STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 25—45

CLUSTERING FINANCIAL DATA USING COPULA-GARCH MODEL IN AN APPLICATION FOR MAIN MARKET STOCK RETURNS

Anna Czapkiewicz[1], Beata Basiura[2]

ABSTRACT

There are many statistical techniques that allow us to find similarities among variables. Cluster analysis discovers structure within sets of data. The choice of a relevant metric is a fundamental problem in the case of clustering financial data. In this paper, the Copula–GARCH model is used to obtain the dependency parameter between time series. The dissimilarity measure based on the maximum likelihood parameter obtained from the Normal or t-Student copula is proposed and applied to classify forty two indices from American, European, and Asian stock markets.

Key words: Clustering stock indices, dependence parameter, Copula–GARCH model, Copula function, Skewed distribution

1. Introduction

The identification of similarities or dissimilarities in financial time series has become an important research area in finance and empirical economics. This problem has been widely studied in the discrimination and clustering literature. It is important to classify time series into groups with similar dependency structures. The choice of a relevant metric is a fundamental problem in clustering time series. Some studies use approaches based on the Euclidean distance. This metric has important limitations of being invariant to transformations that modify the order of observations over time and it does not allow for the correlation structure of time series. Financial time series are often characterized by volatility structures that might be represented by GARCH processes. Piccolo (1990) introduced a metric for ARIMA models based on the autoregressive representation. In this approach, time series are close when the models characterizing them are similar.

[1] Faculty of Management, AGH University of Science and Technology, Krakow, Poland, e-mail: gzrembie@cyf-kr.edu.pl.
[2] Faculty of Management, AGH University of Science and Technology, Krakow, Poland, e-mail: bbasiura@zarz.agh.edu.pl.

The distances between GARCH processed were discussed by Otranto (2004). A closeness between such processes is affirmed when the underlying conditional variance equations are similar in terms of their structure. However, this measure does not take into account the dependence structure between time series.

In this paper, we consider the problem of clustering financial time series. We propose a similarity measure based on the dependency structure between time series. To specify the behaviour of financial data, the GARCH(1,1) model is recommended (Bollerslev, 1986) and widely used. It is well known that for this class of models, the obtained residuals are generally non-normal. The evidence of heavy tails and skewness is also provided. That is why the Pearson correlation coefficient, which describes the dependency only for multivariate normal distribution, is not a good measure of dependence between financial time series. Without the multivariate normality assumption, two pairs of markets can have equal linear correlation coefficient while they can still differ in terms of dependence structure.

Embrechts et al. (2001, 2003) proposed an application of copula function to model the multivariate distribution. Using a copula approach, data modelling can be split into two steps: modelling marginal distributions, and then modelling a dependence structure between marginal distributions. Thus, in the first step we can model the financial data using a GARCH process. Next, the dependency parameter can be obtained from the copula taken as a function connecting the univariate marginal distributions and the multivariate distribution.

A variety of copula functions, already described in the literature (Nelsen, 1999), provides us with a wide range of models of dependence structure, but only some of them are appropriate for financial markets. The most popular are Normal copula and *t*-Student copula. Both Normal and *t*-Student copula functions have different characteristics in terms of tail dependence. The *t*-Student copula is recommended by several authors such as Mashal, Zevi (2002) and Breymann et al. (2003).

There are two important issues arising in describing empirical data using a copula function. Firstly, copula parameters must be accurately estimated. Secondly, the chosen copula specification should be tested. In this context, different statistical tests could be applied. Some, for example Junker and May (2005), used a chi-square test. Breymann et al. (2003) suggested the test based on the probability transform and on projection of the multidimensional problem into its univariate representation, where Anderson-Darling test statistic could be applied. Some other tests have also been discussed in the Chen et al. (2004), Genest et al. (2008)

The aim of this paper is to classify forty two indices coming from different emerging as well as developed, European, American, and Asian stock markets. The empirical data covers period from January 2002 to December 2008. In order to obtain correctly specified marginal distribution, the GARCH models with different conditional distribution of innovations are tested. Next, the dependency

structure between the GARCH processes is estimated. For relatively short time series, we assume that such a dependency is constant over the time.

The elements of the correlation matrix obtained from Normal or *t*-Student copula is proposed as the similarity measure between data. Estimating multidimensional copula would be a challenging task considering the number of data. Therefore, the bivariate copula parameters for all pairs of markets are estimated using the Maximum Likelihood (ML) method. We verify if the two-dimensional Normal or *t*-Student copula is able to capture the dependency structure between two markets.

For a clustering purposes we adapted the Ward (1963) clustering algorithm (see Mirkin (2005)) with the distance based on the estimator of the correlation parameter obtained from the Normal or *t*-Student copula.

The paper is organized as follows: In Section 2, the discussion of the univariate model is presented. Section 3 contains the copula function characteristics, description of the copulas used in the empirical studies, and the tests for models of dependence. The clustering method and the empirical results are presented in Section 4. The main results are summarized in the last section.

2. The return distribution

The advantage of using copulas, as mentioned in the introduction, stems from the fact that marginal distributions can be separated from the underlying dependency structure. Many univariate models have been proposed to describe the dynamics of returns. In this paper we investigate the univariate generalized autoregressive conditional heteroscedastic GARCH(1,1) model proposed by Bollerlav (1986). The GARCH(1,1) model is defined as follows:

$$y_t = \mu + \varepsilon_t, \qquad \varepsilon_t = \sqrt{h_t}\eta_t, \quad \eta_t \sim iid(0,1), \quad h_t = a_0 + a_1\varepsilon_{t-1}^2 + a_2 h_{t-1} \qquad (1)$$

This is the simplest and so far the most popular GARCH parameterization. With the restrictions $a_0 > 0$, $a_1, a_2 > 0$, and $a_1 + a_2 < 1$, the conditional variance process is stationary with finite second moments. In the classical GARCH model, the obtained residuals are assumed to be normally distributed. However, it is well known that in practice such residuals are far from normality. Scrutiny of daily returns led to the introduction of fat-tailed distributions for this residuals. For instance, Nelson (1991) considered the GED (generalized error distribution), while Bollerslev and Wooldridge (1992) focused on Student's *t* distributed innovations. Fat-tails are not the only problem in the context of conditional distribution, the skewness can also be noticed. With the skewness parameter being introduced, we can consider skewed *t*-Student or skewed GED distribution of the η_t (Arellano-Valle et. al. 2004).

Generally, the considered skewed distribution has the form:

$$f_{SKEW}(x;\psi,\nu)=\frac{2}{a(\psi)+b(\psi)}\left[g\left(\frac{x}{a(\psi)}\right)I_{(x<0)}+g\left(\frac{x}{b(\psi)}\right)I_{(x>0)}\right],$$

where ψ and ν denote, respectively, the asymmetry and the degrees-of-freedom parameters. The $a(\psi)$, $b(\psi)$ are some normalizing functions. The most known versions are $a(\xi)=\xi$ and $b(\xi)=\xi^{-1}$ (Fernandez, Steel, 1998) or $a(\lambda)=1-\lambda$ and $b(\lambda)=1+\lambda$ (Hansen, 1994). In our empirical research, we use the first weight. For several selected indices we used both weights, but the results were similar. The $g(\cdot)$ denotes symmetrical function distribution, it is *t*-Student or GED here. The density of Student's *t* distribution has the form:

$$f_S(x,\nu)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\nu-2)\pi}}\left(1+\frac{x^2}{(\nu-2)}\right)^{\frac{\nu+1}{2}},$$

whereas the density of the GED distribution is:

$$f_G(x,\nu)=\frac{\nu\exp\left\{-\frac{1}{2}\left|\frac{x}{\lambda}\right|^{\eta}\right\}}{\lambda\Gamma\left(\frac{1}{\nu}\right)}2^{\frac{\nu+1}{\nu}},\ \text{where}\ \lambda=\left[\frac{\Gamma\left(\frac{1}{\nu}\right)}{\Gamma\left(\frac{3}{\nu}\right)}2^{-\frac{2}{\nu}}\right]^{1/2}.$$

The maximum likelihood method is a general approach to estimate the parameters of GARCH model. Alternatively, a nonparametric approach can be followed by computing the empirical *cdf* of the data. We focus on the parametrical methods due to using chi-square test to verify the adapted model. Several restriction of the model specification are tested in the empirical section. In particular, we test within the class of GARCH models with different conditional distributions: Normal, *t*-Student, GED, skewed *t*-Student and skewed GED.

3. The Copula functions

The copula function is a multivariate distribution defined on the unit cube $[0,1]^d$, with uniformly distributed margins. The function $C:[0,1]^d\to[0,1]$ is a *d*-dimensional copula if it the satisfies the following properties:

1. For all $u_i\in[0,1]$, $C(1,..,1,u_i,1,..,1)=u_i$.
2. For all $u\in[0,1]^d, C(u_1,...,u_d)=0$, if at least one coordinate $u_i=0$.
3. C is *d*-increasing.

The importance of the copula function stems from the fact that it captures the dependence structure of the multivariate distribution. Let $X = (X_1,...,X_d) \in R^d$ be a random vector with a multivariate distribution $F(x_1,...,x_d) = P(X_1 \le x_1,...,X_d \le x_d)$, and continuous margins $F_n(x_n) = P(X_n \le x_n)$, $x_n \in R$, $n = 1,...,d$.

According to Sklar's theorem (Sklar, 1959), there exists a unique copula $C:[0,1]^d \to [0,1]$ of F such that for all $x \in R^d$,

$$F(x_1,...,x_d) = P(X_1 \le x_1,...,X_d \le x_d) = C\big(F_1(x_1),...,F_d(x_d)\big).$$

A fundamental conclusion of this theorem states that in multivariate continuous distribution functions, the dependence structure and the margins can be separated, and the dependence structure can be represented by the copula. This property allows us to fit the marginal distribution of each series of data separately, and to model the dependence structure between different data. Furthermore, the copulas are invariant under strictly increasing transformations of the variables. This property implies that the dependence structure between the variables is captured by the copula, regardless of the margins.

The most important copula in finance is the Normal copula:

$$C(u_1,...,u_d) = \Phi_\Sigma(\Phi^{-1}(u_1),...,\Phi^{-1}(u_d)),$$

where Φ is a cumulative normal distribution function and, Φ_Σ is a d-dimensional cumulative normal distribution function with correlation matrix Σ. The bivariate Normal copula is defined by:

$$C(u_1,u_2;\rho) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{s^2 - 2\rho st + t^2}{2(1-\rho^2)}\right) dsdt, \quad (2)$$

and ρ denotes the correlation coefficient.

Another important copula in finance is the t-Student copula:

$$C\ (u_1,...,u_d) = t_{\Sigma,\eta}(t_\eta^{-1}(u_1),...,t_\eta^{-1}(u_d)),$$

where t_η is the Student's t cumulative distribution with η degrees of freedom and $t_{\Sigma,\eta}$ is the Student's t cumulative distribution with η degrees of freedom and the correlation matrix Σ. The bivariate case of t-Student is given by:

$$C(u_1,u_2;\rho) = \int_{-\infty}^{t_\eta^{-1}(u_1)} \int_{-\infty}^{t_\eta^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \left(1 + \frac{s^2 - 2\rho st + t^2}{\eta(1-\rho^2)}\right)^{-\frac{\eta+2}{2}} dsdt, \quad (3)$$

where ρ is the correlation ratio.

In our empirical work, we focus on the copulas defined above. Both copulas have different characteristics in terms of tail dependence. The Normal copula has no tail dependence. The *t*-Student has symmetric tail dependence: large joint positive realizations have the same probability of occurrence as large negative realizations. Such a difference has important consequences to the modelling of the dependency parameter. In the next research we take up this subject.

3.1. Estimation of copula model

The theory of copulas shows that any joint distribution can be decomposed into its univariate marginal distributions and a copula function, which describes the dependency between the variables. Let $X = (X_1, X_2, ..., X_d)$ be random variables of interest, where X_i has parametric cumulative distribution function $F_i(x_i;\alpha_i)$ and corresponding densities denoted by $f_i(x_i;\alpha_i)$. The parameters of the marginal distribution are $\theta_1 = \left(\alpha_1^T, ..., \alpha_d^T\right)^T$. Let $F(x_1, ..., x_d;\theta) = C\left(F_1(x_1,\alpha_1), ..., F_d(x_d,\alpha_d);\theta_2\right)$, where C is a copula distribution function with parameters θ_2, and let $\theta = \left(\theta_1^T, \theta_2^T\right)^T$. Given a marginal distributions, different multivariate distributions can be formed simply by applying different copula functions. The density is given by

$$f\left(x_1, ..., x_d;\theta\right) = c(F_1(x_1,\alpha_1), ..., F_d(x_d,\alpha_d);\theta_2) f_1(x_1,\alpha_1) ... f_d(x_d,\alpha_d).$$

The log-likelihood function of a sample $x_t = (x_{t1}, x_{t2}, ..., x_{td})$ is

$$l(\theta) = \sum_{t=1}^{T} \ln c\left(F_1(x_{t1};\alpha_1), ..., F_d(x_{t2};\alpha_d);\theta_2\right) + \sum_{t=1}^{T}\sum_{i=1}^{d} \ln f_i(x_{ti};\alpha_i).$$

The log-likelihood function can be decomposed as: $l(\theta) = l_c(\theta_1,\theta_2) + l_m(\theta_1)$ – this decomposition could not hold if some parameters are common across the margins and the copula (Patton, 2006). The ML estimation involves maximizing the log-likelihood function with respect to all the parameters θ simultaneously. Because of complicated form of the likelihood function, it is difficult to accomplish this goal. That is why some simplifications were introduced. In the empirical work we consider IFM strategy (Shih, Louis, 1995; Joe, Xu, 1996). IFM proceeds in two steps. Firstly, $\hat{\theta}_1$ parameter estimators of margins distribution were estimated. Secondly, the estimator of copula dependency parameter was obtained as $\hat{\theta}_2 = \operatorname{argmax}\, l_c\left(\hat{\theta}_1, \theta_2\right)$. Under some regularity conditions, Patton (2006) shows that the IFM procedure yields consistent and asymptotically normal

estimates. However, there is some loss of efficiency in estimation because Step 1 ignores the dependence between margins. The other algorithms were also proposed by Lui, Luger (2009).

3.2. Testing procedure

The maximum likelihood value or other measures derived from the likelihood method, such as the Akaike (AIC) or Schwarz, or Bayesian information criterion are currently used in model selection. However, those criteria are not sufficient to accept a copula as being enough representative. To show that the considered copulas are representative enough, the testing of goodness of fit is required. Generally, the goodness-of-fit testing is equivalent to test the hypothesis

$$H_0 : C(u_1,...,u_d) = C(u_1,...,u_d;\theta) \text{ for } \theta \in \Theta,$$

i.e., the unknown copula $C(u_1,...,u_d)$ is a member of the parametric family, against the alternative hypothesis

$$H_1 : C(u_1,...,u_d) \neq C(u_1,...,u_d;\theta) \text{ for } \theta \in \Theta$$

i.e., the unknown copula is not a member of the parametric family.

3.2.1. A test based on Rosenblatt's transform

This test is based on the probability integral transformation due to Rosenblatt(1952). Let $C_j(u_1,...u_j,\theta)$ denote the joint distribution function of $U_1,...,U_j$ under H_0. It is given by $C_j(u_1,...u_j,\theta) = C(u_1,...,u_j,1,...,1;\theta)$. In addition, let $C_j(u_j;\theta \,|U_1,...,U_{j-1})$ denote the conditional distribution function of U_j given $(U_{j-1},...,U_1)$ under H_0. It can be derived via:

$$C_j(u_j;\theta \,|U_1,...,U_{j-1}) = \frac{\partial^{j-1} C_j(u_1,...u_j,\theta)}{\partial u_1 ... \partial u_{j-1}} \Big/ \frac{\partial^{j-1} C_{j-1}(u_1,...,u_{j-1};\theta)}{\partial u_1 ... \partial u_{j-1}}.$$

Define $Z_1 = U_1$ and $Z_j = C_j(U_j;\theta \,|U_1,...,U_{j-1})$ for $j = 2,...,d$.

For the two-dimensional case,

$$Z_1 = U, \; Z_2 = C_2(v;\theta|U),$$

where

$$C_2(v;\theta|U)=\frac{\partial C\ (u,v;\theta)}{\partial u}.$$

Since the copula function is a multivariate distribution function, the H_0 holds if and only if the probability integral transformed random variables $Z_1,...,Z_d$ are independent and identically distributed as a Uniform$[0,1]$ (Rosenblatt, 1952).

Exploiting that fact, the random variable $W=\sum_{j=1}^{d}\left[\Phi^{-1}(Z_j)\right]^2$ should follow a χ_d^2 distribution. To test this distribution, Breymann et al. (2003) applied the Anderson-Darling statistic. Since W is not observable, the pseudo observations $\hat{W}_t$ are taken into consideration:

$$\hat{W}_t=\sum_{j=1}^{d}\left[\Phi^{-1}(Z_{ij})\right]^2,\ i=1,...,n\ .$$

For more details on this testing procedure see Genest and Remillard (2008), Genest et.al. (2009), Kojdanovic and Yan (2009).

3.2.2. Chi-square test

The chi-square test can be applied only in the case when known margins are estimated using parametric methods. Each marginal distribution model must be accepted which requires some previous estimation and validation procedures. Furthermore, the observations are assumed to be independent and identically distributed. Otherwise, the chi-square test does not hold. The procedure for testing bivariate copulas is as follows. Let A_{ij} be bins of $[0,1]^2$ (the *y*-axis is divided into *r* parts and the *x*-axis is divided into *s* parts), O_{ij} be the observed frequency for bin A_{ij}, and let $E_{ij}(\hat{\theta})=n\cdot p_{ij}(\hat{\theta})$ be the expected frequency for A_{ij}, where $p_{ij}(\hat{\theta})=\iint_{A_{ij}} dC(u,v;\hat{\theta})$.

The chi-square statistic is calculated by formula:

$$\chi^2(\hat{\theta})=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(O_{ij}-E_{ij}(\hat{\theta})\right)^2}{E_{ij}(\hat{\theta})},$$

where $\hat{\theta}$ is the vector of estimated parameters. The test statistic follows, approximately, a chi-square distribution with $(rs-1-d)$ degrees of freedom. This analysis is sensitive to the selection of the number of bins (Roch, Alegree (2006)). In the empirical study, we used two partitions. Firstly, the unit square is divided into bins of equal area, being $r=s=10$. Secondly, the unit square was separated into eight squares, as it was suggested by Patton (2003). This second partition is useful for expressing the tail dependences.

4. Empirical results

4.1. The data

We investigate the relationships between forty two selected stock indices. Table 1 presents the variables under consideration.

Table 1. The indices from forty two countries

	Country	Name	Code		Country	Name	Code
1	Argentine	MERVAL	ME	22	Japan	NIKKEI	NI
2	Australia	AS30	AS	23	Latvia	RGSE	RG
3	Austria	ATX	ATX	24	Malaysia	KLSE	KL
4	Belgium	BEL20	BE	25	Mexico	IPC	IPC
5	Brazil	BOVESPA	BO	26	New Zealand	NZSE50	NZ
6	Bulgarian	SOFIX	SO	27	Norway	OSE	OS
7	Canada	TSE300	TS	28	the Philippines	PSE	PS
8	Chile	IPSA	IP	29	Poland	WIG	WI
9	China	SHANGHAI	SH	30	Romania	BET	BET
10	Czech Republic	PX50	PX	31	Russia	RTS	RT
11	Estonia	TLSE	TL	32	Singapore	STI	ST
12	Finland	HEX	HE	33	Slovakia	SAX	SA
13	France	CAC40	CA	34	South Korea	KOSPI	KO
14	Germany	DAX	DA	35	Spain	IBEX	IB
15	Greece	ATGI	AT	36	Switzerland	SMI Swiss	SM
16	Nederland	AEX	AE	37	Taiwan	TWSE	TW
17	Hong Kong	HSI	HS	38	Thailand	SET	SE
18	Hungary	BUX	BU	39	Turkey	ISE100	IS
19	India	BSE30	BS	40	the UK	FTSE250	FT
20	Indonesia	JKSE	JK	41	the USA	DJIA	DJ
21	Italia	MIB30	MI	42	the USA	NASDAQ	NA

The sample consists of daily frequency data and covers the period from January 2002 to December 2008. Missing data were filled by linear interpolation from the preceding to the following missing quotations. The return of indices is defined as $r_t = \ln P_t / P_{t-1}$ where P_t (t =1, …, T) is an adjusted index value at period t. In our empirical work we used *R-project* program.

4.2. Estimation of marginal model

In a preliminary step of our empirical work, we investigate the structure of the univariate marginal returns. We consider several restrictions as a possible candidates for adjusting the empirical return distribution. The model GARCH(1, 1) with symmetrical and skewed conditional distributions is specified. The unknown parameters of this specification are estimated using the ML method. We have selected one of discussed conditional distributions that fits best to the data in terms of the values of the log-likelihood. The analysis of the log-likelihood or the Akaike criterion implies that skewed distributions fit better to the data than symmetrical ones. Thus, the procedure of testing the goodness-of-fit is carried out only for skewed Student's t distribution and skewed GED distribution. For the testing purposes, we follow the procedure described in Diebold et al. (1998). If a marginal distribution is correctly specified, the margin u_{it} denoting the transformed GARCH(1,1) standardized residuals should be *iid* Uniform $[0, 1]$. The test is performed in two steps. Firstly, we evaluate whether u_{it} is serially independent. Doing so, we separately examine the serial correlation of $(u_{it} - \overline{u}_i)^k$, for $k = 1,...,4$, on 20 own lags. The k-th test statistic is defined as $R(k) = (T-20)R^2$, where R^2 is the coefficient of determination of the regression, and is distributed as a χ^2_{20} under the null hypothesis. Secondly, we test the null hypothesis that u_{it} is Uniform $[0, 1]$. For the testing purposes, we divide the empirical and theoretical distributions into N bins and then we test whether the two distributions significantly differ at each bin. We consider the case with N=10 and we adapt the chi-square test. Table 2 reports goodness-of-fit statistics. The first four columns contain p-values, labeled $p(k)$, of $R(k)$ under the null hypothesis of no serial correlation of the k-th centered moments of the u_{it}. Next column reports p-values for the chi-square test statistics under the null hypothesis that *cdf* of residuals is Uniform [0, 1]. Last column presents the maximum likelihood values for selected conditional distributions.

Table 2. The p-values for test specification and maximum likelihood values

State	*p*(1)	*p*(2)	*p*(3)	*p*(4)	*p*	Max likelihood
Argentina	0.741	0.984	0.509	0.981	0.528	4534.17
Australia	0.880	0.195	0.780	0.379	0.020	5938.64
Austria	0.568	0.949	0.360	0.978	0.056	5343.64
Belgium	0.975	0.155	0.569	0.351	0.001	5448.18
Canada	0.709	0.399	0.349	0.417	0.702	5720.46
China	0.000	0.183	0.000	0.171	0.060	4698.27
Czech R	0.258	0.993	0.057	0.972	0.002	5228.22
Estonia	0.000	0.001	0.000	0.001	0.071	5805.34
Finland	0.332	0.371	0.435	0.273	0.253	4945.83
France	0.029	0.011	0.306	0.022	0.125	5187.11
Germany	0.419	0.101	0.507	0.308	0.932	5070.78
Greece	0.000	0.135	0.000	0.257	0.769	5242.24
Netherlands	0.134	0.045	0.579	0.063	0.411	5187.72
Hong Kong *	0.094	0.090	0.072	0.250	0.503	5170.27
Hungary	0.470	0.935	0.037	0.817	0.031	4922.01
Italy *	0.695	0.409	0.660	0.640	0.382	5471.54
Japan *	0.097	0.589	0.062	0.411	0.186	5018.13
Mexico	0.087	0.880	0.113	0.656	0.044	5154.21
Norway	0.864	0.579	0.420	0.554	0.052	5110.74
Poland*	0.135	0.005	0.025	0.007	0.976	5167.47
Russia	0.001	0.582	0.000	0.889	0.608	4555.85
Slovakia	0.000	0.000	0.000	0.000	0.722	5546.35
South Korea *	0.144	0.344	0.060	0.542	0.813	4850.86
Spain	0.667	0.573	0.744	0.547	0.149	5332.16
Switzerland	0.188	0.035	0.289	0.144	0.071	5405.75
Taiwan	0.008	0.075	0.003	0.018	0.105	5013.62
Turkey	0.165	0.513	0.056	0.306	0.001	4286.35
UK	0.002	0.011	0.023	0.250	0.069	5682.96
USA NASDAQ *	0.710	0.081	0.929	0.127	0.433	5092.67
USA DJIA *	0.265	0.049	0.057	0.972	0.077	5601.52

** GARCH with skewed GED conditional distribution*
Note: p(k) - p-values for the null hypothesis of no serial correlation of the k-th centered moments; p are p-values for the chi-square test statistics for the null hypothesis that cdf of residuals is Uniform [0, 1];

The skewed *t*-Student is the most frequently chosen conditional distribution. Only in the case of a few indices, the skewed GED fits better. We present all indices with at last one searching conditional distribution not being fulfilled at a

1% significance level. The results for data satisfying both conditions, *iid* and uniformity, are written in italics. We can notice that specifications of the margins are quite proper for the well developed countries. For other indices, it is difficult to fix a proper specification. The model is not correctly specified for Romania, Brazil, India, Chile, Indonesia, Malaysia, Latvia, Thailand, Bulgaria, Singapore, the Philippines. For these indices we failed to obtain the independence as well as to select the proper conditional distribution. For Belgium, Turkey and Czech Republic the specification of GARCH(1, 1) is correct, but the null hypothesis of uniformity were rejected. For Poland, Grace, the UK, Taiwan, Russia, Slovakia, China, Estonia, the GARCH(1,1) is not sufficient enough, whereas the conditional distributions were selected properly.

4.3. Estimation of the multivariate model

In the empirical work, we consider the case of constant parameter dependency with respect to time. This assumption is justified by relatively short sample size. It is impossible to estimate multidimensional copula function due to the fact that the number of data to be investigated would be large. Therefore, the bivariate copula parameters for all market pairs are estimated. The combination of the data vectors yields 861 pairs. For almost all market pairs, the estimates of the dependency parameter for the Normal and *t*-Student copula is found to be positive and strongly significant. Strong dependency between European pairs is observed. Table 3 reports some of the parameters estimated.

Table 3. Parameter estimates of the Normal copula and *t*-Student copula

	Canada	Brazil	France	Germany	Nederland	Italy	Japan	South Korea	Switzerland	USA (DJIA)
Canada		0.43	0.48	0.63	0.49	0.49	0.22	0.24	0.40	0.63
Brazil	0.44		0.45	0.43	0.46	0.45	0.32	0.36	0.41	0.32
France	0.49	0.46		0.92	0.93	0.90	0.37	0.33	0.85	0.51
Germany	0.63	0.44	0.92		0.89	0.87	0.33	0.32	0.82	0.55
Nederland	0.49	0.47	0.93	0.89		0.86	0.37	0.33	0.83	0.51
Italy	0.50	0.45	0.90	0.87	0.86		0.29	0.25	0.79	0.50
Japan	0.23	0.33	0.38	0.34	0.38	0.31		0.63	0.37	0.14
South Korea	0.25	0.37	0.33	0.32	0.34	0.30	0.64		0.34	0.14
Switzerland	0.42	0.42	0.85	0.82	0.83	0.80	0.38	0.35		0.43
USA (DJIA)	0.63	0.32	0.53	0.56	0.52	0.51	0.14	0.14	0.45	

Note: This table supply estimated parameters of dependency; for that parameters all p – values are smaller than 0.0001 Estimates for the Normal copula are reported below the main diagonal, while for the t-Student copula above it.

To investigate whether a given copula function is able to capture the dependence structure present in the data, we perform two goodness-of-fit tests. It is known, that chi-square test requires proper specifications of the margins. That's the reason tests were made only for pairs formed from correctly specified margins. The tests were carried out for Normal copula and *t*-Student copula. Table 4 presents the results. To use the chi-square, the unit square $[0, 1]^2$ is divided into 100 bins of equal area.

Table 4. The percentage of cases where we do not to reject the null at a 1% significance level

Copula function	Test based on the probability transform	Chi-square test
Normal	70	74
t-Student	75	80

Note: Only for pairs formed from correctly specified margins.

In both tests, the *t*-Student copula better describes the dependence structure between correctly specified margins. The results indicates that the null hypothesis is not rejected only in approximately eighty percent of cases, which is not a satisfactory result. Hence, we cannot directly assess the adequacy of the models. Thus, we now turn to the hit test suggested by Patton (2003). This test is based on the chi-square statistic associated with a partition of the unit square into eight zones. Instead of the total chi-square value, we examine the contribution of each particular region to the global test statistic to find whether this copulas are unveiled. Such evidence were presented in the Canela, Collazo (2006). In Table 5, we document the average of the contributions of each region to the total value.

Table 5. The average of the contribution to the chi-square value (%)

	Regions							
Copula	1	2	3	4	5	6	7	8
Normal	12	16	13	6	3	21	23	5
t-Student	7	40	10	7	8	8	6	13

Note: Zones: the first zone is a square at lower left top (0, 0) and upper right top (0.1, 0.1); the second square has lower left top (0.9, 0.9) and upper right top (1, 1); the third square has tops (0.1, 0.1); (0.25, 0.25); the fourth - (0.75, 0.75); (0.9, 0.9); the most important is the fifth region - a big square with tops (0.25, 0.25); (0.75, 0.75); the sixth has lower left (0, 0.75) and upper right (0.25, 1); the seventh is a square with lower left top (0, 0.75) and upper right top (1, 0.25); the eight region is the rest area.

Only pairs formed from correctly specified margins; the region 5 determine only 8% of the whole chi-square value (*t*-Student copula), in contrast to region 2 which determine 40% of the whole chi-square value (the same copula).

The region 2 is the worst one for the *t*-Student copula. The other regions are clarified well enough. On the basis of the above results, we come to the conclusion that rejection of null hypothesis is caused by predictably bad performance in the upper-right quadrant. The rejected copulas do not therefore describe upper tail dependence.

5. The cluster analysis

In our study, we use Ward's clustering algorithm. Given a dissimilarity matrix $D = (d_{ij})$, the Ward algorithm can be formulated as follows. In an initial setting, all the entities are considered as singleton clusters. Assuming that there is N elements to classify, begin with N clusters consisting exactly of one entity, search the similarity matrix, reduce the number of clusters by one through merging the most similar pair of clusters. Perform those steps until all clusters are merged. The Ward is different from all other methods because it uses an analysis of variance approach to evaluate the distances between clusters. In short, this method attempts to minimize the Sum of Squares of any two hypothetical clusters that can be formed at each step.

For the choice of relevant metric, some authors used the Pearson correlation coefficient as a similarity measure of a pair of time series. The square of this distance is a popular and widely used dissimilarity index between standardized variables (Krzanowski, 1988; Kaufman, Rousseeuw, 1990) but requires the assumption of the normal margins. Although this metric can be useful to ascertain the structure of stock returns movements, it could not be used in the case of non-elliptical distributions.

As a dissimilarity measure between y_i and y_j time series we propose to take, instead of Pearson correlation, the dependency parameter ρ_{ij} from Normal (2) or *t*-Student (3) copula. (In the case of the multivariate normal distribution, the proposed ρ_{ij} distance is the same as the distance based on the Pearson correlation). For y_i and y_j specified by equation (1), the $\hat{\rho}_{ij}$ parameter is computed via MLM for $u_1 = F_i\left(\varepsilon_{ti} / \sqrt{h_{ti}}; \alpha_i\right)$ and $u_2 = F_j\left(\varepsilon_{tj} / \sqrt{h_{tj}}; \alpha_j\right)$. For the clustering purposes we considered the dissimilarity index between two time series as $1 - \hat{\rho}_{ij}$.

Fig. 1 presents the dendogram obtained with the distance discussed above based on the parameter obtained from Normal copula, whereas Fig. 2 shows the results for the distance based on the parameter obtained from the *t*-Student copula. We can notice that the cluster groups slightly differ.

Figure 1. The dendrogram for Normal copula

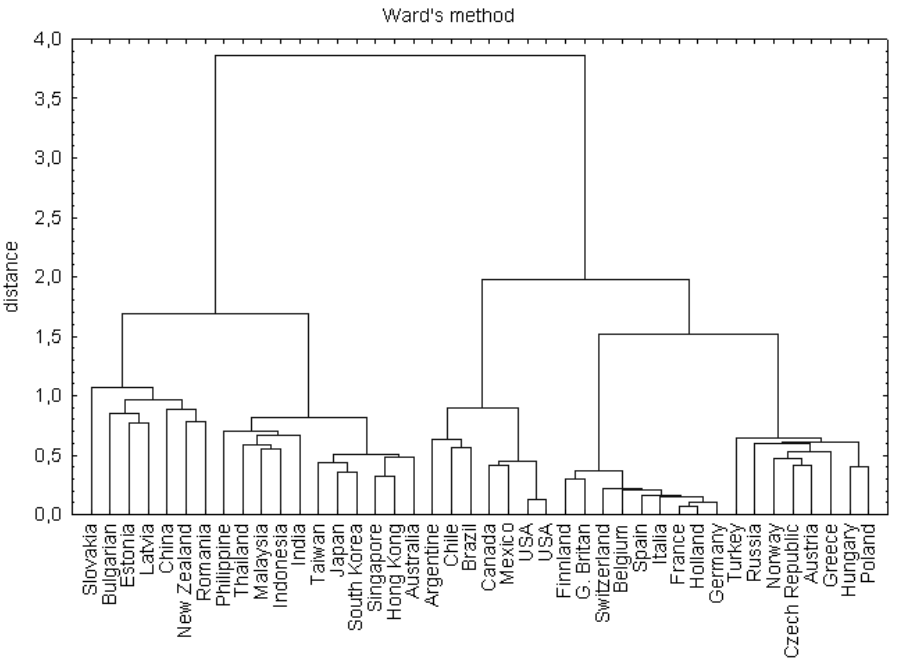
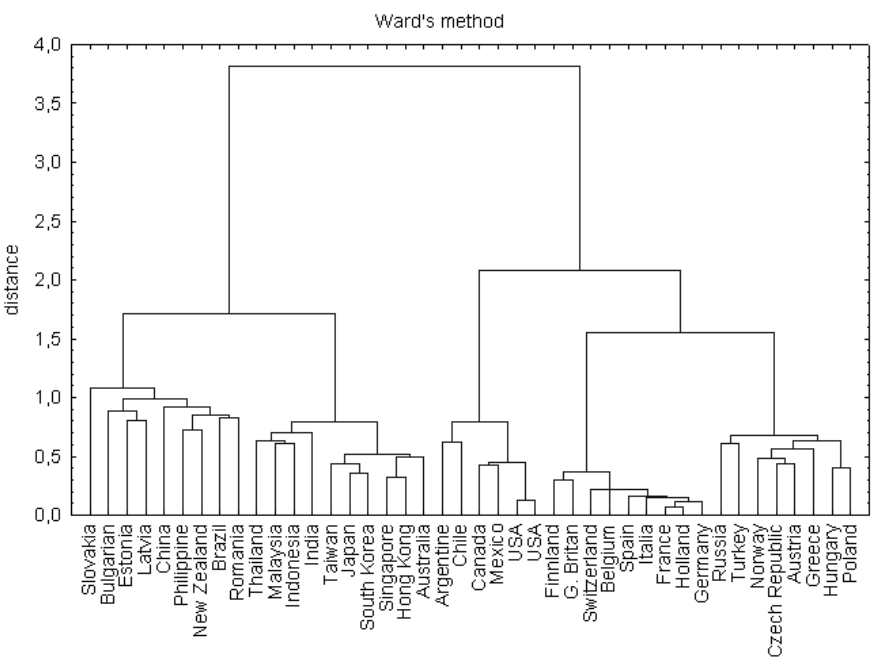


Figure 2. The dendrogram for *t*-Student copula

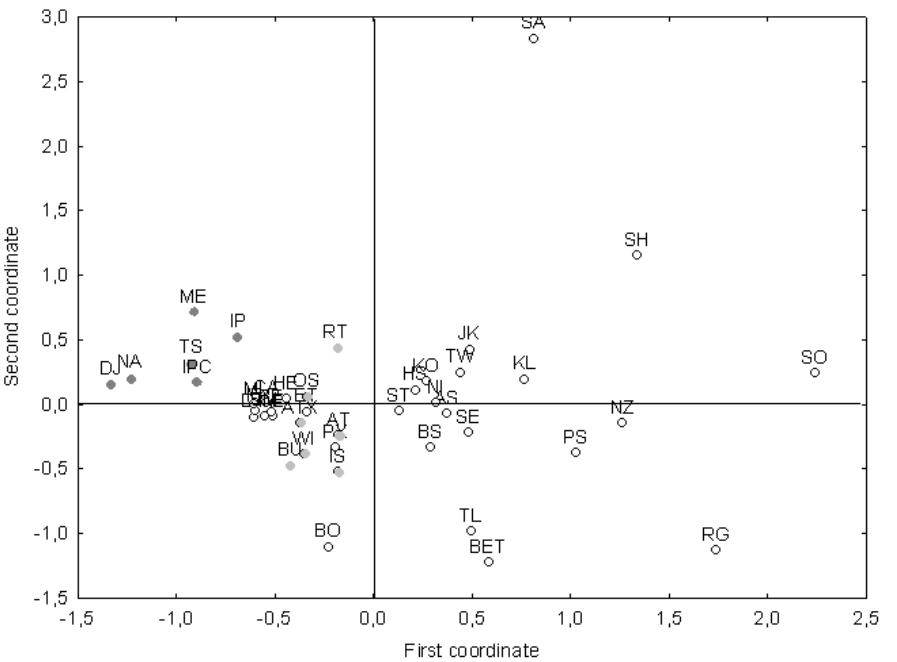


We obtained five groups. First one includes indices from West Europe: Switzerland, Belgium, Spain, Italy, France, the Netherlands, Germany. Finland and the UK also belong to this group. The second group consists of three small subgroups. It is formed by Poland, Hungary, Norway and Czech Republic, Austria and Russia, Turkey. These group includes some countries from East Europe. The third group involves American countries. Next group merges three subgroups of indices from Asia: Thailand, Malaysia, Indonesia, India and Taiwan, Japan, South Korea and the last: Singapore, Hong Kong, Australia. Finally, we have a group of outlier countries with weak connection with another ones. It includes: China, Slovakia, Bulgaria, Estonia, Latvia, the Philippines, New Zealand, Brazil and Romania.

The results of goodness-of-fit testing indicate that for properly specified margins the conclusions should be drawn from Fig. 2. However, as can be seen from the figures, for properly specified margins both copulas gave identical clustering results.

In order to summarize and better interpret the results, the multidimensional scaling map is presented. Fig. 3. presents the map for all discussed indices.

Figure 3. The multidimensional scaling map for all indices



The cluster groups are confirmed by the scaling map results. We can notice the indices concentrations according to the geographical regions, especially very dense concentration for West Europe and Asia indices. The outliers are noticeable

too. The most irrelevant indices are Slovakian (SA), Latvian (RG), Chinese (SH) and Bulgarian (SO).

Figure 4. The multidimensional scaling map for GARCH indices. First distance is based on the dependency structure, second is based on the fitted autoregressive expansions.

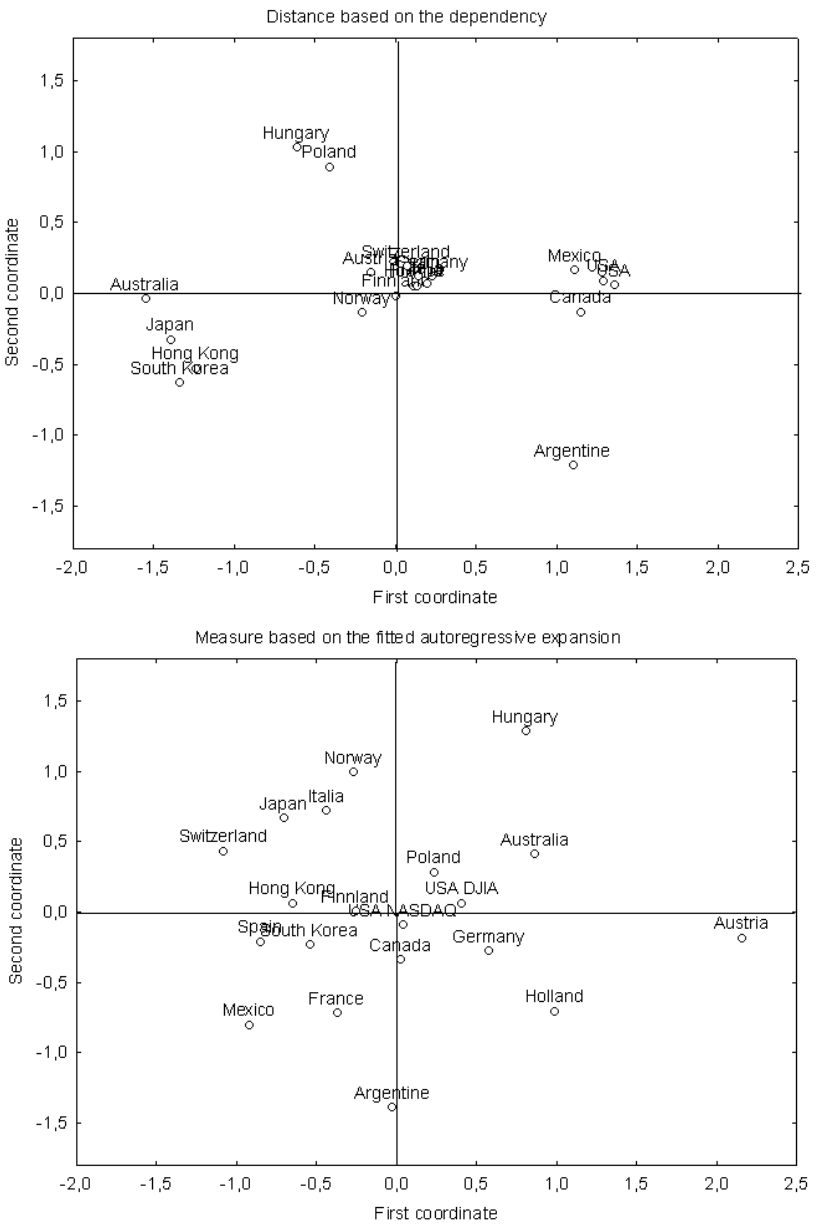


Fig. 4. presents the scaling map for the well specified model GARCH(1,1). The measure discussed in this paper and based on the fitted autoregressive expansions suggested by Otranto (2004) is taken as a dissimilarity measure. There are presented clustered groups in the case of the method discussed in this paper in contrary to dispersed map of indices where the distance is based on the fitted autoregressive expansions.

6. Conclusions

The aim of the work was to separate, by means of cluster analysis, forty two indices returns coming from European, American, and Asian stock markets, into similar groups. Modelling the dependency between stock market returns is a difficult task when returns follow a complicated dynamics. When returns are non-normal, it is often simply impossible to specify the multivariate distribution relating two or more return series. To cope with this problem, the Copula-Garch model is adapted. Firstly, the specification of GARCH(1,1) model for indices returns was verified. We have found this specification to be proper for developed markets, whereas it failed for emerging markets. Secondly, the hypothesis that the Normal and *t*-Student copulas are able to capture the dependency structure of the markets were tested. In the case of *t*-Student copula, this hypothesis was rejected approximately in 30 percent of all pairs, and approximately in 20 percent in the case of properly specified margins. These rejections were due to bad performance in the model upper tail dependence.

For the clustering purposes, we have adapted Ward's algorithm. The similarity measure as well as the dissimilarity index were based on the parameter from Normal or *t*-Student copula obtained by the maximum likelihood method. We have received five clustered groups tied to geographical regions: West, East Europe, American regions and Asia, which confirms our expectations. We have also received the group of countries weakly connected between them and the other markets.

Researches were prepared without taking into consideration area of time. The time dependences were reflected in groups. The lags behind for time series will be included in the next study. We also intend to take into consideration the tail dependence and dynamic copulas for long time series in the future. For the specifications of the margins, the GARCH model in which the first four moments are conditional and time varying, will also be discussed.

REFERENCES

ARELLANO-VALLE, R., and GÓMEZ, H., and QUINTANA, F., 2004, A New Class of Skew-Normal Distributions, *Communications in Statistic*s, Series A, 33(7), 1465–1480.

BOLLERSLEV T., 1986, Generalized Autoregressive Conditional Heteroskedasticity, *Journal of Econometrics*, 31, 307–327.

BOLLERSLEV T., WOOLDRIDGE J.M., 1992, Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariance's, *Econometric Reviews*, 11, 143–172.

BREYMANN W., DIAS A., EMBRECHTS P., 2003, Dependence Structures for Multivariate High-Frequency Data in Finance, *Quantitative Finance*, 3, 1–14.

CAIADO J., CRANTO N., PENA D., 2006, A periodogram-based metric for time series classification, *Computation Statistic & Data Analysis* v50 i10, 2668–2684.

CANELA M.A., COLLAZO E. P., 2005, *Modeling Dependence in Latin American Markets Using Copula Function*, Working Paper.

CHEN X., FAN Y., PATTON, A.J., 2004, *Simple Tests for Models of Dependence Between Multiple Financial Time Series, with Applications to U.S. Equity Returns and Exchange Rates*, FMG Discussion Papers, Financial Markets Group.

DIEBOLD F.X., GUNTHER T.A., TAY A.S., 1998, Evaluating Density Forecasts with Applications to Financial Risk Management, *International Economic Review*, 39(4), 863–883.

EMBRECHTS P., LINDSKOG F., MCNEIL A., 2001a, *Modeling Dependence with Copulas and Applications to Risk Management*, ETHZ, Working Paper.

EMBREECHT P., MCNEIL A.J., STRAUMANN D., 2001b, *Correlation and dependency in risk management: properties and pitfalls*. [In:] M. Dempster, H. Moffant, Risk Management, Cambridge University Press , New York, 176–223.

EMBRECHTS P., LINDSKOG F., MCNEIL A., 2003, *Modeling Dependence with Copulas and Applications to Risk Management*, In: Rachev, S.T. (Ed.), Handbook of Heavy Tailed Distributions in Finance, Elsevier/Noth-Holland, Amsterdam.

FERNANDEZ C., STEEL M., 1998, On Beyesian Modeling of Fat Tails and Skewness, *Journal of the American Statistical Association*, 93, 359–371.

EMBRECHTS P., LINDSKOG F., MCNEIL A., 2001a, Modeling *Dependence with Copulas and Applications to Risk Management*, ETHZ, Working Paper.

EMBREECHT P., MCNEIL A.J., STRAUMANN D., 2001b, *Correlation and dependency in risk management: properties and pitfalls*. [In:] M. Dempster, H. Moffant, Risk Management, Cambridge University Press , New York, 176–223.

EMBRECHTS P., LINDSKOG F., MCNEIL A., 2003, Modeling Dependence with Copulas and Applications to Risk Management, [In:] Rachev, S.T. (Ed.), *Handbook of Heavy Tailed Distributions in Finance*, Elsevier/Noth-Holland, Amsterdam.

FERNANDEZ C., STEEL M., 1998, On Beyesian Modeling of Fat Tails and Skewness, *Journal of the American Statistical Association*, 93, 359–371.

GENEST R., REMILLARD B., 2008, Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models, *Annales de l'Institut Henri Poincare: Probabilites et Statistiques*, 44, 1096–1127.

GENEST R., REMILLARD B., BEAUDOIN D., 2009, Goodness-of-fit tests for copulas: A review and a power study, *Insurance: Mathematics and Economics*, 44, 199–214.

HANSEN B., 1994, Autoregressive Conditional Density Estimation, *International Economic Review*, v.35, no. 3,705–730.

JOE H., XU J.J., 1996, *The estimation method of inference function for margins for multivariate models*, Technical Report, Departments of Statistics, University of British Columbia.

JUNKER M., MAY A., 2005, Measurement of aggregate risk with copulas, *Econometrics Journal, Royal Economic Society*, 8(3): 428–454, December.

KAUFMAN L., ROUSSEEUW P., 1990, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.

KOJADINOVIC I., YAN J., 2009a, *Fast large-sample goodness-of-fit tests for copulas*, Submitted.

KOJADINOVIC I., YAN J., 2009b, A goodness-of-fit test for multivariate multiparameter copulas based on multiplier central limit theorems, *Statistics and Computing*,. In press.

KRZANOWSKI W., LAI Y., 1988, A Criterion For Determining the Number of Groups In Data Set Using Sum of Squares Clustering, *Biometrics* 44(1), 23–34.

LUI Y., LUGER R., 2009, Efficient estimation of copula-GARCH models 1, *Computational Statistics & Data Analysi*s, 53, 2284–2297

MASHAL R., ZEEVI A., 2002, *Beyond Correlation: Extreme co-movements Between Financial Assets*, Mimeo, Columbia Graduate School of Business.

MIRKIN B., 2005, *Clustering for Data Mining: A Data Recovery Approach*, Boca Raton Fl., Chapman and Hall/CRC.

NELSEN B. R., 1999, *An Introduction to Copulas*, Springer Verlag, New York.

NELSON D.B., 1991, Conditional Heteroskedasticity in Asset Returns: A New Approach, *Econometrica*, 59(2), 397–370.

OTRANTO E., 2004, Classifying the Markets Volatility with ARMA Distance Measures, *Quaderni di Statistica*, 6,1–19.

PATTON A.J., 2003, Modeling Asymmetric Exchange Rate Dependence, Working paper, University of California, San Diego.

PATTON A.J., 2004. On the Out-of-Sample Importance of Skewness and Asymmetric Dependence for Asset Allocation, *Journal of Financial Econometrics*, Oxford University Press, 2(1), 130–168.

PATTON A.J., 2006, Estimation of multivariate models for time series of possibly different lengths, *Journal of Applied Econometrics*, John Wiley & Sons, Ltd., 21(2), 147–173.

Piccolo, 1990, A distance measure for classifying ARIMA models, *Journal of Time Series Analysis* v11, 153–164.

R Development Core Team, 2004, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN is 3-900051-07-0 URL <http://www.Rproject.org>.

ROCH O., ALEGRE A., 2006, Testing the bivariate distribution of daily equity returns using copulas. An application to the Spanish stock market, *Computational Statistics& Data Analysis*, 51, 1312–1329.

ROSENBLATT M., 1952, Remarks on a Multivariate Transformation, *The Annals of Mathematical statistics*, 23, 470–472.

SHIH J., LOUIS T.A., 1995, Inference on the Association Parameter in Copula Models for Bivariate Survival Data, *Biometric*s, 51: 1384–1399.

SKLAR A., 1959, Fonction de Repartition a n Dimension et Leur Marges, *Publications de L'Institut de Statistiques de L'Universite de Paris*, 8, 229–231.

WARD J. H., 1963, Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, 58, 236–244.