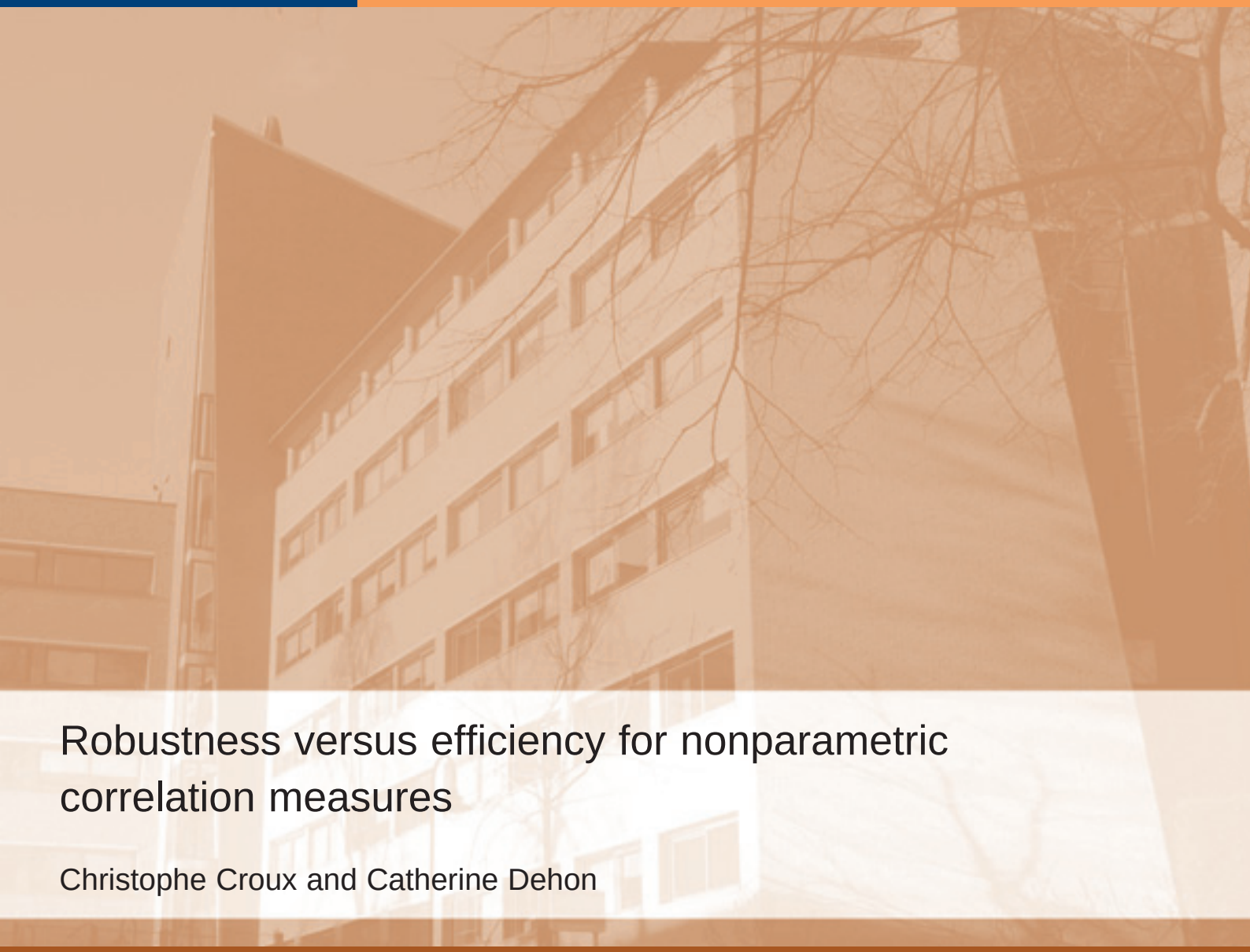
KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics



Robustness versus efficiency for nonparametric correlation measures

Christophe Croux and Catherine Dehon

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0803

Robustness versus efficiency for nonparametric correlation measures

Christophe Croux *
K.U.Leuven

Catherine Dehon†
Université libre de Bruxelles

Abstract

Nonparametric correlation measures at the Kendall and Spearman correlation are widely used in the behavioral sciences. These measures are often said to be robust, in the sense of being resistant to outlying observations. In this note we formally study their robustness by means of their influence functions. Since robustness of an estimator often comes at the price of a loss in precision, we compute efficiencies at the normal model. A comparison with robust correlation measures derived from robust covariance matrices is made. We conclude that both Spearman and Kendall correlation measures combine good robustness properties with high efficiency.

Keywords: Asymptotic Variance, Correlation, Gross-Error Sensitivity, Influence function, Kendall correlation, Robustness, Spearman correlation.

*Faculty of Business & Economics and Leuven Statistics Research Centre, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium. E-Mail:christophe.croux@econ.kuleuven.ac.be.

†ECARES and Institut de Recherche en Statistique, Université libre de Bruxelles, CP-114, Av. F.D. Roosevelt 50, B-1050 Brussels, Belgium. E-Mail:cdehon@ulb.ac.be.

1 Introduction

Pearson's correlation measure is one of the most often used statistical estimators. But its value may be seriously affected in presence of even only one outlier. The effect of an outlier on an estimator can be measured by its influence function. The influence function gives the effect that an outlying observation has on an estimator, and it is an important measure of robustness of an estimator (Hampel et al., 1986). Devlin et al. (1975) showed that the influence function of the classical Pearson correlation is unbounded, proving the lack of robustness of the latter estimator.

In this paper we provide expressions for the influence functions of other measures of correlation, in particular for the popular Spearman and Kendall correlation. We show that their influence function is bounded, hereby formally proving their robustness. This confirms the general belief that these nonparametric measure of correlation are more robust to outliers. Other robust measures of correlation have been introduced in the literature (e.g. Shevlyakov and Vilchevski, 2002; Wilcox, 1998) and a comparison with some of them is made in this paper.

Besides being robust, an estimator should also be precise, in the sense of having a high statistical efficiency. At the normal distribution the Pearson correlation measure is the most efficient. The price of using a more robust estimator is a loss of efficiency, but we would like this loss in precision to be limited. We compute the statistical efficiency at the normal distribution of the Spearman and Kendall correlation estimators, and it turns out to be above 75% for all possible values of the true correlation. Hence they provide a good compromise between robustness and efficiency.

In Section 2 we review several measures of robust correlation with focus on (i) the rank and sign based measures Spearman, Kendall and the Quadrant correlation; (ii) robust correlations derived from robust covariance matrices. Their influence function and gross-error-sensitivity are presented in Section 3. Asymptotic variances are derived in Section 4. Finally, in Section 5 we present a simulation study comparing the performance

of the different estimators of correlation in presence of outliers at finite samples. Section 6 contains the conclusions.

2 Measures of Correlation

Given a bivariate sample $\{(x_i, y_i), 1 \leq i \leq n\}$, the classical Pearson's estimator of correlation is given by

$$r_P = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{x})^2}} \tag{2.1}$$

where $\bar{x}$ and $\bar{y}$ are the sample means. To compute influence functions, it is necessary to consider the associated functional form of the estimator. Let $(X, Y) \sim H$, with H an arbitrary distribution (having second moments). The population version of Pearson's correlation measure is then given by

$$R_P(H) = \frac{E_H[XY] - E_H[X]E_H[Y]}{\sqrt{(E_H[X^2] - E_H[X]^2)(E_H[Y^2] - E_H[Y]^2)}}. \tag{2.2}$$

and the function $H \to R_P(H)$ is the functional representation of this estimator. If the sample $(x_1, y_1), \ldots, (x_n, y_n)$ has been generated according to a distribution H, then the estimator r_P, as defined in (2.1), converges in probability to $R_P(H)$. If we take as model distribution H_ρ, the bivariate normal with population correlation coefficient ρ, then we have that

$$R_P(H_\rho) = \rho.$$

The above property is called the Fisher consistency of R_P at the normal model (e.g. Maronna et al., 2006).

As an alternative to Pearson's correlation, nonparametric measures of correlation using univariate ranks and signs, have been introduced. The Quadrant correlation (Mosteller, 1946) r_Q is computed by dividing the plane in 4 quadrants, with the coordinatewise median as origin. Then r_Q equals the frequency of observations being in the first or third

quadrant, minus the frequency of observations in the second or fourth quadrant:

$$r_Q = \frac{2}{n}\sum_{i=1}^{n} \text{sign}\{(x_i - \text{median}_j(x_j))(y_i - \text{median}_j(y_j))\} - 1. \tag{2.3}$$

Here, the sign function equals 1 for positive and -1 for negative arguments. The associated functional is given by

$$R_Q(H) = 2P_H[(X - \text{median}(X))(Y - \text{median}(Y)) > 0] - 1. \tag{2.4}$$

When comparing a nonparametric correlation measure with the classical Pearson correlation, one needs to realize that they estimate different population quantities. For H_ρ the bivariate normal distribution with correlation ρ, one has (Blomqvist, 1950)

$$\rho_Q := R_Q(H_\rho) = \frac{2}{\pi}\arcsin(\rho)$$

being different from ρ, for any $\rho \neq 0$. To obtain a consistent version of the Quadrant correlation at the normal model, we apply the following transformation

$$\tilde{R}_Q(H) = \sin(\frac{1}{2}\pi R_Q(H)).$$

Another nonparametric correlation measure based on signs is Kendall's correlation (Kendall, 1938), given by

$$r_K = \frac{2}{n(n-1)}\sum_{i<j} \text{sign}\left((x_i - x_j)(y_i - y_j)\right). \tag{2.5}$$

The corresponding functional version is then

$$R_K(H) = E_H[\text{sign}(X_1 - X_2)(Y_1 - Y_2)] \tag{2.6}$$

where (X_1, Y_1) and (X_2, Y_2) are two independent copies from H. At normal distributions, it estimates the same parameter as the Quadrant's correlation (Blomqvist, 1950), so $R_K(H_\rho) = \rho_K = \rho_Q$. Hence, the Fisher consistent version of Kendall's correlation is given by

$$\tilde{R}_K(H) = \sin(\frac{1}{2}\pi R_K(H)).$$

Finally, the most popular nonparametric correlation measure is Spearman's rank correlation (Spearman, 1904), of which the sample version is simply the classical Pearson correlation computed from the ranks of the observations. Take $(X, Y) \sim H$, and denote $F(t) = P_H(X \leq t)$ and $G(t) = P_H(Y \leq t)$ the marginal cumulative distribution functions of X and Y. Then the functional version of Spearman's correlation is given by

$$R_S(H) = \text{Corr}(F(X), G(Y)) = 12E_H[F(X)G(Y)] - 3. \tag{2.7}$$

At the normal model H_ρ, we have

$$\rho_S := R_S(H_\rho) = \frac{6}{\pi} \arcsin(\frac{\rho}{2}),$$

see Moran (1948). Again we see that the Spearman correlation differs from the correlation coefficient ρ of the bivariate normal distribution. To make a comparison between different estimators at the normal model possible, we will therefore consider the transformed version of R_S:

$$\tilde{R}_S(H) = 2 \sin(\frac{1}{6}\pi R_S(H)).$$

In this paper we focus on the above nonparametric correlation measures. Robust correlations, however, are often derived from robust covariance matrix estimates (see Maronna et al., 2006; Croux & Dehon, 2002). If $C(X, Y)$ is a 2×2 robust covariance matrix computed from X and Y, then a robust correlation results immediately as

$$R_C(H) = \frac{C_{12}(X, Y)}{\sqrt{C_{11}(X, Y)C_{22}(X, Y)}}. \tag{2.8}$$

Hence, any robust bivariate covariance matrix C leads to a robust correlation coefficient. We will consider two highly robust covariance matrix estimators for C in (2.8). The S-estimator (e.g. Davies, 1987), leading to the correlation measure R_S, and the Minimum Covariance Determinant (MCD, Rousseeuw and Van Driessen, 1999), resulting in R_{MCD}. We take the MCD and the S-estimator with maximum breakdown point, i.e. 50%. The breakdown measures the maximum fraction of outliers the estimator can withstand. The MCD and S-estimator estimate (a multiple) of the population covariance matrix at the normal distribution, so $R_C(H_\rho) = \rho$.

3 Influence Function and Gross-Error-Sensitivity

As model distribution for (X, Y) we take the bivariate normal H_ρ, with correlation coefficient ρ. We assume that the population means of X and Y are equal to zero, and their variances one. Since all correlation measures considered in this paper are invariant with respect to linear transformation of X, respectively Y, the latter assumption is without loss of generality. The influence function (IF) of a statistical functional R at the model distribution H_ρ is defined as

$$\mathrm{IF}((x,y), R, H_\rho) = \lim_{\varepsilon \downarrow 0} \frac{R((1-\varepsilon)H_\rho + \varepsilon \Delta_{(x,y)}) - R(H_\rho)}{\varepsilon}$$

where $\Delta_{(x,y)}$ is a Dirac measure putting all its mass at (x, y). It can be interpreted as the infinitesimal effect that a small amount of contamination placed at (x, y) has on R, when the data come from the model distribution H_ρ. An estimator is then called B-robust if its influence function is bounded (see Hampel et al., 1986). For the Pearson correlation, Devlin et al. (1975) computed

$$\mathrm{IF}((x,y), R_P, H_\rho) = xy - \rho \frac{x^2 + y^2}{2}, \tag{3.1}$$

which is an unbounded function, showing that R_P is not B-robust. The influence functions associated to the Quadrant, Kendall and Spearman correlation can be derived in a rather straightforward way, and are given by

$$\begin{aligned}
\mathrm{IF}((x,y), R_Q, H_\rho) &= \mathrm{sign}[(x - \mathrm{median}(X))(y - \mathrm{median}(Y))] - \rho_Q & (3.2)\\
\mathrm{IF}((x,y), R_K, H_\rho) &= 2\{2P_{H_\rho}[(X-x)(Y-y) > 0] - 1 - \rho_K\} & (3.3)\\
\mathrm{IF}((x,y), R_S, H_\rho) &= -3\rho_S - 9 + 12\{F(x)G(y) + E_{H_\rho}[F(X)I(Y \geq y)] \\
&\quad + E_{H_\rho}[G(Y)I(X \geq x)]\}, & (3.4)
\end{aligned}$$

where $I(t)$ stands for the indicator function. While the expression for the IF for R_Q appeared in Shevlyakov and Vilchevski (2002), the other expressions for the IF do not seem to have been published in the printed literature, even if they are not difficult to

obtain. There is only an unpublished manuscript of Grize (1978) who listed similar expressions as above. Details on their calculation can be obtained upon request from the authors.

For comparing the numerical values of the different IF, it is important that all considered estimators estimate the same population quantity, i.e. are Fisher consistent. Figure 1 plots the influence function of R_P and of the transformed measures $\tilde{R}_Q, \tilde{R}_K$ and $\tilde{R}_S$, for $\rho = 0.5$. The analytical expressions of their IF are simply given by

$$\begin{aligned} \mathrm{IF}((x,y), \tilde{R}_Q, H_\rho) &= \frac{\pi}{2}\mathrm{sign}(\rho)\sqrt{1-\rho^2}IF((x,y), R_Q, H_\rho) && (3.5) \\ \mathrm{IF}((x,y), \tilde{R}_K, H_\rho) &= \frac{\pi}{2}\mathrm{sign}(\rho)\sqrt{1-\rho^2}IF((x,y), R_K, H_\rho) && (3.6) \\ \mathrm{IF}((x,y), \tilde{R}_S, H_\rho) &= \frac{\pi}{3}\mathrm{sign}(\rho)\sqrt{1-\frac{\rho^2}{4}}IF((x,y), R_S, H_\rho). && (3.7) \end{aligned}$$

INSERT FIGURE 1

As one can see from Figure 1, the IF of the Pearson correlation is indeed unbounded. On the other hand, the influence function for the Quadrant estimator is bounded but has jumps at the coordinate axes. This means that small changes in data points close to the median of one of the marginals, will lead to relatively large changes in the estimator. For Kendall and Spearman the influence functions are both bounded and smooth. The value of the IF for R_K and R_S increases fastest along the first bisection axis. It can be checked that for $\rho = 0$ the influence functions of Spearman and Kendall estimators are exactly the same, but they slightly differ for other values of ρ.

We also compare with the IF of the correlation estimator R_C, based on an affine equivariant covariance matrix estimator C. Croux and Haesbroeck (2000) showed that there exist a function $\gamma_C : [0, \infty[\rightarrow I\!R_+$ such that

$$\mathrm{IF}((x,y), R_C, H_\rho) = \gamma_C(d(z))\mathrm{IF}((x,y), R_P, H_\rho) \qquad (3.8)$$

with $d^2(z) = z^t\Sigma^{-1}z$, and $\Sigma = ((1,\rho)^t, (\rho,1)^t)$. For the MCD estimator, the function γ_C is given by

$$\gamma_{MCD}(t) = \frac{I(t \leq \sqrt{q_\alpha})}{P(\chi^2_6 < q_\alpha)} \text{ with } q_\alpha = \chi^2_{2,1-\alpha},$$

and $\chi^2_{2,1-\alpha}$ the $1-\alpha$ quantile of a chi-square distribution with 2 degrees of freedom, and α the trimming proportion used in the definition of the MCD. In this paper we take $\alpha = 50\%$, corresponding to the estimator C with the highest possible breakdown point. Since γ_{MCD} equals zero for large values of its argument, the IF for the corresponding correlation measure will be bounded, as is confirmed by Figure 1. But is can also be seen that, when using the MCD, the IF contains jumps and is not smooth anymore. Using the S-estimator, however, the IF for R_C will be both bounded and smooth, as can see be seen from Figure 1. For the analytical expression of γ_C for the S estimator, we refer to Lopuhaä (1989).

An influence function can be summarized in a single index, the *gross-error sensitivity* (GES), giving the maximal influence an observation has. Formally, the GES of the functional R at the model distribution H_ρ is given by

$$\text{GES}(R, H_\rho) = \sup_{(x,y)} |\text{IF}((x,y), R, H_\rho)|.$$

For example, since the classical Pearson estimator is not B-robust, $\text{GES}(R_P, H_\rho) = \infty$. The following proposition gives the GES associated to the nonparametric measures of correlation and those based on robust covariance matrices.

Proposition 1 *The gross-error sensitivity (GES) of the three transformed nonparametric correlation measures are given by*

$$\begin{aligned}
(i) \qquad GES(\tilde{R}_Q, H_\rho) &= \frac{\pi}{2}\sqrt{1-\rho^2}[\frac{2}{\pi}\arcsin(|\rho|)+1] \\
(ii) \qquad GES(\tilde{R}_K, H_\rho) &= \pi\sqrt{1-\rho^2}[\frac{2}{\pi}\arcsin(|\rho|)+1] \\
(iii) \qquad GES(\tilde{R}_S, H_\rho) &= \pi\sqrt{1-\frac{\rho^2}{4}}[\frac{6}{\pi}\arcsin(|\frac{\rho}{2}|)+1],
\end{aligned}$$

and the GES of any correlation estimator based on an affine equivariant covariance matrix estimator C by

$$(iv) \qquad GES(R_C, H_\rho) = \frac{(1-\rho^2)}{2} \sup_t \gamma_C(\sqrt{t})t.$$

The gross-error sensitivities depend on the parameter ρ in a non-linear way, and are pictured in Figure 2. A first observation is that the GES for the estimator based on the MCD is extremely large compare to the others. Using the S robust covariance matrix estimator, having a smooth IF, leads to much lower values for the GES. Surprisingly, the GES of the simple nonparametric correlation measures are of the same magnitude as the more complicated S-estimator, the latter being designed for its robustness properties. Note that for lower values of the population correlation ρ, the Quadrant is even more robust than the S-estimator. The Quadrant estimator has uniformly a lower GES than Kendall and Spearman. Kendall's measure is on his turn preferable to Spearman, although the difference in GES is negligible for smaller values of ρ. Finally, note the GES curve for Spearman is increasing in ρ and does not vanish to zero for ρ tending to one.

INSERT FIGURE 2

4 Asymptotic Variance

All considered correlation estimators are asymptotically normal, and their asymptotic variance can be computed from the influence functions derives in Section 2. Let r be the correlation estimator associated with the functional R, then at the model distribution H_ρ

$$\sqrt{n}(r-\rho) \xrightarrow{d} N(0, \mathrm{ASV}(R, H_\rho))$$

with asymptotic variance $\mathrm{ASV}(R,H) = E_H[\mathrm{IF}((X,Y),R,H)^2]$, see (Hampel et al., 1986, p. 226). The next proposition, with the proof in Appendix, presents expressions for the asymptotic variance of several correlation estimators.

Proposition 2 *At the model distribution H_ρ, we have:*

$$
\begin{aligned}
(i) \qquad ASV(R_P, H_\rho) &= (1-\rho^2)^2 && (4.1)\\
(ii) \qquad ASV(\tilde{R}_Q, H_\rho) &= (1-\rho^2)(\frac{\pi^2}{4} - \arcsin^2(\rho)) && (4.2)\\
(iii) \qquad ASV(\tilde{R}_K, H_\rho) &= \pi^2(1-\rho^2)(\frac{1}{9} - \frac{4}{\pi^2}\arcsin^2(\frac{\rho}{2})) && (4.3)\\
(iv) \qquad ASV(\tilde{R}_S, H_\rho) &= \frac{\pi^2}{9}(1-\frac{\rho^2}{4})144\{\frac{1}{144} - \frac{9}{4\pi^2}(\arcsin^2(\frac{\rho}{2})\\
&+ \frac{1}{\pi^2}\int_0^{\arcsin(\frac{\rho}{2})} \arcsin(\frac{\sin(x)}{1+2\cos(2x)})dx\\
&+ \frac{2}{\pi^2}\int_0^{\arcsin(\frac{\rho}{2})} \arcsin(\frac{\sin(2x)}{\sqrt{1+2\cos(2x)}})dx\\
&+ \frac{1}{\pi^2}\int_0^{\arcsin(\frac{\rho}{2})} \arcsin(\frac{\sin(2x)}{2\sqrt{\cos(2x)}})dx\\
&+ \frac{1}{2\pi^2}\int_0^{\arcsin(\frac{\rho}{2})} \arcsin(\frac{3\sin(x)-\sin(3x)}{4\cos(2x)})dx\} && (4.4)\\
(v) \qquad ASV(R_C, H_\rho) &= (1-\rho^2)^2 ASV(C_{12}, H_0). && (4.5)
\end{aligned}
$$

The asymptotic variances of the Pearson, Quadrant, and Kendall correlations are explicit formulas. Most complicated is the expression for Spearman's correlation, requiring standard numerical integration of univariate integrals. Note that a similar result, but expressed more generally in terms of expectations of the joint and marginal distribution functions is given in Borkowf (2002). Result (v) of proposition 2 is known (e.g. Bilodeau and Brenner, 1999, p. 230) and expresses the asymptotic variance of a correlation derived from an affine equivariant robust covariance matrix C as a function of the asymptotic variance of an off-diagonal element of C. For the MCD, for example, the asymptotic variance $\mathrm{ASV}(C_{12}, H_0)$ is computed in (Croux and Haesbroeck, 1999).

It can be verified that all asymptotic variances decrease in ρ, and tend to the value zero for ρ converging to one. In Figure 3 we plot asymptotic efficiencies (relative to Pearson correlation) as a function of ρ. Most striking are the high efficiencies for Kendall and Spearman correlation, being larger than 70% (??) for all possible values of ρ. This means

that Kendall and Spearman are at the same time B-robust, and very efficient. Comparing Kendall's with Spearman's correlation is favorable for Kendall, but the difference in efficiency is rather small, and almost negligible for ρ smaller than 0.2. On the other hand, using the Quadrant correlation leads to a high loss in efficiency.

As can be seen from Figure 3, the efficiency associated to the estimators based on robust covariance matrices is constant in ρ. For MCD, we have an efficiency of only 3.33% and for S an efficiency of 37.65%.

INSERT FIGURE 3

5 Simulation study

By means of a modest simulation experiment, we investigate two different questions. First we verify whether the finite-sample variances of the estimators are close their asymptotic counterparts, derived in Section 4. Secondly, we check how the estimators behave when outliers are introduced in the sample.

We first generate $m = 2000$ samples of size $n = 20, 50, 100, 200$ from a bivariate normal with $\rho = 0$. We did performed the same simulation exercise for several other values of ρ, with similar conclusions. For each sample j, the correlation coefficient is estimated by $\hat{\rho}_j$, one of the estimators introduced in Section 2. The mean squared error (MSE) is then computed as

$$\text{MSE} = \frac{1}{m}\sum_{j=1}^{m}(\hat{\rho}_j - \rho)^2$$

and reported in Table 1. As we can see from Table 1, the finite sample MSE converge rather quickly to the asymptotic variance (reported under the column $n = \infty$). For the S and MCD estimators convergence is slower, and we see that for MCD the finite-sample MSE is substantially smaller than the asymptotic counterpart. The simulation experiment confirms the conclusions from Section 4. Also at finite samples, the precision of the Spearman and Kendall estimators is close to the Pearson correlation. The MSE of

the Quadrant correlation is about twice as large, and the estimates derived from robust correlation measures perform even worse.

INSERT TABLE 1

The second simulation scheme is similar, but now we only generate samples of size $n = 200$, and replace a certain percentage ε of the observations by outliers. The outliers are placed at a distance equal to the square root of the 0.90 quantile of a χ^2_2 distribution, and in the direction of the 45-degree line. Indeed, as we can see from Figure 1, the influence of outliers increases fastest in that direction. The MSEs are reported in Table 2.

INSERT TABLE 2

Although we know that the MSE is smallest for the Pearson correlation if no outliers are present, we see from Table 2 that this does not hold anymore in presence of outliers. The MSE for the Pearson correlation increases quickly with the fraction of outliers, and already for 5% of outliers its MSE is by far the largest of all considered estimators. This confirms the non robustness of the Pearson correlation. A comparison of the other estimators shows that for about 5% of contamination, the MSE for Spearman and Kendall correlation remains small, but for larger, more unrealistic, amounts of contamination, there is also a substantial increase in MSE. The Quadrant estimator perform betters than the two other nonparametric correlation measures under contamination, as we can see from Table 2. The good robustness of the Quadrant correlation was already observed from Figure 2, where it has the smallest value of the gross-error sensitivity. Finally note the high robustness of the S and MCD based estimators, where the MSE remains low for even 20% of contamination. The reason for this good performance is due to the fact that the S and MCD are redescending estimators, meaning that there influence function equals zero for larger values of the observations (see Figure 1). Outliers have little effect on the S and MCD estimators, unless if they are located at very particular positions.

6 Conclusion

In this paper we study the robustness and efficiency of some widely used nonparametric measures of correlation at a bivariate normal distribution. The main conclusion is that the Spearman and Kendall correlation measures are fairly robust, while maintaining a quite high statistical efficiency. They have a bounded and smooth influence functions, and reasonably small values for the gross-error sensitivity. The Kendall correlation measure is at the same time slightly more robust and slightly more efficient than Spearman's rank correlation, making it the preferable estimator from both perspectives. The Quadrant correlation measure was also studied, and shown to be highly robust but at the price of a too low efficiency. The efficiency of the Quadrant correlation even converges to zero if the true correlation is close to one.

Although the nonparametric correlation measures discussed in this paper are well known, and frequently used in psychometrics, this paper is up to our knowledge the first one that gives a more formal treatment of their robustness and efficiency properties. The robustness of an estimator is summarized by its gross-error sensitivity, measuring the maximal effect that a single outlier can have on the estimator. We stress that both the gross-error sensitivity and the efficiencies of the different estimators are depending on the true value of the correlation coefficient, and this in a nonlinear way. We also make a comparison with robust correlation estimators derived from robust covariance matrices, the latter being well studied in the literature. This type of robust estimators is much harder to compute, and it turns out that both their gross-error sensitivity and their asymptotic variance are higher as for the simple Spearman and Kendall measures. We are, however, not claiming that one should discard robust correlation estimators derived from robust covariance matrices, like the MCD or S. From the simulations in Section 5 we could see that these estimators perform well in presence of larger amounts of contamination. Moreover, by decreasing the breakdown point of the considered estimator to 25%, for example, the statistical efficiency of the S-estimator increases from 38% to 84% and of

the MCD estimator from 3% to 16%. Of course, this increase of efficiency goes along with a decrease of robustness.

While this paper focuses on widely used measures of correlation as the Spearman and Kendall coefficient, other proposals for robust estimation of correlation have been made. For example a correlation coefficient based on mad and comedians (Falk, 1998), a correlation coefficient based on the decomposition of the covariance into a difference of variances (Genton & Ma, 1999), and a multiple skipped correlation (Wilcox, 2003) have been proposed. We did not pursued in this paper to cover all previous proposal of robust correlation measures. Another limitation of this paper is that robustness is measured by means of the influence function, which is suitable for measuring the robustness with respect to small amounts of outliers. For measuring robustness in presence of larger amounts of outliers, the breakdown point is more useful. Defining the breakdown point for correlation measures needs to be done with care, and we refer to the rejoinder of (Davies & Gather, 2005) were breakdown points are considered for the Spearman and Kendall correlation measures.

A Appendix

Proof of Proposition 2.

(i) From (3.1) it follows that

$$\begin{aligned} \mathrm{ASV}(R_p, H_\rho) &= E_{H_\rho}[(XY - \frac{\rho}{2}(X^2 + Y^2))^2] \\ &= (1-\rho^2)^2, \end{aligned}$$

since $E_{H_\rho}[X^4] = E_{H_\rho}[Y^4] = 3$, $E_{H_\rho}[X^2Y^2] = 1 + 2\rho^2$ and $E_{H_\rho}[X^3Y] = E_{H_\rho}[XY^3] = 3\rho$.
(ii) For the nonparametric Quadrant measure, using (3.2) and (3.5), we get

$$\begin{aligned} \mathrm{ASV}(\tilde{R}_Q, H_\rho) &= \frac{\pi^2}{4}(1-\rho^2)(1-\rho_Q^2) \\ &= (1-\rho^2)(\frac{\pi^2}{4} - \arcsin^2(\rho)), \end{aligned}$$

since $E[\text{sign}(XY)] = \rho_Q$ and $E[\text{sign}^2(XY)] = 1$.

(iii) From (3.3) and (3.6), we obtain

$$\text{ASV}(\tilde{R}_K, H_\rho) = \pi^2(1-\rho^2)E_{H_\rho}[\left(2P_{H_\rho}[(X-X_1)(Y-Y_1)>0] - 1 - \frac{2}{\pi}\arcsin(\rho)\right)^2]$$

which can be rewritten as

$$\text{ASV}(\tilde{R}_K, H_\rho) = cE[(K(X,Y) - E[K(X,Y)])^2] = c\{E[K^2(X,Y)] - \rho_K^2\}, \qquad \text{(A.1)}$$

where $K(x,y) = 2P_{H_\rho}[(X-x)(Y-y) > 0] - 1 = 1 - 2(\Phi(x) + \Phi(y)) + 4\Phi_\rho(x,y)$ and $c = \pi^2(1-\rho^2)$. Now

$$\begin{aligned} E[K^2(X,Y)] &= E[\text{sign}((X-X_1)(Y-Y_1)(X-X_2)(Y-Y_2))] \\ &= 2P((\frac{X-X_1}{\sqrt{2}})(\frac{Y-Y_1}{\sqrt{2}})(\frac{X-X_2}{\sqrt{2}})(\frac{Y-Y_2}{\sqrt{2}}) > 0) - 1, \end{aligned}$$

where (X_1, Y_1) and (X_2, Y_2) are independent copies of (X, Y). To simplify the above expression, denote $Z_1 = (X - X_1)/\sqrt{2}$, $Z_2 = (Y - Y_1)/\sqrt{2}$, $Z_3 = (X - X_2)/\sqrt{2}$ and $Z_4 = (Y - Y_2)/\sqrt{2}$, yielding

$$E[K^2(X,Y)] = 2P(Z_1Z_2Z_3Z_4 > 0) - 1. \qquad \text{(A.2)}$$

It is now easy to show that

$$\text{Cov}\begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{pmatrix} = \begin{pmatrix} 1 & \rho & \frac{1}{2} & \frac{\rho}{2} \\ \rho & 1 & \frac{\rho}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{\rho}{2} & 1 & \rho \\ \frac{\rho}{2} & \frac{1}{2} & \rho & 1 \end{pmatrix}.$$

By symmetry, we have

$$\begin{aligned} P(Z_1Z_2Z_3Z_4 > 0) &= 2[P(Z_1>0, Z_2>0, Z_3>0, Z_4>0) + P(Z_1>0, Z_2>0, Z_3<0, Z_4<0) \\ &+ P(Z_1>0, Z_3>0, Z_2<0, Z_4<0) + P(Z_1>0, Z_4>0, Z_2<0, Z_3<0)]. \end{aligned}$$

The first term in the above expression is of type (r), the second term of type (w), the third term of type (r) and the fourth term of type (w) where the (r) and (w) types are defined in Appendix 2 in David and Mallows (1961). We then obtain

$$P(Z_1Z_2Z_3Z_4 > 0) = 2[\frac{5}{18} + \frac{1}{\pi^2}(\arcsin^2(\rho) - \arcsin^2(\frac{\rho}{2}))]. \tag{A.3}$$

Combining (A.1), (A.2) and (A.3) yields (4.3).

(iv) For the transformed Spearman measure, one can rewrite (3.7) as

$$\mathrm{IF}((x,y), \tilde{R}_S, H_\rho) = 12c\{k(x,y) - E[k(X,Y)]\}$$

where $k(x,y) = F(x)G(y) + E_{H_\rho}[F(X)I(Y \geq y)] + E_{H_\rho}[G(Y)I(X \geq x)]$ and $c = \frac{\pi}{3}\sqrt{1 - \frac{\rho^2}{4}}$. It follows that

$$\mathrm{ASV}(\tilde{R}_S, H_\rho) = 144\frac{\pi^2}{9}(1 - \frac{\rho^2}{4})\{E[k^2(X,Y)] - 9(\frac{1}{4} + \frac{1}{2\pi}\arcsin(\frac{\rho}{2}))^2\}. \tag{A.4}$$

Now, we must compute the expression $E[k^2(X,Y)]$, with

$$k(x,y) = E[I(X_1 \leq x)I(Y_2 \leq y)] + E[I(X_2 \leq X_1)I(Y_1 \geq y)] + E[I(X_1 \geq x)I(Y_2 \leq Y_1)].$$

Tedious calculations result in

$$\begin{aligned}
E[k(X,Y)^2] &= E[I(X_1 \leq X)I(Y_2 \leq Y)I(X_3 \leq X)I(Y_4 \leq Y)] \\
&+ 2E[I(X_1 \leq X)I(Y_2 \leq Y)I(X_4 \leq X_3)I(Y_3 \geq Y)] \\
&+ 2E[I(X_1 \leq X)I(Y_2 \leq Y)I(X_3 \geq X)I(Y_4 \leq Y_3)] \\
&+ E[I(X_2 \leq X_1)I(Y_1 \geq Y)I(X_4 \leq X_3)I(Y_3 \geq Y)] \\
&+ 2E[I(X_2 \leq X_1)I(Y_1 \geq Y)I(X_3 \geq X)I(Y_4 \leq Y_3)] \\
&+ E[I(X_1 \geq X)I(Y_2 \leq Y_1)I(X_3 \geq X)I(Y_4 \leq Y_3)],
\end{aligned}$$

from which, using Appendix 2 of David and Mallows (1961), we obtain the following sum

of 6 terms

$$
\begin{aligned}
E[k(X,Y)^2] &= \frac{82}{144} + \frac{9}{4\pi}\arcsin(\frac{\rho}{2}) + \frac{1}{\pi^2}\int_0^{\arcsin(\frac{\rho}{2})} \arcsin(\frac{\sin(x)}{1+2\cos(2x)})dx \\
&+ \frac{2}{\pi^2}\int_0^{\arcsin(\frac{\rho}{2})} \arcsin(\frac{\sin(2x)}{\sqrt{1+2\cos(2x)}})dx + \frac{1}{\pi^2}\int_0^{\arcsin(\frac{\rho}{2})} \arcsin(\frac{\sin(2x)}{2\sqrt{\cos(2x)}})dx \\
&+ \frac{1}{2\pi^2}\int_0^{\arcsin(\frac{\rho}{2})} \arcsin(\frac{3\sin(x)-\sin(3x)}{4\cos(2x)})dx.
\end{aligned}
$$

Using the above expression and (A.4) results in (4.4).

References

Bilodeau, M. & Brenner, D.(1999). *Theory of Multivariate Statistics.* Springer, New York.

Blomqvist, N.(1950). On a measure of dependance between two random variables. *Annals of Mathematical Statistics*, 21, 593–600.

Borkowf, C.(2002). Computing the nonnull asymptotic variance and the asymptotic relative efficiency of Spearman's rank correlation. *Computational Statistics and Data Analysis*, 39, 271–286.

Croux, C. & Dehon, C.(2002). Analyse canonique basée sur des estimateurs robustes de la matrice de covariance. *Revue de Statistique Appliquée*, L (2), 5–26.

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

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

David, F.N., & Mallows, C.L.(1961). The variance of Spearman's rho in normal samples. *Biometrika*, 48, 19–28.

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

Davies, P.L. & Gather, U.(2005). Breakdown and Groups (with discussion). *The Annals of Statistics*, 33, 977–1035.

Devlin, S.J., Gnanadesikan, R., & Kettering, J.R.(1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62, 531–545.

Falk, M.(1998). A note on the Comedian for Elliptical Distributions. *Journal of Multivariate Analysis*, 67, 306–317.

Genton, M.G., & Ma Y.(1999). Robustness properties of dispersion estimators. *Statistics and Probability Letters*, 44, 343–350.

Grize, Y.L.(1978). *Robustheitseigenschaften von Korrelations-schätzungen,* Unpublished Diplomarbeit, ETH Zürich.

Hampel, F.R., Ronchetti, E.M., Rousseeuw P.J., & StahelW.A.(1986). *Robust statistics: the approach based on influence functions.* John Wiley and Sons, New York.

Kendall, M.G.(1938). A new measure of rank correlation. *Biometrika*, 30, 81–93.

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.

Maronna, R., Martin, D. & Yohai, V.(2006). *Robust Statistics.* Wiley, New York.

Moran, P.A.P.(1948). Rank Correlation and Permutation Distributions. *Proceedings of the Cambridge Philosophical Society*, 44, 142–144.

Mosteller, F.(1946). On some useful inefficient statistics. *Annals of Mathematical Statistics*, 17, 377.

Rousseeuw, P.J., & Van Driessen, K.(1999). A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, 41, 212–223.

Shevlyakov, G.L., & Vilchevski, N.O.(2002). *Robustness in Data Analysis: Criteria and Methods.* Modern Probability and Statistics, Utrecht.

Spearman, C.(1904). General intelligence objectively determined and measured. *American Journal of Psychology*, 15, 201–293.

Wilcox, R.R.(1998). The goals and strategies of robust methods. *British Journal of Mathematical & Statistical Psychology*, 51, 1–39.

Wilcox, R.R.(2003). Inferences based on multiple skipped correlations. *Computational Statistics and Data Analysis*, 44, 223–236.

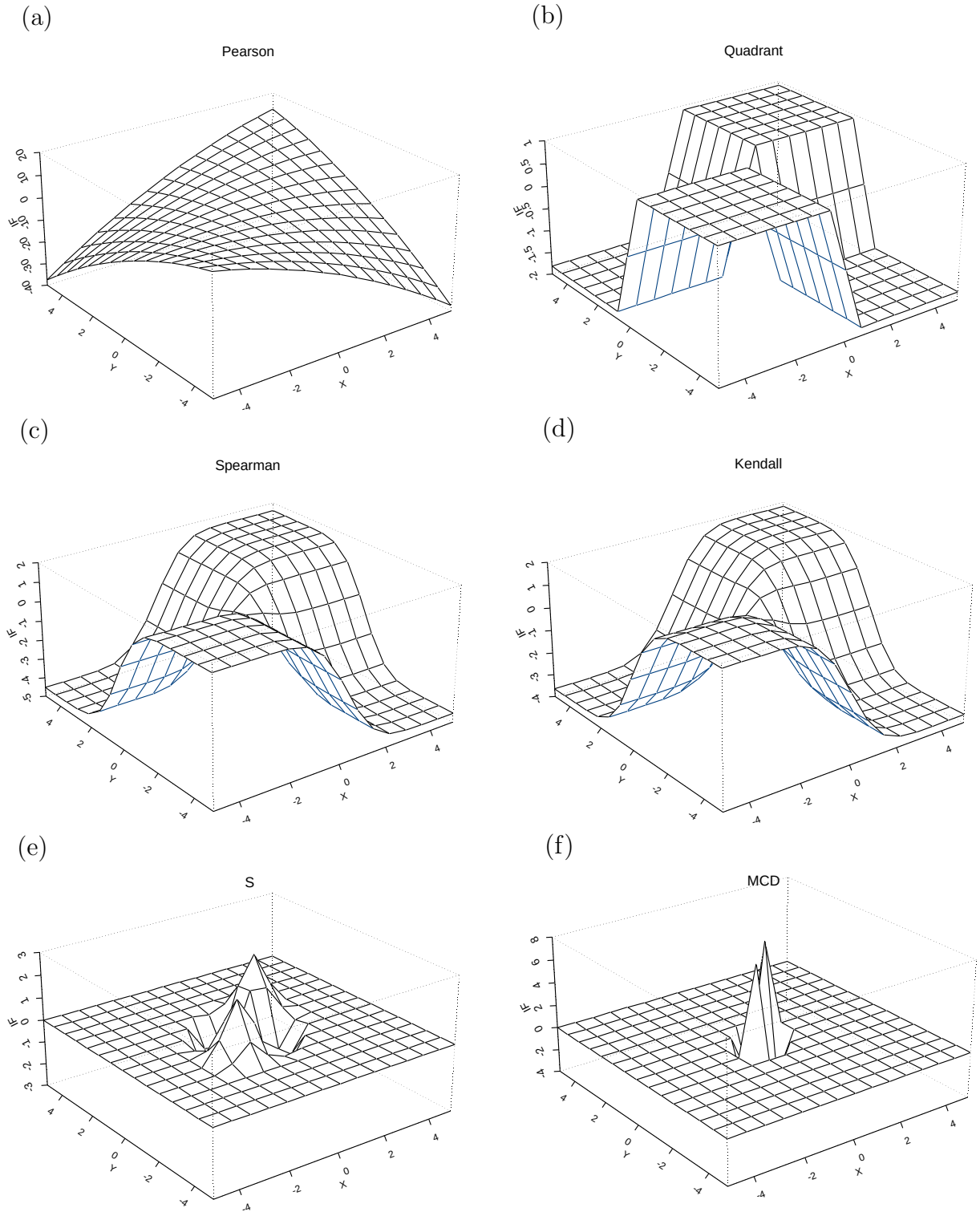


Figure 1: Influences functions for the consistent versions of the Pearson, Spearman, Kendall and Quadrant estimators at a bivariate normal distribution with correlation $\rho = 0.5$. The bottom row presents the IF for the correlation measures based on the MCD and S covariance matrix estimator.

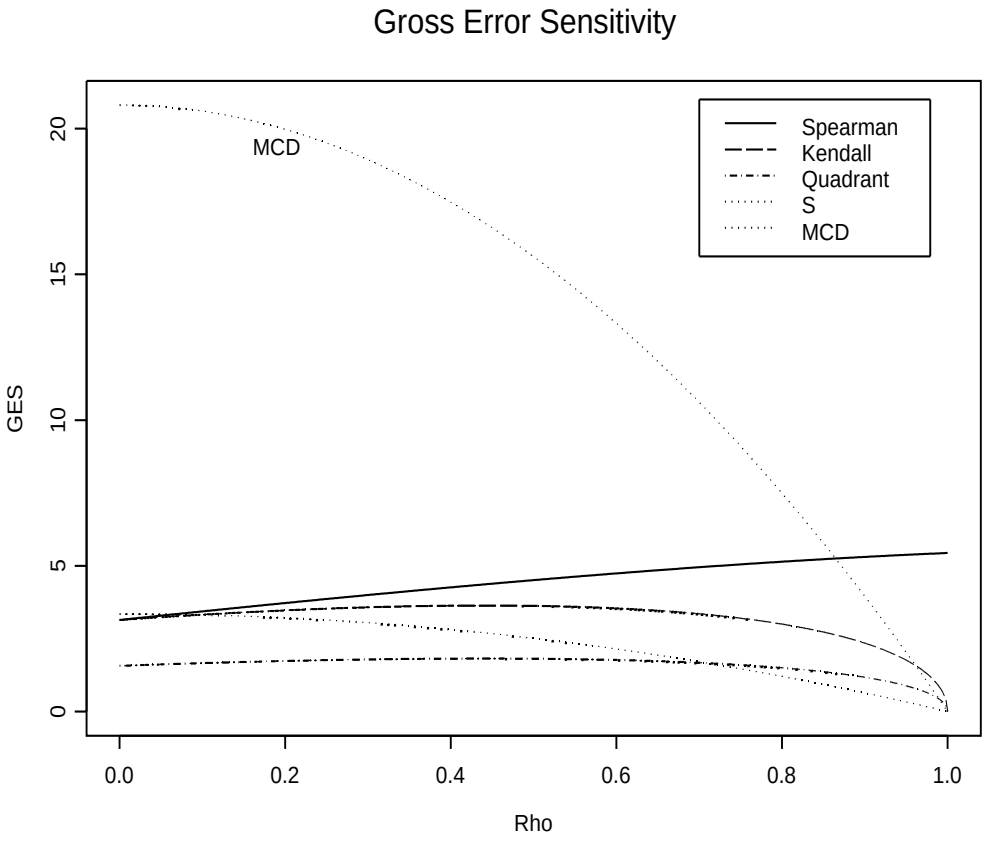


Figure 2: Gross-error sensitivities for the nonparametric correlation measures $\tilde{R}_Q, \tilde{R}_K, \tilde{R}_S$ and correlations based on the MCD and S covariance matrix as a function of ρ, the correlation of the bivariate normal model distribution.

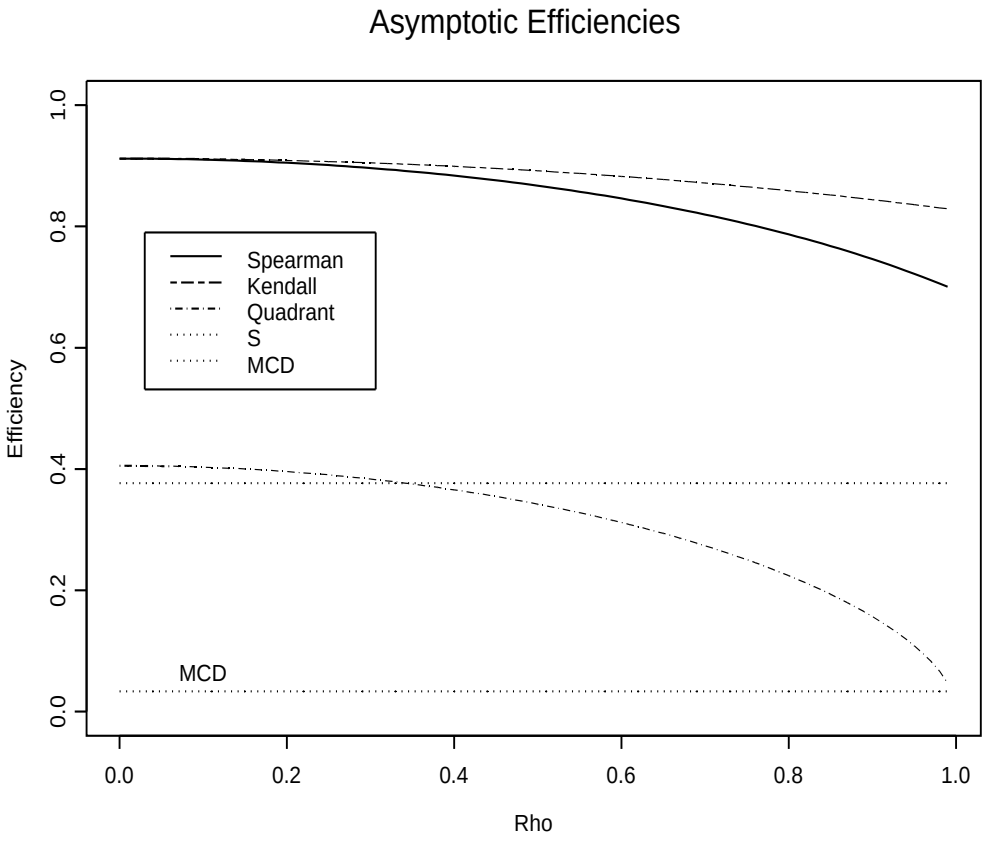


Figure 3: Asymptotic efficiencies for the nonparametric correlation measures $\tilde{R}_Q, \tilde{R}_K, \tilde{R}_S$ and correlations based on the MCD and S covariance matrix as a function of ρ, the correlation of the bivariate normal model distribution.

Table 1: MSE for several estimators of the population correlation $\rho = 0$ at a bivariate normal distribution, for sample sizes n=20, 50, 100 and 200.

$n * \text{MSE}$	n=20	n=50	n=100	n=200	n=∞
Pearson	1.05	1.02	1.00	1.00	1.00
Spearman	1.14	1.11	1.10	1.10	1.09
Kendall	1.22	1.15	1.11	1.11	1.09
Quadrant	2.30	2.40	2.43	2.47	2.46
S	3.39	3.06	2.82	2.80	2.65
MCD	8.09	12.96	18.04	21.53	30.01

Table 2: MSE for several estimators of the population correlation $\rho = 0$ at a bivariate normal distribution for sample size n=100 with a fraction ε of outliers.

MSE	$\varepsilon = 0\%$	$\varepsilon = 5\%$	$\varepsilon = 10\%$	$\varepsilon = 20\%$
Pearson	0.01	0.07	0.19	0.41
Spearman	0.01	0.02	0.07	0.24
Kendall	0.01	0.02	0.08	0.28
Quadrant	0.01	0.01	0.03	0.10
S	0.01	0.01	0.01	0.02
MCD	0.01	0.01	0.02	0.07