On the robustness of non-parametric correlation measures

C. Dehon[1], F. Alqallaf[2], C. Croux[3], and R. Zamar[4]

[1] ECARES, Faculté SOCO, and Institut de Statistique, Université Libre de Bruxelles, CP-114, Av. F.D. Roosevelt 50, B-1050 Brussels, Belgium, Email:cdehon@ulb.ac.be
[2] Department of Statistics and Operations Research, Kuwait University, Faculty of Science, P.O.Box 5969, Safat 13060, Kuwait, Email:fatemah@kuc01.kuniv.edu.kw.
[3] Department of Applied Economics, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium, Email:christophe.croux@econ.kuleuven.ac.be
[4] Department of Statistics, University of British Columbia, 333-6356 Agricultural Road, Vancouver, BC, V6T 1Z2, Email:ruben@stat.ubc.ca.

Keywords: Asymptotic Variance, Correlation coefficient, Influence function, Maxbias Curve, Robustness, Spearman correlation.

1 Introduction

Pearson's correlation is probably one of the most used statistical quantities. But it can seriously be affected by only one outlier. Devlin et al (1975) did show that its influence function is unbounded. Several robust measures of correlation have already been proposed in the literature. We refer to the book of Shevlyakov and Vilchevski (2001) containing a chapter on robust estimation of correlation. In this note focus is on popular nonparametric correlation measures like Spearman, Kendall and the Quadrant correlation, which are widely used in the applied sciences. Recall that for a sample $\{(x_i, y_i), 1 \leq i \leq n\}$ the Quadrant correlation is defined as

$$\hat{r}_Q = \frac{1}{n}\sum_{i=1}^{n} \text{sign}\{(x_i - \text{median}_j(x_j))(y_i - \text{median}_j(y_j))\},$$

the Kendall correlation coefficient as

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

while the Spearman coefficient is simply the standard correlation computed from the univariate ranks of the observations. While it is clear that these popular sign and rank based correlation measures have some intrinsic robustness properties, a formal robustness analysis seems not to have been completed yet.

Local robustness can be measured by means of influence functions (IF). The influence function gives us the effect that an outlying observation may have on an estimator. Computing the IF of the Quadrant, Spearman, and Kendall correlation measures is not difficult. It seems that Grize (1978) was the first one who listed expressions for the IF. From the influence functions, gross-error sensitivities and asymptotic variances can be computed. Expressions for the asymptotic variances are quite complicated in case that the population correlation deviates from zero. Since robustness often comes at the price of a loss in efficiency, a numerical comparison of efficiencies of nonparametric correlation measures compared to robust correlation measures derived from bivariate robust scatter matrices has been made by Croux and Dehon (2005). They conclude that the Spearman and Kendall correlation measures give an excellent compromise between local robustness and high efficiency. They also conducted a simulation study comparing the efficiency of the different estimators of correlation in presence of outliers.

Here we want to focus on the global robustness of the correlation measures by computing their maxbias curves. Note that is not so clear whether the breakdown point, another measure of global robustness, is a useful concept for correlation estimates (see Davies and Gather, 2005).

2 Maxbias Curves

We wish to compare the robustness performance of the robust correlation estimates using the concept of *maxbias,* which is known to provide the most complete description of the robustness properties of an estimate. Roughly speaking, the maxbias gives the worst-case asymptotic bias that can be caused by a certain fraction of contamination, ϵ, in the data. A plot of the contamination fraction, ϵ, versus the maxbias, $B(\epsilon)$, is called maxbias curve.

Little is known about the maxbias of robust correlation estimates. This gap in the literature may be partly due to the technical difficulties caused by the lack of equivariance of correlation estimates. In the absence of key equivariance properties one cannot assume without loss of generality that the "true" correlation ρ has some fixed cannonical value - e.g. $\rho = 0$ - but must work with all possible values of ρ. Comparisons also become trickier in the absence of equivariance because the order of preference of two competing robust estimates may reverse for different values of ρ.

Let H denote the bivariate distribution which is supposed to generate the good data. The correlation measure R can move upwards or downwards in presence of contamination. Therefore we define the correlation explosion curve as

$$B^+(\epsilon; R, H) = \sup_K R((1-\varepsilon)H + \varepsilon K),$$

and the correlation implosion curve as

$$B^-(\epsilon; R, H) = \inf_K R((1-\epsilon)H + \epsilon K),$$

where K can be any contaminating distribution and $0 \leq \epsilon < 1$ is the level of contamination. The final maxbias curve of a correlation functional R at the bivariate distribution H is then computed as

$$\text{Maxbias}(\epsilon; R, H) = g(B^+(\epsilon; R, H), B^-(\epsilon; R, H), R(H)).$$

The choice of the loss function g is a bit tricky, since it is not clear whether upwards or downward bias need to be treated symmetrically. A reasonable choice seems to be $g(b^+, b^-, \rho) = \max(|\tanh^{-1}(\rho) - \tanh^{-1}(b^+)|, |\tanh^{-1}(\rho) - \tanh^{-1}(b^-)|)$. Note that the above setting with an implosion and explosion curve and the choice of a loss function is similar as for maxbias curves for scale functionals (see Martin and Zamar, 1993). In most cases, the worst contaminating distribution K is a point mass distribution or Dirac measure.

Expressions for the maxbias curves of several non-parametric correlation measures have been computed and formally proved. Moreover, a location-corrected version of the Quadrant correlation is proposed, which is shown to attain Hubers' min-max bias bound. Furthermore, the maxbias curve of robust correlations derived from bivariate scatter matrices is obtained. Numerical comparisons between the different estimators have been made at the normal bivariate model.

References

C. Croux and C. Dehon (2005) Robustness versus efficiency for nonparametric correlation measures. Manuscript.

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

P.L. Davies, and U. Gather (2005). Breakdown and Groups (with discussion). *The Annals of Statistics*, to appear.

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

R.D. Martin and R.H. Zamar (1993). Bias robust estimation of scale. *The Annals of Statistics*, 2, 991–1017.

G.L. Shevlyakov and N.O. Vilchevski (2001). *Robustness in Data Analysis: Criteria and Methods.* Modern Probability and Statistics, Utrecht.