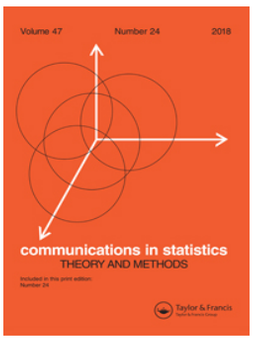

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

S_n covariance

Sajana O. Kunjunni & Sajesh T. Abraham

To cite this article: Sajana O. Kunjunni & Sajesh T. Abraham (2019): S_n covariance, Communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2019.1628275

To link to this article: https://doi.org/10.1080/03610926.2019.1628275

Published online: 16 Jun 2019.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
https://doi.org/10.1080/03610926.2019.1628275

S_n covariance

Sajana O. Kunjunni and Sajesh T. Abraham

Department of Statistics, St Thomas' College (Autonomous), Thrissur, Kerala, India

ABSTRACT

Main purpose of this paper is to study a robust measure of estimating dependence between random variables that can be used as an alternative to classical covariance estimator. An efficient univariate nested L-estimator (repeated median) S_n with high breakdown point is used to define bivariate dispersion. Results regarding in the characteristics of proposed estimator is discussed through this paper.

ARTICLE HISTORY
Received 6 March 2019
Accepted 31 May 2019

KEYWORDS
Robust estimation; scale; covariance; finite efficiency

1. Introduction

Robust inferences have significant role in the field of statistical analysis. It is necessary to estimate location and scatter parameters robustly for noise free analysis. Median is the most extensively known robust estimator for location of a random variable X. Usually, it gives $[n/2]^{th}$ order statistic from n independent observations $x_1, x_2 \ldots x_n$ of X when n is odd. In the case where n is even, median is the average of $[n/2]^{th}$ and $([n/2]+1)^{th}$ order statistic. It is clear that median posses the optimal breakdown point 50%.

Several median based scale estimators are available in literature. A very popular median based robust scale estimator is Median Absolute Deviation from median (MAD) raised by Hampel (1974) and he established that it is an approximation of M estimator of scale. The asymptotic variance and influence function of MAD was derived by Huber (1981). A detailed study on limit theorems and strong consistency of MAD has been developed by Hall and Welsh (1985). MAD has breakdown point which is equal to that of median. Despite of high breakdown value, MAD has only 37% Gaussian efficiency in symmetric distributions. More efficient alternative for MAD with 50% breakdown point is discussed by Rousseeuw and Croux (1993). A pairwise distance estimator $Q_n(X) = 2.2219\{|x_i - x_j|; i < j\}_{(k)}$ where $k \approx \frac{\binom{n}{2}}{4}$ is one of the alternative to MAD. Another reliable substitute for MAD is

$$S_n(X) = 1.1926 \underset{i}{med}\, \underset{j}{med}\, |x_i - x_j|$$

CONTACT Sajana O. Kunjunni sajana.kunjunni@gmail.com Department of Statistics, St Thomas' College (Autonomous), Thrissur, Kerala 680001, India.

where *med* stands for low median ($[\frac{n+1}{2}]^{th}$ order statistic) for outer median and high median ($([\frac{n}{2}]+1)^{th}$ order statistic) for inner median and 1.1926 is the consistency factor for normal distributions. S_n estimator of scale assures bounded influence function and optimal breakdown point 50%. Even though S_n is less efficient than Q_n, S_n is more applicable because of its low gross error sensitivity. Thus, S_n is more robust than Q_n (Rousseeuw and Croux 1993).

Sample covariance is not robust against existence of possible outliers. Falk (1997) proposed a median based robust alternative to the sample covariance between two random variables X and Y called comedian. It turns out that MAD is a special case of comedian. Similar type of estimator which robustly measure the degree of relation between two random variables is discussed in this paper. Location free scale estimator is used to define the dependence among two variables. The scope for location free robust covariance estimator is discussed in Falk (1997). Characteristics of proposed robust covariance is compared to classical covariance and comedian by utilizing theoretical and empirical results.

2. Robust covariance estimator

Consider the bivariate random variable (X, Y) and let $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ be n independent observations of (X, Y). Then $\bar{X} = \sum_{i=1}^{n} X_i$ and $\bar{Y} = \sum_{i=1}^{n} Y_i$ are the sample means of X and Y respectively. Empirical covariance between X and Y is defined as

$$\widehat{COV(X, Y)} = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n) \tag{1}$$

It is clear that, $\widehat{COV(X, Y)}$ is highly influenced by the presence of outliers which decreases its breakdown point $1/n$. Asymptotically it will become zero.

In this paper a robust alternative for covariance estimation is proposed and is denoted by $S_nCov(X, Y)$. It is defined as

$$S_nCov(X, Y) = \underset{i}{med} \left\{ \underset{j \neq i}{med} \left[(x_i - x_j)(y_i - y_j) \right] \right\} \tag{2}$$

where $1 \leqslant i, j \leqslant n$ and *med* stands for low median ($[\frac{n+1}{2}]^{th}$ order statistic). The square of consistency factor (1.1926) of $S_n(x)$ can be multiplied to $S_nCov(X, Y)$ in order to get consistency at normal distribution. The repeated use of median was introduced by Tukey (1977) and it is applied in estimation of linear regression by Siegel (1982). Clearly, the defined robust covariance estimator is designed on the basis of repeated median idea. Due to lemma by Siegel (1982), repeated median values are bounded. Further properties of proposed estimators are discussed below.

Assume that $\{z_i = (x_i, y_i); i = 1, ...n\}$ are independent observations from a Euclidean space $\mathcal{X}$ with common distribution $L = F \times G$ (F and G are distribution functions of X and Y respectively). Define a kernel function $u : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ where $u(z_i, z_j) = (x_i - x_j)(y_i - y_j)$. Let T_1 and T_2 be sample medians. For each z, let $U(z) = T_1(H_z(t))$ and $\theta = T_2(H)$ where $H_z(t)$ and H are distribution functions $u(z, Z)$ and $U(Z)$ respectively. Here, θ be the covariance that need to estimate. For estimating θ, first estimate $U(z_i)$

by $\widehat{U}(z_i) = T_1(H_{z_i,n-1}(t))$, where $H_{z_i,n-1}(t)$ is the empirical distribution of $\{u(z_i, z_j); j \neq i, i \text{ fixed}\}$. Then put $\widehat{\theta}_n = T_2(H_n)$, where H_n is the empirical distribution of $\widehat{U}(z_1), ..., \widehat{U}(z_n)$.

Now, define the distribution function $H_z(t)$ i.e,

$$H_z(t) = P(u(z,Z) \leq t) = 1 - \int_{-\infty}^{x} \int_{-\infty}^{y-\frac{t}{x-x'}} l(x',y')d(y')d(x') - \int_{x}^{\infty} \int_{y-\frac{t}{x-x'}}^{\infty} l(x',y')d(y')d(x') \tag{3}$$

where l is the density function of L

Lemma 1. *If X and Y are independent and continuous with* $F^{-1}(0.5) = G^{-1}(0.5) = 0$, *then* $S_nCov(X,Y) = 0$

Proof. Since X and Y are independent

$$H_z(t) = 1 - \int_{-\infty}^{x} G\left(y - \frac{t}{x-x'}\right) dF(x') - \int_{x}^{\infty} \left[1 - G\left(y - \frac{t}{x-x'}\right)\right] dF(x') \tag{4}$$

Here, $U(z)$ solves for $H_z(U) = 0.5$. By Equation (4), and since

$$sgn(F(t) - 0.5) = sgn(G(t) - 0.5) = sgn(t)$$

where $sgn(t) = -1$ if $t<0, = 0$ if $t = 0$ and $= 1$ if $t>0$

$$\begin{aligned} H_z(0) = 1 - F(x)G(y) - (1-F(x))(1-G(y)) &< 0.5, \text{ if } sgn(xy) > 0 \\ &= 0.5, \text{ if } sgn(xy) = 0 \\ &> 0.5, \text{ if } sgn(xy) < 0 \end{aligned}$$

Hence, based on the similar arguments that of Hössjer, Rousseeuw and Croux (1992), $sgn(U(z)) = sgn(xy)$.

Also

$$H(0) = P(sgn(xy) \leq 0) = 1 - F(0)G(0) - (1-F(0))(1-G(0)) = 0.5$$

This follows,

$$\begin{aligned} H^{-1}(0.5) = \underset{Z \sim L}{med}\, U(Z) &= 0 \\ S_nCov(X,Y) &= 0 \end{aligned}$$

□

Considering the two equalities in lemma 1.1 and 1.3 by Falk (1997), it is clear that $S_n(aX+b) = |a|S_n(X)$ and $S_nCov(X,Y) = aS_n(X)^2$ when $Y = aX + b$ where $a, b \in \mathbb{R}$. Let $X = Y, S_nCov(X,X) = S_n(X)^2$, this states that S_n is a special case of S_nCov. Moreover S_nCov is symmetric, location invariant and scale equivariant, i.e.

$$S_nCov(X, aY+b) = aS_nCov(X,Y) = aS_nCov(Y,X)$$

A robust location free alternative to the coefficient of correlation $\rho = \frac{COV(X,Y)}{\sigma_x \sigma_y}$ is therefore the S_n*correlation* is denoted by $\xi(X,Y)$ is defined as

$$\xi = \xi(X,Y) = \frac{S_nCov(X,Y)}{S_n(X)S_n(Y)}$$

By lemma 1, if X and Y are independent and symmetric $\xi(X, Y) = 0$. Similarly in the case where there is complete dependence i.e $Y = aX + b$ for bivariate normal random variable $\xi(X, Y) = sgn(a), almost\ surely$. Hence, $\xi \in \{-1, 1\}$.

3. Simulation study

Empirical properties of proposed estimator is discussed by Monte Carlo experiments. Nonparametric correlation coefficient estimators are compared by Croux and Dehon (2010) using finite sample variances. Finite sample efficiencies are estimated through Mean Square Error (MSE) and it is defined as

$$MSE = \frac{1}{k}\sum_{i=1}^{k}(\widehat{\rho}-\rho)^2 \tag{5}$$

where $\widehat{\rho}$ is the estimated correlation coefficient.

Survey and comparison of different approaches of robust correlation coefficients are presented by Shevlyakov and Smirnov (2011). Their study includes robust correlation coefficients introduced by Gnanadesikan and Kettenring (1972), which is defined as

$$r_{\tilde{\sigma}} = \frac{\tilde{\sigma}^2(v_1)-\tilde{\sigma}^2(v_2)}{\tilde{\sigma}^2(v_1)+\tilde{\sigma}^2(v_2)}$$

where $v_1 = (X/\tilde{\sigma}(X) + Y/\tilde{\sigma}(Y))/\sqrt{2}, v_2 = (X/\tilde{\sigma}(X)-Y/\tilde{\sigma}(Y))/\sqrt{2}$ and $\tilde{\sigma}$ is the robust estimators of scale. Substituting MAD, S_n and Q_n for $\tilde{\sigma}$ provides robust estimator for $r_{\tilde{\sigma}}$. The corresponding robust estimators of correlation coefficients are denoted by r_{MAD}, r_{S_n} and r_{Q_n}. From the Monte Carlo study among these estimators presented in Shevlyakov and Smirnov (2011), r_{Q_n} shows better performance.

The following simulation aims to give a performance evaluation of robust correlation coefficient defined using proposed covariance estimator. The MSE of proposed estimator is compared with the robust correlation coefficients discussed in Shevlyakov and Smirnov (2011), correlation median established by Falk (1997) and the classical coefficient of correlation r.

The simulation is performed for $k = 10000$, $\rho = 0.8$ and different sample sizes n. The results are presented in Tables 1 and 2. Table 1 shows the empirical $n*$MSEs of various estimators for data sets without outliers. The simulation is performed with varying amount of contamination and the results for 30% contamination is presented in Table 2. The results are similar in other cases as well. Table 1 shows that error of ξ is less for small sample sizes as compared to correlation median. On small and large

Table 1. $n*$MSE in Symmetric distribution.

	n				
	20	50	100	200	1000
Correlation median	1.825	2.17	2.465	4.040	10.333
ξ	1.200	1.54	2.241	5.431	16.312
r_{MAD}	0.778	0.502	0.432	0.389	0.366
r_{S_n}	0.462	0.293	0.246	0.238	0.248
r_{Q_n}	0.326	0.206	0.176	0.175	0.164
r	0.173	0.151	0.135	0.137	0.133

Table 2. $n*$MSE in Symmetric distribution with 30% outlier.

	n				
	20	50	100	300	1000
Correlation median	0.788	1.190	1.485	2.444	5.673
ξ	0.741	0.745	0.767	0.889	1.734
r_{MAD}	0.529	0.598	1.176	4.345	17.038
r_{S_n}	0.439	0.643	1.416	5.330	19.645
r_{Q_n}	0.288	0.608	1.393	4.857	17.203

sample sizes, classical coefficient of correlation r is the best for symmetric samples. From Table 2, it is clear that in contaminated situation the proposed method perform better than the other methods when $n>50$.

4. Conclusion

An efficient alternative for robust covariance estimator on the basis of repeated median is investigated. Purpose of this paper is to find an alternative for comedian using S_n scale estimator which is the more efficient and robust alternatives for MAD. Covariance based on S_n estimator motivates to realize that it is a nested L-estimator. A non-parametric measure of covariance and coefficient of correlation are proposed through this paper. The $S_nCov(X, Y)$ satisfies characteristic of sample covariance in independent and symmetric random variables. The proposed estimator assures location invariant and scale equivariant properties as well. The efficiency of $S_ncorrelation$ is greater than median correlation in terms of MSE. The error estimate of $S_ncorrelation$ is lower than that of r_{Q_n} for large samples in contaminated situations. This specifies that $S_ncorrelation$ is more resistant than other estimates to the presence of outliers in the large sample cases. The direct generalization of proposed method to higher dimension is not be possible as the corresponding covariance matrix may not be positive definite. But an orthogonalization similar to that of developed by Maronna and Zamar (2002) may be helpful in higher dimensional cases.

Acknowledgment

The authors are thankful to the reviewer for their valuable comments and efforts towards improving our manuscript.

References

Croux, C., and C. Dehon. 2010. Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications* 19 (4):497–515. doi:10.1007/s10260-010-0142-z.

Falk, M. 1997. On mad and comedians. *Annals of the Institute of Statistical Mathematics* 49 (4): 615–44. doi:10.1023/A:1003258024248.

Gnanadesikan, R., and J. R. Kettenring. 1972. Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics* 28 (1):81–124. doi:10.2307/2528963.

Hall, P., and A. H. Welsh. 1985. Limit theorems for the median deviation. *Annals of the Institute of Statistical Mathematics* 37 (1):27–36. doi:10.1007/BF02481078.

Hampel, F. R. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69 (346):383–93. doi:10.2307/2285666.

Hössjer, O., P. J. Rousseeuw, and C. Croux. 1992. *Influence function and asymptotic normality of the repeated median slope estimator. Report 1992:2, Department of Mathematics*. Uppsala, Sweden: Uppsala University.

Huber, P. J. 1981. *Robust statistics*. New York: John Wiley and Sons.

Ma, Y., and M. G. Genton. 2001. Highly robust estimation of dispersion matrices. *Journal of Multivariate Analysis* 78 (1):11–36. doi:10.1006/jmva.2000.1942.

Maronna, R. A., and R. H. Zamar. 2002. Robust estimates of location and dispersion of high-dimensional datasets. *Technometrics* 44 (4):307–17. doi:10.1198/004017002188618509.

Rousseeuw, P. J., and C. Croux. 1993. Alternatives to the median absolute deviation. *Journal of the American Statistical Association* 88 (424):1273–83. doi:10.2307/2291267.

Shevlyakov, G., and P. Smirnov. 2011. Robust estimation of the correlation coefficient: An attempt of survey. *Austrian Journal of Statistics* 40:147–56.

Siegel, A. F. 1982. Robust regression using repeated medians. *Biometrika* 69 (1):242–4. doi: 10.2307/2335877.

Tukey, J. W. 1977. *Exploratory data analysis*. Reading: Massachusetts: Addison-Wesley.