ROBUST ESTIMATION OF A CORRELATION COEFFICIENT: AN ATTEMPT OF SURVEY

G.L. Shevlyakov, P.O. Smirnov
St. Petersburg State Polytechnic University
St.Petersburg, RUSSIA
E-mail: Georgy.Shevlyakov@gmail.com

Abstract

Various groups of robust estimators of a correlation coefficient are studied. The performance of most prospective estimators in ε-contaminated normal models is examined both on small and large samples, and the best of the proposed robust estimators are revealed.

1 Introduction

The aim of robust methods is to ensure high stability of statistical inference under the deviations from the assumed distribution model. Less attention is devoted in the literature to robust estimators of association and correlation as compared to robust estimators of location and scale [6, 8, 14]. However, it is necessary to study these problems due to their widespread occurrence (estimation of the correlation and covariance matrices in regression and multivariate analysis, estimation of the correlation functions of stochastic processes, etc.), and also because of the instability of classical methods of estimation in the presence of outliers in the data.

Consider the problem of estimation of the correlation coefficient ρ between the random variables X and Y. Given the observed sample $(x_1, y_1), \ldots, (x_n, y_n)$ of a bivariate random variable (X, Y), the classical estimator of a correlation coefficient ρ is given by *the sample correlation coefficient*

$$r = \sum (x_i - \bar{x})(y_i - \bar{y}) \Big/ \left[\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2 \right]^{1/2}, \tag{1}$$

where $\bar{x} = n^{-1} \sum x_i$ and $\bar{y} = n^{-1} \sum y_i$ are the sample means.

On the one hand, the sample correlation coefficient r is a statistical counterpart of the correlation coefficient ρ. On the other hand, it is the maximum likelihood estimator of ρ for a bivariate normal distribution density

$$\mathcal{N}(x, y; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \right.$$
$$\left. \times \left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\}, \tag{2}$$

where the parameters μ_1 and μ_2 are the means, σ_1 and σ_2 are the standard deviations of the r.v.'s X and Y, respectively [9].

To illustrate the necessity in robust counterparts of the sample correlation coefficient, consider Tukey's gross error model [18] described by the mixture of normal densities ($0 \le \varepsilon < 0.5$)

$$f(x,y) = (1-\varepsilon)\mathcal{N}(x,y;0,0,1,1,\rho) + \varepsilon\mathcal{N}(x,y;0,0,k,k,\rho'), \tag{3}$$

where the first and the second summands generate "good" and "bad" data, respectively; $0 \le \varepsilon < 0.5,\ k > 1,\ \mathrm{sgn}(\rho') = -\mathrm{sgn}(\rho)$. In general, the characteristics of "bad" data, namely their component means, standard deviations and especially the correlation ρ' may significantly differ from their counterparts in the first summand.

Further, we are mostly interested in estimation of the correlation coefficient ρ of "good" data regarding "bad" data as outliers. In model (3), the sample correlation coefficient is strongly biased with regard to ρ so that the presence of outliers in the data can completely destroy the sample correlation coefficient of "good" data up to the change of its sign [3, 5].

The paper pursues two main goals: first, a brief overview of various approaches to robust estimation of correlation is given; second, some results on the performance of a selected subset of robust estimators generated by those approaches are displayed.

The paper is organized as follows. In Section 2, we successively describe various groups of robust estimators of a correlation coefficient focusing mostly on the precise results concerned with Huber's minimax approach to robust estimation. In Section 3, the Monte Carlo performance of several prospective robust estimators on small and large samples is presented. In Section 4, some conclusions are drawn.

2 The Main Approaches to Robust Correlation

2.1 Robust correlation via direct robust counterparts of the sample correlation coefficient

A natural approach to robustifying the sample correlation coefficient is to replace the linear procedures of averaging by the corresponding nonlinear robust counterparts [3, 5]

$$r_\alpha(\psi) = \Sigma_\alpha\psi(x_i - \widehat{x})\psi(y_i - \widehat{y}) / \left(\Sigma_\alpha\psi^2(x_i - \widehat{x})\Sigma_\alpha\psi^2(y_i - \widehat{y})\right)^{1/2}, \tag{4}$$

where $\widehat{x}$ and $\widehat{y}$ are some robust estimators of location, for example, the sample medians $\mathrm{med}x$ and $\mathrm{med}y$; $\psi = \psi(z)$ is a monotonic function, for instance, Huber's ψ-function: $\psi(z,k) = \max\left[-k, \min(z,k)\right]$; Σ_α is a robust analogue of a sum.

The latter transformation is based on trimming the outer order statistics with subsequent summation of the remaining ones:

$$\Sigma_\alpha z_i = nT_\alpha(z) = n(n-2r)^{-1}\sum_{i=r+1}^{n-r} z_{(i)}, \quad 0 \le \alpha \le 0.5, \quad r = [\alpha\,(n-1)],$$

where $[\cdot]$ stands for the integer part. For $\alpha = 0$, the operations of ordinary and of robust summation coincide: $\Sigma_0 = \Sigma$. The following versions of estimator (4)

$$r_\alpha = \Sigma_\alpha(x_i - \mathrm{med}x)(y_i - \mathrm{med}y) \Big/ \left[\Sigma_\alpha(x_i - \mathrm{med}x)^2\Sigma_\alpha(y_i - \mathrm{med}y)^2\right]^{1/2}$$

with $\alpha = 0.1, 0.2$ were used in [3, 5, 16] For $\alpha = 0.5$, $\widehat{x} = \mathrm{med}x$, $\widehat{y} = \mathrm{med}y$, $\psi(z) = z$, formula (4) yields *the correlation median estimator* [4, 10]

$$r_{0.5} = r_{\mathrm{COMED}} = \mathrm{med}\{(x - \mathrm{med}x)(y - \mathrm{med}\}\Big/ (\mathrm{MAD}x\,\mathrm{MAD}y)\,,$$

where $\mathrm{MAD}z = \mathrm{med}\{|z - \mathrm{med}z|\}$ stands for the median absolute deviation.

2.2 Robust correlation via nonparametric measures

An estimation procedure can be endowed with robustness properties by the use of rank statistics. The best known of them are *the quadrant (sign) correlation coefficient* [2]

$$r_{\mathrm{Q}} = n^{-1} \sum \mathrm{sgn}(x_i - \mathrm{med}x)\mathrm{sgn}(y_i - \mathrm{med}y), \tag{5}$$

that is the sample correlation coefficient between the signs of deviations from medians, and *the Spearman rank correlation coefficient* r_S that is the sample correlation coefficient between the observation ranks [17].

2.3 Robust correlation via robust regression

The problem of estimation of the correlation coefficient is directly related to the linear regression problem of fitting the straight line of the conditional expectation [9]

$$\mathrm{E}(X \mid Y = y) = \mu_1 + \beta_1(y - \mu_2), \quad \mathrm{E}(Y \mid X = x) = \mu_2 + \beta_2(x - \mu_1).$$

For the bivariate normal distribution (2), $\rho^2 = \beta_1\beta_2$. Hence, using robust estimators of slope, we arrive at the robust estimator of the form [10]

$$r_{\mathrm{REG}} = \sqrt{\widehat{\beta}_1\,\widehat{\beta}_2}\,. \tag{6}$$

For instance, we may use the least absolute values (LAV) estimators and the least median squares (LMS) estimators [11]. The corresponding estimators are referred as r_{LAV} and r_{LMS}, respectively.

2.4 Robust correlation via robust principal variables

Consider the following identity for the correlation coefficient ρ [5]

$$\rho = [\mathrm{var}(U) - \mathrm{var}(V)]\Big/[\mathrm{var}(U) + \mathrm{var}(V)], \tag{7}$$

where $U = (X/\sigma_1 + Y/\sigma_2)/\sqrt{2}$, $\quad V = (X/\sigma_1 - Y/\sigma_2)/\sqrt{2}$ are the principal variables such that $\mathrm{cov}(U, V) = 0$, $\quad \sigma_U^2 = 1 + \rho$, $\quad \sigma_V^2 = 1 - \rho$.

Following Huber [8], introduce a robust scale functional $S(X) : S(aX+b) = |a|S(X)$ and write $S^2(\cdot)$ for a robust counterpart of variance. Then a robust counterpart for (7) is given by

$$\rho^*(X, Y) = [S^2(U) - S^2(V)]\Big/[S^2(U) + S^2(V)]. \tag{8}$$

By substituting the sample robust estimates for S into (8), we obtain robust estimates for ρ [5]

$$\widehat{\rho} = [\widehat{S}^2(U) - \widehat{S}^2(V)] \Big/ [\widehat{S}^2(U) + \widehat{S}^2(V)]. \tag{9}$$

The choice of the median absolute deviation $\widehat{S} = \mathrm{MAD}x$ in (9) yields a remarkable robust estimator called *the* MAD *correlation coefficient* [10]

$$r_{\mathrm{MAD}} = (\mathrm{MAD}^2 u - \mathrm{MAD}^2 v) \Big/ (\mathrm{MAD}^2 u + \mathrm{MAD}^2 v), \tag{10}$$

where u and v are the robust principal variables

$$u = \frac{x - \mathrm{med}x}{\sqrt{2}\,\mathrm{MAD}x} + \frac{y - \mathrm{med}y}{\sqrt{2}\,\mathrm{MAD}y}, \qquad v = \frac{x - \mathrm{med}x}{\sqrt{2}\,\mathrm{MAD}x} - \frac{y - \mathrm{med}y}{\sqrt{2}\,\mathrm{MAD}y}. \tag{11}$$

Choosing Huber's trimmed standard deviation estimators as $\widehat{S}$ (see [8], p. 121), we obtain the *trimmed correlation coefficient*:

$$r_{\mathrm{TRIM}} = \left(\sum_{i=n_1+1}^{n-n_2} u_{(i)}^2 - \sum_{i=n_1+1}^{n-n_2} v_{(i)}^2 \right) \Bigg/ \left(\sum_{i=n_1+1}^{n-n_2} u_{(i)}^2 + \sum_{i=n_1+1}^{n-n_2} v_{(i)}^2 \right), \tag{12}$$

where $u_{(i)}$ and $v_{(i)}$ are the ith order statistics of the corresponding robust principal variables, n_1 and n_2 are the numbers of trimmed observations.

The general formula (12) yields the following limit cases: (i) the sample correlation coefficient r with $n_1 = 0$, $n_2 = 0$ and with the classical estimators (the sample means for location and the standard deviations for scale) in its inner structure; (ii) *the median correlation coefficient* with $n_1 = n_2 = [0.5(n-1)]$

$$r_{\mathrm{MED}} = (\mathrm{med}^2|u| - \mathrm{med}^2|v|) \Big/ (\mathrm{med}^2|u| + \mathrm{med}^2|v|). \tag{13}$$

Note that the estimators r_{MAD} (10) and r_{MED} (13) are asymptotically equivalent.

The other possibilities are connected with the use in (9) of the highly efficient and robust estimators of scale S_n and Q_n proposed by Rousseeuw and Croux in [13]: $S_n \sim \mathrm{med}_i\, \mathrm{med}_j |x_i - x_j|$, $\quad Q_n \sim \{|x_i - x_j|;\, i < j\}_{(k)}$, where $k = C_h^2$ and $h = [n/2] + 1$. The corresponding robust estimators of correlation are denoted by r_{Sn} and r_{Qn}.

2.5 Minimax variance robust estimation of correlation

The class of robust estimators of correlation (9) based on robust principal variables (11) turned out to be one of most advantageous: Huber's minimax variance approach to robust estimation [8] is realized just in this class of estimators.

In [15], it is shown that the trimmed correlation coefficient r_{TRIM} (12) is asymptotically minimax with respect to variance for ε-contaminated bivariate normal distributions

$$f(x,y) \geq (1-\varepsilon)\,\mathcal{N}(x,y;0,0,1,1,\rho), \qquad 0 \leq \varepsilon < 1. \tag{14}$$

This result holds under rather general conditions of regularity imposed on joint distribution densities $f(x, y)$ similar to the conditions under which Huber's M-estimators of scale are consistent and minimax (for details, see [15]), and under the following additional condition: the underlying joint distribution should be independent with respect to its principal variables taking the form

$$f(x,y) = \frac{1}{\sqrt{1+\rho}}\, g\left(\frac{u}{\sqrt{1+\rho}}\right) \frac{1}{\sqrt{1-\rho}}\, g\left(\frac{v}{\sqrt{1-\rho}}\right), \tag{15}$$

where the principal variables u, v are given by $u = (x+y)/\sqrt{2}$, $v = (x-y)/\sqrt{2}$, and $g(x)$ is a symmetric density $g(-x) = g(x)$ with a bounded variance. In this case ρ is just the correlation coefficient of distribution (15). The class (15) contains the standard bivariate normal density $f(x, y) = \mathcal{N}(x, y; 0, 0, 1, 1, \rho)$ if $g(x) = (2\pi)^{-1/2} \exp(-x^2/2)$.

The idea of introducing class (15) is quite plain: for any random pair (X, Y) the transformation $U = X + Y$, $V = X - Y$ yields the uncorrelated random principal variables (U, V) (actually independent for densities (15)), and robust estimation of their scales solves the problem of robust estimation of correlation between X and Y with the use of the estimators of Subsection 2.4. Thus, class (9) of estimators entirely corresponds to class (15) of distributions, and this allows to extend Huber's results on minimax M-estimators of location and scale to estimation of a correlation coefficient.

The levels of trimming n_1 and n_2 of the minimax trimmed correlation coefficient r_{TRIM} depend on the contamination parameter ε: $n_1 = n_1(\varepsilon)$ and $n_2 = n_2(\varepsilon)$ [15]. In particular, the minimax variance estimator r_{TRIM} takes the following limit forms:

- as $\varepsilon \to 1$, it tends to the median correlation coefficient r_{MED};
- if $\varepsilon = 0$, it is equivalent to the sample correlation coefficient r.

Thus, the trimmed correlation coefficient r_{TRIM} may be regarded as a correlation analog of the classical Huber's robust estimators of location and scale, namely, the trimmed mean and standard deviation.

2.6 Minimax bias robust estimation of correlation

Monte Carlo experiments [3, 5, 10] show that estimator's bias under contamination seems to be even a more informative characteristic of robustness than estimator's variance. The first result in constructing minimax bias robust estimators belongs to Huber [8]: the quadrant correlation coefficient r_{Q} (5) is asymptotically minimax with respect to bias at the mixture $F = (1-\varepsilon)G + \varepsilon H$ with G and H symmetric in $\mathbb{R}^2$.

Here we announce the following result confirming the high robustness of the median correlation coefficient.

Theorem *Under standard regularity conditions (for details, see [6], pp. 125-127), the median correlation coefficient r_{MED} (13) is asymptotically minimax with respect to bias at ε-contaminated bivariate normal distributions (14) satisfying the condition (15).*

2.7 Robust correlation via the rejection of outliers

The preliminary rejection of outliers from the data with the subsequent application of a classical estimator (for example, the sample correlation coefficient) to the rest of

the observations defines the two-stage group of robust estimators of correlation. Their variety wholly depends on the variety of the rules for detection and/or rejection of multivariate outliers based on using discriminant, component, factor analysis, canonical correlation analysis, projection pursuit, etc. (for instance, see [1, 7, 12]).

Each robust procedure of estimation inherently possesses the rule for rejection of outliers (for example, see [6, 8]), and it may seem that then there is no need for any independent procedure for rejection, at least if to aim at estimation, and therefore no need for two-stage procedures of robust estimation. However, a rejection rule may be quite informal, for example, based on a prior knowledge about the nature of outliers, and, in this case, its use can improve the efficiency of estimation.

3 Performance Evaluation: Monte Carlo Study

In this section, we compare the Monte Carlo performance (50,000 trials) of the most prospective robust estimators of a correlation coefficient from Subsection 2.5 and Subsection 2.6 at the bivariate normal distribution and at the ε-contaminated bivariate normal distribution (3) both on small ($n = 20$) and large samples ($n = 1000$).

The estimator's efficiency is defined as the ratio of the asymptotic variance of the sample correlation coefficient and the experimental estimator's variance [9]:

$$\mathrm{Eff}(\widehat{\rho}) = (1-\rho^2)^2/(n\,\mathrm{var}(\widehat{\rho})),$$

so it has sense on large samples being close to the asymptotic relative efficiency.

Table 1: Normal distribution: $\rho = 0.9$.

$n = 20$	$\mathrm{E}(r)$	$n\,\mathrm{var}(r)$	Eff	$n = 1000$	$\mathrm{E}(r)$	$n\,\mathrm{var}(r)$	Eff
r	0.895	0.049	–	r	0.899	0.036	1.000
r_{TRIM}	0.873	0.123	–	r_{TRIM}	0.899	0.069	0.522
r_{MAD}	0.852	0.292	–	r_{MAD}	0.899	0.101	0.359
r_{MED}	0.832	0.311	–	r_{MED}	0.899	0.101	0.356
r_{Sn}	0.871	0.164	–	r_{Sn}	0.900	0.062	0.580
r_{Qn}	0.881	0.103	–	r_{Qn}	0.900	0.045	0.801

4 Concluding Remarks

Normal distribution. From Table 1 it follows that

1) on small and large samples, the best is the sample correlation coefficient r both in bias and variance;

2) the r_{MAD} and r_{MED} estimators practically repeat each other in behavior;

Table 2: Contaminated normal distribution: $\rho = 0.9$, $\rho' = -0.9$, $k = 10$.

$n = 20$	E(r)	$n\,\mathrm{var}(r)$	Eff	$n = 1000$	E(r)	$n\,\mathrm{var}(r)$	Eff
r	-0.330	8.771	–	r	-0.747	1.435	0.025
r_{TRIM}	0.810	0.210	–	r_{TRIM}	0.812	0.104	0.345
r_{MAD}	0.838	0.322	–	r_{MAD}	0.887	0.124	0.290
r_{MED}	0.795	0.434	–	r_{MED}	0.887	0.125	0.288
r_{Sn}	0.844	0.189	–	r_{Sn}	0.880	0.100	0.362
r_{Qn}	0.844	0.191	–	r_{Qn}	0.874	0.084	0.430

3) on large samples, the biases of estimators can be neglected, but not their variances;

4) the best estimator among robust estimators is the r_{Qn}.

Contaminated normal distribution. From Table 2 it follows that

1) the sample correlation coefficient r is catastrophically bad under contamination;

2) on small samples, the best estimators are the r_{Sn} and r_{Qn} both with respect to bias and to variance;

3) on large samples, the r_{Sn} and r_{Qn} are superior in variance and efficiency, but the MAD and median correlation coefficients are better in bias confirming their asymptotic minimax bias properties;

4) under heavy contamination, the bias of an estimator is a more informative characteristic than its variance.

The former Monte Carlo studies of robust estimators of correlation [3, 5, 10, 16] show that the estimators based on robust principal variables, namely, r_{TRIM}, r_{MAD} and r_{MED}, generally dominate over all the other robust estimators aforementioned in Section 2 including a direct robust counterpart of the sample correlation coefficient r_{COMED}, nonparametric r_Q and r_S, regression r_{LAV} and r_{LMS}, a two-stage r_{ELL}. In our Monte Carlo study, we have obtained that the r_{Qn} estimator is better than the others, so we may conclude that it is generally the best over the initially chosen set of estimators. However, its computation is much more time consuming than of its competitors.

References

[1] Atkinson A., Riani M. (2000). *Robust Diagnostics Regression Analysis*. Springer, New York.

[2] Blomqvist N. (1950). On a Measure of Dependence between Two Random Variables. *The Annals of Mathematical Statistics*. Vol. **21**, pp. 593-600.

[3] Devlin S.J., Gnanadesikan R., Kettenring J.R. (1975). Robust Estimation and Outlier Detection with Correlation Coefficient. *Biometrika*. Vol. **62**, pp. 531-545.

[4] Falk M. (1998). A Note on the Correlation Median for Elliptical Distributions. *Journal of Multivariate Analysis*. Vol. **67**, pp. 306-317.

[5] Gnanadesikan R., Kettenring J.R. (1972). Robust Estimates, Residuals and Outlier Detection with Multiresponse Data. *Biometrics*. Vol. **28**, pp. 81-124.

[6] Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (1986). *Robust Statistics. The Approach Based on Influence Functions*. John Wiley, New York.

[7] Hawkins D.M. (1980). *The Identification of Outliers*. Chapman & Hall, London.

[8] Huber P.J. (1981). *Robust Statistics*. John Wiley, New York.

[9] Kendall M.G., Stuart A. (1963). *The Advanced Theory of Statistics. Inference and Relationship*. Griffin, London.

[10] Pasman V.R., Shevlyakov G.L. (1987). Robust Methods of Estimation of a Correlation Coefficient. *Automation and Remote Control*. pp. 332-340.

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

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

[13] Rousseeuw P.J., Croux C. (1993). Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*. Vol. **88**, pp. 1273-1283.

[14] Maronna R., Martin D., Yohai V. (2006). *Robust Statistics. Theory and Methods*. John Wiley, New York.

[15] Shevlyakov G.L., Vilchevsky N.O. (2002). Minimax Variance Estimation of a Correlation Coefficient for Epsilon-Contaminated Bivariate Normal Distributions. *Statistics and Probability Letters*. Vol. **57**, pp. 91-100.

[16] Shevlyakov G.L., Vilchevski N.O. (2002). *Robustness in Data Analysis: criteria and methods*. VSP, Utrecht.

[17] Spearman C. (1904). The Proof and Measurement of Association between Two Things. *American Journal of Psychology*. Vol. **15**, pp. 88-93.

[18] Tukey J.W. (1960). A Survey of Sampling from Contaminated Distributions. In *Contributions to Probability and Statistics*. (I. Olkin, Ed.), Stanford University Press, Stanford, pp. 448-485.