Dynamic Models for Multivariate Time Series of Count Data

Ruey S. Tsay*

Abstract

We consider a grouped autoregressive intensity with momentum (GAIM) model for analysis of multivariate time series of count data. The model may include observed or latent common factors. We demonstrate the importance of the momentum factor and application of the proposed GAIM model via empirical data analysis.

Key Words: Factor Model, Intensity, Momentum, Negative Binomial, Poisson distribution

1. Introduction

Analysis of count data has a long history, e.g., Cameron and Trivedi (2013), Winkelmann (2010) and the references therein. For serially correlated count data, the literature is also extensive; see, for example, Zeger (1988), Kedem and Fokianos (2002), Davis, Dunsmuir and Streett (2003), Davis and Wu (2005), and Jung, Kukuk and Liesenfeld (2006). In recent years, much work in financial econometrics focuses on autoregressive intensity models for analysis of positive-valued time series; see Engle and Russell (1998), Heinen (2003), Hautsch (2012) and the references therein. On the other hand, multivariate time series of count data is less investigated. Lung, Liesenfeld, and Richard (2011) consider a dynamic factor model with latent common factors and use efficient important sampling (EIS) in estimation.

In this paper, we propose a grouped autoregressive intensity with momentum (GAIM) model for modeling multivariate count data. Our goal is to suggest a multivariate model that is easy to estimate yet widely applicable. To begin, we start with a univariate autoregressive intensity model. Let y_t be a scalar time series of counts. For instance, y_t may represent the number of trades in a 5-minute time interval of an asset. As another example, y_t may represent the number of hourly visitors to a particular web site. An effective model for y_t is the conditional autoregressive intensity model of Heinen (2003). Let F_{t-1} denote the information available at time $t-1$ and $\lambda_t = E(y_t|F_{t-1})$ be the conditional expectation of y_t given F_{t-1}, then a simple autoregressive intensity model for y_t is

$$y_t|F_{t-1} \sim p(\lambda_t|\theta) \quad (1)$$

$$\lambda_t = \omega + \alpha_1 y_{t-1} + \gamma_1 \lambda_{t-1}, \quad (2)$$

where $p(\lambda_t|\theta)$ denotes a probability mass function (pmf) over the non-negative integers such that $E(y_t|F_{t-1}) = \lambda_t$, θ represents a constant parameter, $\omega > 0$, $0 \leq \alpha_1, \gamma_1 \leq 1$, and $\alpha_1 + \gamma_1 < 1$. The model in Equation (2) can be extended to higher-order ones, but we shall focus on the (1,1) model for simplicity. The condition $\alpha_1 + \gamma_1 < 1$ is needed so that the unconditional expectation of λ_t is finite. We entertain three discrete distributions for $p(\lambda_t|\theta)$ in this paper. They are

1. Poisson: $y_t|F_{t-1} \sim Po(\lambda_t)$ with pmf

$$Po(y_t|\lambda_t) = \frac{\exp(-\lambda_t)\lambda_t^{y_t}}{y_t!}, \quad \lambda_t > 0.$$

*Booth School of Business, The University of Chicago, 5807 S. Woodlawn Avenue, Chicago, IL 60637. This research is supported in part by the Booth School of Business, University of Chicago.

Here $E(y_t|F_{t-1})$ = Var$(y_t|F_{t-1}) = \lambda_t$.

2. Negative Binomial: $y_t|F_{t-1} \sim NB(\theta, \lambda_t)$ with pmf

$$NB(y_t|\theta,\lambda_t) = \frac{\Gamma(y_t+\theta)}{y_t!\Gamma(\theta)}\left(\frac{\theta}{\lambda_t+\theta}\right)^{\theta}\left(\frac{\lambda_t}{\lambda_t+\theta}\right)^{y_t},$$

where $\lambda_t > 0$ and $\theta > 0$. Here $E(y_t|F_{t-1}) = \lambda_t$ and Var$(y_t|F_{t-1}) = \lambda_t + \frac{\lambda_t^2}{\theta}$.

3. Double Poisson: See Efron (1986)

$$\begin{aligned} DP(y_t|\theta,\lambda_t) &= c(\theta,\lambda_t)\theta^{1/2}[Po(y_t|\lambda_t)]^{\theta}[Po(y_t|y_t)]^{1-\theta} \\ &= c(\theta,\lambda_t)\theta e^{-\theta\lambda_t}\left(\frac{e^{-y_t}y_t^{y_t}}{y_t!}\right)\left(\frac{e\lambda_t}{y_t}\right)^{\theta y_t}, \end{aligned}$$

where $\lambda_t > 0$, $\theta > 0$, and $1/c(\theta,\lambda_t) \approx 1 + \frac{1-\theta}{12\lambda_t\gamma}\left(1+\frac{1}{\lambda_t\theta}\right)$. Here $E(y_t|F_{t-1}) = \lambda_t$ and Var$(y_t|F_{t-1}) \approx \lambda_t/\theta$.

Clearly, negative Binomial allows for over-dispersion whereas double Poisson can describe both under- and over-dispersion. If some explanatory variables $\boldsymbol{x}_t$ are available, then the model can be modified so that

$$E(y_t|F_{t-1},\boldsymbol{x}_t) = \exp(\boldsymbol{x}_t'\boldsymbol{\beta}) \times \lambda_t$$

where λ_t is given in Equations (1) and (2). For high-frequency stock transaction data, $\boldsymbol{x}_t$ may contain some indicator variables or trigonometric series to describe the diurnal pattern of trading activity. See, for example, Hautsch (2012).

2. Momentum Factors

While the intensity model in Equations (1) and (2) is widely used, it may encounter some difficulty in capturing details of the dynamic dependence of y_t. To illustrate, let y_t be the number of trades in 5-minute time interval of the stock of Glatfelter Company (GLT) from January 3 to March 31, 2005. This series is used in Lung, Liesenfeld, and Richard (2011). Following the prior work, we omit the trading activity from 9:30 AM to 9:45 AM so that there are 75 observations in each trading day. To handle the diurnal pattern of trading activity, we employed eight explanatory variables: they are the indicators for the first four 5-minute intervals from market open and the last four 5-minute intervals to market close. That is, x_{1t} is the indicator for time interval from 9:45:01 to 9:50 AM and x_{5t} is the indicator for time interval from 15:55:01 to 16:00 PM. The fitted autoregressive model for y_t with negative Binomial distribution is

$$\lambda_t = 0.483(0.07) + 0.171(0.01)y_{t-1} + 0.741(0.02)\lambda_{t-1},$$

and $\hat{\theta}$ = 4.91(0.20), where the number in parentheses denotes standard error. This model fits the data reasonably well except that there remain some serial correlations in the standardized residuals. For instance, the Ljung-Box statistics show $Q(20)$ = 33.97 with p-value 0.027.

To improve the model, we consider a momentum factor. For a given integer $m > 1$, define the momentum factor

$$f_{m,t-1} = (y_{t-1} + \cdots + y_{t-m})/m, \quad m > 1.$$

We refer to $f_{m,t-1}$ as a momentum factor because it represents a smoothed version of local trend. The intensity model is then modified to

$$\lambda_t = \omega + \alpha_1 y_{t-1} + \delta f_{m,t-1} + \gamma_1 \lambda_{t-1}, \tag{3}$$

where δ signifies the momentum effect. Properties of the model in Equation (3) can be obtained from those of higher-order autoregressive intensity models because the momentum model is simply a constrained model with higher-order coefficients being equal. Such an idea has been used in the literature, e.g. Duan (2013) and Hosseini, Takemura and Hosseini (2014) for the conventional time series analysis. In applications, the choice of m can be selected by information criteria or fixed *a priori*.

Using m = 15, we obtain a refined model for the y_t series of GLT trading:

$$\lambda_t = 0.832(0.14) + 0.197(0.01)y_{t-1} + 0.13(0.04)f_{m,t-1} + 0.519(0.07)\lambda_{t-1},$$

and $\hat{\theta} = 4.93(0.20)$. The Ljung-Box statistics give $Q(20)$ = 21.60 with p-value 0.36. It is interesting to see that the momentum coefficient is statistically significant and the standardized residuals have no serial correlations.

3. The Proposed Model

Turn to multivariate case. Let $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{kt})'$ be a k-dimensional vector of count data. In most applications, the components of $\boldsymbol{y}_t$ can be classified into subgroups, depending on their characteristics. Let $\boldsymbol{y}_{g,t}$ be the gth subgroup and k_g be the dimension of $\boldsymbol{y}_{g,t}$. Then, $\sum_{g=1}^{G} k_g = k$ and $k_g > 0$ and we write the ith element of $\boldsymbol{y}_{g,t}$ as $y_{g,it}$. In addition, let $\boldsymbol{f}_t = (f_{1t}, \ldots, f_{rt})'$ be a r-dimensional vector of common factors. Here f_{it} may be latent or observable. An example of observable common factor is the momentum factor of the $y_{g,it}$ component as defined in Section 2. A dynamic multivariate model for $\boldsymbol{y}_t$ is then

$$E(\boldsymbol{y}_t|F_{t-1}, \boldsymbol{x}_t, \boldsymbol{f}_t) = \exp(\boldsymbol{\beta x}_t + \boldsymbol{\delta f}_t) \odot \boldsymbol{\lambda}_t \tag{4}$$

$$\lambda_{g,it} = \omega_{g,i} + \alpha_{g,i} y_{g,i,t-1} + \gamma_{g,i} \lambda_{g,i,t-1}, \tag{5}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are matrices of parameters with dimensions $k \times d$ and $k \times r$, respectively, with d being the dimension of $\boldsymbol{x}_t$, $\odot$ denotes component multiplication. Furthermore, we assume that conditioned on F_{t-1}, $\boldsymbol{x}_t$, and $\boldsymbol{f}_t$, elements of $\boldsymbol{y}_{g,t}$ are independent and follow the same marginal distribution. The marginal distributions may differ for different groups, but they are conditionally independent between groups. The model in Equations (4) and (5) is relatively hard to estimate when $\boldsymbol{f}_t$ contains latent common factors. In such situations, one needs to apply either efficient important sampling methods or Markov chain Monte Carlo methods in estimation because the likelihood function of the data involves high-dimensional integration over the latent variables.

To simply the estimation, we propose a grouped autoregressive intensity with momentum (GAIM) model for $\boldsymbol{y}_t$. The GAIM model is a special case of the dynamic model in Equations (4) and (5) with $\boldsymbol{f}_t$ consisting of the momentum variables of each component of $\boldsymbol{y}_t$. More specifically, for a given $m > 1$, we define

$$\boldsymbol{f}_t = (f_{m,1,t-1}, \ldots, f_{m,k,t-1})', \quad f_{m,i,t-1} = (y_{i,t-1} + \cdots + y_{i,t-m})/m.$$

Under the GAIM model, the cross-dependence among elements of $\boldsymbol{y}_t$ is governed by the explanatory variables in $\boldsymbol{x}_t$ and the momentum variables in $\boldsymbol{f}_t$. On the other hand, the individual intensity $\lambda_{g,it}$ of Equation (5) does not depend on other components of $\boldsymbol{y}_t$. Since $\boldsymbol{f}_t$ consists of momentum factors, the GAIM model is a special case of multivariate autoregressive intensity model as such properties of the GAIM can be deduced from those of the latter.

Table 1: Descriptive Statistics of the Numbers of Transactions within 5-Minute Intervals for Five Stocks on New York Stock Exchange During the First Quarter of 2005.

Statistics	GLT	WPP	EDE	NU	WR
Mean	5.80	7.90	3.47	10.41	9.61
Var	16.76	34.18	9.66	34.42	34.90
Median	5	7	3	9	9
Min	0	0	0	0	0
Max	54	43	25	48	59
Skew	1.70	1.23	1.70	1.29	1.36
Kurt	7.46	1.98	4.84	2.80	4.16
Q(20)	2549	5026	1909	3927	5942

4. Application

In this section, we demonstrate the proposed GAIM model via analyzing the 5-dimensional count data of Lung, Liesenfeld, and Richard (2011). As mentioned in Section 2, the sample period of the data is from January 3 to March 31, 2005 for 61 trading days. Since each trading day consists of 75 observations, the sample size is 4575. The five stocks used are

1. GLT: Glatfelter Company
2. WPP: Wausau Paper Corporation
3. EDE: Empire District Electric Company
4. NU: Northeast Utilities
5. WR: Westar Energy, Inc.

Figure 1 shows the time plots of the first three components of $\boldsymbol{y}_t$ whereas selected summary statistics of $\boldsymbol{y}_t$ are given in Table 1. Clearly, the Poisson distribution is not adequate for the data due to over-dispersion. It turns out that negative Binomial distribution fares better so that we only provide results for GAIM models with negative Binomial distributions.

We start with a 2-dimensional case in which $\boldsymbol{y}_t = (\text{NU}_t, \text{WR}_t)'$. First, we assume that the two stocks have the same conditional marginal distribution, i.e. the number of group is $G = 1$. With m = 15, the fitted GAIM model shows that the fitted loading matrix of the common factors is

$$\boldsymbol{\delta} = \begin{bmatrix} 0.0051(0.0063) & 0.0005(0.0021) \\ -0.0060(0.0023) & 0.0065(0.0076) \end{bmatrix}, \quad \theta = 7.71(0.21),$$

where standard errors are in parentheses, and the individual equations for λ_{it} are

$$\begin{aligned} \lambda_{1t} &= 1.20 + 0.19y_{1,t-1} + 0.66\lambda_{1,t-1} \\ \lambda_{2t} &= 0.70 + 0.21y_{2,t-1} + 0.71\lambda_{2,t-1}, \end{aligned}$$

where all estimates are highly significant. The estimates of $\boldsymbol{\beta}$ are not shown for simplicity. To check the model, we consider the individual standardized residuals. The Ljung-Box statistics give $Q(50)$ = 61.13 (0.13) and $Q(50)$ = 55.48(0.28), respectively, for the two standardized residual series, where the number in parentheses denotes p-value. It is interesting to see that WR_t depends significantly on the momentum of NU_t.

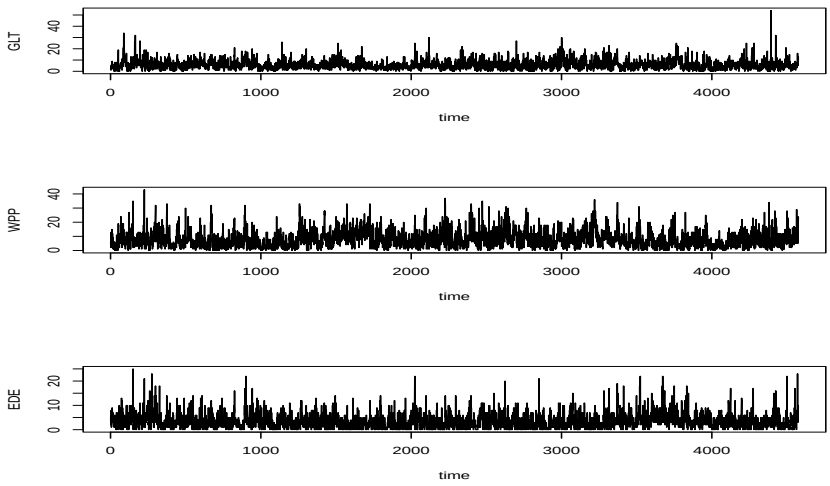

Figure 1: Number of transactions in 5-minute intervals for stocks GLT, WPP, and EDE. The sample period is from January 03 to March 31, 2005 and the intraday time span is from 9:45AM to 16:00 PM, Eastern Time.

Next, we assume that the two stocks have their own conditional marginal distributions. Again, with $m = 15$, the fitted GAIM model shows that the loading matrix of the common factors is

$$\boldsymbol{\delta} = \begin{bmatrix} 0.0048(0.0060) & 0.0005(0.0021) \\ -0.0061(0.0025) & 0.0052(0.0076) \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} 8.83(0.35) \\ 6.70(0.25) \end{bmatrix}$$

and the individual equations for λ_{it} are

$$\begin{aligned} \lambda_{1t} &= 1.18 + 0.20 y_{1,t-1} + 0.66 \lambda_{1,t-1} \\ \lambda_{2t} &= 0.66 + 0.21 y_{2,t-1} + 0.72 \lambda_{2,t-1}. \end{aligned}$$

The Ljung-Box statistics for the individual standardized residuals show $Q(50)$ = 61.72(012) and $Q(50)$ = 56.48(0.25), respectively. Again, the model fits the data well and the two series shows some minor cross-dependence. In this particular instance, the fitted values of $\hat{\theta}$ for the negative Binomial distributions are close so that there is no significant difference between use of one or two subgroups.

Finally, we entertain simultaneously all five series but divide the components into two groups with Group 1 consisting of GLT, WPP, and EDE. The fitted GAIM model provides a loading matrix as

$$\boldsymbol{\delta} = \begin{bmatrix} .044(.008) & .003(.003) & .008(.007) & -.004(.004) & .002(.003) \\ .004(.005) & .026(.006) & -.013(.007) & -.014(.003) & .003(.003) \\ .001(.006) & -.010(.004) & .049(.017) & -.003(.004) & .008(.003) \\ -.006(.004) & -.005(.002) & .014(.005) & .017(.004) & .003(.002) \\ -.011(.004) & -.001(.002) & .012(.006) & -.006(.003) & .031(.006) \end{bmatrix},$$

where the number in parentheses denotes standard error. The parameters for the marginal negative Binomial distributions are

$$\boldsymbol{\theta} = \begin{bmatrix} 4.62(0.12) \\ 8.66(0.27) \end{bmatrix}.$$

Table 2: Parameter Estimates for Individual Intensity Equations and Ljung-Box Statistics of Standardized Residuals for the 5-Dimensional GAIM model.

Par.	GLT	WPP	EDE	NU	WR
ω	1.22	1.38	0.61	1.80	1.60
α_1	0.13	0.21	0.16	0.16	0.17
γ_1	0.54	0.57	0.60	0.58	0.57
Q(50)	47.0(.59)	67.5(.05)	44.2(.71)	67.1(.05)	44.9(.68)

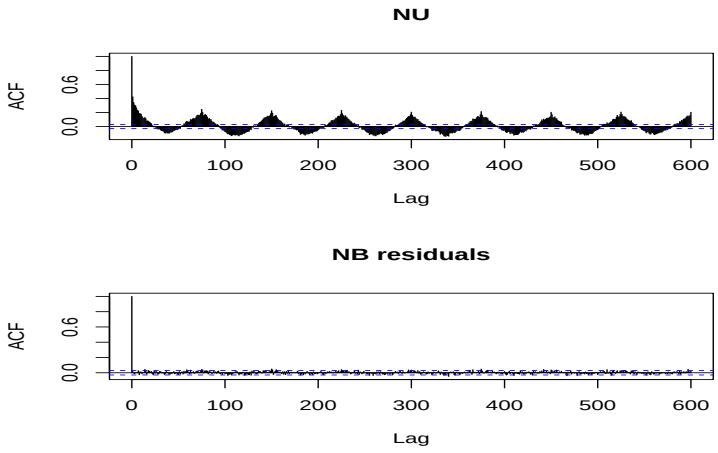

Figure 2: ACF of the observed and standardized residuals for NU stock. The residuals are from a 5-dimensional GAIM model.

Parameters for the individual intensity equation and the Ljung-Box statistics for the standardized residuals are in Table 2.

Figures 2 and 3 show that sample autocorrelation functions of the observed series and the standardized residuals for NU and WR stocks, respectively. The residuals are from the 5-dimensional GAIM model. Similar plots hold for the other three components of $\boldsymbol{y}_t$. Clearly, the fitted GAIM model is capable of handling the dynamic dependence between the 5-dimensional series.

Finally, the GAIM model is estimated by R code available from the author's web: faculty.chicagobooth.edu/ruey.tsay/research/.

REFERENCES

Cameron, A. C. and Trivedi, P. K. (2013). *Regression Analysis of Count Data*, 2nd Edition, New York: Cambridge University Press.

Davis, R.A., Dunsmuir, W.T.M., and Streett, S. (2003), "Observation Driven Models for Poisson Counts," *Biometrika*, 90, 770–790.

Davis, R.A. and Wu, R. (2009), "A Negative Binomial Model for Time Series of Counts," *Biometrika*, 96, 1–15.

Duan, J. C. (2013), "Time Series Models with Momentum," Working paper, Risk Management Institute, National University of Singapore.

Efron, B. (1986), "Double Exponential Families and Their Use in Generalized Linear Regression, *Journal of the American Statistical Association*, 81, 709–721.

Engle, R. F. and Russell, J. R. (1998), "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data," *Econometrica*, 66, 1127–1162.

Hautsch, N. (2012), *Econometrics of Financial High-Frequency Data*, Berlin: Springer-Verlag.

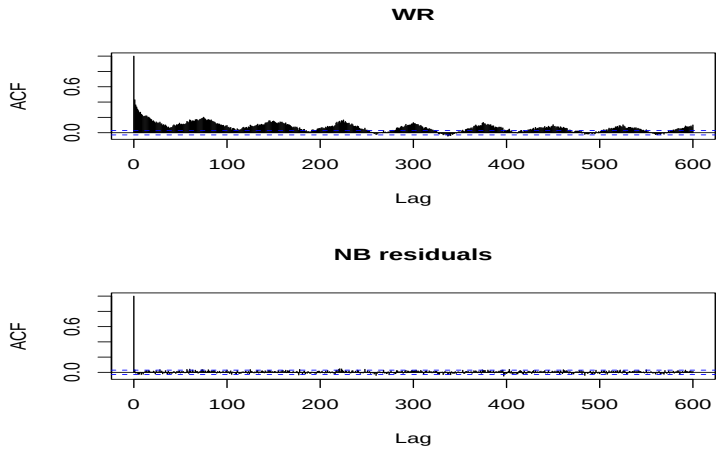

Figure 3: ACF of the observed and standardized residuals for WR stock. The residuals are from a 5-dimensional GAIM model.

Heinen, A. (2003), "Modelling Time Series Count Aata: An Autoregressive Conditional Poisson Model,"CORE Discussion Paper 2003062, Universit catholique de Louvain.

Hosseini, R., Takemura, A. and Hosseini, A. (2014), "Nonlinear Time-Varying Stochastic Models for Agroclimate Risk Assessment," *Environmental and Ecological Statistics*, 21, 1–20.

Jung, R. C., Kukuk, M., and Liesenfeld, R. (2006), "Time Series of Count Data: Modeling, Estimation and Diagnostics. *Computational Statistics and Data Analysis*, 51, 2350–2364.

Jung, R. C. Liesenfeld, R. and Richard, J. F. (2011), "Dynamic Factor Models for Multivariate Count Data: An Application to Stock-Market Trading Activity," *Journal of Business & Economic Statistics*, 29, 73–85.

Kedem, B. and Fokianos, K. (2002), *Regression Models for Time Series Analysis*, New Jersey: Wiley.

Winkelmann, R. (2010), *Econometric Analysis of Count Data*, 5th Edition. Berlin: Springer-Verlag.

Zeger, S. L. (1988), "A Regression Model for Time Series of Counts," *Biometrika*, 75, 621-629.