*Journal of Korean*
*Data & Information Science Society*
*2006, Vol. 17, No. 1, pp. 115 ~ 122*

# A Simultaneous Test for Multivariate Normality and Independence with Application to Univariate Residuals[1)]

Cheolyong Park[2)]

**Abstract**

A test is suggested for detecting deviations from both multivariate normality and independence. This test can be used for assessing the normality and independence of univariate time series residuals. We derive the limiting distribution of the test statistic and a simulation study is conducted to study the accuracy of the limiting distribution in finite samples. Finally, we apply our method to a real data of time series.

***Keywords*** : Chi-squared test, Independence, Normality, Time series

## 1. 서론

주어진 단변량 시계열자료 $\{X_t: t=1,2,\cdots,n\}$에 대해 설정되는 모형

$$X_t = f(X_{t-1},\cdots,X_{t-p}) + \varepsilon_t$$

에서는 통상적으로 오차 $\{\varepsilon_t: t=1,2,\cdots,n\}$가 서로 무상관이라고 가정된다. 또한 모형과 관련된 모수에 대한 검정에서는 일반적으로 오차의 정규성이 추가적으로 가정된다. 정규성이 가정될 경우 무상관성은 독립성과 동일하게 되기 때문에 이 경우 무상관성을 독립성이라고 표시하여도 상관없음을 알 수 있다. 앞에서 설정된 시계열모형의 적절성은 잔차

$$e_t = X_t - \hat{f}(X_{t-1},\cdots,X_{t-p})$$

---

1) This research was supported by the Program for the Training of Graduate Students in Regional Innovation which was conducted by the Ministry of Commerce Industry and Energy of the Korean Government.

2) Associate Professor, Department of Statistics, Keimyung University, Taegu 704-701
E-mail : cypark1@kmu.ac.kr

의 무상관성과 정규성의 탐색을 통해 간접적으로 이루어지는 것이 일반적이다. 잔차의 무상관을 탐색하는 고전적인 방법은 표본 자기상관함수(sample autocorrelation function)과 표본 편자기상관함수(sample partial autocorrelation function)를 사용하는 고전적인 Box-Jenkins 방법과 Ljung-Box (1978) 퍼트맨토 통계량(portmanteau statistic)을 이용하는 방법이 있다. 또한 정규성을 탐색하는 고전적인 방법으로는 그림을 통해 시각적으로 알아보는 정규확률그림(normal probability plot)과 단변량 정규성 검정으로 Shapiro-Wilk (1965) 검정, Kolmogorov-Smirnov 검정 (Conover(1971)의 6장 참조) 등이 있다.

이 연구는 현재 두 번에 나뉘어 실시되고 있는 단변량 시계열잔차의 무상관성과 정규성에서의 이탈을 동시에 탐색할 수 있는 하나의 검정방법을 제공하고자 하는 목적으로 시작되었다. 그 기본적인 절차는 간단하다. 먼저 주어진 잔차를 서로 겹치지 않게 다변량 자료로 변환한 후 각 변수를 표준화시킨다. 이 표준화된 변수에 대해 표준정규분포의 분위수를 이용하여 각 범주에 속할 가능성이 동일한 범주형 변수를 만든다. 이렇게 만든 범주형 변수들에서 다차원 분할표를 만들고 카이제곱 독립성 검정을 수행한다.

이 연구에서 제안하는 방법은 결국 다변량 정규성과 독립성에서의 이탈을 동시에 탐색하는 검정법이다. 먼저 개개의 원변수를 범주형 변수로 만드는 과정에 표준정규분포의 분위수를 사용함으로써 정규성에서의 이탈에 검정력을 가질 수 있게 만들었으며, 이 범주형 변수에서 만들어지는 분할표의 독립성 검정을 통해 독립성에서의 이탈을 탐색할 수 있게 만들었다. 따라서 단변량 시계열 자료를 서로 겹치지 않게 다변량 자료로 변환한 후 이 자료에 대해 본 연구에서 제안한 방법을 사용하면 단변량 시계열 자료의 정규성과 독립성을 동시에 검정하는 방법이 될 수 있는 것이다.

이 논문은 다음과 같은 순서로 구성되어 있다. 2절에서는 제안된 검정의 절차를 간략히 소개하며, 관찰도수 벡터 및 카이제곱 통계량의 독립적 다변량 정규성 하의 점근분포(asymptotic distribution)를 유도한다. 3절에서는 2절에서 유도한 카이제곱 통계량의 점근분포가 유한표본(finite samples)에서 얼마나 정확한지 살펴보는 모의실험을 수행한다. 마지막으로 실제 시계열자료에 적절한 모형을 적합한 후 나오는 잔차에 이 연구에서 제안한 방법을 적용하는 실제 적용 예제를 제시한다.

## 2. 검정통계량과 점근분포

먼저 이 논문에서 사용할 표기법에 대해서 간단히 소개하도록 하겠다. $I, e$ 및 $0$은 각각 단위행렬, 모든 원소가 1인 벡터 및 모든 원소가 0인 벡터나 행렬을 나타낸다. 차수를 나타낼 때는 첨자로서 나타내고 차수가 문맥상 분명하면 생략하기로 한다. 또한 모든 벡터는 열벡터이지만 편의상 논문에서 원소를 나열할 때는 행벡터로 나타내기도 한다.

평균이 $\mu$이고 공분산행렬이 $\Sigma$인 $p$-차원 다변량 정규분포를 $N_p(\mu, \Sigma)$로, 자유도가 $f$인 카이제곱분포를 $\chi^2(f)$로 나타내기로 하자. 또한 표준정규분포의 확률밀도함수와 누적분포함수를 각각 $\phi$와 $\Phi$로 나타내기로 하며, 표준정규분포의 $p$-분위수

를 $\xi_p = \Phi^{-1}(p)$로 나타내기로 한다.

원자료 $Y_1, Y_2, \cdots, Y_n$은 서로 독립인 다변량 정규분포에서의 확률표본이다. 다시 말해 원자료는 공분산행렬이 대각행렬 $D = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)$인 $N_p(\mu, D)$에서의 확률표본인 것이다. 여기서 각 관찰벡터의 원소는 $Y_i = (y_{i1}, \cdots, y_{ip})$로 나타내기로 하자.

이 연구에서 제안하는 절차를 소개하면 다음과 같다. 우선 원자료 벡터에서 각 변수를 표준화시킨다. 다시 말해 각 변수의 표본평균이 0이고 표본분산이 1이 되도록 변환시켜

$$z_{ij} = (y_{ij} - \overline{y_j})/s_j \quad (i = 1, \cdots, n; j = 1, \cdots, p)$$

표준화벡터 $Z_i = (z_{i1}, \cdots, z_{ip})$를 계산한다. 여기서 $\overline{y_j}, s_j$ $(j = 1, \cdots, p)$는 각각 $j$번째 변수 $y_{1j}, \cdots, y_{nj}$의 표본평균과 표본표준편차이다. 이 때 편의상 표본분산에는 분모가 $n$을 사용하는데 이것은 정규분포에서 모분산의 최대우도추정량이다. 다음으로 이 표준화벡터를 이용하여 그룹화를 통해 각 범주에 속할 확률이 근사적으로 동일한 범주형 변수를 만든다. 다시 말해 각 범주형 변수가 취하는 범주의 수를 $d$로 나타낼 때 표준정규분포의 분위수 $\xi_p = \Phi^{-1}(p)$를 이용하여 각 범주에 속할 확률이 근사적으로 동일하게 다음과 같이 변환시킨다.

$$\text{만약 } \xi_{(k-1)/d} < z_{ij} \le \xi_{k/d} \text{이면 } t_{ij} = k \text{이다. (단, } k = 1, \cdots, d)$$

여기서 $P(t_{ij} = k) \approx 1/d$가 만족되는 것을 쉽게 알 수 있다. (그룹화 과정에서 각 범주형 변수가 범주의 수를 다르게 가지도록 할 수도 있지만 이론전개와 적용의 편의를 위해 동일한 범주의 수를 가지는 것으로 한정하도록 한다.) 이 범주형 벡터 $T_i = (t_{i1}, \cdots, t_{ip})$ $(i = 1, \cdots, n)$으로부터 $\pi = (\pi_1, \cdots, \pi_p)$ $(\pi_i = 1, \cdots, d)$와 일치되는 관찰도수

$$u_\pi = \sum_{i=1}^{n} I(T_i = \pi)$$

를 계산할 수 있다. 여기서 $I(A)$은 $A$가 참이면 1이고 거짓이면 0인 표시함수(indicator function)이다. 이 관찰도수로 구성되는 $p$-차원 분할표의 독립성을 검정하는 피어슨-피셔(Pearson-Fisher) 카이제곱 검정통계량

$$X^2 = \sum_{\pi} \frac{(u_\pi - n/d^p)^2}{n/d^p}$$

을 통해 다변량 정규성과 독립성을 검정하게 된다. 여기서는 $E(u_\pi) \approx n/d^p$가 되는 것을 이용하였기 때문에 아주 계산이 쉬운 검정통계량이 되었다.

이제 서로 독립인 다변량 정규성 가정 하에서 관찰도수 $u_\pi$의 벡터와 카이제곱 검정통계량의 점근분포를 유도하도록 하겠다. 먼저 점근분포를 표시할 때 필요한 여러 벡터 및 행렬을 정의하도록 하겠다. 먼저 $U_n=(u_{n\pi})$는 차수가 $d^p\times 1$인 관찰도수 벡터이다. 이 벡터는 표준화된 벡터의 함수이며 이것을 명확히 할 필요가 있을 때는 $U_n(Z_1,\cdots,Z_n)$라고 표기하도록 하겠다. 점근분포를 쉽게 표시하기 위해서 $U_n$의 원소 $u_\pi$는 표준순서에 의해 나열되었다고 가정한다. 다시 말해, 해당 셀 벡터 $\pi=(\pi_1,\pi_2,\cdots,\pi_p)$의 첫 번째 원소 $\pi_1$이 1에서 $d$까지 가장 빨리 변하고, 두 번째 원소 $\pi_2$가 두 번째 빨리 변하고, $\pi_p$가 가장 느리게 변하는 순서에 따라 $u_\pi$가 순서대로 나열되어 $U_n$가 구성된다고 가정한다.

다음으로, $j=1,\cdots,d$에 대해서 $\zeta_j=\Phi^{-1}(j/d)$,

$$\psi_j=\phi(\zeta_{j-1})-\phi(\zeta_j),\quad \omega_j=\zeta_{j-1}\phi(\zeta_{j-1})-\zeta_j\phi(\zeta_j),$$

라고 정의하고 여기서 $\pm\infty\phi(\pm\infty)\equiv 0$라는 극한값을 관례대로 사용하도록 하겠다. $D_1$은 차수가 $d^p\times p$인 행렬로서 $i$번째 열이 벡터 $(d\psi_1 e_i, d\psi_2 e_i,\cdots,d\psi_p e_i)$을 $d^{p-i}$번 되풀이해서 만들어지는 벡터이다. 여기서 $e_i$는 1이 $d^{i-1}$개인 벡터이다. $D_2$는 행렬 $D_1$에서 $\psi_i$를 $\omega_i$로 대체해서 만들어지는 행렬이다.

관찰도수 벡터 $U_n$과 카이제곱 통계량 $X^2$의 점근분포를 유도하면 다음과 같다.

**정리 1.** $Y_1, Y_2,\cdots, Y_n$이 공분산행렬이 대각행렬 $D=diag(\sigma_1^2,\cdots,\sigma_p^2)$인 정규분포 $N_p(\mu, D)$에서의 확률표본이면, $n\to\infty$일 때

$$n^{-1/2}(U_n-n e/d^p)\xrightarrow{d} N_{d^p}(0,\Sigma)$$

가 성립한다. 여기서

$$\Sigma=I/d^p-ee^t/d^{2p}-D_1D_1^t/d^{2p}-D_2D_2^t/(2d^{2p})$$

이다.

**증명**: 관찰도수 벡터 $U_n(Z_1,\cdots,Z_n)$은 표준화된 벡터 $Z_1,\cdots,Z_n$의 함수이므로 위치-척도 불변통계량(location-scale invariant statistic)이다. 따라서 이 벡터의 분포는 위치-척도 모수인 $\theta=(\mu, D)$에 의존하지 않기 때문에 보조통계량(ancillary statistic)이며 $\theta$의 완비충분통계량(complete and sufficient statistic)인 $(\overline{y_1},\cdots,\overline{y_p}, s_1^2,\cdots,s_p^2)$과 서로 독립이 된다.

따라서 모든 $\theta$에 대해

$$\begin{aligned}&\mathcal{L}_{\theta}(U_n(Z_1,\cdots,Z_n)) = \mathcal{L}_{\theta_0}(U_n(Z_1,\cdots,Z_n))\\ &= \mathcal{L}_{\theta_0}\left(U_n(Z_1,\cdots,Z_n)\,|\,\overline{y_i}=0, s_i^2=1\ \forall i\right)\\ &= \mathcal{L}_{\theta_0}\left(U_n(Y_1,\cdots,Y_n)\,|\,\overline{y_i}=0, s_i^2=1\ \forall i\right)\end{aligned} \tag{1}$$

가 성립한다. 여기서 $\theta_0=(0,I)$는 고정된 값이며, 마지막 등식은 표본평균이 0이고 표본분산이 1이면 표준화된 변수는 원변수와 같아지기 때문이다. Park (1995)의 조건부 극한이론(conditional limit theorem)을 적용하기 위해서는 앞의 식 (1)의 조건 $\overline{y_i}=0, s_i^2=1\ \forall i$를 정준충분통계량(canonical sufficient statistic)으로 나타낼 필요가 있다. 임의의 $p$-벡터 $y=(y_1,\cdots,y_p)$에 대해 $s(y)=(y, d(y))$라고 정의하자. 단 여기서 $d(y)=(y_1^2,\cdots,y_p^2)$로서 제곱으로 구성된 벡터이다. 그러면 정준충분통계량은 $\sum_{i=1}^{n}s(y_i)$로 주어지면 조건은 $\sum_{i=1}^{n}s(y_i)/n=(0_p,e_p)$가 된다. 따라서 식 (1)은

$$\mathcal{L}_{\theta_0}\left(U_n(Y_1,\cdots,Y_n)\,|\,\sum_{i=1}^{n}s(Y_i)/n=(0_p,e_p)\right)$$

가 된다. 다변량 정규분포에서는 Park (1995)의 Corollary 1의 조건을 쉽게 만족하는 것을 보일 수 있다. (예를 들어 Park (1999)의 Theorem 1을 참조할 수 있다.) 따라서 $n\to\infty$일 때

$$n^{-1/2}(U_n - n\,e/d^p) \xrightarrow{d} N_{d^p}(0, A-BC^{-1}B^t)$$

가 성립한다. 여기서

$$A=\mathrm{Cov}_{\theta_0}(U_1(Y_1)),\ B=\mathrm{Cov}_{\theta_0}(U_1(Y_1),s(Y_1)),\ C=\mathrm{Cov}_{\theta_0}(s(Y_1))$$

이다. 그런데

$$A=I/d^p-ee^t/d^{2p},\ B=(D_1,D_2)/d^p,\ C=\mathrm{diag}(I_p,2I_p)$$

가 되는 것을 쉽게 확인할 수 있기 때문에

$$\Sigma=A-BC^{-1}B^t=I/d^p-ee^t/d^{2p}-D_1D_1^t/d^{2p}-D_2D_2^t/(2d^{2p})$$

가 성립하며 증명이 완료된다. □

**정리 2.** 정리 1의 조건 하에서 $n\to\infty$일 때

$$X^2 \xrightarrow{d} W_1+(1-da)W_2+(1-db/2)W_3$$

를 만족한다. 여기서 $W_1, W_2, W_3$은 서로 독립이며 각각 자유도가 $d^p-1-2p$, $p$, $p$인 카이제곱 확률변수이며, $a=\sum_{j=1}^{d}\psi_j^2$, $b=\sum_{j=1}^{d}\omega_j^2$이다.

**증명**: $X^2=(U_n-ne/d^p)^t(U_n-ne/d^p)/(n/d^p)$가 만족하기 때문에 $X^2$의 점근분포는 $\sum_{i=1}^{d^p}\lambda_i W_i$가 된다. 여기서 $\lambda_i$는

$$E=d^p\Sigma=I-ee^t/d^p-D_1D_1^t/d^p-D_2D_2^t/(2d^p)$$

의 고유값이며 $W_i$는 서로 독립인 자유도가 1인 카이제곱 확률변수이다.

간단한 연산에 의해

$$\sum_{j=1}^{d}\psi_j=\sum_{j=1}^{d}\omega_j=\sum_{j=1}^{d}\psi_j\omega_j=0$$

이 만족되는 것을 쉽게 알 수 있다. 따라서 $e^tD_1=0$, $e^tD_2=0$, $D_1^tD_2=0$이며

$$e^te=d^p,\ D_1^tD_1=d^{p+1}aI, D_2^tD_2=d^{p+1}bI$$

가 만족된다. 따라서 $e^te/d^p$, $D_1^tD_1/(d^{p+1}a), D_2^tD_2/(d^{p+1}b)$는 서로 직교하며 계수가 각각 $1, p, p$인 멱등행렬(idempotent matrix)이 된다. 따라서 $E$의 고유값은 1이 $d^p-1-2p$번, $1-da$가 $p$번, 그리고 $1-db$가 $p$번 반복되며 0이 한번 나타난다. 따라서 증명이 완료된다. □

## 3. 모의실험과 적용 예제

이 절에서는 먼저 정리 2에 주어진 카이제곱 검정통계량의 점근분포가 유한표본에서 얼마나 정확한지 알아보는 간단한 모의실험을 실행하고, 다음으로 실제 시계열자료에 모형을 적용하고 나오는 잔차에 이 연구에서 제안한 방법을 적용해보고 기존의 고전적인 방법과 비교하는 예제를 제공한다.

카이제곱 검정통계량의 점근분포에 대한 모의실험 결과는 한 가지 경우만 제공하는데, 아주 많은 경우에 대해 모의실험을 해 보았지만 셀의 숫자인 $d^p$가 너무 작지 않으면 모두 비슷한 결과를 제시하기 때문이다. 이 모의실험에서 사용한 경우는 $p=d=3$인 경우로서 카이제곱 검정통계량의 점근분포는

$$\chi^2(20)+0.207\chi^2(3)+0.779\chi^2(3)$$

가 된다. 이 모의실험에서는 네 가지의 표본크기 $n=27,54,135$ 및 $270$을 고려하였

는데 이 경우 각 셀의 평균관찰도수(average cell count)가 1, 2, 5 및 10이 된다.

모의실험의 순서는 다음과 같다. 우선 각 표본크기 $n$에 대해 $N_3(0, I)$에서 표본크기 $n$인 표본을 500개 생성하여 각각의 카이제곱 검정통계량을 계산한다. 이 500개의 카이제곱 통계량의 순서통계량과 점근분포에서 계산되는 기대값을 산포도로 그리는 카이제곱확률도(chi-squared probability plot)를 그려 점근분포의 정확성을 간접적으로 판단한다. 여기서 점근분포가 카이제곱 확률분포의 선형결합형태로 나타나기 때문에 수리적으로 순서통계량의 기대값을 구하지 않고 모의실험 방법에 의존하여 계산하였다. 구체적으로 점근분포에서 표본크기 500인 표본을 100개 생성하여 100개 표본의 순서통계량의 평균값을 계산하여 기대값으로 사용한 것이다. 모의실험의 결과인 카이제곱확률도를 정리한 것이 <그림 1>에 주어져 있다.

<그림 1>의 카이제곱확률도에는 원점을 지나고 기울기가 1인 직선을 그려 넣었는데 이 직선은 카이제곱 통계량의 표본분포가 점근분포와 정확하게 일치하게 되는 이상적인 경우에 해당되는 선이다. 카이제곱확률도를 관찰하여 보면 평균관찰도수가 2, 5 및 10인 경우는 점근분포가 아주 좋은 근사분포라는 것을 알 수 있다. 그러나 평균관찰도수가 1인 경우는 이산성(discreteness)이 너무 심해 점근분포가 좋은 근사분포라고 판단하기 힘든 것을 알 수 있다.

다음으로 실제 시계열자료에 모형을 적합시켜 나오는 잔차의 정규성과 무상관성을

<그림 1> 카이제곱확률도

평가하는 실제 적용 예제를 살펴보도록 하자. 이 연구에서 사용하는 시계열 자료는

미국 Wyoming 주의 Yellowstone 국립공원의 간헐천 자료(geyser data)이다. 이 자료는 S-Plus에 자료집합으로 나와 있으며 분출사이의 대기시간과 분출지속시간의 두 가지 시계열이 399개 있는데 여기서는 대기시간을 이용하도록 하겠다.

대기시간에 S-Plus의 AR이라는 절차를 적용하였는데 이것은 아카이케 정보기준(Akaike information criterion)에 의해 최적의 자기회귀모형을 찾아주게 된다. 자기상관계수를 추정하는 방식으로 율-워커 방정식(Yule-Walker equation)을 사용하였더니 AR(2)이 최적의 모형으로 선택되었다. 차수 25까지의 자기상관계수와 편자기상관계수는 모두 95% 신뢰구간 안에 위치하고 있어 Box-Jenkins 방법에 의한 잔차의 무상관성에는 문제가 없는 것으로 나타났다. 보다 공식적으로 6, 12, 18, 24개의 표본 자기상관계수를 사용하여 행한 Ljung-Box (1978) 검정에서도 (근사적) 유의확률이 모두 0.3 이상으로 나와 무상관성에는 문제가 없는 것으로 나타났다. 또한 Shapiro-Wilk (1965) 검정 및 Kolmogorov-Smirnov 검정에서 모두 (근사적) 유의확률이 0.15보다 크게 나타나 잔차의 정규성을 의심할 만한 증거 역시 제시되지 못했다.

이제 앞의 대기시간 시계열잔차에 이 연구에서 제안한 방법을 적용해 보도록 하겠다. 먼저 3개의 잔차를 겹치지 않도록 묶어서 99개의 3차원 다변량 관찰치로 만들어 $d=3$을 적용하였더니 카이제곱 통계량이 41.45로서 z점수(z-score)가 2.79가 나왔다. z점수가 2.79이기 때문에 정규성이나 독립성에서 다소간의 문제가 있음을 암시한다. 실제로 바로 전시점과 현시점 잔차사이의 산포도를 그려 보면 전시점 잔차의 값이 커짐에 따라 현시점 잔차의 분산이 커지는 경향이 발생하고 있어 독립성을 만족하고 있다고 보기 힘들다는 것을 발견할 수 있다. 따라서 고전적인 무상관성이나 정규성 탐색 및 검정 방법에서 쉽게 찾아내지 못한 것을 이 연구에서 제안한 동시 검정방법에서 찾아낸 것이다.

## 참고문헌

1. Conover, W.J. (1980). *Practical Nonparametric Statistics*, Second Edition, John Wiley & Sons, New York.
2. Ljung, G.M. and Box, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, 297-303.
3. Park, C. (1995). Some remarks on the chi-squared test with both margins fixed, *Communications in Statistics - Theory and Methods,* 24, 653-61.
4. Park, C. (1999). A note on the chi-square test for multivariate normality based on the sample Mahalanobis distances, *Journal of the Korean Statistical Society,* 28, 479-488.
5. Shapiro, S.S. and Wilk, M.B. (1965). An analysis of variance test for normality (complete samples), *Biometrika,* 52, 591-611.

[ 2005년 12월 접수, 2006년 2월 채택 ]