Robust Estimation of Cronbach's Alpha

Andreas Christmann and Stefan Van Aelst

August 26, 2002

Andreas Christmann is statistician at the University of Dortmund, HRZ, 44421 Dortmund, Germany. The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged. Stefan Van Aelst is professor at the Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, B-9000 Gent, Belgium. E-mail: Stefan.VanAelst@rug.ac.be (corresponding author)

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.*

1 Introduction

We consider the problem of constructing a measure of reliability for a set of items such as in a test. Cronbach (1951) proposed the coefficient alpha as a lower bound to the reliability coefficient in classical test theory (see also Kuder and Richardson, 1937). This popular measure has been investigated further by e.g. Feldt (1965), Ten Berge and Zegers (1978), Kraemer (1981), and Bravo and Potvin (1991).

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

$$\begin{aligned} \alpha_n^C &= \frac{p}{p-1} \frac{\operatorname{Var}\left(\sum_{j=1}^p Y_j\right) - \sum_{j=1}^p \operatorname{Var}(Y_j)}{\operatorname{Var}\left(\sum_{j=1}^p Y_j\right)} \\ &= \frac{p}{p-1} \frac{\sum\sum_{i \neq j} \operatorname{Cov}(Y_i, Y_j)}{\operatorname{Var}\left(\sum_{j=1}^p Y_j\right)} \\ &= \frac{p}{p-1} \left[1 - \frac{\sum_{j=1}^p \sigma_j^2}{\sum_{j=1}^p \sum_{k=1}^p \sigma_{jk}} \right] \quad , \end{aligned} \tag{1}$$

where σ_j^2 is the variance of item Y_j and σ_{jk} is the covariance of the pair (Y_j, Y_k). It has been shown that Cronbach's alpha is always a lower bound of reliability (Gutman 1953).

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. Hampel et al. 1986). 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 (1992) 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 (Maronna 1976, Kent and Tyler 1991), the minimum volume ellipsoid and minimum covariance determinant estimator (Rousseeuw 1984), and S-estimators (Davies 1987, Rousseeuw and Leroy 1987, Lopuhaä 1989).

Recently, robust multivariate statistical methods based on robust estimation of location and scatter have been developed and investigated such as factor analysis (Pison et al. 2002a), principal component analysis (Croux and Haesbroeck 2000), canonical correlation analysis (Croux and Dehon 2001) and multivariate regression (Rousseeuw et al. 2001). 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 a clear advantage over the proposal of Wilcox (1992) 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 is given by $C_n^0 = c_n \frac{1}{h}\sum_{j=1}^h (y_{i_j} - t_n^0)(y_{i_j} - t_n^0)^t$, which is essentially the classical covariance estimator based on the data of the optimal subset. The factor c_n makes the MCD consistent and unbiased at finite-samples (see Pison et al. 2002b).

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. Rousseeuw and Leroy (1987) 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 (Butler, Davies, and Jhun 1993, Croux and Haesbroeck 1999) which means that a small amount of contamination at a certain place can only have a bounded effect on the MCD estimates, see Hampel et al. (1986) 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

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

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 $\delta = 0.025$ (Rousseeuw and van Zomeren 1990). Similarly as for the initial MCD, the factor d_n makes the MCD consistent and unbiased at finite-samples (Pison et al. 2002b). The reweighted MCD estimators (RMCD) preserve the breakdown value (Lopuhaä and Rousseeuw 1991) and the bounded influence function (Lopuhaä 1999) of the initial MCD estimators but have a higher efficiency as shown by Croux and Haesbroeck (1999). Recently, Rousseeuw and Van Driessen (1999) 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

$$\frac{1}{n} \sum_{i=1}^n \rho(\sqrt{(y_i - t_n)^t C_n^{-1} (y_i - t_n)}\,) \leq b, \tag{3}$$

over all $t_n \in \mathrm{I\!R}^p$ 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. Lopuhä (1989) 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 Lopuhaä 1989). The most popular choice of ρ function is Tukey's biweight function which is given by

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

Its derivative is given by

$$\psi_c(t) = t\left(1 - \frac{t^2}{c^2}\right)^2 I(|t| < c),\ , \quad t \in \mathrm{I\!R}. \tag{5}$$

The tuning constant c in the ρ function (4) 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 (Davies 1987, Lopuhaä 1989). Efficient algorithms to compute S-estimators have been constructed by Ruppert (1992) and Rocke and Woodruff (1993). 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 by Kent and Tyler (1991). 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 $Y_n = \{y_i; i = 1, \ldots, n\} \subset \mathrm{I\!R}^p$ 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

$$\alpha_n^R = \frac{p}{p-1} \frac{\sum\sum_{j \neq k} c_{jk}}{\sum\sum_{j,k} c_{jk}} \tag{6}$$

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

$$f_{\mu,\Sigma}(y) = \frac{g(y-\mu)^t \Sigma^{-1}(y-\mu)}{\sqrt{\det(\Sigma)}} \tag{7}$$

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

$$\alpha = \frac{p}{p-1} \frac{\sum \sum_{j \neq k} \sigma_{jk}}{\sum \sum_{j,k} \sigma_{jk}}. \tag{8}$$

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 (6) is a consistent estimator of α (in probability or almost surely). Consistency of robust location/scatter estimators at elliptically symmetric distributions has been shown by Butler, Davies, and Jhun (1993) for the MCD, by Lopuhaä (1999) for the RMCD and by Davies (1987) and Lopuhaä (1989) 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 (6) 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

$$\begin{aligned} IF(y; \alpha_C^R, F_{\mu,\Sigma}) &= \lim_{\varepsilon \downarrow 0} \frac{\alpha_C^R((1-\varepsilon)F_{\mu,\Sigma} + \varepsilon \Delta_y) - \alpha_C^R(F_{\mu,\Sigma})}{\varepsilon} \\ &= \frac{\partial}{\partial \varepsilon} \alpha_C^R((1-\varepsilon)F_{\mu,\Sigma} + \varepsilon \Delta_y)\Big|_{\varepsilon=0}. \end{aligned} \tag{9}$$

See Hampel et al. (1986) for further details. For scatter matrix estimators possessing an influence function the following result can easily be derived from (6) by computing the derivate of α_C^R with respect to ε as in (9).

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 (4) for a multivariate standard normal distribution $F = N(\mathbf{0}, \mathbf{I})$ which is given by

$$\mathrm{IF}(y; C^S, F) = \frac{2}{\gamma_3}\,(\rho(||y||) - b_0) \,+\, \frac{1}{\gamma_1}\, p\, \psi(||y||)\, ||y|| \left(\frac{yy^t}{||y||^2} - \frac{1}{p}\mathbf{I} \right), \tag{10}$$

where

$$\gamma_1 = (p+2)^{-1} \mathrm{E}_F \left[\psi'(||Y||)\, ||Y||^2 + (p+1)\psi(||Y||)\, ||Y|| \right], \tag{11}$$

$$\gamma_3 = \mathrm{E}_F \left[\psi(||Y||)\, ||Y|| \right]. \tag{12}$$

(see Lopuhaä (1989), 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 empirical versions of the influence function for finite sample sizes. Here, we consider the *empirical influence function* EIF_n and the *sensitivity curve* SC_n, c.f. Hampel et al. (1986, p. 93). The empirical influence function and the sensitivity

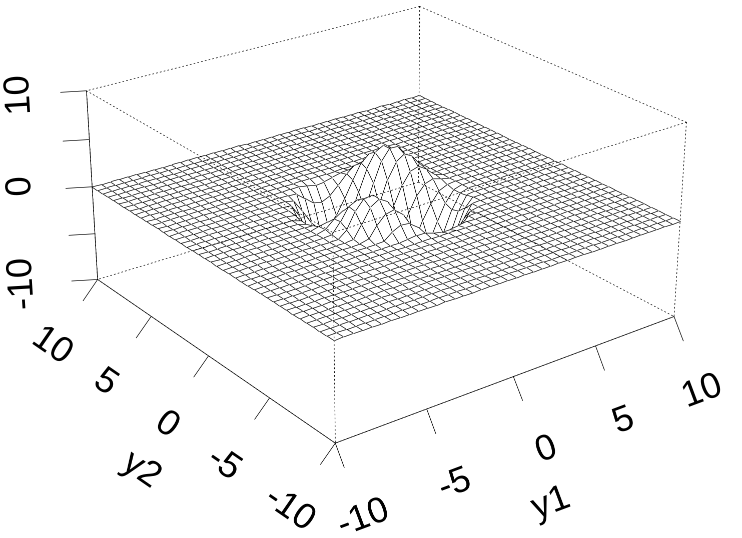

Figure 1: Influence function of Cronbach's alpha based on the S-estimator S_{bw} at the bivariate normal distribution.

curve of Cronbach's alpha α_n given a multivariate data set $(y_1, \ldots, y_{n-1})$ are defined by

$$\mathrm{EIF}_n(y) = \alpha_n(y_1, \ldots, y_{n-1}, y)\,, \quad y \in \mathbb{R}^p, \tag{13}$$

and

$$\mathrm{SC}_n(y) = n\,\left[\alpha_n(y_1, \ldots, y_{n-1}, y) - \alpha_{n-1}(y_1, \ldots, y_{n-1})\right]\,, \quad y \in \mathbb{R}^p. \tag{14}$$

Hence, EIF_n describes the behavior of the estimate if one arbitrary data point y is added, whereas SC_n is a scaled version of EIF_n.

Empirical influence functions and sensitivity curves of Cronbach's alpha based on empirical (co)variances and its robustifications based on robust estimates of the covariance matrix are given in the upper left subplots of Figures 2 and 3 for the bivariate standard normal distribution, respectively. Note that due to different magnitudes of the empirical influence function and of the sensitivity curves the scaling of the z-axes in the plots are not identically for all four estimates. Besides the classical Cronbach's alpha based on the empirical covariance matrix S, we also consider robust Cronbach's alpha based on RMCD and the S-estimator S_{bw} (both with an asymptotical breakdown point of 25%), and the M-estimator T3. Both figures show that the impact of even one single additional observation can be extremely large for the original definition of Cronbach's alpha, whereas the robustifications behave much more stable. From the empirical influence functions shown in Figure

2 we see that even the extreme values of -1 or $+1$ for the classical Cronbach's alpha are possible, although all data points with the exception of a single outlier are generated from the bivariate standard normal distribution for which the theoretical value of Cronbach's alpha coefficient is of course equal to zero. In contrast to that, the three robust measures behave much more reliable in this respect. Especially the sensitivity curves based on RMCD and S_{bw} are very stable for observations far away from the bulk of the data, cf. Figure 3. Note that the influence function of Cronbach's alpha based on the S-estimator S_{bw} given in Figure 1 and the corresponding sensitivity curve shown in Figure 3 are very similar, although we used only a moderate sample size of $n = 100$ to construct the latter. Cronbach's alpha based on Kent and Tyler's M-estimator T3 shows a smooth and more robust behavior than the classical estimator, but it is not as robust as the other two estimators based on RMCD and S_{bw} for extreme outliers. In contrast to Figure 3, the sensitivity curves for Cronbach's alpha and its robustifications are shown in Figure 4 at a bivariate normal distribution with mean vector $\mathbf{0}$, both variances equal to 1, and a covariance of 0.5. The corresponding sensitivity curves are qualitatively similar in both figures. Please note, that the sensitivity curves of the robust Cronbach's alpha coefficient based on RMCD or on the S-estimator are constant outside a *circle* with midpoint approximately equal to the true mean vector $\mathbf{0}$ in Figure 3, whereas the sensitivity curves of these robust Cronbach's alpha coefficients in Figure 4 are constant outside an *ellipse*. This is of course due to the non-zero correlation in the latter situation.

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

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

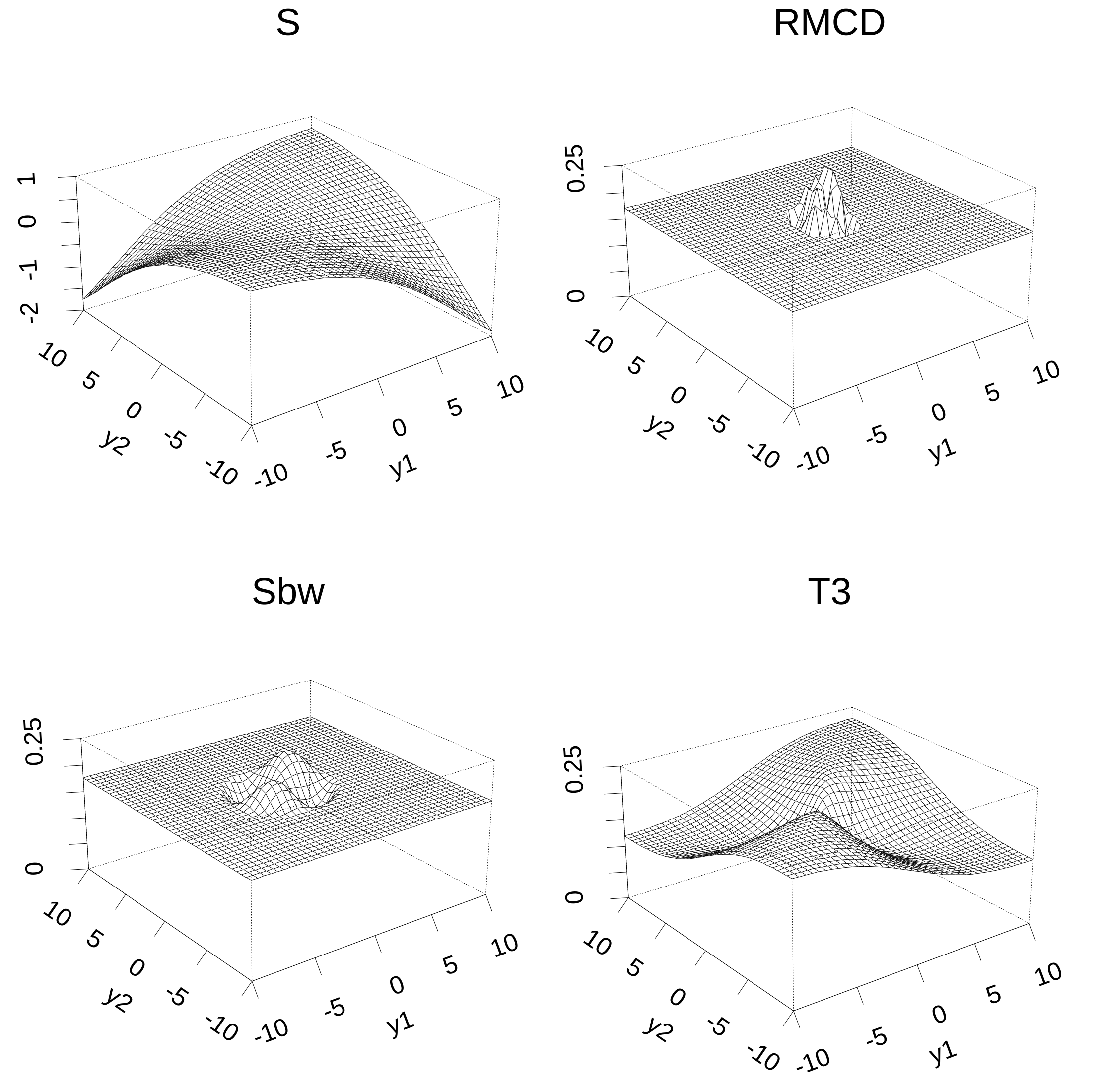

Figure 2: Empirical influence functions for a 2−dimensional data set with $n = 100$ observations simulated from $F = \mathrm{N}(\mathbf{0}, \mathbf{I})$.

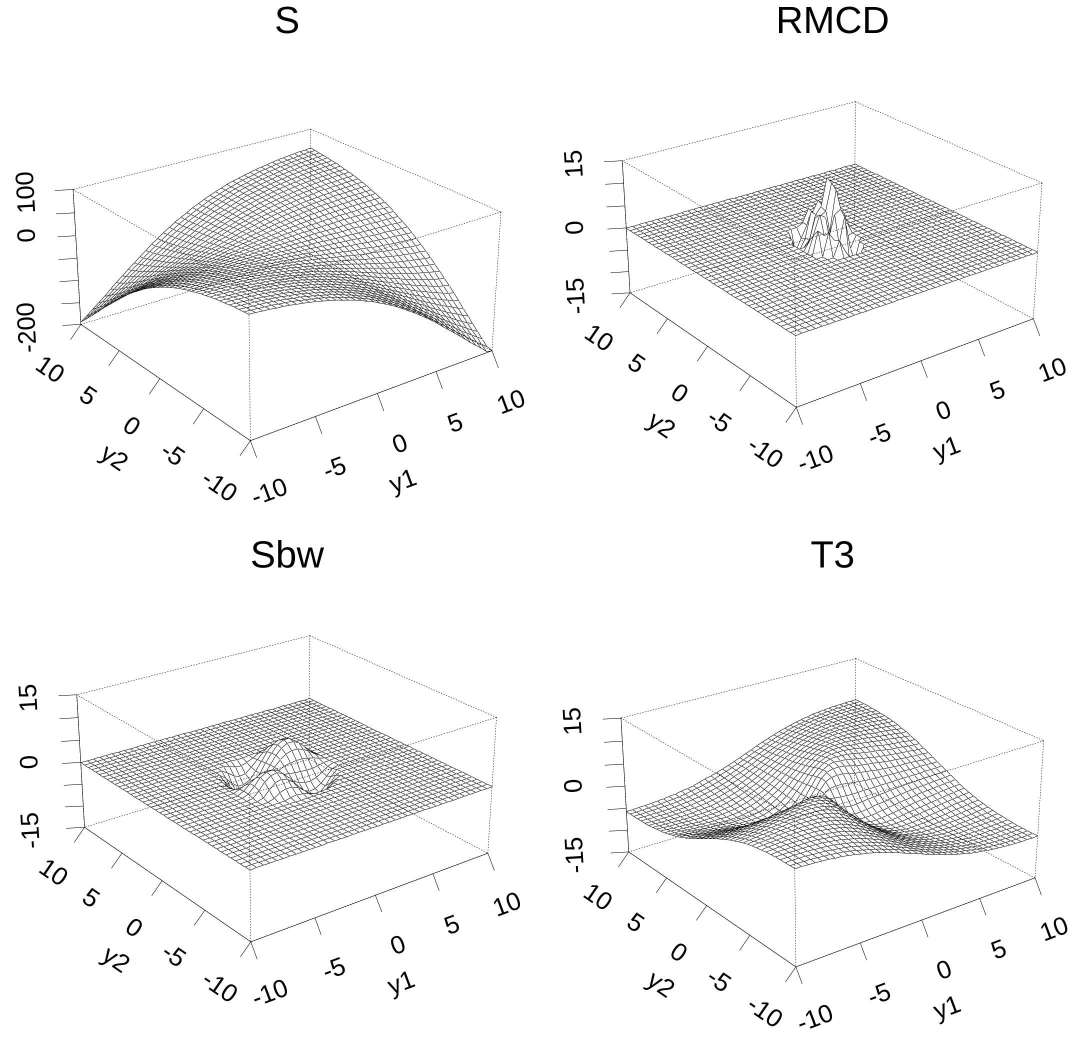

Figure 3: Sensitivity curves for a 2−dimensional data set with $n = 100$ observations simulated from $F = \mathrm{N}(\mathbf{0}, \mathbf{I})$.

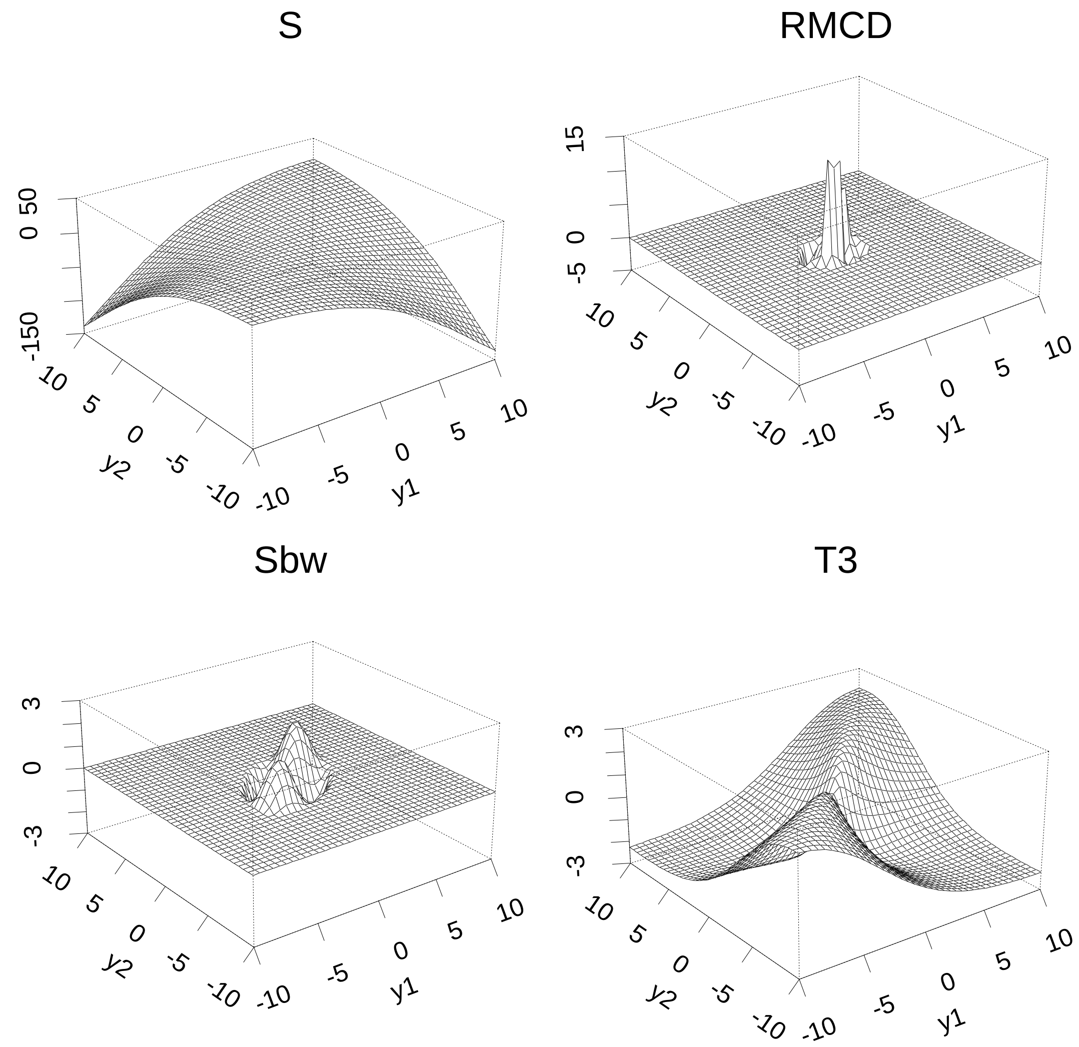

Figure 4. Sensitivity curves for a 2−dimensional data set with $n = 100$ observations from $F = \mathrm{N}(\mathbf{0}, \Sigma)$, where $\mathrm{Var}(Y_1) = \mathrm{Var}(Y_2) = 1$, and $\mathrm{Cov}(Y_1, Y_2) = \rho = 0.5$.

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 $X_1, \ldots, X_n$ 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 $\{i = 1 \text{ or } j = 1 \text{ and } i \neq j\}$, and $\sigma_{ij} = \rho$, if $\{i > 1,\ j > 1 \text{ and } i \neq j\}$. We use $\delta = 0.05, 0.10$, and 0.20 as 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 $F = t_3(\mu, \Sigma)$
- δ% M1: contamination model 1 with different covariance matrix:
  $F = (1 - \delta)N(\mu, \Sigma) + \delta N(\mu, \Sigma_1)$
- δ% M2: contamination model 2 with different location parameter and covariance 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 5. The data points generated from the contamination part of the distributions are marked as dots. For each of the sample sizes 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} and T_3. The main results of the simulations are summarized in Tables 1 to 4 and Figures 6 and 7. 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, Table 1) or underestimate (contamination models 1 and 2, Table 3) the reliability of a score. 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. Hampel et al. (1986, 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

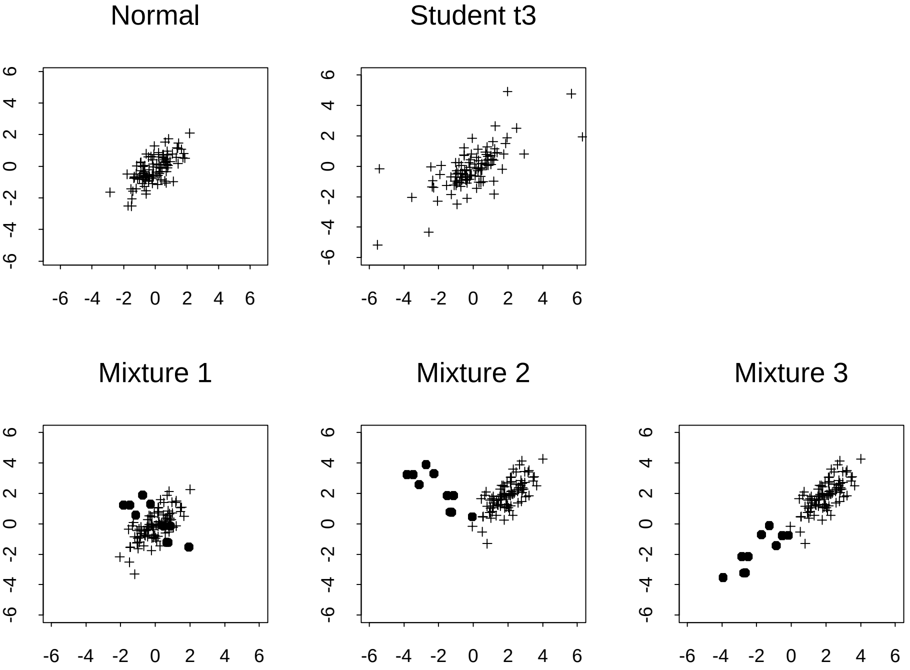

Figure 5. Scatterplots of simulated data for $p = 2$, $n = 100$, $\rho = 0.8$, and $\delta = 10\%$.

the same mean vector but a different covariance matrix than the main part of the mixture distribution, see Tables 3 and 4. 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 measure the reliability of a score in a more stable manner 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 MCD estimator, which often has a higher bias and a higher mean squared error. Hence, we will not consider the MCD approach 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 $\delta = 20\%$. Moreover, this robust method performed almost as good as the classical estimator, if the assumption of normality is fulfilled. The application of 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 (1990) (see Figure 2). This behavior of T3 coincides with the properties of the sensitivity curves and empirical influence curves given in section 3.

Table 1: Bias for several estimators of Cronbach's α, $p = 2$. True value under classical normality assumption is 0. All values are multiplied by 10^3.

ρ	n	model	S	MCD	RMCD	$\mathrm{S_{bw}}$	T3
0	40	N	−61	−273	−122	−67	−66
		t_3	−159	−249	−177	−66	−58
		5% M1	−48	−322	−127	−57	−58
		5% M2	−16	−260	−115	−60	−42
		5% M3	602	−261	−112	−45	226
		10% M1	−64	−288	−143	−83	−78
		10% M2	−32	−187	−115	−51	−39
		10% M3	741	−185	−111	29	463
		20% M1	−75	−312	−137	−78	−79
		20% M2	−24	−145	−90	−31	−29
		20% M3	836	−127	−85	806	765
0	100	N	−16	−119	−40	−17	−17
		t_3	−132	−69	−32	−17	−14
		5% M1	−25	−117	−49	−26	−27
		5% M2	−18	−94	−46	−31	−26
		5% M3	598	−93	−45	−11	252
		10% M1	−23	−90	−37	−31	−29
		10% M2	−16	−94	−42	−26	−22
		10% M3	739	−92	40	87	480
		20% M1	−37	−113	−49	−38	−37
		20% M2	−20	−64	−53	−23	−24
		20% M3	834	−54	−34	806	766
0	500	N	−6	−26	−12	−9	−9
		t_3	−45	−3	−1	1	−3
		5% M1	−7	−32	−12	−9	−8
		5% M2	−7	−27	−13	−8	−8
		5% M3	602	−27	−11	8	266
		10% M1	3	−16	0	4	3
		10% M2	0	−8	1	0	0
		10% M3	743	−7	4	121	495
		20% M1	−4	−26	−4	−3	−4
		20% M2	−1	−6	−2	−2	−1
		20% M3	837	−3	8	809	771

Table 2: Square root of mean squared error for several estimators of Cronbach's α, $p = 2$. All values are multiplied by 10^3.

ρ	n	model	S	MCD	RMCD	S_{bw}	T3
0	40	N	367	1064	611	397	397
		t_3	845	976	723	476	429
		5% M1	347	1085	628	393	391
		5% M2	305	941	590	377	344
		5% M3	612	936	586	394	357
		10% M1	361	1074	663	415	414
		10% M2	308	810	576	367	324
		10% M3	745	798	565	429	501
		20% M1	377	1079	653	405	404
		20% M2	277	637	500	302	298
		20% M3	837	563	495	808	768
0	100	N	204	545	293	221	218
		t_3	688	424	323	261	236
		5% M1	214	556	310	233	234
		5% M2	208	492	295	238	224
		5% M3	603	491	293	237	303
		10% M1	210	560	302	234	233
		10% M2	188	459	275	217	200
		10% M3	741	447	274	268	492
		20% M1	224	531	289	235	233
		20% M2	180	363	288	192	189
		20% M3	835	334	279	807	768
0	500	N	8	51	14	10	10
		t_3	302	189	144	119	108
		5% M1	91	236	121	102	100
		5% M2	89	214	119	98	93
		5% M3	603	214	117	102	276
		10% M1	87	219	111	93	94
		10% M2	82	180	109	89	83
		10% M3	744	179	106	161	497
		20% M1	92	230	118	99	99
		20% M2	74	146	109	78	76
		20% M3	837	136	108	810	771

Table 3: Bias for several estimators of Cronbach's α, $p = 2$. True value under classical normality assumption is 0.667. All values are multiplied by 10^3.

ρ	n	model	S	MCD	RMCD	$\mathrm{S_{bw}}$	T3
0.5	40	N	−14	−79	−36	−18	−17
		t_3	−59	−67	−52	−23	−23
		5% M1	−62	−110	−59	−52	−53
		5% M2	−201	−64	−28	−16	−62
		5% M3	163	−65	−27	6	73
		10% M1	−114	−138	−93	−88	−94
		10% M2	−309	−39	−9	−30	−113
		10% M3	220	−38	−6	72	147
		20% M1	−222	−245	−191	−187	−193
		20% M2	−465	−28	−19	−272	−285
		20% M3	258	−24	4	247	237
0.5	100	N	−5	−42	−10	−5	−6
		t_3	−38	−30	−21	−10	−8
		5% M1	−53	−74	−41	−39	−44
		5% M2	−184	−33	−5	−6	−52
		5% M3	166	−33	−5	18	80
		10% M1	−102	−113	−81	−80	−86
		10% M2	−302	−26	−2	−28	−112
		10% M3	218	−25	2	90	150
		20% M1	−216	−206	−176	−181	−191
		20% M2	−446	−6	2	−254	−267
		20% M3	257	−9	23	247	237
0.5	500	N	1	−7	0	0	1
		t_3	−9	−7	−5	−2	−2
		5% M1	−44	−38	−30	−32	−35
		5% M2	−173	−5	3	0	−44
		5% M3	168	−5	3	25	85
		10% M1	−93	−75	−66	−70	−76
		10% M2	−287	−1	7	−21	−103
		10% M3	220	−2	9	102	153
		20% M1	−202	−163	−155	−165	−175
		20% M2	−434	9	16	−241	−254
		20% M3	258	4	36	249	239

Table 4: Square root of mean squared error for several estimators of Cronbach's α, $p = 2$. All values are multiplied by 10^3.

ρ	n	model	S	MCD	RMCD	S_{bw}	T3
0.5	40	N	119	313	192	134	132
		t_3	282	268	230	151	134
		5% M1	152	349	209	157	155
		5% M2	277	285	178	130	152
		5% M3	172	284	179	132	120
		10% M1	206	371	250	190	193
		10% M2	378	244	164	139	189
		10% M3	222	240	167	148	162
		20% M1	301	488	334	274	273
		20% M2	523	199	176	355	349
		20% M3	258	194	178	249	239
0.5	100	N	67	193	96	73	74
		t_3	263	168	129	93	86
		5% M1	97	210	111	89	92
		5% M2	218	168	90	73	94
		5% M3	169	168	91	77	96
		10% M1	138	248	146	123	126
		10% M2	328	142	88	86	143
		10% M3	219	142	90	116	154
		20% M1	248	326	226	214	222
		20% M2	467	105	81	283	289
		20% M3	257	107	97	248	238
0.5	500	N	31	75	39	33	33
		t_3	103	61	46	39	36
		5% M1	58	91	54	50	52
		5% M2	180	72	38	33	57
		5% M3	169	72	39	43	89
		10% M1	103	118	83	81	87
		10% M2	292	65	39	43	110
		10% M3	220	65	40	107	154
		20% M1	209	200	168	174	183
		20% M2	438	50	39	247	259
		20% M3	258	50	54	249	239

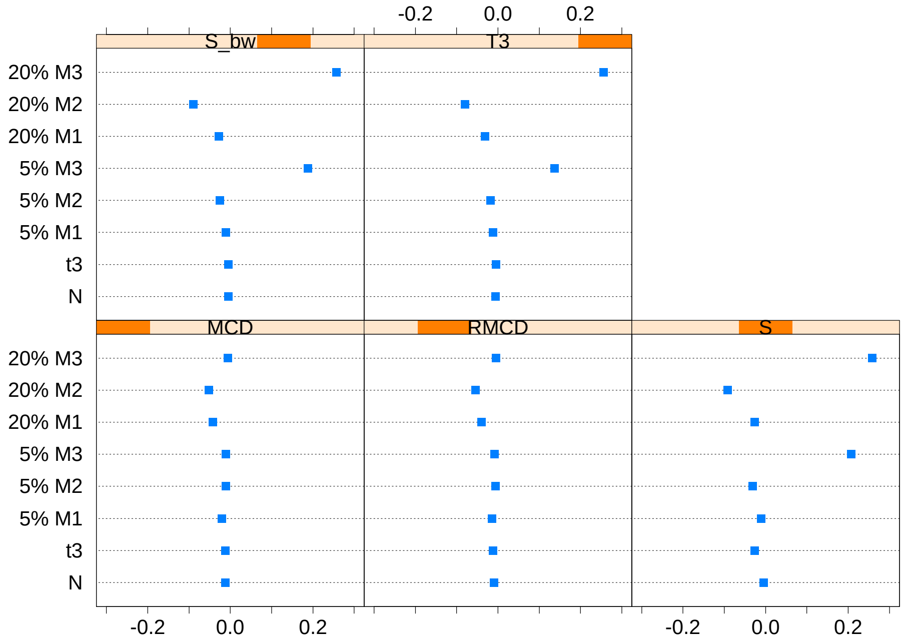

Figure 6. Bias for several estimators of Cronbach's α for $p = 10$, $\rho = 0.2$, and $n = 100$. The true value of CR_α under classical normality assumptions is 0.714.

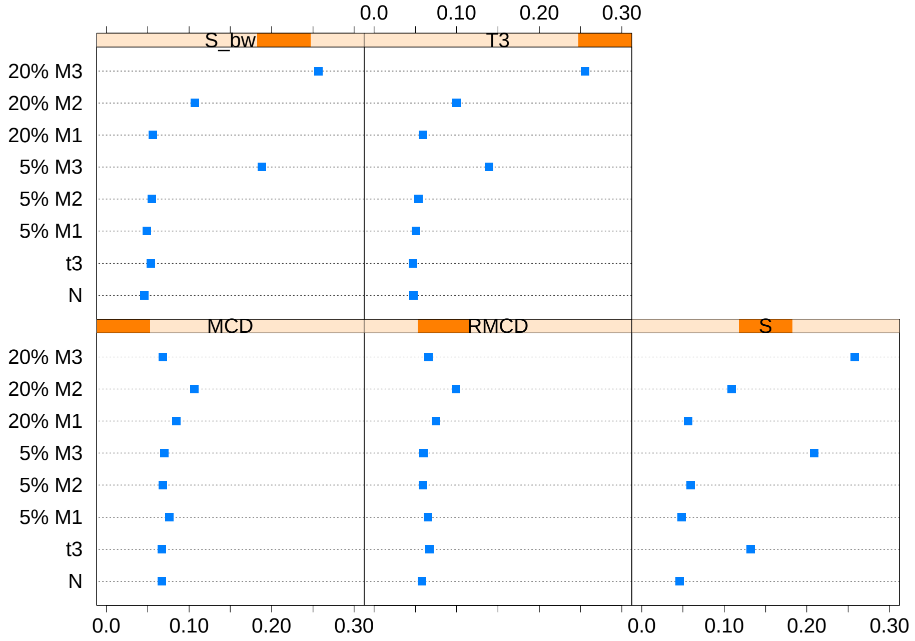

Figure 7. Square root of the mean squared error for several estimators of Cronbach's α for $p = 10$, $\rho = 0.2$, and $n = 100$. The true value of CR_α under classical normality assumptions is 0.714.

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

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

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 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.

Table 5: 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

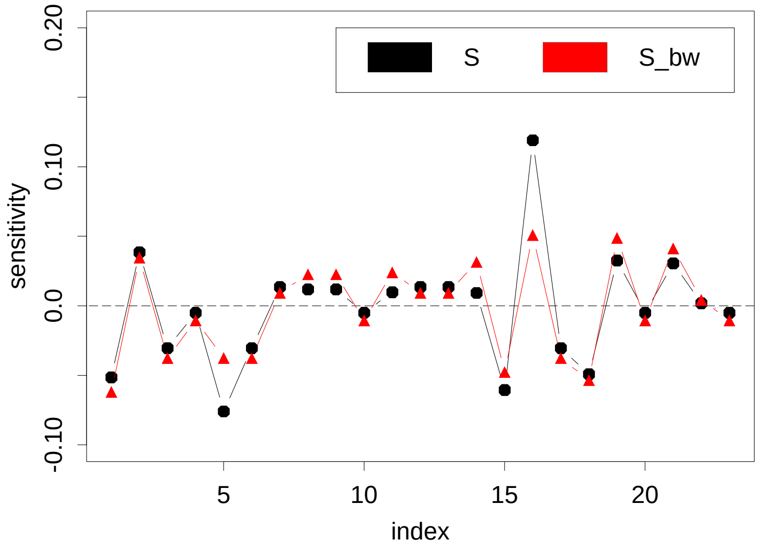

Figure 8. Indexplot of sensitivities for the data set of bavarian teachers.

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

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.

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

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

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.

Croux, C., and Haesbroeck, G. (1999), "Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator," *Journal of Multivariate Analysis*, 71, 161–190.

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

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

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

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

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.

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

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

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

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.

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

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.

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

Pison, G., Rousseeuw, P.J., Filzmoser, P., and Croux, C. (2002a), "Robust Factor Analysis," *Journal of Multivariate Analysis*, to appear.

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

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

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

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

Rousseeuw, P.J., Van Aelst, S., Van Driessen, K., and Agullò, J. (2001) "Robust Multivariate Regression," submitted.

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

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

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

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.

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