응용통계연구 (2008)
21(6), 1037-1044

멱변환 이분산성 시계열 모형을 이용한 인터넷 트래픽 예측 기법 연구

하명호[1] · 김삼용[2]

[1]중앙대학교 통계학과, [2]중앙대학교 통계학과

(2008년 7월 접수, 2008년 8월 채택)

요 약

본 연구에서는 재무시계열 자료의 변동성을 분석하는데 유용하게 쓰이는 멱변환 시계열 모형을 인터넷 트래픽 자료 특성 분석에 적용하여 효용성을 보이고자 한다. 트래픽의 특성인 장기기억(long memory)특성을 설명하기 위하여 멱변환 GARCH(PGARCH) 모형을 소개하고 기존의 GARCH 모형보다 더 유용함을 시뮬레이션과 실제 인터넷 트래픽 자료에 적합시켜 입증하였다.

주요용어: GARCH 모형, 멱변환 GARCH 모형, 인터넷 트래픽, 장기기억.

1. 서론

인터넷의 급속한 활용과 이에 상응하는 인터넷 트래픽의 기하급수적인 증가는 통신망의 운용을 목표로 하는 통신 사업자에게는 시급하게 해결 되야 하는 문제로 대두되고 있다. 이러한 요구에 상응하여 많은 연구자들이 인터넷 트래픽의 효율적인 분석과 예측을 위하여 다양한 통계적 모형을 도입하여 사용하고 있다. 먼저 정상성(stationarity)을 만족하는 인터넷 트래픽의 예측을 위하여 Basu 등 (1996)는 자기회귀(AR) 모형을 도입하였고 Kim (2007)은 이분산 자기회귀(ARCH) 모형을 이용하여 네트워크 트래픽을 예측하여 AR 모형보다 성능이 우수함을 보였다. 최근에 Shu 등 (2005)은 인터넷 트래픽의 장기기억 특성을 설명할 수 있는 Fractional ARMA(FARIMA) 모형보다 GARCH 모형이 더 예측력이 뛰어남을 보였다. 한편 인터넷 트래픽과 특성이 유사하다고 알려져 있는 재무 시계열 자료(financial time series data)의 장기억특성을 설명하기 위하여 Ding 등 (1993)은 새로운 모형을 제시하여 자료의 특성을 분석하고 예측하였다. 이 모형은 멱변환 ARCH 모형(Power ARCH model)이라 부른다.

본 연구에서는 기존의 GARCH 모형 하에서 인터넷 트래픽을 예측하는 것을 발전시켜 멱변환 GARCH 모형을 적용하여 그 효율성을 보이고자 한다.

2. GARCH 모형과 멱변환(PGARCH) 모형

2.1. GARCH(1, 1) 모형

다음과 같은 AR(1)-GARCH(1, 1) 모형을 고려하기로 한다.

이 연구는 2007년도 산학 협동재단 학술 연구비 지원에 위한 것임.
[1](156-756) 서울시 동작구 흑석동 221, 중앙대학교 통계학과, 석사과정. E-mail: ggosma@hanmail.net
[2]교신저자: (281-207) 서울시 동작구 흑석동 221, 중앙대학교 통계학과, 부교수. E-mail: sahm@cau.ac.kr

$$
\begin{aligned}
y_t &= \phi y_{t-1} + \epsilon_t, \\
\epsilon_t &= \sqrt{h_t} e_t, \\
h_t &= \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \beta_1 h_{t-1}, \\
e_t &\sim iid(0,\, \sigma^2).
\end{aligned} \tag{2.1}
$$

GARCH(1, 1) 모형은 재무 자료의 변동성을 기존의 선형모형 보다는 효율적으로 탐지 할 수 있으나 이상치의 존재 시 추정치에 약점이 있음을 알 수 있다 (Dijk 등, 1999).

2.2. 멱변환 이분산성(PGARCH) 모형

Ding 등 (1993)은 1928년 1월 3일부터 1991년 8월 30일까지 17055개의 자료를 분석하기 위해 새로운 모형을 제안하였는데 그 모형은 아래와 같다.

$$
\begin{aligned}
\epsilon_t &= \sigma_t e_t, \quad e_t \sim iid(0, 1), \\
\sigma_t^2 &= \alpha_0 + \alpha_1 \epsilon_{t-1}^d, \quad d > 0.
\end{aligned} \tag{2.2}
$$

Ding 등 (1993)은 이 모형을 이용하여 주가(stock price)의 장기 기억 특성을 규명하였고 이러한 특성은 d가 1에 가까울 때 가장 크게 나타난 것을 실증적으로 보여 주었다. 한편 Ding과 Granger (1996)은 Power GARCH(1, 1) 모형을 제안하였고 $d = 0.4$일 때 장기기억 특성이 가장 크다는 것을 보여 주었다. 네트워크 트래픽의 가장 큰 특성 중 하나는 장기기억이고 전통적으로 이러한 성질은 Hurst 모수(H)로 판별을 하였다. 이 모수는 다음의 식에서 구할 수 있다.

$$
\log Q = a + H \log k, \tag{2.3}
$$

여기서 $Q = R(t,\, k)/s(t,\, k)$이고

$$
R(t,\, k) = \max_{0 \le i \le k} \left[Y_{t+i} - Y_t - \frac{i}{k} \left(Y_{t+k} - Y_t \right) \right] - \min_{0 \le i \le k} \left[Y_{t+i} - Y_t - \frac{i}{k} \left(Y_{t+k} - Y_t \right) \right], \tag{2.4}
$$

$$
S(t,\, k) = \sqrt{\frac{1}{k} \sum_{i=k+1}^{t+k} \left(X_i - \overline{X_{t,k}} \right)^2}, \tag{2.5}
$$

$$
\overline{X_{t,k}} = \frac{1}{k} \sum_{i=k+1}^{t+k} X_i \tag{2.6}
$$

이다. 여기서 y_t는 t시점에서의 시계열 자료이다.

만일 $1/2 < H < 1$이라면 이 시계열 자료는 장기기억 특성을 가진다고 할 수 있다 (Beran, 1994). 한편 Zhou 등 (2005)는 장기기억 모수를 가지는 Fractional ARIMA 모형보다 ARIMA/GARCH 모형의 예측의 정확도가 높다는 것을 보였다. 여기서 다음의 모형을 고려 해 보자.

$$
\begin{aligned}
y_t &= \phi y_{t-1} + \sqrt{h_t} e_t, \quad e_t \sim iid(0, 1), \\
h_t^d &= \alpha_0 + \alpha_1 \left(\epsilon_{t-1}^d \right) + \beta_1 h_{t-1}^d.
\end{aligned} \tag{2.7}
$$

위에서 제시한 AR(1)-Power GARCH(1, 1)(PGARCH) 모형을 네트워크 트래픽 예측에 응용하기로 하며 이 모형은 다음과 같다.

표 **3.1.** 모수 추정 결과($\phi = 0.3$, $\alpha_0 = 0.1$, $\alpha_1 = 0.1$, β_1=0.7에서 생성)

모형		GARCH(1,1)				PGARCH				
d		$\widehat{\phi}$	$\widehat{\alpha_0}$	$\widehat{\alpha_1}$	$\widehat{\beta}$	$\widehat{\phi}$	$\widehat{\alpha_0}$	$\widehat{\alpha_1}$	$\widehat{\beta}$	$\widehat{d}$
0.3	$\bar{X}$	0.2935	0.1400	0.0502	0.2947	0.2957	0.2249	0.0609	0.6345	0.2257
	S	0.0320	0.1299	0.0390	0.6081	0.0349	0.0702	0.0238	0.1105	0.1013
0.5	$\bar{X}$	0.3096	0.1194	0.0892	0.4851	0.3088	0.1767	0.0961	0.6038	0.4800
	S	0.0321	0.1122	0.0421	0.3998	0.0330	0.0565	0.0327	0.0928	0.1704
0.7	$\bar{X}$	0.2903	0.0844	0.0917	0.6705	0.2906	0.1172	0.0970	0.6784	0.6976
	S	0.0266	0.0323	0.0299	0.1049	0.0268	0.0475	0.0289	0.0970	0.1514

표 **3.2.** SHAPIRO-WILK TEST

data	통계량	p-value
1초	0.99803	0.0285
10초	0.26961	<0.0100

$$
\begin{aligned}
&y_t = \phi y_{t-1} + \sqrt{h_t} e_t, \quad e_t \sim iid(0,1),\\
&h_t = \left\{ \alpha_0 + \alpha_1 \left(\epsilon_{t-1}^d \right) + \beta_1 h_{t-1}^d \right\}^{\frac{1}{d}}, \quad d > 0,\\
&\alpha_0 > 0, \quad \alpha_1 + \beta_1 < 1, \quad \alpha_1 > 0,
\end{aligned} \tag{2.8}
$$

여기서 $d = 1$이면 GARCH(1,1) 모형이 된다.

3. 시뮬레이션 및 실제자료 분석

시뮬레이션은 PGARCH 모형을 토대로 하여 $\phi = 0.3$, $\alpha_0 = 0.1$, $\alpha_1 = 0.1$, $\beta = 0.7$에서 d값을 0.3에서 0.7까지 0.2씩 증가시켜 1000개를 생성하였다. 이러한 과정을 30회 반복하여 각 모형에서의 모수의 추정치에 대한 평균($\bar{X}$)과 표본표준편차(S)를 구하였다.

표 3.1에서는 모수의 값($\phi = 0.3$, $\alpha_0 = 0.1$, $\alpha_1 = 0.1$, $\beta = 0.7$)을 주고 d를 변화시켜 각 모수의 추정치를 계산한 것이다. 먼저 GARCH(1,1) 모형에서는 d를 1로 고정되어 있고 이러한 연유로 모든 d 값에서 PGARCH 모형이 GARCH(1,1) 모형보다 실제 모수에 근접한 결과가 나타났으며 또한 각 모형에서 추정치의 오차의 크기를 보면 PGARCH 모형에서의 오차가 GARCH(1,1) 모형의 것보다 작다. 이것은 PGARCH 모형이 GARCH(1,1) 모형보다 훨씬 안정적이라는 것을 보여준다.

다음으로는 실제자료로서 이 자료는 2007년 4월 11일 23시 25분 05초부터 2007년 4월 11일 23시 55분 05초까지 30분간의 트래픽 자료를 1초, 10초 단위로 측정한 트래픽의 평균 패킷수이며 각각의 자료의 수는 1800개와 180개이다. 이러한 방법으로 측정한 원 자료를 다음과 같은 방법으로 변수 변환하여 실제 모수를 추정하였다.

$$
Z_t = \log\left(\frac{y_t}{y_{t-1}} \right) \times 100. \tag{3.1}
$$

그림 3.1과 3.2는 원자료의 시계열 그림이고, 그림 3.3과 3.4는 로그 변환 후 차분한 자료의 시계열 그림이다. 또한 그림 3.5와 3.6는 원자료의QQ-Plot이다. 표 3.2은 Shapiro-Wilk test의 통계량과 p-value이다. Shapiro-Wilk test 결과 본 논문에서 사용한 트래픽 자료는 정규 분포를 따른다고 할 수 없다. 이것은 멱변환을 통하여 자료 분석 및 모수추정 시 개선할 수 있는 여지를 보여준다 할 수 있다.

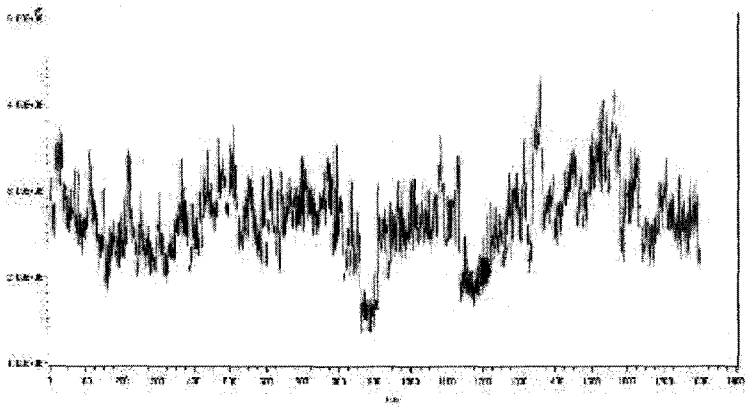

그림 3.1. 1초 자료의 원 시계열 그림

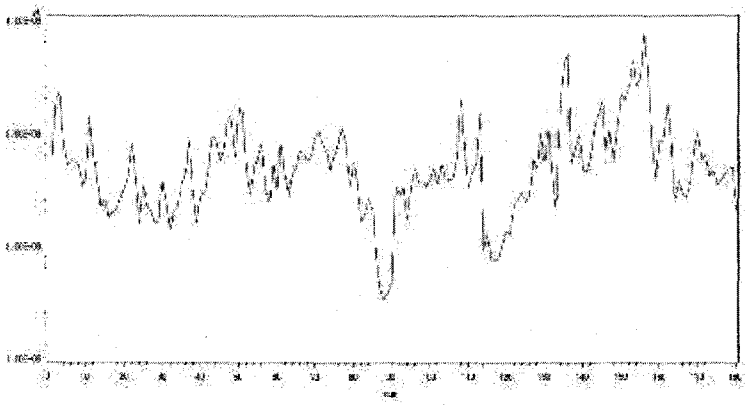

그림 3.2. 10초 자료의 원 시계열 그림

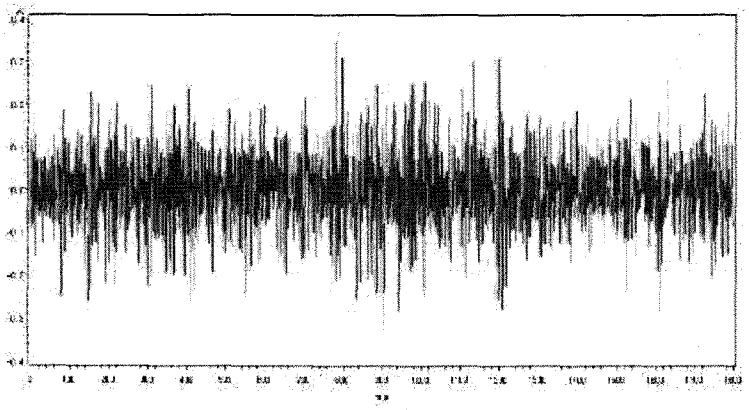

그림 3.3. 1초 자료의 로그 변환 후 차분한 시계열 그림

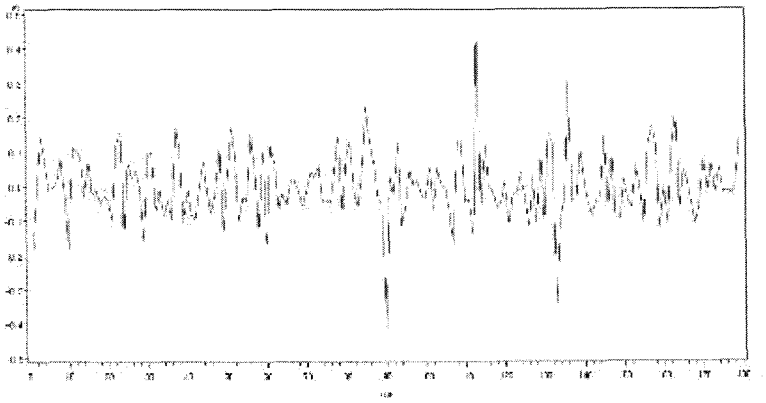

그림 3.4. 10초 자료의 로그 변환 후 차분한 시계열 그림

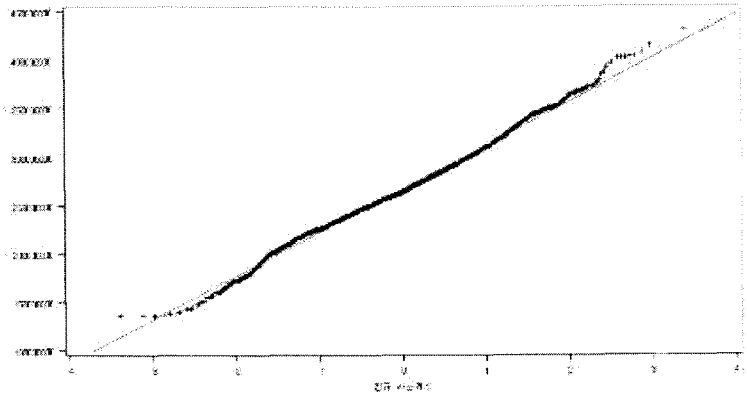

그림 3.5. 1초 자료의 QQ-Plot

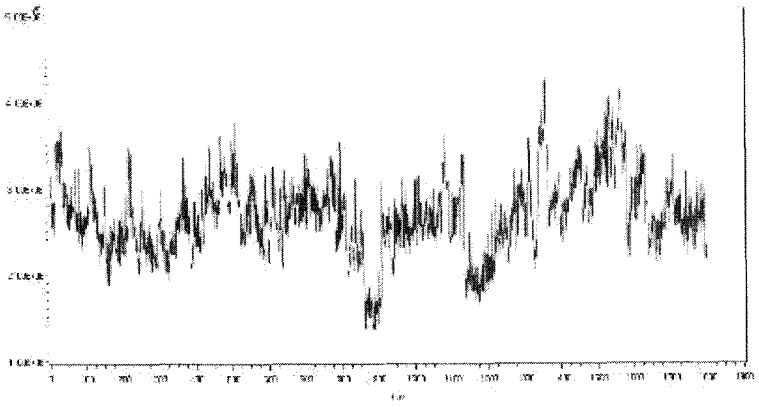

그림 3.6. 10초 자료의 QQ-Plot

표 3.3. HURST 모수 추정치

Hurst 모수 추정치(1초)	Hurst 모수 추정치(10초)
0.963	0.956

표 3.4. 실제자료 추정치

자료	GARCH(1,1)					PGARCH				
	$\widehat{\phi}$	$\widehat{\alpha_0}$	$\widehat{\alpha_1}$	$\widehat{\beta}$	$\widehat{d}$	$\widehat{\phi}$	$\widehat{\alpha_0}$	$\widehat{\alpha_1}$	$\widehat{\beta}$	$\widehat{d}$
1초	−0.2989	0.0008	0.0479	0.8453	1	−0.2932	0.0000	0.0133	0.7801	2.5925
10초	−0.0200	0.0003	0.0805	1.0444	1	−0.1419	0.0550	0.1691	0.6396	0.3090

표 3.5. 실제자료 RMSE

자료	GARCH(1,1)	PGARCH
1초	7.05024	7.0498
10초	7.71086	7.6006

한편 표 3.3은 각각의 변환된 자료에 대한 Hurst 모수 추정치를 보여주는데 값들이 모두 0.5보다 크고 1에 매우 가까운 값들로서 이것은 매우강한 장기기억 특성을 보여주고 있다. Ding과 Granger (1996)은 이것을 표본자기상관함수가 매우 큰 시차까지 상당히 큰 값을 가지는 것으로 장기기억 특성을 보여주었지만 트래픽자료에서는 Hurst 모수 추정치로서 자료의 장기기억 특성을 간단히 보여주고 있다.

표 3.4에서는 실제자료에 대한 추정치를 나타내고 있으며 여기서 d의 추정치가 1이 아닌 값들을 가지는 것을 볼 수 있다. 다음으로 표 3.5에서는 예측치와 실측치의 차이를 보여주는 RMSE(Root Mean Squared Error)를 나타내주고 있다. RMSE를 계산하기 위하여 각각 자료의 90퍼센트는 모형의 설정을 위하여 사용하였고 나머지 10퍼센트는 예측의 정확도를 구하기 위하여 사용하였다. 위의 결과를 종합하여 보면 트래픽 자료를 추정하고 예측하는데 PGARCH 모형이 기존의 GARCH 모형보다 예측 성능이 우수하다는 것을 보여준다.

4. 결과

본 연구에서는 최근에 많은 관심을 받고 있는 통신망 트래픽 자료의 예측을 위하여 기존의 GARCH 모형을 소개하고 PGARCH 모형의 예측도를 성능 실제자료를 통하여 비교하여 보았다. Feng 등 (2004)은 ARMAX/GARCH 모형이 기존의 multi-fractal wavelet 모형보다 예측의 정확도가 높음을 보였는데 본 연구에서는 GARCH 모형보다 PGARCH 모형이 예측의 성능이 우수함을 보였다. 향후 Threshold PARCH 모형과 같은 더욱 정교한 모형을 가지고 예측의 정확도를 따지는 성능평가가 요구되어 진다.

참고문헌

Basu, A., Mukherjee, A. and Klivansky, S. (1996). Time series models for internet traffic, In *Proceedings IEEE IEEE INFOCOM 96, Fifteenth Annual Joint Conference of the IEEE Computer Societies*, **4**, 24–28.

Beran, J. (1994). *Statistics for Long-Memory Processes*, Chapman & Hall/CRC, New York.

Dijk, D. V., Franses, P. H. and Lucas, A. (1999). Testing for ARCH in the presence of additive outliers, *Journal of Applied Econometrics*, **14**, 539–562.

Ding, Z. and Granger, C. W. J. (1996). Modeling volatility persistence of speculative returns: A new approach, *Journal of Econometrics*, **73**, 185–215.

Ding, Z., Granger, C. W. J. and Engle, R. F. (1993). A long memory property of stock market returns and a new model, *Journal of Empirical Finance*, **1**, 83–106.

Feng, C., He, D. and Sun, Z. (2004). IP traffic trace modelling with ARMAX/GARCH, *HET-NET'03, Ilkley*, 26–28.

Kim, S. (2007). Time series models for performance evaluation of network traffic forecasting, *The Korean Journal of Applied Statistics*, **20**, 219–227.

Shu, Y., Yu, M., Yang, O., Liu, J. and Feng, H. (2005). Wireless traffic modeling and prediction using seasonal ARIMA models, *IEICE-Transactions on Communications*, **10**, 3992–3999.

Zhou, B., He, D. and Sun, Z. (2005). *Network Traffic Modeling and Prediction with ARIMA/GARCH*, CiteULike.

Internet Traffic Forecasting Using Power Transformation Heteroscadastic Time Series Models

M.H. Ha[1] · S. Kim[2]

[1]Dept. of Statistics, Chung-Ang University; [2]Dept. of Statistics, Chung-Ang University

(Received July 2008; accepted August 2008)

Abstract

In this paper, we show the performance of the power transformation GARCH(PGARCH) model to analyze the internet traffic data. The long memory property which is the typical characteristic of internet traffic data can be explained by the PGARCH model rather than the linear GARCH model .Small simulation and the analysis of the real internet traffic show the out-performance of the PARCH MODEL over the linear GARCH one.

Keywords: GRACH model, PGARCH model, internet traffic, long memory.

[1]Graduate student, Dept. of Statistics, Chung-Ang University, Dongjack-Gu, Seoul 156-756, Korea. E-mail: ggosma@hanmail.net

[2]Corresponding author: Associate Professor, Dept. of Statistics, Chung-Ang University, Dongjack-Gu, Seoul 156-756, Korea. E-mail: sahm@cau.ac.kr