Local Characterization of a Time Series Model Using Generalized Tukey Depth

Kosiorowski, Daniel
Cracow University of Economics, Department of Statistics
Ul. Rakowicka 27
Cracow 31-510, Poland
E-mail:daniel.kosiorowski@uek.krakow.pl

I. Introduction

From a practical view, the primary aims of any economic time series analysis would be to provide an insight into the short term probabilistic features of the possible underlying model on base of constantly updated sample path of a moderate length. On base of such a generally imprecise knowledge are made various economic decisions or predictions. We mean here for example evaluation of a portfolio of securities, options for sale of purchased shares management.

In practice main tools for time series analysis and model identification are mean, autocovariance, partial autocovariance or cross-correlation functions. They are extremely sensible to various kinds of outliers that may occur in time series. Their estimates critically depends on stationarity, ergodicity of the underlying model. Several authors stress that observed time series almost always consist of atypical observations (see (Pena (1990) or Maronna et all (2006)). These atypical points can be produced by nonsystematic changes in the variables that are driving the series or affecting them. Since the forecast from any time series model are based on the extrapolation of the historical patterns, if the parameters of the model are very dependent on a few atypical observations resulting from isolated or nonrepeatable events, then the quality of the forecasts can be expected to be poor. Moreover, when these parameters have or economic interpretations, the presence of undetected influential observations can lead the economist to wrong decisions.

In this paper we study certain properties of the generalized Tukey depth (location, location-scale, regression depths) induced procedures and look into the probabilistic information of the underlying time series model carried by the procedures. We focus our attention on short term multivariate quantile based description of the possible time series model. We give several examples of easy and user friendly depth induced statistical procedures for robust short term economic decision making.

II. Data Depth Procedures

Statistical depth functions originated with the notion of halfspace depth which has became much studied as a tool in nonparametric multivariate location inference. Tukey and Donoho and Gasko (see Zuo and Serfling (2000)) defined the halfspace depth of a point $\mathbf{x} \in \mathbb{R}^d$ with respect to an empirical distribution P_n on $\mathbb{R}^d$ based on data $\{\mathbf{y}_1,...,\mathbf{y}_n\}$ as the smallest proportion of data points in any closed halfspace with $\mathbf{x}$ on the boundary. In detail let $\mathbf{u}$ be a vector on unit sphere S^{d-1} of $\mathbb{R}^d$ then the Tukey depth of a point $\mathbf{x}$ can be written as

$$(1) \qquad d(\mathbf{x},P_n)=\min_{\mathbf{u}\in S^{d-1}}\# \;\; i:\mathbf{u}^T\mathbf{y}_i\geq\mathbf{u}^T\mathbf{x} \;\; \Big/ n=\min_{\mathbf{u}\in S^{d-1}}\# \;\; i:(\mathbf{y}_i-\mathbf{x})\in H_u \;\; \Big/ n \quad ,$$

where P_n is the empirical distribution based on data $\{\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_n\}$, $\#\{\cdot\}$ denotes the number of data points in $\{\cdot\}$, and $H_{\mathbf{u}}=\{\mathbf{x}:\mathbf{u}^T\mathbf{x}\geq 0\}$ is the closed halfspace containing 0 on its boundary with $\mathbf{u}$ pointing inside the halfspace and orthogonal to the boundary.

The Tukey depth is independent of the coordinate system, that is it is affine invariant. The point(s) with the maximum Tukey depth provides a measure of centrality known as Tukey median. For $p\in(0,1)$, the p-th Tukey depth contour $D(p)$ is the collection of $\mathbf{x}\in\mathbb{R}^d$ such that $d(\mathbf{x})\geq p$; it means $D(p)=\;\; \mathbf{x}\in\mathbb{R}^d : d(\mathbf{x})\geq p \;\;$. Contours (some authors use term central regions) form a sequence of nested convex sets (for details see Rousseeuw and Struyf (1999) or Zuo and Serfling (2000)). One useful application of the contours is to provide a nonparametric description of the dispersion of distribution using the volumes of the enclosed regions. An example concerning a relation between inflation and unemployment rate in Poland is presented on Figure 1 and Figure 2. Struyf and Rousseeuw (1999) proved that the Tukey depth completely determines empirical distributions by actually reconstructing the data points from the depth contours. Also Kong and Zuo (2010) studied properties of the Tukey depth contours and looked into the probabilistic interpretation carried by the contours and show that any distribution with smooth depth contours is completely described by its Tukey depth.

Innovative extension of the Tukey depth to univariate multiple regression was proposed by Rousseeuw and Hubert (1999). Mizera (2002) encompases notion of halfspace depth and regression depth within a general framework “tangent depth” defined with respect to “gradient probability fields” and equipped with differential calculus. His definition of the depth in general models is motivated by theoretical considerations with a decision – theoretic flavor. General halfspace depth can be defined as a measure of data – analytic admissibility – the simplest version of this principle defines depth of a parameter θ as the proportion of the data points whose omission causes θ to become a nonfit, a fit that can be uniformly dominated by another one. Mizera and Muller (2004) apply the “tangent depth” to the classical univariate location–scale problem through a “location – scale” depth defined on a bivariate parameter space. Mizera and Muller introduced not one but a family of depths depending on the choice of the underlying density f. In a context of robust short-term analysis of relations between the mean and the dispersion of the economic time series their Student – depth seems to be especially interesting.

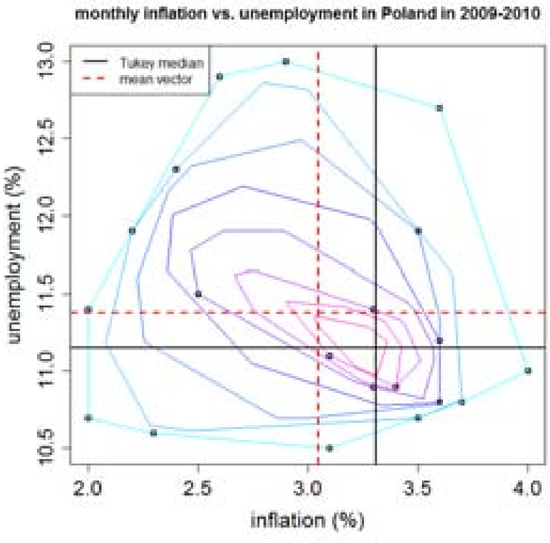
The Student depth of $(\mu,\sigma)\in\mathbb{R}\times[0,\infty)$ with respect to a probability measure P on $\mathbb{R}$ is defined

$$(2) \qquad d(\mu,\sigma,P)=\inf_{(u_1,u_2)^T\neq 0} P \;\; y:u_1(y-\mu)+u_2((y-\mu)^2-\sigma^2)\geq 0 \;\; ,$$

the Student depth with respect to the data $y_1,...,y_n$ is obtained by applying the definition to the empirical probability measure P_n supported by the data points.

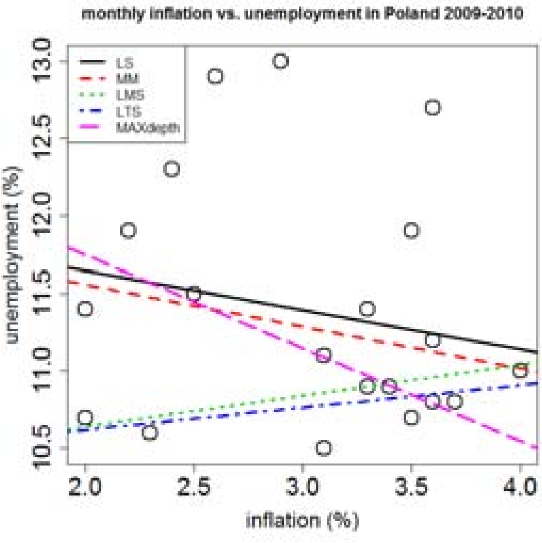
The location μ of the Student median lies relatively close to the sample median – in particular for data exhibiting symmetry. For asymmetric unimodal distributions, we may observe that the Student median location μ shrinks from the sample median toward the mode. We observed also that the student median scale σ is usually shrunk down from the MAD. Results presented in Mizera (2002) imply for the Student depth that breakdown point of the Student median is not less than $\lceil n/3 \rceil$. This means considerable robustness. We can say that the Student depth plots indicate asymmetry including that present in the core of the data rather than just in the tails, but they are capable of detecting heavy–tailed behavior too.

Fig. 1: Tukey depth contour plot – inflation vs. unemployment in Poland.



Source*: Own calculations, data GUS*

Fig. 2: Linear regression fits – inflation vs. unemployment in Poland



Source*: Own calculations, data GUS*

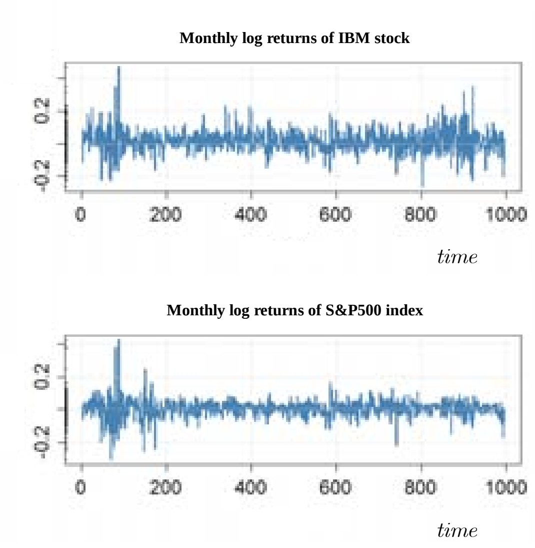
III. Propositions

Depth–based statistical methods are providing short term multivariate quantile based description of the possible time series model. Although such the description is rather imprecise but very often gives us a base for a decision making. Data depth concept offers a variety of easy and user – friendly analytic tools for a preliminary analysis of time series and economic decision making. We mean here in particular:

A. We can use moving multidimensional median as an alternative to one-dimensional moving mean or median filter. In a contrary to the mean and the median, the multidimensional median takes into account multidimensional geometry of the data and hence the natural dependence of points in time series analysis. We advocate here using a moving projection median which has very good properties in the context of a balance between robustness and efficiency (for details see Zuo (2003)). In case of linear autoregression estimation we strongly recommend using maximum regression depth estimator (Maxdepth) instead of least squares, maximum likelihood based methods. We underline here relatively high breakdown (BP) point of Maxdepth estimator but also relatively small sensitivity of the maximum depth estimator for a data subset – for a majority of the data (see Visek (2002)). Using the autoregression estimator with relatively high BP we protect our analysis against an effect of propagation of an outlier.

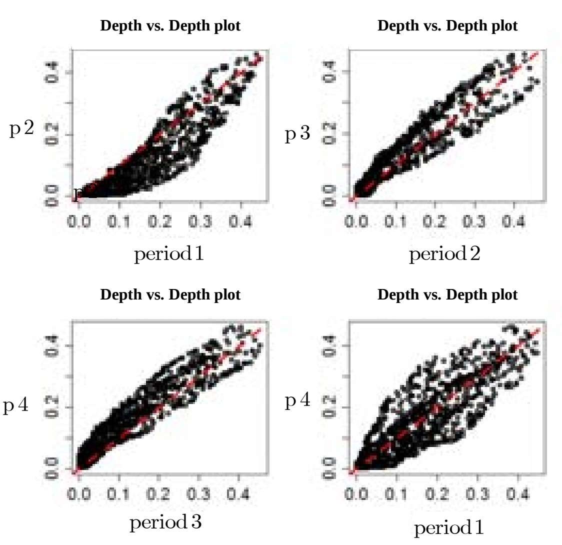
B. We can use simple Depth vs. Depth plot (see Liu et all (1999)) for a preliminary analysis of stationarity of multidimensional time series. We calculate sample depths of points assuming say first 25% and last 25% points of the considered time series. Next we compare calculated depths for each point using scatter plot. Departures from diagonal line of the scatter plot should inform us about differences of the probability distribution generating time series. Figure 3 presents four depth vs. depth plots prepared on base of two-dimensional time series of the monthly log returns of IBM stock and the S&P 500 index from January 1926 to December 1999 with 888 observations (see Tsay (2010)). We divided series into four approximately equal size parts following each another. We can notice a significant differences between first and second and first and fourth period.

Fig. 3: Tukey depth contour plot – inflation vs. unemployment in Poland.



Source: *Own calculations, data Tsay (2010)*

Fig. 4: Tukey depth contour plot – inflation vs. unemployment in Poland.

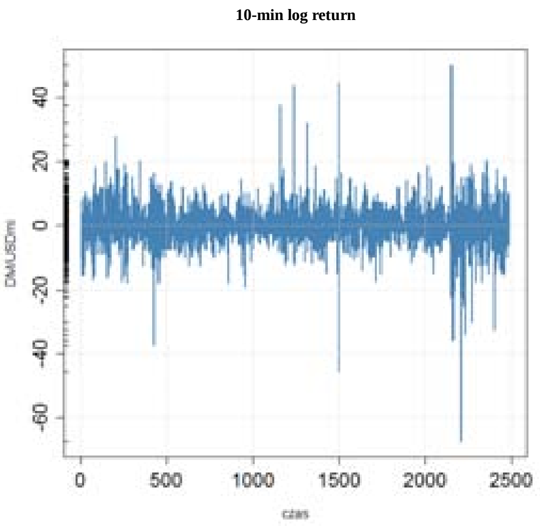


Source: *Own calculations, data Tsay (2010)*

C. In order to indicate a model generating time series it is useful to analyze a behavior of a moving Student median or the Student median calculated for short following each another periods. Scatter diagrams of the location and the scale coordinates of the Student medians could be very helpful tools for an investigation of relations between the mean and the dispersion of the underlying process generating series. We recommend this tool for a preliminary discrimination between simple GARCH, SV and ARMA models in cases of the samples of a short or moderate length consisting outliers. Figure 5 presents 10-minute FX log returns of Mark-Dollars exchange rate. Figure 6 presents the locations and the scales for Student medians calculated on base of 6-hours periods following one another subtracted from the original series (30 observations for each median calculation). Figure 7 presents scatter diagram of the Student median scale in the period t versus the Student median scale in the preceding period (t-1) with maximal regression fit represented by red line. Figure 8 presents scatter diagram of the Student median location in the period t versus Student median scale in the preceding period (t-1) with maximal regression fit represented by red line. Figure 7 and Figure 8 together focus our further attention on MGARCH (GARCH in mean) class of models generating the considered time series.

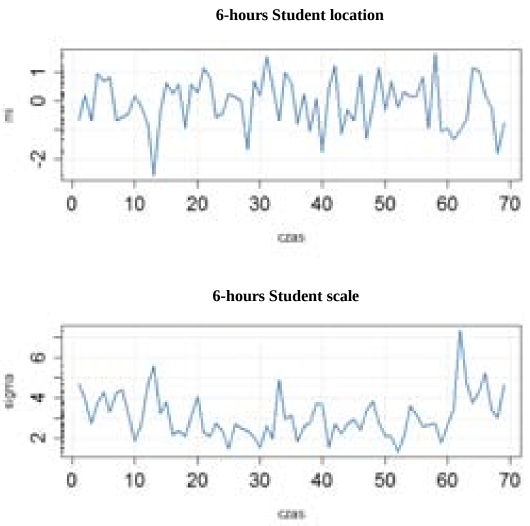
D. In a context of the autocorrelation coefficient estimation we recommend using slope calculated by means of maximal regression depth method adjusted by means of median of absolute deviation from the median. For autocorrelation of a moderate order we avoid an effect of propagation of an outlier.

Fig. 5: DM/USD 10-min log returns.



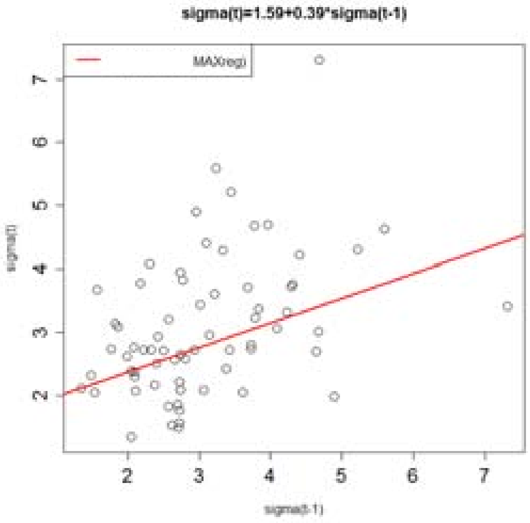
Source: *Own calculations, data Tsay (2010)*

Fig. 6: 6 – hours moving Student median.



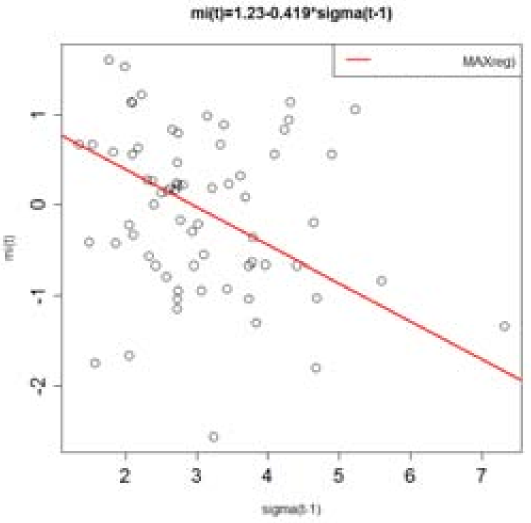
Source: *Own calculations, data Tsay (2010)*

Fig. 7: Student median scale in the period t vs. Student median scale in the period (t-1).



Source: *Own calculations, data Tsay (2010)*

Fig. 8: Student median location in the period t vs. Student median scale in the period (t-1).



Source: *Own calculations, data Tsay (2010)*

IV. Conclusions

In the short term forecasting a behavior of an economic system the goal is to predict future values of a time series based on the data collected to the present. Very often time series contain influential outliers misleading the economist about the properties of the considered process. In these situations data depth based exploratory techniques could provide us sufficient basis for the decision making. The Location – scale depth proposed by Mizera & Muller (2003) seems to be especially worth further studies in the context of a robust time series analysis.

REFERENCES

Kong L., Zuo Y. (2010), Smooth Depth Contours Characterize the Underlying Distribution, Journal of Multivariate Analysis 101, 2222–2226

Kosiorowski D. (2010). Depth Based Procedures for Estimation ARMA and GARCH Models, Y. Lechevallier, G. Saporta (Red.) Proceedings of COMPSTAT'2010, 1207 – 1214, Physica - Verlag.

Liu, R. Y., Parelius, J. M., Singh, K. (1999) Multivariate Analysis by Data Depth: Descriptive statistics, Graphics and Inference (with discussion). The Annals of Statistics 27, 783 – 858

Maronna, R. A., Martin, R. D., Yohai, V. J. (2006). Robust Statistics - Theory and Methods. Chichester: John Wiley & Sons Ltd.

Mizera, I. (2002). On Depth and Depth Poins: a Calculus. The Annals of Statistics (30), 1681 - 1736.

Mizera, I., Muller, C. H. (2004). Location – Scale Depth (with Discussion and Rejoinder). Journal of the American Statistical Association 99(4), 981 – 989

Pena, D. (1990). Influential Observations in Time Series, Journal of Business & Economic Statistics 8(2), 235 - 241

Rousseeuw, J. P., Hubert, M. (1998). Regression Depth. Journal of The American Statistical Association (94), 388 – 433

Tsay R. S. (2010), Analysis of Financial Time Series, Wiley – Interscience, Hoboken, New – Yersey

Tukey, J. (1975). Mathematics and Picturing Data. R. James (Red.), Proceedings of the 1974 International Congress of Mathematicians. 2, strony 523–531. Canadian Math. Congress.

Struyf and Rousseeuw (1999), Halfspace Depth and Regression Depth Characterize the Empirical Distribution, Journal of Multivariate Analysis 69, 135153

Visek, J. A. (2002) Sensitivity Analysis of M – estimates of Nonlinear Regression Model: Influence of Data Subsets. The Annals of the Institute of Statistical Mathematics 54(2), 261 – 290

Zuo, Y. (2003). Projection-based Depth Functions and Associated Medians, Annals of Statistics 31, 1460 – 1490

Zuo, Y., Serfling, R. (2000). General Notions of Statistical Depth Function. The Annals of Statistics (28), 461 - 482.