**Central European Journal of Economic Modelling and Econometrics**

# A Bivariate Copula-based Model for a Mixed Binary-Continuous Distribution: A Time Series Approach

Katarzyna Bień-Barkowska*

Submitted: 13.11.2012, Accepted: 28.12.2012

**Abstract**

In this paper we present a copula-based model for a binary and a continuous variable in a time series setup. Within this modeling framework both marginals can be equipped with their own dynamics whereas the contemporaneous dependence between both processes can be flexibly captured via a copula function. We propose a method for testing the goodness-of-fit of such a time series model using probability integral transforms (PIT). This verification procedure allows not only a verification of the goodness-of-fit of the estimated marginal distribution for a continuous variable but also the conditional distribution of a continuous variable given the outcome of its binary counterpart (i.e. the adequacy of the copula choice). We test the model on an empirical example: investigating the relationship between trading volume and the indicators of arbitrarily 'large' price movements on the interbank EUR/PLN spot market.

**Keywords:** copula function, mixed binary-continuous distribution, ACD models, market microstructure

**JEL Classification**: C18, G15

*Warsaw School of Economics, National Bank of Poland; e-mail: katarzyna.bien@sgh.waw.pl

# 1 Introduction

A description of multivariate distributions within the framework of copula models has gained a vast amount of interest in the econometric literature throughout the past two decades. Moreover, the ubiquity of studies that apply the copula theory to empirical finance studies is simply enormous. As Paul Embrechts states: "... copulas has taken the world of finance and insurance, and well beyond, by storm" (Embrechts, 2009). The most popular focus of these financial inquires lies in an accurate depiction of multivariate distributions for continuous variables (cf., Cherubini, 2004; Patton, 2005a, 2005b, 2009; Doman, 2006, 2007, 2011; Gurgul and Syrek, 2006). The popularity of these approaches for financial applications is still growing because copula-based models facilitate a flexible investigation of the time-varying dependence between marginal distributions. These marginal distributions may, quite obviously, refer to financial returns from different assets or various financial markets, which is of crucial interest in the context of market risk management. Relationships between discrete marginals (i.e., between count variables or even binary outcomes) have been much less popular within the context of copula-based models (see Cameron *et al.*, 2004; Zimmer and Trivedi, 2006; Trivedi and Zimmer, 2007; Bień *et al.*, 2007; Bień *et al.*, 2011 for applications of copula models to count variables and Bhat and Sener, 2009; Winkelmann, 2012 for applications of copula models to binary variables).
The aim of this paper is to contribute to the recent literature in two dimensions. First, we will present an easy specification of the copula-based bivariate model for a mixed binary-continuous distribution. To our knowledge this is a novel approach; at least within economics- or finance-oriented studies; although a very similar model has already been proposed in medical sciences by de Leon and Wu (2011). Thus, we adjust the de Leon-Wu model, so that it can fit into a time-series application and serve as a general model depicting the time-varying joint distribution for a pair of variables consisting of a dichotomous one and a continuous one. The advantage of this model lies in is its complexity; it can account for different parametric families of marginal distributions and 'glue' them together using a flexible copula function that is solely responsible for the dependence between marginal processes. Second, we suggest a method for testing the appropriateness of a copula-based time series model using probability integral transforms (PIT), primarily proposed by Diebold *et al.* (1998) for univariate continuous distributions. This verification procedure makes it possible to check the goodness-of-fit of the time-varying marginal distribution for a continuous variable (conditional on the information set until $t-1$), as well as to test the conditional distribution of a continuous variable given the contemporaneous realization of its binary counterpart (which is also conditional on the information set until $t-1$). The conditional distribution of a continuous variable (given the one or the other outcome of the binary variable) depends on the strength of dependence between two marginals. Accordingly, verification of statistical properties inherent to these distribution estimates can tell us something about the goodness-of-fit of a copula-based model with respect to the original time series.

We will also present a simple and intuitive econometric exercise where we apply the adjusted de Leon-Wu copula-based model in order to investigate the relationship between trading volume and an indicator of a 'large' price change on the interbank EUR/PLN spot market. Intraday movements of the EUR/PLN rate cluster on values being multiples of five pips (whereas one pip denotes one hundredth of a Polish grosz). This phenomenon is one of stylized features of high-frequency financial data which was initially evidenced by Harris (1994). Therefore, it may make sense to consider how much volume is required in order to move the FX rate by a certain level (i.e., 5, 10 or 20 pips). Volume and volatility are also significantly positively correlated, which has also been justified by many market microstructure models (cf. Copeland, 1976; Jennings *et al.*, 1981; Easley and O'Hara, 1987; Blume *et al.*, 1994; Easley *et al.* 1997; and others) as well as shown in numerous empirical applications (see Karpoff, 1987 and Louhichi, 2011 for the vast surveys of such studies). Moreover, the trading volume and the indicator of a sufficiently 'large' price change have many of the stylized properties of the trading marks (i.e. tick-by-tick data). Both variables are strongly autocorrelated, both are prone to intraday seasonality patterns, one variable is discrete and the other is defined on the strictly positive domain. Such a 'bundle' of different characteristics inherent to market microstructure data as: different distribution families, contemporaneous dependence and the pronounced dynamic features make from the pair defined as "price change indicator - trading volume" kind of an excellent laboratory for presenting our copula-based model specification and functioning of the PIT verification procedures.

# 2 General model

## 2.1 The bivariate probit and the Winkelmann's copula bivariate probit model

The copula-based bivariate model for mixed binary-continuous distribution is very much related to the work of Winkelmann (2012), who proposed a copula-based probit model for a bivariate distribution of two binary variables. Thus, in order to maintain a clear exposition, we will begin from the detailed description of this approach.
It is important to note that the Winkelmann's copula-based probit model allows for correlation between two random, normally distributed, latent variables that underlie the process of corresponding observed binary outcomes, whereas this specification does not impose a restriction of the bivariate normality on the joint bivariate distribution of two latent factors. Additionally, application of a copula function facilitates a much more complex dependence structure than a linear correlation that is a measure of codependence inherent to elliptical distributions.
In the standard framework of a bivariate probit model defined for two binary variables, $Y_1$ and $Y_2$, their underlying latent and normally distributed counterparts can be defined as follows: $Y_{1,i}^* = z_{1,i}^T\alpha + \varepsilon_{1,i}$ ($Y_{1,i} = 0$ if $Y_{1,i}^* \leq 0$) and $Y_{2,i}^* = z_{2,i}^T\beta + \varepsilon_{2,i}$

($Y_{2,i} = 0$ if $Y^*_{2,i} \leq 0$), where $z_{1,i}$ and $z_{2,i}$ denote explanatory variables, $\alpha$ and $\beta$ are the corresponding parameters and $\varepsilon_{1,i} \sim I.I.D.\ N\left(0, \sigma_1^2\right)$, $\varepsilon_{2,i} \sim I.I.D.\ N\left(0, \sigma_2^2\right)$. Since $Y_1$ and $Y_2$ have only two outcomes, the joint distribution of $Y_1$ and $Y_2$ can be fully characterized by four distinct probabilities:

$$P\left(Y_{1,i} = 0, Y_{2,i} = 0 | z_{1,i}, z_{2,i}\right) = P\left(\varepsilon_{1,i} \leq -z_{1,i}^T\alpha, \varepsilon_{2,i} \leq -z_{2,i}^T\beta\right) \tag{1}$$

$$P\left(Y_{1,i} = 1, Y_{2,i} = 0 | z_{1,i}, z_{2,i}\right) = P\left(\varepsilon_{1,i} > -z_{1,i}^T\alpha, \varepsilon_{2,i} \leq -z_{2,i}^T\beta\right) \tag{2}$$

$$P\left(Y_{1,i} = 0, Y_{2,i} = 1 | z_{1,i}, z_{2,i}\right) = P\left(\varepsilon_{1,i} \leq -z_{1,i}^T\alpha, \varepsilon_{2,i} > -z_{2,i}^T\beta\right) \tag{3}$$

$$P\left(Y_{1,i} = 1, Y_{2,i} = 1 | z_{1,i}, z_{2,i}\right) = P\left(\varepsilon_{1,i} > -z_{1,i}^T\alpha, \varepsilon_{2,i} > -z_{2,i}^T\beta\right) \tag{4}$$

Assuming that the joint distribution of $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$ is bivariate normal, each of the probabilities (1)-(4) can be obtained with the help of a bivariate normal cumulative distribution function $\Phi_2\left(\cdot;\rho\right)$, where $\rho$ denotes the linear correlation parameter between $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$.

$$P\left(Y_{1,i} = 0, Y_{2,i} = 0 | z_{1,i}, z_{2,i}\right) = \Phi_2\left(-z_{1,i}^T\alpha, -z_{2,i}^T\beta; \rho\right) \tag{5}$$

$$P\left(Y_{1,i} = 1, Y_{2,i} = 0 | z_{1,i}, z_{2,i}\right) = \Phi\left(-z_{2,i}^T\beta\right) - \Phi_2\left(-z_{1,i}^T\alpha, -z_{2,i}^T\beta; \rho\right) \tag{6}$$

$$P\left(Y_{1,i} = 0, Y_{2,i} = 1 | z_{1,i}, z_{2,i}\right) = \Phi\left(-z_{1,i}^T\alpha\right) - \Phi_2\left(-z_{1,i}^T\alpha, -z_{2,i}^T\beta; \rho\right) \tag{7}$$

$$P\left(Y_{1,i} = 1, Y_{2,i} = 1 | z_{1,i}, z_{2,i}\right) = \Phi_2\left(z_{1,i}^T\alpha, z_{2,i}^T\beta; \rho\right) \tag{8}$$

This exposition of a standard bivariate probit model, that was initially proposed by Ashford and Sowden (1970), served Winkelmann (2012) as a starting point for deriving a more general, copula-based bivariate model for two mutually dependent binary variables. The concept of a copula function, initially proposed by Sklar (1959), is currently an extremely popular statistical tool for building multivariate distributions from the marginal distributions and the copula function linking these marginal distributions together into a joint multivariate distribution and being responsible for the dependence structure between the marginals. Within the financial literature there are plenty of valuable surveys of this theory as well as applications of copula-based models (cf. Cherubini *et al.*, 2004; Nelsen, 2006; Trivedi and Zimmer, 2007; Doman, 2011). In an attempt to conserve space, in this paper we refrain from presenting the details of this concept. Nevertheless, to maintain a clear and understandable exposition, we will remind the reader that from a statistical viewpoint a copula function is simply an n-dimensional cumulative distribution function $C : [0,1]^n \rightarrow [0,1]$. For a 2-dimentional case, the bivariate distribution of two random continuous variables $A$ and $B$ has its unique copula representation, i.e., $C_{A,B} : [0,1]^2 \rightarrow [0,1]$, as follows:

$$P\left(A \leq a, B \leq b\right) = C_{A,B}\left(F_A(a), F_B(b); \theta\right), \tag{9}$$

where $F_A(a)$ and $F_B(b)$ denote the cumulative distribution functions for variables $A$ and $B$ respectively and $\theta$ denotes the copula parameter, called the 'dependence

parameter', which accounts for a dependence structure between both marginals.
In the context of the bivariate probit model, maintaining the assumption about the univariate normality of $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$, but relaxing the assumption about joint bivariate normality led Winkelmann (2012) to the following copula-based representation of equation (1):

$$P\left(Y_{1,i}=0,Y_{2,i}=0|z_{1,i},z_{2,i}\right)=C\left(\Phi\left(-z_{1,i}^{T}\alpha\right),\Phi\left(-z_{2,i}^{T}\beta;\theta\right)\right). \tag{10}$$

Analogously, the remaining components of the joint probability function for $Y_1$ and $Y_2$ (see equations (2)-(4)) could be depicted as:

$$\begin{aligned}&P\left(Y_{1,i}=1,Y_{2,i}=0|z_{1,i},z_{2,i}\right)\\&\quad=P\left(\varepsilon_{2,i}\leq-z_{2,i}^{T}\beta\right)-P\left(\varepsilon_{1,i}\leq-z_{1,i}^{T}\alpha,\varepsilon_{2,i}\leq-z_{2,i}^{T}\beta\right)\\&\quad=\Phi\left(-z_{2,i}^{T}\beta\right)-C\left(\Phi\left(-z_{1,i}^{T}\alpha\right),\Phi\left(-z_{2,i}^{T}\beta\right);\theta\right)\end{aligned} \tag{11}$$

$$\begin{aligned}&P\left(Y_{1,i}=0,Y_{2,i}=1|z_{1,i},z_{2,i}\right)\\&\quad=P\left(\varepsilon_{1,i}\leq-z_{1,i}^{T}\alpha\right)-P\left(\varepsilon_{1,i}\leq-z_{1,i}^{T}\alpha,\varepsilon_{2,i}\leq-z_{2,i}^{T}\beta\right)\\&\quad=\Phi\left(-z_{1,i}^{T}\alpha\right)-C\left(\Phi\left(-z_{1,i}^{T}\alpha\right),\Phi\left(-z_{2,i}^{T}\beta\right);\theta\right)\end{aligned} \tag{12}$$

$$\begin{aligned}&P\left(Y_{1,i}=1,Y_{2,i}=1|z_{1,i},z_{2,i}\right)\\&\quad=1-P\left(\varepsilon_{1,i}\leq-z_{1,i}^{T}\alpha\right)-P\left(\varepsilon_{2,i}\leq-z_{2,i}^{T}\beta\right)+P\left(\varepsilon_{1,i}\leq-z_{1,i}^{T}\alpha,\varepsilon_{2,i}\leq-z_{2,i}^{T}\beta\right)\\&\quad=1-\Phi\left(-z_{1,i}^{T}\alpha\right)-\Phi\left(-z_{2,i}^{T}\beta\right)+C\left(\Phi\left(-z_{1,i}^{T}\alpha\right),\Phi\left(-z_{2,i}^{T}\beta\right);\theta\right).\end{aligned} \tag{13}$$

In case of the Gaussian copula that is given by $C\left(u,v,\theta\right)=\Phi_2\left(\Phi^{-1}\left(u\right),\Phi^{-1}\left(v\right);\theta\right)$, the copula-based model boils down to the standard bivariate probit model of Ashford and Sowden (1970) where the dependence parameter is simply a linear correlation between both marginals. On the other hand, other popular copula functions that are typically used in financial applications (i.e., the Clayton copula or the Gumbel copula) allow for capturing the lower or upper tail dependence, respectively, i.e. more pronounced dependence between the most extremely small realizations (lower tail dependence) or large realizations (upper tail dependence) of two underlying latent processes (check Table A1 in the Appendix for the exposition of some basic properties of the Gumbel, Clayton and Frank copulas). Another very popular copula (i.e. the Frank copula), has a relatively weak dependence in tails and strongest dependence at the middle of the distribution when compared to the Gaussian copula (see Cherubini *et al.*, 2004, p. 124-126, for a detailed exposition of the properties of these copulas).
The unknown parameters $\alpha$, $\beta$ and $\theta$ of the bivariate copula-based probit model can be easily estimated with the maximum likelihood method. The log likelihood function for $N$ observations can be derived from (10)-(13) as follows:

$$\begin{aligned}\ln L\left(\Theta\right) \quad=\quad &\textstyle\sum_{i=1}^{N}\left[\left(1-Y_{1,i}\right)\left(1-Y_{2,i}\right)\ln\left(P\left(Y_{1,i}=0,Y_{2,i}=0\right)\right)+\right.\\&+Y_{1,i}\left(1-Y_{2,i}\right)\ln\left(P\left(Y_{1,i}=1,Y_{2,i}=0\right)\right)+\\&+\left(1-Y_{1,i}\right)Y_{2,i}\ln\left(P\left(Y_{1,i}=0,Y_{2,i}=1\right)\right)+\\&\left.+Y_{1,i}Y_{2,i}\ln\left(P\left(Y_{1,i}=1,Y_{2,i}=1\right)\right)\right].\end{aligned} \tag{14}$$

## 2.2 Copula-based model for a binary and a continuous variable

### 2.2.1 General modeling setup

After becoming acquainted with the copula-based probit model, this section will focus on the copula-based model for a continuous and a binary variable. Additionally, we will also turn our attention to a time series framework in an effort to deal with the sequence $\{X_t, Y_t\}_{t=1}^{N}$, where $X_t$ denotes a continuous variable and $Y_t$ denotes a binary variable. We also assume that both variables are characterized by an autocorrelation and that some significant lead-lag relationships between the two variables can also exist. For the dichotomous variable $Y_t$, we assume an existence of the underlying continuous latent factor $Y_t^*$ as follows:

$$Y_t^* = z_t^T \alpha + \varepsilon_t \tag{15}$$

where $z_t$ denotes the $(k \times 1)$ vector of explanatory variables and $\alpha$ is the $(k \times 1)$ vector of corresponding parameters. The error term $\varepsilon_t$ has the following properties: $E(\varepsilon_t | z_t) = 0$ and $\varepsilon_t \sim I.I.D.\ N\left(0, \sigma^2\right)$. If we additionally assumed that $\varepsilon_t$ is normally distributed, for the univariate model of $Y_t$ we would obtain the well-known probit model, whereas if $\varepsilon_t$ had a logistic distribution, the univariate model would boil down to a standard logistic regression.
For a continuous $X_t$ we could hypothetically assume a lot of well-established parametric distribution families depending on the nature of the process under study. For example, for financial returns one could use the normal distribution, the skewed Student's $t$-distribution or the general error distribution, among many other choices. If we assumed that $X_t$ denotes financial durations (i.e. time spells between subsequent events defined by appropriate thinning of the tick-by-tick data), bid-ask spreads or trading volumes, we could apply the exponential, the Weibull, the Burr or the generalized gamma distribution. All of these were typically used in the context of Autoregressive Conditional Duration (ACD) models proposed by Russell and Engle (1998) for significantly autocorrelated variables defined on the positive domain.
In agreement with the framework of the bivariate copula-based probit model developed by Winkelmann (2012), the joint probability that $X_t$ would be lower than or equal to $x_t$ and that $Y_t$ would be equal to 0 can be derived as follows:

$$\begin{aligned} P(X_t \leq x_t, Y_t = 0 | z_t) &= P(X_t \leq x_t, Y_t^* \leq 0 | z_t) \\ &= P\left(X_t \leq x_t, \varepsilon_t \leq -z_t^T \alpha\right) \\ &= C\left(F_{X_t}(x_t), F_{\varepsilon_t}\left(-z_t^T \alpha\right); \theta\right) \end{aligned} \tag{16}$$

Analogously, the probability that the variable $X_t$ would be lower than or equal to $x_t$ and the variable $Y_t$ would be equal to 1 is as follows:

$$\begin{aligned} P(X_t \leq x_t, Y_t = 1 | z_t) &= P(X_t \leq x_t, Y_t^* > 0 | z_t) \\ &= P\left(X_t \leq x_t, \varepsilon_t > -z_t^T \alpha\right) \\ &= C\left(F_{X_t}(x_t), 1; \theta\right) - C\left(F_{X_t}(x_t), F_{\varepsilon_t}\left(-z_t^T \alpha\right); \theta\right) \\ &= F_{X_t}(x_t) - C\left(F_{X_t}(x_t), F_{\varepsilon_t}\left(-z_t^T \alpha\right); \theta\right) \end{aligned} \tag{17}$$

The bivariate density function for the pair $(X_t, Y_t)$ can be obtained from the two equations above as follows:

$$\begin{aligned} f(x_t, Y_t|z_t) &= [f(x_t, Y_t = 0|z_t)]^{1-Y_t} [f(x_t, Y_t = 1|z_t)]^{Y_t} = \\ &= \left[\frac{\partial C\left(F_{X_t}(x_t), F_{\varepsilon_t}\left(-z_t^T\alpha\right);\theta\right)}{\partial F_{X_t}(x_t)} \cdot \frac{\partial F_{X_t}(x_t)}{\partial x_t}\right]^{1-Y_t} \cdot \\ &\quad \cdot \left[\left(\frac{\partial F_{X_t}(x_t)}{\partial F_{X_t}(x_t)} - \frac{\partial C\left(F_{X_t}(x_t), F_{\varepsilon_t}\left(-z_t^T\alpha\right);\theta\right)}{\partial F_{X_t}(x_t)}\right) \cdot \frac{\partial F_{X_t}(x_t)}{\partial x_t}\right]^{Y_t} \\ &= \left(\frac{\partial C\left(F_{X_t}(x_t), F_{\varepsilon_t}\left(-z_t^T\alpha\right);\theta\right)}{\partial F_{X_t}(x_t)}\right)^{1-Y_t} \cdot \\ &\quad \cdot \left(1 - \frac{\partial C\left(F_{X_t}(x_t), F_{\varepsilon_t}\left(-z_t^T\alpha\right);\theta\right)}{\partial F_{X_t}(x_t)}\right)^{Y_t} \cdot f_{X_t}(x_t) \end{aligned} \tag{18}$$

The log-likelihood function can be derived from what is above as follows:

$$\begin{aligned} \ln L(\Theta) &= \sum_{t=1}^{N} (1 - Y_t) \ln\left(\frac{\partial C\left(F_{X_t}(x_t), F_{\varepsilon_t}\left(-z_t^T\alpha\right);\theta\right)}{\partial F_{X_t}(x_t)}\right) + \\ &\quad + \sum_{t=1}^{N} Y_t \ln\left(1 - \frac{\partial C\left(F_{X_t}(x_t), F_{\varepsilon_t}\left(-z_t^T\alpha\right);\theta\right)}{\partial F_{X_t}(x_t)}\right) + \sum_{t=1}^{N} \ln\left(f_{X_t}(x_t)\right) \end{aligned} \tag{19}$$

### 2.2.2 Marginal distribution of an autocorrelated binary variable

We assume that the $Y_t$ variable is characterized by a strong autocorrelation that can be partially attributed to an intraday seasonality pattern. This is a very common feature of market microstructure variables, such as: indicators of price changes, indicators of buy (sell) order submissions or indicators of hidden order placements. In order to account for this dynamics, we suggest the application of the Generalized Linear Autoregressive Moving Average (henceforth GLARMA) model of Shephard (1995), augmented with the diurnality pattern depicted by the Flexible Fourier Form (FFF). The background of a GLARMA model is the autologistic model given as follows:

$$g_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j Y_{t-j}, \tag{20}$$

where $g_t$ denotes a logistic link function, i.e. $g_t = \ln\left(\frac{P(Y_t=1|\Im_{t-1})}{P(Y_t=0|\Im_{t-1})}\right)$, where $\Im_{t-1}$ denotes the information set until $t-1$. The autologistic specification has been widely used in several medical studies (cf. Huffer and Wu, 1998; Gumpertz *et al.* 1997, among others). Rydberg and Shephard (2003) use this specification to capture the dynamics of price change indicators. However, they also show that this model is not

sufficiently flexible to account for a strong persistence in the probability of a price change and demands a very high parameterization (i.e., a high number of lags $p$).
The GLARMA model is intended to enhance the dynamic properties of an autologistic model. The intuition behind this specification is following. Rydberg and Shephard (2003) consider introducing a given function $H$ of lagged realizations of $Y_t$ into a logistic link specification:

$$g_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j H\left(Y_{t-j}\right). \tag{21}$$

Because the dependent variable in equation (21) is defined as a transformation of a conditional expectation $E\left(Y_t = 1|\Im_{t-1}\right) = F_{\log}\left(g_t|\Im_{t-1}\right)$ (where $F_{\log}\left(\cdot\right)$ denotes the cumulative distribution function of a logistic distribution); i.e. $F_{\log}^{-1}\left(P\left(Y_t = 1|\Im_{t-1}\right)\right) = \ln\left(\frac{P(Y_t=1|\Im_{t-1})}{1-P(Y_t=1|\Im_{t-1})}\right)$; the explanatory factors defined as past realizations of $Y_t$ should be transformed accordingly:

$$F_{\log}^{-1}\left(P\left(Y_t = 1|\Im_{t-1}\right)\right) = \alpha_0 + \sum_{j=1}^{p} \alpha_j F_{\log}^{-1}\left(Y_{t-j}\right), \tag{22}$$

Thus, the conditional expectation $E\left(Y_t = 1|\Im_{t-1}\right)$ and the lagged values of $Y_t$ would be measured on the same scale. Because the transformation:

$$F_{\log}^{-1}\left(P\left(Y_t = 1|\Im_{t-1}\right)\right) = \ln\left(\frac{P\left(Y_t = 1|\Im_{t-1}\right)}{1 - P\left(Y_t = 1|\Im_{t-1}\right)}\right)$$

is not valid for binary variables, Shephard (1995) proposes replacing it with the Taylor expansion around $P\left(Y_t = 1|\Im_{t-1}\right)$ as follows:

$$\begin{aligned} F_{\log}^{-1}\left(Y_t\right) &\approx F_{\log}^{-1}\left(P\left(Y_t = 1|\Im_{t-1}\right)\right) + \frac{\partial F_{\log}^{-1}\left(P\left(Y_t = 1|\Im_{t-1}\right)\right)}{\partial P\left(Y_t = 1|\Im_{t-1}\right)}\left(Y_t - P\left(Y_t = 1|\Im_{t-1}\right)\right) \\ &= F_{\log}^{-1}\left(P\left(Y_t = 1|\Im_{t-1}\right)\right) + \frac{Y_t - P\left(Y_t = 1|\Im_{t-1}\right)}{P\left(Y_t = 1|\Im_{t-1}\right)\left(1 - P\left(Y_t = 1|\Im_{t-1}\right)\right)} \end{aligned} \tag{23}$$

In the empirical application Rydberg and Shephard (2003) showed that the specification with a modified MA term has better properties:

$$c_t = \frac{Y_t - P\left(Y_t = 1|\Im_{t-1}\right)}{\sqrt{P\left(Y_t = 1|\Im_{t-1}\right)\left(1 - P\left(Y_t = 1|\Im_{t-1}\right)\right)}}.$$

As the denominator:

$$\sqrt{P\left(Y_t = 1|\Im_{t-1}\right)\left(1 - P\left(Y_t = 1|\Im_{t-1}\right)\right)}$$

is equal to the square root of the conditional variance of $Y_t$, the sequence $c_t$ has the conditional expectation of 0 and the conditional variance of 1. Therefore, it can be interpreted as 'standardized' values of $Y_t$.
Taking this into account, the GLARMA(p,q) model can be written as follows:

$$g_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j^g g_{t-j} + \sum_{j=1}^{q} \alpha_j^c c_{t-j}. \tag{24}$$

In order to account for the additive intraday seasonality in the dynamics of $Y_t$, the GLARMA specification can be enriched with the Flexible Fourier Form (FFF) components:

$$S(\nu, \tau) = \nu_0 \tau + \sum_{l=1}^{k} \left[\nu_{2l-1} \sin(2\pi l\tau) + \nu_{2l} \cos(2\pi l\tau)\right] \tag{25}$$

where $\tau$ denotes an intraday time standardized to $[0, 1]$ and $\nu$ denotes a $2k + 1$ parameter vector (cf. Andersen and Bollerslev, 1997).
To sum it all up, the explanatory variables that drive the dynamics of the latent factor $Y_t^* = z_t^T \alpha + \varepsilon_t$ (compare with section 2.2.1) and the corresponding parameters can be defined as:

$$z_t^T = [1, g_{t-1}, \cdots, g_{t-p}, c_{t-1}, \cdots, c_{t-q}, \tau, \sin(2\pi\tau), \\ , \cos(2\pi\tau), \cdots, \sin(2k\pi\tau), \cos(2k\pi\tau)]$$

$$\alpha_t^T = \left[\alpha_0, \alpha_1^g, \cdots, \alpha_p^g, \alpha_1^c, \cdots, \alpha_q^c, \nu_0, \cdots, \nu_{2k}\right]$$

The residuals from the GLARMA specification can easily be computed as follows:

$$\widehat{c}_t = \frac{Y_t - \widehat{P}(Y_t = 1|\Im_{t-1})}{\sqrt{\widehat{P}(Y_t = 1|\Im_{t-1})\left(1 - \widehat{P}(Y_t = 1|\Im_{t-1})\right)}}.$$

### 2.2.3 Marginal distribution of an autocorrelated continuous variable

In order to describe the dynamics inherent to a continuous variable, i.e. trading volume, we will apply the Autoregressive Conditional Duration (ACD) model of Engle and Russell (1998). Such a model was initially applied to a highly autocorrelated time series of durations between selected events. More recently the model was also used in order to describe other financial variables as transaction volumes (Manganelli, 2005) or bid-ask spread (Nolte, 2008). The ACD model can explicitly capture two specific features of financial variables measured at high frequency. First, it is designed for variables defined on a strictly positive domain. Second, it can flexibly describe processes that are strongly autocorrelated, often with a high degree of persistence.

Here we will apply the ACD model with the Burr distribution for the error term proposed by Grammig and Maurer (2000). The model for the variable $X_t$ is as follows:

$$x_t = \Psi_t \xi_t \tag{26}$$

where $\Psi_t = E\left(x_t|\Im_{t-1}\right)$ and $\Im_{t-1}$ denotes an information set up at the time point $t$ and $\xi_t$: *i.i.d.* $Burr\left(\kappa, \sigma^2\right)$ (where $\sigma^2$ and $\kappa$ denote parameters of the Burr distribution and $0 < \sigma^2 < \kappa$). The conditional expectation of the dependent variable $x_t$ is described as:

$$\Psi_t = \beta_0 + \sum_{i=1}^{p} \beta_{\Psi,i} \Psi_{t-i} + \sum_{j=1}^{q} \beta_{x,j} x_{t-j}. \tag{27}$$

In order to specify the conditional bivariate density of $\{X_t, Y_t\}$, it is crucial to derive the conditional density and the conditional survival function for the $X_t$ under given assumptions about the distribution of $\xi_t$. If the hazard and the survival function of an error term $\xi_t$ were denoted as $f_\xi(\cdot)$ and $F_\xi(\cdot)$, respectively, under the necessary assumption that $E\left(\xi_t\right) = 1$, the conditional hazard and the conditional survival function of $x_t = \Psi_t \xi_t$ can be given as:

$$f_x(x_t|\Im_{t-1}) = \frac{1}{\Phi_t} f_\xi \left(\frac{x_t}{\Phi_t}\right), \tag{28}$$

$$F_x(x_t|\Im_{t-1}) = F_\xi \left(\frac{x_t}{\Phi_t}\right), \tag{29}$$

where $\Phi_t = \frac{\Psi_t}{\mu}$ and $\mu$ is an expectation of a Burr-distributed random variable (see Appendix 2 for some basic properties of the Burr distribution).

## 2.3 Copula-based model verification with PIT

The appropriateness of the distributional assumptions underlying the copula-based model can be tested with the help of probability integral transforms (PIT) proposed by Diebold *et al.* (1998). This verification procedure has been widely used to verify the adequacy of the distribution choice and the quality of the conditional mean specification in numerous applications of the ACD models (e.g., Bauwens et al., 2004; Grammig & Mauer, 2000; Hautsch, 2004; Bień-Barkowska 2011). In this paper we show how to apply the PIT in order to check the goodness-of-fit of the dynamic bivariate copula-based model for mixed binary-continuous distribution. Because the continuous variable is described with the help of the ACD model with the Burr distribution, the PIT can quite naturally be used to test the adequacy of the distribution choice as well as the quality of the conditional mean specification of this marginal distribution. However, one should bear in mind, that even if the models for marginal distributions (i.e. GLARMA and ACD) fit the data well, this does not give enough information to make a judgment on the goodness-of-fit of the joint distribution

that accounts for the dependence between both marginal processes. Therefore, we suggest the method for testing the goodness-of-fit of the joint distribution with the help of the probability integral transforms for conditional distribution of the continuous variable given the realization of the binary outcome.
In short, the approach can be presented as following. If $\{f_t(x_t|\Im_{t-1})\}_1^m$ denotes a sequence of one-step-ahead density forecasts from the model of a continuous variable $X_t$ and $\{p_t(x_t|\Im_{t-1})\}_1^m$ is a sequence of conditional densities for the corresponding true data generating process, the model for the marginal density (but still conditional on $\Im_{t-1}$) will be correctly specified if the following equation holds true:

$$\{f_t(x_t|\Im_{t-1})\}_1^m = \{p_t(x_t|\Im_{t-1})\}_1^m. \tag{30}$$

Although the sequence $\{p_t(x_t|\Im_{t-1})\}_1^m$ cannot be observed, Diebold *et al.* (1998) show that under the null hypothesis (30), the sequence of density transforms $\{z_t\}_1^m$ corresponding to the sequence $\{x_t\}_1^m$ should be i.i.d. uniformly distributed on $(0,1)$:

$$z_t = \int_{-\infty}^{x_t} f_t(u)\,du, \quad z_t \sim i.i.d.\, U(0,1). \tag{31}$$

Accordingly, in order to compute the sequence $\{\widehat{z}_t\}_1^m$ we simply need to evaluate the cumulative distribution function at $x_t$. As the cumulative distribution function of the Burr distribution has a closed parametric form it is not difficult to calculate the sequence $\{\widehat{z}_t\}_1^m$ (see equation (29)). It should be stressed, however, that the application of this verification procedure is particularly suited for the time-series applications, because it checks the goodness-of-fit of the conditional distribution derived at different time points $t$, i.e. $\widehat{z}_t = \widehat{F}_{x_t}(x_t|\Im_{t-1})$. Accordingly, it tests both: (1) the appropriateness of $\xi_t$ distribution choice and (2) the quality of the conditional mean specification $\Psi_t = E(x_t = 1|\Im_{t-1})$. Diebold *et al.* (1998) suggest the visual inspection of the histogram as well as the autocorrelation function of $\widehat{z}_t$.
In order to check the goodness-of-fit of the joint bivariate model for $X_t$ and $Y_t$, and thus the adequacy of the chosen copula function, we propose to perform the above-mentioned procedure with respect to the conditional distributions: $f(x_t|Y_t = 0, \Im_{t-1})$ and $f(x_t|Y_t = 1, \Im_{t-1})$. If $X_t$ and $Y_t$ are not independent, conditional distribution of $X_t$ given the realization of $Y_t$ will depend on the specification of the copula function. The application of the PIT procedure could be easily performed in this case, because the binary variable can have only two values: 0 or 1. Accordingly, for the null hypothesis: $\{f_t(x_t|Y_t = 0, \Im_{t-1})\}_1^m = \{p_t(x_t|Y_t = 0, \Im_{t-1})\}_1^m$, we can use the probability integral transform estimate:

$$\begin{aligned} \widehat{z}_{t,Y_t=0} &= \int_{-\infty}^{x_t} \widehat{f}(u|Y_t = 0, \Im_{t-1})\,du = \frac{\widehat{P}(X_t \le x_t, Y_t = 0|z_t, \Im_{t-1})}{\widehat{P}(Y_t = 0|z_t)} \\ &= \frac{\widehat{C}\left(\widehat{F}_{x_t}(x_t|\Im_{t-1}), \widehat{F}_{\varepsilon_t}\left(-z_t^T\alpha\right); \theta\right)}{\widehat{F}_{\varepsilon_t}\left(-z_t^T\alpha\right)}, \end{aligned} \tag{32}$$

while for a null hypothesis: $\{f_t(x_t|Y_t=1,\Im_{t-1})\}_1^m = \{p_t(x_t|Y_t=1,\Im_{t-1})\}_1^m$, the corresponding PIT estimate can be derived:

$$\begin{aligned}\widehat{z}_{t,Y_t=1} &= \int_{-\infty}^{x_t}\widehat{f}(u|Y_t=1,\Im_{t-1})\,du = \frac{\widehat{P}(X_t\le x_t, Y_t=1|z_t,\Im_{t-1})}{\widehat{P}(Y_t=1|z_t)} \\ &= \frac{\widehat{C}\left(\widehat{F}_{x_t}(x_t|\Im_{t-1}),1\right)-\widehat{C}\left(\widehat{F}_{x_t}(x_t|\Im_{t-1}),\widehat{F}_{\varepsilon_t}\left(-z_t^T\alpha\right);\theta\right)}{1-\widehat{F}_{\varepsilon_t}\left(-z_t^T\alpha\right)}.\end{aligned} \tag{33}$$

# 3 Empirical example: the volume – return relationship

In this section we apply our copula-based model to study the relationship between trading volume and the indicator of a 'large' EUR/PLN rate change. We use trade data from the Reuters Dealing 3000 Spot Matching System. It is a fully automated (i.e. orders are automatically matched if they arrive to opposite market sides and if their prices agree) order-driven market where the interbank trading of the EUR/PLN currency pair takes place and which accounts for over 40% of the whole turnover (on the offshore market, i.e. between London banks, and in Poland). The EUR/PLN rate is quoted as an amount of Zlotys per one Euro. The transaction currency (base currency) is Euro. The smallest trade size is 1 million Euro. We used the data from the whole year 2007, whereas the variables of interest, i.e. trading volumes and price changes were measured on the 15-minute frequency. Trading of the EUR/PLN currency pair is characterized by a strong intraday seasonality, therefore we consider periods when trading intensity is relatively high, i.e. after 8:00 CET and before 18:00 CET.

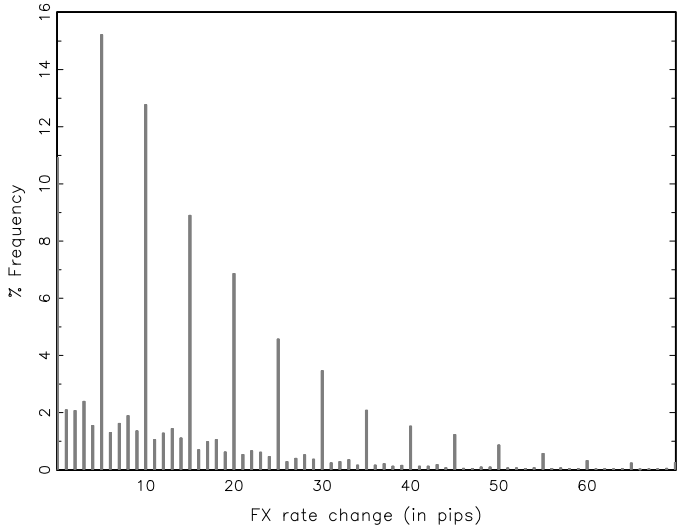
In our empirical study we are most interested in the measure of dependence between the volume and the probability that the EUR/PLN rate exceeds some arbitrarily given level. The boundaries set for such a price change can be different. For example, we may be interested in the value of a turnover that is necessary for the FX rate to move more than 20 pips up or down, which is in line with the fact that prices measured at a very high frequency usually stick to a grid of predefined values, i.e. multiples of 5 pips. The bivariate density constructed for the pair "volume - price change indicator" can be used for example to construct measures of market liquidity, which approximate the price impact of a given trade. The very popular question "how much volume does it take to move the price?" has a certain meaning in this context, since we could state how much volume does it take to move prices by more than 5 or 30 pips. The model can be further developed at a later stage in order to describe some measures of loss (price changes) that should not be incurred (i.e. if a trader sets a price change limit that should not be exceeded).

In the first step of our empirical exercise we checked for the presence of potential deterministic or stochastic trends in the volume variable. We performed the

Augmented Dickey-Fuller test and rejected the null about the unit root, thus the volume is stationary and the application of the ACD model is allowed. We also deseasonalized trading volumes. We assumed a multiplicative intraday seasonality factor $s_t$, such as $x_t = s_t\overline{x_t}$. The intraday seasonality factor $s_t$ has been estimated with the application of the kernel regression of $x_t$ on a time-of-day variable (we use quartic kernel with the bandwidth computed as $2.78sN^{-1/5}$, where $s$ is the standard deviation of the data. For details of the estimation procedure please refer to (Bauwens, Veredas 2004)). Further estimation has been performed on diurnally adjusted volumes $\overline{x_t}$.

In Figure 1 we depict the histogram of the EUR/PLN rate changes. We can see here that the price changes cluster on multiples of 5 pips and this feature seems to be quite striking. In line with this, we defined a dense grid of threshold price changes as $\Delta P_t \in \{0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$. For each of these values a corresponding time series of the indicator variables $Y_t$ has been computed.

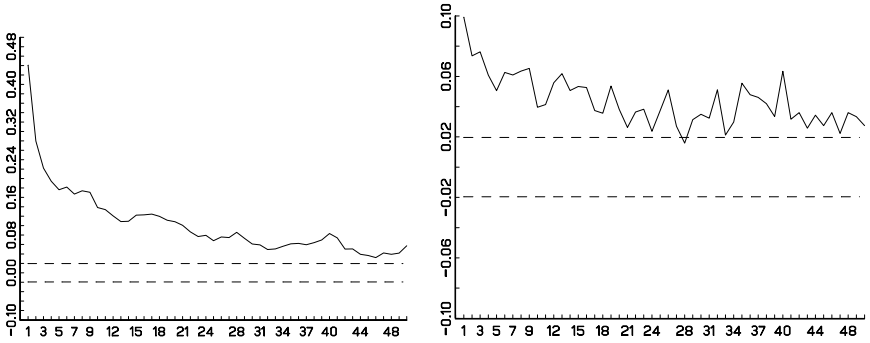
Figure 1: Histogram of the EUR/PLN rate changes in pips (15-minute intervals).




In Figure 2 we depict the autocorrelation function of the deseasonalized trading volumes and price change indicators for 30 pips (i.e., $Y_t = 1$, if $|\Delta P_t| > 30$ and $Y_t = 0$ if $|\Delta P_t| \leq 30$). We can see that both processes are not only significantly autocorrelated, but also very persistent. Therefore, an application of the autoregressive structure of the ACD or the GLARMA models seems to be very well justified.

For each of 11 defined threshold price changes we built three different models characterized by three different copulas: Clayton, Frank and Gumbel. In total we constructed 33 models. We have chosen the Clayton, Frank and Gumbel copulas

Figure 2: Autocorrelation function of the deseasonalized trading volume (left panel) and the binary indicator of a price change larger than 30 pips (right panel). Horizontal lines depict 95% confidence intervals.




as they are very popular in financial applications and allow for different dependency structure between the marginal distributions (see section 2.1). As far as the marginals are concerned, trading volume has been described using the Burr-ACD(2,2) model and the indicator variable $Y_t$ has been described with the GLARMA(1,1) model augmented with the FFF (for $k = 1$) in order to account for a possible diurnality in the probability of a 'arbitrarily large' price change. In Table 1 we depict exemplary estimates of the Gumbel copula model for trading volumes and price change indicators of the EUR/PLN rate movements larger than 30 pips. In Table 2 we present the log-likelihood values and dependence parameter estimates corresponding to all 33 models (estimation of the models has been performed using the maxlik library of the Gauss 7.0). Our results allow to formulate two interesting conclusions. First, the dependence parameter estimates increase with the size of a price change $|\Delta P_t|$. This regularity holds true for all three copula-based models and seems to be intuitively easy to understand. The larger the price movement, the more it relates to an increase in trading volume, thus the old well-known Wall Street adage "it needs volume to move prices" holds true with respect to the EUR/PLN market as well. Clearly, the probability that an FX price changes at least 40 pips demands much more trading volume (or it is codependent with much more trading volume) than the rise or fall of FX rate of 5 pips only. An increase in trading activity often signals the arrival of new information, which leads to increased price volatility (Easley, O'Hara, 1987; Blume *et al.*, 1994). However, the copula-based model does not assume any form of causal relationship between two marginal processes. The dependence parameter simply depicts an 'interdependence'; a measure of mutual association between both processes. Second, we observe that the increments in the size of the dependence parameter estimates are larger for price changes that are small in value (i.e. an increase between dependence coefficients corresponding to the price change of 5 and 10 pips, respectively, is much more pronounced than the difference in parameter

estimates corresponding to price change of 40 and 45 pips). This means that the difference in a EUR/PLN rate jump between 5 and 10 pips is associated with much more volume than the difference in a EUR/PLN rate jump between 40 and 45 pips. Thus, the relationship between volume and price changes is not linear. From Table 2 we also see that the model with the Gumbel copula wins our 'horse race' as far as some primary measures of the goodness-of-fit are concerned - in case of all eleven models it achieves the highest values of the log-likelihood.

Table 1: Estimation results of the Gumbel copula GLARMA(1,1)-ACD(2,2) model for the trading volume and the indicator of an EUR/PLN FX rate movement larger than 30 pips.

| GLARMA model | | ACD model | |
|---|---|---|---|
| parameter | estimate (p-value) | parameter | estimate (p-value) |
| $\alpha_0$ | 0.011 (0.759) | $\beta_0$ | 0.022 (0,000) |
| $\alpha_1^g$ | 0.985 (0.000) | $\beta_{\Psi,1}$ | 1.016 (0.000) |
| $\alpha_2^g$ | 0.085 (0.000) | $\beta_{\Psi,2}$ | -0.102 (0.002) |
| $\nu_0$ | -0.078 (0.252) | $\beta_{x,1}$ | 0.354 (0.000) |
| $\nu_1$ | -0.016 (0.018) | $\beta_{x,2}$ | -0.291 (0.000) |
| $\nu_2$ | -0.092 (0.000) | $\kappa$ | 1.855 (0.000) |
| | | $\sigma^2$ | 0.485 (0.000) |
| dependency parameter | | | |
| $\theta$ | 1.356 (0.0000) | | |

Table 2: Comparison of LogL values and the dependence parameter estimates for the bivariate copula-based GLARMA(1,1)-ACD(2,2) models.

| $\Delta P_t$ (pips) | 0 | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Gumbel | | | | | | | | | | | |
| $\widehat{\theta}$ | 1.135 | 1.215 | 1.273 | 1.279 | 1.317 | 1.346 | 1.356 | 1.364 | 1.379 | 1.401 | 1.359 |
| LogL | -1.134 | -1.343 | -1.445 | -1.418 | -1.304 | -1.176 | -1.126 | -1.035 | -1.004 | -0.946 | -0.909 |
| Clayton | | | | | | | | | | | |
| $\widehat{\theta}$ | 0.124 | 0.231 | 0.439 | 0.521 | 0.799 | 1.175 | 1.337 | 1.711 | 1.992 | 2.588 | 2.913 |
| LogL | -1.134 | -1.346 | -1.454 | -1.429 | -1.319 | -1.191 | -1.142 | -1.050 | -1.017 | -0.958 | -0.917 |
| Frank | | | | | | | | | | | |
| $\widehat{\theta}$ | 0.808 | 1.334 | 1.820 | 1.971 | 2.552 | 3.190 | 3.470 | 4.071 | 4.477 | 5.339 | 5.681 |
| LogL | -1.135 | -1.345 | -1.449 | -1.423 | -1.310 | -1.183 | -1.134 | -1.043 | -1.011 | -0.953 | -0.913 |

The choice of a suitable copula has serious implications for the obtained relationship between marginals and, obviously, for the quality of forecasting one variable given the value of the other. In order to present a more clear exposition of this problem, in Figure 3 we plot a conditional probability of observing an FX rate change larger

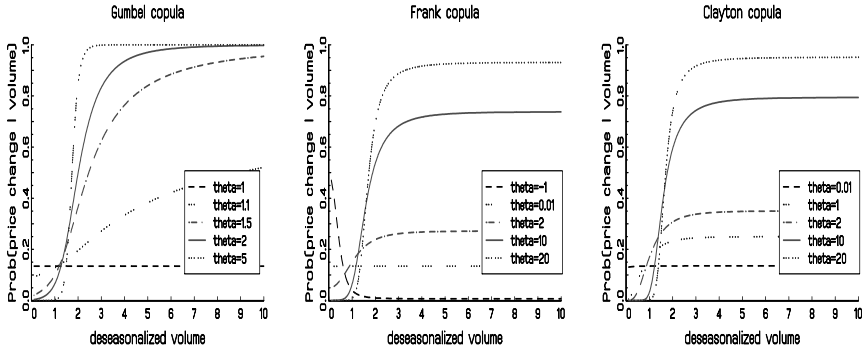
than 30 pips (within 15 minutes) given different realizations of the volume variable. The corresponding conditional probabilities can be derived as:

$$\begin{aligned} f(Y_t = 1|x_t, z_t) &= \frac{f\left(x_t, Y_t = 1|\Im_{t-1}, z_t\right)}{f\left(x_t|\Im_{t-1}\right)} \\ &= 1 - \frac{\partial C\left(F_{x_t}\left(x_t|\Im_{t-1}\right), F_{\varepsilon_t}\left(-z_t^T\alpha\right); \theta\right)}{\partial F_{\varepsilon_t}\left(x_t|\Im_{t-1}\right)}. \end{aligned} \tag{34}$$

In Figure 3 we present different probabilities of a price change (larger than 30 pips) given different values of the volume variable, different copulas (Gumbel, Frank, Clayton) and different dependence parameters (we set $\Psi_t$ equal to the unconditional expectation of $x_t$ and $F_{\varepsilon_t}\left(-z_t^T\alpha\right)$ equal to the unconditional probability of $Y_t = 0$). Some interesting observations can be found. An independence between the probability that $Y_t = 1$ and $X_t$ is achieved if $\theta = 1$ for the Gumbel copula, and $\theta$ approaching 0 for the Frank and the Clayton copula. In these cases we obtain the flat line indicating that the trading volume is not related to the probability of a price change. The Gumbel copula does not allow for a negative dependence between the marginals. The strength of positive dependence rises with the value of the dependence parameter. For $\theta = 1.5$ we get approximately linear relationship, whereas the higher the value of the coefficient, the more S-shaped relationships are obtained. For large values of the dependence parameter, the probability of $Y_t = 1$ is quickly approaching one. This is in line with a stronger dependence in the upper tail of the bivariate distribution of $Y_t^*$ and $X_t$ as predicted by the features of the Gumbel copula. As far as the Frank and Gumbel copulas are concerned, we see that the shapes of the obtained conditional probability curves are quite different. Unless the values of the dependence parameter are very large (i.e., for example for $\theta = 20$), the probability of $Y_t = 1$ does not approach 1, as in the Gumbel case (for the same range of volume values). This result can potentially be justified with a lack of upper tail dependence inherent to the Frank or the Clayton copula functions. From a very intuitive point of view, if the values of (deseasonalized) trading volumes are extremely large, this does not necessary imply that the values of $Y_t^*$ must also be extremely large. Hence, the probability of $Y_t = 1$ is not as high as in case of the Gumbel copula model.
Analogously, the conditional density function for the volume variable (given a realization of the binary outcome) can be derived as:

$$\begin{aligned} f(x_t|Y_t = 1, \Im_{t-1}) &= \frac{f\left(x_t, Y_t = 1|\Im_{t-1}, z_t\right)}{f\left(Y_t = 1|z_t\right)} \\ &= \left[f_{x_t}\left(x_t|\Im_{t-1}\right) - \frac{f_{x_t}(x_t|\Im_{t-1})\partial C\left(F_{x_t}(x_t|\Im_{t-1}), F_{\varepsilon_t}\left(-z_t^T\alpha\right); \theta\right)}{\partial F_{\varepsilon_t}(x_t|\Im_{t-1})}\right] \cdot \\ &\quad \cdot \left[1 - F_{\varepsilon_t}\left(-z_t^T\alpha\right)\right]^{-1}, \end{aligned} \tag{35}$$

Figure 3: Conditional probability that the EUR/PLN rate moves more than 30 pips (up or down) during 15 minutes.

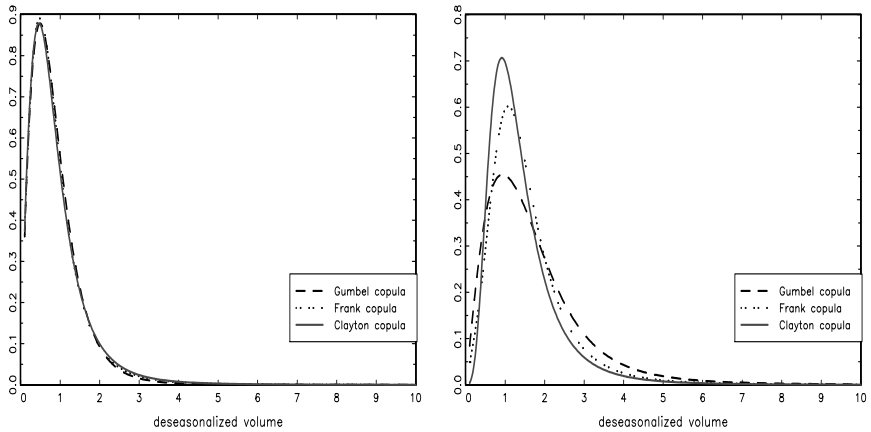



$$
\begin{aligned}
f(x_t|Y_t=0,\Im_{t-1}) &= \frac{f\left(x_t, Y_t=0|\Im_{t-1}, z_t\right)}{f\left(Y_t=0|z_t\right)} \\
&= \left[\frac{\partial C\left(F_{x_t}\left(x_t|\Im_{t-1}\right), F_{\varepsilon_t}\left(-z_t^T\alpha\right);\theta\right)}{\partial F_{\varepsilon_t}\left(x_t|\Im_{t-1}\right)} \cdot f_{x_t}\left(x_t|\Im_{t-1}\right)\right] \cdot \\
&\quad \cdot \left[F_{\varepsilon_t}\left(-z_t^T\alpha\right)\right]^{-1}.
\end{aligned}
\tag{36}
$$

We depict these two density functions in Figure 4 (for a price change larger than 30 pips). The values of the dependence parameters correspond to their estimates, i.e. $\theta = 1.356$ for the Gumbel copula, $\theta = 1.337$ for the Clayton copula and $\theta = 3.47$ for the Frank copula (we also set $\Psi_t$ equal to the unconditional expectation of $X_t$ and $F_{\varepsilon_t}\left(-z_t^T\alpha\right)$ equal to the unconditional probability of $Y_t = 0$). We see that all three obtained conditional density functions of the trading volume (given the realization of $Y_t$) have different shapes, whereas the discrepancies are striking given $Y_t = 1$. Expectations of all three distributions in the right panel are shifted to the right when compared with expected values of distributions depicted in the left panel, which is obvious because the 'large' price movements are associated with the higher trading volume. The GLARMA(1,1)-ACD(2,2) model with the Clayton copula generates the distribution that is most concentrated around its mode, once the model with the Gumbel copula allows for the most dispersed (flat) distribution.

In order to compare the goodness-of-fit of the three copula-based models, we calculated the PIT estimates with respect to: (1) marginal distribution of the trading volume ($\widehat{z}_t$, see equation (31)) (2) conditional distribution of the trading volume given $Y_t = 10$ ($\widehat{z}_{t,Y_t=0}$, see equation (32)), (3) conditional distribution of trading volume, given price $Y_t = 1$ ($\widehat{z}_{t,Y_t=1}$, see equation (33)). We later checked to

Figure 4: Conditional density function of trading volume given that no 'large' price movement is observed, i.e. price change bigger than 30 pips (left panel) and the conditional density function of trading volume given a 'large' price movement is observed (right panel).




see whether the $\widehat{z}_t$, $\widehat{z}_{t,Y_t=0}$ and $\widehat{z}_{t,Y_t=1}$ series are uniformly distributed on $(0,1)$. P-values for the $\chi^2$ test of uniformity for all the models estimated on the grid $\Delta P_t \in \{0,5,10,15,20,25,30,35,40,45,50\}$ are given in Table 3. The results confirm our previous finding (formulated on the basis of log-likelihood values), that the GLARMA(1,1)-ACD(2,2) model with the Gumbel copula allows for the best goodness-of-fit. We cannot reject the null that the $\widehat{z}_t$ is uniformly distributed on $(0,1)$, thus the marginal specification of the Burr-ACD(2,2) model for the trading volume seems to be appropriate. The estimated conditional density of $X_t$ given $Y_t = 0$ fits the true data generating process also quite well. We cannot reject the null of uniformity only in case of one model out of eleven estimated specifications (at a 1% significance level). However, the goodness-of-fit of the conditional density of $X_t$ given $Y_t = 1$ is a bit worse. With respect to this criterion, only eight of the eleven models prove to have an adequate specification (at a 1% significance level). All in all, the models with the Gumbel copula are doing considerably better than the other two.

In order to better understand the reasons for the good fit or lack of thereof, let us once more stick to the model where the hurdle for the price change has been set equal to 30 pips (i.e. $Y_t = 1$ if the price moves more than 30 pips and $Y_t = 0$ in other cases). For the model with the Gumbel copula we have obtained quite good results (we cannot reject the null of uniformity for $\widehat{z}_t$, $\widehat{z}_{t,Y_t=0}$ and $\widehat{z}_{t,Y_t=1}$). However, models with the Frank or Clayton copula seem not to fit at all. The reasons for this lack of fit can be understood from a visual inspection of PIT histograms (see Figure 5). We see that models with the Frank or Clayton copula are clearly misspecified. The lack of fit is particularly striking for the tails of the conditional distribution of $X_t$

Table 3: P-values of the $\chi^2$ test results for the uniformity of the PIT estimates.

| $\Delta P_t$ (pips) | 0 | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Gumbel | | | | | | | | | | | |
| $\widehat{z_t}$ | 0.039 | 0.045 | 0.274 | 0.243 | 0.171 | 0.285 | 0.131 | 0.146 | 0.222 | 0.161 | 0.135 |
| $\widehat{z}_{t,Y_t=0}$ | 0.000 | 0.350 | 0.495 | 0.073 | 0.016 | 0.285 | 0.148 | 0.537 | 0.537 | 0.207 | 0.150 |
| $\widehat{z}_{t,Y_t=1}$ | 0.025 | 0.017 | 0.058 | 0.121 | 0.049 | 0.070 | 0.055 | 0.043 | 0.006 | 0.009 | 0.005 |
| Clayton | | | | | | | | | | | |
| $\widehat{z_t}$ | 0.076 | 0.045 | 0.012 | 0.007 | 0.014 | 0.053 | 0.032 | 0.016 | 0.006 | 0.019 | 0.017 |
| $\widehat{z}_{t,Y_t=0}$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.007 | 0.030 | 0.153 |
| $\widehat{z}_{t,Y_t=1}$ | 0.062 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Frank | | | | | | | | | | | |
| $\widehat{z_t}$ | 0.066 | 0.035 | 0.029 | 0.025 | 0.064 | 0.082 | 0.034 | 0.044 | 0.085 | 0.081 | 0.011 |
| $\widehat{z}_{t,Y_t=0}$ | 0.000 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.047 | 0.098 | 0.078 | 0.166 |
| $\widehat{z}_{t,Y_t=1}$ | 0.012 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

given $Y_t = 1$ In this case, we can see that very 'small' values of trading volume (i.e. corresponding to the 0.01-quantile of these distributions), as well as 'large' values (i.e. corresponding to 0.95-quantile of the distribution) occur much more frequently than it is assumed by these theoretical conditional density functions. This observation agrees with Figure 4 where we showed that the Gumbel copula allows for a much more dispersed distribution of the volume variable (given $Y_t = 1$) when compared to the Frank or the Clayton copula. The latter two specifications forced a much more concentrated distribution, which does not fit the true data generating process.
Visual inspection of obtained histograms for PIT estimates is not enough to infer about the dynamic properties of the estimated model. To this end, we plotted the autocorrelation functions for the residuals obtained from both sub-models (ACD(2,2) and GLARMA(1,1)) as well as autocorrelation functions for three probability integral transform estimates (for $\widehat{z}_t$, $\widehat{z}_{t,Y_t=0}$ and $\widehat{z}_{t,Y_t=1}$) in Figures 6 and 7. The residuals are not significantly autocorrelated, which means that the dynamic specification of both marginal models is able to capture a very strong and persistent autocorrelation inherent to both: trading volumes and price change indicators (when compared to the ACF for raw data shown in Figure 2). The ACF for the probability integral transforms leads to the same conclusion.

Figure 5: Histograms of $\widehat{z}_t$ (left column), $\widehat{z}_{t,Y_t=0}$ (middle column) and $\widehat{z}_{t,Y_t=1}$ (right column). The horizontal lines are approximate 99% confidence intervals for the individual bin heights under the null that $\widehat{z}_t$ ($\widehat{z}_{t,Y_t=0}$ or $\widehat{z}_{t,Y_t=1}$, respectively) is U(0,1).

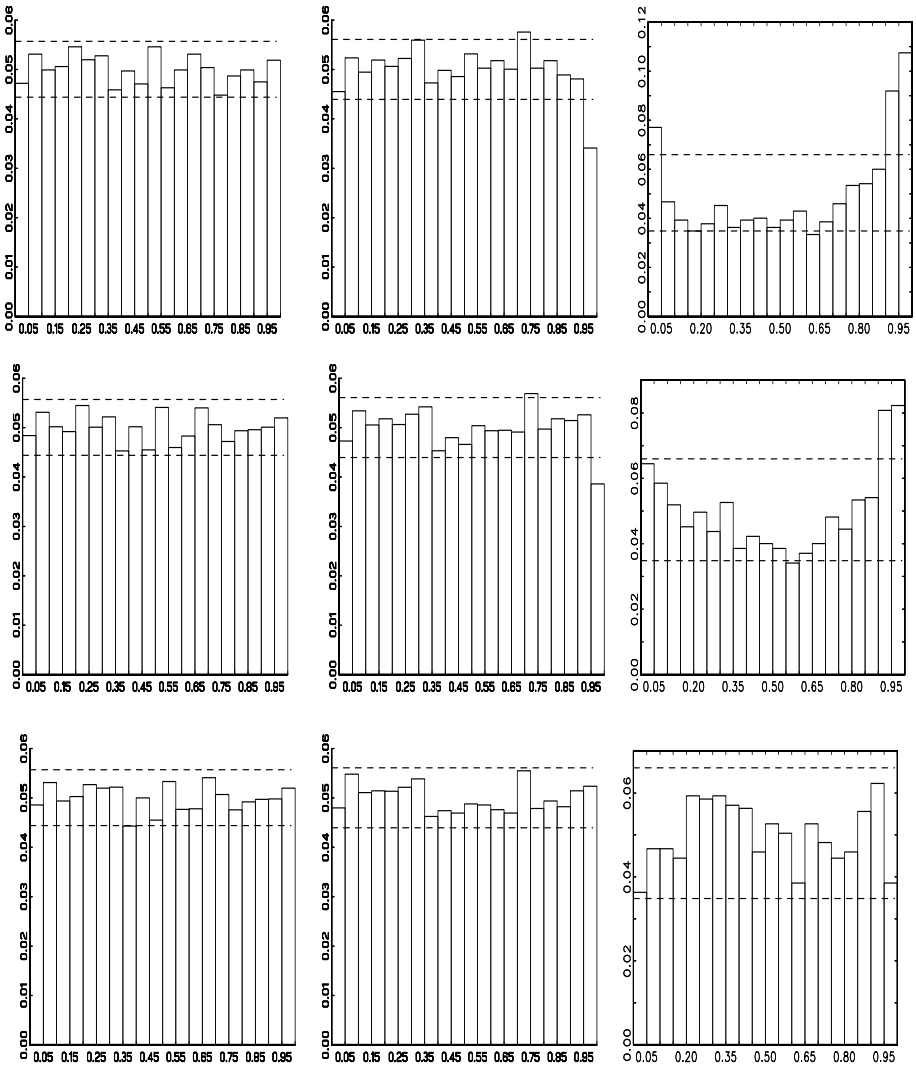


Figure 6: Autocorrelation function of the ACD model residuals (left panel) and the GLARMA model residuals (right panel). Horizontal lines depict 95% confidence intervals.

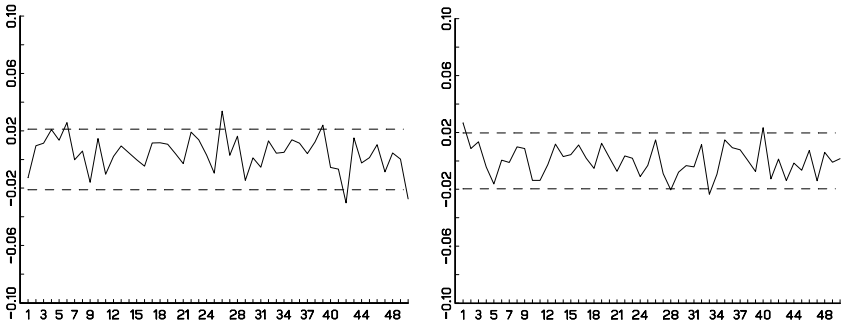
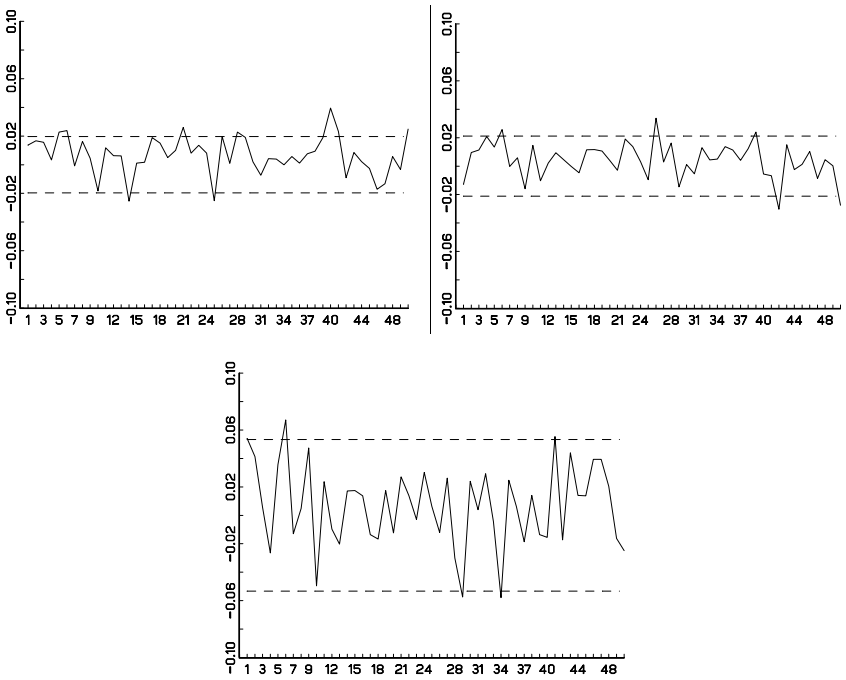



Figure 7: Autocorrelation function of $\widehat{z}_t$ (upper left panel), $\widehat{z}_{t,Y_t=0}$ (upper right panel) and $\widehat{z}_{t,Y_t=1}$ (bottom panel). Horizontal lines depict 95% confidence intervals.



# 4 Conclusions

In this paper we have developed the de Leon and Wu (2011) copula-based model for mixed binary-continuous distribution to a time-series setup. We have also proposed a method of testing the goodness-of-fit of this specification by means of the probability integral transforms defined for a marginal distribution of the continuous variable and for two conditional distributions: conditional distribution of continuous variable given $Y_t = 0$ and conditional distribution of continuous variable given $Y_t = 1$. If the marginals are not independent, the shape of these conditional distributions depends on the selected copula function, capturing the dependence between both processes.
The practical applications of the proposed methods have been presented with a short market microstructure study of the relationship between trading volume (continuous variable) and corresponding price change indicators (for selected grid of hurdle values). This econometric exercise proves that there is a significant dependence between both variables as predicted by the literature on market microstructure. Additionally, we show that this copula-based model is not indifferent to the choice of the copula function, which can be quite easily tested using the PIT procedures. In this analysis the best fitting copula is the Gumbel copula. As there is a common believe about the positive relationship between volatility and volume, it is the only copula among the considered ones that allows for the positive tail dependence. This study can be easily developed further by implementing different copulas (for example Gaussian copula, Student's t copula) or a kind of observation-driven dynamics governing the dependence parameter as suggested by Patton (2005b). Moreover, the PIT verification procedures can be easily applied to a copula-based model of a mixed discrete-continuous distribution where, instead of binary outcomes, the ordered outcomes of the discrete variable would be used. The copula-based model and its verification method can have several further applications. The bivariate volume-price change indicator model can be used as a starting point for constructing some price impact measures (liquidity measures). Moreover, it would be easy to construct a more complex model for the trading volume and price changes using the decomposition of the price change variable in the spirit of Rydberg and Shephard (2003).

# Acknowledgements

The author wants to thank the Thomson Reuters for providing the data from the Reuters Dealing 3000 Spot Matching system and Ryszard Doman for valuable comments on the previous version of this paper. Any remaining errors are the responsibility of the author. The views and opinions presented herein are those of the author and do not have to necessarily reflect those of the National Bank of Poland.

## References

[1] Andersen, T., Bollerslev, T., (1997), Heterogeneous Information Arrivals and Return Volatility Dynamics: Uncovering the Long Run in High Frequency Returns, *The Journal of Finance* 52, 975-1005.

[2] Ashford, J.R., Sowden, R.R., (1970), Multivariate Probit Analysis, *Biometrics* 26, 535-546.

[3] Bauwens, L., Veredas, D., (2004), The Stochastic Conditional Duration Model: A Latent Variable Model for the Analysis of Financial Durations, *Journal of Econometrics* 119, 381-412.

[4] Bhat, C.R., Sener, I.N., (2009), A Copula-based Closed Form Binary Logit Choice Model for Accommodating Spatial Correlation Across Observational Units, *Journal of Geographic Systems* 11, 243-272.

[5] Bień, K., Nolte, I., Pohlmeier, W. (2007), A Multivariate Integer Count Hurdle Model: Theory and Application to Exchange Rate Dynamics [in:] *Recent Developments in High Frequency Financial Econometrics*, [ed.:] Bauwens, L., Pohlmeier W. and Veredas D., Springer, Berlin, 31-48.

[6] Bień, K., Nolte, I., Pohlmeier, W. (2011), An Inflated Multivariate Integer Count Hurdle Model: An Application to Bid and Ask Quote Dynamics, *Journal of Applied Econometrics* 26, 549-714.

[7] Bień-Barkowska, K. (2011), Distribution Choice for the Asymmetric ACD Models, *Dynamic Econometric Models* 11, 55-72.

[8] Blume, L., Easley, D., O'Hara, M., (1994), Market Statistics and Technical Analysis: The Role of Volume, *Journal of Finance*, 49, 153-181.

[9] Cameron C., Li, T., Trivedi, P., Zimmer, D., (2004), Modelling the Differences in Counted Outcomes Using Bivariate Copula Models with Application to Mismeasured Counts, *Econometrics Journal* 7, 566-584.

[10] Cherubini, U., Luciano, E., Vecchiato, W., (2004), *Copula Methods in Finance*, John Wiley & Sons, Chichester.

[11] Copeland, T.E., (1976), A Model of Asset Trading under Assumption of Sequential Information Arrival, *Journal of Finance*, 31, 1149-1168.

[12] Diebold, F.X., Gunther, T.A., Tay, A.S., (1998), Evaluating Density Forecasts with Applications to Financial Risk Management, *International Economic Review* 39, 863-883.

[13] Doman, R., (2006), Measuring Conditional Dependence of Polish Financial Returns, *Dynamic Econometric Models* 7, 21-28, 59-67.

[14] Doman, R., (2008), Modeling Conditional Dependencies Between Polish Financial Returns with Markov-Switching Copula Models, *Dynamic Econometric Models* 8, 21-28.

[15] Doman, R., (2011), *Zastosowania kopuli w modelowaniu dynamiki zależności na rynkach finansowych*, Wydawnictwo UE, Poznań.

[16] Easley, D., Kiefer, N., O'Hara, M., (1997), One Day in the Life of a Very Common Stock, *Review of Financial Studies* 10, 805-835.

[17] Easley, D., O'Hara M., (1987), Price, Trade Size and Information in Securities Markets, *Journal of Financial Economics* 19, 69-90.

[18] Embrechts, P., (2009), Copulas: a Personal View, *Journal of Risk and Insurance* 76, 639-650.

[19] Engle, R. F., Russell, J. R., (1998), Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data, *Econometrica* 66, 1127-1162.

[20] Genest, C., Gendron, M., Bourdeau-Brien, M., (2009), The Advent of Copulas in Finance, *The European Journal of Finance* 15, 609-618.

[21] Grammig, J., Maurer, K., (2000), Non-monotonic Hazard Functions and the Autoregressive Conditional Duration Model, *Econometrics Journal* 3, 16-38.

[22] Gumpertz, M.L., Graham, J.M., Ristano, J.B., (1997), Autologistic Model of Spatial Pattern of Phytophthora Epidemic in Bell Paper: Effects of Soil Variables on Disease Presence, *Journal of Agricultural, Biological, and Environmental Statistics*, 2, 131-156.

[23] Gurgul, H., Syrek, R., (2006), Archimedean Copulas for Price-Volume Dependencies of DAX companies, *Systems Science* 32, 63-90.

[24] Harris, L.E., (1994), Minimum Price Variations, Discrete Bid-ask Spreads and Quotation Sizes, *Review of Financial Studies* 7, 149-178.

[25] Hautsch, N., (2004), *Modelling Irregularly Spaced Financial Data*, Springer, Berlin.

[26] Huffer, F., Wu, H., (1998), Markov Chain Monte Carlo for Autologistic Regression Models with Application to the Distribution of Plant Species, *Biometrics*, 54(2), 509-524.

[27] Jennings, R.H., Starks, L.T., Fellingham, J.C., (1981), An Equilibrium Model of Asset Trading with Sequential Information Arrival, *Journal of Finance*, 36, 143-161.

[28] Karpoff, J.M., (1987), What Drives the Volume-Volatility Relationship on Euronext Paris?, *The Journal of Financial and Quantitative Analysis*, 22(1), 109-126.

[29] de Leon, A.R., Wu, B., (2011), Copula-based Regression Models for a Bivariate Mixed Discrete and Continuous Outcome, *Statistics in Medicine* 30, 175-185.

[30] Louhichi, W., (2011), The Relation between Price Changes and Trading Volume: a Survey, *International Review of Financial Analysis*, 200-206.

[31] Patton, A., (2005a) Estimation of Multivariate Models for Time Series of Possibly Different Lengths, *Journal of Applied Econometrics*, 21 (2), 147-173.

[32] Patton, A., (2005b), Modelling Asymmetric Exchange Rate Dependence, *International Economic Review*, 2, 527-556.

[33] Patton, A., (2009), Copula-based Models for Financial Time Series, *Handbook of Financial Time Series*, 5, 767-785.

[34] Manganelli, S., (2005), Duration, Volume and Volatility Impact of Trades, *Journal of Financial Markets* 8, 377-399.

[35] Mikosch, T., (2006), Copulas: Tales and Facts, *Extremes* 9, 3-20, Discussion, 21-53. Rejoinder 55-62.

[36] Nelsen, R.B., (2006), *An Introduction to Copulas*, 2nd edition, New York: Springer.

[37] Nikoloulopoulos, A., K., Karlis, D., (2008), Multivariate Logit Model with an Application to Dental Data, *Statistics in Medicine* 27, 6393-6406.

[38] Nolte, I., (2008), Modeling a Multivariate Transaction Process, *Journal of Financial Econometrics* 6, 143-170.

[39] Rydberg, T.H., Shephard, N., (2003), Dynamics of Trade-by-Trade Price Movements: Decomposition and Models, *Journal of Financial Econometrics* 1, 2-25.

[40] Shephard, N., (1995), Generalized Linear Autoregressions, *unpublished paper*, Nuffield College, Oxford.

[41] Sklar, A., (1959), *Fonctions de répartition à n dimensions et leurs marges*, Public Institute of Statistics at the University of Paris, 8, 229-231.

[42] Trivedi, P.K., Zimmer, D.M., (2007), *Copula Modelling: An Introduction for Practitioners*, now Publishers Inc., Hanower.

[43] Winkelmann, R., (2012), Copula Bivariate Probit Models: with an Application to Medical Expenditures, *Health Economics* 21, 1444-1455.

[44] Zimmer, D.M., Trivedi, P.K., (2006), Using Trivariate Copulas to Model Sample Selection and Treatment Effects, *Journal of Business and Economics Statistics* 24, 63-76.

# Appendix 1

Table 4: Selected properties of Frank, Clayton and Gumbel copulas. Formulae for the Frank and the Gumbel copula functions are taken from Cherubini *et al.* (2004) p. 124 and formulae for the Clayton copula function from Trivedi and Zimmer (2005), p. 16.

| | |
|---|---|
| Frank copula: | |
| Copula function: | $C(u,v) = -\frac{1}{\theta}\ln\left(1+\frac{(\exp(-\theta u)-1)(\exp(-\theta v)-1)}{\exp(-\theta)-1}\right)$ |
| First order partial derivative: | $\frac{\partial C(u,v)}{\partial u} = \frac{\exp(-\theta u)\exp(-(\theta v)-1)}{\exp(-\theta)-1+(\exp(-\theta u)-1)(\exp(-\theta v)-1)}$ |
| Range of $\theta$: | $(-\infty,0)\cup(0,\infty)$ |
| Clayton copula: | |
| Copula function: | $C(u,v) = \left(u^{-\theta}+v^{-\theta}-1\right)^{-\frac{1}{\theta}}$ |
| First order partial derivative: | $\frac{\partial C(u,v)}{\partial u} = u^{-\theta-1}\left(u^{-\theta}+v^{-\theta}-1\right)^{-1-\frac{1}{\theta}}$ |
| Range of $\theta$: | $(0,\infty)$ |
| Gumbel copula: | |
| Copula function: | $C(u,v) = \exp\left\{-\left[(-\ln(u))^{\theta}+(-\ln(v))^{\theta}\right]^{\frac{1}{\theta}}\right\}$ |
| First order partial derivative: | $\frac{\partial C(u,v)}{\partial u} = \frac{C(u,v)}{u}\left(\frac{\ln(u)}{\ln(C(u,v))}\right)^{\theta-1}$ |
| Range of $\theta$: | $[1,\infty)$ |

# Appendix 2

*Burr distribution* with parameters $\kappa > 0$, $\sigma^2 > 0$ and scale parameter is set to 1:
Survival: $S(\varepsilon) = \left(1+\sigma^2\varepsilon^{\kappa}\right)^{-\frac{1}{\sigma^2}}$
Density: $f(\varepsilon) = \frac{\kappa\varepsilon^{\kappa-1}}{\left(1+\sigma^2\varepsilon^{\kappa}\right)^{1+\frac{1}{\sigma^2}}}$
Hazard: $h(\varepsilon) = \frac{\kappa\varepsilon^{\kappa-1}}{1+\sigma^2\varepsilon^{\kappa}}$
Expectation: $\mu = \frac{\Gamma\left(1+\kappa^{-1}\right)\Gamma\left(\frac{1}{\sigma^2}-\kappa^{-1}\right)}{\sigma^{2\left(1+\kappa^{-1}\right)}\Gamma\left(\frac{1}{\sigma^2}+1\right)}$ if $\kappa > \sigma^2$