Jung, Robert; Liesenfeld, Roman; Richard, Jean-François

Working Paper

Dynamic Factor Models for Multivariate Count Data: An Application to Stock-Market Trading Activity

Economics working paper / Christian-Albrechts-Universität Kiel, Department of Economics, No. 2008,12

Provided in cooperation with:
Christian-Albrechts-Universität Kiel (CAU)

Dynamic Factor Models for Multivariate Count Data: An Application to Stock-Market Trading Activity

by Robert C. Jung, Roman Liesenfeld and Jean-François Richard

Christian-Albrechts-Universität Kiel

Department of Economics

Economics Working Paper

No 2008-12

Dynamic Factor Models for Multivariate Count Data: An Application to Stock-Market Trading Activity

Robert C. Jung
Universität Erfurt, Department of Economics, Germany

Roman Liesenfeld*
Christian-Albrechts-Universität Kiel, Department of Economics, Germany

Jean-François Richard
University of Pittsburgh, Department of Economics, USA

(July 22, 2008)

Abstract

We propose a dynamic factor model for the analysis of multivariate time series count data. Our model allows for idiosyncratic as well as common serially correlated latent factors in order to account for potentially complex dynamic interdependence between series of counts. The model is estimated under alternative count distributions (Poisson and negative binomial). Maximum Likelihood estimation requires high–dimensional numerical integration in order to marginalize the joint distribution with respect to the unobserved dynamic factors. We rely upon the Monte–Carlo integration procedure known as Efficient Importance Sampling which produces fast and numerically accurate estimates of the likelihood function. The model is applied to time series data consisting of numbers of trades in 5 minutes intervals for five NYSE stocks from two industrial sectors. The estimated model accounts for all key dynamic and distributional features of the data. We find strong evidence of a common factor which we interpret as reflecting market–wide news. In contrast, sector–specific factors are found to be statistically insignificant.

Keywords: Dynamic latent variables, Importance sampling, Mixture of distribution models, Poisson distribution, Simulated Maximum Likelihood;

*Corresponding author. Tel.: +49-431-8803810; fax: +49-431-8807605. *E-mail address:* liesenfeld@stat-econ.uni-kiel.de (R. Liesenfeld)

1. Introduction

Modelling of dispersion and serial correlation for univariate count series has received much attention over recent years. Existing approaches can be broadly classified as either observation– or parameter–driven. The monographs of Kedem and Fokianos (2002) and McKenzie (2003) provide excellent overviews. More recent contributions include Jung et al. (2006), Neal and Subba Rao (2007) and Jung and Tremayne (2008).

Multivariate dynamic models for count data remain few. As discussed by Cameron and Trivedi (1998, Section 8.1), this might be explained by the fact that classical inference in multivariate count data models has proven to be analytically as well as computationally very demanding. This is particularly relevant for models attempting to capture the complex correlation structure characterizing many multivariate count time series. Three pioneering multivariate applications are found in Jørgensen et al. (1999), Held et al. (2005), and Heinen and Rengifo (2007). The specification proposed by Jørgensen et al. (1999) belongs to the class of parameter–driven models. It is a multivariate Poisson state–space model with a common factor following a gamma Markov process. These specific distributional assumptions produce a model which can be analyzed by a Kalman filter. The model is used to assess the impact of air pollution on daily emergency admission counts in an hospital for four sickness categories. Held et al. (2005) propose an observation–driven multivariate model which imposes a simple vector–autoregressive structure for the means. This model can be estimated by standard Maximum Likelihood (ML). It is applied to infectious disease surveillance counts from a measle epidemic. Heinen and Rengifo (2007) also adopt an observation–driven approach extending the univariate autoregressive conditional Poisson model of Heinen (2003). A copula approach is used to represent contemporaneous correlations among time series counts. Since efficient joint ML estimation is not feasible, the authors rely upon a consistent though less efficient two–stage ML approach for separate estimation of the parameters of the marginal distributions and those of the copula. Their model is then used to analyze co–movements in the number of trades for stocks traded at the New York Stock Exchange (NYSE). Other multivariate count models rely upon panel data techniques, with emphasis on unobserved heterogeneity in the individual series. See Winkelmann (2008) for a recent survey.

In the present paper we adopt a parameter–driven approach and propose a new flexible, parsimonious and easy to interpret dynamic factor model for multivariate count series. It builds upon and generalizes earlier models by Jørgensen et al. (1999) and Wedel et al. (2003). The former model includes a single dynamic common factor only and no dynamic idiosyncratic components. The latter model is a static multivariate Poisson factor model for cross-sectional analyses. Our model allows for serially correlated common as well as idiosyncratic factors driving the conditional means of the count distributions. Therefore, it can represent non–trivial contemporaneous and temporal interactions across count series. It can also accommodate different distributional assumptions for the conditional distribution of the counts given the factors. This can be critical since the commonly used Poisson distribution has an index of dispersion equal to one (the latter being defined as the ratio between the variance and the mean). However, count data often exhibit strong over–dispersion (index significantly larger than one) which can not be fully captured by a conditional Poisson distribution even if a varying conditional mean generates by itself an over–dispersed unconditional distribution. Hence, it is important to allow for conditional distributions which can accommodate over–dispersion, such as the negative binomial (here after Negbin) and the double Poisson.

Our model depends non–linearly upon its dynamic latent factors. Whence, likelihood evaluation requires high–dimensional numerical integration, for which we use the Efficient Importance Sampling (hereafter EIS) procedure developed by Richard and Zhang (2007). EIS is a generic, flexible and easy to implement Monte Carlo integration procedure specifically designed to maximize numerical accuracy. It also facilitates exploring alternative model specifications which typically require only minor modifications of a baseline EIS implementation. Last but not least, EIS can be used to compute filtered and/or smoothed estimates of the latent factors themselves. Several diagnostic test statistics are based upon such estimates.

Our model is then applied to a multivariate time series consisting of numbers of trades in 5–minutes intervals for five stocks traded at the NYSE. We implicitly adopt the information flow interpretation associated with the mixture-of-distribution model of Tauchen and Pitts (1983). See also Andersen (1996) and Liesenfeld (2001). In this context, numbers of trades are directly influenced by the arrival of new information, whether specific to a single stock (idiosyncratic factor), to an industry (sector factor), or to the market (market factor).

The paper is organized as follows. The multivariate dynamic factor model is introduced in Section 2, Section 3 discusses ML estimation, filtering and smoothing based upon EIS. The application to NYSE data is presented in Section 4. Section 5 concludes. Technical derivations are regrouped in an Appendix.

2. Dynamic Factor Model for Multivariate Count Data

The econometric model we propose consists of a dynamic extension of the static multivariate Poisson factor model introduced by Wedel et al. (2003). Consider a J–dimensional vector of counts $y_t = (y_{t1}, ..., y_{tJ})'$ recorded at time t, $(t = 1, ..., T)$. Dynamics will be introduced at the level of the latent factors. Whence, counts are assumed to be conditionally independently distributed with Poisson distributions

$$p(y_{tj}|\theta_{tj}) = \frac{\exp(-\theta_{tj})\theta_{tj}^{y_{tj}}}{y_{tj}!}, \quad t = 1, .., T, \quad j = 1, ..., J, \tag{1}$$

whose means θ_{tj} are latent random variables. We assume the existence of a link function $b(\cdot)$, whereby the mean vector $\theta_t = (\theta_{t1}, ..., \theta_{tJ})'$ can be expressed as a linear function of a P–dimensional vector of latent random factors f_t, say

$$b(\theta_t) = \mu + \Gamma f_t, \tag{2}$$

where μ denotes a vector of fixed intercepts and Γ a $(J \times P)$ matrix of factor loadings. The P latent factors in f_t are assumed to be independent of each other. A log-link function $b(\theta_t) = \ln(\theta_t)$ is convenient since it implies positivity of θ_t without parametric restrictions on (μ, Γ). Alternative link functions will not be considered here.

In the context of our NYSE application considering the joint behavior of the number of trades for different stocks, we allow for a single common market factor λ_t, $S < J$ industry–specific factors $\tau_t = (\tau_{t1}, ..., \tau_{tS})'$, and J stock–specific factors $\omega_t = (\omega_{t1}, ..., \omega_{tJ})'$. Whence, f_t is partitioned into $f_t = (\lambda_t, \tau_t', \omega_t')'$ and $P = J + S + 1$. The matrix of factor loadings is partitioned conformably with f_t into $\Gamma = (\Gamma_\lambda, \Gamma_\tau, \Gamma_\omega)$, where $\Gamma_\lambda = (\gamma_j^\lambda)$ is a J–dimensional vector, $\Gamma_\tau = (\gamma_j^{\tau_s})$ a $J \times S$ matrix with zero entries for any firm j which does not belong to sector s, and $\Gamma_\omega = \mathrm{diag}(\gamma_j^{\omega_j})$ a $(J \times J)$ diagonal matrix. Whence, the log-mean function for stock j, belonging to industry s is given by

$$\ln \theta_{tj} = \mu_j + \gamma_j^{\lambda} \lambda_t + \gamma_j^{\tau_{s_j}} \tau_{ts_j} + \gamma_j^{\omega_j} \omega_{tj}, \tag{3}$$

where the index s_j denotes the industry of firm j.

In order to account for possible serial and cross-correlation in the counts, we assume that the factors follow independent gaussian AR(1) processes, say

$$\lambda_t | \lambda_{t-1} \sim \mathrm{N}(\kappa^{\lambda} + \delta^{\lambda} \lambda_{t-1} \ , \ [\nu^{\lambda}]^2) \tag{4}$$

$$\tau_{ts} | \tau_{t-1s} \sim \mathrm{N}(\kappa^{\tau_s} + \delta^{\tau_s} \tau_{t-1s} \ , \ [\nu^{\tau_s}]^2), \tag{5}$$

$$\omega_{tj} | \omega_{t-1j} \sim \mathrm{N}(\kappa^{\omega_j} + \delta^{\omega_j} \omega_{t-1j} \ , \ [\nu^{\omega_j}]^2). \tag{6}$$

To ensure stationarity of the factors, it is assumed that $|\delta^{\lambda}| < 1$, $|\delta^{\tau_s}| < 1$, and $|\delta^{\omega_j}| < 1$. Other distributional and dynamic specifications for the factors are easily accommodated. Under an identity link $b(\cdot)$, for example, a Gamma transition distribution or a log-normal transition distribution would be suitable factor specifications (see, Jørgensen et al., 1999, and Jung and Liesenfeld, 2001).

The model as specified is unidentified. Identification for the static case with i.i.d. factors is discussed in Wedel et al. (2003) and can be extended to the dynamic model introduced here. We impose the restrictions that $\kappa^{\lambda} = \kappa^{\tau_s} = \kappa^{\omega_j} = 0$ for $s = 1, ..., S$ and $j = 1, ..., J$ in order to identify the μ_j's (see Equations 4–6). Furthermore, we set $\gamma_1^{\lambda} = 1$, $\gamma_j^{\omega_j} = 1$ for $j = 1, ..., J$, and $\gamma_j^{\tau_s} = 1$ for one arbitrarily selected stock j in industry s for $s = 1, ..., S$ (see Equation 3). This eliminates indeterminacies in the factor scales.

Under the assumed Poisson distribution, whose dispersion index equals one, over–dispersion of the counts can only originate from the unconditional variances of the factors, which themselves critically depend on the persistence parameters $(\delta^{\lambda}, \delta^{\tau_s}, \delta^{\omega_j})$. In order to relax this close relationship between over–dispersion and persistence, we can substitute a more flexible distribution for the Poisson. One such distribution which we shall apply below is the negative binomial (Negbin), which is given by

$$p(y_{tj}|\theta_{tj}) = \frac{\Gamma(y_{tj} + 1/\sigma_j^2)}{\Gamma(1/\sigma_j^2)\Gamma(y_{tj}+1)} \left(\frac{1}{1+\sigma_j^2\theta_{tj}}\right)^{1/\sigma_j^2} \left(\frac{\theta_{tj}}{\theta_{tj}+1/\sigma_j^2}\right)^{y_{tj}}, \tag{7}$$

where $\Gamma(\cdot)$ denotes the Gamma function. Its mean and variance are given by θ_{tj} and $\theta_{tj}(1 + \sigma_j^2\theta_{tj})$, respectively. The over–dispersion is a monotone increasing function of $\sigma_j > 0$ and the Poisson distribution in Equation (1) obtains as the limit for $\sigma_j \to 0$. The double Poisson distribution proposed by Efron (1994) or the generalized Poisson distribution proposed by Consul (1989) offer alternatives to capture (conditional) over–dispersion but will not be considered here.

3. EIS Based Inference

3.1 EIS

The evaluation of the likelihood function for the model described by Equations (1) to (6) requires integrating the joint density of counts and factors with respect to the $T \cdot P$ latent factor variables (in our application below $T \cdot P$ ranges from 22,875 to 36,600!). For likelihood evaluation counts are kept fixed at their observed values and are, therefore, omitted from notation except for the fact that densities need to be time indexed to reflect their dependence on the data.

The likelihood integral to be evaluated is of the following form:

$$L(\psi) = \int \cdots \int \prod_{t=1}^{T} \varphi_t(f_t, f_{t-1}; \psi) df_T \cdots df_1, \tag{8}$$

where ψ regroups the parameters of the model. φ_t denotes the product of the time t densities for y_t given f_t and for f_t given f_{t-1} as defined by Equations (1) to (6). The initial condition f_0 is assumed to be a known constant, which we set in our application to $f_0 = \mathrm{E}(f_t) = 0$. If all relevant integrals had analytical solutions, $L(\psi)$ would obtain from the following (backward) recursive sequence of P–dimensional integrals

$$L_t(f_{t-1}; \psi) = \int \varphi_t(f_t, f_{t-1}; \psi)\, L_{t+1}(f_t; \psi) df_t, \tag{9}$$

with $L_{T+1}(f_T;\psi) \equiv 1$, and $L(\psi) \equiv L_1(f_0;\psi)$. When these integrals are analytically intractable, EIS, as proposed by Richard and Zhang (2007), essentially amounts to constructing a sequence of auxiliary parametric density kernels $\{k_t(f_t, f_{t-1}; a_t)\,,\ a_t \in A_t\}_{t=1}^T$, which (i) are analytically integrable in f_t given f_{t-1}, and (ii) are amenable to MC simulation. The corresponding importance samplers are then given by

$$m_t(f_t|f_{t-1};a_t) = \frac{k_t(f_t, f_{t-1}; a_t)}{\chi_t(f_{t-1};a_t)}, \quad \text{with} \quad \chi_t(f_{t-1};a_t) = \int k_t(f_t, f_{t-1}; a_t) df_t. \tag{10}$$

The integral in Equation (8) is then rewritten as

$$L(\psi) = \chi_1(f_0;a_1) \int \cdots \int \prod_{t=1}^{T} \left[\frac{\varphi_t(f_t, f_{t-1};\psi)\chi_{t+1}(f_t;a_{t+1})}{k_t(f_t, f_{t-1};a_t)} \right] m_t(f_t|f_{t-1};a_t) df_T \cdots df_1, \tag{11}$$

with $\chi_{T+1}(\cdot) \equiv 1$. Here χ_{t+1} essentially substitutes for the analytically intractable L_{t+1} in Equation (9). EIS then aims at selecting $\{\hat{a}_t\}_{t=1}^T$ which minimizes the MC sampling variances of the ratios $\varphi_t \cdot \chi_{t+1}/k_t$ as functions of f_t *and* f_{t-1}, not just f_t. An MC-EIS approximate solution of this minimization problem obtains from the following backward sequence of auxiliary Least Squares (LS) problems:

$$\begin{aligned}(\hat{c}_t, \hat{a}_t) \quad = \quad \arg \min_{c_t \in \mathbb{R}, a_t \in A_t} \sum_{i=1}^{N} \Bigg\{ & \ln \left[\varphi_t\big(\tilde{f}_t^{(i)}, \tilde{f}_{t-1}^{(i)};\psi\big) \cdot \chi_{t+1}\big(\tilde{f}_t^{(i)};\hat{a}_{t+1}\big) \right] \\ & -c_t - \ln k_t\big(\tilde{f}_t^{(i)}, \tilde{f}_{t-1}^{(i)};a_t\big) \Bigg\}^2,\end{aligned} \tag{12}$$

where $\{\tilde{f}_t^{(i)}\}_{t=1}^T$ denotes a trajectory drawn from the (forward) sequence of auxiliary samplers $\{m_t(f_t|\tilde{f}_{t-1}^{(i)};\hat{a}_t)\}_{t=1}^T$ with $i = 1, ..., N$ (i.i.d.). In order to account for the fact that the $\{\tilde{f}_t^{(i)}\}$ in Equation (12) also depends on $\{\hat{a}_t\}$, the latter obtains as fixed-point solutions of the following iterated sequences of auxiliary backward LS problems:

$$\cdots \rightarrow \quad \{\hat{a}_t^{(k-1)}\}_{t=1}^T \quad \rightarrow \quad \begin{array}{c}\text{forward}\\ \text{draws}\end{array} : \{\{\tilde{f}_t^{(i),(k-1)}\}_{t=1}^T\}_{i=1}^N\} \quad \rightarrow \quad \begin{array}{c}\text{backward}\\ \text{LS}\end{array} : \{\hat{a}_t^{(k)}\}_{t=1}^T \quad \rightarrow \quad \cdots .$$

At convergence the EIS estimate of $L(\psi)$ is given by:

$$\bar{L}_N(\psi) = \chi_1(f_0; \hat{a}_1)\, \frac{1}{N} \sum_{i=1}^{N} \prod_{t=1}^{T} \left[\frac{\varphi_t(\tilde{f}_t^{(i)}, \tilde{f}_{t-1}^{(i)}; \psi)\chi_{t+1}(\tilde{f}_t^{(i)}; \hat{a}_{t+1})}{k_t(\tilde{f}_t^{(i)}, \tilde{f}_{t-1}^{(i)}; \hat{a}_t)} \right]. \tag{13}$$

For smooth convergence of the EIS fixed-point sequence as well as subsequent continuity of $\bar{L}_N(\psi)$ w.r.t. ψ, it is critical that all i-th trajectories $\{\tilde{f}_t^{(i),(k)}\}_{t=1}^T$ be obtained by transforming a single set of Common Random Numbers (CRNs), say $\{\tilde{u}_t^{(i)}\}_{t=1}^T$. CRNs are $\mathrm{N}(0,1)$ for gaussian EIS samplers and $U(0,1)$ for EIS-sampling densities simulated by cdf inversion. Most importantly, EIS-density kernels within the exponential family of distributions are linear in the auxiliary parameters a_t under their natural parametrization as well as closed under multiplication. As detailed in the Appendix, these two properties considerably simplify the application of EIS to our model. Note finally that $\{\hat{a}_t\}_{t=1}^T$ is an implicit function of ψ. Therefore, maximal numerical efficiency requires complete reruns of the EIS algorithm for any new value of ψ. See Richard and Zhang (2007) for details.

3.2 EIS likelihood for the dynamic count data model

EIS estimation of the likelihood function of the model defined by Equations (1) to (6) turns out to be conceptually straightforward and numerically accurate though notationally tedious. In this section we only outline the EIS implementation. All relevant algebraic details are regrouped in the Appendix.

Under the log–link function, the Poisson density in Equation (1) is rewritten as

$$p\left(y_{tj} \mid \phi_{tj}\right) = \frac{\exp\left(y_{tj}\phi_{tj} - e^{\phi_{tj}}\right)}{y_{tj}!}, \tag{14}$$

with $\phi_{tj} = \ln \theta_{tj}$. Equations (2) and (3) are rewritten in matrix form as

$$\phi_t = \mu + \Gamma f_t, \tag{15}$$

with $\phi_t' = (\phi_{t1}, \ldots \phi_{tJ})$, $f_t' = (\lambda_t, \tau_{t1}, \ldots \tau_{tS}, \omega_{t1}, \ldots, \omega_{tJ},)$, and

$$\Gamma = \left(\begin{array}{c|cccc|cccc} \gamma_1^\lambda & \gamma^{\tau_1} & 0 & \ldots & 0 & \gamma_1^{\omega_1} & 0 & \ldots & 0 \\ \gamma_2^\lambda & 0 & \gamma^{\tau_2} & \ldots & 0 & 0 & \gamma_2^{\omega_2} & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \gamma_J^\lambda & 0 & 0 & & \gamma^{\tau_S} & 0 & 0 & & \gamma_J^{\omega_J} \end{array} \right), \tag{16}$$

where $\gamma^{\tau_s} = (\gamma_j^{\tau_{s_j}})$, for $j = J_s + 1, \ldots, J_{s+1}$ ($J_1 = 0$, $J_{S+1} = J$) denotes the vector of factor loadings on the industry factor τ_{ts} for all stocks which belong to sector s. Equations (4) to (6) imply that

$$p\left(f_t \mid f_{t-1}\right) \sim \mathrm{N}\left(\Delta f_{t-1}, H^{-1}\right), \tag{17}$$

where Δ and H are both diagonal and H denotes the inverse of the covariance matrix of f_t given f_{t-1}.

In order to apply sequential EIS to this model, we first note that the factor $\varphi_t(f_t, f_{t-1}; \psi)$ in the likelihood integral (8) and (11) is given by

$$\varphi_t\left(f_t, f_{t-1}; \psi\right) = p\left(f_t \mid f_{t-1}\right) \cdot \left[\prod_{j=1}^{J} p\left(y_{tj} \mid \phi_{tj}\right)\right], \tag{18}$$

where $p\left(f_t \mid f_{t-1}\right)$ is linear gaussian and ϕ_{tj} is a linear function of f_t. Next, note that if $k_t\left(f_t, f_{t-1}; a_t\right)$ is a gaussian kernel in both f_t and f_{t-1}, then its integrating constant w.r.t. f_t given by $\chi_t\left(f_{t-1}; a_t\right)$ is a gaussian kernel in f_{t-1}. By recursion this implies that the sole non–gaussian term in the product $\varphi_t \chi_{t+1}$ to be approximated by k_t is the product of the J densities $p\left(y_{tj} \mid \phi_{tj}\right)$. It follows that all we have to do is to construct gaussian approximates in ϕ_{tj} to the latter densities in order to produce a gaussian kernel k_t for (f_t, f_{t-1}). The kernel k_t then consists of the product of $p\left(f_t \mid f_{t-1}\right)$ by J univariate gaussian kernels in the ϕ_{tj}'s and by χ_{t+1}. Moreover, the factors $p\left(f_t \mid f_{t-1}\right)$ and χ_{t+1} appear in logs on both sides of the auxiliary EIS regressions in Equation (12) and cancel out. All in all, the EIS auxiliary regression for the approximation of $\varphi_t \chi_{t+1}$ by k_t simplifies into J independent bivariate linear LS regressions of $\{\ln p(y_{tj} \mid \tilde{\phi}_{tj}^{(i)})\}_{i=1}^N$ on $\{(\tilde{\phi}_{tj}^{(i)}\ ,\ [\tilde{\phi}_{tj}^{(i)}]^2)\}_{i=1}^N$ and a constant. These auxiliary regressions run fast and produce numerically very accurate evaluations of the likelihood function, rendering ML-EIS estimation of the model fully operational.

The corresponding matrix algebra, which essentially consists of regrouping three gaussian kernels in (f_t, f_{t-1}) and integrating out f_t, is conceptually straightforward. Details are regrouped in the Appendix.

Last but not least, note that if we replace the Poisson density by the Negbin density in Equation (7), we only need to modify accordingly the dependent variables in the auxiliary EIS regressions, a trivial adjustment all together.

3.3 Filtering and smoothing

In many state-space applications such as the one analyzed here, interest lies also in the estimation of the latent states (i.e. in our application the factors) whether for diagnostic checking, interpretation and/or forecasting. Since, however, factors are one-time occurrences (incidental in statistical jargon) they obviously cannot be consistently estimated. Nevertheless, their moments conditional upon alternative information sets are functions of the parameters of the model and can therefore be consistently estimated.

The filtered moments of f_t are defined as being conditional upon information available up to time $t-1$ denoted by Y_{t-1}. In the present paper we shall compute means and variances of $\exp(\gamma_j' f_t)$, where γ_j' denotes the jth row of Γ. These moments are instrumental in the computation of the standardized Pearson residuals

$$z_{tj} = \frac{y_{tj} - \mathrm{E}(y_{tj}|Y_{t-1})}{\mathrm{Var}(y_{tj}|Y_{t-1})^{1/2}}. \tag{19}$$

These residuals are critical components of a variety of diagnostic statistics since they should have zero mean and unit variance and should be serially uncorrelated if the model is correctly specified. Under the Poisson model the relevant conditional moments of y_{tj} are given by

$$\mathrm{E}(y_{tj}|Y_{t-1}) = \exp\{\mu_j\} \cdot \mathrm{E}(\exp\{\gamma_j' f_t\}|Y_{t-1}), \tag{20}$$

and

$$\mathrm{Var}(y_{tj}|Y_{t-1}) = \exp\{\mu_j\} \cdot \mathrm{E}(\exp\{\gamma_j' f_t\}|Y_{t-1}) + \exp\{2\mu_j\} \cdot \mathrm{Var}(\exp\{\gamma_j' f_t\}|Y_{t-1}), \tag{21}$$

respectively. The filtered moments of $\exp(\gamma_j' f_t)$ take the form of ratios of integrals in $\{f_r\}_{r=1}^{t}$ which are functionally similar to the likelihood integral in Equation (8) with products running only up to period $t-1$. Both numerator and denominator can be accurately approximated by EIS. Moreover, both EIS approximations should use the same set of CRNs in order to induce positive correlation between numerator and denominator, resulting in additional efficiency gains.

Smoothed moments of f_t are defined as being conditional on the entire sample Y_T and are also computed by EIS (and are typically very close to the moments of the EIS samplers since the latter can be interpreted as approximations of the posterior densities of the factors). Smoothed moments provide, therefore, an ex-post image of the factor history over the sample period.

4. Application to Stock-Market Trading Volume

4.1 The data

Different versions of the dynamic factor model introduced in Section 2 are applied to the number of trades in 5-minute intervals between 9:45 AM and 4:00 PM for $J = 5$ stocks traded at the NYSE: Two companies - P.H. Glatfelter Company (GLT) and Wausau Paper Corporation (WPP) – belong to the industry subsector paper; three companies – Empire District Electric Company (EDE), Northeast Utilities (NU) and Westar Energy, Inc. (WR) – belong to the industry subsector conventional electricity. Data are taken from the TAQ (Trades and Quotes) data set, provided by the NYSE. The time period covered is the first quarter of 2005 (January 3, 2005 – March 31, 2005) with 61 trading days. As there are 75 5-minute intervals per day, the sample size is $T = 4575$. See the top panel of Figure 3 for time series plots of the number of trades. Descriptive statistics are provided in Table 1, and Table 2 reports the sample correlations across the five stocks. As one can see, the empirical distribution of the number of trades is clearly over–dispersed. The Ljung–Box statistics for the number of trades Q_{10} and Q_{20} including 10 and 20 lags, respectively, indicate strong serial correlation. As shown in Table 2, contemporaneous correlation between the trading activities of the five stocks are all positive.

4.2 Daily trading pattern

It is well–known that daily trading activity has a distinctive U-shape pattern (see, e.g., Admati and Pfleiderer, 1988). In order to capture it we introduce a Fourier series for the intercept of the log–mean function (see Equation 3). Specifically, μ_j is replaced by a cyclical term μ_{tj} defined as

$$\mu_{tj} = \mu_j + \alpha_j' x_t, \tag{22}$$

with $\alpha_j' = (\alpha_{1j}, \ldots, \alpha_{4j})$ and $x_t' = (\cos(2\pi t/75), \sin(2\pi t/75), \cos(4\pi t/75), \sin(4\pi t/75))$, accounting for the fact that there are 75 5-minutes intervals in a trading day. EIS trivially accommodates this extension. The filtering equations (20) and (21) are modified as follows:

$$\mathrm{E}(y_{tj}|Y_{t-1}, x_t) = \exp\{\mu_j + \alpha_j' x_t\} \cdot \mathrm{E}(\exp\{\gamma_j' f_t\}|Y_{t-1}, x_t), \tag{23}$$

$$\begin{aligned} \mathrm{Var}(y_{tj}|Y_{t-1}, x_t) \quad = \quad & \exp\{\mu_j + \alpha_j' x_t\} \cdot \mathrm{E}(\exp\{\gamma_j' f_t\}|Y_{t-1}, x_t) \\ & + \exp\{2(\mu_j + \alpha_j' x_t)\} \cdot \mathrm{Var}(\exp\{\gamma_j' f_t\}|Y_{t-1}, x_t). \end{aligned} \tag{24}$$

4.3 Univariate Analysis

As initial step, we first estimate a univariate dynamic Poisson model for each of the five stocks separately. This model is defined by Equations (1) and (6) – with $\kappa^{\omega_j} = 0$ – together with

$$\ln \theta_{tj} = \mu_{tj} + \omega_{tj}, \tag{25}$$

and Equation (22). This univariate (parameter–driven) dynamic Poisson model was introduced by Zeger (1988) and analyzed by Chan and Ledolter (1995), Kuk and Cheng (1997), Jung and Liesenfeld (2001), and Jung et al. (2006). The ML-EIS estimation results based on a MC sample size of $N = 50$ are found in Table 3. Most importantly, we find that the parameters governing the stochastic latent processes $\{\omega_{tj}\}_{j=1}^{5}$ and those characterizing the diurnal patterns are quite similar across the five stocks. In particular, estimates of δ^{ω_j} range from 0.60 to 0.72 and are indicative of strong persistence, while the estimates of ν^{ω_j} range from 0.30 to 0.50. These findings motivate our subsequent multivariate analysis where we shall aim at identifying common factors. They also allow us to impose in Equation (22) a common diurnal pattern to the five stocks obtained by setting the vectors $\alpha_j' = \alpha' = (\alpha_1, \ldots, \alpha_4)$ for $j = 1, ..., J$, thereby preserving parsimony in the multivariate specification.

4.4 Multivariate Factor Models

4.4.1 Poisson Model with one common factor

Allowing for a single common factor λ_t in addition to the idiosyncratic factors ω_{tj}, the log-mean function in the conditional Poisson distribution (1) for stock j is now given by

$$\ln \theta_{tj} = \mu_{tj} + \gamma_j^{\lambda} \lambda_t + \omega_{tj}, \tag{26}$$

with $\gamma_1^{\lambda} = 1$, together with Equations (4), (6) and (22) – under the restriction $\alpha_j = \alpha$. Joint ML-EIS estimates based upon $N = 50$ trajectories are found in Table 4. ML-EIS estimation requires approximately 65 BFGS iterations and takes of the order of 100 minutes on a Core 2 Duo Intel 2.7 GHz processor using GAUSS on Windows XP. (We also experimented with the Nelder-Mead simplex method for maximizing the log-likelihood functions (see, e.g., Press et al., 1988). It turned out that it produces the same results and requires about the same computing time as the BFGS algorithm. However, for higher dimensional factor models and/or less well-behaved likelihood functions we advise the use of this gradient free simplex algorithm.) MC numerical standard deviations of the ML-EIS parameter estimates used as measures of numerical precision are obtained from 20 i.i.d. ML-EIS estimations conducted under different CRN seeds (see Richard and Zhang, 2007 for details). They indicate that the parameter estimates are numerically very accurate. The fact that such high accuracy obtains with as little as $N = 50$ trajectories indicates that the likelihood integrands in Equation (11) are very well-behaved functions of the 27,450 latent factor variables, which are accurately approximated by the EIS-sampler (using 54,900 auxiliary parameters). In particular, the R^2s of the EIS auxiliary LS-problems (12) are typically larger than 0.99.

All parameter estimates are reasonable and apart from α_4 significant at the 1% significance level. The estimates of the parameters for the common process δ^λ and ν^λ indicate a substantial variation and a slight, yet significant, persistence. The estimates of the factor loadings ranging from 0.53 to 1.24 suggest that the trading activity of all stocks load significantly on the common factor, which is not surprising as trading is positively correlated across stocks. The estimates of the parameters characterizing the idiosyncratic factors indicate substantially more persistence than for the common factor as well as uniformly more persistence than their univariate counterparts in Table 3. Hence, the idiosyncratic factors capture the persistent movements of the trading process, whilst the common factor accounts for the more transitory variation. Note, furthermore, that the estimated α-parameters governing the deterministic seasonal effects are similar in magnitude to those obtained under the univariate models (see Table 3). Figure 1 shows the estimated diurnal seasonal effects for the number of trades obtained under the dynamic factor model. They exhibit the well-documented U-shape pattern. The sum of the individual log-likelihood values for the five independent univariate models equals -62,134 (see Table 3) which is substantially smaller than the log-likelihood value of -61,631 for the multivariate factor model. This large difference reflects the fact that, as shown below, the common factor model fully accounts for observed correlations between trading activities, in sharp contrast with the univariate models which ignore them.

In order to assess the reliability and the statistical properties of the ML-EIS estimator in this multivariate factor model we conducted a small simulation experiment, in which we drew 20 fictitious samples of size 4575 from that model setting the parameters equal to their estimates obtained from the real data. MC mean and standard deviation of the ML-EIS estimates obtained for the fictitious samples are provided in Table 5 and indicate that the ML-EIS estimation procedure is statistically very well behaved. Figure 2 shows the time series of the true log conditional mean $\ln \theta_{tj}$ of the first count data series for the first 500 time periods together with its smoothed estimates $\mathrm{E}(\ln \theta_{tj}|Y)$ obtained for simulated data. Unsurprisingly, the series of smoothed estimates closely follows the true value.

The parameter estimates in Table 4 can be used to compute the implied estimates of the means and the covariance of the unconditional distribution for the number of trades, to be compared with their sample counterparts. In the presence of deterministic diurnal effects, the unconditional means and variances for the trades of stock j under the factor model are

$$\mathrm{E}(y_{tj}) = E_T\left(\exp\{\mu_{tj} + 0.5\gamma_j'\Sigma_f\gamma_j\}\right) \tag{27}$$

and

$$\begin{aligned} \mathrm{Var}(y_{tj}) \quad = \quad & \mathrm{Var}_T\left(\exp\{\mu_{tj} + 0.5\gamma_j'\Sigma_f\gamma_j\}\right) \\ & + E_T\left(\exp\{2\mu_{tj} + \gamma_j'\Sigma_f\gamma_j\}\left[\exp\{\gamma_j'\Sigma_f\gamma_j\} - 1\right] + \exp\{\mu_{tj} + 0.5\gamma_j'\Sigma_f\gamma_j\}\right), \end{aligned} \tag{28}$$

where γ_j represents the vector of the factor loadings for stock j and Σ_f the covariance matrix of the vector of factors f_t. The notation E_T and Var_T indicate sample mean and sample variance computed w.r.t. the deterministic variation of the diurnal seasonal effects. The corresponding covariance between trading of stock j and stock k is obtained from the cross moments

$$\mathrm{E}(y_{tj}y_{tk}) = E_T\left(\exp\{\mu_{tj} + \mu_{tk} + 0.5(\gamma_j + \gamma_k)'\Sigma_f(\gamma_j + \gamma_k)\}\right), \qquad j \neq k. \tag{29}$$

The estimates of the unconditional mean and covariance matrix of y_t are given by

$$\hat{\mathrm{E}}(y_t) = \begin{pmatrix} 5.83 \\ 7.93 \\ 3.45 \\ 10.41 \\ 9.72 \end{pmatrix} \quad \text{and} \quad \hat{\mathrm{Var}}(y_t) = \begin{pmatrix} 16.93 & & & & \\ 5.46 & 38.70 & & & \\ 3.11 & 3.76 & 9.55 & & \\ 5.94 & 7.40 & 4.02 & 33.11 & \\ 4.99 & 6.27 & 3.34 & 7.19 & 39.07 \end{pmatrix},$$

respectively. The corresponding sample moments are given by

$$\bar{y} = \begin{pmatrix} 5.80 \\ 7.90 \\ 3.47 \\ 10.41 \\ 9.61 \end{pmatrix} \quad \text{and} \quad \Sigma_{\bar{y}} = \begin{pmatrix} 16.75 & & & & \\ 5.68 & 34.17 & & & \\ 3.26 & 4.65 & 9.65 & & \\ 4.95 & 5.97 & 5.17 & 34.41 & \\ 4.09 & 6.32 & 3.37 & 7.47 & 34.89 \end{pmatrix},$$

respectively. The close match between those two sets of moments indicates that the common factor model provides an excellent parsimonious representation of the contemporaneous correlation across the five stocks.

In order to assess the relative importance of the common factor, idiosyncratic factors and the diurnal component we computed their relative contribution to the overall variance of the log-link function for the individual stocks $\ln\theta_{tj}$. The implied estimates of the contributions to the variation of the log-link function are reported in the upper panel of Table 8. The fraction of variation explained by the common factor varies between 8% (WR) and 31% (EDE) while that of the idiosyncratic factors range between 56% (GLT) and 76% (WR). Even though the idiosyncratic factors explain a larger fraction of variation in $\ln\theta_{tj}$ than the common factor, it is important to remember that the latter is indispensable to capture observed contemporaneous correlations.

For diagnostic checking we computed the standardized Pearson residuals z_{tj} as defined by Equations (19), (23) and (24). The conditional moments of $\exp\{\gamma_j' f_t\}$ appearing in the Equations (23) and (24) are filtered moments and are evaluated by EIS as described in Section 3.3 above. The upper panel of Table 9 summarizes the properties of the Pearson residuals. Their sample means are all close to zeros. However, their standard deviations are all substantially larger than 1. This indicates that there is more variation and over–dispersion in the data than the model accounts for. Furthermore, the Ljung-Box statistics for the residuals including 10 and 20 lags indicates that the model does not fully capture the dynamic behavior of the trading activity even though it dramatically reduces the Ljung-Box statistic for the raw data as given in Table 1. The bottom row of Figure 3 displays the time series of the standardized residuals for the 5 stocks. The comparison with the time series plot of the raw data (see top row of Figure 3) reveals that the model accounts for a substantial fraction of the variation in the data.

4.4.2 Negbin Model with one common factor

In order to better account for over–dispersion and to allow for more flexibility to capture the serial correlation we replace the conditional Poisson density in Equation (1) by the more flexible Negbin density in Equation (7). As explained in Section 3.2 above, this substitution only requires minor modifications of the baseline EIS-ML algorithm.

The results of the ML-EIS estimation of the Negbin factor model with one common and J idiosyncratic factors are reported in Table 6. Note that the substitution of the Negbin for the Poisson distribution increases the value of the maximized likelihood function by 381, indicative of a much better fit. Moreover, the additional σ-parameters measuring the deviation from the Poisson distribution are in each case statistically significantly larger than zero at any conventional significance level. On the other hand the estimates of the intercept parameters (μ_j), the factor loadings (γ_j^λ), and of the seasonal parameters (α_i) obtained under the Poisson and the Negbin specification are very similar to each other, indicating a fairly robust factor structure. Note, however, that the δ-coefficients have increased under the Negbin model while the ν-parameters have decreased. This indicates that the factors, whether common or idiosyncratic evolve more smoothly over time and with greater persistence under the Negbin. These differences also suggest that a substantial part of the variation in trading activities, which was attributed to persistent shocks in the factor processes under the Poisson model, is now interpreted as transitory and attributed to conditional over–dispersion ($\sigma_j > 0$) under the Negbin model. (Such a substitution effect can also be observed for applications of the stochastic volatility models when the usual gaussian density for the conditional distribution of the returns is, in the presence of outliers, substituted by a fat-tailed student-t assumption, see, e.g., Liesenfeld and Jung, 2000). Note also that the MC numerical standard errors for the parameters common to both models are much smaller under the Negbin model, consistent with the fact that the latter better accounts for observed over–dispersion.

The relative contributions of the individual factors to the overall variance of the log-link function for the individual stocks obtained under the Negbin model are reported in the middle panel of Table 8. Note that the relative contribution of the common factor has increased slightly for all five stocks under the Negbin.

The middle panel of Table 9 summarizes the properties of the Pearson residuals z_{tj} obtained from the Negbin factor model. The values shown for the means (close to zero) and standard deviations (close to one) for all five stocks are indicative of a valid specification. Whence, the Negbin factor model appears to fully account for over–dispersion in the data, in contrast to the Poisson specification. Moreover, the Ljung-Box statistics for the residuals indicate that the Negbin model successfully accounts for serial correlation in the number of trades for GLT, EDE and NU. However, there remain difficulties capturing the full dynamics in trading activity for WPP and WR.

4.4.3 Negbin Model with a common and industry–specific factors

Since the five stocks considered here belong to two different industries, their trading volumes might be affected by industry–specific news in addition to market–wide news which are already captured by the common factor. Therefore, we now add two sector-specific factors to the Negbin model introduced in Section 4.4.2. The log-mean function for stock j is now given by

$$\ln \theta_{tj} = \mu_{tj} + \gamma_j^{\lambda} \lambda_t + \gamma_j^{\tau_{s_j}} \tau_{ts_j} + \omega_{tj}, \tag{30}$$

with $s_j = 1$ for $j = \{1, 2\}$ (GLT,WPP) and $s_j = 2$ for $j = \{3, 4, 5\}$ (EDE,NU,WR). Accounting for normalization ($\gamma_1^{\tau_1} = \gamma_3^{\tau_2} = 1$) this amounts to introducing seven additional parameters in the model: three loading factors ($\gamma_2^{\tau_1}$,$\gamma_4^{\tau_2}$,$\gamma_5^{\tau_2}$) and four factor parameters (δ^{τ_1},ν^{τ_1},δ^{τ_2},ν^{τ_2}).

The ML-EIS estimation results are summarized in Table 7. Note that the addition of the two industry factors (seven parameters) produces an additional significant increase of the likelihood function by 40. Moreover, most of the parameters common to both Negbin models are very similar, providing evidence of robustness in the factor specifications (see Tables 6 and 7). Two significant differences are observed for σ_1 (GLT) and ν^{ω_5} (WR). The addition of industry factors has noticeably decreased these two parameters. The estimates of the factor loadings for the industry factors suggest that the paper–industry factor τ_{t1} is essentially a GLT factor while that of the electric–service industry τ_{t2} is dominated by WR. This is confirmed by the finding that the industry factors account for 29% (GLT) and 27% (WR) of the variation in the conditional means of these two stocks, and only for 3% or less for the other three stocks (see bottom panel of Table 8). The negative signs of the NU and WR loadings on the electric–service factor reported in Table 7 are indicative of a 'substitution effect' in the trading activities of this sector, though both coefficients are statistically insignificant. All in all, our analysis suggests that the two industry–specific factors we have introduced actually capture additional firm–specific variations for GLT and WR, rather than genuine industry-specific factors. In particular, τ_{t1} captures mostly transitory movements in GLT trading, while τ_{t2} reflects mostly persistent movements in WR trading. Such interpretation is supported further by the finding that the over–dispersion parameter σ_1 (GLT) and the idiosyncratic volatility parameter ν^{ω_5} (WR) have both decreased following the addition of the two industry factors.

The properties of the Pearson residuals z_{tj} for the factor model with industry–specific factors are summarized in the bottom panel of Table 9. The model clearly accounts for most of the observed over–dispersion except possibly for GLT with a standard deviation of 1.14. However, closer inspection of the GLT–residuals indicate that this large standard deviation is essentially due to a single outlier. The Ljung–Box statistics indicate that the model successfully accounts for most of the observed serial correlation except possibly for WPP, whose dynamics have been difficult to parsimoniously capture under all three model specifications.

5. Conclusion

We can draw three sets of conclusions from the application we have presented in this paper. With respect to modelling multivariate time series of counts, we have illustrated that our proposed parsimonious and easy to interpret dynamic factor model is able to represent non-trivial contemporaneous and temporal interdependencies across count series. Hence, we expect that it provides a useful framework for the analysis of high-dimensional time series of counts.

In regard to our application to the number of trades for five NYSE stocks itself, we found robust evidence for a common factor reflecting market–wide news and accounting for observed co–movements in trading activities across the five individual stocks analyzed here. While the two industry–specific factors we added to the model do capture additional variations in trading activities, they appear to represent additional firm–specific factors for two stocks rather than genuine industry–specific factors. Last but not least, the Negbin clearly dominates the Poisson distribution in terms of accounting for observed over–dispersion and serial correlation of trading volumes.

From a numerical viewpoint, we have demonstrated that EIS enables one to analyze complex factor structures in the context of dynamic multivariate discrete models, at least as long as the dynamics of the model is specified in the form of gaussian autoregressive factors. (Work in progress should allow us to relax such restrictions but goes beyond the objectives of the present paper.) In the current application, EIS simplifies into a sequence of $J \cdot T$ bivariate auxiliary linear LS regressions, irrespective of the number P of factors and of the complexity of the factor structure. Moreover, numerically highly accurate ML-EIS parameter estimates obtain under very small numbers of draws ($N = 50$ trajectories for the present application). Last but not least, the baseline EIS algorithm requires only minor adjustments to accommodate alternative specifications (factor structure and/or discrete distribution) providing thereby unparalleled flexibility for the analysis of complex dynamic factor structures.

Acknowledgements

The second author acknowledges support by the Deutsche Forschungsgemeinschaft (Grant HE 2188/1-1), and the third author acknowledges support by the National Science Foundation (Grants SES-9223365 and SES-0516642).

Appendix: EIS Implementation

This appendix details the functional forms of the EIS implementation for the dynamic count data model given by Equations (1) and (3)–(6).

Let the integrating constant of the EIS gaussian kernel $k_t(f_t, f_{t-1}; \hat{a}_t)$ w.r.t. f_t be parameterized as

$$\chi_{t+1}(f_t; \hat{a}_{t+1}) = \exp -\frac{1}{2}\left(f_t' P_{t+1} f_t - 2 f_t' q_{t+1} + r_{t+1}\right), \tag{A.1}$$

where $(P_{t+1}, q_{t+1}, r_{t+1})$ denote appropriate functions of the EIS auxiliary parameter $\hat{a}_{t+1}$, to be obtained by backward recursions as described below. (Since $\chi_{T+1} \equiv 1$, the 'initial values' are $P_{T+1} = 0$, $q_{T+1} = 0$, $r_{T+1} = 0$.) Let the EIS–LS approximation of the product $\prod_{j=1}^{J} p(y_{tj} \mid \phi_{tj})$ in Equation (18) be denoted as

$$k_t^1(f_t; \hat{a}_t) = \exp -\frac{1}{2}\left(\phi_t' \hat{B}_t \phi_t - 2\phi_t' \hat{c}_t\right), \tag{A.2}$$

with $\phi_t = \mu + \Gamma f_t$. $\hat{B}_t = \mathrm{diag}(\hat{b}_{tj})$ denotes a $J \times J$ positive definite diagonal matrix and $\hat{c}_t = (\hat{c}_{tj})$ a J–dimensional vector. The EIS auxiliary parameter $\hat{a}_t$ is defined as $\hat{a}_t' = (\mathrm{vech}(\hat{B}_t)', \hat{c}_t')$. The EIS gaussian kernel k_t is then given by

$$k_t(f_t, f_{t-1}; \hat{a}_t) = k_t^1(f_t; \hat{a}_t)\, p(f_t \mid f_{t-1})\, \chi_{t+1}(f_t; \hat{a}_{t+1}). \tag{A.3}$$

Combining together Equations (17), (A.1) and (A.2) we have

$$\begin{aligned} -2\ln k_t(f_t, f_{t-1}; \hat{a}_t) &= (\mu + \Gamma f_t)' \hat{B}_t (\mu + \Gamma f_t) - 2(\mu + \Gamma f_t)' \hat{c}_t \\ &\quad + (f_t - \Delta f_{t-1})' H (f_t - \Delta f_{t-1}) + f_t' P_{t+1} f_t - 2 f_t' q_{t+1} + r_{t+1}. \end{aligned} \tag{A.4}$$

Completing the quadratic form in f_t (given f_{t-1}) we rewrite $-2\ln k_t$ as

$$\begin{aligned} -2\ln k_t(f_t, f_{t-1}; \hat{a}_t) &= [f_t - (d_t + G_t f_{t-1})]' M_t [f_t - (d_t + G_t f_{t-1})] \\ &\quad + \left(\mu' \hat{B}_t \mu - 2\mu' \hat{c}_t + f_{t-1}' \Delta' H \Delta f_{t-1} + r_{t+1}\right) \\ &\quad - (d_t + G_t f_{t-1})' M_t (d_t + G_t f_{t-1}), \end{aligned} \tag{A.5}$$

with

$$
\begin{aligned}
M_t &= \Gamma'\hat{B}_t\Gamma + H + P_{t+1}, && \text{(A.6)}\\
d_t &= M_t^{-1}\left[q_{t+1} + \Gamma'\left(\hat{c}_t - \hat{B}_t\mu\right)\right], && \text{(A.7)}\\
G_t &= M_t^{-1}H\Delta. && \text{(A.8)}
\end{aligned}
$$

It immediately follows that the gaussian EIS sampler for $f_t \mid f_{t-1}$ is given by

$$
m_t\left(f_t \mid f_{t-1}; \hat{a}_t\right) \sim \mathrm{N}\left(d_t + G_t f_t\ ,\ M_t^{-1}\right). \tag{A.9}
$$

The log integrating constant $-2\ln\chi_t\left(f_{t-1}; \hat{a}_t\right)$ obtains by regrouping all remaining factors in Equation (A.5) and is, therefore, of the form introduced in Equation (A.1) together with

$$
\begin{aligned}
P_t &= \Delta' H\Delta - G_t' M_t G_t, && \text{(A.10)}\\
q_t &= G_t' M_t d_t, && \text{(A.11)}\\
r_t &= \mu'\hat{B}_t\mu - 2\mu'\hat{c}_t + r_{t+1} - d_t' M_t d_t. && \text{(A.12)}
\end{aligned}
$$

Hence, Equations (A.6)–(A.8) and (A.10)–(A.12) fully characterize the EIS recursion whereby the coefficients $(P_{t+1}, q_{t+1}, r_{t+1})$ are combined with the period t EIS coefficients $(\hat{B}_t, \hat{c}_t)$ in order to produce (back recursively) the coefficients (M_t, d_t, G_t) characterizing the EIS-sampling densities.

Based on these functional forms the computation of the EIS estimate of the likelihood requires the following simple steps:

Step (1). Generate N independent P–dimensional trajectories $\{\{\tilde{f}_t^{(i)}\}_{t=1}^{T}\}_{i=1}^{N}$ from a sequence of initial samplers $\{m(f_t|f_{t-1}, a_t^{(0)})\}$. Such a sequence is obtained, e.g., by using as k_t^1 in Equation (A.2) a second-order Taylor-series approximation (TSA) in ϕ_{tj} to $\prod_{j=1}^{J} p(y_{tj} \mid \phi_{tj})$ around $\mathrm{E}(\phi_{tj}) = \mu_j$. The resulting TSA values of the auxiliary parameters $(a_t^{(0)})' = (\mathrm{vech}(B_t^{(0)})', (c_t^{(0)})')$ can be used to construct according to Equation (A.9) together with the recursions (A.6)–(A.8) and (A.10)–(A.12) an initial EIS sampler, which obtains by setting $\hat{B}_t = B_t^{(0)}$ and $\hat{c}_t = c_t^{(0)}$.

Step (2). Transform the simulated f_t–trajectories from the previous sampler into the corresponding

N independent J–dimensional ϕ_t–trajectories according to $\phi_t = \mu + \Gamma f_t$. Use the latter trajectories to solve for each period t the LS–problem defined in Equation (12). This requires to run for each period t the following J independent linear auxiliary regressions:

$$
\begin{aligned}
\ln p(y_{t1} \mid \tilde{\phi}_{t1}^{(i)}) &= \text{constant} - \frac{1}{2} b_{t1} [\tilde{\phi}_{t1}^{(i)}]^2 + c_{t1} \tilde{\phi}_{t1}^{(i)} + \zeta_{1t}^{(i)}, \qquad i = 1, ..., N, && \text{(A.13)} \\
&\vdots \\
\ln p(y_{tJ} \mid \tilde{\phi}_{tJ}^{(i)}) &= \text{constant} - \frac{1}{2} b_{tJ} [\tilde{\phi}_{tJ}^{(i)}]^2 + c_{tJ} \tilde{\phi}_{tJ}^{(i)} + \zeta_{Jt}^{(i)}, \qquad i = 1, ..., N, && \text{(A.14)}
\end{aligned}
$$

where $\zeta_{jt}^{(i)}$ denotes the regression error term of regression j.

Step (3). Use the LS estimates $\hat{B}_t = \text{diag}(\hat{b}_{tj})$ and $\hat{c}_t = (\hat{c}_{tj})$ obtained in Step (2) to construct back-recursively the sequence of EIS-sampling densities $\{m(f_t | f_{t-1}, \hat{a}_t)\}$ as given by Equation (A.9) together with the recursions (A.6)–(A.8) and (A.10)–(A.12).

Step (4). Generate N independent trajectories from the sequence of EIS samplers constructed in Step (3) and use them either to repeat Step (2) and (3) or (at convergence) to compute the EIS-MC estimate of the likelihood according to Equation (13).

References

Admati, A.R., Pfleiderer, P., 1988. A theory of intraday patterns: volume and price variability. The Review of Financial Studies 1, 3–40.

Andersen, T., 1996. Return volatility and trading volume: an information flow interpretation of stochastic volatility. Journal of Finance 51, 146–153.

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge.

Consul, P.C., 1989. Generalized Poisson distribution: Properties and Applications. Marcel Dekker, New York.

Chan, K.S., Ledolter, J., 1995. Monte Carlo EM estimation for time series models involving counts. Journal of the American Statistical Association 90, 242–251.

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

Heinen, A., 2003. Modelling time series count data: An autoregressive conditional Poisson model. Working paper.

Heinen, A., Rengifo, E., 2007. Multivariate autoregressive modeling of time series count data using copulas. Journal of Empirical Finance 14, 564–583.

Held, L., Höhle, M., Hofmann, M., 2005. A statistical framework for the analysis of multivariate infectious disease surveillance counts. Statistical Modelling 5, 187-199.

Jørgensen, B., Lundbye-Christensen, S., Song, P., Sun, L., 1999. A state space model for multivariate longitudinal count data. Biometrika 86, 169–181.

Jung, R.C., Liesenfeld, R., 2001. Estimating time series models for count data using efficient importance sampling. Allgemeines Statistisches Archiv 85, 387–407.

Jung, R.C., Kukuk, M., Liesenfeld, R., 2006. Time series of count data: modeling estimation and diagnostics. Computational Statistics & Data Analysis 51, 2350–2364.

Jung R.C., Tremayne, A.R., 2008. Count time series with overdispersed data. Working paper.

Kedem, B., Fokianos, K., 2002. Regression models for time series analysis. John Wiley & Sons, Hoboken, NJ.

Kuk, A.Y.C., Chen, Y.W., 1997. The Monte Carlo Newton-Raphson algorithm. Journal of Statistical and Computational Simulation 59, 233–250.

Liesenfeld, R., 2001. A generalized bivariate mixture model for stock price volatility and trading volume. Journal of Econometrics 104, 141–178.

Liesenfeld, R., Jung, R.C., 2000, Stochastic volatility models: conditional normality versus heavy-tailed distributions. Journal of Applied Econometrics 15, 137–160.

McKenzie, E., 2003. Discrete variate time series, in Shanbhag, D.N., Rao, C.R. (eds) Handbook of Statistics, Vol. 21, Elsevier Science, Amsterdam, Chapter 16.

Neal, P., Subba Rao, T., 2007. MCMC for integer-valued ARMA processes. Journal of Time Series Analysis 28, 92-110.

Press, H.W., Flannary, B.P., Teukolsky, S.A., Vettereing, W.T., 1988. Numerical Recipes in C. Cambridge University Press, Cambridge, UK.

Richard, J.-F., Zhang, W., 2007. Efficient high-dimensional importance sampling. Journal of Econometrics 141, 1385–1411.

Tauchen, G.E., Pitts, M. 1983. The price variability-volume relationship on speculative markets. Econometrica 51, 485–505.

Wedel, M., Böckenholt, U., Wagner, A.K., 2003. Factor models for multivariate count data. Journal of Multivariate Analysis 87, 356–369.

Winkelmann, R., 2008. Econometric analysis of count data. Springer, Heidelberg.

Zeger, S.L., 1988. A regression model for time series of counts. Biometrika 75, 621–629.

Table 1. Descriptive Statistics for the Number of Trades.

	GLT	WPP	EDE	NU	WR
Mean	5.80	7.90	3.47	10.41	9.61
Median	5	7	3	9	9
Standard dev.	4.09	5.85	3.11	5.87	5.91
Minimum	0	0	0	0	0
Maximum	54	43	25	48	59
Q_{10}	$2,086$	$4,057$	$1,575$	$3,452$	$4,150$
Q_{20}	$2,549$	$5,026$	$1,908$	$3,927$	$5,942$

NOTE: The number of observations per stock is $T = 4575$. The Ljung-Box Statistics Q_{10} and Q_{20} include 10 and 20 lags.

Table 2. Sample Correlation Matrix of the Number of Trades.

	GLT	WPP	EDE	NU	WR
GLT	1				
WPP	.238	1			
EDE	.256	.256	1		
NU	.206	.174	.283	1	
WR	.168	.183	.184	.216	1

Table 3. ML-EIS Estimates for the Univariate Dynamic Poisson Count Data Model for the TAQ Data.

	GLT	WPP	EDE	NU	WR
μ_j	1.619	1.869	1.030	2.246	2.140
	(.015)	(.018)	(.019)	(.013)	(.017)
δ^{ω_j}	.596	.616	.598	.619	.723
	(.021)	(.017)	(.022)	(.021)	(.016)
ν^{ω_j}	.400	.469	.496	.304	.330
	(.011)	(.010)	(.013)	(.008)	(.009)
α_{1j}	.250	.320	.339	.286	.201
	(.012)	(.007)	(.015)	(.017)	(.019)
α_{2j}	$-.024$	$-.125$	$-.091$	.006	$-.088$
	(.016)	(.007)	(.009)	(.008)	(.014)
α_{3j}	.009	.065	.045	.064	.034
	(.011)	(.009)	(.021)	(.009)	(.007)
α_{4j}	.044	$-.065$	$-.028$	$-.032$	$-.009$
	(.012)	(.022)	(.012)	(.009)	(.008)
Log-likelihood	$-11,833.5$	$-13,219.5$	$-10,264.6$	$-13,386.0$	$-13,430.2$
Sum of the Log-likelihood values			$-62,133.9$		

NOTE: The reported numbers are the ML-EIS estimates for the parameters, asymptotic standard errors obtained from a numerical approximation to the Hessian are in parentheses. ML-EIS estimates are based on a MC sample size of $N = 50$ and three EIS iterations.

Table 4. ML-EIS for the Poisson Factor Model with Firm Specific Factors and a Common Factor for the TAQ Data.

	GLT	WPP	EDE	NU	WR
μ_j	1.622	1.871	1.033	2.248	2.140
	(.015)	(.018)	(.018)	(.012)	(.016)
	[.0021]	[.0025]	[.0022]	[.0009]	[.0020]
γ_j^{λ}	1.000	.855	1.242	.635	.525
		(.048)	(.052)	(.039)	(.039)
		[.0105]	[.0114]	[.0071]	[.0084]
δ^{ω_j}	.814	.730	.786	.766	.786
	(.014)	(.012)	(.018)	(.018)	(.014)
	[.0043]	[.0034]	[.0031]	[.0029]	[.0022]
ν^{ω_j}	.232	.373	.306	.219	.281
	(.010)	(.008)	(.013)	(.009)	(.009)
	[.0030]	[.0030]	[.0026]	[.0013]	[.0017]
Factor param.	δ^{λ}	ν^{λ}			
	.152	.283			
	(.026)	(.007)			
	[.0051]	[.0017]			
Seasonal param.	α_1	α_2	α_3	α_4	Log-likelihood
	.275	$-.050$	.043	$-.016$	-61,630.7
	(.005)	(.009)	(.004)	(.010)	
	[.0019]	[.0014]	[.0014]	[.0010]	[2.354]

NOTE: The reported numbers are the ML-EIS estimates for the parameters, asymptotic standard errors obtained from a numerical approximation to the Hessian are in parentheses and MC (numerical) standard deviations obtained from 20 ML-EIS estimations conducted under different sets of CRNs are in brackets. ML-EIS estimates are based on a MC sample size of $N = 50$ and three EIS iterations.

Table 5. ML-EIS for the Poisson Factor Model with Firm Specific Factors and a Common Factor for Simulated Data.

		GLT	WPP	EDE	NU	WR
μ_j	true	1.622	1.871	1.033	2.248	2.140
	mean	1.625	1.940	1.055	2.252	2.142
	std.	.018	.055	.030	.013	.020
γ_j^λ	true	1.000	.855	1.242	.635	.525
	mean	1.000	.822	1.243	.630	.529
	std.		.051	.054	.031	.039
δ^{ω_j}	true	.814	.730	.786	.766	.786
	mean	.814	.744	.793	.759	.785
	std.	.015	.019	.013	.014	.013
ν^{ω_j}	true	.232	.373	.306	.219	.281
	mean	.229	.353	.294	.220	.278
	std.	.010	.016	.009	.007	.007
Factor param.		δ^λ	ν^λ			
	true	.152	.283			
	mean	.182	.288			
	std	.017	.008			
Seasonal param.		α_1	α_2	α_3	α_4	
	true	.275	$-.050$	.043	$-.016$	
	mean	.289	$-.051$	.045	$-.018$	
	std	.020	.022	.016	.016	

NOTE: The reported numbers are the MC mean and standard deviation of 20 repeated ML-EIS estimates for the parameters for different simulated data set and a fixed set of CRNs for ML-EIS estimation. ML-EIS estimates are based on a MC sample size of $N = 50$ and three EIS iterations.

Table 6. ML-EIS for the Negbin Factor Model with Firm Specific Factors and a Common Factor for the TAQ Data.

	GLT	WPP	EDE	NU	WR
μ_j	1.646	1.929	1.080	2.269	2.166
	(.022)	(.020)	(.023)	(.016)	(.021)
	[.0001]	[.0001]	[.0002]	[<.0001]	[<.0001]
γ_j^λ	1.000	.915	1.331	.681	.583
		(.053)	(.083)	(.037)	(.040)
		[.0006]	[.0014]	[.0005]	[.0004]
δ^{ω_j}	.939	.946	.927	.916	.957
	(.008)	(.008)	(.017)	(.008)	(.007)
	[.0001]	[.0001]	[.0003]	[.0002]	[<.0001]
ν^{ω_j}	.112	.135	.142	.110	.106
	(.008)	(.012)	(.023)	(.007)	(.008)
	[.0001]	[.0001]	[.0004]	[.0001]	[.0001]
σ_j	.281	.365	.341	.218	.288
	(.012)	(.011)	(.024)	(.009)	(.009)
	[.0003]	[.0002]	[.0005]	[.0001]	[.0001]
Factor param.	δ^λ	ν^λ			
	.351	.260			
	(.023)	(.009)			
	[.0002]	[.0001]			
Seasonal param.	α_1	α_2	α_3	α_4	Log-likelihood
	.271	$-.050$	.044	$-.015$	-61,249.9
	(.004)	(.009)	(.003)	(.004)	
	[<.0001]	[<.0001]	[<.0001]	[<.0001]	[0.138]

NOTE: The reported numbers are the ML-EIS estimates for the parameters, asymptotic standard errors obtained from a numerical approximation to the Hessian are in parentheses and MC (numerical) standard deviations obtained from 20 ML-EIS estimations conducted under different sets of CRNs are in brackets. ML-EIS estimates are based on a MC sample size of $N = 50$ and three EIS iterations.

Table 7. ML-EIS for the Negbin Factor Model with Firm- and Industry Specific Factors and a Common Factor for the TAQ Data.

	GLT	WPP	EDE	NU	WR	
μ_j	1.619	1.931	1.072	2.269	2.140	
	(.027)	(.033)	(.039)	(.018)	(.047)	
	[.0034]	[.0003]	[.0024]	[.0001]	[.0005]	
γ_j^{λ}	1.000	.955	1.767	.735	.686	
		(.067)	(.202)	(.062)	(.105)	
		[.0022]	[.0536]	[.0077]	[.0229]	
$\gamma_j^{\tau_1}$	1.000	.224	0.000	0.000	0.000	
		(.096)				
		[.0230]				
$\gamma_j^{\tau_2}$	0.000	0.000	1.000	$-.504$	-2.386	
				(.334)	(1.408)	
				[.0884]	[.4317]	
δ^{ω_j}	.951	.944	.965	.914	.986	
	(.009)	(.009)	(.015)	(.043)	(.005)	
	[.0006]	[.0003]	[.0034]	[.0004]	[.0007]	
ν^{ω_j}	.095	.138	.086	.111	.052	
	(.010)	(.013)	(.025)	(.010)	(.011)	
	[.0008]	[.0004]	[.0056]	[.0003]	[.0018]	
σ_j	.160	.370	.294	.219	.226	
	(.054)	(.013)	(.041)	(.011)	(.020)	
	[.0221]	[.0009]	[.0116]	[.0005]	[.0026]	
Factor param.	δ^{λ}	ν^{λ}	δ^{τ_1}	ν^{τ_1}	δ^{τ_2}	ν^{τ_2}
	.372	.229	.263	.273	.608	.083
	(.033)	(.016)	(.092)	(.036)	(.086)	(.044)
	[.0037]	[.0032]	[.0216]	[.0134]	[.0169]	[.0128]
Seasonal param.	α_1	α_2	α_3	α_4	Log-likelihood	
	.262	$-.056$	.042	$-.014$	-61,210.0	
	(.012)	(.011)	(.009)	(.009)		
	[.0005]	[.0001]	[.0001]	[.0001]	[1.001]	

NOTE: The reported numbers are the ML-EIS estimates for the parameters, asymptotic standard errors obtained from a numerical approximation to the Hessian are in parentheses and MC (numerical) standard deviations obtained from 20 ML-EIS estimations conducted under different sets of CRNs are in brackets. ML-EIS estimates are based on a MC sample size of $N = 50$ and three EIS iterations.

Table 8. Relative Contributions of the Factors to the Overall Variance of the Log Conditional Mean.

	seasonal	unique	common	industry
	Poisson Model: Firm Specific Factors & Common Factor			
GLT	.14	.56	.29	
WPP	.10	.74	.15	
EDE	.10	.59	.31	
NU	.21	.61	.18	
WR	.15	.76	.08	
	Negbin Model: Firm Specific Factors & Common Factor			
GLT	.18	.48	.35	
WPP	.14	.62	.23	
EDE	.12	.45	.43	
NU	.26	.50	.24	
WR	.19	.67	.13	
	Negbin Model: Firm- and Industry Specific Factors & Common Factor			
GLT	.14	.35	.22	.29
WPP	.14	.64	.21	.01
EDE	.11	.30	.55	.03
NU	.25	.51	.22	.02
WR	.16	.43	.13	.27

Table 9. Diagnostics Based on the Pearson Residuals.

	GLT	WPP	EDE	NU	WR
			Poisson Model: Firm Specific Factors & Common Factor		
Mean	.05	.07	.05	.04	.04
Standard dev.	1.24	1.29	1.21	1.20	1.22
Q_{10}	25.0 (.005)	60.9 (.000)	6.9 (.732)	29.6 (.001)	48.1 (.000)
Q_{20}	56.0 (.000)	128.1 (.000)	29.3 (.080)	62.1 (.000)	95.4 (.000)
			Negbin Model: Firm Specific Factors & Common Factor		
Mean	.01	.00	.00	.00	.01
Standard dev.	1.05	1.01	1.01	1.03	1.02
Q_{10}	12.9 (.228)	27.5 (.002)	10.6 (.391)	12.8 (.237)	31.8 (.000)
Q_{20}	21.2 (.383)	48.2 (.000)	21.8 (.353)	29.3 (.081)	38.1 (.008)
			Negbin Model: Firm- and Industry Specific Factors & Common Factor		
Mean	.03	.01	.02	.02	.02
Standard dev.	1.14	1.06	1.09	1.07	1.09
Q_{10}	10.1 (.428)	20.0 (.029)	14.8 (.140)	11.9 (.229)	15.5 (.116)
Q_{20}	27.1 (.133)	37.5 (.010)	28.2 (.104)	31.5 (.049)	20.0 (.454)

NOTE: The Ljung-Box Statistic Q_{20} includes 20 lags. Probability values are given in parentheses.

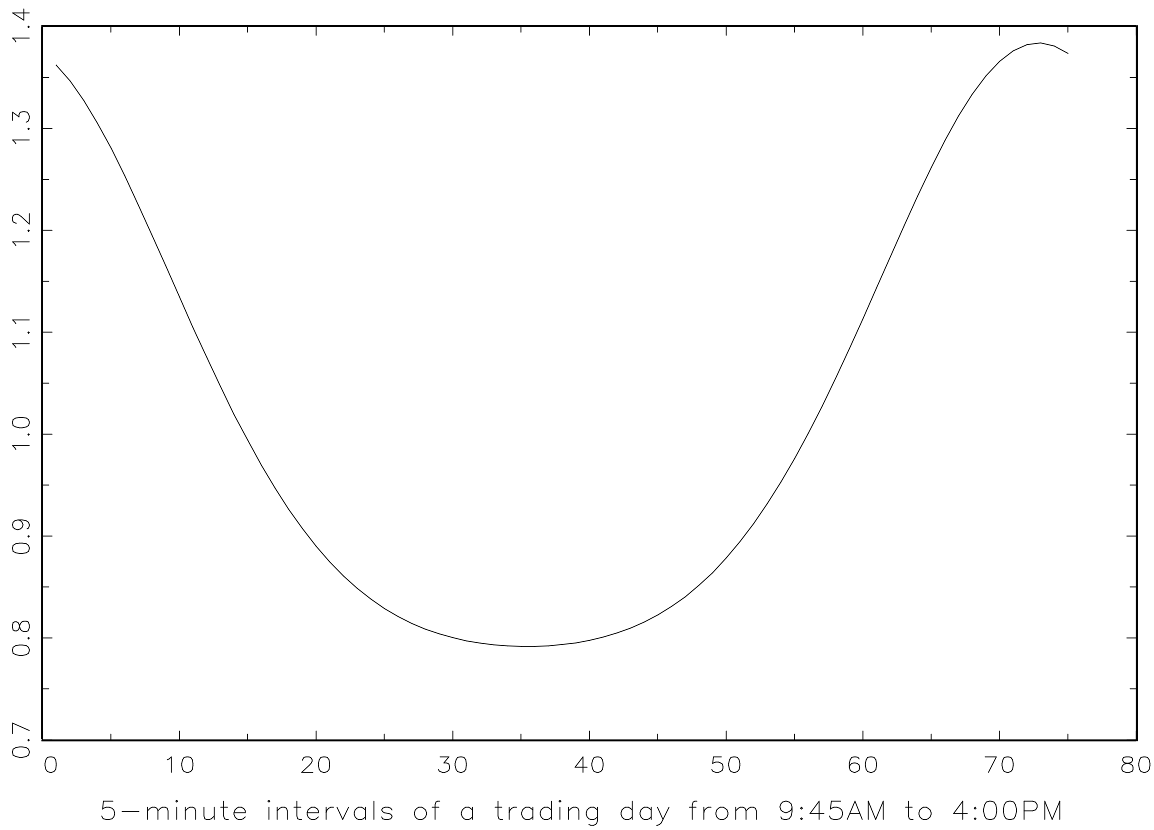

Figure 1. Estimated diurnal seasonal effects for the number of trades obtained under the dynamic Poisson factor model with one common factor given by $\exp\{0.275\cos(2\pi t/75) - 0.050\sin(2\pi t/75) + 0.043\cos(4\pi t/75) - 0.016\sin(4\pi t/75)\}$.

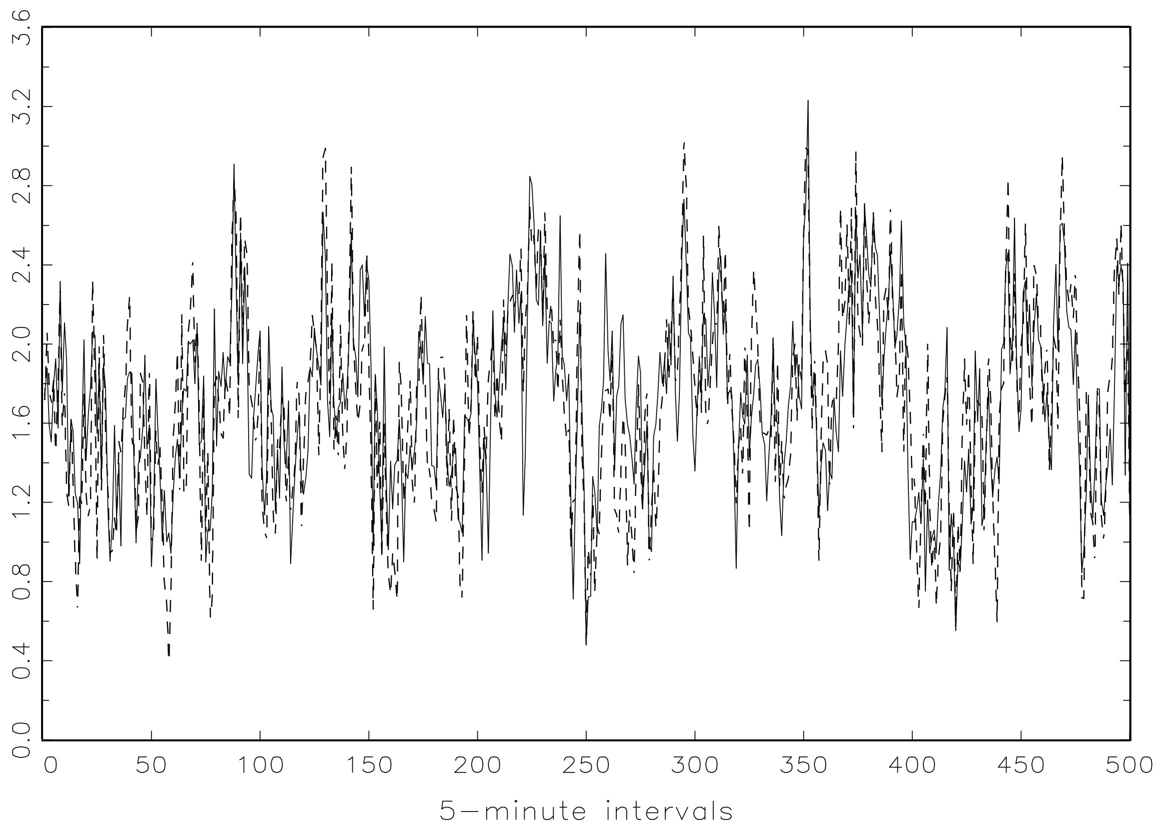

Figure 2. True $\ln \theta_{t1}$ (dashed line) and its smoothed estimates (solid line) for the first 500 time periods obtained for simulated data of the Poisson factor model with one common factor.

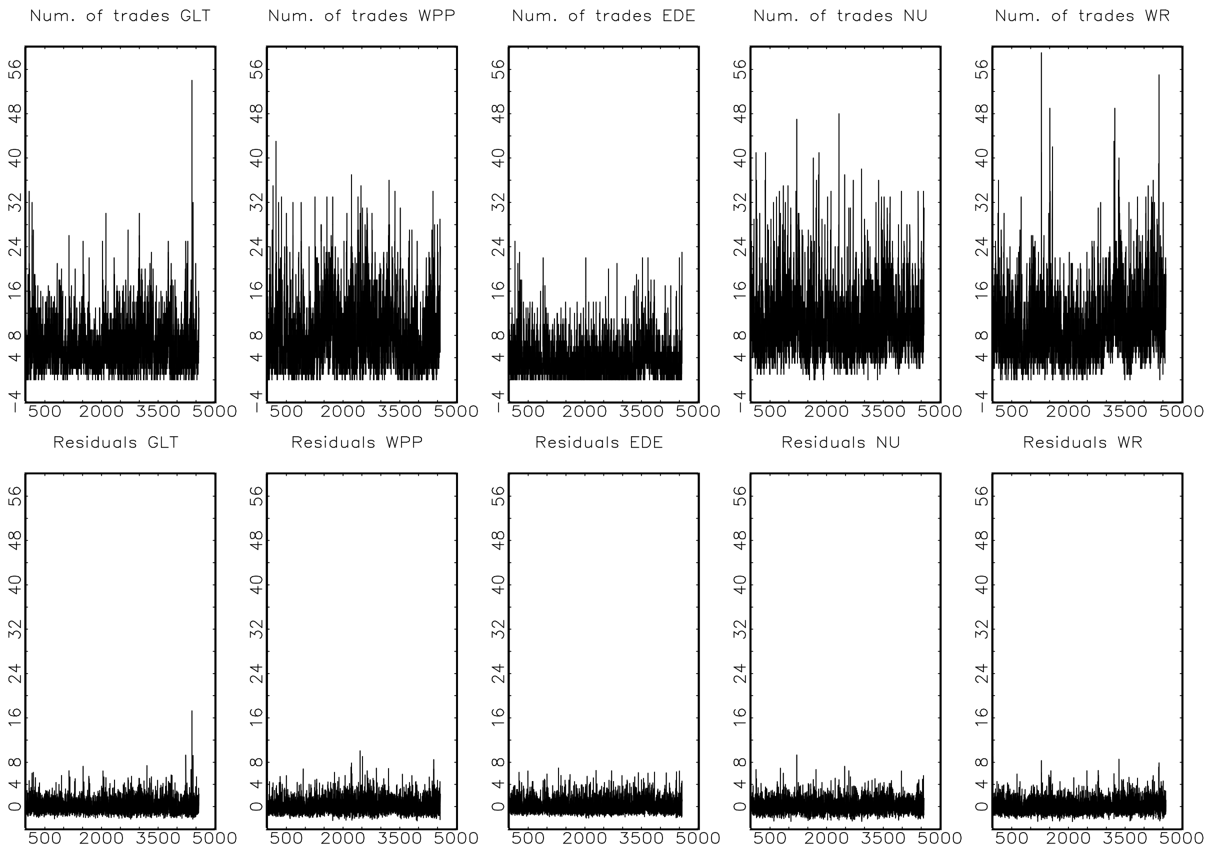

Figure 3. Time series of the number of trades in 5-minute intervals y_{tj} (top row) and of the standardized residuals z_{tj} obtained from the dynamic Poisson factor model with one common factor (bottom row).