depends solely on the scalar $\sigma_1$. From (4), we immediately obtain that the asymptotic variances of the MM estimators for shape and covariance are given by

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

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 $\rho_0$ and $\rho_1$. In the next section we provide some numerical efficiency results.

3. PRINCIPAL COMPONENTS ANALYSIS BASED ON THE MM ESTIMATOR

Now suppose that $\Gamma$ has distinct eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$ with corresponding eigenvectors $v_1, v_2, \ldots, v_p$. 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 $\Sigma_n$ instead of $\hat{\Gamma}_n$ would yield the same method, because $\Sigma_n$ 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, because in this way we can avoid the potentially damaging effect of the bias of the scale component of $\Sigma_n$.

Throughout this article, we let $\lambda_j(A)$ and $v_j(A)$ denote the $j$th eigenvalue and eigenvector of the matrix $A$, and 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 $F$, let $\lambda_{\Gamma_{\text{MM}},j}(F)$ and $v_{\Gamma_{\text{MM}},j}(F)$ denote the $j$th eigenvalue and corresponding eigenvector of the MM functional of shape and set $\lambda_{\Sigma_{\text{MM}},j}(F)$ denote the $j$th eigenvalue of the MM functional of covariance ($j = 1, \ldots, p$). Note that $\lambda_{\Sigma_{\text{MM}},j}(F) = |\Sigma_{\text{MM}}(F)|^{1/2}\lambda_{\Gamma_{\text{MM}},j}(F)$. These eigenvector and eigenvalue functionals inherit the Fisher consistency of the MM functionals, which implies that $\lambda_{\Gamma_{\text{MM}},j}(F_{\mu, \Sigma}) = \lambda_j$ and $v_{\Gamma_{\text{MM}},j}(F_{\mu, \Sigma}) = v_j$.

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

$$IF(x; \lambda_{\Gamma_{\text{MM}},j}, F_{\mu, \Sigma}) = v_j IF(x; \Gamma_{\text{MM}}, F_{\mu, \Sigma}) v_j$$

and

$$IF(x; v_{\Gamma_{\text{MM}},j}, F_{\mu, \Sigma})$$

$$= \sum_{k=1; k \neq j}^{p} \frac{1}{\lambda_j - \lambda_k} (v_j^p IF(x; \Gamma_{\text{MM}}, F_{\mu, \Sigma}) v_j) v_k,$$

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

$$IF(x; \lambda_{\Gamma_{\text{MM}},j}, F_{\mu, \Sigma})$$

$$= \alpha_{\text{MM}}(d(x)) \left( z_j^2 |\Sigma|^{-1/p} - \frac{d^2(x)}{p} \lambda_j \right),$$

$$IF(x; v_{\Gamma_{\text{MM}},j}, F_{\mu, \Sigma})$$

$$= \alpha_{\text{MM}}(d(x)) |\Sigma|^{-1/p} \sum_{k=1; k \neq j}^{p} \frac{z_k z_j}{\lambda_j - \lambda_k} v_k,$$

and

$$IF(x; \lambda_{\Sigma_{\text{MM}},j}, F_{\mu, \Sigma})$$

$$= \alpha_{\text{MM}}(d(x)) z_j^2 - \beta_{\text{MM}}(d(x)) \lambda_j |\Sigma|^{1/p},$$

where $z_j = v_j^p (x - \mu)$ for $j = 1, \ldots, p$ and

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

Figure 1(a) 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 $\rho_1$ is Tukey’s biweight with tuning constant $c_1$ chosen so that the shape estimator attains 95% shape efficiency at the normal model. For comparison, Figure 1(b) plots the influence function corresponding to the initial S estimator of shape, where $\rho_0$ again is the biweight function but is 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 due solely to the choice of the constant $c$ in the $\rho$ functions. This choice is directly responsible for the

![Figure 1](image-url)
size of the ellipse and in this way determines the efficiency of the estimator.

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

$$\text{ASV}(\lambda_{\Gamma_{\text{MM},j}}, F_{\mu, \Sigma}) = \lambda_j^2 \text{ASV}(\Gamma_{\text{MM},11}, F_{0,t})$$

(14)

and

$$\text{ASV}(\nu_{\Gamma_{\text{MM},j}}, F_{\mu, \Sigma}) = \text{ASV}(\Gamma_{\text{MM},12}, F_{0,t}) \sum_{k=1, k \neq j}^{p} \frac{\lambda_j \lambda_k}{(\lambda_j - \lambda_k)^2} v_k v_k^T.$$  

(15)

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

Asymptotic relative efficiencies (AREs) with regard to the classical estimators are defined as

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

$$\text{ARE}(\nu_{\Gamma_{\text{MM},j}}, F_{\mu, \Sigma}) = \left( \frac{\text{ASV}(\nu_{\Gamma_{\text{MM},j}}, F_{\mu, \Sigma})}{\text{ASV}(\nu_{\Gamma_{\text{MM},j}}, F_{\mu, \Sigma})} \right)^{1/p} = \frac{1}{\text{ASV}(\Gamma_{\text{MM},12}, F_{0,t})},$$

and

$$\text{ARE}(\lambda_{\Sigma_{\text{MM},j}}, F_{\mu, \Sigma}) = \frac{\text{ASV}(\lambda_{\Sigma_{\text{MM},j}}, F_{\mu, \Sigma})}{\text{ASV}(\lambda_{\Sigma_{\text{MM},j}}, F_{\mu, \Sigma})} = \frac{2}{\text{ASV}(\Sigma_{\text{MM},11}, F_{0,t})},$$

where $\Sigma_{\text{Cov}}$ and $\Gamma_{\text{Cov}}$ denote the classical covariance and corresponding shape matrix. 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

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

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

and

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

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

To obtain a highly robust and highly efficient method, we need to choose the (Tukey biweight) $\rho$ functions accordingly. An obvious choice for the $\rho_0$ function is the one that corresponds to maximal breakdown point (and yields consistency at the normal model). The choice for $\rho_1$ is somewhat less obvious. For MM estimators in a univariate setting, a quite natural choice would be the $\rho_1$ function that yields 95% relative efficiency at the normal model for the location (or regression) estimates. In this way we would obtain a considerable improvement over the S estimator. In the multivariate situation, S estimators become more efficient with increasing dimension (see, e.g., Croux and Haesbroeck 1999; Lopuhaä 1989). Therefore, choosing $\rho_1$ such that the estimators attain 95% efficiency is a good choice only up to a certain dimension. Furthermore, there are three different efficiency values on which to base the choice, which can be called location efficiency, shape efficiency, and variance efficiency. The location efficiency, which for MM equals that of the S estimator using the $\rho_1$ function, is denoted by $\text{Eff}_{\mu}$. The shape efficiency, denoted by $\text{Eff}_{\gamma}$, 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 with $\rho_1$ chosen to achieve 95% shape efficiency. It can be seen that for dimensions $p \geq 5$, the MM estimator no longer yields an improvement over the initial S estimator and even has an adverse effect. Instead of settling for 95% efficiency, we could design the MM so that it always attains, say, 99% efficiency. However, it must be noted that although the maximal breakdown value is ensured by $\rho_0$, the choice of the constant $c$ in $\rho_1$ 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 because 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 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, with a choice of $\rho_1 = \rho_0$, the resulting MM estimator would be just the initial S estimator.

In the simulations and the example in this article, we use the 95% shape-efficiency design, because the dimensions do not exceed $p = 15$. Note that if we wished to use the MM estimator in other settings than PCA, then it might be desirable to choose

<table>
<thead>
<tr>
<th>$p$</th>
<th>$\text{Eff}_{\mu}$ (MM)</th>
<th>$\text{Eff}_{\gamma}$ (MM)</th>
<th>$\text{Eff}<em>{\Sigma</em>{11}}$ (MM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.970 .967 .966</td>
<td>.964 .960 .958</td>
<td>.957</td>
</tr>
<tr>
<td>3</td>
<td>.580 .722 .800</td>
<td>.846 .933 .958</td>
<td>.970</td>
</tr>
<tr>
<td>4</td>
<td>.950 .950 .950</td>
<td>.950 .950 .950</td>
<td>.950</td>
</tr>
<tr>
<td>5</td>
<td>.377 .579 .702</td>
<td>.778 .915 .950</td>
<td>.965</td>
</tr>
<tr>
<td>10</td>
<td>.878 .921 .937</td>
<td>.944 .952 .952</td>
<td>.962</td>
</tr>
<tr>
<td>15</td>
<td>.515 .652 .743</td>
<td>.803 .920 .952</td>
<td>.966</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Asymptotic Relative Efficiencies for the 50% Breakdown MM and S Estimators at the Normal Model; MM Designed to Have 95% Shape Efficiency
\(\rho_1\) to achieve a certain location efficiency instead of shape efficiency, however, as shown in Table 1, the difference between the location and shape strategies is rather small.

We end this section by reporting 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 \(\Gamma\). 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 reweighted version (RMCD) (see, e.g., Croux and Haesbroeck 1999).

We first generated 5,000 random samples of size \(n = 50\) and dimension \(p = 5\) following a \(N_p(\mu, \Sigma)\) distribution, where \(\Sigma\) is as in (27) (see Sec. 4.3). Table 2 gives, for each estimator and for each eigenvector \(v_j\), \(j = 1, \ldots, 5\), the average angle (with standard error) in radians between \(v_j\) and its estimates (see Sec. 4.3.2). We see that the MM estimator attains a finite-sample accuracy that is close to that of the classical estimator. As would be expected, it yields an improvement over the initial S estimator. The MM estimator also outperforms both the MCD and the RMCD.

Table 3 shows the simulation results for samples generated with 20% outliers as in Section 4.3, that is, 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 Principal Components Analysis Based on MM Estimators

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, however, and the nonparametric bootstrap may give better results in this case. Examples of bootstrap applied to classical PCA have been given by Diaconis and Efron (1983), Daudin, Duby, and Trecourt (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 Sec. 1), we investigated an extension of the fast and robust bootstrap of Salibian-Barrera and Zamar (2002) to multivariate MM estimators.

#### 4.1 Fast and Robust Bootstrap for Multivariate MM Estimators

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

\[
\hat{\mu}_n = \left( \sum_{i=1}^{n} \rho_i^\prime \left( \frac{d_i}{|\hat{\Sigma}_n|^{1/2}} \right) \right)^{-1} \times \left( \sum_{i=1}^{n} \rho_i^\prime \left( \frac{d_i}{|\hat{\Sigma}_n|^{1/2}} \right) \frac{x_i - \hat{\mu}_n}{\sqrt{n}} \right),
\]

(16)

\[
\hat{\Sigma}_n = G \left( \sum_{i=1}^{n} \rho_i^\prime \left( \frac{d_i}{|\hat{\Sigma}_n|^{1/2}} \right) \frac{x_i - \hat{\mu}_n}{\sqrt{n}} \frac{x_i - \hat{\mu}_n}{\sqrt{n}} \right) + \left( \sum_{i=1}^{n} \hat{w}_i \right) \hat{\Sigma}_n,
\]

(17)

and

\[
\bar{\mu}_n = \left( \sum_{i=1}^{n} \rho_i^\prime \left( \frac{d_i}{d_i^*} \right) \right)^{-1} \left( \sum_{i=1}^{n} \rho_i^\prime \left( \frac{d_i}{d_i^*} \right) \frac{x_i^* - \hat{\mu}_n}{\sqrt{n}} \right),
\]

(19)

where for \(p \times p\) matrices \(A\), \(G(A) = |A|^{-1/2} A\), and \(d_i = [(x_i - \hat{\mu}_n)\hat{\Sigma}_n^{-1}(x_i - \hat{\mu}_n)]^{1/2}\), \(\hat{d}_i = [(x_i - \hat{\mu}_n)\hat{\Sigma}_n^{-1}(x_i - \hat{\mu}_n)]^{1/2}\) and \(\hat{w}_i = \rho_0(\hat{d}_i) - \rho_0(\hat{d}_i)\hat{d}_i\). 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 \(X^*_n = \{x^*_1, \ldots, x^*_n\}\), an intuitive way to obtain fast approximated recalculations would be

\[
\bar{\mu}_n = \left( \sum_{i=1}^{n} \rho_i^\prime \left( \frac{d_i^*}{|\hat{\Sigma}_n|^{1/2}} \right) \right)^{-1} \times \left( \sum_{i=1}^{n} \rho_i^\prime \left( \frac{d_i^*}{|\hat{\Sigma}_n|^{1/2}} \right) \frac{x^*_i - \hat{\mu}_n}{\sqrt{n}} \right),
\]

(20)

\[
\hat{\Sigma}_n = G \left( \sum_{i=1}^{n} \rho_i^\prime \left( \frac{d_i^*}{|\hat{\Sigma}_n|^{1/2}} \right) \frac{x_i^* - \hat{\mu}_n}{\sqrt{n}} \frac{x_i^* - \hat{\mu}_n}{\sqrt{n}} \right) + \left( \sum_{i=1}^{n} \hat{w}_i \right) \hat{\Sigma}_n,
\]

(21)

\[
\bar{\Sigma}_n = \frac{1}{nb} \left( \sum_{i=1}^{n} \rho_0^\prime \left( \frac{\hat{d}_i}{\hat{d}_i^*} \right) \left( x^*_i - \hat{\mu}_n \right) \left( x^*_i - \hat{\mu}_n \right)^t \right) + \left( \sum_{i=1}^{n} \hat{w}_i \right) \hat{\Sigma}_n.
\]

(22)
and
\[
\hat{\mu}_n = \left( \sum_{i=1}^{n} \frac{\rho'(\tilde{d}_i^*)}{d_i^*} \right)^{-1} \left( \sum_{i=1}^{n} \frac{\rho'(\tilde{d}_i^*)}{d_i^*} x_i^* \right),
\] (23)
where \( d_i^* = [(x_i^* - \hat{\mu}_n)^T (x_i^* - \hat{\mu}_n)]^{1/2} \), \( \tilde{d}_i^* = [(x_i^* - \hat{\mu}_n)^T \Sigma_n^{-1} (x_i^* - \hat{\mu}_n)]^{1/2} \), and \( \tilde{w}_i^* = \rho_0(d_i^*) - \rho_0(\tilde{d}_i^*) d_i^* \). Note that because we are keeping the estimators \( \hat{\mu}_n, \hat{\Sigma}_n, \) and \( \tilde{\mu}_n \) fixed on the right side of (20)–(23), these approximations will likely underestimate the variability of the MM estimator. To remedy this deficiency, a linear correction can be applied as follows. Denote (16)–(19) by means of a function \( f: \mathbb{R}^{2(p+p^2)} \rightarrow \mathbb{R}^{2(p+p^2)} \) such that
\[
f(\hat{\Theta}_n) = \hat{\Theta}_n,
\]
where \( \hat{\Theta}_n := ((\hat{\mu}_n)^T, \text{vec}(\hat{\Sigma}_n)^T, \text{vec}(\Sigma_n)^T, (\hat{\mu}_n)^T)^T \). Given the smoothness of \( f \), we can calculate a Taylor expansion about the limiting value of \( \hat{\Theta}_n \),
\[
\hat{\Theta}_n = f(\Theta) + \nabla f(\Theta)(\hat{\Theta}_n - \Theta) + R_n,
\] (24)
where \( \Theta = (\mu^T, \text{vec}(\Sigma)^T, \text{vec}(\Sigma)^T, \mu^T)^T \), \( R_n \) is the remainder term, and \( \nabla f(\cdot) \) is the matrix of partial derivatives. If the remainder term is sufficiently small, then we can rewrite (24) as
\[
\sqrt{n}(\hat{\Theta}_n - \Theta) \approx [I - \nabla f(\Theta)]^{-1} \sqrt{n}(f(\Theta) - \Theta).
\] (25)
Because both sides of this equation are asymptotically equivalent, the distribution of the bootstrapped statistics will also converge to the same limit. Moreover, we can estimate the matrix \([I - \nabla f(\Theta)]^{-1} \) by the sample version \([I - \nabla f(\hat{\Theta}_n)]^{-1} \) and obtain
\[
\sqrt{n}(\hat{\Theta}_n - \Theta) \approx [I - \nabla f(\hat{\Theta}_n)]^{-1} \sqrt{n}(f(\hat{\Theta}_n) - \hat{\Theta}_n).
\] (26)
We then obtain the fast and robust bootstrap by computing the right side instead of the left side for each bootstrap sample. Note that \( \Gamma'(\hat{\Theta}_n) \) corresponds to (20)–(23) and that the matrix of partial derivatives provides the linear correction. We denote the approximations to \( \Gamma_n' \), obtained by (26) by \( \hat{\Gamma}_n = ((\hat{\mu}_n)^T \text{vec}(\hat{\Sigma}_n)^T \text{vec}(\Sigma_n)^T (\hat{\mu}_n)^T)^T \).

Now, for bootstrapping the eigenvalues and eigenvectors of \( \hat{\Gamma}_n \), we propose recalculating the shape estimates \( \hat{\mu}_n \) using the fast and robust bootstrap, and taking \( \lambda_j(\hat{\Gamma}_n) \) and \( v_j(\hat{\Gamma}_n) \) as recalculated versions of the \( j \)th eigenvalue and eigenvector estimates.

Remark 2. Because of the linear correction in (26), the recalculated shape estimates \( \hat{\mu}_n \) may not be positive definite, and hence the bootstrapped eigenvalue estimates can be negative. This seems to occur very rarely in practice, but nonetheless care must be taken, especially for small sample sizes. A simple solution is to discard those bootstrap samples for which this happens. Alternatively, we could consider transformations of \( \hat{\mu}_n \) such as those described by Rousseau and Molenberghs (1993). In our simulations, we used the first approach.

To illustrate why the fast bootstrap is more robust than the classical bootstrap method, suppose that outliers are present in the data and that the MM estimator is not severely affected by them. Because 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 reestimated 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 is 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 \( \rho_0 \) and \( \rho_1 \) satisfy (R1) and (R2), then we have
\[
\frac{\rho_1(d)}{d} \rightarrow 0 \quad \text{and} \quad \frac{\rho_0(d)}{d} \rightarrow 0.
\]
Furthermore, these weights vanish for \( d \) outside some bounded interval. As for the weights \( \tilde{w}_i \), it can be seen that
\[
\rho_0(d) - \rho_0(d) \frac{d \rightarrow \infty}{d \rightarrow \infty} \rho_0(c).
\]
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 is given in the example in Section 5 (see Fig. 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. We first show this consistency for the estimator \( \hat{\Gamma}_n \), after which the property for the eigenvalues and eigenvectors follows fairly easily.

Both \( \rho_0 \) and \( \rho_1 \) need to satisfy the following regularity conditions:

(A.1) The following functions are bounded and almost everywhere continuous:
\[
\rho'(x) \quad \rho''(x) - \rho'(x) \quad \rho'''(x) - 3 \rho''(x) + \frac{\rho'(x)}{x}.
\]

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

Some additional conditions are needed for the function \( \rho_1 \):

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

(A.2a) \( E_F\left[\frac{\rho_1(d)}{d}\right] (X - \mu) (X - \mu)^T \) exists.

Remark 3. Tukey’s biweight satisfies (A.1) and (A.1a). (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 \( \rho_0 \) and \( \rho_1 \) be real functions defined as before and assume that (A.1) and (A.1a) are satisfied. Let \( \hat{\mu}_n \) be the corresponding multivariate MM estimators and \( \hat{\Sigma}_n \) be the initial S estimators. Assume that \( \hat{\mu}_n \rightarrow \mu, \hat{\Sigma}_n \rightarrow \Sigma, \hat{\mu}_n \rightarrow \mu, \) and \( \hat{\Gamma}_n \rightarrow \Gamma \). Then, given that assumptions (A.2) and (A.2a) are satisfied, conditional on the first \( n \) observations,
along all sample sequences the distributions of $\sqrt{n}(\hat{\mu}_n^R - \mu_n)$ and $\sqrt{n}(\hat{\Gamma}_n^R - \Gamma_n)$ converge weakly to the same limit distributions as those of $\sqrt{n}(\hat{\mu}_n - \mu)$ and $\sqrt{n}(\hat{\Gamma}_n - \Gamma)$.

Bickel and Freedman (1981) showed 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, because otherwise we do not have the necessary smoothness conditions.

**Theorem 3.** Let $\lambda_1$ be a simple eigenvalue of $\Gamma$ with normalized eigenvector $v_1$. Then in some neighborhood $N(\Gamma)$ of $\Gamma$ there exists a real-valued function $\lambda$ and vector function $v$ such that $\lambda(\Gamma) = \lambda_1$ and $v(\Gamma) = v_1$, as well as $Cv = \lambda v$ and $v'v = 1$ for all $C \in N(\Gamma)$. With the assumptions from the previous theorem, we have that the distributions of $\sqrt{n}(\lambda(\hat{\Gamma}_n^R) - \lambda(\Gamma_n))$ and $\sqrt{n}(v(\hat{\Gamma}_n^R) - v(\Gamma_n))$ converge weakly to the same limit distributions as those of $\sqrt{n}(\lambda(\hat{\Gamma}_n) - \lambda(\Gamma))$ and $\sqrt{n}(v(\hat{\Gamma}_n) - v(\Gamma))$.

### 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 (Sec. 4.3.1), estimating the distribution of the angles between the eigenvector estimators and their population versions (Sec. 4.3.2), and constructing confidence intervals for the percentage of variance explained by the first $k$ principal components ($k = 1, \ldots, p - 1$) (Sec. 4.3.3). For each application, we present the results of a simulation study in which we investigated the finite-sample performance of the fast and robust bootstrap. A comparison with the classical bootstrap is not included in the simulations due to the high computational cost of the latter.

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

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

(27)

The eigenvalues of the corresponding shape matrix $\Gamma_1 = |\Sigma_1|^{-1/2}\Sigma_1$ are $[7.92, 2.41, .59, .35, .25]$. Next, we introduced contamination by replacing a specified percentage of the observations by outliers randomly scattered within the subspace spanned by the last two eigenvectors of $\Sigma_1$. 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 cannot be identified by univariate methods. We considered four degrees of contamination: $0\%$, $10\%$, $20\%$, and $30\%$ outliers. In this article we present results only for the cases of $0\%$ and $20\%$ outliers; simulation results for $10\%$ and $30\%$ outliers were very similar. For each combination of $n$ and proportion of outliers, we generated $m = 1,000$ 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 $B = 1,000$ recalculation. 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 (EAV) approach cannot be used to assess the stability of the eigenvector, however.

#### 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 $\Gamma$. 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, p. 202). The nominal confidence level was 95%. The intervals based on the EAV and the asymptotic normal approximation are of the form

$$\left[\hat{\lambda}_j - \frac{1}{1 + \Phi^{-1}(0.975)}\sqrt{\hat{V}_1/(n - 1)}, \hat{\lambda}_j - \frac{1}{1 - \Phi^{-1}(0.975)}\sqrt{\hat{V}_1/(n - 1)}\right],$$

where $\hat{\lambda}_j$ denotes the empirical version of $\lambda_j$.

Table 4 gives the observed coverage levels (among the $m = 1,000$ 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. In Tables 4 and 5, the average lengths of the intervals is given in brackets.

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 $\Gamma_1$. Indeed, reasonably large ratios between successive eigenvalues are needed to obtain good estimates and accurate bootstrap inference results. In our case we have that $\lambda_1/\lambda_2$ and $\lambda_2/\lambda_3$ are quite large, whereas the other ratios are comparatively small. Also note that the intervals based

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\lambda_1$</th>
<th>$\lambda_2$</th>
<th>$\lambda_3$</th>
<th>$\lambda_4$</th>
<th>$\lambda_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>96.2(6.584)</td>
<td>95.4(2.021)</td>
<td>91.9(4.311)</td>
<td>89.5(2.424)</td>
<td>86.5(1.30)</td>
</tr>
<tr>
<td>100</td>
<td>94.7(4.362)</td>
<td>96.0(1.324)</td>
<td>92.1(3.04)</td>
<td>89.9(1.70)</td>
<td>91.3(1.14)</td>
</tr>
<tr>
<td>200</td>
<td>95.6(2.966)</td>
<td>96.1(1.904)</td>
<td>95.9(2.215)</td>
<td>92.4(1.123)</td>
<td>93.5(0.89)</td>
</tr>
</tbody>
</table>

EAV

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\lambda_1$</th>
<th>$\lambda_2$</th>
<th>$\lambda_3$</th>
<th>$\lambda_4$</th>
<th>$\lambda_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>91.5(7.286)</td>
<td>93.3(2.152)</td>
<td>94.5(2.526)</td>
<td>97.5(3.02)</td>
<td>94.9(1.86)</td>
</tr>
<tr>
<td>100</td>
<td>92.4(4.533)</td>
<td>96.0(1.355)</td>
<td>94.3(3.377)</td>
<td>97.9(1.96)</td>
<td>96.8(1.32)</td>
</tr>
<tr>
<td>200</td>
<td>94.3(3.048)</td>
<td>95.7(1.915)</td>
<td>95.8(2.224)</td>
<td>96.8(1.133)</td>
<td>97.1(0.93)</td>
</tr>
</tbody>
</table>
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 because they have a negative bias, but works adversely for the positively biased largest eigenvalues.

Finally, note that there is little difference between the contaminated (Table 5) and “clean” (Table 4) datasets. 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, although not dramatically so.

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. In this case a performance measure for an estimator 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 components estimates by examining the bootstrap distribution of the angles between the calculated eigenvectors and the originally estimated eigenvector. The angle between the normalized eigenvectors \( \hat{v}_j \) and \( v_j \) is given by \( \arccos(\langle \hat{v}_j, v_j \rangle) \). The bootstrap distribution of these angles is then an estimate of the distribution of the angles \( \arccos(\langle \hat{v}_j, v_j \rangle) \) 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 Bootstrap estimates \( \hat{v}_j(\hat{\Gamma}_n) \) and the original MM estimate \( v_j(\hat{\Gamma}_n) \). The average and standard deviation of the \( m = 1,000 \) values are displayed in Table 6 for the normal data and in Table 7 for the data with 20% outliers. The average is compared with 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.

**Table 6. Average Bootstrap Estimates (with standard deviations) of the Mean Angle Between MM Eigenvectors and Distribution Eigenvectors; Normal Data**

<table>
<thead>
<tr>
<th>( n )</th>
<th>( v_1 )</th>
<th>( v_2 )</th>
<th>( v_3 )</th>
<th>( v_4 )</th>
<th>( v_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>.121</td>
<td>.169</td>
<td>.369</td>
<td>.567</td>
<td>.462</td>
</tr>
<tr>
<td>Monte Carlo</td>
<td>.124</td>
<td>.193</td>
<td>.395</td>
<td>.566</td>
<td>.434</td>
</tr>
<tr>
<td>(SD)</td>
<td>(.041)</td>
<td>(.055)</td>
<td>(.124)</td>
<td>(.133)</td>
<td>(.148)</td>
</tr>
<tr>
<td>100</td>
<td>.084</td>
<td>.115</td>
<td>.235</td>
<td>.382</td>
<td>.327</td>
</tr>
<tr>
<td>Monte Carlo</td>
<td>.087</td>
<td>.124</td>
<td>.270</td>
<td>.428</td>
<td>.341</td>
</tr>
<tr>
<td>(SD)</td>
<td>(.016)</td>
<td>(.013)</td>
<td>(.100)</td>
<td>(.125)</td>
<td>(.139)</td>
</tr>
<tr>
<td>200</td>
<td>.057</td>
<td>.080</td>
<td>.166</td>
<td>.258</td>
<td>.221</td>
</tr>
<tr>
<td>Monte Carlo</td>
<td>.059</td>
<td>.083</td>
<td>.176</td>
<td>.292</td>
<td>.245</td>
</tr>
<tr>
<td>(SD)</td>
<td>(.007)</td>
<td>(.005)</td>
<td>(.051)</td>
<td>(.091)</td>
<td>(.102)</td>
</tr>
</tbody>
</table>

Concerning the bootstrap performance, we mostly have a slight overestimation of the variability. Nevertheless, the bootstrap estimates of the mean angles seem 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:

\[
\hat{\rho}_k = \frac{\sum_{j=1}^{k} \hat{\lambda}_j}{\sum_{j=1}^{p} \lambda_j} \quad \text{for } k = 1, \ldots, p - 1.
\]

These statistics estimate the percentage of variance explained by the first \( k \) principal components. Often the statistics \( \hat{\rho}_k \), \( k = 1, \ldots, p - 1 \), are used to determine how many principal components should be used in further analyses. Therefore, measuring the variability and stability of \( \hat{\rho}_k \) is of interest. 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 \( \hat{\rho}_k \) for choosing the number of components, we might use the lower limit of a confidence interval.

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 \( \hat{\rho}_k \) at the distribution \( F_{\mu, \Sigma} \) is given by

\[
\text{ASV}(\hat{\rho}_k) = \frac{(1 - \rho_k)^2}{\left(\sum_{j=1}^{k} \lambda_j^2\right)^2} \left(\sum_{j=1}^{k} \lambda_j^2 + \rho_k^2 \sum_{j=k+1}^{p} \lambda_j^2 \right). \tag{28}
\]

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

Table 8 gives the coverage and average length of the 95% confidence intervals for the normal data. Note that the true values, corresponding to \( \Gamma_1 \), 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 where \( n = 50 \) and \( k = 4 \). The asymptotic variance intervals are unduly short, resulting in poor coverage. Apparently larger samples are required for the asymptotic variance to produce accurate inference results for \( \hat{\rho}_k \). 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.
Table 8. Coverage and Length of 95% Confidence Intervals for Percentage of Variance Explained by the First k Principal Components; Normal Data

<table>
<thead>
<tr>
<th>n</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bootstrap</td>
<td>Bootstrap</td>
<td>Bootstrap</td>
<td>Bootstrap</td>
</tr>
<tr>
<td>50</td>
<td>94.7(0.220)</td>
<td>96.8(0.085)</td>
<td>95.0(0.047)</td>
<td>90.6(0.020)</td>
</tr>
<tr>
<td>100</td>
<td>94.1(0.155)</td>
<td>95.4(0.058)</td>
<td>95.3(0.032)</td>
<td>93.4(0.016)</td>
</tr>
<tr>
<td>200</td>
<td>96.8(0.108)</td>
<td>95.4(0.039)</td>
<td>95.3(0.022)</td>
<td>94.8(0.011)</td>
</tr>
<tr>
<td>EASV</td>
<td>90.5(0.189)</td>
<td>90.2(0.067)</td>
<td>83.9(0.035)</td>
<td>73.9(0.016)</td>
</tr>
</tbody>
</table>

Table 9. Coverage and Length of 95% Confidence Intervals for Percentage of Variance Explained by the First k Principal Components; 20% Outliers

<table>
<thead>
<tr>
<th>n</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bootstrap</td>
<td>Bootstrap</td>
<td>Bootstrap</td>
<td>Bootstrap</td>
</tr>
<tr>
<td>50</td>
<td>94.2(0.237)</td>
<td>94.4(0.124)</td>
<td>93.1(0.079)</td>
<td>89.7(0.045)</td>
</tr>
<tr>
<td>100</td>
<td>94.2(0.167)</td>
<td>94.7(0.065)</td>
<td>91.2(0.037)</td>
<td>90.3(0.017)</td>
</tr>
<tr>
<td>200</td>
<td>93.8(0.119)</td>
<td>93.4(0.045)</td>
<td>91.4(0.026)</td>
<td>90.5(0.013)</td>
</tr>
<tr>
<td>EASV</td>
<td>90.4(0.066)</td>
<td>89.7(0.075)</td>
<td>86.5(0.040)</td>
<td>75.7(0.018)</td>
</tr>
</tbody>
</table>

Even though the intervals are slightly longer now, 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.

5. EXAMPLE

For an illustration on a real data example, we consider the measurements on \( n = 100 \) forged old Swiss 1,000 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 over all coefficients in the eigenvectors. The cutoff lines are obtained through simulation (as described in Pison and Van Aelst, 2004). Observation 13 can be regarded as a non-outlying point with high influence, although it is a boundary case and should probably be seen as a regular point.

The MM estimates for the eigenvalues of the shape matrix are given by \( \hat{\lambda} = [10.25, 1.94, 1.05, .51, .39, .24] \). The weights in the first principal component are \( \hat{v}_1 = [-0.070, .028, -0.019, .813, -0.569, -0.094] \). Hence the first principal component can be interpreted as the difference between the fourth and the fifth variables. These correspond to the distance from the inner frame on the bill to the lower and upper borders.

Figure 3 shows histograms for the weights in the first principal component, obtained by performing fast and robust bootstrap with \( B = 1,000 \). This is another application of bootstrap for PCA not considered in the previous section. Some care is needed here, because the coefficients of the principal components are not uniquely defined. To obtain meaningful inference results, we imposed that for each \( \hat{v}_j, j = 1, \ldots, 6 \), 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.

Figure 2. Swiss Bank Notes Data, Diagnostic Plot of 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.
The bootstrap result in Figure 3 indicates that the coefficients of \( \hat{v}_1 \) (i.e., the weights associated with the first principal component) 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 principal component.

As an alternative way of assessing the stability of the first principal component, we look at the bootstrap distribution of the angles \( \cos(\| \hat{v}_1 \|, \hat{v}_1^* \) where the upper panels in Figure 4 compare the result from the classical bootstrap (a) with that from the fast bootstrap (b).

Whereas the fast bootstrap recalculations yield angles no larger than about .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, which plot 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. Figure 4(c) corresponds to the classical bootstrap; Figure 4(d), to the fast and robust bootstrap. Clearly, in the classical procedure higher proportions of outlying observations give rise to a larger variability concerning the eigenvector angles. On the other hand, in the fast bootstrap procedure the group of 15 outliers is severely downweighted, because of 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. Figure 5(a) shows the bootstrap intervals (BCA) and for comparison also gives the intervals based on the empirical version of the asymptotic variance. The classical bootstrap intervals, not shown in 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 optimistic. Note that the simulation study demonstrated that bootstrap intervals usually have better coverage. The intervals in general are somewhat long, although no overlap is present. Concerning the number of principal components to retain, suppose that 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 were set at 80%, then the choice for \( k = 2 \) would be obvious, because both the point estimate and almost the whole interval would exceed 80%. In the case of a 90% cutoff, the choice of \( k = 3 \) based on \( \hat{p}_3 \) would be 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 \( \pi \) would then equal, for example, 80%.
Figure 5. Swiss Bank Notes Data. (a) 95% confidence intervals for the percentage of variance explained by the first $k$ principal components; fast bootstrap (—–) compared to intervals based on asymptotic normality (−−−−−). (b) 95% one-sided confidence intervals for the percentage of variance explained by the first $k$ principal components.

or 90%. We can decide to accept $H_0$ at the level $\alpha$ whenever $\pi$ is contained in a $(1-\alpha)100\%$ one-sided confidence interval for $p_k$. As an example, Figure 5(b) plots the 95% one-sided intervals for $p_k$ based on the fast bootstrap, where we again used the BCA method. From this plot we can immediately see which hypotheses would be accepted on the 5% significance level; for example, $H_0:p_k \geq 80\%$ would be accepted for $k \geq 2$, whereas $H_0:p_k \geq 90\%$ would be accepted for $k \geq 3$.

Remark 4. We have illustrated the robustness of the fast bootstrap in this section, and found it in the simulation study as well. We would like to stress that its speed is also a very important feature. As an example, on a 1.9-GHz Pentium IV, computation of the classical bootstrap on the Swiss bank notes data took about 20 minutes, whereas the fast bootstrap only took a few seconds. Software code in MATLAB and R/S–PLUS 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/MMPCAbout.html](http://users.ugent.be/~svaelst/software/MMPCAbout.html).

6. CONCLUSION

In this article we have 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 Salibian-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 in general slightly outperforms the approach based on asymptotic normality. However, one should be aware of the intrinsic difficulties in bootstrapping a PCA; that is, care must be taken when interpreting bootstrap results, in the event that eigenvalue ratios are small (see, e.g., Jolliffe 2002, p. 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 dataset.

APPENDIX: PROOFS

Proof of Theorem 1

As given by Van Aelst and Willems (2005), for the S estimators, we have that

$$e_n^\ast(\mu_n, \Sigma_n; \mathcal{X}_n) = \min\{[nr], [n-nr] - k(\mathcal{X}_n)\}/n.$$  (A.1)

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

$$\frac{1}{n} \sum_{x \in \mathcal{X}_n'} \rho_1(d(x; \hat{\mu}_n', \hat{\Sigma}_n')) \leq \frac{1}{n} \sum_{x \in \mathcal{X}_n} \rho_1(d(x; \mu_n, \Sigma_n)) \leq \frac{1}{n} \sum_{x \in \mathcal{X}_n} \rho_0(d(x; \mu_n, \Sigma_n)) = b = r\rho_0(\infty).$$

The first of these inequalities holds because $(\hat{\mu}_n, \hat{\Sigma}_n^\ast)$ form the solution to the M minimization; the second is due to the assumptions on the $\rho$ functions. Because $r\rho_1(\infty) = r\rho_0(\infty)$, we furthermore obtain

$$\frac{1}{n} \sum_{x \in \mathcal{X}_n} \rho_1([(x - \mu_n)^T \hat{\Sigma}_n^{-1} (x - \mu_n)]^{1/2}) \leq r\rho_1(\infty).$$

It follows that there exists some $d_1 < \infty$, not depending on $\mathcal{X}_n'$, such that $[(x - \mu_n)^T \hat{\Sigma}_n^{-1} (x - \mu_n)]^{1/2} < d_1$ for at least $[n-nr]$ points in $\mathcal{X}_n'$. We know that the S scale $\hat{\sigma}_n$ is bounded for $m < \min([nr], [n-nr] - k(\mathcal{X}_n))$, and hence there also exists some $d_2 < \infty$ such that $[(x - \mu_n)^T \hat{\Sigma}_n^{-1} (x - \mu_n)] < d_2$ for the same $[n-nr]$ points. Now, because $m < [n-nr] - k(\mathcal{X}_n)$, 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 $\hat{\Sigma}_n^\ast$ must be bounded away from 0, and because its determinant is constant, its largest eigenvalue is bounded from above. It immediately follows that $\|\hat{\mu}_n\|$ must be bounded as well. Furthermore, the eigenvalues of $\hat{\Sigma}_n^\ast$ are bounded because $\hat{\sigma}_n$ and the eigenvalues of $\hat{\Sigma}_n^\ast$ are bounded.

Proof of Theorem 2

We need the following lemma to prove Theorem 2.

**Lemma A.1.** Let $Y_1, \ldots, Y_n \sim F$ be a sequence of iid random vectors. Let $(\hat{\mu}_n, \hat{\Sigma}_n)$ be consistent estimators for $(\mu, \Sigma)$. Let $k: \mathbb{R} \to \mathbb{R}$
be a function that is bounded and almost everywhere continuous. If \( \kappa(Y, \mathbf{t}, S) = \kappa((Y - \mathbf{t})S^{-1}(Y - \mathbf{t})) \), then
\[
\frac{1}{n} \sum_{i=1}^{n} \kappa(Y_i, \mathbf{u}_i, S_n) \xrightarrow{p} E\kappa(Y, \mu, \Sigma).
\]

**Proof.** The proof is based on an argument used by Davies (1987, proof of thm. 3). Denote \( \kappa_n(y) := \kappa(y, \mathbf{u}_i, S_n) \) and \( \kappa(y, \mathbf{u}, \Sigma) \). For any \( y \) such that \( \kappa \) is continuous at \( (y - \mu)^T\Sigma^{-1}(y - \mu) \), and for any sequence \( (y_n)_n \) such that \( y_n \rightarrow y \), we have that
\[
\kappa_n(y_n) \xrightarrow{n \rightarrow \infty} \kappa(y).
\]
Because \( \kappa \) is almost everywhere continuous, the convergence in (A.2) 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 \( Y_1, \ldots, Y_n \). Then we obtain from the theorem that
\[
\int \gamma(\kappa_n(y))dF_n \rightarrow \int \gamma(\kappa(y))dF,
\]
because \( \gamma \) is bounded and uniformly continuous.

**Proof of Theorem 2.** We mostly follow along the lines of the approach of Salibian-Barrera and Zamar (2002). We can write the estimating equations (16)–(19) as
\[
\begin{align*}
\mu_n &= a_n(\tilde{\mu}_n, \tilde{\gamma}_n, \tilde{\Sigma}_n)^{-1}B_n(\tilde{\mu}_n, \tilde{\gamma}_n, \tilde{\Sigma}_n), \\
\tilde{\gamma}_n &= \frac{1}{n} \sum_{i=1}^{n} \gamma(Y_i - \mu_n) \tilde{\gamma}_n, \\
\tilde{\Sigma}_n &= \tilde{\Sigma}_n, \\
\tilde{\mu}_n &= \tilde{\mu}_n(\tilde{\mu}_n, \tilde{\gamma}_n, \tilde{\Sigma}_n).
\end{align*}
\]
and
\[
\begin{align*}
\tilde{\mu}_n &= a_n(\tilde{\mu}_n, \tilde{\gamma}_n, \tilde{\Sigma}_n)^{-1}B_n(\tilde{\mu}_n, \tilde{\gamma}_n, \tilde{\Sigma}_n), \\
\tilde{\gamma}_n &= \tilde{\gamma}_n, \\
\tilde{\Sigma}_n &= \tilde{\Sigma}_n, \\
\tilde{\mu}_n &= \tilde{\mu}_n(\tilde{\mu}_n, \tilde{\gamma}_n, \tilde{\Sigma}_n).
\end{align*}
\]
with properly defined functions \( a_n, B_n, V_n, \tilde{\gamma}_n, \tilde{\mu}_n, \tilde{\Sigma}_n \).

Consider the function \( f: \mathbb{R}^{2(p+p^2)} \rightarrow \mathbb{R}^{2(p+p^2)} \) for \( M, M_0 \in \mathbb{R}^p \) and \( G, C \in \mathbb{R}^{p \times p} \),
\[
f \left( \begin{array}{c} M \\
vec(G) \\
vec(C) \end{array} \right) := \left( \begin{array}{c} a_n(M, G, C)^{-1}B_n(M, G, C) \\
vec(V_n(M, G, C)^{-1/2}V_n(M, G, C)) \\
\vec(V_n(M_0, G, C)) \\
\vec(M_0 G, C) \end{array} \right).
\]
Let \( \tilde{\Theta}_n := (\tilde{\mu}_n, \tilde{\gamma}_n, \tilde{\Sigma}_n) \). We have that \( f(\tilde{\Theta}_n) \). Because \( \rho_0 \) and \( \rho_1 \) are sufficiently smooth, the function \( f \) allows a Taylor expansion around \( \Theta := (\mu^T (\gamma^T \Sigma^{-1} \mu)^T) \),
\[
\Theta_n = f(\Theta) + \sqrt{f(\Theta)(\tilde{\Theta}_n - \Theta)}
\]
Here \( \nabla f(\Theta) \in \mathbb{R}^{2(p+2p^2)} \times \mathbb{R}^{2(p+p^2)} \) is the Jacobian and \( H(f)(\Theta) \in \mathbb{R}^{2(p+2p^2)} \times \mathbb{R}^{2(p+p^2)} \). Thus, we can rewrite (A.7) as
\[
\sqrt{n}(\tilde{\Theta}_n - \Theta) = \left[ I - \nabla f(\Theta) \right]^{-1} \sqrt{n}(f(\Theta) - \Theta) + o_p(1). \]


