ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 216, 2008

Małgorzata Kobylińska, Wiesław Wagner***

BOOTSTRAP CONFIDENCE REGIONS BASED ON THE MAHALANOBIS DEPTH MEASURE OF TWO-DIMENSIONAL SAMPLES

ABSTRACT. Construction of confidence regions for multi-dimensional samples is usually performed with a known stochastic distribution of a random vector in question. However, for multidimensional studies of socio-economic phenomena, such an assumption is difficult to make. Bootstrap methods can be helpful. The main problem with its application is the aligning of respective vectors. To this end, depth measures are used which express the vector distance from the central vector system cluster. Among many such depth measures, the Mahalanobis measure is one of the easiest from a numerical point of view.

This paper presents a bootstrap region creation algorithm. It was illustrated for a two-dimensional sample.

Key words: depth measure, measure of depth by Mahalanobis, bootstrap methods.

I. INTRODUCTION

Bootstrap methods belong to the statistical inference methods. Their aim is to aproximate such distributions of statistics from a sample that are either estimators or test functions. Moreover, they also aim at the evaluation of parameters or function characteristics on the grounds of the specified bootstrap distribution, see Efron (1979), Domański and other (1999) and Domański and Pruska (2000).

A method of bootstrap constructions of confidence regions for the R^2 case without the loss of generality on any freely dimentional case will be given. The main question concerning the constructuion of such intervals is ordering of multi-dimentional data. For that reason the Mahalanobis depth measure was used, although other measures are also possible here (for example the Tukey depth measure). The review of such measures can be found, among others, in the works by Liu (1990), Liu and Singh (2001), and also in the work by Kobylińska and

* Ph.D., University of Warmia-and-Mazury in Olsztyn.
** Professor, Higher School of Information Technology and Management in Rzeszów.

Wagner (2000). Such ordering of bootstrap vectors (points) allows for their removal to a given level of confidence

In this study we present an algorithm for bootstrap constructions of confidence regions and give a numerical example with the stages of calculations conducted in EXCEL.

II. THEORETICAL BASIS

Let us assume that we examine a parent population because of the X two-dimentional random variable and let $X_1, X_2, ..., X_n$ be an n-element simple chance sample drawn from that population of the distribution specified by an unknown two-dimentional distribution function F_2, and let the $x_1, x_2, ..., x_n$ arrangement stand for n independent realizations of a two-dimentional chance sample. Let $\theta \in R^2$ be a two-dimentional vector of the unknown distribution parameters F_2, for which the confidence region is constructed.

In case $p = 1$, the construction of bootstrap confidence intervals amounts only to determining of confidence intervals for θ (see e.g. Domanski & others (1999).

The construction of bootstrap confidence regions in R^2 requires ordering of an arrangement of two-dimentional bootstrap vectors. This is done by using depth measure that allow for their ordering with regard to the distance form the sample centre. The measure decreases monotonically when we go away from the sample centre in any direction. This means that the vectors of the lowest depth measure values are furthest from the sample centre and at the same time they determine a convex hull. The removal of a specified number of vectors from the sample characterized by low depth measure values allows for obtaining the planned confidence level and the convex hull of the other points may be used for constructing two-dimentional bootstrap confidence regions.

The construction of a two-dimentional bootstrap confidence region is quoted after Yeh and Singh (1997). The following symbols will be used: n – size of a two-dimentional sample, $P_n^2 = \{x_1, x_2, ..., x_n)'$ – sample of n two-dimentional observations, N - number of two-dimentional bootstrap samples, $\theta_n = (\theta_1, \theta_2)'$ - vector of F_2 distribution parameters $\hat{\theta}_n = (\hat{\theta}_1, \hat{\theta}_2)'$ – vector consistent estimator θ_n from the P_n^2 sample, $S_{\hat{\theta}_n}$ – consistent estimator of variance-covariance matrix of $\hat{\theta}_n$ vector, $S_{\hat{\theta}_n}^{-1}$ - inverse matrix to $S_{\hat{\theta}_n}$ matrix, $S_{\hat{\theta}_n}^{1/2}$, $S_{\hat{\theta}_n}^{-1/2}$ - square roots of variance-covariance matrix and its inverse matrix.

The required square root matrix is determined in the following way. For a given square symmetric matrix $A = \begin{bmatrix} a_{11} & a_{21} \\ a_{21} & a_{22} \end{bmatrix}$ we determine matrix $A^{1/2}$ in the following stages:

1^0 We determine matrix $B = A^{1/2} = \begin{bmatrix} b_{11} & 0 \\ b_{21} & b_{22} \end{bmatrix}$, so as to satisfy an equation of $BB' = A$,

2^0 From the matrix equation given in 1^0, we determine the conditions: $b_{11}^2 = a_{11}$, $b_{11}^2 = a_{11}$, $b_{11}b_{21} = a_{21}$, $b_{21}^2 + b_{22}^2 = a_{22}$,

3^0Solutions of conditions in 2^0 are: $b_{11} = \sqrt{a_{11}}$, $a_{11} > 0$, $b_{21} = \dfrac{a_{21}}{\sqrt{a_{11}}}$,

$b_{22} = \sqrt{a_{22} - \dfrac{a_{21}^2}{a_{11}}}$.

Analogous symbols are used for the bootstrap samples, introducing only an additional symbol of *, thus we have vector $\hat{\theta}_n^*$ and matrix $S_{\hat{\theta}_n}^*$. An adequate symbol for the root matrix and its converse is used.

Suppose $T_n = n^{1/2} S_{\hat{\theta}_n}^{-1/2}(\hat{\theta}_n - \theta)$, where $\hat{\theta}_n$ and $S_{\hat{\theta}_n}$ are consistent estimators for the vector of θ parameter and its variance-covariance matrix in F_2 distribution. A bootstrap equivalent of T_n sample vector is $T_n^* = n^{1/2} S_{\hat{\theta}_n}^{*-1/2}(\hat{\theta}_n^* - \hat{\theta}_n)$. There are N of such two-dimentional vectors. They are constructed in the same way as for the initial sample P_n. Using the arrangement of N two-dimesional vectors, we determine for them the vector of averages and its variance-covariance matrix and then the Mahalanobis distances d_n^*, $d_n^* = (T_n^* - \overline{T}^*) S_{T_n^*}^{-1} (T_n^* - \overline{T}^*)^T$. For each T_n^* vector the Mahalanobis depth measure $z_M(d_n^*) = 1/(1 + d_n^*)$ is determined. These measures allow to order the T_n^* bootstrap vectors in a non-decreasing sequence and reject $m = [N\alpha]$ of the lowest values. The rest of the $N - m$ vectors will construct a set of $V_{n,1-\alpha}^*$, which makes a convex hull. Finally a $100(1-\alpha)\%$ bootstrap confidence region is described by the following set:

$$A_{n,1-\alpha}^* = \left\{ \theta_n - \frac{1}{\sqrt{n}} S_{\hat{\theta}_n}^* \cdot \omega;\ \omega \in V_{n,1-\alpha}^* \right\}. \tag{1}$$

Thus the determined region takes a form of some $\hat{\theta}_n$ environment of a point on the plane which corresponds to the sample estimation of the θ parameters vector.

III. ALGORITHM OF BOOTSTRAP CONSTRUCTIONS OF CONFIDENCE REGIONS

The construction of the bootstrap confidence regions will be illustrated with an example of the expected values vector in $N_2(\mu,\Sigma)$, a two-dimensional normal distribution whose vectors were determined according to the following algorithm:

1^0 We generate random numbers of space R_1, R_2 from the $J(0, 1)$ uniform distribution, and obtain pairs of $(R_{1i}, R_{2i}) \in J(0,1)$ for $i = 1,2,...,n$,

2^0 We transfer the (R_{1i}, R_{2i}) number pairs to the $N_2(0,I)$ standard two-dimentional normal distribution using the Box – Muller transformation (see e.g. Wieczorkowski and Zieliński 1997)

$$U_{1i} = \sqrt{-2\ln R_{1i}}\cos(2\pi R_{2i}), \quad U_{2i} = \sqrt{-2\ln R_{2i}}\sin(2\pi R_{2i}).$$

Vectors $(R_{1i}, R_{2i})'$ arrange themselves in a unitary square of $(0, 1)^2$, and the $(U_{1i}, U_{2i})'$ vectors belong to the region that is placed centrally in relation to the beginning of the co-ordinate system and whose radius is 3,

3^0 We give vector of the expected values of $\mu = (\mu_1, \mu_2)$ and variance-covariance matrix of $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$, where $\sigma_{11} = \sigma_1^2 = D^2(X_1)$, $\sigma_{22} = \sigma_2^2 = D^2(X_2)$ and $\sigma_{12} = \sigma_{21} = Cov(X_1, X_2)D(X_1)D(X_2) = \rho\sigma_1\sigma_2$, and $\rho = Corr(X_1, X_2)$,

4^0 We determine the values of $X_{1i} = c_{11}U_{1i}$ and $X_{2i} = c_{21}U_{1i} + c_{22}U_{2i}$ of the random variable of $X = [X_1, X_2]$ of two-dimensional normal distribution of $N_2(\mu,\Sigma)$ with given parameters of μ and Σ,

5^0 We conduct the distribution of the Σ na $\Sigma = CC'$ matrix, where matrix $C = \begin{bmatrix} c_{11} & 0 \\ c_{21} & c_{22} \end{bmatrix}$ is determined according to the formulas supplied in chapter two,

6^0 We calculate N of the T_n^* two-dimensional bootstrap vectors, and then we calculate for them the vector of average values, covariance matrix and the Mahalanobis depth measure values,

7^0 We arrange depth measures into a decreasing sequence and cut the left tail, which corresponds to $m = [N \cdot \alpha]$ of the lowest values,

8^0 The $N - m$ subset of the $\omega = [\omega_1, \omega_2]$ vectors obtained from stage 7^0 is used for determining the $V_{n,1-\alpha}^*$ region, where $\omega \in V_{n,1-\alpha}^*$, and is finally used for determining of the $A_{n,1-\alpha}^*$. bootstrap confidence region.

IV. NUMERICAL EXAMPLE

The construction of the bootstrap confidence region is illustrated with the example of the two-dimensional random variable distribution (X_1, X_2) of N_2 distribution with zero vector of expected values of μ and covariance matrix of $\Sigma = \begin{bmatrix} 1 & 1,6 \\ 1,6 & 4 \end{bmatrix}$.

From the population of the above specified distribution, a random sample characterized by the following aspect was drawn:

No	X_1	X_2	No	X_1	X_2
1	0,597489	0,081561	9	–1,53101	–1,99833
2	–0,89497	–0,85039	10	0,995763	0,806522
3	–0,07614	1,951707	11	–0,42281	0,392776
4	–0,12811	–0,37291	12	–0,01294	0,614983
5	–0,29082	–2,37729	13	1,756857	2,753824
6	–0,46499	0,034676	14	–0,66034	–1,59509
7	0,583153	1,655271	15	0,105686	2,114101
8	–0,59035	–3,86067			

The contour of point dispersion on the correlation diagram points to its eliptical shape with a positive inclination in the first and third quarters of co-ordinate system. The points arrange themselves around *(0, 0)*. The results for this sample are: vector of averages of (–0,0689, –0,04328) and covariance martix of

$\begin{bmatrix} 0,609772 & 0,95149 \\ 0,95149 & 3,18139 \end{bmatrix}$. The square root for this matrix is the matrix of $\begin{bmatrix} 0,78089 & 0 \\ 1,22829 & 1,29332 \end{bmatrix}$.

The determination of $N = 1000$ of two-dimensional bootstrap samples, each of $n = 15$, was carried out. The calculation was done according to the algorithm given in chapter 3. Recalculation for the one first bootstrap samples are given in table 1. The next columns show numbers R_1 and R_2 from the uniform distribution (stage 1), vectors (U_1, U_2) of the standard two-dimensional normal distribution determined with the Box-Muller transformation (stage 2) and the X_1 and X_2 values from the two-dimensional normal distribution with the given parameters of μ and Σ (stage 4). Vecotrs of average values of the given two bootstrap samples, vector statistics of T_n^* and corresponding to them values of the Mahalanobis depth measures are included in table 2 for the earlier specified two bootstrap samples. The values taken from the table were ordered non-decreasingliy according to the depth measures (tabl. 3). For the three levels of confidence of $1-\alpha = 0{,}95,\ 0{,}90,\ 0{,}85$, and 50, 100 and 150 respectively were rejected, up to the lowest values of depth measures. The other vectors were used to determine the $A_{n,1-\alpha}^*$ bootstrap confidence regions whose coordinates for the illustration at $\alpha = 0{,}05$ include columns c and d of table 4.

The diagrams of the bootstrap confidence regions with given levels of confidence are presented in figure 2.

Table 1

Values of one bootstrap samples

1						2					
R1	R2	U1	U2	X1	X2	R1	R2	U1	U2	X1	X2
0,9846	0,6079	-0,1372	-0,1104	-0,1372	-0,3520	0,0837	0,3991	-1,7948	1,3193	-1,7948	-1,2886
3						4					
R1	R2	U1	U2	X1	X2	R1	R2	U1	U2	X1	X2
0,2848	0,0309	1,5552	0,3056	1,5552	2,8552	0,7870	0,2665	-0,0717	0,6884	-0,0717	0,7114
5						6					
R1	R2	U1	U2	X1	X2	R1	R2	U1	U2	X1	X2
0,5594	0,8767	0,7701	-0,7542	0,7701	0,3271	0,9138	0,2353	0,0393	0,4229	0,0393	0,5703

Table 1 (cont.)

7						8					
R1	R2	U1	U2	X1	X2	R1	R2	U1	U2	X1	X2
0,7287	0,5182	-0,7903	-0,0909	-0,7903	-1,3735	0,6406	0,0904	0,7955	0,5078	0,7955	1,8822
9						10					
R1	R2	U1	U2	X1	X2	R1	R2	U1	U2	X1	X2
0,7413	0,9141	0,6638	-0,3977	0,6638	0,5848	0,1245	0,8168	0,8314	-1,8643	0,8314	-0,9069
11						12					
R1	R2	U1	U2	X1	X2	R1	R2	U1	U2	X1	X2
0,3356	0,1268	1,0328	1,0570	1,0328	2,9209	0,9209	0,9832	0,4036	-0,0427	0,4036	0,5946
13						14					
R1	R2	U1	U2	X1	X2	R1	R2	U1	U2	X1	X2
0,0452	0,9285	2,2418	-1,0809	2,2418	2,2898	0,0397	0,3631	-1,6573	1,9258	-1,6573	-0,3408
15											
R1	R2	U1	U2	X1	X2						
0,2712	0,1576	0,8863	1,3506	0,8863	3,0388						

Source: own calculations.

Table 2

Vectors $\hat{\theta}_n^*$, T_n^* and d_n^*, $z_M(d_n^*)$ values

No	$\hat{\theta}_1^*$	$\hat{\theta}_2^*$	T_1^*	T_2^*	d_n^*	$z_M(d_n^*)$
1	0,3179	0,76754	1,18043	0,69872	2,22408	0,31017
2	–0,3161	0,03373	–3,1836	0,63965	3,44269	0,22509

Source: own calculations.

Table 3

Values arranged according to depth measure

$\hat{\theta}_1^*$	$\hat{\theta}_2^*$	T_1^*	T_2^*	d_n^*	$z_M(d_n^*)$
0,69569	1,23837	7,68826	-0,944	16,2437	0,05799
–0,7656	–1,483	–5,5685	–2,0197	15,8519	0,05934

Source: own calculations.

Table 4

Vector coordinates of bootstrap confidence regions for $1-\alpha=0{,}95$

No	$\frac{1}{\sqrt{n}}S^{*}_{\hat{\theta}_n}\cdot\omega$		$\theta_n-\frac{1}{\sqrt{n}}S^{*}_{\hat{\theta}_n}\cdot\omega$	
	a	b	c	D
1	1,550122	2,12305	−1,619025	−2,16633
2	−1,12272	−2,44044	1,05382	2,397158

Source: own calculations.

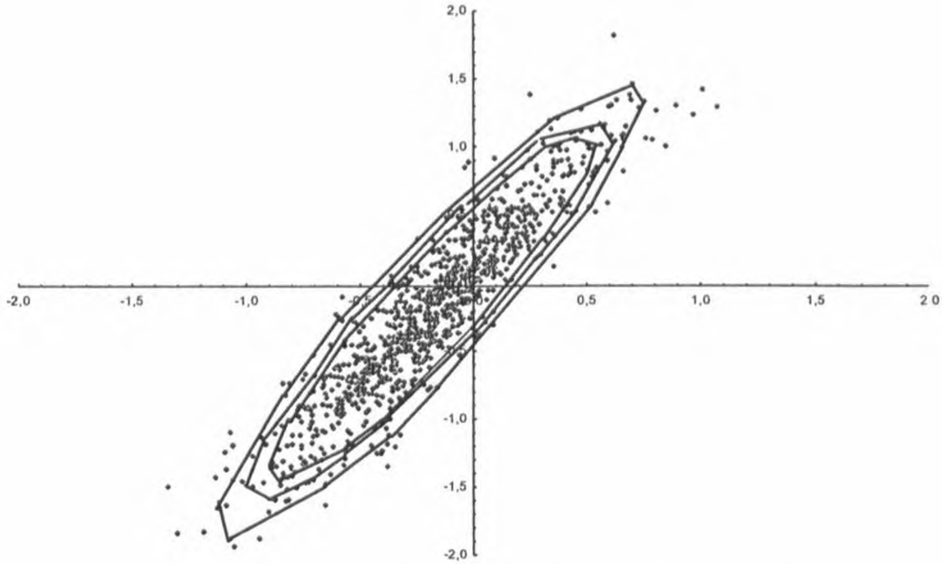

Fig. 2. Bootstrap confidence region at confidence levels of 0,95, 0,90 and 0,8

V. SUMMARY

In the above paper the method of the construction of the bootstrap confidence regions based on the Mahalanobis depth measure of observation in a sample was presented. By assigning the observation to their releveant depth measures, it is possible to order them in relation to the distance from the central cluster and to eliminate a specific number of observations to which the lowest values of depth measure are corresponding up to the moment when the assumed level of confidence is reached. The suggested method may be used for any any dimentions.

REFERENCES

Domański Cz., Pruska K., (2000), *Nieklasyczne metody statystyczne*, PWE, Warszawa.

Domański Cz., Pruska K., Wagner W., (1998), *Wnioskowanie statystyczne przy nieklasycznych założeniach* , Wyd. Uniwersytet Łódzki, Łódź.

Efron B., (1993), An Introduction to the Bootstrap, *Chapman and Hall*, New York.

Kobylińska M., Wagner W., (2000), Miary i kontury zanurzenia w opisie statystycznym próby dwuwymiarowej, *Wyzwania i dylematy statystyki XXI wieku,* AE.

Liu R., (1990), On a notation of data depth based on random simplices, *Ann. Statist.*, 18, 405–414.

Liu R.., Singh K., (1997): A quality index based on data depth and multivariate rank Tests, *J. Am. Statist. Ass.*, 88, 252–260.

Wieczorkowski R., Zieliński R., (1997), *Komputerowe generatory liczb losowych*, WNT, Warszawa.

Yeh B., Singh K., (1997), Balanced confidence regions baser on Tukey's depth and the bootstrap, *Journal Royal Statistical Society*, 59, 639–65.

Małgorzata Kobylińska, Wiesław Wagner,

BOOTSTRAPOWE OBSZARY UFNOŚCI OPARTE NA ZANURZANIU MAHALANOBISA DLA PRÓB DWUWYMIAROWYCH

W pracy przedstawiony został algorytm tworzenia obszarów bootstrapowych. Do konstrukcji tych obszarów wykorzystano miary zanurzania obserwacji w próbie. Konstrukcję zaprezentowano dla przypadku dwuwymiarowego.