The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.1.115
28(1), 115–122

Integer-Valued GARCH Models for Count Time Series: Case Study

J.E. Yoon[a] · S.Y. Hwang[a,1]

[a]Department of Statistics, Sookmyung Women's University

(Received January 20, 2015; Revised February 2, 2015; Accepted February 3, 2015)

Abstract

This article is concerned with count time series taking values in non-negative integers. Along with the first order mean of the count time series, conditional variance (volatility) has recently been paid attention to and therefore various integer-valued GARCH(generalized autoregressive conditional heteroscedasticity) models have been suggested in the last decade. We introduce diverse integer-valued GARCH(INGARCH, for short) processes to count time series and a real data application is illustrated as a case study. In addition, zero inflated INGARCH models are discussed to accommodate zero-inflated count time series.

Keywords: Count time series, integer-valued GARCH(INGARCH), over-dispersion, zero-inflated INGARCH.

1. 서론

최근 들어 정수 값을 갖는(integer valued) 계수 시계열(count time series) 분석이 활발히 이루어지고 있다. 계수 시계열의 일차 적률인 조건부 평균(conditional mean)을 분석하기 위한 표준적인 모형은 'binomial thinning operator'를 이용한 INAR(Integer-valued AR)모형 (Hwang과 Basawa, 2011)이며 이외에도 조건부 포아송, 조건부 이항모형, 조건부 자기로지스틱 모형 등이 있다. 이에 대한 자세한 내용은 Fokianos (2011)를 참고하기 바란다.

본 연구에서는 이차 적률인 조건부 분산, 즉, 계수 시계열의 변동성(volatility)을 연구하고자 한다. 대표적인 시계열 분석 모형인 ARMA 모형이나 GARCH 모형은 시간의 흐름에 따라 관측된 실수 값(real valued) 시계열을 다루는 모형으로 정수 값만을 갖는 계수 시계열 자료에 적용시키는 경우 근사적인 모형으로서는 모르겠으나 정확한 분석결과를 기대하는 것은 무리가 있다. 따라서 시간의 흐름에 따라 조사된 특정 질병의 감염자 수, 특정 범죄의 횟수, 가격 변동이 일어난 횟수 등 음이 아닌 정수 값을 갖는 계수 시계열(count time series)의 변동성을 위한 적절한 모형의 필요성이 제기되어 왔다. 실제 음이 아닌 정수 값을 취하는 계수 시계열 자료에서는 시간에 따라 조건부 분산(변동성)이 변화하는 경우가 많으므로 (조건부) 등분산을 가정하고 다루는 것은 적절하지 못한 경우가 많다. 이러한 계수 시계열 자료에 대한 조건부 이분산(conditionally heteroscedastic) 모형의 필요성을 반영하여 Ferland 등 (2006)는 계수 시계열의 조건부 분포로 포아송(Poisson) 분포를 이용한 Integer-valued GARCH 모형,

[1]Corresponding author: Department of Statistics, Sookmyung Women's University, Yongsan-Gu, Seoul 140-742, Korea. E-mail: shwang@sookmyung.ac.kr

즉, INGARCH 모형을 제안하였다. 이 모형은 대표적인 이분산 시계열 모형인 GARCH 모형과 유사한 수학적 성질을 갖는 모형으로서 조건부 평균 모형인 INAR을 조건부 이분산 모형으로 확장시킨 모형이라 할 수 있다.

계수 시계열 자료들은 음이 아닌 정수값을 가지며, 관측된 값들 간의 양의 상관관계가 존재하는 경우가 많기에 과산포(over-dispersion, Zhu (2012))의 경향을 보이는 경우가 많다. 계수시계열의 변동성 분석 분야에서 이러한 과산포 문제를 다루기 위하여 Zhu (2011)는 조건부 분포로 포아송 분포 대신 음이항 분포를 적용시킨 Negative Binomial INGARCH, 즉, NBINGARCH 모형을 제안하였다. 본 논문에서는 다양한 계수형 GARCH 모형들을 소개하고 이들을 국내 풍진 발생건수 (계수 시계열) 자료에 적용하고 있으며, 효과적인 계수 시계열 자료 분석을 위한 향후 연구 과제를 제안해 보고자 한다.

2. 계수형 GARCH 모형: INGARCH(p,q) 모형과 NBINGARCH(p,q) 모형

시계열의 변동성 분석을 위한 모형인 표준적인 GARCH 모형은 다음과 같이 정의된다.

$$X_t|F_{t-1} : N\left(0, \sigma_t^2\right),$$
$$\sigma_t^2 = \alpha_0 \ + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2,$$

여기서 자료는 $\{X_t\}$이고 F_{t-1}은 $t-1$ 시점까지의 정보 집합이다. 위와 같은 GARCH 모형과 유사한 수학적 성질을 가지며, 정수 값을 갖는 계수 시계열 자료를 다루기 위한 모형으로 Ferland 등 (2006)가 제안한 integer-valued GARCH, 즉, INGARCH 모형은 다음과 같다.

$$X_t|F_{t-1} : \text{Poisson}(\lambda_t),$$
$$\lambda_t = \alpha_0 \ + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j},$$

여기서 $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, $i = 1, \ldots, p$, $j = 1, \ldots, q$, $p \geq 1$, $q \geq 0$이다. 위 모형에서 $q = 0$인 경우 INARCH(p) 모형이 된다. 이 모형은 조건부 분포로 정규분포를 가정하던 GARCH 모형과 달리 X_t의 조건부 분포로 포아송 분포를 가정하고 있다. INGARCH(p,q) 모형에서는 조건부 평균과 조건부 분산이 관측된 시계열에서의 과거 값과 자신의 과거 값에 의존한다. 이 모형에서는 평균과 분산이 같다는 특징을 가진 포아송분포를 조건부 분포로 가정하여 실제 자료 분석에서 흔히 발생하는 과산포(over-dispersion, Zhu (2012))문제를 설명하지 못한다는 단점이 있다. 이러한 단점을 보완하기 위하여 Zhu (2011)는 다음과 같은 음이항 INGARCH, 즉, NBINGARCH(p,q)를 제안하였다.

$$X_t|F_{t-1} : NB(r, p_t),$$
$$\lambda_t = \frac{1-p_t}{p_t} = \alpha_0 \ + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j},$$

여기서 NB는 음이항 분포(negative binomial distribution)를 나타내며 r은 양의 정수이고 $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, $i = 1, \ldots, p$, $j = 1, \ldots, q$, $p \geq 1$, $q \geq 0$이다. NBINGARCH(p,q) 모형은 INGARCH(p,q) 모형과 다르게 X_t의 조건부 분포로 포아송분포 대신 음이항분포를 가정하였다. 이 모형은 계수 시계열 자료에서 흔히 발생하는 과산포 문제와 극단적인 관측값을 INGARCH(p,q) 모형보다 더 효과적으로 다룰 수 있다고 알려져 있다 (Zhu, 2011, 2012). 위 모형에서 $q = 0$인 경우에는 NBINARCH(p) 모형이 된다. 이제 차수가 $(1,1)$인 모형을 알아보도록 한다.

2.1. INGARCH$(1,1)$

INGARCH(p,q)의 일차모형인 INGARCH$(1,1)$ 모형은 다음과 같다.

$$X_t|F_{t-1} : \text{Poisson}(\lambda_t),$$
$$\lambda_t = \alpha_0 \ + \alpha_1 X_{t-1} + \beta_1 \lambda_{t-1},$$

여기서 $\alpha_0 > 0$이고 정상성(stationarity)을 만족하기 위해 $\alpha_1 + \beta_1$은 1보다 작다. t시점에서의 (조건부)평균은 한 시점 전의 관측 값과 (조건부)평균에 의존한다. 특수한 경우로서 $q = 0$인 경우인 INARCH(1) 모형은 다음과 같다.

$$X_t|F_{t-1} : \text{Poisson}(\lambda_t),$$
$$\lambda_t = \alpha_0 \ + \alpha_1 X_{t-1},$$

여기서 α_1은 1보다 작다.

2.2. NBINGARCH$(1,1)$

NBINGARCH(p,q)의 일차모형인 NBINGARCH$(1,1)$ 모형은 다음과 같다.

$$X_t|F_{t-1} : NB(r, p_t),$$
$$\lambda_t = \frac{1-p_t}{p_t} = \alpha_0 \ + \alpha_1 X_{t-1} + \beta_1 \lambda_{t-1},$$

여기서 $\alpha_0 > 0$이고 α_1, β_1은 1보다 작으며 r은 양의 정수이다. 이 모형의 정상성 조건은 $(r\alpha_1 + \beta_1)^2 + r\alpha_1^2 < 1$이다 (Zhu, 2011). 조건부 평균과 조건부 분산은 음이항 분포의 특징으로부터 다음과 같이 유도된다.

$$E(X_t|F_{t-1}) = r\lambda_t,$$
$$\text{Var}(X_t|F_{t-1}) = r\lambda_t(1+\lambda_t).$$

위의 식을 보면 조건부 분산이 조건부 평균보다 더 큰 값을 가짐을 알 수 있다. 이러한 성질을 이용하여 포아송 분포에서는 설명하기 어려운 과산포 문제를 적절히 설명할 수 있다. 하지만 INGARCH 모형과 비교해서 추가적으로 모수 r을 고려해야 한다는 단점이 있다. 특별히 $q = 0$인 경우인 NBINARCH(1) 모형은 다음과 같다.

$$X_t|F_{t-1} : NB(r, p_t),$$
$$\lambda_t = \frac{1-p_t}{p_t} = \alpha_0 \ + \alpha_1 X_{t-1},$$

여기서 α_1은 1보다 작다. 이제 여러 가지 계수형 변동성 모형을 실제 자료에 적용시켜 보도록 한다.

3. 사례분석

본 절에서는 2001년 1월부터 2012년 12월까지 보건복지부에 보고된 법정감염병발생보고 자료 중 풍진 자료에 다음의 4가지 계수형 변동성 모형인 INARCH(1), INGARCH$(1,1)$, NBINARCH(1)과 NBINGARCH$(1,1)$을 적합해 보았다.

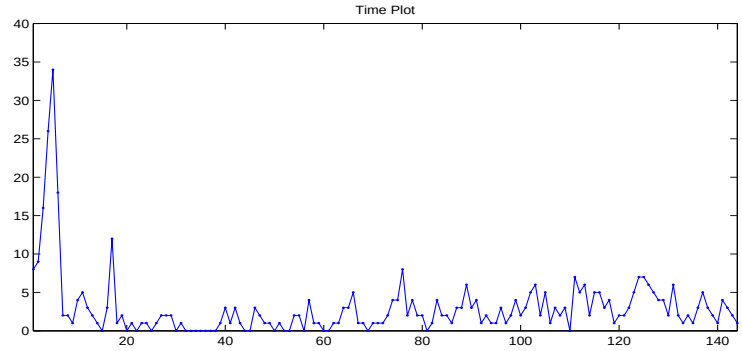

Figure 3.1. Count time series plot

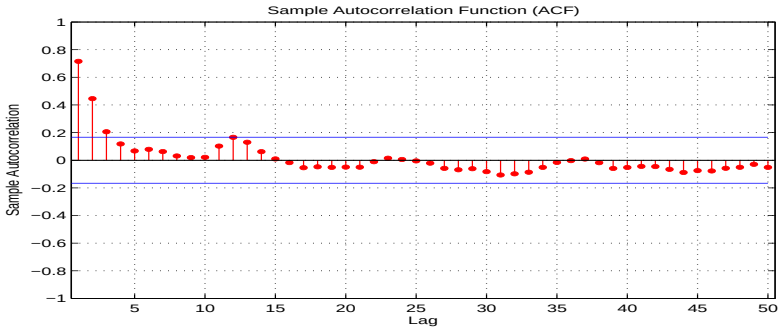

Figure 3.2. Autocorrelation function(ACF)

이 자료는 매월 보고된 발생환자 수에 대한 기록으로 12년간 총 144개의 관측 값들로 구성된 계수 시계열이다. Figure 3.1은 144개 자료에 대한 시도표이다. 관측 초기에 급격하게 발생 환자 수가 증가한 후 감소하였고 일정 기간 환자가 발생하지 않은 경우도 있다. 증가와 감소를 반복하고 있으며 변동 폭이 일정하지 않은 것으로 보인다. Figure 3.2와 Figure 3.3은 각각 자기상관함수(ACF)와 부분자기상관함수(PACF)에 대한 그래프이다. 두 그림을 보면 관측값들 간에 자기상관관계가 존재함을 볼 수 있다. 자료의 평균과 분산은 각각 3과 17.9858로 과산포(over-dispersion) 문제가 의심된다. 여기서 과산포란 분산 17.9858가 평균 3보다 큼을 의미한다. INARCH(1) 모형과 INGARCH$(1,1)$ 모형을 적합 시킨 결과는 Table 3.1과 같다. 괄호 안의 값은 추정량에 대한 표준오차(standard error)이다. 추정방법은 최우추정법(ML)을 사용하였으며 음이항 분포인 경우 닫힌 형태의 추정량(closed form)을 가지고 있지 않으므로 수치적으로 최적화시키는 방법(numerical optimizaion method)을 이용하였으며 Matlab의 fmincon 함수를 사용하였다.

Table 3.1로부터 다음 식을 얻을 수 있다.

$$\begin{aligned} \text{INARCH}(1);\ & X_t|F_{t-1} : \text{Poisson}(\lambda_t), \\ & \lambda_t = 1.0201 + 0.6436X_{t-1}, \\ \text{INGARCH}(1,1);\ & X_t|F_{t-1} : \text{Poisson}(\lambda_t), \\ & \lambda_t = 0.8391 + 0.6138X_{t-1} + 0.0912\lambda_{t-1}. \end{aligned}$$

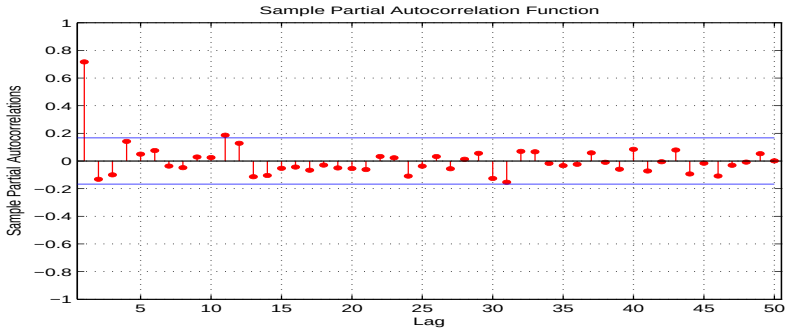

Figure 3.3. Partial autocorrelation function(PACF)

Table 3.1. Parameter estimates for INARCH(1) and INGARCH(1, 1)

	α_0	α_1	β_1	AIC	BIC
INARCH(1)	1.0201	0.6436		609.1711	615.0827
	(0.0489)	(0.0138)			
INGARCH(1, 1)	0.8391	0.6138	0.0912	609.5972	618.4435
	(0.0839)	(0.0133)	(0.0105)		

Table 3.2. AIC and BIC

	initial values of r	1	2	3	4	5
NBINARCH(1)	AIC	596.4749	572.6968	567.9695	567.668	568.7603
	BIC	605.3212	581.5431	576.8158	576.5143	577.6066
NBINGARCH(1, 1)	AIC	598.2789	572.7161	567.0869	566.1943	566.8483
	BIC	610.0455	584.4826	578.8535	577.9609	578.6149

Table 3.3. Parameter estimates for NBINARCH(1) and NBINGARCH(1, 1)

	α_0	α_1	β_1	AIC	BIC
NBINARCH(1)	0.272	0.1499		567.668	576.5143
$\hat{r} = 4$	(0.0029)	(0.0006)			
NBINGARCH(1, 1)	0.1697	0.1311	0.2195	566.1943	577.9609
$\hat{r} = 4$	(0.0059)	(0.0009)	(0.0314)		

NBINARCH(1) 모형과 NBINGARCH(1, 1) 모형을 적합 시킨 결과는 Table 3.3과 같다. 모형을 적합하기에 앞서 새로운 모수인 r에 대한 추정이 필요하다. 1부터 5까지의 초기 r값을 주고 모형을 적합 시킨 결과 가장 작은 AIC와 BIC값을 갖게 하는 r값으로 추정하였다. 그 결과 $\hat{r} = 4$로 추정되었으며 모형을 적합 시킨 결과는 이 값에 대한 결과만 제시하였다. 추정식은 다음과 같다.

$$\begin{aligned}
&\text{NBINARCH}(1);\ X_t|F_{t-1} : NB(r, p_t),\\
&\qquad\qquad \lambda_t = \frac{1-p_t}{p_t} = 0.272 + 0.1499X_{t-1},\\
&\text{NBINGARCH}(1,1);\ X_t|F_{t-1} : NB(r, p_t),\\
&\qquad\qquad \lambda_t = \frac{1-p_t}{p_t} = 0.1697 + 0.1311X_{t-1} + 0.2195\lambda_{t-1}.
\end{aligned}$$

두 모형을 적합 시킨 결과를 비교해보면 AIC와 BIC값을 비교해보았을 때 INGARCH(1, 1)보다는 NBINGARCH(1, 1) 모형이 더 작은 값을 갖는 것을 알 수 있다. 따라서 이 자료에 대해서는 NBINGARCH(1, 1) 모형이 자료를 더 잘 설명한다고 볼 수 있다. 이는 자료에서 보이고 있는 과산포 문제와 극단적인 관측값의 존재로 인한 것으로 생각된다. 자료가 보이고 있는 이러한 특징들로 인해 평균과 분산이 같은 값을 갖는다고 가정하는 포아송 분포 즉, INGARCH(1, 1)보다는 과산포 현상을 상대적으로 잘 설명하는 음이항 분포를 조건부 분포로 설정하는 NBINGARCH(1, 1) 모형이 더 효과적이라고 할 수 있다. 적합값 $\hat{X}_t = E(X_t|F_{t-1})$은 조건부분포로부터 쉽게 추정치를 얻을 수 있으며 분석에서는 생략했지만 잔차 $\{X_t - \hat{X}_t\}$가 백색잡음임을 확인하는 잔차분석도 모형 선정에 도움이 될 것이다.

4. 향후 연구 과제

본 절에서는 실증자료 분석에서 과산포 문제와 더불어 발생하는 많은 0이 관측되는 문제를 다루기 위한 모형을 소개하고 향후 연구 과제를 제안하고자 한다.

정수 값을 갖는 계수 시계열 자료가 특정한 범죄의 발생, 특정한 질병의 발병 등 일반적으로 흔히 일어나지 않는 사건들에 대해 관심을 가지고 조사하는 경우에는 많은 수의 0이 포함된 자료가 얻어지는 경우가 발생한다. 이러한 자료를 영과잉 자료(zero-inflated count data)로 볼 수 있다. 영과잉이란 포아송 모형을 가정하는 경우 예상되는 0값보다 더 많은 0값이 관측되는 현상을 의미한다 (Zhu, 2012). 계수 시계열 자료에서 영과잉 자료인지 파악하기 위해 Puig와 Valero (2006)가 제안한 다음과 같은 zero-inflation index(ZI index)를 이용할 수 있다.

$$\text{ZI index} = 1 + \frac{\log(p_0)}{\mu},$$

여기서 p_0는 0값의 비율(proportion)이고 μ는 평균이다. ZI index로부터 관측값에 0이 많은 영과잉 자료로 판단되는 경우, 과산포 문제는 더욱 심화될 것이며 앞에서 살펴본 INGARCH(p, q) 또는 NBINGARCH(p, q) 모형을 그대로 적합 시키는 경우 모형의 적합도가 떨어질 것으로 예상된다. 계수 시계열의 변동성 분석에서 영과잉 문제를 해결하는 방안으로 Zhu (2012)는 자료에 포함된 0에 대해 적절한 가중치를 주어 처리하는 영과잉 변동성 모형을 제안하였다.

기존의 INGARCH(p, q) 모형에서 조건부 분포로 사용한 포아송분포를 영과잉 포아송분포(Zero-inflated Poisson; ZIP)로 대체하여 다음과 같은 ZIP-INGARCH(p, q) 모형을 제안하였다.

$$\begin{aligned} X_t|F_{t-1} &: \text{ZIP}(\lambda_t, w), \\ \lambda_t &= \alpha_0 \ + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j}, \end{aligned}$$

여기서 $0 < w < 1$, $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, $i = 1, \ldots, p$, $j = 1, \ldots, q$, $p \geq 1$, $q \geq 0$이며, 조건부분포인 ZIP(λ, w)는 다음과 같다.

$$P(X = k) = w\delta_{k,0} + (1 - w)\frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots,$$

여기서 $0 < w < 1$이며 $\delta_{k,0}$는 $k = 0$인 경우에는 1이 되고 $k \neq 0$인 경우에는 0이 된다.

이와 같은 ZIP-INGARCH 모형에서 조건부 평균과 조건부 분산은 각각 다음과 같다.

$$\begin{aligned} E(X_t|F_{t-1}) &= (1 - w)\lambda_t, \\ \text{Var}(X_t|F_{t-1}) &= (1 - w)\lambda_t(1 + w\lambda_t), \end{aligned}$$

여기서 $\mathrm{Var}(X_t|F_{t-1}) > E(X_t|F_{t-1})$임을 알 수 있으며, 이를 통해 INGARCH(p,q) 모형에서 해결하지 못하였던 과산포 문제를 다룰 수 있게 되었음을 알 수 있다 (Zhu, 2012). 특별한 경우로서 ZIP-INGARCH 모형에서 $w = 0$인 경우에는 INGARCH 모형이 된다. 같은 방식으로 NBINGARCH 모형에서 조건부 분포로 이용하였던 음이항 분포에서 영과잉 음이항분포를 도입하여 이를 Zero-inflated Negative Binomial; ZINB로 명명하고 조건부 분포로 ZINB을 이용한 ZINB-INGARCH(p,q) 모형은 다음과 같다.

$$X_t|F_{t-1} : \mathrm{ZIPNB}(\lambda_t, a, w),$$
$$\lambda_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j}.$$

조건부 분포인 ZIPNB(λ, a, w)은 다음과 같다 (Zhu, 2012).

$$P(X=k) = w\delta_{k,0} + (1-w)\frac{\Gamma\left(k + \frac{\lambda^{1-c}}{a}\right)}{k!\Gamma\left(\frac{\lambda^{1-c}}{a}\right)}\left(\frac{1}{1+a\lambda^c}\right)^{\frac{\lambda^{1-c}}{a}}\left(\frac{a\lambda^c}{1+a\lambda^c}\right)^k, \quad k = 0,1,2,\ldots,$$

여기서 $\lambda > 0$, $0 < w < 1$, $a \geq 0$이며 a는 퍼진 정도를 나타내는 모수(scale parameter)이며 $\delta_{k,0}$은 $k = 0$인 경우에는 1이 되고 $k \neq 0$인 경우에는 0이 된다. ZINB-INGARCH 모형에서 조건부 평균과 조건부 분산은 각각 다음과 같다.

$$E(X_t|F_{t-1}) = (1-w)\lambda_t,$$
$$\mathrm{Var}(X_t|F_{t-1}) = (1-w)\lambda_t\left(1 + w\lambda_t + a\lambda_t^c\right).$$

이 식으로부터 $\mathrm{Var}(X_t|F_{t-1}) > E(X_t|F_{t-1})$이 성립하므로 ZINB-INGARCH$(p,q)$ 모형은 과산포 성질이 있음을 알 수 있다.

위에서 살펴본 ZIP-INGARCH(p,q) 모형과 ZINB-INGARCH(p,q) 모형은 관측값 0을 다룰 때 이에 대해 고려하는 w값을 0과 1사이의 특정 값으로 고정된 상수모수로 간주하고 있다. 시간대 별로 0의 비율이 달라지는 계수 시계열의 변동성 분석을 위해 w가 시간의 흐름에 따라 변화하는 확률변수로서의 w_t값을 고려하는 것에 대한 연구도 흥미로울 것으로 생각된다.

References

Ferland, R., Latour, A. and Oraichi, D. (2006). Integer-valued GARCH process, *Journal of Time Series Analysis*, **27**, 923–942.

Fokianos, K. (2011). Some recent progress in count time series, *Statistics*, **45**, 49–58.

Hwang, S. Y. and Basawa, I. V. (2011). Asymptotic optimal inference for multivariate branching-Markov processes via martingale estimating functions and mixed normality, *Journal of Multivariate Analysis*, **102**, 1018–1031.

Puig, P. and Valero, J. (2006). Count data distributions: Some characterizations with applications, *Journal of the American Statistical Society*, **101**, 332–330.

Zhu, F. (2011). A negative binomial integer-valued GARCH model, *Journal of Time Series Analysis*, **32**, 54–67.

Zhu, F. (2012). Zero-inflated Poisson and negative binomial integer-valued GARCH models, *Journal of Statistical Planning and Inference*, **142**, 826–839.

계수 시계열을 위한 정수값 GARCH 모델링: 사례분석

윤재은[a] · 황선영[a,1]

[a]숙명여자대학교 통계학과

(2015년 1월 20일 접수, 2015년 2월 2일 수정, 2015년 2월 3일 채택)

요 약

본 연구에서는 정수값을 갖는 계수 시계열의 조건부 이차적률인 변동성(volatility)을 다루고 있다. 여러 가지 정수값 GARCH, 즉, INGARCH 모형들을 소개하고 계수 시계열인 국내 풍진발생건수에 적용시켜 보았다. 과산포(over-dispersion)와 영과잉(zero-inflation)현상을 계수 시계열의 변동성 분석 입장에서 살펴보았고 향후 분석 모형으로서 영과잉(zero-inflation) INGARCH 모형인 ZI-INGARCH 모형을 살펴보았다.

주요용어: 계수 시계열, 정수값 GARCH (INGARCH), 과산포, 영과잉 GARCH.

[1]교신저자: (140-742) 서울특별시 용산구 청파로47길 100, 숙명여자대학교 통계학과.
E-mail: shwang@sookmyung.ac.kr