Robust Estimation of Cronbach's Alpha

A. Christmann

University of Dortmund, Fachbereich Statistik, 44421 Dortmund, Germany.

S. Van Aelst *

Ghent University, Department of Applied Mathematics and Computer Science, Krijgslaan 281 S9, B-9000 Gent, Belgium.

Abstract

Cronbach's alpha is a popular method to measure reliability, e.g. in quantifying the reliability of a score to summarize the information of several items in questionnaires. The alpha coefficient is known to be non-robust. We study the behavior of this coefficient in different settings to identify situations, which can easily occur in practice, but under which the Cronbach's alpha coefficient is extremely sensitive to violations of the classical model assumptions. Furthermore, we construct a robust version of Cronbach's alpha which is insensitive to a small proportion of data that belong to a different source. The idea is that the robust Cronbach's alpha reflects the reliability of the bulk of the data. For example, it should not be possible that some small amount of outliers makes a score look reliable if it is not.

Key words: Cronbach's alpha, MCD, M-estimator, Robustness, S-estimator.

* Corresponding author.
Email addresses: Christmann@statistik.uni-dortmund.de (A. Christmann), Stefan.VanAelst@UGent.be (S. Van Aelst).

1 Introduction

We consider the problem of constructing a measure of reliability for a set of items such as in a test. Cronbach [3] proposed the coefficient alpha as a lower bound to the reliability coefficient in classical test theory (see also [13]). This popular measure has been investigated further in e.g. [8,27,12,1].

Consider a series of items $Y_j = T_j + \varepsilon_j$ for $j = 1, \ldots, p$, where T_j are the true unobservable test scores and ε_j are the associated errors which are independent from the true test scores and distributed with zero mean. The score Z of the p items is defined as the sum, i.e. $Z = Y_1 + \ldots + Y_p$. Then Cronbach's alpha is given by

$$\alpha_n^C = \frac{p}{p-1} \frac{\mathrm{Var}\left(\sum_{j=1}^p Y_j\right) - \sum_{j=1}^p \mathrm{Var}\left(Y_j\right)}{\mathrm{Var}\left(\sum_{j=1}^p Y_j\right)}$$

$$= \frac{p}{p-1} \frac{\sum \sum_{j \neq k} \sigma_{jk}}{\sum \sum_{j,k} \sigma_{jk}} \tag{1}$$

$$\tag{2}$$

where σ_{jk} is the covariance of the pair (Y_j, Y_k). It has been shown in [9] that Cronbach's alpha is always a lower bound of reliability.

Cronbach's alpha can be estimated by substituting empirical variances and covariances in expression (1) above. However it is well known that classical estimators such as empirical variances and covariances can be heavily influenced by a few erroneous observations (see e.g. [10]). Therefore the resulting estimate of Cronbach's alpha can be completely misleading as soon as some mistaken observations are present. We want to avoid this problem and aim to construct a robust version of Cronbach's alpha in the sense that this reliability measure is able to resist some outlying observations. The robust Cronbach's alpha will

thus measure the reliability of the most central part of the observations while not being affected by some outlying observations. A robust measure of reliability was already proposed by Wilcox [28] who used the midvariance and midcovariance as robust estimates for the variances and covariances in (1). In this paper we propose to estimate the covariance matrix of $Y = (Y_1, \ldots, Y_p)^t$ using a robust estimator and then we substitute the elements of this robust covariance estimate into (1).

Many robust estimators of multivariate location and scatter have been investigated in the literature, such as M-estimators [17,11], the minimum volume ellipsoid and minimum covariance determinant estimator [21], and S-estimators [7,22,14].

Recently, robust multivariate statistical methods based on robust estimation of location and scatter have been developed and investigated such as factor analysis [18], principal component analysis [6], canonical correlation analysis [4] and multivariate regression [23]. An advantage of constructing a robust Cronbach's alpha as proposed in this paper is that it can be obtained immediately from the robust scatter matrix estimate computed for the robust multivariate analysis without any additional computational load. This is a clear advantage over the proposal of Wilcox [28] that has to be computed separately and does not take into account the multivariate nature of the data.

In Section 2 we review robust estimators of multivariate location and scatter. The robust Cronbach's alpha is introduced in Section 3 where we also investigate some important properties. Section 4 contains some simulation studies that show that the robust Cronbach's alpha performs well in situations with some outlying observations. A real data example is given in Section 5 while

Section 6 summarizes the conclusions.

2 Robust estimators of location and scatter

The robust Cronbach's alpha can be computed from any robust scatter estimate. For the simulations and examples in this paper we will mainly use the reweighted minimum covariance determinant (RMCD) estimator and S-estimators which are highly robust estimators that can be computed with standard statistical software packages, e.g. S-PLUS.

Consider a multivariate data set $\{y_i; 1 \leq i \leq n\}$ with $y_i = (y_{i1}, \ldots, y_{ip})^t \in \mathbb{R}^p$. Fix $\lceil n/2 \rceil \leq h \leq n$, then the MCD looks for the subset $\{y_{i_1}, \ldots, y_{i_h}\}$ of size h which is the most concentrated subset of size h in the sense that its covariance matrix has the smallest determinant. The estimate for the center is then defined as the mean $t_n^0 = \frac{1}{h}\sum_{j=1}^{h} y_{i_j}$ of the optimal subset and the covariance estimate C_n^0 is a multiple of $\frac{1}{h}\sum_{j=1}^{h}(y_{i_j} - t_n^0)(y_{i_j} - t_n^0)^t$, the classical covariance estimator based on the data in the optimal subset. For a specific model distribution, e.g. a multivariate normal, the multiplication factor can be selected to make the MCD consistent and unbiased at finite-samples [18].

The breakdown value of an estimator is the smallest fraction of observations that has to be replaced by arbitrary values to make the estimator useless (i.e. its norm goes to infinity). See e.g. [22] for more information about the breakdown value. We will denote $\gamma = (n - h)/n$ so that $0 \leq \gamma \leq 0.5$. It then follows that the MCD has breakdown value equal to γ. This means that a fraction γ of the data points may contain errors without having an unbounded effect on the MCD estimates of the location and scatter. Moreover, the MCD

location and scatter estimators are asymptotically normal and have a bounded influence function [2,5] which means that a small amount of contamination at a certain place can only have a bounded effect on the MCD estimates, see [10] for more information on the influence function. Two common choices for the subset size h are $h = [(n+p+1)/2] \approx n/2$ (so $\gamma \approx 0.5$) which yields the highest possible breakdown value, and $h \approx 3n/4$ (i.e. $\gamma \approx 0.25$) which gives a better compromise between efficiency and breakdown.

To increase the performance of the MCD it is customary to compute the reweighted MCD estimates (t_n^1, S_n^1) which are defined as

$$t_n^1 = \frac{\sum_{i=1}^n w(d_i^2) y_i}{\sum_{i=1}^n w(d_i^2)} \quad \text{and} \quad C_n^1 = d_n \frac{\sum_{i=1}^n w(d_i^2)(y_i - t_n^1)(y_i - t_n^1)^t}{\sum_{i=1}^n w(d_i^2)}. \qquad (3)$$

The weights $w(d_i^2)$ are computed as $w(d_i^2) = I(d_i^2 \leq q_\delta)$ where $q_\delta = \chi^2_{p,1-\delta}$ and $d_i^2 = (y_i - t_n^0)^t (C_n^0)^{-1} (y_i - t_n^0)$ is the squared robust distance of observation y_i based on the initial MCD estimates (t_n^0, C_n^0). It is customary to take $\delta = 0.025$ [25]. As for the initial MCD, the factor d_n can be chosen to make the reweighted MCD consistent and unbiased for a specific model distribution [18]. The reweighted MCD estimators (RMCD) preserve the breakdown value [16] and the bounded influence function [15] of the initial MCD estimators but have a higher efficiency as shown in [5]. Recently, Rousseeuw and Van Driessen [24] constructed a fast algorithm to compute the RMCD.

The S-estimates of location and scatter are defined as the couple (t_n^S, C_n^S) that minimizes $\det(C_n)$ under the constraint

$$\frac{1}{n} \sum_{i=1}^n \rho(\sqrt{(y_i - t_n)^t C_n^{-1} (y_i - t_n)}) \leq b, \qquad (4)$$

over all $t_n \in \mathbb{R}^p$ and $C_n \in \mathrm{PDS}(p)$, where $\mathrm{PDS}(p)$ is the set of all positive definite symmetric matrices of size p. See e.g. [14] for important conditions

on the ρ function. The constant b satisfies $0 < b < \rho(\infty)$ and determines the breakdown value of the estimator which equals $\min(\frac{b}{\rho(\infty)}, 1 - \frac{b}{\rho(\infty)})$ (see [14]). The most popular choice of ρ function is Tukey's biweight function which is given by

$$\rho_c(t) = \min\left(\frac{t^2}{2} - \frac{t^4}{2c^2} + \frac{t^6}{6c^4}\ ,\ \frac{c^2}{6}\right), \quad t \in \mathbb{R}. \tag{5}$$

Its derivative is given by

$$\psi_c(t) = t\left(1 - \frac{t^2}{c^2}\right)^2 I(|t| < c),\ , \quad t \in \mathbb{R}. \tag{6}$$

The tuning constant c in the ρ function (5) can be selected such that consistency at a specific model distribution is obtained. It is customary to choose c such that $E_H[\rho(\|y\|)] = b$ for $H = N(0, I_p)$. This implies that the S-estimators are consistent for the parameters (μ, Σ) of the normal distribution $N(\mu, \Sigma)$. S-estimators are asymptotically normal and have a bounded influence function [7,14]. Efficient algorithms to compute S-estimators have been constructed in [26,20]. The S-estimators based on Tukey's biweight function will be denoted S_{bw}.

Another class of robust scatter matrix estimators are M-estimators. We will consider the M-estimator based on the assumption of Student's t_3 distribution which will be denoted by T3. It has reasonable robustness and efficiency properties, but also some additional advantages. There exists a unique solution of the objective criterion under very weak assumptions and there exists an always converging iterative algorithm to compute the estimate, as was shown in [11]. Furthermore, this estimator is intuitively appealing as it is a maximum likelihood estimator if the errors follow a multivariate t_3 distribution. However, the main disadvantage of T3 is its low breakdown point.

3 Robust Cronbach's alpha

Consider a dataset $Y_n = \{y_i; i = 1, \ldots, n\} \subset \mathbb{R}^p$ and denote by t_n and C_n the corresponding robust estimates of location and scatter such as the RMCD estimates or S-estimates defined above. Then the robust Cronbach's alpha estimate is defined as

$$\alpha_n^R = \frac{p}{p-1} \frac{\sum \sum_{j \neq k} c_{jk}}{\sum \sum_{j,k} c_{jk}} \tag{7}$$

where c_{ij}, $i, j = 1, \ldots, p$, are the elements of the matrix C_n. Hence, instead of substituting the empirical variances and covariances in (1) we now use their robust counterparts to obtain a robust estimate of Cronbach's alpha.

Let us now consider the class of unimodal elliptically symmetric distributions $F_{\mu,\Sigma}$ with density function

$$f_{\mu,\Sigma}(y) = \frac{g(y-\mu)^t \Sigma^{-1} (y-\mu)}{\sqrt{\det(\Sigma)}} \tag{8}$$

with $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathrm{PDS}(p)$ and where the function g has a strictly negative derivative. Multivariate normal distributions obviously belong to this class of distributions. With $\Sigma = (\sigma_{ij})$, we then focus on estimating the quantity

$$\alpha = \frac{p}{p-1} \frac{\sum \sum_{j \neq k} \sigma_{jk}}{\sum \sum_{j,k} \sigma_{jk}} \,. \tag{9}$$

If the scatter estimator C_n is consistent in probability or almost surely, then it follows immediately from Slutsky's theorem that the corresponding Cronbach's alpha estimator given by (7) is a consistent estimator of α (in probability or almost surely). Consistency of robust location/scatter estimators at elliptically symmetric distributions has been shown in [2] for the MCD, in [15] for the RMCD and in [7,14] for S-estimators.

The influence function (IF) describes the local robustness of the functional

version of an estimator. A statistical functional corresponding to an estimator C_n is a map C which maps any p-variate distribution G on $C(G) \in \mathrm{PDS}(p)$ such that $C(F_n) = C_n$ for any possible empirical distribution function F_n. The functional version of the robust Cronbach's alpha associated with a scatter functional C will be denoted by α_C^R. Hence, by using the elements of $C(G)$ into (7) we obtain $\alpha_C^R(G)$. It follows immediately that $\alpha_C^R(F_{\mu,\Sigma}) = \alpha$ whenever $C(F_{\mu,\Sigma}) = \Sigma$, that is, C is Fisher-consistent for Σ at elliptical distributions $F_{\mu,\Sigma}$.

The influence function of the functional α_C^R at the distribution $F_{\mu,\Sigma}$ measures the effect on $\alpha_C^R(F_{\mu,\Sigma})$ of adding a small mass at a certain point y. Such a perturbation mimics the occurrence of isolated outliers, e.g. due to typing errors. Hence, a robust method should have a bounded influence function such that contamination at any point can only have a limited effect on the estimate. If we denote by Δ_y the distribution putting all its mass on y, then the influence function is given by

$$\begin{aligned} IF(y; \alpha_C^R, F_{\mu,\Sigma}) &= \lim_{\varepsilon \downarrow 0} \frac{\alpha_C^R((1-\varepsilon)F_{\mu,\Sigma} + \varepsilon\Delta_y) - \alpha_C^R(F_{\mu,\Sigma})}{\varepsilon} \\ &= \frac{\partial}{\partial \varepsilon} \alpha_C^R((1-\varepsilon)F_{\mu,\Sigma} + \varepsilon\Delta_y)\big|_{\varepsilon=0}. \end{aligned} \tag{10}$$

See [10] for further details. For scatter matrix estimators possessing an influence function the following result can easily be derived from (7) by computing the derivate of α_C^R with respect to ε as in (10).

Theorem 3.1 *If the scatter matrix estimator C possesses an influence function then the influence function of α_C^R at elliptically symmetric distributions*

$F := F_{\mu,\Sigma}$ is given by

$$IF(y; \alpha_C^R, F) = \frac{\frac{p}{p-1} \sum \sum_{j \neq k} \mathrm{IF}(y; c_{jk}, F) - \alpha_C^R(F) \sum \sum_{j,k} \mathrm{IF}(y; c_{jk}, F)}{\sum \sum_{j,k} \sigma_{jk}}.$$

It follows that the influence function of the robust Cronbach's alpha is bounded as soon as the influence function of the robust scatter matrix estimator is bounded which is the case for RMCD, T3, and S-estimators. Therefore, our approach based on a robust estimate of the scatter matrix indeed yields a robust estimate of Cronbach's alpha.

As an example, let us consider the influence function of the S-estimator of scatter based on Tukey's biweight function (5) for a multivariate standard normal distribution $F = N(\mathbf{0}, \mathbf{I})$ which is given by

$$\mathrm{IF}(y; C^S, F) = \frac{2}{\gamma_3} \left(\rho(||y||) - b_0\right) + \frac{1}{\gamma_1} p \, \psi(||y||) \, ||y|| \left(\frac{yy^t}{||y||^2} - \frac{1}{p}\mathbf{I}\right), \quad (11)$$

where

$$\gamma_1 = (p+2)^{-1} \mathrm{E}_F \left[\psi'(||Y||) \, ||Y||^2 + (p+1)\psi(||Y||) \, ||Y||\right]$$
$$\gamma_3 = \mathrm{E}_F \left[\psi(||Y||) \, ||Y||\right].$$

(see [14, Corollary 5.2]). The influence function of Cronbach's alpha based on the S-estimator S_{bw} for the bivariate standard normal distribution is given in Figure 1. Note that the influence function is smooth and bounded. Furthermore, for points with large euclidean norm $||y||$ it is constant, but not necessarily equal to zero for general multivariate normal distributions. Hence, data points lying far away from the bulk of the data cloud only have small impact on this robust version of Cronbach's alpha.

As the influence function is an asymptotical concept, it is also interesting to consider an empirical version of the influence function for finite sample sizes.

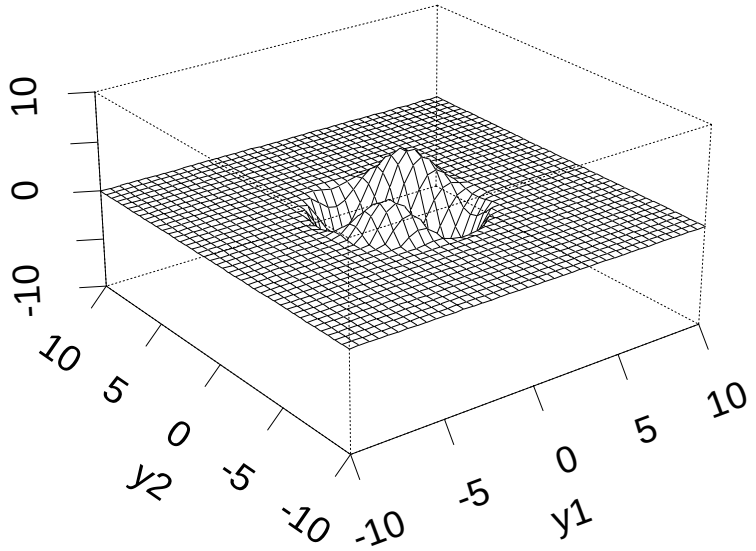


Fig. 1. Influence function of Cronbach's alpha based on the S-estimator S_{bw} at the bivariate normal distribution.

Here, we consider the *sensitivity curve* SC_n, c.f. [10, p. 93]. The sensitivity curve of Cronbach's alpha α_n given a multivariate data set $(y_1, \ldots, y_{n-1})$ is defined by

$$SC_n(y) = n\,[\alpha_n(y_1, \ldots, y_{n-1}, y) - \alpha_{n-1}(y_1, \ldots, y_{n-1})]\,, \quad y \in \mathbb{R}^p. \tag{12}$$

Hence, SC_n describes the standardized behavior of the estimate if one arbitrary data point y is added to the dataset.

Sensitivity curves of Cronbach's alpha based on empirical (co)variances and its robust alternatives are given in Figure 2 for the bivariate standard normal distribution. Note that due to different magnitudes of the sensitivity curves the scaling of the vertical axis in the plots is not identical for all four estimates. In Figure 2, we consider the classical Cronbach's alpha based on the empirical covariance matrix S, and robust Cronbach's alpha based on RMCD and the S-estimator S_{bw} (both with a 25% breakdown point), and the M-estimator T3. We see that the impact of even one single additional observation can be

extremely large for the classical Cronbach's alpha, whereas the robustifications behave much more stable. Especially the sensitivity curves based on RMCD and S_{bw} are very stable for observations far away from the bulk of the data. Note that the sensitivity curve of Cronbach's alpha based on the S-estimator S_{bw} is very similar to the influence function shown in Figure 1, although we used only a moderate sample size of $n = 100$ to construct SC_n. Cronbach's alpha based on T3 shows a smooth and more robust behavior than the classical estimator, but it is not as robust as the estimators based on RMCD and S_{bw} for extreme outliers.

Software code to compute our robust versions of Cronbach's alpha in SAS and S-PLUS is available from

http://www.statistik.uni-dortmund.de/sfb475/berichte/cronbach.zip.

4 Simulations

We investigate the behavior of the classical and robust Cronbach's alpha estimators for finite sample sizes via simulations for sample sizes of $n = 40, 100,$ and 500. Let $Y_1, \ldots, Y_n$ be independent and identically distributed random vectors with multivariate distribution F. For dimension $p = 2$ we define location vectors $\mu = (0, 0)'$, $\mu_1 = (2, 2)'$, and $\mu_2 = (-2, 2)'$. For dimension $p = 10$ we define location vectors $\mu = \mathbf{0} \in \mathbb{R}^p$, $\mu_1 = (2, \ldots, 2)'$, and $\mu_2 = (-2, 2, \ldots, 2)'$. As scatter matrices we use $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{p \times p}$, where $\sigma_{ij} = 1$, if $i = j$, and $\sigma_{ij} = \rho$, if $i \neq j$, and $\Sigma_1 = (\sigma_{ij}) \in \mathbb{R}^{p \times p}$, where $\sigma_{ij} = 1$, if $i = j$. If $p = 2$ the off-diagonal elements of Σ_1 are $\sigma_{12} = \sigma_{21} = -\rho$. If $p = 10$ we set the off-diagonal elements of Σ_1 equal to $\sigma_{ij} = -\rho$, if $\{i = 1$ or $j = 1$ and $i \neq j\}$, and $\sigma_{ij} = \rho$, if $\{i > 1, j > 1$ and $i \neq j\}$. We use $\delta = 0.05, 0.10,$ and 0.20 as

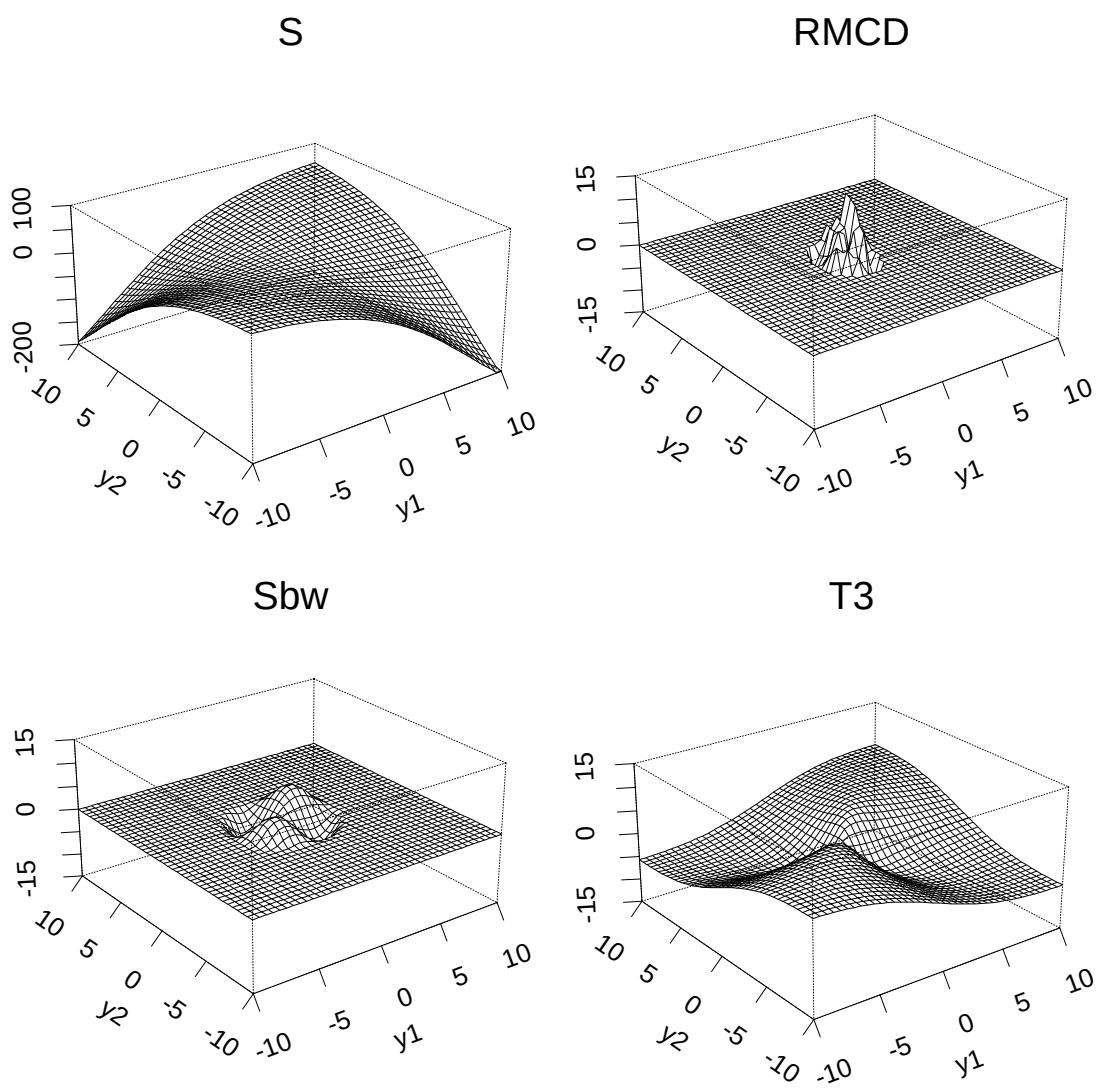


Fig. 2. Sensitivity curves for a 2−dimensional data set with $n = 100$ observations simulated from $F = \mathrm{N}(\mathbf{0}, \mathbf{I})$.

contamination proportions, and study correlations of $\rho = 0, 0.5$, and 0.8. In the simulations the following five probability models are considered:

- N: multivariate normal $F = N(\mu, \Sigma)$
- t_3: multivariate Student's t with 3 df $F = t_3(\mu, \Sigma)$
- $\delta\%$ M1: contamination model 1 with different covariance matrix:
  $F = (1 - \delta)N(\mu, \Sigma) + \delta N(\mu, \Sigma_1)$
- $\delta\%$ M2: contamination model 2 with different location parameter and co-

variance matrix: $F = (1-\delta)N(\mu, \Sigma) + \delta N(\mu_1, \Sigma_1)$

- δ% M3: contamination model 3 with different location parameter:
  $F = (1-\delta)N(\mu, \Sigma) + \delta N(\mu_1, \Sigma)$

To allow a visual comparison of these probability models, scatterplots of data sets simulated according to these five models for $p = 2$, $n = 100$, $\rho = 0.8$, and $\delta = 10\%$ are given in Figure 3. The data points generated from the contamination part of the distributions are marked as dots. For each of the sample sizes

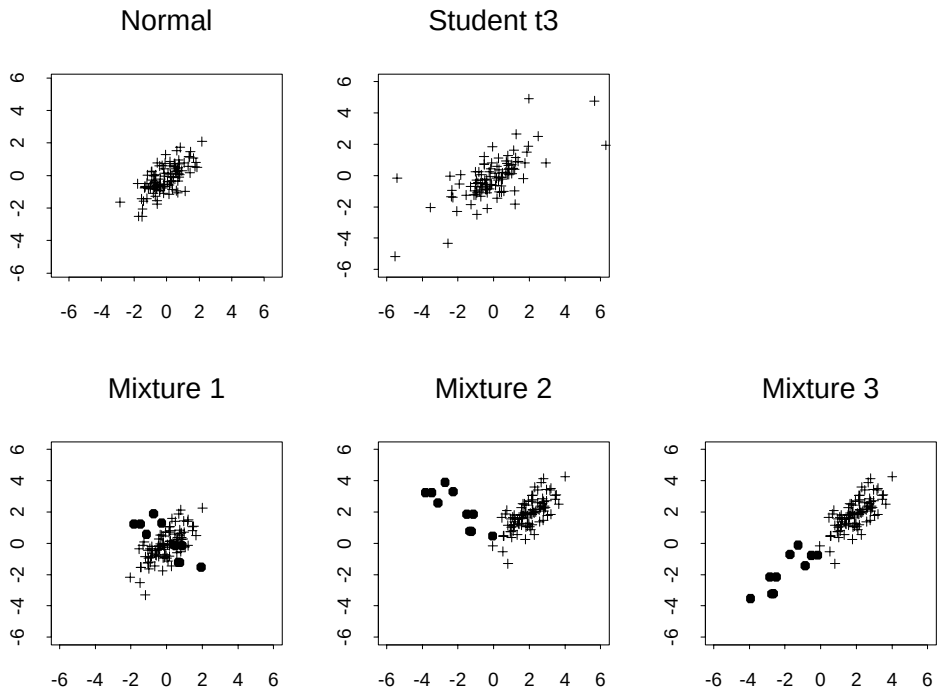


Fig. 3. Scatterplots of simulated data for $p = 2$, $n = 100$, $\rho = 0.8$, and $\delta = 10\%$.

we generated 1000 datasets and computed bias and mean squared error of the Cronbach's alpha based on the classical covariance matrix estimator S and based on the robust alternatives MCD, RMCD, S_{bw}, all with 25% breakdown point, and T3. The main results of the simulations are summarized in Figures 4 to 6. The simulations results for the other situations were very similar.

First, note that these simulations confirm that the classical Cronbach's alpha is non-robust with respect to violations of the model assumptions. It can seriously overestimate (contamination model 3, Figure 4a) or underestimate (contamination models 1 and 2, Figure 5a) its population value. Student's distribution t_3 is elliptically symmetric with heavier tails than the normal distribution and is often a good approximation to the distribution of high quality data, c.f. [10], p. 23). However, even in this situation the bias and the MSE of Cronbach's alpha is often much larger than under the classical assumption. The same is true for contamination model 1 where the contaminating distribution is a normal with the same mean vector but a different covariance matrix than the main part of the mixture distribution, see Figure 5. If the contamination is asymmetric as in the other two contamination models, the behavior of Cronbach's alpha can be even worse.

The robust Cronbach's alpha coefficients based on all three robust covariance estimators yield more stable estimates than the classical approach. In most cases Cronbach's alpha based on the RMCD estimator gives better result than the Cronbach's alpha based on the initial MCD estimator, which often has a higher bias and a higher mean squared error. Hence, we will not consider the MCD results in more detail. Cronbach's alpha coefficient based on RMCD is the only estimator under consideration which still gives reasonable results if the mixing proportion is as high as $\delta = 20\%$. Furthermore, this estimator often gives already better results with respect to bias and mean squared error than Cronbach's alpha under a multivariate t_3 distribution.

When the assumption of normality is not valid, Cronbach's alpha based on the Tukey biweight S-estimator, i.e. S_{bw}, performed best except for contamination models with contamination proportion $\delta = 20\%$. This amount of contamina-

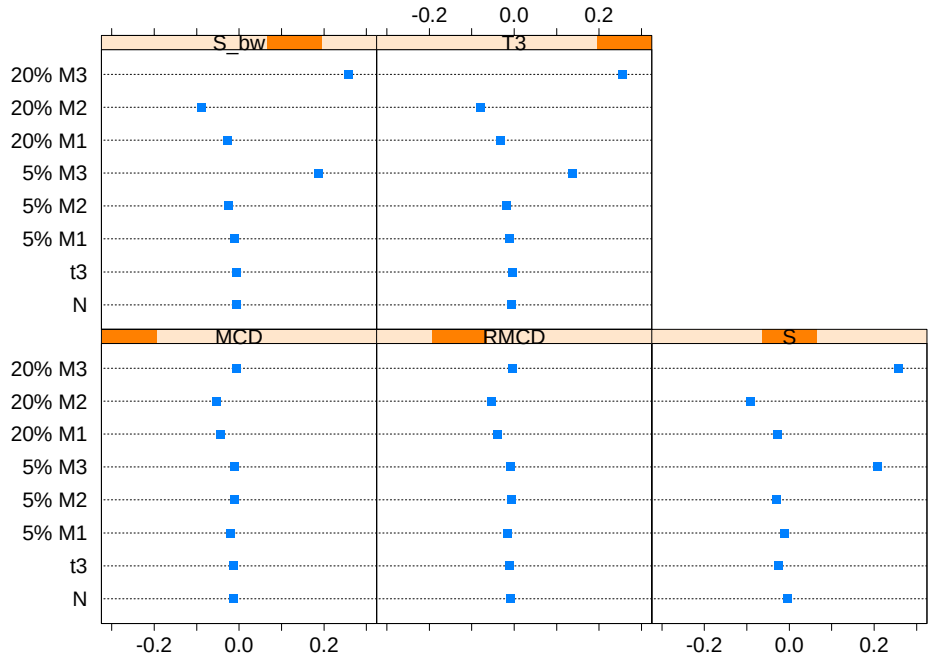
tion is close to the breakdown point of the estimator and causes a large (but bounded) bias which affects its perfomrance. Finally, S_{bw} performs almost as good as the classical estimator if the assumption of normality is fulfilled.

The M-estimator T3 yields more robust results than the classical approach based on the empirical covariance matrix, but even for models with 5% of contamination it often gives worse results than the estimators based on RMCD or S_{bw}, especially for contamination model 3 where the outlying observations can be interpreted as good leverage points in the sense of Rousseeuw and van Zomeren [25] (see Figure 5). This behavior of T3 coincides with the properties of the sensitivity curves shown in section 3.

Fig. 4. (a) Bias and (b) square root of mean squared error for several estimators of Cronbach's α for $p = 2$, $\rho = 0$, and $n = 100$. The true value of CR_α under classical normality assumptions is 0.

Fig. 5. (a) Bias and (b) square root of mean squared error for several estimators of Cronbach's α for $p = 2$, $\rho = 0.5$, and $n = 100$. The true value of CR_α under classical normality assumptions is 0.667.

(a)



(b)

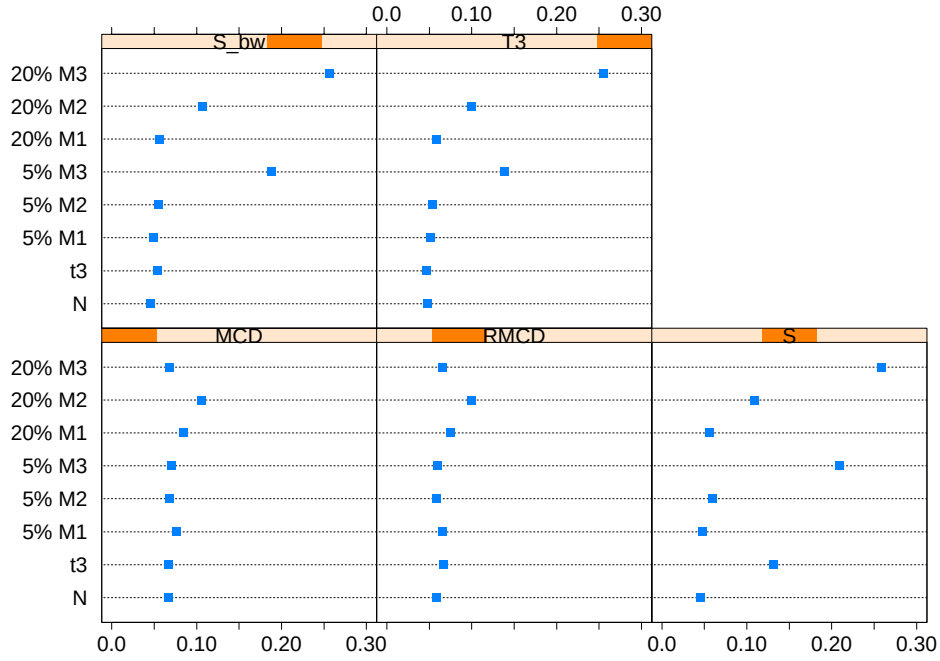


Fig. 6. (a) Bias and (b) square root of the mean squared error for $p = 10$, $\rho = 0.2$, and $n = 100$. The true value of CR_α under classical normality assumptions is 0.714.

5 Example

To illustrate the usefulness of a robust Cronbach's alpha coefficient for a real data set, let us consider a subset of a larger data set collected by A. Nolle from the University of Dortmund. The data set listed in Table 5 gives the answers of 23 bavarian teachers for the following three items.

- Item 1: "I possess knowledge of the basic principles of education."
- Item 2 "I can define education and knowledge and can distinguish them from each other."
- Item 3 "I can list basic theories of socialization."

The items were measured on an ordinal scale with 5 values (1=good knowledge, ..., 5=unknown). Hence, the classical assumption of normality is surely not fulfilled here. The Cronbach's alpha coefficients based on S, RMCD, S_{bw}, and T3 are 0.55, 0.70, 0.62, and 0.65 for this data set, respectively. From a data analytic point of view, simple sensitivity measures are often useful, as they describe the impact of a single observation onto the quantity one is studying.

An indexplot of the sensitivities for Cronbach's alpha coefficient defined by

$$\alpha_n(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n) - \alpha_{n-1}(y_1, \ldots, y_n)$$

based on the classical approach (S) and Tukey's S-estimator (S_{bw}) is given in Figure 8. It is obvious, that the answers for teacher 16 − who has not much knowledge with respect to item 1, but reasonable knowledge w.r.t. to items 2 and 3 − have much higher impact on the estimation of Cronbach's alpha coefficient than on its robust alternative. In contrast to that, the other sensitivity values were very similar for both approaches. Just for comparison

Table 1
Data set: bavarian teachers.

ID No.	Item 1	Item 2	Item3
1	1	2	2
2	2	3	2
3	3	3	4
4	2	2	3
5	1	2	1
6	3	3	4
7	2	2	4
8	3	2	4
9	3	2	4
10	2	2	3
11	3	3	3
12	2	2	4
13	2	2	4
14	2	3	5
15	3	4	4
16	4	2	2
17	3	3	4
18	1	1	3
19	1	2	4
20	2	2	3
21	1	3	3
22	2	3	4
23	2	2	3

reasons, the Cronbach's alpha coefficients based on S, RMCD, S_{bw}, and T3 are 0.67, 0.74, 0.67, and 0.70 for the data set without observation 16. As 0.70 is often used as a cut-off value for Cronbach's alpha this data set illustrates that even a single observation may have a high impact on the estimation of Cronbach's alpha but only a much smaller impact if the estimation is based on a robust method. Of course, we do not propose to bluntly drop out any outliers, but a robust method is helpful to identify observations which are far

away from the bulk of the data and it also allows to assess their impact on the data analysis.

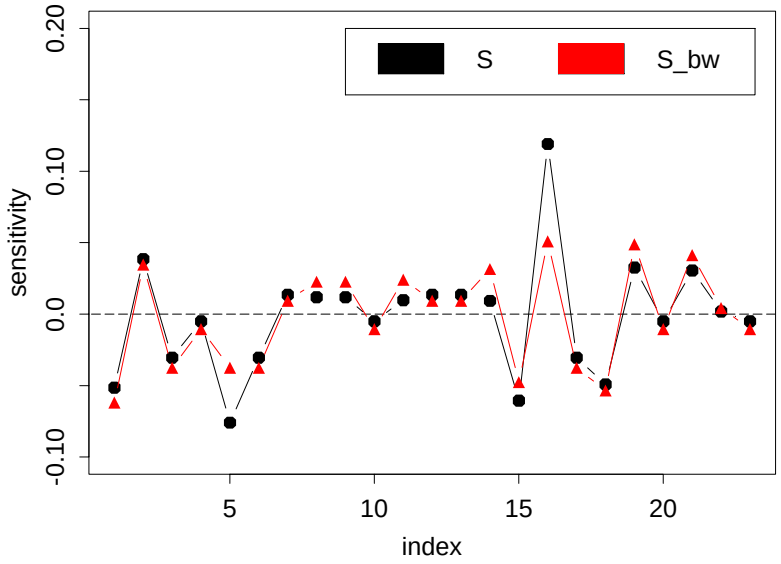


Fig. 7. Indexplot of sensitivities for the data set of bavarian teachers.

6 Discussion

The reliability measure Cronbach's alpha is non-robust and even a single observation can have a high impact on this coefficient. Therefore, we proposed robust alternatives, which have good robustness properties, e.g. a bounded influence function, perform well in a simulation study with respect to bias and mean squared error, and are easy to compute with common statistical software packages as SAS, S-PLUS or R.

Acknowledgments

We like to thank Prof. David M. Rocke (University of California, Davis) for making available his program to compute the S-estimator and Alexander Nolle (University of Dortmund, IFS) for making available the data set we used in section 5.

References

[1] Bravo, G. and Potvin, L. (1991), Estimating the Reliability of Continuous Measures with Cronbach's Alpha or the Intraclass Correlation Coefficient: Toward the Integration of Two Traditions, *J. Clin. Epidemiol.*, **44**, 381–390.

[2] Butler, R.W., Davies, P.L., and Jhun, M. (1993), Asymptotics for the Minimum Covariance Determinant Estimator, *The Annals of Statistics*, **21**, 1385–1400.

[3] Cronbach, L.J. (1951), Coefficient Alpha and the Internal Structure of Tests, *Psychometrika*, **16**, 297–334.

[4] Croux, C. and Dehon, C. (2002), Analyse Canonique basèe sur des Estimateurs Robustes de la Matrice de Covariance, *La Revue de Statistique Apliquée*, **2**, 5–26.

[5] 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.

[6] Croux, C. and Haesbroeck, G. (2000), Principal Component Analysis based on Robust Estimators of the Covariance or Correlation Matrix: Influence Function and Efficiencies, *Biometrika*, **87**, 603–618.

[7] Davies, L. (1987), Asymptotic Behavior of S-Estimators of Multivariate Location Parameters and Dispersion Matrices, *The Annals of Statistics*, **15**, 1269–1292.

[8] Feldt L.S. (1965), The Approximate Sampling Distribution of Kuder-Richardson Reliability Coefficient Twenty, *Psychometrika*, **30**, 357–370.

[9] Guttman, L. (1953), Reliability Formulas That Do Not Assume Experimental Independence, *Psychometrika*, **18**, 225–239.

[10] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics: the Approach based on Influence Functions*, New York: John Wiley.

[11] Kent, J.T., and Tyler, D.E. (1991), Redescending M-estimates of Multivariate Location and Scatter, *The Annals of Statistics,* **19**, 2102–2119.

[12] Kraemer, H.C. (1981), Extension of Feldt's Approach to Testing Homogeneity of Coefficients of Reliability, *Psychometrika*, **46**, 41–45.

[13] Kuder, G.F. and Richardson, M.W. (1937), The Theory of the Estimation of Test Reliability," *Psychometrika*, **2**, 151–160.

[14] Lopuhaä, H.P. (1989), On the Relation Between S-Estimators and M-Estimators of Multivariate Location and Covariance, *The Annals of Statistics*, **17**, 1662–1683.

[15] Lopuhaä, H.P. (1999), Asymptotics of Reweighted Estimators of Multivariate Location and Scatter, *The Annals of Statistics*, **27**, 1638–1665.

[16] Lopuhaä, H.P. and Rousseeuw, P.J. (1991), Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices, *The Annals of Statistics*, **19**, 229–248.

[17] Maronna, R.A. (1976), Robust M-Estimates of Multivariate Location and Scatter, *The Annals of Statistics*, **4**, 51–67.

[18] Pison, G., Van Aelst, S., and Willems, G. (2002), Small Sample Corrections for LTS and MCD, *Metrika*, **55**, 111-123.

[19] Pison, G., Rousseeuw, P.J., Filzmoser, P., and Croux, C. (2003), Robust Factor Analysis, *Journal of Multivariate Analysis*, **84**, 145-172.

[20] Rocke, D.M., and Woodruff, D.L. (1993), Computation of Robust Estimates of Multivariate Location and Shape, *Statistica Neerlandica,* **47**, 27–42.

[21] Rousseeuw, P.J. (1984), Least Median of Squares Regression, *Journal of the American Statistical Association*, **79**, 871–880.

[22] Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* New York: John Wiley.

[23] Rousseeuw, P.J., Van Aelst, S., Van Driessen, K., and Agullò, J. (2004) Robust Multivariate Regression, *Technometrics*, **46**, 293-305.

[24] Rousseeuw, P.J., and Van Driessen, K. (1999), A Fast Algorithm for the Minimum Covariance Determinant Estimator, *Technometrics*, **41**, 212–223.

[25] Rousseeuw, P.J., and van Zomeren, B.C. (1990), Unmasking Multivariate Outliers and Leverage Points, *Journal of the American Statistical Association*, **85**, 633–651.

[26] Ruppert, D. (1992), Computing S-estimators for Regression and Multivariate Location/Dispersion, *Journal of Computational and Graphical Statistics,* **1**, 253–270.

[27] Ten berge, J.M.F. and Zegers F.E. (1978), A Series of Lower Bounds to the Reliability of a Test, *Psychometrika*, **43**, 575–579.

[28] Wilcox, R.R. (1992), Robust Generalizations of Classical Test Reliability and Cronbach's Alpha, *British Journal of Mathematical and Statistical Psychology*, **45**, 239–254.