Comovements in Trading activity: A Multivariate Autoregressive Model of Time Series Count Data Using Copulas [*]

Andréas Heinen[†] **Erick Rengifo**[‡]

November 2003

JEL Classification codes: C32, C35, G10.

Keywords: Continuousation; Factor model; Market microstructure.

Abstract

This paper introduces the Multivariate Autoregressive Conditional Poisson model to deal with issues of discreteness, overdispersion and both auto- and cross-correlation, arising with multivariate counts. We model counts with a double Poisson and assume that conditionally on past observations the means follow a Vector Autoregression. We resort to copulas to introduce contemporaneous correlation. We advocate the use of our model as a feasible alternative to multivariate duration models and apply it to the study of sector and stock specific news related to the comovements in the number of trades per unit of time of the most important US department stores traded on the New York Stock Exchange. We show that the market leaders inside an specific sector, in terms of more sectorial information conveyed by their trades, are related to their size measured by their market capitalization.

[*] The authors would like to thank Luc Bauwens, Philippe Lambert and David Veredas for helpful discussions and suggestions. The usual disclaimers apply.

[†] Corresponding author. University of California, San Diego and Center of Operations Research and Econometrics, Catholic University of Louvain, 34 Voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium, e-mail: heinen@core.ucl.ac.be

[‡] Center of Operations Research and Econometrics, Catholic University of Louvain.

1 Introduction

In empirical studies of market microstructure, the Autoregressive Conditional Duration (ACD) model, introduced by Engle & Russell (1998), has been used widely to test theories with tick-by-tick data in a univariate framework. This model is designed specifically to deal with the irregularly-spaced nature of financial time series of durations. However, extensions to more than one series have proven to be very difficult. The difficulty comes from the very nature of the data, which are by definition not aligned in time, i.e. the times at which an event of any type happens are random. Engle & Lunde (2003) suggest a model for the bivariate case, but the specification is not symmetric in the two processes. They analyse jointly the duration between successive trades and the duration between a trade and the next quote arrival. This is done in the framework of competing risks. **?** model bivariate durations using a univariate model for the duration between the arrival of all events, regardless of their type, and a probit specification which determines the type of event that occurred. These models become intractable when the number of series is greater than two.

In this paper we suggest working with counts instead of durations, especially when there are more than two series. Any duration series can easily be made into a series of counts by choosing an appropriate interval, which depends on the applications at hand, and counting the number of events that occur every period. The loss of information from considering counts is largely compensated for by the possibility of flexibly modelling interactions between several series. Moreover, most applications involve relatively rare events, which makes the use of the normal distribution questionable. Thus, modelling this type of series requires one to deal explicitly with the discreteness of the data as well as its time series properties and correlation. Neglecting either of these characteristics would lead to potentially serious misspecification.

We introduce a new multivariate model for time series count data and apply it to the study of comovements in trading activity of several stocks belonging to the same sector.

The Multivariate Autoregressive Conditional Double Poisson model (MDACP) makes it possible to deal with issues of discreteness, over- and underdispersion (variance greater or smaller than the mean) and both cross- and serial correlation. This paper constitutes a multivariate extension to the univariate time series of counts model developed in Heinen (2003). We take a fully parametric approach where the counts have the Double Poisson distribution proposed by Efron (1986) and their mean, conditional on past observations, is autoregressive. In order to introduce contemporaneous correlation we use a multivariate normal copula. This copula is very flexible, since it makes it possible to accommodate both positive and negative correlation, something that is impossible in most existing multivariate count distributions. The models are estimated using maximum likelihood, which makes the usual tests available. In this framework autocorrelation can be tested with a straightforward likelihood ratio test, whose simplicity is in sharp contrast with test procedures in the latent variable time series count model of Zeger (1988). We apply a two-stage estimation procedure developed in Patton (2002), which consists in estimating first the marginal models and then the copula, taking the parameters of the marginal models as given. This considerably eases estimation of the model. In order to capture the dynamic interactions between the series we model the conditional mean as a VARMA-type structure, focusing our attention to the (1,1) case, motivated largely by considerations of parsimony.

It is well documented in market microstructure that the trading process conveys information. According to Admati & Pfleiderer (1988) and **?** frequent trading implies that news is arriving to the market. Thus a higher number of trades in a given time interval is a signal for the arrival of news. Information in trading activity can be either stock-specific or sector-wide. How much sector-specific information a stock's trading activity contains has important implications from the point of view of identifying sectorial leaders. **?** study this question using a duration-based approach for pairs of assets. As a feasible alternative to multivariate duration models, we apply the MDACP to the study of sector and stock specific news of the most important US department stores traded on the New York Stock Exchange during the year 1999. We model the dynamics of the number of transactions of all stocks simultaneously using an intuitive and parsimonious factor structure, whereby the conditional mean of every series depends on one lag of itself, one lag of the count and one factor of the cross-section of lagged counts. We show that the assets that contain more sector information correspond to assets with larger market capitalizations. This is in contradiction with the findings of **?**, who find that it is most frequently traded stocks that contain most sector-specific information.

The paper is organised as follows. Section 2 introduces the Multivariate Double Autoregressive Poisson Model, shows how we use copulas in the present context, and describes the conditional mean and the marginal distribution of the model. Section 3 presents the empirical application. Section 4 concludes.

2 The Multivariate Double Autoregressive Poisson

In this section we discuss the way in which we use copulas and continusousation to generate a multivariate discrete distribution. Then we present the conditional distribution and the conditional mean of the Multivariate Double Autoregressive Poisson. Next we summarise the features of our model and establish its properties.

2.1 A General Multivariate Model Using Copulas

In order to generate richer patterns of contemporaneous cross-correlation, we resort to copulas. Copulas provide a very general way of introducing dependence among several series with known marginals. Copula theory goes back to the work of Sklar (1959), who showed that a joint distribution can be decomposed into its K marginal distributions and a copula, that describes the dependence between the variables. This theorem provides an easy way to form valid multivariate distributions from known marginals that need not be necessarily of the same distribution, i.e. it is possible to use normal, student or any other marginals, combine them with a copula and get a suitable joint distribution, which reflects the kind of dependence present in the series. A more detailed account of copulas can be found in Joe (1997) and in Nelsen (1999).

Let $H(y_1, \ldots, y_K)$ be a continuous K-variate cumulative distribution function with univariate margins $F_i(y_i)$, $i = 1, \ldots, K$, where $F_i(y_i) = H(\infty, \ldots, y_i, \ldots, \infty)$. According to Sklar (1959), there exists a function C, called copula, mapping $[0,1]^K$ into $[0,1]$, such that:

$$H(y_1, \ldots, y_K) = C(F_1(y_1), \ldots, F_K(y_K)) \ . \tag{2.1}$$

The joint density function is given by the product of the marginals and the copula density:

$$\frac{\partial H(y_1, \ldots, y_K)}{\partial y_1 \ldots \partial y_K} = \prod_{i=1}^{K} f_i(y_i) \frac{\partial C(F_1(y_1), \ldots, F_K(y_K))}{\partial F_1(y_1) \ldots \partial F_K(y_K)} \ . \tag{2.2}$$

With this we can define the copula of a multivariate distribution with Uniform $[0,1]$ margins as:

$$C(z_1, \ldots, z_K) = H(F_1^{-1}(z_1), \ldots, F_K^{-1}(z_K)) \ , \tag{2.3}$$

where $z_i = F_i(y_i)$, for $i = 1, \ldots, K$.
As we can see with the use of the copulas we are able to map the univariate marginal distributions of K random variables, each supported in the $[0,1]$ interval, to their K-variate distribution, supported on $[0,1]^K$, something that holds, no matter what the dependence among the variables is (including if there is none).

Most of the literature on copulas is concerned with the bivariate case. However, we are trying to specify a general type of multivariate count model, not limited to the bivariate case. Whereas there are many alternative formulations in the bivariate case, the number of possibilities for multi-parameter multivariate copulas is rather limited. We choose to work with the most intuitive one, which is arguably the Gaussian copula, obtained by the inversion method (based on Sklar (1959)). This is a K-dimensional copula such that:

$$C(z_1, \ldots, z_K; \Sigma) = \Phi^K(\Phi^{-1}(z_1), \ldots, \Phi^{-1}(z_K); \Sigma) \ , \tag{2.4}$$

and its density is given by,

$$c(z_1, \ldots, z_K; \Sigma) = \mid \Sigma \mid^{-1/2} \exp\left(\frac{1}{2}(q'(I_K - \Sigma^{-1})q\right) \, , \tag{2.5}$$

where Φ^K is the K-dimensional standard normal multivariate distribution function, Φ^{-1} is the inverse of the standard univariate normal distribution function and $q = (q_1, \ldots, q_K)'$ with normal scores $q_i = \Phi^{-1}(z_i)$, $i = 1, \ldots, K$. Furthermore, it can be seen that if $Y_1, \ldots, Y_K$ are mutually independent, the matrix Σ is equal to the identity matrix I_K and the copula density is then equal to 1.

In the present paper we are using a discrete marginal, the Double Poisson, whose support is the set of integers, instead of continuous ones, which are defined for real values. If the marginal distributions functions are all continuous then C is unique. However when the marginal distributions are discrete, this is no longer the case and the copula is only uniquely identified on $\overset{K}{\underset{i=1}{\otimes}} Range(F_i)$, a K-dimensional set, which is the Cartesian product of the range of all marginals. Moreover, a crucial assumption, which underlies the use of copulas, is that the marginal models are well specified and that the probability integral transformation (PIT) of the variables under their marginal distribution is distributed uniformly on the $[0,1]$ interval. The problem with discrete distributions is that the Probability Integral Transformation Theorem (PITT) of Fisher (1932) does not apply, and the uniformity assumption does not hold, regardless of the quality of the specification of the marginal model. The PITT states that if Y is a continuous variable, with cumulative distribution F, then

$$Z = F(Y)$$

is uniformly distributed on $[0,1]$.

We use a continuousation argument to overcome these difficulties and apply copulas with discrete marginals. The main idea of continuousation is to create a new random variable Y^* by adding to a discrete variable Y a continuous variable U valued in $[0,1]$, independent of Y, with a strictly increasing cdf, sharing no parameter with the distribution of Y, such as the Uniform $[0,1]$ for instance:

$$Y^* = Y + (U - 1) \ .$$

Continuousation does not alter the concordance between pairs of random variables; intuitively, two random variables Y_1 and Y_2 are concordant, if large values of Y_1 are associated with large values of Y_2. Concordance is an important concept, since it underlies many measures of association between random variables, such as Kendall's tau for instance. It is easy to see that continuousation does not affect concordance, since $Y_1^* > Y_2^* \iff Y_1 > Y_2$. Using continuousation, we state a discrete analog of the PITT. If Y is a discrete random variable with domain χ, in $\mathbf{N}$, such that $f_y = P(Y = y), \ y \in \chi$, continuoused by U, then

$$Z^* = F^*(Y^*) = F^*(Y + (U - 1)) = F\left([Y^*]\right) + f_{[Y^*]+1}U = F(Y - 1) + f_y U$$

is uniformly distributed on $[0,1]$, and $[Y]$ denotes the integer part of Y. Then, we will use Z^* as an argument in the copula, instead of Z, since, provided that the marginal model is well specified, this will ensure that the conditions for the use of the copula are met. This amounts to replacing the expression above for $z_{i,j}$ and $z_{i,j}^0$ by their continuoused versions:

$$z_{i,j} = F^*(Y_{i,j}^*) = F(Y_{i,j} - 1) + f(Y_{i,j}) * U_{i,j}$$

and

$$z_{i,j}^0 = F^{*,0}(Y_{i,j}^*) = F^0(Y_{i,j} - 1) + f^0(Y_{i,j}) * U_{i,j}$$

where $Y_{i,j}^*$ are the continuoused version of the original data $Y_{i,j}$:

$$Y_{i,j}^* = Y_{i,j} + (U_{i,j} - 1)$$

F^*, F and f are, respectively the continuoused c.d.f. and the p.d.f. of Y, and $F^{*,0}$, F^0 and f^0 are the same for the unconditional distribution. $U_{i,j}$ are independent uniform random variables on $[0, 1]$.

In this paper, we will use the continuoused version of the probability integral transformation in order to test the correct specification of the marginal models. If the marginal models are well-specified, then Z^*, the PIT of the series under the estimated distribution and after continuousation, is uniformly distributed. We will also use Z^* as an argument in the copula, since, provided that the marginal model is well specified, this will ensure that the conditions for use of a copula are met.

One remark needs to be made concerning the use of continuousation in the present context. In a sense the lack of identifiability of the copula outside of the range of the cumulative distribution of the marginal model is less acute in the time-varying distribution case, as the number of points at which the copula is observed increases, relative to the static case. In order to illustrate this point, let's consider the case of Bernoulli variables, which are in a sense, the 'most discrete' possible random variables. The problem we describe is the same with the Poisson or the Double Poisson distribution. We consider the Bernoulli variables Y_i, for $i = 1, \ldots, K$, whose cumulative density functions F_i can only take 3 possible values:

$$Z_{i,t} = F_i(Y_{i,t}) = \begin{cases} 0 & \text{if } y_{i,t} \leq 0 \\ p_i & \text{if } 0 < y_{i,t} < 1 \\ 1 & \text{otherwise} \end{cases}$$

The copula is then only identified on the set $S = \underset{i=1}{\overset{K}{\otimes}} \{0, p_i, 1\}$. Therefore it is impossible to distinguish two copulas which have the same values on S, but are different on $[0, 1]^n \bigcap \overline{S}$. In the case where the distributions are time-varying, we have:

$$F_{i,t}(Y_{i,t}) = \begin{cases} 0 & \text{if } y_{i,t} \leq 0 \\ p_{i,t} & \text{if } 0 < y_{i,t} < 1 \\ 1 & \text{otherwise} \end{cases}$$

The copula is now identified on the set $\bigcup_{t=1}^{T} \overset{K}{\underset{i=1}{\otimes}} \{0, p_{i,t}, 1\}$, which is obviously much larger a set than S. Nonetheless, it remains true in the time-varying case, that the non-corrected Z-statistic is not uniformly distributed, which, alone, justifies the use of continuousation.

2.2 The conditional distribution and the conditional mean

In order to extend the Autoregressive Conditional Double Poisson model to a $(K \times 1)$ vector of counts N_t, we build a VARMA-type system for the conditional mean. In a first step, we assume that conditionally on the past, the different series are uncorrelated. This means that there is no contemporaneous correlation and that all the dependence between the series is assumed to be captured by the conditional mean. Even though the Poisson distribution with autoregressive means is the natural starting point for counts, one of its characteristics is that the mean is equal to the variance, property referred to as equidispersion. However, by modelling the mean as an autoregressive process, we generate overdispersion in even the simple Poisson case.

In some cases one might want to break the link between overdispersion and serial correlation. It is quite probable that the overdispersion in the data is not attributable solely to the autocorrelation, but also to other factors, for instance unobserved heterogeneity. It is also imaginable that the amount of overdispersion in the data is less than the overdispersion resulting from the autocorrelation, in which case an underdispersed marginal distribution might be appropriate. In order to account for these possibilities we consider the double Poisson distribution introduced by Efron (1986) in the regression context, which is a natural extension of the Poisson model and allows one to break the equality between conditional mean and variance. The advantages of using this distribution are that it can be both under-

and overdispersed, depending on whether ϕ is larger or smaller than 1. We write the model as:

$$N_{i,t}|\mathcal{F}_{t-1} \sim DP(\mu_{i,t}, \phi_i) \ , \ \forall i = 1, \ldots, K. \tag{2.6}$$

where $\mathcal{F}_{t-1}$ designates the past of all series in the system up to time $t-1$[1]. With the double Poisson, the conditional variance is equal to:

$$V[N_{i,t}|\mathcal{F}_{t-1}] = \sigma^2_{i,t} = \frac{\mu_{i,t}}{\phi_i} \tag{2.7}$$

The coefficient ϕ_i of the conditional distribution will be a parameter of interest, as values different from 1 will represent departures from the Poisson distribution. The Double Poisson generalises the Poisson in the sense of allowing more flexible dispersion patterns.

The conditional means μ_t are assumed to follow a VARMA-type process:

$$E[N_t|\mathcal{F}_{t-1}] = \mu_t = \omega + \sum_{j=1}^{p} A_j N_{t-j} + \sum_{j=1}^{q} B_j \mu_{t-j} \tag{2.8}$$

For reasons of simplicity, in most of the ensuing discussion, we will focus on the most common $(1,1)$ case and for notational simplicity, we will denote $A = \sum_{j=1}^{p} A_j$ and $B = \sum_{j=1}^{q} B_j$ and drop the index whenever there is no ambiguity.

In most empirical applications, most especially when the number of series analysed jointly is large, some additional restrictions might have be be imposed on A and B.

In systems with large K, which could be found, for instance when analysing a large group of stocks like the constituents of an index, the full approach would not be feasible, as the number of parameters would get too large. If we assume that A and B are of full rank, the number of parameters that has to be estimated in this model would be $2K^2 + K$. In situations where this is not an option, we propose to impose some additional structure on

[1] It is shown in Efron (1986) (Fact 2) that the mean of the Double Poisson is μ and that the variance is approximately equal to $\frac{\mu}{\phi}$. Efron (1986) shows that this approximation is highly accurate, and we will use it in our more general specifications.

the process of the conditional mean. The most interesting structure is the reduced rank and own effect model. In this formulation it is assumed, that for every series the conditional mean depends on one lag of itself, one lag of the count and r factors of the cross-section of lagged counts. $A = diag(\alpha_i) + \gamma\delta'$ where γ and δ are (K, r) matrices. This is suited for large systems, where imposing a reduced rank is necessary for practical reasons, but there is reason to believe that every series' own past has explanatory power beyond the factor structure. In particular, the conditional mean can be specified as:

$$\mu_t = \omega + (diag(\alpha_i) + \gamma\delta')N_{t-1} + diag(\beta_i)\mu_{t-1} \ . \tag{2.9}$$

Moreover, in some cases one might want to assume that the dynamics of all the series under consideration is common, and that one factor explains the dynamics of the whole system. This can be obtained as a special case of our specification under the following set of assumptions: $A = \alpha\,\gamma\,\delta^{'}$, $B = \beta\, I$, $\omega = c\,\gamma$, where α, β are scalars, $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_K)^{'}$ and $\delta = (1, \delta_2, \ldots, \delta_K)^{'}$, where we impose the normalisation $\delta_1 = 1$ in order to identify the model. This means that if we denote $f_t = \delta^{'} N_t$, we have an autoregressive process for the factor:

$$\mu_t^0 = c + \alpha f_{t-1} + \beta\mu_{t-1}^0 \ ,$$

and $\mu_t = \gamma\mu_t^0$.

It is easy to show that the MDACP is stationary as long as the roots of the sum of the autoregressive coefficient matrices are within the unit circle, or equivalently, the eigenvalues of $(I - A - B)$ lie within the unit circle. In that case, the unconditional mean of the MDACP(p,q) is identical to the one of a VARMA process:

$$E[N_t] = \mu = (I - A - B)^{-1}\omega \tag{2.10}$$

2.3 The Model

Having dealt with the problems due to the discreteness and the time-varying nature of the marginal density in earlier sections, we proceed with the estimation of the model. The joint density of the counts in the Double Poisson case with the Gaussian copula is:

$$h(N_{1,t},\ldots,N_{K,t},\theta,\Sigma) = \prod_{i=1}^{K} f_{DP}(N_{i,t},\mu_{i,t},\phi_i) \cdot c(q_t;\Sigma) \ ,$$

$f_{DP}(N_{i,t},\mu_{i,t},\phi_i)$ denotes the Double Poisson density as a function of the observation $N_{i,t}$, the conditional mean $\mu_{i,t}$ and the dispersion parameter ϕ_i. c denotes the copula density of a multivariate normal and $\theta = (\omega, vec(A), vec(B))$.

The $q_{i,t}$, gathered in the vector q_t are the normal quantiles of the $z_{i,t}$:

$$q_t = (\Phi^{-1}(z_{1,t}),\ldots,\Phi^{-1}(z_{K,t}))' \ ,$$

where the $z_{i,t}$ are the PIT of the continuoused count data, under the marginal densities:

$$z_{i,t} = F^*(N^*_{i,t}) = F(N_{i,t}-1) + f(N_{i,t}) * U_{i,t} \ ,$$

The $N^*_{i,t}$ are the continuoused version of the original count data $N_{i,t}$:

$$N^*_{i,t} = N_{i,t} + (U_{i,t} - 1) \ .$$

Finally the $U_{i,t}$ are uniform random variable, on $[0,1]$.

Taking logs, one gets:

$$log(h_t) = \sum_{i=1}^{K} log(DP(N_{i,t},\mu_{i,t},\phi_i)) + log(c(q_t;\Sigma))$$

We consider a two-stage estimator as in Patton (2002). Given that we use the multivariate normal copula, the second step of the two-stage procedure does not require any optimisation, as the MLE of the variance-covariance matrix of a multivariate normal with a zero mean, is simply the sample counterpart:

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T} q_t q_t^{'}$$

It is important to realise that correct specification of the density in the marginal models is crucial to the specification of the copula, as any mistake would have as a consequence the fact that the uniformity assumption is violated which would invalidate the use of copulas. We evaluate models on the basis of their log-likelihood, but also on the basis of their Pearson residuals, which are defined as: $\epsilon_t = \frac{N_t - \mu_t}{\sigma_t}$. If a model is well specified, the Pearson residuals will have variance one and no significant autocorrelation left. The $z_{i,t}$'s are another tool for checking the specification. If the model is well specified, the $z_{i,t}$'s should be uniformly distributed and serially uncorrelated. We will check this for all the models we estimate.

We now establish two properties about the unconditional variance and the autocorrelation of the MDACP with an ARMA(1,1) structure, which we denote the MDACP(1,1).

Proposition 2.1 (Unconditional variance of the MDACP(1,1) Model). *The unconditional variance of the MDACP(1,1) model, when the conditional mean is given by 2.8, is equal to:*

$$vec(V[N_t]) = \left(I_{K^2} + \left(\left(I_{K^2} - (A+B)\otimes(A+B)^{'}\right)^{-1}\cdot\left(A\otimes A^{'}\right)\right)\right)\cdot vec(\Omega)\ , \qquad (2.11)$$

where $\Omega = E\left(Var[N_t|\mathcal{F}_{t-1}]\right)$. Under small dispersion asymptotics of Jorgensen (1987), $\Omega \simeq V^{\frac{1}{2}}\Sigma V^{\frac{1}{2}}$, where Σ is the copula covariance and V is the variance of the marginal models: $V = diag(\frac{\mu_i}{\phi_i})$.

This is a multivariate extension of Proposition 3.2 of Heinen (2003). The proof is shown in the Appendix. The variance is equal to the ratio of the mean to the dispersion parameter, the covariances are zero and therefore the variance-covariance matrix is diagonal. It can be seen that the variance-covariance of the counts is the product of a term reflecting the autoregressive part of the model, a term capturing the variance of the marginal models and a copula term responsible for the part of the contemporaneous cross-correlation which does not go through the time-varying mean.

Proposition 2.2 (Autocovariance of the MDACP(1,1) Model). *The autocovariance of the MDACP(1,1) model, when the conditional mean is given by 2.8, is equal to:*

$$vec(Cov[N_t, N_{t-s}]) = \left[I \otimes A^{-1}\left((A+B)^s - B(A+B)\right)\right] \cdot vec\left(V[N_t] - \Omega\right) \tag{2.12}$$

where Ω and $V[N_t]$ are as defined in Proposition 2.1.

The proof is shown in the Appendix.

3 Sector- and Stock-Specific News

Much of the microstructure literature is based on the existence of asymmetric information and consequently of two types of traders: the uninformed who trade for liquidity reasons and informed traders who possess superior information. This superior information can be macroeconomic, sector- or stock-specific information. Through the trading process this information is disseminated to the public, therefore trading conveys information. According to Admati & Pfleiderer (1988) and **?** frequent trading implies that news is arriving to the market. Thus a higher number of trades in a given time interval is a signal for the arrival of news.

The trading activity of one asset does not only convey information about that specific asset, but can also contain information about the whole sector that this asset belongs to. In order to model comovement in trading activity within a sector, **?** propose a duration model for the trading intensities of pairs of stocks of department stores. Their model consists of a univariate duration model for the pooled trades of two stocks and a probit specification which determines in which stock a transaction took place. They classify stocks according to how much sector-wide information they contain, based on a series of ratios of the sample variance of the conditional intensity of the pooled and univariate ACD models for each pair of stocks. In recent years, the focus of empirical microstructure has shifted from the study of an individual asset to the analysis of the cross-sectional interactions amongst stocks. Hasbrouck & Seppi (2001) document the existence of commonalities in order flow that are responsible for about two thirds of the commonalities in returns, using principal components analysis and canonical correlations on the stocks of the Dow Jones Industrial Average.

We analyse the same data as **?**, but the MDACP allows us to take into account the interaction amongst all stocks simultaneously as in Hasbrouck & Seppi (2001), which is helpful for the purpose of identifying leaders from the point of view of dissemination of sectorial information, while at the same time modelling the dynamics in a very general framework.

3.1 Data description

We are working with the five most important US department stores traded on the New York Stock Exchange during the year 1999: May Department Stores (MAY), Federated Department Stores (FD), J.C. Penney Company, Inc (JCP), Dillar's INC (DDS) and Saks Inc (SKS). We work with the number of trades in 5-minute intervals. The data we use was taken from the Trades and Quotes (TAQ) data set, produced by the New York Stock Exchange (NYSE). This data set contains every trade and quote posted on the NYSE, the American Stock Exchange and the NASDAQ National Market System for all securities listed on NYSE. We first remove any trades that occurred with non-standard correction or G127 codes (both of these are fields in the trades data base on the TAQ CD's), such as trades that were cancelled, trades that were recorded out of time sequence, and trades that were called for delivery of the stock at some later date. Any trades that were recorded to have occurred before 9:45 AM or after 4 PM (the official close of trading) were removed. The reason for starting at 9:45 instead of 9:30 AM, the official opening time, is that we wanted to make sure that none of the opening transactions were accidentally included in the sample, or that there would not be artificially low numbers of events at the start of the day, due to the fact that part of the first interval was taking place before the opening transaction. This could have biased estimates of intradaily seasonality.

The data used was from January 2nd 1999 to December 30th 1999. This means that the sample covers 252 trading days, that represent $18,900$ observations, as there are 75 5-minute intervals every day between 9:45 AM and 4 PM. The descriptive statistics are given in Table 1. The means of the series are relatively small, which makes the use of a continuous distribution like the normal problematic. As can be seen, the data exhibits significant overdispersion (the variance is greater than the mean), which could be due alternatively to autocorrelation or to overdispersion in the marginal distribution. The presence of overdispersion is confirmed by looking at the histogram of the data in Figure 1, which shows that, whereas the probability mass is fairly concentrated around the mean, there exist large outliers. There is significant autocorrelation in each series, as can be seen from the Ljung-Box Q-statistic shown here at order 20. Table 2 presents the contemporaneous correlation matrix among the five series we analyze, obtained using the Gaussian copula on the data which marginals are assume to be Double Poisson distributed. Figure 3 shows the auto- and cross-correlations of the vector of market-events, up to 375 5-minute intervals, which corresponds to 5 trading days. A very striking pattern of seasonality could be appreciated. Clearly looking only at contemporaneous correlation does not reveal the full picture, there is a very significant and systematic link across time between the various trading events. The correlations move from positive to negative in a systematic way, which seems to be due to the presence of diurnal seasonality of the U-shape type, which is commonly found in time series based on high-frequency data.

3.2 Estimation results

In the present subsection we discuss the estimates of two different specifications of the model, one based on the idea of a common factor and the second based on a mean structure based on a common factor, a series-specific lagged term in the moving average part and a diagonal autoregressive part. In order to fit the dispersion we use the double Poisson

Table 1: Descriptive statistics

	DDS	FD	JCP	MAY	SKS
No.trades	55,399	100,928	108,392	90,881	59,725
Mean	2.93	5.34	5.73	4.81	3.16
Median	2.00	5.00	5.00	4.00	3.00
Std.Dev.	2.57	3.56	3.89	3.04	2.84
Dispersion	2.25	2.38	2.64	1.92	2.55
Maximum	37	35	38	22	32
Minimum	0	0	0	0	0
$Q(20)$	11,560	15,504	34,482	8,531.7	33,679

Descriptive statistics for the number of trades. The number of observations is $18,900$. $Q(20)$ is the Ljung-Box Q-statistic of order 20 on the series. The dispersion refers to the ratio of the variance to the mean.

Table 2: Correlation Matrix of the trades data

	DDS	FD	JCP	MAY	SKS
DDS	1.00				
FD	0.27	1.00			
JCP	0.24	0.29	1.00		
MAY	0.25	0.30	0.31	1.00	
SKS	0.12	0.10	0.15	0.12	1.00

distribution and we model seasonality using a series of half hourly dummy variables. The results are shown in table 3. Note that the estimates of the MDACP are quasi-maximum likelihood estimates (QMLE), which deliver consistent parameters, even in the case of a misspecified distribution. The eigenvalues of $A+B$ are smaller than 1, which means that the model is stationary. A likelihood ratio test shows that the seasonality variables (the estimates are not shown) are jointly significant. The coefficients on the seasonality shown in Figure 2 exhibit the well-documented U-shape, which means that there is more activity at the beginning and end of the trading day and less at lunch time. The dispersion parameter ϕ of the double Poisson is also very significantly different from 1, which corresponds to the Poisson case. This means that the Poisson distribution is strongly rejected and that we now have a much better model for the conditional distribution. Furthermore, if the model is well specified, the Pearson residuals will have variance one and neither significant auto- nor cross-correlation left, something that could be appreciated in the estimations presented here.

Visual inspection of the Q-Q plots of the Z statistic of the factor plus own model in figure 4 reveals that indeed the distribution is well specified, since the Q-Q plots nearly coincide with the 45-degree line. This means that with the use of the double Poisson we satisfy the uniformity assumption, which is the theoretical basis for using copulas. The same results hold for the factor only model. The autocorrelations of the Z statistic, shown in figure 5 are essentially not significant, which indicates that the dynamics of the series is well accounted for. The correlations of the Pearson residuals of the series (not shown) confirm that there is no more seasonal pattern left and the correlations are well below significance. This however is not the case for the factor only model, for which there still remains autocorrelation in the residuals.

In order to model the contemporaneous correlations we estimate a multivariate normal copula. As this model is somewhat involved in terms of the number of parameters, we use the two-step procedure of Patton (2002). Table 4 shows the copula correlation matrix Σ of Proposition 2.1, which is responsible for the part of the contemporaneous and lagged cross-correlation which does not go through the time-varying mean.

Our first results are based on the factor only model (left panel of Table 3), in which we assume that the dynamics of all the series under consideration is common, and that one factor explains the dynamics of the whole system. To see the influence on the factor of each of the assets involved we just need to take a look at the vector of factor weights (the δ's). According to this the ranking of sectorial influence is JCP, SKS, DDS, FD and MAY. These results are closely related to **?** who find that the assets that contain more sectorial information are, in descending order, JCP, FD, SKS, DDS and MAY. As they mention, this ranking is related to the average number of transactions (see Table 1). This amounts to saying that the stocks with most sectorial information are the most frequently traded ones. However, if instead we rely on the intuitive idea that every stock's past trading activity plays a special role for that asset, in addition to an effect through a common factor, we find quite a different result. The results in the right side of Table 3, obtained with a series-specific lagged term, a common factor in the moving average part and a diagonal autoregressive part, we find a quite different ranking: MAY, FD, DDS, JCP and SKS. This ranking does no longer match the ranking based on the average trading activity of the asset, but it is instead highly related with the market capitalizations of the stocks. Indeed, ranking the most important US department stores by their size we have: MAY, FD, JCP, DDS and SKS [2].

Based on these results, we can conclude that indeed, within a sector there exist two kinds of information that matter for traders: stock specific information, related to the series-specific autocorrelation coefficients (the $\alpha_{i,i}$'s) and sector specific news, captured by the common factor (the δ's and γ's). Unlike traders trying to benefit from a stock-specific information, a trader with sector-specific information who is trying to conceal it, has the choice of which asset to trade in. He should naturally chose the asset with the least amount of sectorial information in its trading activity, as this would allow him to hide his private information without impacting the market too much. Based on our system of five department stores, trading in SKS and JCP could be appropriate. Of course, to obtain more general results, one would need to incorporate all stocks of that sector and not only the biggest ones as we do.

The comparison of our results with the ones of duration-based models suggests that taking into consideration all the assets simultaneously does make a difference. We are able to capture cross-sectional interactions with an intuitive factor-structure, commonly used in finance since the CAPM and also used more recently in the context of liquidity and order flow by Hasbrouck & Seppi (2001). This advantage of our multivariate specification over bivariate duration models compensates for the loss of information due to the aggregation from durations to counts. We have estimated our models for different time intervals (10 and 15 minutes) and we obtain the same results (available upon request). This robustness

[2] their Market capitalization in millions of US Dollars were: $11,226$, $8,945$, $3,538$, $1,647$, and $1,612$ respectively

over time aggregation and the accordance of our results with economic intuition increases our confidence in the findings.

4 Conclusion

In this paper we introduce new models for the analysis of multivariate time series of count data with many possible specifications. These models have proved to be very flexible and easy to estimate. We discuss how to adapt copulas to the case of time series of counts and show that the Multivariate Autoregressive Conditional Double Poisson model (MDACP) can accommodate many features of multivariate count data, such as discreteness, over- and underdispersion (variance greater and smaller than the mean) and both auto- and cross-correlation. Hypothesis testing in this context is straightforward, because all the usual likelihood-based tests can be applied. Another important advantage of this model is that it can accommodate both positive and negative correlation among variables, which most multivariate count models cannot do, and this is shown to be important in our financial application.

As a feasible alternative to multivariate duration models, the model is applied to the study of sector and stock specific news related to the comovements in the number of trades per unit of time of the most important US department stocks traded on the New York Stock Exchange. We show that the informational leaders inside a specific sector are related to their size measured by their market capitalization rather than to their trading activity.

We advocate the use of the Multivariate Autoregressive Conditional Double Poisson model for the study of multivariate point processes in finance, when the number of variables considered simultaneously exceeds two and looking at durations becomes too difficult. Plans for further research include evaluating the forecasting ability of these models, both in terms of point and density forecasts and we left more empirical applications for further work with more detailed tick-by-tick data sets.

5 Appendix

Proof of Proposition 2.1. Upon substitution of the mean equation in the autoregressive intensity, one obtains:

$$\mu_t - \mu = A(N_{t-1} - \mu) + B(\mu_{t-1} - \mu) \tag{5.1}$$

$$\mu_t - \mu = A(N_{t-1} - \mu_{t-1}) + (A + B)(\mu_{t-1} - \mu) \tag{5.2}$$

Squaring and taking expectations gives:

$$V[\mu_t] = AE\left[(N_{t-1} - \mu_{t-1})(N_{t-1} - \mu_{t-1})^{'}\right] A + (A + B)V[\mu_{t-1}](A + B)^{'} \tag{5.3}$$

Using the law of iterated expectations and denoting $\Omega = V\left[N_t|\mathcal{F}_{t-1}\right]$, one gets:

$$V[\mu_t] = A\Omega A + (A + B)V[\mu_{t-1}](A + B)^{'} \tag{5.4}$$

Vectorialising and collecting terms, one gets:

$$vec(V[\mu_t]) = \left(I_{K^2} - (A + B) \otimes (A + B)^{'}\right)^{-1} \cdot (A \otimes A^{'}) \cdot vec(\Omega) \tag{5.5}$$

Now, applying the following property on conditional variance

$$V[y] = E_x\left[V_{y|x}(y|x)\right] + V_x\left[E_{y|x}(y|x)\right] \tag{5.6}$$

to the counts and vectorialising, one obtains:

$$vec(V\left[N_t\right]) = vec(\Omega) + vec(V[\mu_t]) \tag{5.7}$$

Again using the law of iterated expectations, substituting the conditional variance σ_t for its expression, then making use of the previous result, and after finally collecting terms, one gets the announced result.

$$vec(V[N_t]) = \left(I_{K^2} + \left(\left(I_{K^2} - (A+B) \otimes (A+B)^{'}\right)^{-1} \cdot (A \otimes A^{'}\right)\right) \cdot vec(\Omega) \tag{5.8}$$

Based on Song (2000), and on tail area approximations (Jorgensen (1997), we can approximate the Pearson residual as follows:

$$F(N_{i,t}, \mu_{i,t}, \phi) \simeq \Phi \left(\frac{N_{i,t} - \mu_{i,t}}{\sqrt{\frac{\mu_{i,t}}{\phi_i}}} \right) , \tag{5.9}$$

Equivalently, we have:

$$q_{i,t} \equiv \Phi^{-1}(F(N_{i,t}, \mu_{i,t}, \phi)) \simeq \frac{N_{i,t} - \mu_{i,t}}{\sqrt{\frac{\mu_{i,t}}{\phi_i}}} \equiv \epsilon_{i,t} , \tag{5.10}$$

Therefore we can approximate the variance-covariance of the Pearson residuals with the copula covariance:

$$\Sigma = Cov(q_t) \simeq Cov\left(\epsilon_{i,t}\right) \tag{5.11}$$

Now the average conditional variance-covariance matrix Ω can be obtained simply from Σ as:

$$\Omega \simeq V^{\frac{1}{2}} \Sigma V^{\frac{1}{2}} \tag{5.12}$$

□

Proof of Proposition 2.2. As a consequence of the martingale property, deviations between the time t value of the dependent variable and the conditional mean are independent from the information set at time t. Therefore:

$$E[(N_t - \mu_t)(\mu_{t-s} - \mu)^{'}] = 0 \qquad \forall\, s \geq 0 \tag{5.13}$$

By distributing $N_t - \mu_t$, one gets:

$$Cov[N_t, \mu_{t-s}] = Cov[\mu_t, \mu_{t-s}] \qquad \forall\ s \geq 0 \tag{5.14}$$

By the same "non-anticipation" condition as used above, it must be true that:

$$E[(N_t - \mu_t)(N_{t-s} - \mu)^{'}] = 0 \qquad \forall\ s \geq 0 \tag{5.15}$$

Again, distributing $N_t - \mu_t$, one gets:

$$Cov[N_t, N_{t-s}] = Cov[\mu_t, N_{t-s}] \qquad \forall\ s \geq 0 \tag{5.16}$$

Now,

$$\begin{aligned} Cov[\mu_t, \mu_{t-s}] &= ACov[N_t, \mu_{t-s+1}] + BCov[\mu_t, \mu_{t-s}] \\ &= (A+B)Cov[\mu_t, \mu_{t-s}] \\ &= (A+B)^s V[\mu_t] \end{aligned} \tag{5.17}$$

The first line was obtained by replacing μ_t by its expression, the second line by making use of 5.14, the last line follows from iterating line two.

$$Cov[\mu_t, \mu_{t-s+1}] = ACov[\mu_t, N_{t-s}] + BCov[\mu_t, \mu_{t-s}] \tag{5.18}$$

Rearranging and making use of 5.16, one gets:

$$\begin{aligned} ACov[N_t, N_{t-s}] &= Cov[\mu_t, \mu_{t-s+1}] - BCov[\mu_t, \mu_{t-s}] \\ &= ((A+B)^s - B(A+B))\, V[\mu_t] \end{aligned} \tag{5.19}$$

Under the condition that A is invertible, which is not an innocuous assumption, as it excludes the pure factor model, we get after vectorialising:

$$vec(Cov[N_t, N_{t-s}]) = \left[I \otimes \left(A^{-1}(A+B)^s - A^{-1}B(A+B)\right)\right] vec(V[\mu_t]) \tag{5.20}$$

After substituting in 5.8, we get:

$$\begin{aligned} vec(Cov[N_t, N_{t-s}]) = &\left[I \otimes A^{-1}\left((A+B)^s - B(A+B)\right)\right] \cdot \\ &\left(I_{K^2} + \left(\left(I_{K^2} - (A+B) \otimes (A+B)^{'}\right)^{-1} \cdot (A \otimes A^{'}\right)\right) \cdot vec(\Omega) \end{aligned} \tag{5.21}$$

□

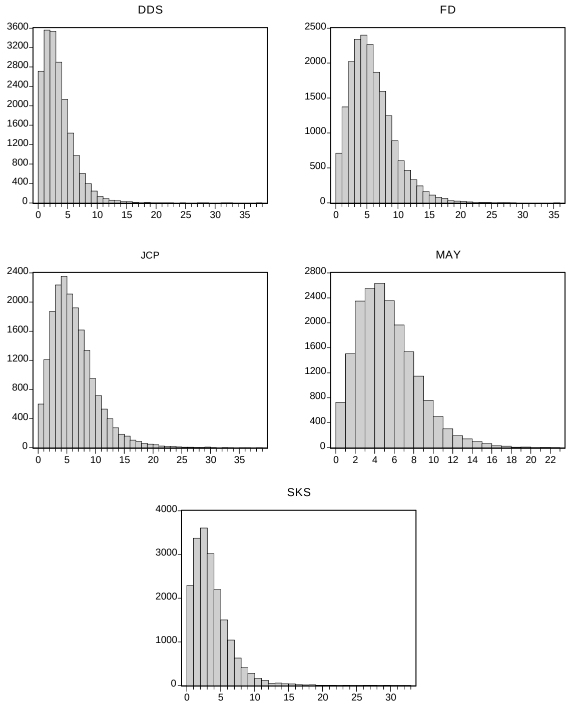

Figure 1: Histogram of the data

References

Admati, A. & Pfleiderer, P. (1988), 'A theory of intraday patterns: Volume and price variability', *Review of Financial Studies* **1**, 3–40.

Efron, B. (1986), 'Double exponential families and their use in generalized linear regression', *Journal of the American Statistical Association* **81**(395), 709–721.

Engle, R. F. & Lunde, A. (2003), Trades and quotes: A bivariate process.

Engle, R. F. & Russell, J. (1998), 'Autoregressive conditional duration: A new model for irregularly spaced transaction data', *Econometrica* **66**(5), 1127,1162.

Fisher, R. A. (1932), *Statistical Methods for Research Workers*.

Hasbrouck, J. & Seppi, D. J. (2001), 'Common factors in prices, order flows and liquidity', *Journal of Financial Economics* **59**(3), 383–411.

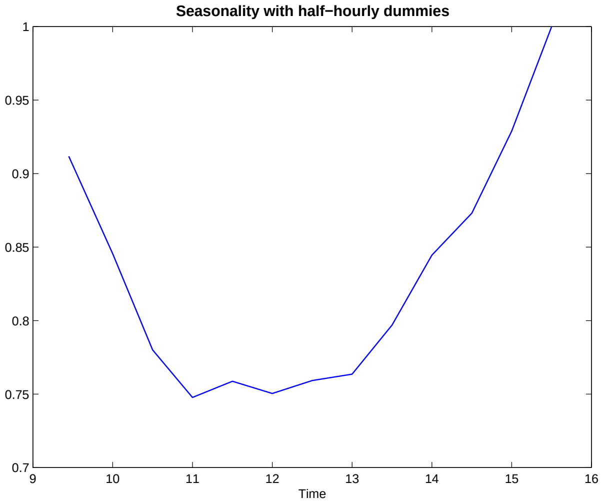

Figure 2: Seasonality plot of the data

Heinen, A. (2003), Modeling time series count data: An autoregressive conditional poisson model.

CORE Discussion Paper 2003/62.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Chapman and Hall, London.

Jorgensen, B. (1987), 'Exponential dispersion model', *Journal of the Royal Statistical Society* **49**(2), 127–162.

Jorgensen, B. (1997), *The Theory of Dispersion Models*, Chapman and Hall, London.

Nelsen, R. B. (1999), *An introduction to Copulas*, Springer, New York.

Patton, A. (2002), Skewness, asymmetric dependence, and portfolios.

mimeo.

Sklar, A. (1959), 'Fonctions de répartitions à n dimensions et leurs marges', *Public Institute of Statistics of the Univeristy of Paris* **8**, 229–231.

Song, P. X.-K. (2000), 'Multivariate dispersion models generated from gaussian copulas', *Scandinavian Journal of Statistics* **27**, 305–320.

Zeger, S. L. (1988), 'A regression model for time series of counts', *Biometrika* **75**(4), 621–629.

Table 3: **Maximum Likelihood Estimates of the MDACP models.**

The table presents the Maximum Likelihood Estimates of the Multivariate Autoregresive Conditional Double Poisson (MDACP) models on counts based on data of the 5 most important retail department stores: DDS, FD, JCP, MAY and SKS at intervals of 5 minutes for the period January 1999 to the end of December 1999. These models consider the seasonality presented in the data and solved it by the use of 30 minutes dummies. The t-statistics are presented in parenthesis. We impose the normalisation $\delta_1 = 0.25$ in order to identify the model. $\epsilon_t = \frac{N_t - \mu_t}{\sigma_t}$ are the Pearson residuals from the model. The equation of the factor only model is:

$\mu_t^0 = c + \alpha f_{t-1} + \beta \mu_{t-1}^0$

with $\mu_t = \gamma \mu_t^0$

and the model with a factor and an own effect:

$\mu_t = \omega + (diag(\alpha_i) + \gamma\delta')N_{t-1} + diag(\beta_i)\mu_{t-1}$

θ	MDACP factor only model					MDACP with factor and own effect				
	DDS	FD	JCP	MAY	SKS	DDS	FD	JCP	MAY	SKS
ω_i	0.325	0.773	0.098	0.600	0.043	0.136	0.331	0.216	0.275	0.094
	(8.73)	(12.62)	(5.47)	(11.40)	(4.04)	(7.97)	(11.98)	(10.30)	(10.67)	(7.46)
α_{1i}						0.137				
						(25.80)				
α_{2i}							0.151			
							(17.51)			
α_{3i}								0.178		
								(36.50)		
α_{4i}									0.108	
									(12.19)	
α_{5i}										0.161
										(33.73)
γ	0.119	0.217	0.213	0.124	0.100	0.022	0.050	0.008	0.043	0.011
	(14.01)	(14.57)	(17.50)	(12.97)	(15.97)	(3.22)	(3.68)	(1.838)	(3.87)	(3.43)
δ	0.250	0.247	0.375	0.218	0.335	0.250	0.382	0.122	0.461	0.072
		(15.44)	(18.90)	(13.20)	(17.39)		(3.32)	(2.89)	(3.20)	(1.77)
β	0.694	0.664	0.790	0.761	0.839	0.811	0.777	0.814	0.820	0.825
	(35.26)	(38.06)	(106.25)	(48.28)	(111.66)	(110.05)	(97.78)	(155.75)	(103.36)	(143.00)
ϕ	0.504	0.496	0.514	0.571	0.498	0.546	0.542	0.575	0.600	0.584
	(94.09)	(96.24)	(95.95)	(95.58)	(99.52)	(92.78)	(98.59)	(101.26)	(97.26)	(95.21)
LogL	-219,072					-214,463				
Eigenval	0.99	0.68	0.71	0.83	0.77	0.94	0.97	0.99	0.99	0.93
$Var(\epsilon_t)$	1.00	0.99	1.00	0.97	1.04	0.97	0.98	0.99	0.97	0.98

Table 4: **Correlation Matrix of the Q estimated by the MDACP model.**
The table presents the correlation matrix of Q, based on the probability integral transformation, Z, of the continuoused count data under the marginal densities estimated using the MDACP models by the two-step procedure

	$COPULA - MDACP4S$				
	DDS	FD	JCP	MAY	SKS
DDS	1.00				
FD	0.16	1.00			
JCP	0.17	0.18	1.00		
MAY	0.15	0.17	0.20	1.00	
SKS	0.02	0.02	0.04	0.03	1.00

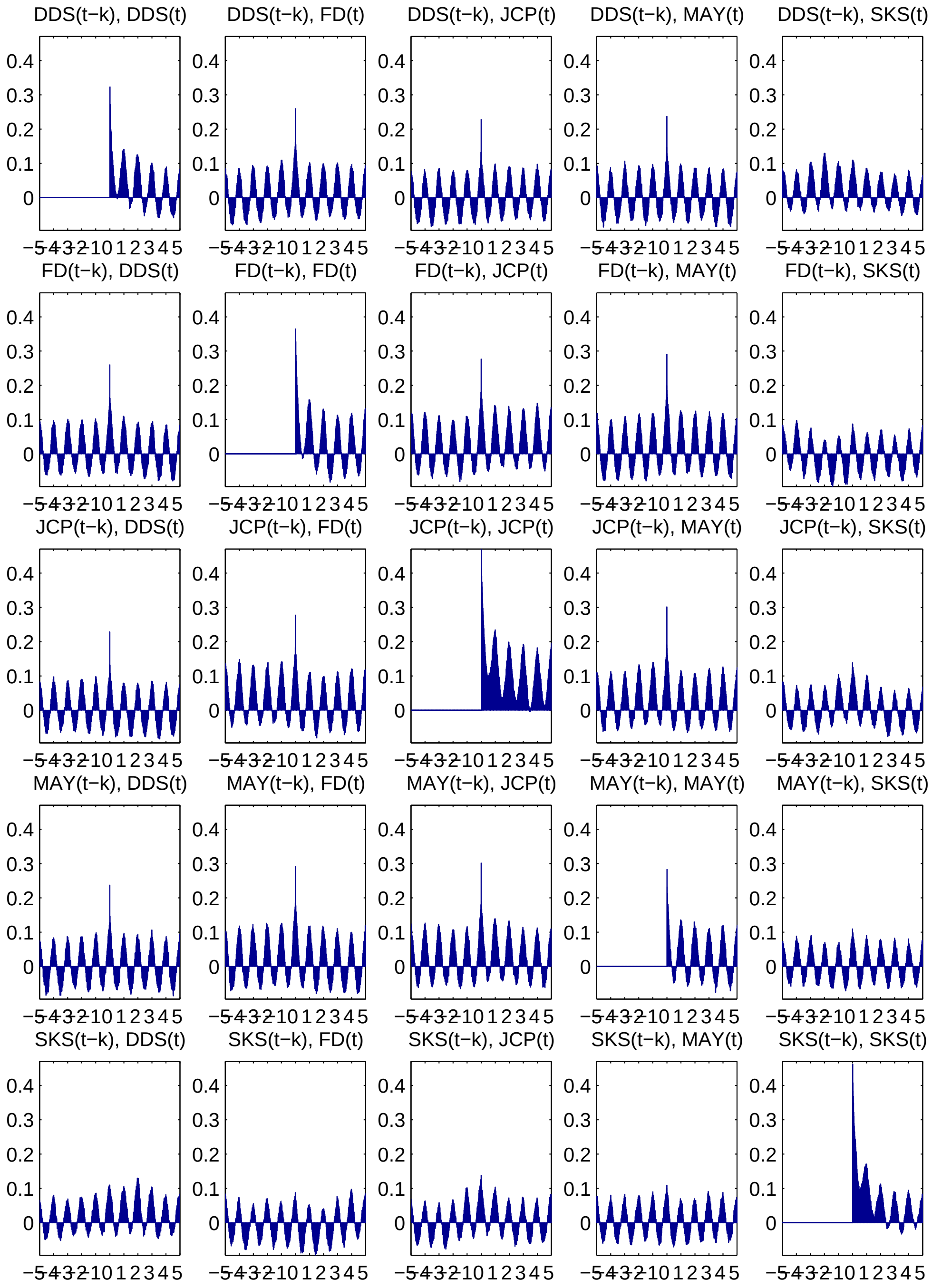

Figure 3: Auto- and cross-correlogram of the data

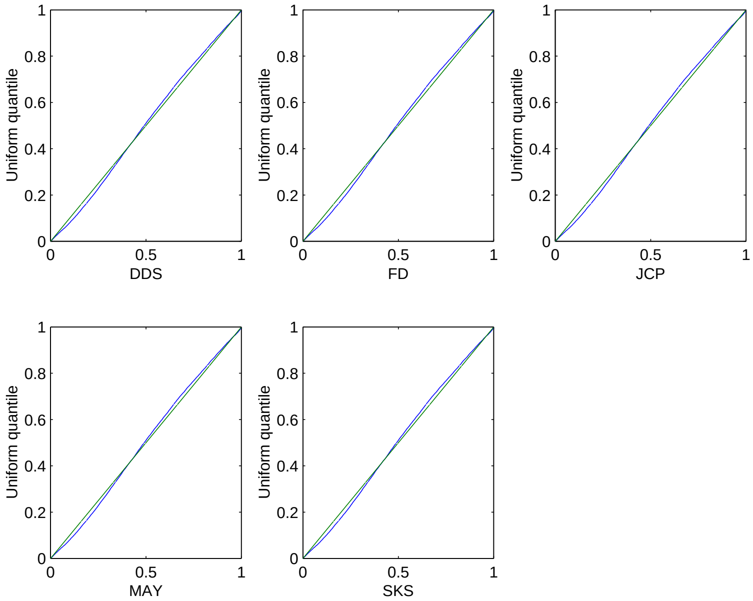

Figure 4: Quantile Plots of the Z statistics of the MDACP model with factor and own effect

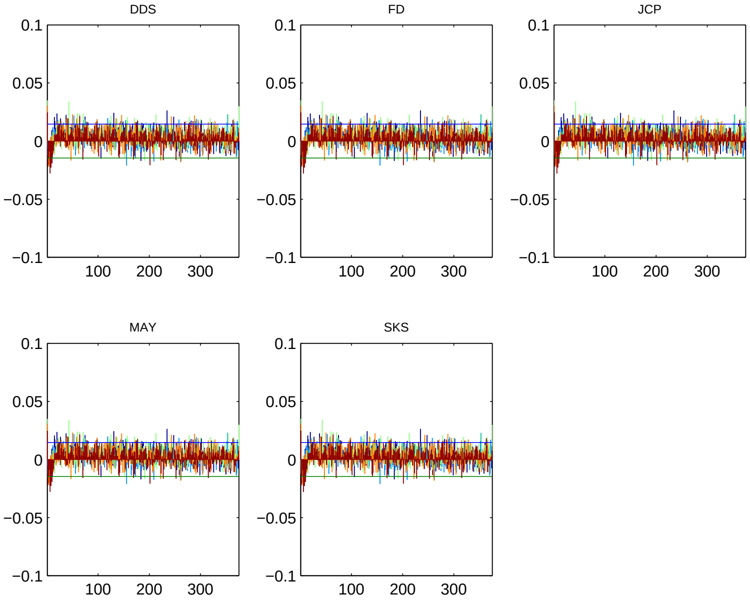

Figure 5: Autocorrelation of the Z statistics of the MDACP model with factor and own effect