myjournal manuscript No.
(will be inserted by the editor)

Influence functions of the Spearman and Kendall correlation measures

Christophe Croux · **Catherine Dehon**

Received: date / Accepted: date

Abstract Nonparametric correlation estimators as the Kendall and Spearman correlation are widely used in the applied sciences. They are often said to be robust, in the sense of being resistant to outlying observations. In this paper we formally study their robustness by means of their influence functions and gross-error sensitivities. Since robustness of an estimator often comes at the price of an increased variance, we also compute statistical efficiencies at the normal model. We conclude that both the Spearman and Kendall correlation estimators combine a bounded and smooth influence function with a high efficiency. In a simulation experiment we compare these nonparametric estimators with correlations based on a robust covariance matrix estimator.

Keywords Asymptotic Variance · Correlation · Gross-Error Sensitivity · Influence function · Kendall correlation · Robustness · Spearman correlation

Mathematics Subject Classification (2000) 62G35 · 62F99

1 Introduction

The Pearson correlation is one of the most often used statistical estimators. But its value may be seriously affected by only one outlier. An important tool to measure

Christophe Croux
Faculty of Business and Economics, & K.U.Leuven, Naamsestraat 69, B-3000 Leuven, Belgium
Tilburg University, The Netherlands
Tel.: +32-(0)16-326958
Fax: +32-(0)16-326732
E-mail: Christophe.Croux@econ.kuleuven.be

Catherine Dehon
Université libre de Bruxelles, ECARES, and Institut de Recherche en Statistique, , CP-114, Av. F.D. Roosevelt 50, B-1050 Brussels, Belgium
Tel.: +32-(0)2-650-3858
Fax: +32-(0)2-650-4012
E-mail: cdehon@ulb.ac.be

robustness of a statistical measure is the influence function (Hampel et al., 1986). It measures the influence of contamination at a given value on the statistical measure; see Section 3 for a formal definition. Devlin et al. (1975) showed that the influence function of the classical Pearson correlation is unbounded, proving its lack of robustness. We refer to Morgenthaler (2007) for a survey on robust statistics.

In this paper we provide expressions for the influence functions of the popular Spearman and Kendall correlation. We show that their influence function is bounded. This confirms the general belief that these nonparametric measures of correlation are robust to outliers. Besides being robust, it is desirable that an estimator has a high statistical efficiency. At the normal distribution the Pearson correlation estimator is the most efficient, but the statistical efficiency of the Spearman and Kendall correlation estimators remains above 70% for all possible values of the population correlation. Hence they provide a good compromise between robustness and efficiency.

The paper is organized as follows. In Section 2 we review the definitions of the Spearman, Kendall and Quadrant correlation. Their influence functions are presented in Section 3 and gross-error-sensitivities are given in Section 4. The asymptotic variances are computed in Section 5. Section 6 presents a simulation study comparing the performance of the different estimators at finite samples. A comparison with a robust correlation measure derived from a bivariate robust covariance matrix estimator is made. The conclusions are in Section 7.

2 Measures of Correlation

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

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

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

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

and the function $H \longrightarrow R_P(H)$ is called the functional representation of this estimator. If the sample $(x_1, y_1), \ldots, (x_n, y_n)$ has been generated according to the distribution H, then the estimator r_P converges in probability to $R_P(H)$. For the bivariate normal distribution with population correlation coefficient ρ, denoted by Φ_ρ, we have

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

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 ranks and signs have been introduced. We first consider the Quadrant correlation, r_Q (Mosteller, 1946). It is computed by first centering the data by the coordinatewise median. Then r_Q equals the frequency of observations in the first or third quadrant, minus the frequency of observations in the second or fourth quadrant

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

Here, the sign function equals 1 for positive arguments, -1 for negative arguments, and $\text{sign}(0) = 0$. The associated functional is given by

$$R_Q(H) = E_H[\text{sign}\{(X - \text{median}(X))(Y - \text{median}(Y))\}].$$

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

$$\rho_Q := R_Q(\Phi_\rho) = \frac{2}{\pi}\arcsin(\rho), \tag{1}$$

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

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

The corresponding functional representation is then

$$R_K(H) = E_H[\text{sign}\{(X_1 - X_2)(Y_1 - Y_2)\}] \tag{2}$$

where (X_1, Y_1) and (X_2, Y_2) are two independent copies of H. At normal distributions, r_K estimates the same parameter as the Quadrant correlation (Blomqvist, 1950), so

$$\rho_K = \rho_Q = R_K(\Phi_\rho), \tag{3}$$

and the Fisher consistent version of Kendall's correlation is

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

Finally, the most popular nonparametric correlation measure is Spearman's rank correlation (Spearman, 1904), which equals the 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 representation of Spearman's correlation is given by

$$R_S(H) = \mathrm{Corr}(F(X), G(Y)) = 12E_H[F(X)G(Y)] - 3. \tag{4}$$

At the normal model Φ_ρ we have

$$\rho_S := R_S(\Phi_\rho) = \frac{6}{\pi}\arcsin(\frac{\rho}{2}), \tag{5}$$

see Moran (1948). For reasons of completeness, we briefly outline a proof of this old result in the Appendix, together with proofs of (1) and (3). Again we see that the Spearman correlation differs from the correlation coefficient ρ of the bivariate normal distribution, and the Fisher consistent version is given by

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

3 Influence Function

Assume that the bivariate random variable (X,Y) follows a distribution H. The influence function (IF) of a statistical functional R at a distribution H is defined as

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

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, for data coming from the distribution H. Note that the influence function is defined at the population level, and that the IF of an estimator refers to the IF of the associated functional representation of the estimator. In most cases, we will compute the influence function at the bivariate normal distribution Φ_ρ, having correlation coefficient ρ. We assume that the population means of the marginal distribution are equal to zero, and their variances equal to one. Since all correlation measures considered in this paper are invariant with respect to linear transformations of X, respectively Y, the latter assumption is without loss of generality.

An estimator is called B-robust if its influence function is bounded (see Hampel et al., 1986). For the Pearson correlation, Devlin et al. (1975) derived

$$\mathrm{IF}((x,y),R_P,\Phi_\rho) = xy - \left(\frac{x^2+y^2}{2}\right)\rho, \tag{7}$$

which is an unbounded function, showing that R_P is not B-robust. The influence functions associated to the Quadrant correlation is given by

$$\mathrm{IF}((x,y),R_Q,\Phi_\rho) = \mathrm{sign}(xy) - \rho_Q, \tag{8}$$

see Shevlyakov and Vilchevski (2002). The IFs of the Kendall and Spearman correlation do not seem to have been published in the printed literature, even if they are not difficult to obtain.

Proposition 1 *The influence function of the Kendall correlation is given by*

$$IF((x,y),R_K,H) = 2\{2P_H[(X-x)(Y-y)>0]-1-R_K(H)\}, \tag{9}$$

for any distribution H. At the bivariate normal model distribution Φ_ρ we have

$$IF((x,y),R_K,\Phi_\rho) = 2\{4\Phi_\rho(x,y)-2\Phi(x)-2\Phi(y)+1-\rho_K\}. \tag{10}$$

Proposition 2 *The influence function of the Spearman correlation is given by*

$$\begin{aligned} IF((x,y),R_S,H) &= -3R_S(H)-9+12\{F(x)G(y)+E_H[F(X)I(Y\geq y)] \\ &\quad +E_H[G(Y)I(X\geq x)]\}, \end{aligned} \tag{11}$$

for any distribution H, with F and G the marginal distributions of H, and where $I(\cdot)$ stands for the indicator function. At the bivariate normal model distribution Φ_ρ we have

$$\begin{aligned} IF((x,y),R_S,\Phi_\rho) &= -3\rho_S-9+12\{\Phi(x)\Phi(y)+E_\Phi[\Phi(X)\Phi(\frac{\rho X-y}{\sqrt{1-\rho^2}})]+ \\ &\quad +E_\Phi[\Phi(Y)\Phi(\frac{\rho Y-x}{\sqrt{1-\rho^2}})]\}, \end{aligned} \tag{12}$$

with Φ the cumulative distribution function of a standard normal, and for $|\rho|<1$.

In an unpublished manuscript of Grize (1978), similar expressions are obtained. Proofs of the propositions 1 and 2 can be found in the Appendix.

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 the IF of the transformed measures are given by

$$\mathrm{IF}((x,y),\tilde{R}_Q,\Phi_\rho) = \frac{\pi}{2}\mathrm{sign}(\rho)\sqrt{1-\rho^2}\,\mathrm{IF}((x,y),R_Q,\Phi_\rho) \tag{13}$$

$$\mathrm{IF}((x,y),\tilde{R}_K,\Phi_\rho) = \frac{\pi}{2}\mathrm{sign}(\rho)\sqrt{1-\rho^2}\,\mathrm{IF}((x,y),R_K,\Phi_\rho) \tag{14}$$

$$\mathrm{IF}((x,y),\tilde{R}_S,\Phi_\rho) = \frac{\pi}{3}\mathrm{sign}(\rho)\sqrt{1-\frac{\rho^2}{4}}\,\mathrm{IF}((x,y),R_S,\Phi_\rho). \tag{15}$$

One can see from Figure 1 that 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 lead to relatively large changes in the estimator. For Kendall and Spearman the influence functions are 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, i.e. $\mathrm{IF}((x,y),\tilde{R}_K,\Phi_0) = \mathrm{IF}((x,y),\tilde{R}_S,\Phi_0) = 4\pi(\Phi(x)-0.5)(\Phi(y)-0.5)$, but they differ slightly for other values of ρ.

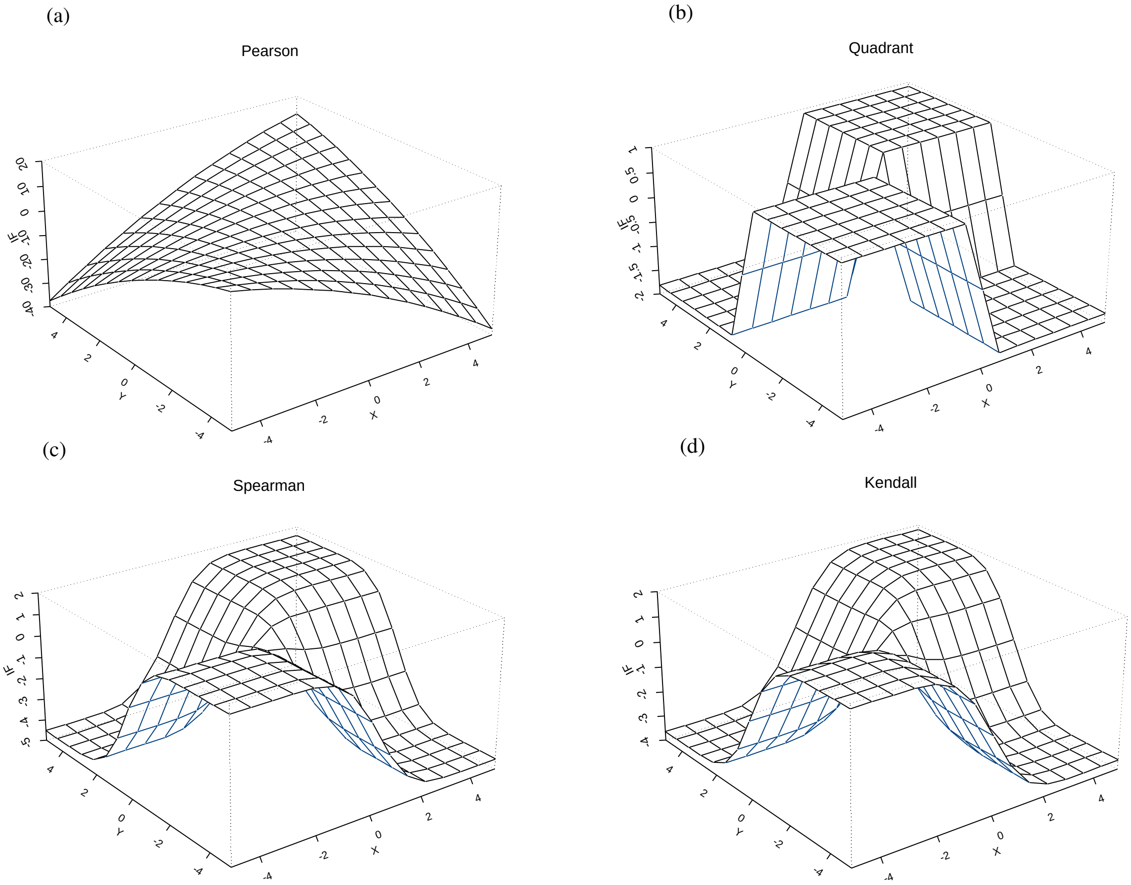

Fig. 1 Influence functions of the Pearson, Quadrant, Spearman and Kendall measures, evaluated at the bivariate normal distribution with $\rho = 0.5$.

4 Gross-error sensitivity

An influence function can be summarized in a single index, the *gross-error sensitivity* (GES), giving the maximal influence an observation may have. Formally, the GES of the functional R at the model distribution Φ_ρ is defined as

$$\mathrm{GES}(R, \Phi_\rho) = \sup_{(x,y)} |\mathrm{IF}((x,y), R, \Phi_\rho)|, \tag{16}$$

see Hampel et al. (1986). For example, since the classical Pearson estimator is not B-robust, $\mathrm{GES}(R_P, \Phi_\rho) = \infty$. The following proposition gives the GES associated with the nonparametric measures of correlation we consider.

Proposition 3 *The gross-error sensitivities (GES) of the three transformed nonparametric correlation measures are given by*

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

where Φ_ρ is a bivariate normal distribution with correlation ρ, for $-1 \leq \rho \leq 1$.

The proof of proposition 3 is elementary. For a positive value of ρ, the sup in definition (16) is attained for x tending to infinity, and y to minus infinity (or, equivalently, x tending to $-\infty$ and y to $+\infty$). For a negative value of ρ, the largest influence corresponds to contamination at (∞,∞) or $(-\infty,-\infty)$. The gross-error sensitivities depend on the parameter ρ in a non-linear way, and are plotted in Figure 2. The Quadrant estimator has uniformly a lower GES than Kendall and Spearman, and is exactly half of the GES of Kendall. On the other hand, Kendall's measure is preferable to Spearman, although the difference in GES is negligible for smaller values of ρ.

A striking feature of Figure 2 is that the GES converges to zero, if ρ tends to one, for the transformed Quadrant and Kendall correlation, but not for Spearman. The reason is that the transformation function $g(r) = \sin(\pi r/2)$ has derivative zero at $r = 1$, which is not true for the transformation needed for the consistency of the Spearman correlation, $g(r) = 2\sin(\pi r/6)$.

5 Asymptotic Variance

Let r be the correlation estimator associated with the functional R. All considered correlation estimators are asymptotically normal at the model distribution

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

where the asymptotic variances are given by $\text{ASV}(R,H) = E_H[\text{IF}((X,Y),R,H)^2]$. This result holds for the Quadrant, Spearmann, and Kendall correlation since they can be expressed as regular U-statistics, see Moran (1948) and Blomqvist (1950). The next proposition lists the expressions for $\text{ASV}(R, \Phi_\rho)$. The proof is in the Appendix.

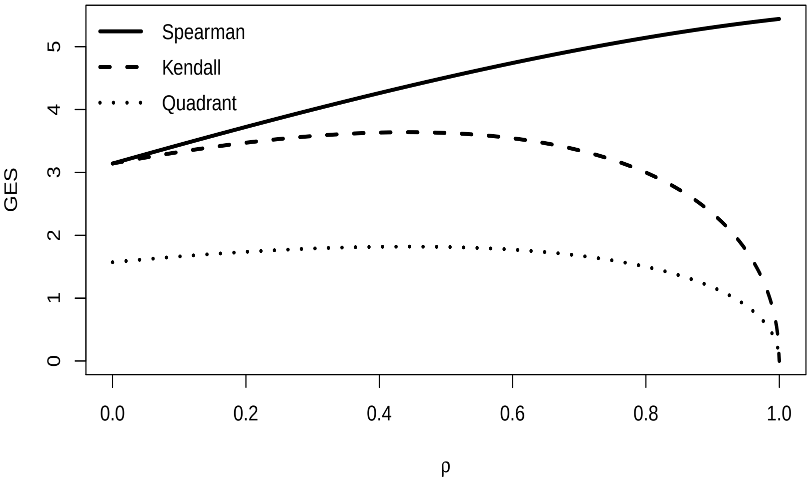

Fig. 2 Gross-error sensitivities of the Quadrant, Spearman, and Kendall correlation, as a function of the correlation ρ of the bivariate normal model distribution.

Proposition 4 *At the model distribution Φ_ρ, and for any $-1 \leq \rho \leq 1$, we have:*

$$
\begin{aligned}
&(i) \qquad && ASV(R_P, \Phi_\rho) = (1-\rho^2)^2 \\
&(ii) \qquad && ASV(\tilde{R}_Q, \Phi_\rho) = (1-\rho^2)(\frac{\pi^2}{4} - \arcsin^2(\rho)) \\
&(iii) \qquad && ASV(\tilde{R}_K, \Phi_\rho) = \pi^2(1-\rho^2)(\frac{1}{9} - \frac{4}{\pi^2}\arcsin^2(\frac{\rho}{2})) \qquad (17)
\end{aligned}
$$

$$
\begin{aligned}
(iv) \qquad ASV(\tilde{R}_S, \Phi_\rho) = {} & \frac{\pi^2}{9}(1-\frac{\rho^2}{4})144\{\frac{1}{144} - \frac{9}{4\pi^2}\arcsin^2(\frac{\rho}{2}) \\
& + \frac{1}{\pi^2}\int_0^{\arcsin(\frac{\rho}{2})} \arcsin(\frac{\sin(x)}{1+2\cos(2x)})dx \\
& + \frac{2}{\pi^2}\int_0^{\arcsin(\frac{\rho}{2})} \arcsin(\frac{\sin(2x)}{\sqrt{1+2\cos(2x)}})dx \\
& + \frac{1}{\pi^2}\int_0^{\arcsin(\frac{\rho}{2})} \arcsin(\frac{\sin(2x)}{2\sqrt{\cos(2x)}})dx \\
& + \frac{1}{2\pi^2}\int_0^{\arcsin(\frac{\rho}{2})} \arcsin(\frac{3\sin(x)-\sin(3x)}{4\cos(2x)})dx\} \qquad (18)
\end{aligned}
$$

The results in the previous proposition are not new, since expressions for the asymptotic variances can be derived from Blomqvist (1950) for the Quadrant and Kendall correlation, and in David and Mallows (1961) for the Spearman correlation

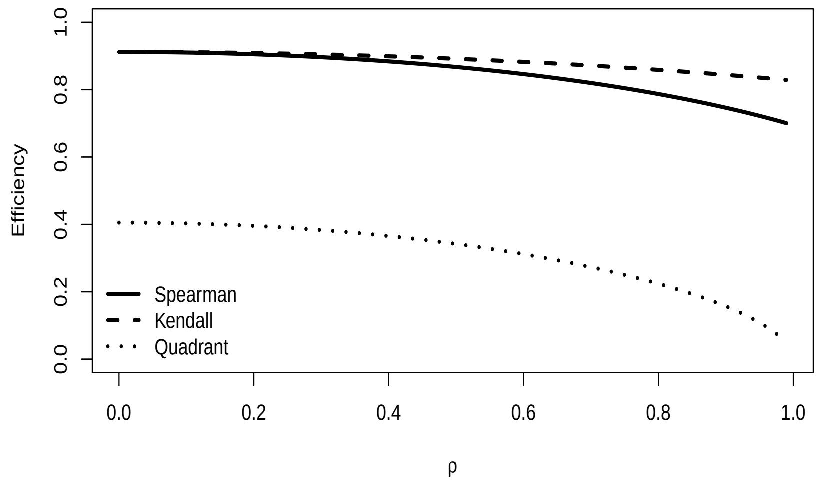

Fig. 3 Asymptotic efficiencies of the Quadrant, Spearman, and Kendall correlation measures, as a function of the correlation ρ of the bivariate normal model distribution.

at normal samples. In these older papers asymptotic expansions of $Var(r)$ as a function of the sample size are given. From these, the same asymptotic variances listed above result. It is surprising, however, that in more recent literature not much attention is given to the asymptotic variances of nonparametric correlation estimators. In Bonett and Wright (2000), for example, confidence intervals for the Spearman and Kendall correlation are constructed using approximations of the asymptotic variances, while Proposition 4 provides the closed form expressions. Most complicated is the expression for $\mathrm{ASV}(\tilde{R}_S, \Phi_\rho)$, requiring numerical integration of univariate integrals. A similar result, but expressed in more general terms, is given in Borkowf (2002).

In Figure 3 we plot asymptotic efficiencies (relative to the Pearson correlation) as a function of ρ, with $0 \leq \rho < 1$. All asymptotic variances are decreasing in ρ, and converge to zero for ρ tending to one. The case $\rho = 1$ is degenerate; the data are then lying on a straight line, and estimators always return one, without any estimation error. 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 quite efficient. Comparing Kendall's with Spearman's correlation is favorable for Kendall, but the difference in efficiency is rather small. The Quadrant correlation has a much lower efficiency. Its efficiency even converges to zero if the true correlation is close to one. Hence, the variance of the Quadrant estimator is decreasing much slower to zero as that of the Pearson correlation.

6 Simulation Study

By means of a simulation experiment, we try to answer two questions. First we verify whether the finite-sample variances of the estimators are close to their asymptotic counterparts, derived in Section 5. Secondly, we study how the estimators behave when outliers are introduced in the sample. We make a comparison with a robust correlation estimator derived from a robust covariance matrix. If $C(X,Y)$ is a 2×2 robust covariance matrix, then the associated robust correlation measure equals

$$R_C(H) = \frac{C_{12}(X,Y)}{\sqrt{C_{11}(X,Y)C_{22}(X,Y)}}. \tag{19}$$

Hence, any robust bivariate covariance matrix C leads to a robust correlation coefficient. We take for C in (19) the Minimum Covariance Determinant (MCD, Rousseeuw and Van Driessen, 1999) with 50% breakdown point, and additional reweighting step[1]. Since the MCD estimator estimates a multiple of the population covariance matrix at the normal distribution, we have $R_C(\Phi_\rho) = \rho$. The asymptotic variance of the MCD estimator is given in Croux and Haesbroeck (1999).

Simulation design without outliers. We generate $m = 10000$ samples of size $n = 20, 50, 100, 200$ from a bivariate normal with $\rho = 0.8$ (simulations for other values of ρ result in similar conclusions). For each sample j, the correlation coefficient is estimated by $\hat{\rho}_j$, and the mean squared error (MSE) is computed as

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

Table 1 reports the MSE for the different estimators we considered. Each MSE is multiplied by the corresponding sample size n, and these quantities should converge to the asymptotic variances given in proposition 4. As we can see from Table 1, the finite sample MSE converges rather quickly to the asymptotic counterpart (reported under the column $n = \infty$). The simulation experiment confirms the findings of Section 5; the precision of the Spearman and Kendall estimators is quite close to that of the Pearson correlation, and Kendall has slightly smaller MSEs than Spearman. On the other hand, the MSE of the Quadrant correlation is much larger. Finally, note that the correlation measure derived from the robust MCD covariance matrix is more efficient than the Quadrant correlation, but much less efficient than the Kendall and Spearmann measures.

To gain insight in the distribution of the (transformed) Quadrant, Spearman, and Kendall correlation, we present the boxplot of the $m = 10000$ simulated estimates, for $n = 200$. From Figure 4 we see that all correlation estimators are nearly unbiased. The distributions are almost symmetric, where the lower tail is slightly more pronounced than the upper tail.

[1] We use the R-command *covMcd* from the “robustbase” package, with default options, for computing the MCD.

Table 1 Simulated MSE (multiplied by the sample size) of several correlation estimators at a bivariate normal distribution with $\rho = 0.8$, for sample sizes n =20, 50, 100 and 200. The column $n = \infty$ refers to the asymptotic variance.

n * MSE	n =20	n =50	n =100	n =200	$n = \infty$
Pearson	0.17	0.14	0.14	0.13	0.13
Spearman	0.26	0.21	0.18	0.18	0.16
Kendall	0.25	0.19	0.17	0.16	0.15
Quadrant	0.67	0.66	0.60	0.60	0.58
MCD	0.85	0.53	0.42	0.37	0.32

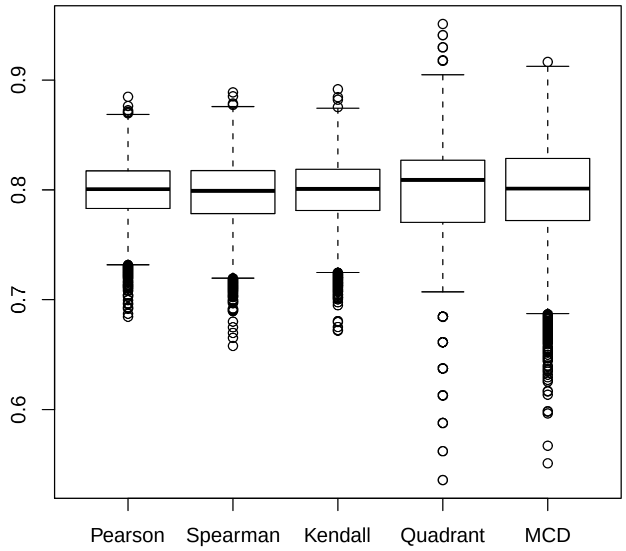

Fig. 4 Boxplots of 10000 correlation estimates for samples of size $n = 200$ from a bivariate normal model distribution with $\rho = 0.8$, for several correlation measures

Simulation designs with outliers. In the second simulation scheme we generate $m = 10000$ samples of size $n = 50, 100$ and 200 from the distribution Φ_ρ, with $\rho = 0.8$. A certain percentage ε of the observations is then replaced by outliers. We consider (i) outliers at position $(5, -5)$, where the influence function is close to its most extreme value, see Figure 1; (ii) correlation outliers, i.e. outliers that are not visible in the marginal distributions, generated from the distribution $\Phi_{-\rho}$, which has the opposite correlation structure as the model distribution. The MSEs are reported in Tables 2 and 3.

Table 2 Simulated MSE (multiplied by the sample size) of several correlation estimators at a bivariate normal distribution with $\rho = 0.8$, for sample sizes $n = 50, 100, 200$ and a fraction ε of outliers at position $(5, -5)$.

	$n*$MSE	$\varepsilon = 0\%$	$\varepsilon = 1\%$	$\varepsilon = 5\%$	$\varepsilon = 10\%$
$n = 50$	Pearson	0.14	19.73	60.48	84.47
	Spearman	0.19	0.82	4.94	12.28
	Kendall	0.18	0.44	2.39	6.50
	Quadrant	0.65	0.85	1.67	3.56
	MCD	0.51	0.51	0.44	0.38
$n = 100$	Pearson	0.14	13.89	102.04	167.48
	Spearman	0.18	0.49	6.67	24.03
	Kendall	0.17	0.28	2.99	12.46
	Quadrant	0.62	0.69	1.74	5.77
	MCD	0.42	0.41	0.35	0.32
$n = 200$	Pearson	0.13	26.87	201.96	331.82
	Spearman	0.17	0.75	12.97	47.35
	Kendall	0.16	0.37	5.69	24.22
	Quadrant	0.61	0.70	2.49	10.42
	MCD	0.37	0.36	0.31	0.28

Although the MSE is the smallest for the Pearson correlation if no outliers are present, this does not hold anymore in presence of outliers. As we see from Table 2, and already for 1% of outliers, the MSE for Pearson is by far the largest of all considered estimators. This confirms the non robustness of the Pearson correlation. For 1% of contamination, the MSE of the Spearman and Kendall correlation remains within bounds, with Kendall being more resistant to outliers. But for larger amounts of contamination, a substantial increase in MSE is observed for these two estimators. For $\varepsilon = 5\%$, the Quadrant estimator performs better than the two other nonparametric correlation measures. Finally note the high robustness of the MCD based estimator, where the MSE remains low for even 10% of contamination. We conclude that the correlation estimator associated to a highly robust covariance matrix estimator is the most resistant in presence of clusters of large outliers,

Correlation outliers do not show up in the marginal distributions, but may still have an important effect on the sample correlation coefficient, see Table 3. The Kendall correlation has consistently a smaller MSE than the Spearmann measure, and for $\varepsilon \geq 5\%$ it also beats the Pearson correlation. It is interesting to notice that the Quadrant correlation yields the highest MSE of the three nonparametric correlation estimators we considered (for $\varepsilon \leq 0.10$), showing that is copes more easily with extreme outliers than with correlation outliers. For larger levels of contamination the MCD is the better estimator, although it looses performance at small sample sizes.

7 Conclusion

In this paper we compute the influence functions of some widely used nonparametric measures of correlation. The Spearman and Kendall correlation have a bounded and

Table 3 Simulated MSE (multiplied by the sample size) of several correlation estimators at a bivariate normal distribution with $\rho = 0.8$, for sample size $n = 50, 100, 200$ and a fraction ε of correlation outliers.

	$n*$MSE	$\varepsilon = 0\%$	$\varepsilon = 1\%$	$\varepsilon = 5\%$	$\varepsilon = 10\%$
	Pearson	0.14	0.20	0.39	0.69
	Spearman	0.20	0.25	0.43	0.71
$n = 50$	Kendall	0.18	0.21	0.33	0.54
	Quadrant	0.67	0.73	0.88	1.14
	MCD	0.53	0.57	0.69	0.85
	Pearson	0.14	0.16	0.40	0.99
	Spearman	0.18	0.20	0.42	0.97
$n = 100$	Kendall	0.17	0.18	0.32	0.72
	Quadrant	0.63	0.65	0.81	1.25
	MCD	0.42	0.45	0.52	0.74
	Pearson	0.13	0.17	0.56	1.61
	Spearman	0.17	0.21	0.57	1.54
$n = 200$	Kendall	0.16	0.18	0.42	1.14
	Quadrant	0.58	0.63	0.91	1.57
	MCD	0.37	0.40	0.51	0.76

smooth influence function, and reasonably small values for the gross-error sensitivity. The gross-error sensitivity, as well as the efficiencies, are depending on the true value of the correlation in a nonlinear way. The Kendall correlation measure is 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 very robust but at the price of a low Gaussian efficiency.

Although the nonparametric correlation measures discussed in this paper are well known, and frequently used in applications, there are few papers presenting a formal treatment of their robustness properties. This paper focusses on studying the influence that observations have on the estimators, as is common in the robustness literature (e.g. Atkinson et al 2004, Olkin and Raveh 2009). We did not consider breakdown properties. The rejoinder of Davies and Gather (2005) discusses the difficulties of finding an appropriate definition of breakdown point for correlation measures. Breakdown properties of the test statistics for independence using Spearman and Kendall correlation are studied in Caperaa and Garralda Guillem (1997).

The correlation measures studied in this paper measure association between two random variables. However, robust correlation measures can be used to construct multivariate covariance matrices, based on pairwise covariances (see Gnadesikan and Kettering, 1972, and Maronna and Zamar, 2002). For instance, Alqallaf et al (2002) use the Quadrant correlation to get a robust scatter matrix in very high dimensions. The resulting multivariate is highly robust and very fast to compute. Khan, J.A., Van Aelst, and Zamar (2007) use a pairwise correlation matrix as input for a robust least angle regression estimator. One might conjecture that the robustness and efficiency properties of the correlation measures derived in this paper will be inherited by the pairwise covariance matrices constructed from them, though this should be confirmed by future research.

While this paper focuses on the Spearman and Kendall coefficient, other proposals for robust estimation of correlation have been made. For example a correlation coefficient based on MAD and co-medians (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. In the simulation study we make a comparison with the robust correlation estimator associated to the Minimum Covariance Determinant, a standard robust covariance matrix estimator (see also Cerioli 2010). For small amounts of outliers, the Kendall correlation can compete in terms of robustness with the MCD, while being much simpler to compute. But in presence of multiple outliers, the MCD is preferable.

Acknowledgment. We would like to thank the two reviewers for their careful reading of our manuscript and their useful comments.

Appendix

Proof of equations (1)**,** (3)**, and** (5)**.** We use the notation $R(X,Y) = R(H)$ if (X,Y) is distributed as H. Let $(X,Y) \sim \Phi_\rho$, and assume without loss of generality that $\rho \geq 0$. We can write $Y = \rho X + \varepsilon\sqrt{1-\rho^2}$, with (X,ε) a bivariate standard normal distribution. Then $(X,\varepsilon) = (R\cos\theta, R\sin\theta)$, with θ uniformly distributed on $[0,2\pi]$ and with R^2 following a chi-squared distribution with two degrees of freedom. Furthermore, there exists an $\alpha = \arcsin(\rho)$ in $[0,\pi/2]$ such that $\sin(\alpha) = \rho$. Then $Y = R\sin(\alpha+\theta)$.

For the Quadrant correlation, we have $R_Q(H_\rho) = 2(P(X>0,Y>0) - P(X>0,Y>0))$. Now

$$\begin{aligned} P(X>0,Y>0) &= P(\cos\theta > 0, \sin(\alpha+\theta) > 0) \\ &= P(\theta \in [-\alpha, \pi/2]) \\ &= (\pi/2+\alpha)/(2\pi) = 1/4 + \arcsin(\rho)/(2\pi) \qquad (20) \end{aligned}$$

since θ is uniform on $[0,2\pi]$. Similarly $P(X>0,Y<0) = (\pi/2-\alpha)/(2\pi)$. We conclude that (1) holds. Notice that this result does not depend on the distribution of R.

For the Kendall correlation, equation (2) shows that $R_K(X,Y) = R_Q(\tilde{X},\tilde{Y})$, with $(\tilde{X},\tilde{Y}) \overset{d}{=} (X_1 - X_2, Y_1 - Y_2)$ following again a bivariate normal distribution. The variances of $\tilde{X}$ and $\tilde{Y}$ are equal to 2, but the correlation between them remains ρ. Hence (3) follows.

Finally, for the Spearmann correlation, we need to compute

$$E_{\Phi_\rho}[\Phi(X)\Phi(Y)] = E[I(U \leq X)I(V \leq Y)] = P(X-U \geq 0, Y-V \geq 0)$$

with (U,V) bivariate standard normal, independent of $(X,Y) \sim \Phi_\rho$. We can readily check that $(X-U, Y-V)$ follows again a bivariate normal distribution, but with correlation $\rho/2$. Then it follows from (20) that $E_{\Phi_\rho}[\Phi(X)\Phi(Y)] = 1/4 + \arcsin(\rho/2)/(2\pi)$. Combined with (4), we obtain (5). □

Proof of Proposition 1. Let $H_\varepsilon = (1-\varepsilon)H + \varepsilon\Delta_{(x,y)}$ be the contaminated distribution. It follows from (2) that

$$\begin{aligned} R_K(H_\varepsilon) &= (1-\varepsilon)^2 E_H[\text{sign}(X_1-X_2)(Y_1-Y_2)] + 2\varepsilon(1-\varepsilon)E_H[\text{sign}(X-x)(Y-y)] \\ &+ \varepsilon^2 \text{sign}(x-x)(y-y) \end{aligned}$$

from which it follows that

$$\begin{aligned} \text{IF}((x,y),R_K,H) &= -2\rho_K + 2E_H[\text{sign}(X-x)(Y-y)] \\ &= -2R_K(H) + 2P_H[(X-x)(Y-y) > 0] - 2P_H[(X-x)(Y-y) < 0], \end{aligned}$$

confirming (9). At continuous distributions H the above expression simplifies further into

$$\text{IF}((x,y),R_K,H) = 2\{-\rho_K + 2P_H[(X-x)(Y-y) > 0] - 1\}.$$

Using

$$P_{\Phi_\rho}[(X-x)(Y-y) > 0] = 2\Phi_\rho(x,y) - \Phi(x) - \Phi(y) + 1$$

yields then the expression (10). □

Proof of Proposition 2. Let $H_\varepsilon = (1-\varepsilon)H + \varepsilon\Delta_{(x,y)}$ be the contaminated model distribution. Then H has marginal distributions $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_x$ and $G_\varepsilon = (1-\varepsilon)G + \varepsilon\Delta_y$. It follows from (4) that

$$R_S(H_\varepsilon) = 12(1-\varepsilon)E_H[F_\varepsilon(X)G_\varepsilon(Y)] + 12\varepsilon F_\varepsilon(x)G_\varepsilon(y) - 3.$$

from which it follows that

$$\text{IF}((x,y),R_S,H) = 12A - 12E_H[F(X)G(Y)] + 12F(x)G(y), \tag{21}$$

with A the derivative w.r.t. ε and evaluated at $\varepsilon = 0$ of

$$\begin{aligned} E_H[F_\varepsilon(X)G_\varepsilon(Y)] &= (1-\varepsilon)^2 E_H[F(X)G(Y)] + \varepsilon(1-\varepsilon)E_H[F(X)I(Y \geq y)] \\ &+ \varepsilon(1-\varepsilon)E_H[G(Y)I(X \geq x)] + \varepsilon^2 E_H[I(Y \geq y)I(X \geq x)]. \end{aligned}$$

But then

$$A = -2E_H[F(X)G(Y)] + E_H[F(X)I(Y \geq y)] + E_H[G(Y)I(X \geq x)].$$

Using the above formula, (21) becomes

$$\begin{aligned} \text{IF}((x,y),R_S,H) &= 12\{E_H[F(X)I(Y \geq y)] + E_H[G(Y)I(X \geq x)] \\ &\quad -3E_H[F(X)G(Y)] + F(x)G(y)\}, \end{aligned}$$

from which, using that $R_S(H) = 12E_H[F(X)G(Y)] - 3$, result (11) follows.

For the bivariate normal distribution, the marginals are given by $F = G = \Phi$. Furthermore, one can write $Y = \rho X + \sqrt{1-\rho^2}Z$, with Z independent of X and standard normal. Then

$$E_{\Phi_\rho}[\Phi(X)I(Y \geq y)] = E_\Phi[\Phi(X)\Phi(\frac{\rho X - y}{\sqrt{1-\rho^2}})].$$

□

Proof of Proposition 4. (i) From (7) it follows that

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

since $E_{\Phi_\rho}[X^4] = E_{\Phi_\rho}[Y^4] = 3$, $E_{\Phi_\rho}[X^2Y^2] = 1+2\rho^2$ and $E_{\Phi_\rho}[X^3Y] = E_{\Phi_\rho}[XY^3] = 3\rho$.
(ii) For the nonparametric Quadrant measure, using (8) and (13), we get

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

since $E[\mathrm{sign}(XY)] = \rho_Q$ and $E[\mathrm{sign}^2(XY)] = 1$.
(iii) From (9) and (14), we obtain

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

which can be rewritten as

$$\mathrm{ASV}(\tilde{R}_K, \Phi_\rho) = cE[(K(X,Y) - E[K(X,Y)])^2] = c\{E[K^2(X,Y)] - \rho_K^2\}, \quad (22)$$

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

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

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

$$E[K^2(X,Y)] = 2P(Z_1Z_2Z_3Z_4 > 0) - 1. \quad (23)$$

It is now easy to show that

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

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

$$P(Z_1Z_2Z_3Z_4 > 0) = 2[\frac{5}{18} + \frac{1}{\pi^2}(\arcsin^2(\rho) - \arcsin^2(\frac{\rho}{2}))]. \tag{24}$$

Combining (22), (23) and (24) yields (17).
(iv) For the transformed Spearman measure, one can rewrite (15) as

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

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

$$\mathrm{ASV}(\tilde{R}_S, \Phi_\rho) = 144\frac{\pi^2}{9}(1 - \frac{\rho^2}{4})\{E[k^2(X,Y)] - 9(\frac{1}{4} + \frac{1}{2\pi}\arcsin(\frac{\rho}{2}))^2\}. \tag{25}$$

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 (25) results in (18). □

References

1. Alqallaf, F. A., Konis, K. P., Martin, R. D. & Zamar, R. H. Scalable robust covariance and correlation estimates for data mining. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton (2002)
2. Atkinson, A.C., Riani M., & Cerioli A. *Exploring multivariate data with the forward search.*, Springer, New York (2004)
3. Blomqvist, N. On a measure of dependance between two random variables. *Annals of Mathematical Statistics*, 21, 593–600 (1950)
4. Bonett, D.G., & Wright, T.A. Sample size requirements for estimating Pearson, Kendall and Spearman correlation *Psychometrika*, 65, 23-28 (2000)
5. Borkowf, C. Computing the nonnull asymptotic variance and the asymptotic relative efficiency of Spearman's rank correlation. *Computational Statistics and Data Analysis*, 39, 271–286. (2002)
6. Caperaa, P., and Garralda Guillem, A.I. Taux de resistance des tests de rang d'independance. *The Canadian Journal of Statistics*, 25, 113-124 (1997)
7. Cerioli, A., Multivariate Outlier Detection with High-Breakdown Estimators. *Journal of The American Statistical Association*, to appear (2010)
8. Croux, C. & Haesbroeck G. Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator. *The Journal of Multivariate Analysis*, 71, 161–190 (1999)
9. David, F.N., & Mallows, C.L. The variance of Spearman's rho in normal samples. *Biometrika*, 48, 19–28 (1961)
10. Davies, P.L. & Gather, U. Breakdown and Groups (with discussion). *The Annals of Statistics*, 33, 977–1035 (2005)
11. Devlin, S.J., Gnanadesikan, R., & Kettering, J.R. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62, 531–545 (1975)
12. Falk, M. A note on the Comedian for Elliptical Distributions. *Journal of Multivariate Analysis*, 67, 306–317 (1998)
13. Genton, M.G., & Ma Y. Robustness properties of dispersion estimators. *Statistics and Probability Letters*, 44, 343–350 (1999)
14. Gnanadesikan, R., & Kettering, J. R. Robust estimates, Residuals, and Outlier Detection wit Multiresponse Data. *Biometrics*, 28, 81–124 (1972)
15. Grize, Y.L. *Robustheitseigenschaften von Korrelations-schätzungen,* Unpublished Diplomarbeit, ETH Zürich (1978)
16. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., & Stahel, W.A. *Robust statistics: the approach based on influence functions.* John Wiley and Sons, New York (1986)
17. Kendall, M.G. A new measure of rank correlation. *Biometrika*, 30, 81–93 (1938)
18. Khan, J.A., Van Aelst, S & Zamar, R.H. Robust linear model selection based on least angle regression *Journal of the American Statistical Association*, 480, 1289–1299 (2007).
19. Maronna, R., Martin, D. & Yohai, V. *Robust Statistics.* Wiley, New York (2006)
20. Maronna, R.A., & Zamar, R.H. Robust estimates of location and dispersion of high-dimensional datasets. *Technometrics*, 44, 307-317 (2002)
21. Moran, P.A.P. Rank Correlation and Permutation Distributions. *Biometrika*, 44, 142–144 (1948)
22. Morgenthaler, S. A survey of robust statistics. *Statistical Methods and Applications*, 15, 271–293 (2007)
23. Mosteller, F. On some useful inefficient statistics. *Annals of Mathematical Statistics*, 17, 377 (1946).
24. Olkin, I., & Raveh, A. Bounds for how much influence an observation can have. *Statistical Methods and Applications*, 18, 1–11 (2009)
25. Rousseeuw, P.J., & Van Driessen, K. A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, 41, 212–223 (1999)
26. Shevlyakov, G.L., & Vilchevski, N.O. *Robustness in Data Analysis: Criteria and Methods.* Modern Probability and Statistics, Utrecht (2002)
27. Spearman, C. General intelligence objectively determined and measured. *American Journal of Psychology*, 15, 201–293 (1904)
28. Wilcox, R.R. Inferences based on multiple skipped correlations. *Computational Statistics and Data Analysis*, 44, 223–236 (2003)