ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 225, 2009

*Daniel Kosiorowski**

ABOUT ROBUST ESTIMATORS OF AVERAGE SHAPE AND VARIANCE OF SHAPE

Abstract

In this paper we investigate whether in the statistical inference for shapes of economic systems it is useful to modify least squares methods (commonly used Procrustes approach) by applying data depth concept. In the paper theoretical considerations are illustrated with examples of multidimensional financial time series.

Key words: Procrustes analysis, Procrustes mean, variance of shape estimator, projection depth, capital flow.

1. Introduction

A statistical analysis of shape could be a valuable approach both in a practice and in theory of economics. Namely notions of an average shape and a variance of shape could be adequate for an indirect verification of theoretical concept according to which capital stored in a certain system is described by the ability of this system to perturb a certain space of values, the flow of capital is connected with internal stresses in the substance of the capital carrier. The notion of average shape could correspond to stress operator and the notion of variance of shape could correspond to an amount of energy stored in the economic system.

According to D. G. Kendall the shape of an object is all the geometrical information that remains invariant when location, scale and rotational effects are filtered out from an object. Within statistical analysis of shape, the

* M. Sc., Department of Statistics, Cracow University of Economics.

objects of a considered population are studied on base of m, k-dimensional indicators (landmarks, markers), which are points placed on objects, corresponding with certain essential mathematical or content-related properties of these objects[1].

Relatively well known[2] and valuable for applications in economics is analysis of planar shapes. Considering two centered configurations (centered coordinates of landmarks):

$$y=(y_1,...,y_k)^T \text{ and } w=(w_1,...,w_k)^T \text{ both in } \mathbb{C}^k \text{ and } y^*\mathbf{1}_k=0=w^*\mathbf{1}_k,$$

it is convenient to use the following complex regression model allowing introducing of Prokrustes distances between shapes:

$$y=(a+ib)\mathbf{1}_k+\beta e^{i\theta}w+\varepsilon \tag{1}$$

where $a+ib$ translation, $\beta>0$ scale, $0\le\theta\le 2\pi$ angle of rotation, ε $k\times 1$ complex error vector.

Full Procrustes distance between complex configuration **w** and **y** is given as:

$$d_F(\mathbf{w},\mathbf{y})=\inf_{\beta,\theta,a,b}\left\|\frac{\mathbf{y}}{\|\mathbf{y}\|}-\frac{\mathbf{w}}{\|\mathbf{w}\|}\beta e^{i\theta}-a-ib\right\| \tag{2}$$

Let us consider a situation, where a random sample of a configuration $\mathbf{w}_1,...,\mathbf{w}_n$ is available from the point of view of a perturbational model:

$$\mathbf{w}_i=\gamma_i\mathbf{1}_k+\beta_i e^{i\theta_i}(\boldsymbol{\mu}+\boldsymbol{\varepsilon}_i),\ (i=1,...,n) \tag{3}$$

$\gamma_i\in\mathbb{C}$ translation vectors, $\beta_i\in\mathbb{R}_+$ scale parameters, $0\le\theta\le 2\pi$ angle of rotation, $\boldsymbol{\varepsilon}_i\in\mathbb{C}$ – independent errors of zero average, $\boldsymbol{\mu}$ average shape in the population.

We obtain the estimator of the **Procrustes average shape** $[\hat{\boldsymbol{\mu}}]$ by minimizing, relative to $\boldsymbol{\mu}$, the sum of the squares of full Prokrust distances from each $\mathbf{w}_i$ to unknown average having a unit quantity:

---

[1] Detailed introduction into the problem could be found in e.g. K o s i o r o w s k i (2004).

[2] Details of the problem can be found in the following inspiring work D r y d e n, M e r i d a, (1998).

$$[\hat{\boldsymbol{\mu}}] = \arg\inf_{\mu} \sum_{i=1}^{n} d_F^2(\mathbf{w}_i, \boldsymbol{\mu}) \tag{4}$$

Note that Procrustes mean is estimator of least squares method.

By estimating average *Prokrust's* shape we receive so called **Procrust's cordinates (Procrustes fit)** corresponding to the values which the estimated perturbational model (3) takes on. In the planar case, for $\mathbf{w}_1, ..., \mathbf{w}_n$ Prokrust's coordinates are given:

$$\mathbf{w}_i^P = \mathbf{w}_i^* \hat{\boldsymbol{\mu}} \mathbf{w}_i / (\mathbf{w}_i^* \mathbf{w}_i), \ (i = 1, ..., n) \tag{5}$$

To obtain a general measure of shape variability, it is convenient to use the root of the average square of the distance between each configuration and *Prokrust's* average $[\hat{\boldsymbol{\mu}}]$. We denote this measure as $RMS(d_F)$:

$$RMS(d_F) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} d_F^2(\mathbf{w}_i, \hat{\boldsymbol{\mu}})} \tag{6}$$

A limitations of statistical models commonly used within statistical analysis of shape are high restrictions concerning the assumptions for the examined phenomenon. Namely we mean a multivariate normality or even isotropy assumption for probability distributions generating configurations. The proposed estimators are not robust (e.g. Procrustes mean is least squares estimator and it is well known that least squares estimators have very low replacement breakdown point[3]).

That facts motivate author to propose robust modification of Procrustes analysis referring to data depth concept.

The notion of the depth of a point $\mathbf{x} \in \mathbb{R}^d$, $d > 1$, being a realization of some d-dimensional random vector $\mathbf{X}$ with the probability distribution P, is introduced basing on a special function called the **depth** or the **depth function**[4]. The function of depth assigns to every point a real positive number from interval $[0, 1]$ being the measure of that point's "centrality" in regard to the P distribution. Usually, a point, for which the function of depth assumes the maximum value is called ***d-dimensional median*** (an average, if there are more

[3] A rich overview of issues connected with that matter with references to original works can be found for instance in Rousseeuw, Leroy (1987).

[4] Details of the problem can be found in the following inspiring work Zuo, Serfling (2000).

such points than one). In case we do not know the form of the F distribution, but we have an n-element sample from $\mathbf{X}$, – $\mathbf{x}_1, ..., \mathbf{x}_n$ then we can replace the P distribution with its empirical version of $\hat{P}_n$.

As an example of depth let us consider **symmetric projection depth** (*PD*) that defines the outlyingness of a point $\mathbf{x}$ to be the worst case outlyingness of $\mathbf{x}$ with respect to one dimensional median in any one-dimensional projection that is:

$$PD(\mathbf{x};P)=\left(1+\sup_{\|u\|=1}\frac{\left|\mathbf{u}^T\mathbf{x}-Med(\mathbf{u}^T\mathbf{X})\right|}{MAD(\mathbf{u}^T\mathbf{X})}\right)^{-1} \tag{7}$$

where $\mathbf{X}$ has distribution P, Med denotes the univariate median, *MAD* denotes the univariate median absolute deviation $MAD(Y)=Med(|Y-Med(Y)|)$.

For further considerations it would be useful to introduce following notions: The set $\{\mathbf{x}\in\mathbb{R}^d : PD(\mathbf{x})=\alpha\}$ is called α projection level set or contour of projection depth α. The set $PD_\alpha(\mathbf{X})=\{\mathbf{x}\in\mathbb{R}^d : PD(\mathbf{x},P)\geq\alpha\}$ is referred to as the region enclosed by contour of projection depth α, α – **projection trimmed region** or α – central region.

The projection depth has very high Huber's replacement breakdown point, has also bounded Hempel's influence function on α trimmed region for certain α greater than zero. Under mild conditions on probability distributions in a population an empirical projection depth median is unbiased and strongly consistent estimator of a center of distribution for centrally symmetric distributions[5].

2. Propositions of robust procedures of analysis of shape

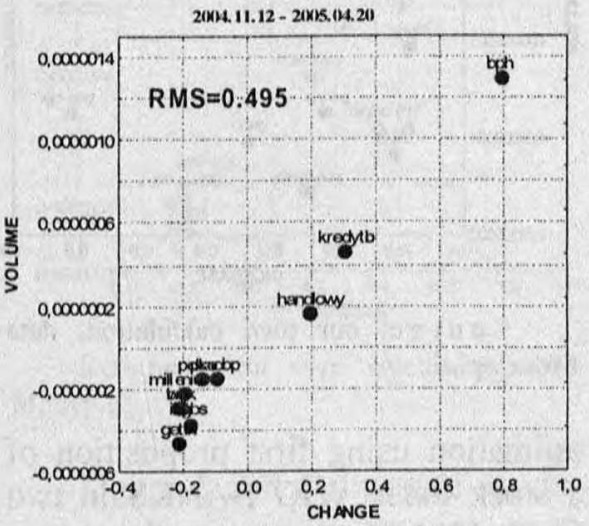
In order to give some arguments for the relation between the capital flow and the internal stresses in the substance of capital carrier we interpret an average shape of stock branch index as a stress operator which enables us to express an activity of an inner force resulting capital flow. We interpret $RMS(d_F)$ – the measure of variability of shape as an amount of that force. Namely during n stock sessions we analyze k stocks of certain stock index

[5] Interesting theoretical considerations relating to the issues can be found in e.g. Zuo (2004).

considered with respect to a daily increase/decrease (change %) of price X_i, $(i=1,...,n)$ and a daily volume Y_i, $(i=1,...,n)$.

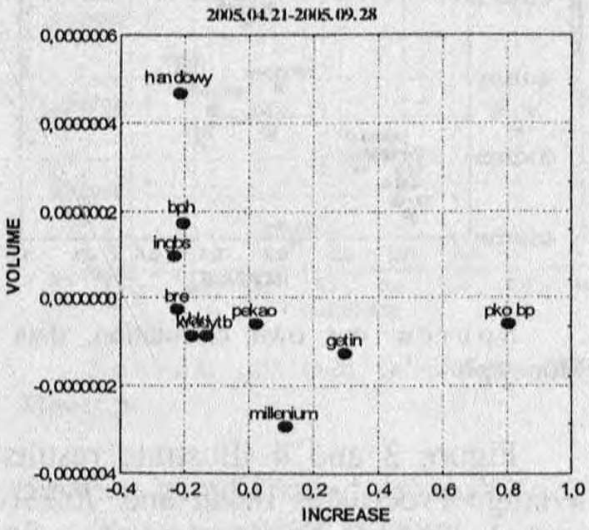
Figure 1 and 2 illustrate result of estimation of average Procrustes mean and $RMS(d_F)$ of stock index WIG BANKS in two periods: 2004.11.12–2005.04.20 and 2005.04.21–2005.09.28. The estimation was made on base of raw data.

Fig. 1. Icon of Procrustes mean shape – WIG BANKS – raw data – 2004.11.12–2005.04.20



Source: our own calculation, data Money.pl.

Fig. 2. Icon of Procrustes mean shape – WIG BANKS – raw data – 2005.04.21–2005.09.28



Source: our own calculation, data Money.pl.

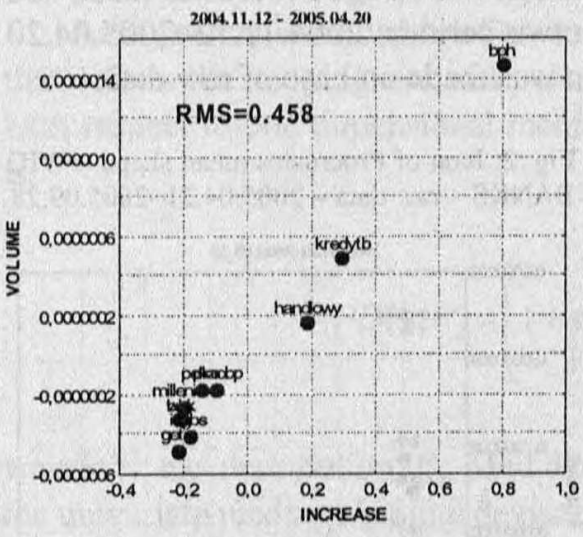
In case of financial markets both assumption of multinormality for coordinates of markers (stocks) and nonexistence of outlying observations are in general not fulfilled. That facts motivates us to following propositions:

First proposition – robust estimation of average shape and variance of shape

Let us consider $k\times 2$ dimensional time series consisting n-elements e.g. k stocks of certain stock index considered with respect to daily increase of price X_i $(i=1,...,n)$ and daily volume Y_i $(i=1,...,n)$. Let us assume $k\times 2 \ll n$.

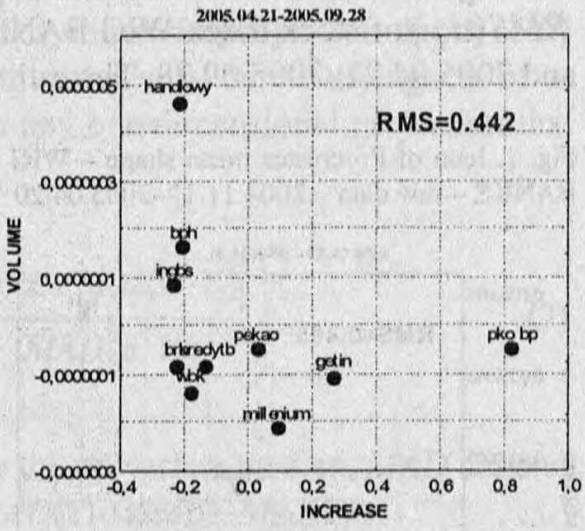
1. We treat observations as n-realizations of $k\times 2$ dimensional random vector $\mathbf{Z}_j=\left(X_{1j},Y_{1j},X_{2j},Y_{2j},...,X_{kj},Y_{kj}\right)$, $(j=1,...,n)$.
2. We calculate the empirical projection depth $PD(\mathbf{z}_j,\hat{F}_n)$ for each $\mathbf{z}_j=\left(x_{1j},y_{1j},x_{2j},y_{2j},...,x_{kj},y_{kj}\right)$, $(j=1,...,n)$.
3. We omit say 10% of the observations with minimal empirical projection depth values $PD(\mathbf{z}_j,\hat{F}_n)$.
4. For the rest of observations we do standard generalized Procrustes analysis.

Fig. 3. Icon of Procrustes mean shape – WIG BANKS – 10% PD trimming – 2004.11.12–2005.04.20



Source: our own calculation, data Money.pl.

Fig. 4. Icon of Procrustes mean shape – WIG BANKS –10% PD trimming – 2005.04.21–2005.09.28



Source: our own calculation, data Money.pl.

Figure 3 and 4 illustrate results of estimation using first proposition of average Procrustes mean and $RMS(d_F)$ of stock index WIG BANKS in two periods: 2004.11.12–2005.04.20 and 2005.04.21–2005.09.28.

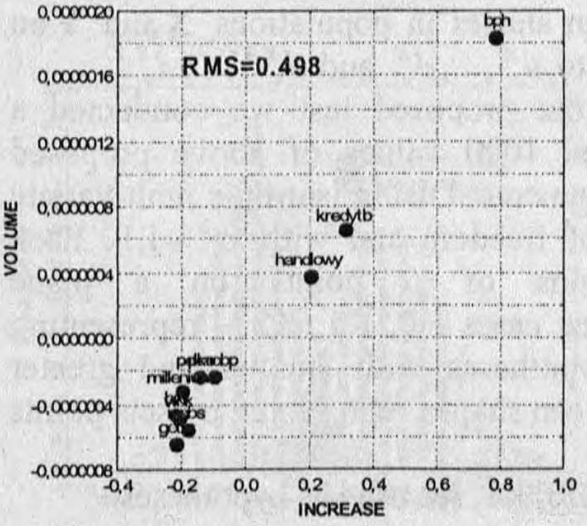
Second proposition – robust estimation of average shape and variance of shape

Assume that we have a sample $\mathbf{x}_1, ..., \mathbf{x}_n$ configuration matrices from population $\mathbf{X}$ e.g. k stocks of certain stock index considered with respect to daily increase of price is $k \times 2$ configuration matrix.

1. For sample $\mathbf{x}_1, ..., \mathbf{x}_n$ we calculate Procrustes mean.
2. We calculate approximate tangent coordinates $\mathbf{v}_1, ..., \mathbf{v}_n$ e.g. Procrustes residuals for sample $\mathbf{x}_1, ..., \mathbf{x}_n$ (the pole of tangent projection is Procrustes mean for that sample).
3. We calculate empirical projection depth $d_1, ..., d_n$ for tangent coordinates $\mathbf{v}_1, ..., \mathbf{v}_n$.
4. We omit say 10% observations $\mathbf{x}_1, ..., \mathbf{x}_n$ with minimal values of tangent coordinates depths $d_1, ..., d_n$.
5. For the rest of observations we do once again standard generalized Procrustes analysis – we calculate Procrustes mean.

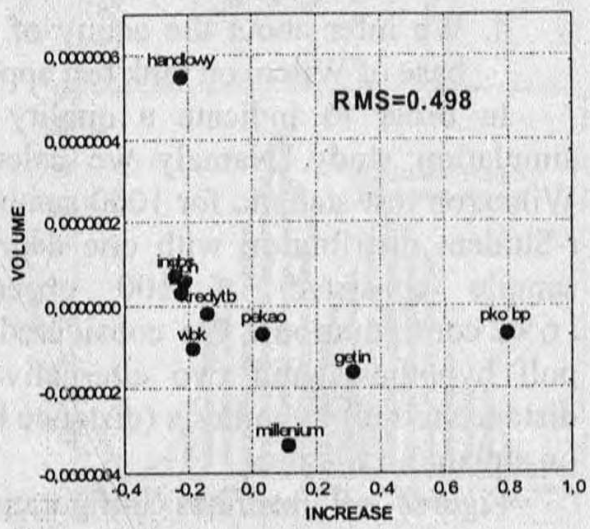
Remark: We could consider successive trimming – in step four we omit 5% observations next we go to step five and then we come back to step one etc. We continue the procedure until omitting appropriate fraction of observations. The remark is tied with discrimination between outlying observation and influential observation in LS method.

Fig. 5. Icon of Procrustes mean shape – WIG BANKS – 10% PD trimming of observations with outlying residuals – 2004.11.12–2005.04.20



Source: our own calculation, data Money.pl.

Fig. 6. Icon of Procrustes mean shape – WIG BANKS –10% PD trimming of observations with outlying residuals – 2005.04.21–2005.09.28



Source: our own calculation, data Money.pl.

Figure 5 and 6 illustrate results of estimation using second proposition of average Procrustes mean and $RMS(d_F)$ of stock index WIG BANKS in two periods: 2004.11.12–2005.04.20 and 2005.04.21–2005.09.28.

Third proposition – induced by depth function rank test equity of two average shapes

We have two samples $\mathbf{x}_1, ..., \mathbf{x}_n$ and $\mathbf{y}_1, ..., \mathbf{y}_n$ of configuration matrices from populations $\mathbf{X}$ and $\mathbf{Y}$. We are going to test following hypotheses (average shape in population $\mathbf{X}$ equals average shape in population $\mathbf{Y}$):

$$H_0 : [\boldsymbol{\mu}_X] = [\boldsymbol{\mu}_Y] \text{ vs. } H_0 : [\boldsymbol{\mu}_X] \neq [\boldsymbol{\mu}_Y]$$

1. We do two generalized Procrustes analyzes for the sample $\mathbf{x}_1, ..., \mathbf{x}_n$ and for the pooled sample $\mathbf{x}_1, ..., \mathbf{x}_n, \mathbf{y}_1, ..., \mathbf{y}_n$.
2. We calculate twice approximate tangent coordinates e.g. Procrustes residuals – for the sample $\mathbf{x}_1, ..., \mathbf{x}_n$ (the pole of tangent projection is Procrustes mean for that sample) and for the pooled sample $\mathbf{x}_1, ..., \mathbf{x}_n, \mathbf{y}_1, ..., \mathbf{y}_n$ $\mathbf{x}_1, ..., \mathbf{x}_n$ (the pole of tangent projection is Procrustes mean for that sample). Let us denote them correspondingly: $\mathbf{v}_1^X, ..., \mathbf{v}_n^X$ and $\mathbf{v}_1^{X,Y}, ..., \mathbf{v}_n^{X,Y}, ..., \mathbf{w}_1^{X,Y}, ..., \mathbf{w}_m^{X,Y}$.

3. We calculate the empirical projection depth for tangent coordinates $\mathbf{v}_1^X, ..., \mathbf{v}_n^X$ - let us denote them $d_{\mathbf{v}_1}^X, ..., d_{\mathbf{v}_n}^X$ and for tangent coordinates $\mathbf{v}_1^{X,Y}, ..., \mathbf{v}_n^{X,Y}, ..., \mathbf{w}_1^{X,Y}, ..., \mathbf{w}_m^{X,Y}$ – let us denote them $d_{\mathbf{v}_1}^{X,Y}, ..., d_{\mathbf{v}_n}^{X,Y}, d_{\mathbf{w}_1}^{X,Y}, ..., d_{\mathbf{w}_m}^{X,Y}$.
4. We infer about the equity of mean shapes in populations $\mathbf{X}$ and $\mathbf{Y}$ on base of Wilcoxon rank test applied to $d_{\mathbf{v}_1}^X, ..., d_{\mathbf{v}_n}^X$ and $d_{\mathbf{v}_1}^{X,Y}, ..., d_{\mathbf{v}_n}^{X,Y}$.

In order to indicate a quality of the proposed test we conducted a simulation study. Namely we calculated 1000 values of above proposed Wilcoxon test statistic for 1000 samples generated using isotropic multivariate t-Student distribution with one degree of freedom and with $\sigma^2 = 1{,}1$. Each sample consisted of 100 observations of 6 points on a plane (6×2 configuration). We considered three cases H0, K1, K2 – representing null hypothesis and two alternative hypotheses with smaller and greater distance to null hypothesis (distance between shapes represented by sets points on a plane).

Figures 7–9 illustrates configurations H0, K1, K2 used as hypotheses.

Fig. 7. Simulation study – null hypothesis

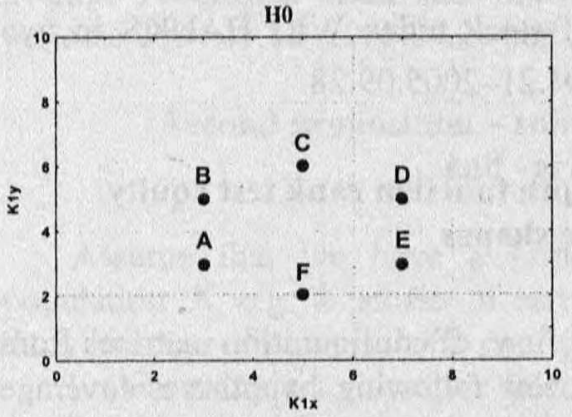


Fig. 8. Simulation study – alternative hypothesis

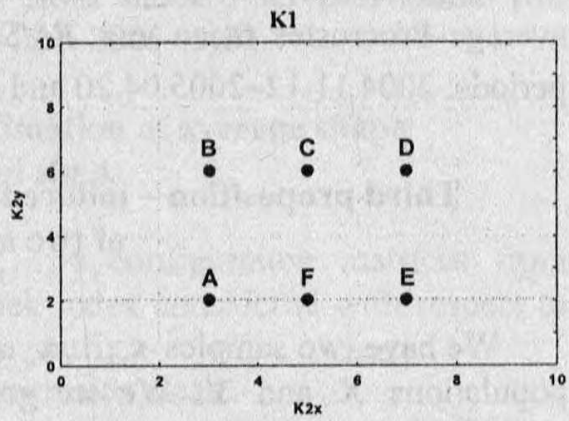


Fig. 9. Simulation study – alternative hypothesis

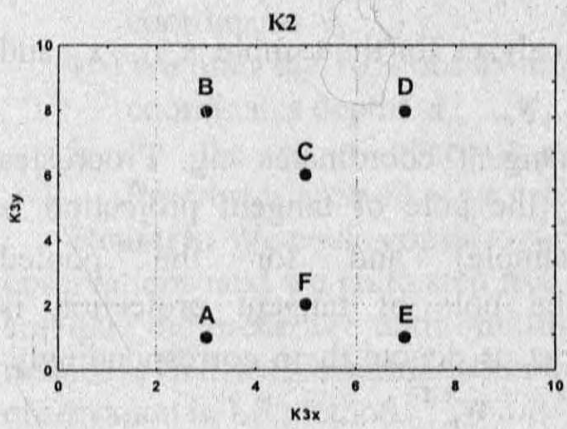
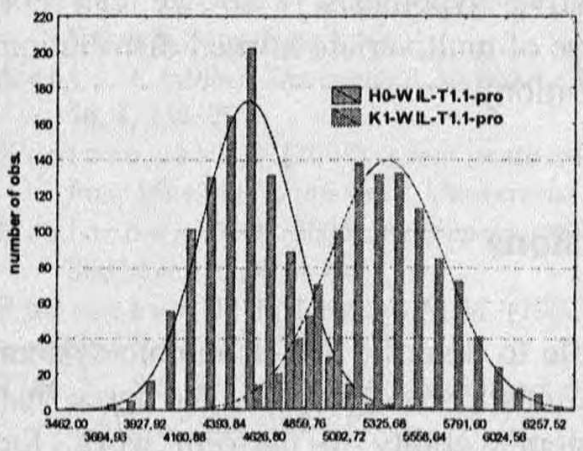
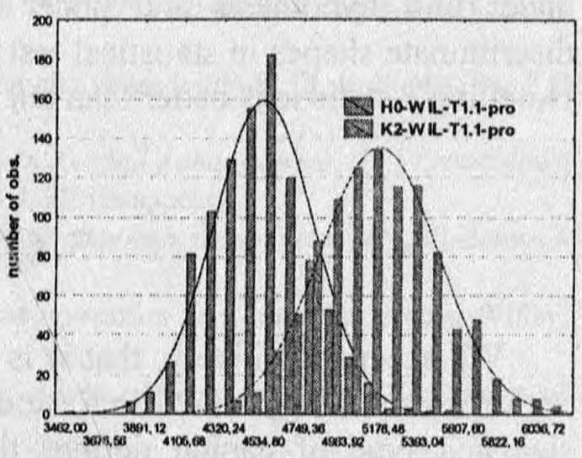


Fig. 10. Empirical density of proposed test statistic under H0 and K1 – 1000 simulation from isotropic multivariate t-Student distribution with one degree of freedom and $\sigma^2 = 1.1$



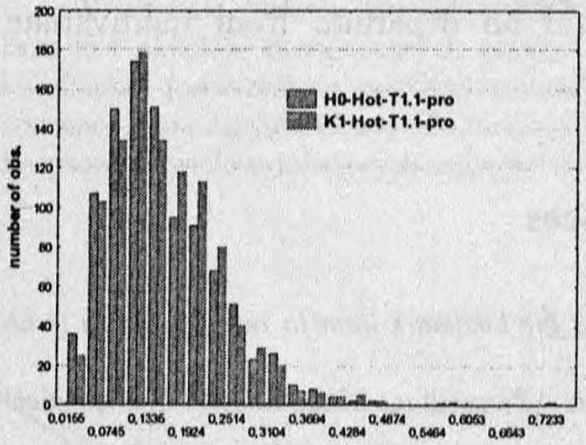
Source: our own calculation, data Money.pl.

Fig. 11. Empirical density of proposed test statistic under H0 and K2 – 1000 simulation from isotropic multivariate t-Student distribution with one degree of freedom and $\sigma^2 = 1.1$



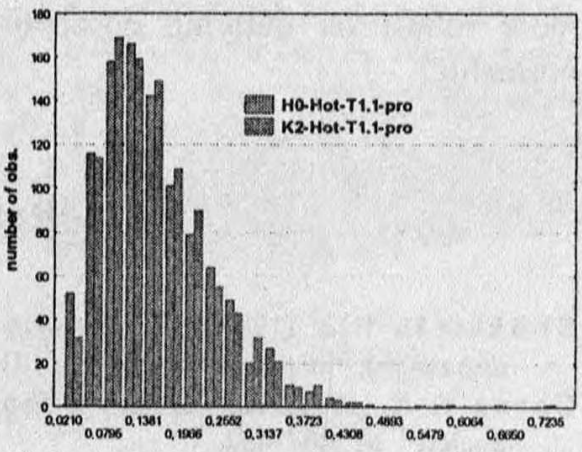
Source: our own calculation, data Money.pl.

Fig. 12. Empirical density of Hotelling T^2 test statistic under H0 and K1 – 1000 simulation from isotropic multivariate t-Student distribution with one degree of freedom and $\sigma^2 = 1.1$



Source: our own calculation, data Money.pl.

Fig. 13. Empirical density of Hotelling T^2 test sttistic under H0 and K2 – 1000 simulation from isotropic multivariate t-Student distribution with one degree of freedom and $\sigma^2 = 1.1$



Source: our own calculation, data Money.pl.

Figures 10–11 illustrates an empirical density of the proposed test statistic under H0, K1 and K2. Pictures 12–13 illustrates an empirical density of commonly used Hotelling T^2 statistic in case of generated in simulation samples.

It is easy to see that for considered alternatives proposed test is better that Hotelling T^2 test – empirical densities of Hotelling's statistic are nearly the same under null hypothesis and under alternative hypothesis – so we can not discriminate shapes in statistical test (in case of multivariate normal distribution Hotelling's statistic is better than our proposition).

3. Conclusions

We accept intuitively, that it is possible to describe any economic system with the use energy stored in a state of various kinds of capital. We agree that size and type of capital defines the system's ability to perform work, for example, the ability of an enterprise to expand to new markets. One could, however, ask: "can general intuition be useful for building stochastic models of any particular economic system?" The notions of statistical analysis of shape allows as to express activity of forces connected with internal stresses in substance of capital carrier – this could be a starting point for further studies.

Stochastic models existing within statistical analysis of shape[6] are too restrictive for economic applications. Proposed in the paper procedures are based on projection depth function and in general on data depth concept. That approach is often described as nonparametric and robust alternative to classical methods. Thanks to application of depth function proposed in the paper procedures seems to be more accurate for economic problems because they are more robust on outlying observations and on departure from multivariate normality.

References

Bookstein F. L. (1986), *Size and shape spaces for landmark data in two dimension (with discussion)*, "Statistical Science", **1**, 181–242.

Carne T. K. (1990), *The geometry of shape spaces*, "Proceedings of the London Mathematical Society", **61**, 407–432.

Dryden I. L., Merida K. V., (1998), *Statistical shape analysis*, John Wiley & Sons, New York.

[6] Various aspects of the statistical analysis of shape could be found e.g. in Bookstein (1986), Carne (1990), Goodall, Marida (1993), Kendall (1989).

Goodall C. R. Marida K. V. (1991), *A geometrical derivation of the shape densities*, "Advances in Applied Probability", **23**, 496–514.

Goodall C. R., Marida K. V. (1993), *Multivariate aspects of shape theory*, "Annals of Statistics", **21**, 848–866.

Kendall D. G. (1989), *A survey of the statistical theory of shape*, "Statistical Science", **4**, 87–120.

Kendall D. G., Barden D., Carne T. K., Le H. (1999), *Shape and shape theory*, John Wiley & Sons, New York.

Kent J. T. (1994), *The complex bingham distribution and shape analysis*, "J. R. Statist. Soc." B 56, **2**, 285–299.

Kosiorowski D. (2004), *About phase transitions in Kendall's shape space*, [in:] *Proceedings from MSA2004 Conference*, Uniwersytet Łódzki, Łódź (to appear).

Kosiorowski D. (2005), *Koncepcja głębi danych w badaniach ekonomicznych*, „Wiadomości Statystyczne", **8**, 1–14.

Rousseeuw P. J., Leroy A. M. (1987), *Robust regression and outlier detection*, Willey, New York.

Zuo Y., Serfling R. (2000), *General notions of statistical depth function*, "The Annals of Statistics", **28**, 461–482.

Zuo Y. (2004), *Robustness of Weighted L^p – Depth and L^p – Median*, AStA, **88**, 215–234.

Calculations have been made by means of Ian Dryden's *The shapes Package* made available under GNU license R project pages.

Daniel Kosiorowski

O odpornych estymatorach przeciętnego kształtu i wariancji kszstałtu

W artykule badamy czy w zagadnieniach wnioskowania statystycznego odnośnie do kształtów układów ekonomicznych użytecznym jest zmodyfikować estymatory najmniejszych kwadratów (powszechnie wykorzystywaną metodę Prokrusta) przez zastosowanie metod koncepcji głębi danych. W artykule rozważania teoretyczne ilustrujemy przykładami dotyczącymi finansowych wielowymiarowych szeregów czasowych.