F. Schmid and R. Schmidt, Bootstrapping Spearman's Multivariate Rho. *Proc. Compstat* (in print), 2006.

Bootstrapping Spearman's Multivariate Rho

Friedrich Schmid[1] and Rafael Schmidt[12]

Summary. Spearman's rho can be generalized to the multivariate, i.e. d-dimensional case. Nonparametric estimation of Spearman's multivariate rho has recently been considered and the asymptotic normality for the estimator has been established. Though a closed and compact formula for the asymptotic variance exists, it is not suitable for practical application. Therefore a bootstrap procedure was suggested. This note investigates the performance of the bootstrap in finite samples by Monte Carlo simulation.

Key words: Copulas, spearman's multivariate rho, asymptotic normality of estimators, asymptotic variance, bootstrapping.

1 Introduction

Spearman's rho is a widely used measure for the amount of association between two random variables X and Y. It does not depend on the marginal distributions of X and Y but can be written as a function of their copula, which represents their dependence structure. Spearman's rho can be generalized to a (multivariate) measure of association or measures of dependence between d random variables $X_1, ..., X_d$ in various ways (see [Ken70] §6, [Wol80], [Joe90], [Nel96], [SS05]). This is of interest in many fields of application, e.g. in the multivariate analysis of financial asset returns where one wants to express the amount of dependence in a portfolio by a single number.

Nonparametric estimation of Spearman's multivariate rho has been considered in [Joe90], [Ste03], [SS05]. Using empirical process theory, the latter authors derived the asymptotic normality for various types of nonparametric estimators and established compact expressions for the asymptotic variances which are determined by the copula and its partial derivatives. They are, however, of limited use for practical application since the copula is not known in general, but has to be estimated. Therefore a bootstrap algorithm was suggested and it was proven that the bootstrap works well asymptotically.

[1] Department of Economic and Social Statistics, University of Cologne, Germany
[2] Department of Statistics, London School of Economics, UK. The author gratefully acknowledges financial support by the Deutsche Forschungsgemeinschaft (DFG).

The aim of this note is to investigate the performance of the bootstrap for one particular estimator of Spearman's multivariate rho in finite samples. The investigation is carried out via a Monte Carlo simulation utilizing special copulas.

The structure of the paper is as follows. Section 2 introduces some notation. Section 3 defines Spearman's multivariate rho and presents some asymptotic theory regarding its nonparametric estimation. Section 4 investigates the performance of the corresponding bootstrap for special copulas.

2 Preliminary

Throughout the paper we write bold letters for vectors, e.g., $\mathbf{x} := (x_1, ..., x_d) \in \mathbb{R}^d$. Inequalities $\mathbf{x} \leq \mathbf{y}$ are understood componentwise, i.e, $x_i \leq y_i$ for all $i = 1, ..., d$. The indicator function on a set A is denoted by 1_A . Let $X_1, X_2, ..., X_d$ be $d \geq 2$ random variables with joint distribution function

$$F(\mathbf{x}) = P(X_1 \leq x_1, ..., X_d \leq x_d), \quad \mathbf{x} = (x_1, ..., x_d) \in \mathbb{R}^d,$$

and marginal distribution functions $F_i(x) = P(X_i \leq x)$ for $x \in \mathbb{R}$ and $i = 1, ..., d$. We will always assume that the F_i are continuous functions. Thus, according to Sklar's theorem [Skl59], there exists a unique *copula* C: $[0,1]^d \longrightarrow [0,1]$ such that

$$F(\mathbf{x}) = C(F_1(x_1), ..., F_d(x_d)) \text{ for all } \mathbf{x} \in \mathbb{R}^d.$$

The copula C is the joint distribution function of the random variables $U_i = F_i(X_i)$, $i = 1, ..., d$. Moreover, $C(\mathbf{u}) = F(F_1^{-1}(u_1), ..., F_d^{-1}(u_d))$ for all $\mathbf{u} \in [0,1]^d$ where the generalized inverse function F^{-1} is defined via $F^{-1}(u) := \inf \{x \in \mathbb{R} \cup \{\infty\} \,|F(x) \geq u\}$ for all $u \in [0,1]$. A detailed treatment of copulas is given in [Nel99] and [Joe97].

It is well known that every copula C is bounded in the following sense:

$$\begin{aligned} W(\mathbf{u}) &:= \max \{u_1 + ... + u_d - (d-1), 0\} \\ &\leq C(\mathbf{u}) \leq \min \{u_1, ..., u_d\} =: M(\mathbf{u}) \text{ for all } \mathbf{u} \in [0,1]^d, \end{aligned}$$

where M and W are called the upper and lower *Fréchet-Hoeffding bounds*, respectively. The upper bound M is a copula itself and is also known as the comonotonic copula. It represents the copula of $X_1, ..., X_d$ if $F_1(X_1) = ... = F_d(X_d)$ with probability one, i.e., where there is (with probability one) a strictly increasing functional relationship between X_i and X_j $(i \neq j)$. Another important copula is the independence copula

$$\Pi(\mathbf{u}) := \prod_{i=1}^{d} u_i, \quad \mathbf{u} \in [0,1]^d,$$

describing the dependence structure of stochastically independent random variables $X_1, ..., X_d$.

3 Spearman's Multivariate Rho and its Estimation

Spearman's rho for a two dimensional random vector $\mathbf{X} = (X_1, X_2)$ with copula C can be written as

$$\rho = \frac{cov(U_1, U_2)}{\sqrt{var(U_1)}\sqrt{var(U_2)}} = \frac{\int_0^1\int_0^1 uv\ dC(u,v) - (\frac{1}{2})^2}{1/12}$$
$$= \frac{\int_0^1\int_0^1 C(u,v)\ dudv - 1/4}{1/3 - 1/4} = \frac{\int_0^1\int_0^1 C(u,v)\ dudv - \int_0^1\int_0^1 \Pi(u,v)\ dudv}{\int_0^1\int_0^1 M(u,v)\ dudv - \int_0^1\int_0^1 \Pi(u,v)\ dudv},$$

because of $\int_0^1\int_0^1 M(u,v)\ dudv = 1/3$ and $\int_0^1\int_0^1 \Pi(u,v)\ dudv = 1/4$. Thus, ρ can be interpreted as the normalized average distance between the copula C and the independence copula $\Pi(u,v) = uv$. The following d-dimensional extension of ρ is now straightforward

$$\rho = \frac{\int_{[0,1]^d} C(\mathbf{u})\ d\mathbf{u} - \int_{[0,1]^d} \Pi(\mathbf{u})\ d\mathbf{u}}{\int_{[0,1]^d} M(\mathbf{u})\ d\mathbf{u} - \int_{[0,1]^d} \Pi(\mathbf{u})\ d\mathbf{u}} = \frac{d+1}{2^d - (d+1)}\left\{2^d \int_{[0,1]^d} C(\mathbf{u})\ d\mathbf{u} - 1\right\}.$$

Consider a random sample $(\mathbf{X}_j)_{j=1,...,n}$ from a d -dimensional random vector $\mathbf{X}$ with joint distribution function F and copula C which are completely unknown. It is further assumed, that the marginal distribution functions F_i are unknown. They are estimated by their empirical counterparts

$$\hat{F}_{i,n}(x) = \frac{1}{n}\sum_{j=1}^{n} 1_{\{X_{ij} \leq x\}}, \quad \text{for } i = 1,...,d \text{ and } x \in \mathbb{R}.$$

Further, set $\hat{U}_{ij,n} := \hat{F}_{i,n}(X_{ij})$ for $i = 1,...,d$, $j = 1,...,n$, and $\hat{\mathbf{U}}_{j,n} = (\hat{U}_{1j,n},...,\hat{U}_{dj,n})$. Note that $\hat{U}_{ij,n} =$ (rank of X_{ij} in $X_{i1},...,X_{in})/n$. The estimation of ρ will therefore be based on ranks (and not on the observations itself). In other words, we consider rank order statistics. The copula C is estimated by the empirical copula which is defined as

$$\hat{C}_n(\mathbf{u}) = \frac{1}{n}\sum_{j=1}^{n}\prod_{i=1}^{d} 1_{\{\hat{U}_{ij,n} \leq u_i\}} \quad \text{for } \mathbf{u} = (u_1,...,u_d) \in [0,1]^d.$$

A nonparametric estimator of ρ is now given by

$$\hat{\rho}_n = h(d)\left\{2^d \int_{[0,1]^d} \hat{C}_n(\mathbf{u})d\mathbf{u} - 1\right\} = h(d)\left\{\frac{2^d}{n}\sum_{j=1}^{n}\prod_{i=1}^{d}(1 - \hat{U}_{ij,n}) - 1\right\},$$

where $h(d) = (d+1)/(2^d - d - 1)$. Asymptotic normality of $\hat{\rho}_n$ is stated next.

Proposition 1. *Let F be a d-dimensional distribution function with copula C and continuous marginal distribution functions F_i . Further assume that the partial derivatives $D_i C(\mathbf{u})$ exist and are continuous for $i = 1, ..., d$. Then*

$$\sqrt{n}(\hat{\rho}_n - \rho) \xrightarrow{d} Z \sim N(0, \sigma^2)$$

where

$$\sigma^2 = 2^{2d} h(d)^2 \int\limits_{[0,1]^d} \int\limits_{[0,1]^d} E\left\{\mathbb{G}_C(\mathbf{u})\mathbb{G}_C(\mathbf{v})\right\} d\mathbf{u} d\mathbf{v}$$

and

$$\mathbb{G}_C(\mathbf{u}) = \mathbb{B}_C(\mathbf{u}) - \sum_{i=1}^{d} D_i C(\mathbf{u}) \mathbb{B}_C(\mathbf{u}^{(i)})$$

with D_i denoting the i-th partial derivative. The vector $\mathbf{u}^{(i)}$ denotes the vector where all coordinates, except the i-th coordinate of $\mathbf{u}$, are replaced by 1. The process $\mathbb{B}_C$ is a tight centered Gaussian process on $[0,1]^d$ with covariance function

$$E\left\{\mathbb{B}_C(\mathbf{u})\mathbb{B}_C(\mathbf{v})\right\} = C(\mathbf{u} \wedge \mathbf{v}) - C(\mathbf{u})C(\mathbf{v}),$$

i.e., $\mathbb{B}_C$ is a d-dimensional Brownian Bridge.

Even if the copula C is known, computation of σ^2 is nearly impossible as it involves $2d$-dimensional integration over $(d+1)^2$ terms (see however [SS05] for special cases such as independence). The next proposition justifies that σ^2 can be determined asymptotically by the following bootstrap.

Proposition 2. *Let $(\mathbf{X}_j^B)_{j=1,...,n}$ denote a bootstrap sample which is obtain by sampling from $(\mathbf{X}_j)$ with replacement and denote the corresponding bootstrap estimator for ρ by $\hat{\rho}_n^B$. Then, under the assumptions of Proposition 1, $\sqrt{n}(\hat{\rho}_n^B - \hat{\rho}_n)$ converges weakly to the same Gaussian random variable as $\sqrt{n}(\hat{\rho}_n - \rho)$ with probability one.*

4 Performance of the Bootstrap in Finite Samples

Since the bootstrap for $\hat{\rho}_n$ is justified asymptotically only, its performance in finite samples should be investigated. This is done in the present section for selected copulas in various dimensions d.

The d-dimensional Cook-Johnson copula (also called Clayton copula) is defined by

$$C(u_1, ..., u_d; \alpha) = \left(\sum_{i=1}^{d} u_i^{-\frac{1}{\alpha}} - d + 1 \right)^{-\alpha}$$

where $\alpha > 0$ is a shape parameter. Random number generation from the Cook-Johnson copula is described in [Dev86].

The d-dimensional equi-correlated Gaussian copula is defined by

$$C(u_1, ..., u_d; \varrho) =$$
$$= \int_{-\infty}^{\Phi^{-1}(u_1)} \ldots \int_{-\infty}^{\Phi^{-1}(u_d)} (2\pi)^{-\frac{d}{2}} det\{\Sigma(\varrho)\}^{-\frac{1}{2}} \exp\Big(- \frac{1}{2}\mathbf{x}'\Sigma(\varrho)^{-1}\mathbf{x}\Big)\, dx_d \ldots dx_1$$

where $\Sigma(\varrho) = \varrho\mathbf{1}\mathbf{1}' + (1-\varrho)I_d$ with identity matrix I_d and $\frac{1}{d-1} < \varrho < 1$.

Tables 1 and 2 summarize simulation results for these two copulas for $d = 2$, 5, and 10. The first and second column in every table contain the values of the parameter and the sample sizes, respectively. The third column contains approximation to the true value of Spearman's multivariate rho. This approximation has been derived by computing $\hat{\rho}$ - the index n will be suppressed for notational convenience - from samples of length 500,000. The first two digits behind the decimal point are accurate. Note that for the Gaussian copula and $d = 2$, Spearman's rho can be exactly computed by utilizing the relationship

$$\rho = \frac{6}{\pi} \arcsin\left(\frac{\varrho}{2}\right).$$

The fourth and sixth columns contain the empirical means $m(\hat{\rho})$ and the standard deviations $\hat{\sigma}(\hat{\rho})$ of $\hat{\rho}$ over 300 Monte Carlo replications.

Comparing the third and fourth column in every table, we observe a considerable bias for small sample sizes, such as $n = 100$, in every dimension under study. This bias is lower for $d = 5$ and 10 than it is for $d = 2$. There is a good agreement between the fourth and fifth column, i.e. between $m(\hat{\rho})$ and $m\left(\hat{\rho}^B\right)$.

The sixth column shows that the standard error $\hat{\sigma}(\hat{\rho})$ of $\hat{\rho}$ decreases with sample size n in a reasonable way. The amount of $\hat{\sigma}(\hat{\rho})$, however, heavily depends on the copula, its parameters, and the dimension d.

The seventh column contains the empirical means of the bootstrap estimations for the standard error of $\hat{\rho}$. The good agreement between the sixth and seventh column indicates that the bootstrap for the determination of the standard error of $\hat{\rho}$ works well under every parameter constellation, for both copulas and for every dimension under study.

Column 8 shows that the standard deviation of $\hat{\sigma}^B$ over 300 Monte Carlo replications is small, especially for $n = 500$ and 1000.

Finally, Column 9 provides bootstrap estimates for the asymptotic standard deviation σ, as given in Proposition 1. It can be seen that σ is well estimated by $\hat{\sigma}^B\sqrt{n}$ for both copulas and every parameter constellation under study, even for small sample size $n = 100$.

Table 1. Cook-Johnson copula. Simulation results for bootstrapping Spearman's multivariate rho $\hat{\rho}_n$ (the index n is suppressed). Results are based on 300 samples with sample sizes n generated from a d-variate Cook-Johnson copula with parameter α. The columns provide the empirical means - denoted by $m()$ - and the empirical standard deviations - denoted by $\hat{\sigma}$ - based on the simulated data and the respective bootstrap samples. The statistics with superscript B refer to the bootstrap sample. 250 bootstrap replications were drawn from each sample. The empirical standard deviation of the bootstrapped statistics is abbreviated by $\hat{\sigma}^B = \hat{\sigma}(\hat{\rho}^B)$.

α	n	ρ	$m(\hat{\rho})$	$m(\hat{\rho}^B)$	$\hat{\sigma}(\hat{\rho})$	$m(\hat{\sigma}^B)$	$\hat{\sigma}(\hat{\sigma}^B)$	$m(\hat{\sigma}^B)\sqrt{n}$
Dimension $d = 2$								
0.5	100	.681	.622	.616	.065	.063	.011	.633
	500	.681	.671	.669	.026	.028	.002	.622
	1000	.681	.677	.677	.020	.020	.001	.626
1	100	.479	.424	.419	.077	.084	.009	.838
	500	.479	.466	.465	.038	.038	.002	.839
	1000	.479	.472	.471	.025	.026	.001	.838
5	100	.135	.072	.071	.101	.100	.007	1.003
	500	.135	.121	.121	.044	.045	.002	.995
	1000	.135	.129	.128	.032	.031	.001	.993
Dimension $d = 5$								
0.5	100	.736	.698	.690	.054	.051	.008	.514
	500	.736	.729	.727	.023	.023	.002	.517
	1000	.736	.732	.731	.017	.016	.001	.519
1	100	.499	.475	.469	.067	.067	.007	.665
	500	.499	.496	.495	.030	.031	.002	.684
	1000	.499	.497	.496	.022	.022	.001	.689
5	100	.118	.106	.105	.045	.047	.009	.467
	500	.118	.116	.115	.020	.021	.002	.481
	1000	.118	.119	.118	.015	.015	.001	.487
Dimension $d = 10$								
0.5	100	.715	.656	.642	.065	.066	.011	.656
	500	.715	.701	.698	.029	.030	.003	.680
	1000	.715	.708	.706	.021	.022	.002	.683
1	100	.417	.386	.376	.082	.079	.013	.786
	500	.417	.414	.412	.039	.039	.003	.863
	1000	.417	.413	.411	.028	.028	.002	.876
5	100	.048	.045	.044	.027	.022	.014	.216
	500	.048	.048	.048	.012	.012	.004	.264
	1000	.048	.047	.047	.008	.008	.002	.269

Table 2. Gaussian copula. Simulation results for bootstrapping Spearman's multivariate rho $\hat{\rho}_n$ (the index n is suppressed). Results are based on 300 samples with sample sizes n generated from a d-variate Gaussian copula with equi-correlation parameter ϱ. The columns provide the empirical means - denoted by $m()$ - and the empirical standard deviations - denoted by $\hat{\sigma}$ - based on the simulated data and the respective bootstrap samples. The statistics with superscript B refer to the bootstrap sample. 250 bootstrap replications were drawn from each sample. The empirical standard deviation of the bootstrapped statistics is abbreviated by $\hat{\sigma}^B = \hat{\sigma}(\hat{\rho}^B)$.

ϱ	n	ρ	$m(\hat{\rho})$	$m(\hat{\rho}^B)$	$\hat{\sigma}(\hat{\rho})$	$m(\hat{\sigma}^B)$	$\hat{\sigma}(\hat{\sigma}^B)$	$m(\hat{\sigma}^B)\sqrt{n}$
Dimension $d = 2$								
0.5	100	.483	.418	.414	.076	.081	.009	.809
	500	.483	.472	.471	.032	.035	.002	.789
	1000	.483	.475	.475	.025	.025	.001	.791
0.2	100	.191	.129	.127	.097	.098	.007	.976
	500	.191	.174	.174	.043	.043	.002	.968
	1000	.191	.184	.183	.030	.031	.001	.968
-0.1	100	-.096	-.147	-.147	.097	.101	.006	1.006
	500	-.096	-.109	-.108	.045	.044	.002	.991
	1000	-.096	-.103	-.103	.030	.031	.001	.996
Dimension $d = 5$								
0.5	100	.439	.407	.403	.057	.056	.005	.556
	500	.439	.433	.432	.025	.025	.001	.566
	1000	.439	.437	.437	.019	.018	.001	.572
0.2	100	.158	.138	.137	.045	.044	.007	.437
	500	.158	.158	.158	.021	.021	.002	.463
	1000	.158	.158	.158	.014	.015	.001	.462
-0.1	100	-.069	-.080	-.079	.017	.017	.004	.170
	500	-.069	-.071	-.071	.008	.008	.001	.181
	1000	-.069	-.070	-.070	.006	.006	.001	.180
Dimension $d = 10$								
0.5	100	.285	.261	.256	.062	.054	.012	.539
	500	.285	.281	.280	.027	.027	.003	.599
	1000	.285	.282	.281	.019	.019	.002	.601
0.2	100	.063	.057	.056	.023	.021	.009	.209
	500	.063	.061	.061	.011	.011	.003	.239
	1000	.063	.062	.062	.008	.008	.001	.247
-0.1	100	-.009	-.010	-.009	.000	.000	.000	.002
	500	-.009	-.009	-.009	.000	.000	.000	.002
	1000	-.009	-.009	-.009	.000	.000	.000	.003

References

[Bor02] Borkowf, C.B.: Computing the nonnull asymptotic variance and the asymptotic relative efficiency of Spearman's rank correlation. Computational Statistics and Data Analysis, **39**, 271–286 (2002)

[DH97] Davidson, A.C., Hinkley, D.V.: Bootstrap Methods and their Applications. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press (1997)

[Dev86] Devroye, L.: Non-Uniform Random Variate Generation. Springer, Heidelberg (1986)

[Joe90] Joe, H.: Multivariate Concordance. Journal of Multivariate Analysis, **35**, 12–30 (1990).

[Joe97] Joe, H.: Multivariate Models and Dependence Concepts. Chapman and Hall, London (1997)

[Ken70] Kendall, M.G.: Rank Correlation Methods. Griffin, London (1970)

[Nel96] Nelsen, R.B.: Nonparametric measures of multivariate association. In: Distribution with Fixed Marginals and Related Topics, IMS Lecture Notes - Monograph Series 28, 223-232 (1996)

[Nel99] Nelsen, R.B.: An Introduction to Copulas. Springer, New York (1999)

[SS05] Schmid, F., Schmidt, R.: On the Asymptotic Behaviour of Spearman's Rho and Related Multivariate Extensions. Preprint (2005)

[Skl59] Sklar, A.: Fonctions de répartition à n dimensions et leurs marges, Publ. Inst. Statist. Univ. Paris **8**, 229–231 (1959)

[Ste03] Stepanova, N.A.: Multivariate rank tests for independence and their asymptotic efficiency. Math. Methods Statist. **12(2)**, 197–217 (2003).

[Wol80] Wolff, E.F.: N-dimensional measures of dependence. Stochastica **4(3)**, 175–188 (1980).