The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.4.579
30(4), 579–590

Multivariate empirical distribution plot and goodness-of-fit test

Chong Sun Hong[a,1] · Yongho Park[a] · Jun Park[a]

[a]Department of Statistics, Sungkyunkwan University

(Received June 14, 2017; Revised July 12, 2017; Accepted July 26, 2017)

Abstract

The multivariate empirical distribution function could be defined when its distribution function can be estimated. It is known that bivariate empirical distribution functions could be visualized by using Step plot and Quantile plot. In this paper, the multivariate empirical distribution plot is proposed to represent the multivariate empirical distribution function on the unit square. Based on many kinds of empirical distribution plots corresponding to various multivariate normal distributions and other specific distributions, it is found that the empirical distribution plot also depends sensitively on its distribution function and correlation coefficients. Hence, we could suggest five goodness-of-fit test statistics. These critical values are obtained by Monte Carlo simulation. We explore that these critical values are not much different from those in text books. Therefore, we may conclude that the proposed test statistics in this work would be used with known critical values with ease.

Keywords: Empirical distribution plot, Quantile vector, normal mixture

1. 서론

일변량 확률표본 $\underline{X} = (X_1, \ldots, X_n)$에서는 X_i와 X_j의 크기순서를 정할 수 있어 일변량 경험분포함수(empirical distribution function)를 쉽게 정의할 수 있다. 그러나 이변량 이상인 k변량의 확률표본 $\underline{X^k} = (\underline{X_1^k}, \ldots, \underline{X_n^k})$에서는 $\underline{X_i^k} = (X_{1i}, \ldots, X_{ki})$와 $\underline{X_j^k} = (X_{1j}, \ldots, X_{kj})$의 순위를 결정하기 쉽지 않다. 다변량 확률표본에서의 크기 순서는 공분산행렬을 포함한 분포함수에 의존한다는 사실을 기반으로 Hong 등 (2017)은 다변량 확률표본의 특정한 분포함수를 추정하면서 다변량 경험분포함수를 정의하고 특징을 연구하였다. 특히 이변량 확률표본에서 추정한 분포함수 하에서 이변량 경험분포함수를 구하면서 이를 시각화하는 계단그림(step plot)과 분위그림(quantile plot)을 제안하였다.

본 연구에서는 이변량 이상의 경험분포함수를 이차원 평면에 표현할 수 있는 경험분포그림(empirical distribution plot)을 제안하고 특징을 살펴본다. 경험분포그림은 변량의 수와 분포함수에 따라 다양한 형태로 표현되는 것을 탐색할 수 있으므로 본 연구에서는 경험분포그림을 다변량 경험분포함수의 바탕이 되는 분포함수의 추정이 적절한지를 검정할 수 있는 적합도 검정방법을 제안한다.

다변량 분포함수의 적합성 검정에 관한 문헌 중에서 다음과 같이 세 종류의 방법으로 구분할 수 있다. 첫 번째로 확률변수들의 선형결합을 이용하여 일변량 확률변수로 변환하는 방법을 사용하고

[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: cshong@skku.edu

(D'Agostino와 Stephens, 1986; Kim, 2005; Kim, 2006; Roy, 1953; Royston, 1983; Thode, 2002; Zhu 등, 1997), 두 번째로 변수변환을 이용하여 균일분포로 변환하여 사용한다 (Justel 등, 1997; Meintanis와 Hlávka, 2010; Rosenblatt, 1952). 그리고 마지막 방법으로는 표준화된 확률변수의 제곱한 카이제곱분포를 이용한다 (Gnanadesikan와 Kettenring, 1972; Gnanadesikan 등, 1977; Kim, 2004; Koziol, 1982; Malkovich와 Afifi, 1973; Moore와 Stubblebine, 1981; Singh, 1993). 본 연구에서는 다변량 경험분포함수에서 추정한 확률표본의 분포함수의 적합성 검정은 본 연구에서 제안한 경험분포그림을 바탕으로 하며, 일변량 분포함수에 대한 대표적인 다섯 종류의 적합도 검정인 Kolmogorov-Smirnov 검정방법 (Kolmogorov, 1933; Smirnov, 1933), Kuiper 검정방법 (Kuiper, 1960), Cramer-Von Mises 검정방법 (Anderson, 1962), Watson 검정방법 (Watson, 1961), 그리고 Anderson-Daring 검정방법 (Anderson과 Darling, 1952, 1954)을 사용한다.

본 논문의 구성은 다음과 같다. 2장에서는 이변량 이상의 경험분포함수를 이차원 평면에 표현할 수 있는 경험분포그림을 제안하고 특징을 살펴본다. 우선 상관계수가 양수와 음수인 두 종류의 이변량 경험분포함수를 구하고 이를 바탕으로 다변량 경험분포그림을 구현한다. 다양한 상관계수를 갖는 이변량 정규분포와 특정한 형식의 분산공분산 행렬을 갖는 삼변량 정규분포 그리고 정규혼합분포를 따르는 확률표본에 대한 경험분포그림을 표현하면서 특징을 토론한다. 다변량 경험분포그림은 경험분포함수의 바탕이 되는 분포함수에 따라 다양한 형태로 표현되므로 다변량 경험분포그림을 통하여 분포함수의 특징을 파악할 수 있다. 따라서 경험분포그림을 통해 추정한 분포함수의 적합도 검정방법을 3장에서 제안한다. 대표적인 다섯 종류의 일변량 분포함수의 적합도 검정통계량을 사용하고, 이변량 이상 그리고 다양한 종류의 분산공분산행렬을 가진 분포함수들에 대하여 다섯 종류의 검정통계량의 기각역을 구한다. 본 연구에서 구한 각각의 검정통계량의 기각역과 일반 문헌에서 쉽게 구할 수 있는 기각역을 비교하면서 검정방법을 토론한다. 4장에서는 이변량 이상의 분포함수를 따르는 실증 예제를 통하여 본 연구에서 제안한 경험분포그림을 작성하고 이를 바탕으로 적절한 분포함수의 적합도 검정 수행하고 설명한다. 그리고 마지막 결론을 5장에서 유도한다.

2. 다변량 경험분포그림

2.1. 이변량 정규분포의 경험분포그림

Hong 등 (2017)은 표본크기 n의 확률표본 $\{(X_{1i}, X_{2i}, \ldots, X_{ki}),\ i = 1, \ldots, n\}$의 누적분포함수를 $\hat{F}(\cdot, \ldots, \cdot)$으로 추정하면서 다변량 경험분포함수를 다음과 같이 제안하였다.

$$\hat{F}_n = \hat{F}_n(x_1, x_2, \ldots, x_k) = \frac{1}{n}\sum_{i=1}^{n} I\left(\hat{F}_i \le f\right), \tag{2.1}$$

여기서 확률변수 $\hat{F}_i = \hat{F}(X_{1i}, X_{2i}, \ldots, X_{ki})$이며, f는 0과 1사이의 상수이고 $\hat{F}(x_1, \ldots, x_k)$값이다. 예를 들어 $\hat{F}(\cdot, \ldots, \cdot)$를 평균이 0벡터이고 분산공분산행렬은 $\hat{\Sigma}$인 표준정규분포함수로 추정한다면, $\hat{F}_n = (1/n)\sum_{i=1}^{n} I(\hat{\Phi}_i \le f)$, 여기서 $\hat{\Phi}_i = \Phi(X_{1i}, \ldots, X_{ki}; \underline{0}, \hat{\Sigma})$이다.

이변량 확률표본에 대하여 경험분포함수를 구한 Table 2.1을 살펴보자. 표본 $n = 20$개를 Table 2.1의 각각의 2열에 (x, y)로 정리하였다. 각 자료의 표본상관계수를 포함한 이변량 표준정규분포를 가정하고, $\Phi(x_1, x_2; r = -0.61)$와 $\Phi(x_1, x_2; r = 0.69)$을 구하고 이를 $\hat{\Phi}_i$로 표기하여 3열에 각각 나타내고, $\hat{\Phi}_i$의 크기 순서를 4열에 각각 나타내었다. 그리고 $\hat{\Phi}_i$에 대응하는 이변량 경험분포함수를 $\hat{F}_n$을 Table 2.1의 마지막 열에 각각 정리하였다. 예를들어 첫 번째 관찰값은 $\Phi_1 = \Phi(-0.67, 0.69; r = -0.61) = 0.117$와 $\Phi_1 = \Phi(0.92, 0.62; r = 0.69) = 0.678$이다. 이에 대응하는 크기순서는 각각 9와 16이며, 이변량 경험분포함수 $\hat{F}_{20}$값은 각각 $9/20 = 0.45$와 $16/20 = 0.80$이다.

Table 2.1. Two bivariate random samples from normal distributions ($r = -0.61, 0.69$)

	$(x_{1i}\ ,\ x_{2i})$	$\hat{\Phi}_i$	(i)	$\hat{F}_n(x_{1i}, x_{2i})$		$(x_{1i}\ ,\ x_{2i})$	$\hat{\Phi}_i$	(i)	$\hat{F}_n(x_{1i}, x_{2i})$
1	(−0.67, 0.69)	0.117	9	0.45	1	(0.92, 0.62)	0.678	16	0.80
2	(1.50, 0.42)	0.598	20	1.00	2	(1.45, 0.40)	0.648	15	0.75
3	(0.45, 0.30)	0.333	18	0.90	3	(0.26, 0.30)	0.485	12	0.60
4	(2.27, −1.65)	0.044	5	0.25	4	(−0.15, 0.21)	0.371	10	0.50
5	(−1.51, −0.21)	0.003	2	0.10	5	(0.84, 0.87)	0.717	17	0.85
6	(0.40, −0.46)	0.122	11	0.55	6	(−0.90, −0.63)	0.125	4	0.20
7	(0.09, 0.29)	0.232	15	0.75	7	(1.07, 1.62)	0.841	18	0.90
8	(1.74, −0.24)	0.368	19	0.95	8	(0.61, −1.16)	0.122	3	0.15
9	(−0.85, 1.09)	0.121	10	0.50	9	(0.08, 0.94)	0.512	13	0.65
10	(0.33, −1.25)	0.020	3	0.15	10	(−1.08, −1.02)	0.075	2	0.10
11	(−0.55, 0.33)	0.095	8	0.40	11	(0.07, −0.74)	0.206	8	0.40
12	(−0.28, 0.37)	0.156	12	0.60	12	(−0.76, −0.26)	0.179	7	0.35
13	(−0.52, 1.37)	0.237	16	0.80	13	(1.44, 1.20)	0.852	19	0.95
14	(0.25, 1.21)	0.308	17	0.85	14	(−0.49, 0.35)	0.288	9	0.45
15	(0.42, −0.03)	0.229	14	0.70	15	(0.39, 0.88)	0.609	14	0.70
16	(−2.41, 0.98)	0.002	1	0.05	16	(−1.20, −1.82)	0.023	1	0.05
17	(0.60, −0.98)	0.058	7	0.35	17	(0.11, 0.33)	0.456	11	0.55
18	(0.98, −0.45)	0.210	13	0.65	18	(0.25, −1.00)	0.150	6	0.30
19	(−0.81, 0.03)	0.035	4	0.20	19	(−0.92, −0.36)	0.142	5	0.25
20	(0.97, −1.22)	0.052	6	0.30	20	(1.93, 1.29)	0.892	20	1.00

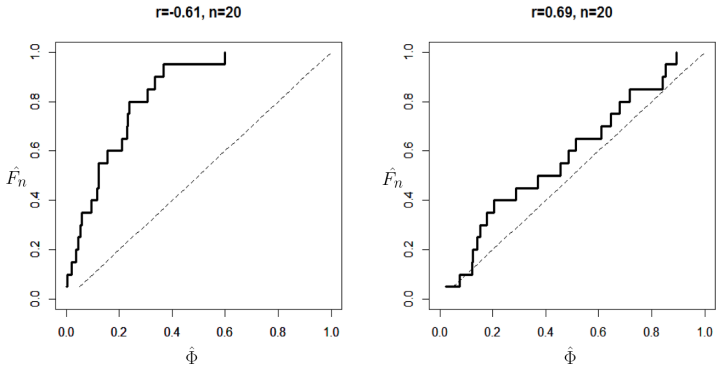

Figure 2.1. Empirical distribution plots.

Table 2.1에서 이변량 경험분포함수 $\hat{F}_n$는 0.05 (=1/20)부터 1.0까지의 값으로 나타나지만, 왼쪽 표에서 정규분포함수 $\hat{\Phi}_i$의 최대값은 0.598이며 오른쪽 표에서는 0.892임을 탐색할 수 있다. 또한 상관관계가 음수인 경우에 $\hat{\Phi}_i$는 작은 값들이며 서로 밀집하게 나타나고, 상관관계가 양수인 경우는 $\hat{\Phi}$값들이 0부터 1까지 넓게 퍼져 경험분포함수 $\hat{F}_n$값과 비슷한 추세를 보인다. 이와 같은 특징들을 시각적으로 표현하기 위해 경험분포함수 $\hat{F}_n$값을 이차원 평면의 수직축에 나타내고 수평축에는 확률표본의 분포함수로 추정한 정규분포함수값 $\hat{\Phi}_i$으로 표현하여 좌표 $(\hat{\Phi}_i, \hat{F}_n)$를 연결하는 곡선으로 나타내는 Figure 2.1과 같은 방법을 제안하며, 이를 경험분포그림(empirical distribution plot)을 제안한다. 일반적인 분포함수 가정 하에서의 경험분포그림은 $\hat{\Phi}_i$를 $\hat{F}_i$로 대체하여 $(\hat{F}_i, \hat{F}_n)$의 좌표로 표현할 수 있다.

Figure 2.1은 좌표 $(\hat{\Phi}_i, \hat{F}_n)$을 연결하여 계단 형태로 표현된다. 그리고 정규확률그림(normal probabil-

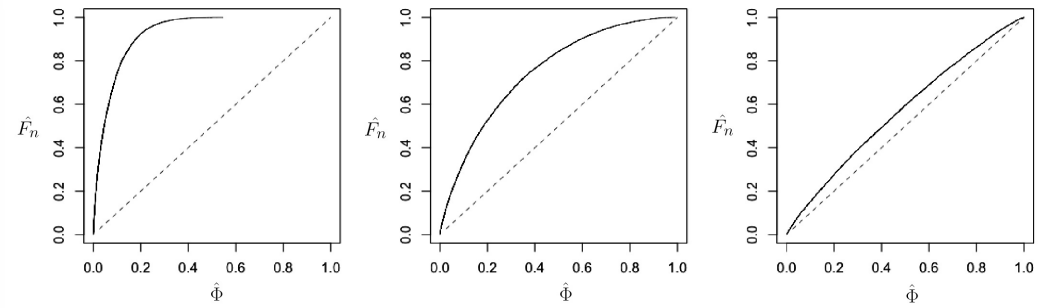

Figure 2.2. Various empirical distribution plots ($\rho = -0.9, 0, 0.9$).

ity plot)과 같은 참조선(reference line)을 대각선으로 나타내었다. 만약 일변량 경험분포그림은 정규확률그림과 유사하게 표현되고, 일변량 확률표본이 정규분포를 따른다면 곡선이 대각선인 참조선과 일치하게 됨을 확인할 수 있다. 표본상관계수가 음수인 Figure 2.1의 왼쪽 경험분포그림은 $\hat{\Phi}_i$값이 $\hat{F}_n$에 비해 작은 값들이 밀집하여 나타나므로 그림의 참조선보다 왼쪽 그리고 상단부분에 치우쳐져 형성되며 $\hat{\Phi}$값이 1에 근접하지 않는다. 그러나 표본상관계수가 양수인 Figure 2.1의 오른쪽 경험분포그림은 $\hat{\Phi}_i$값과 $\hat{F}_n$가 서로 비슷한 추세로 왼쪽 아랫부분부터 오른쪽 윗부분까지 동시에 증가하며 참조선의 바로 왼쪽에 위치한다.

이변량 표준정규분포의 상관계수를 $\rho = -0.9,\ 0,\ +0.9$인 분포로부터 표본크기 10,000을 추출하여 경험분포그림을 Figure 2.2에 구현하였다. 이변량 표준정규분포의 상관계수가 음수일 경우에 $\hat{\Phi}_i$는 상대적으로 1보다 작은 값들로 밀집해서 나타나므로 경험분포그림에서의 $\hat{F}_n$가 $\hat{\Phi}_i$에 비해 급격히 상승하여 1에 먼저 접근하여 $(\hat{\Phi}_i, \hat{F}_n)$을 표현한 곡선은 참조선보다 왼쪽에 치우쳐져 있고, $\hat{F}_n$와 $\hat{\Phi}_i$의 차이 간격이 크다. 그러나 상관계수가 -0.9에서 0.9으로 증가함에 따라 $\hat{\Phi}_i$가 0부터 1까지 넓게 퍼져있으며 차이 간격은 점점 줄어들고 $\hat{F}_n$와 유사한 형태로 증가하여 참조선과 점점 비슷한 추세를 보인다. 이는 상관계수가 증가함에 따라 뚜렷하게 나타나는데, 상관계수가 0.9인 경우에는 $\hat{\Phi}_i$와 $\hat{F}_n$가 큰 차이없이 유사하게 나타난다. 따라서 이변량 표준정규분포함수 하에서의 경험분포그림은 상관계수에 따라 서로 다르게 표현되며 참조선과의 차이가 뚜렷하게 나타남을 발견하였다.

삼변량 경험분포함수에 대한 경험분포그림의 특징을 살펴보기 위해 우선 가장 간단한 경우인 모평균이 영벡터이고 분산공분산행렬이 $\Sigma = \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}$인 삼변량 정규분포함수를 고려하고 이에 대응하는 경험분포그림을 구현하였다. 삼변량 경험분포그림은 이변량과 유사하게 ρ가 증가할수록 참조선에 접근하지만 이변량보다 접근속도는 조금 떨어지는 것을 확인할 수 있다. 그리고 4.1절 실증예제에서는 복잡한 분산공분산행렬을 포함하는 삼변량 정규분포함수에 대한 경험분포그림을 구현하고 살펴본 결과, 분산공분산행렬의 형태와 다변량 정규분포에 따라 참조선과의 접근 정도가 다르며 경험분포그림의 차이가 발생하는 것을 탐색할 수 있다.

2.2. 특정한 다변량분포의 경험분포그림

정규분포가 아닌 특정한 분포함수를 고려하기 위하여 다음과 같은 삼변량 정규혼합분포를 고려한다.

$$F(X_1, X_2, \ldots, X_k) = 0.5\Phi(X_1, X_2, X_3; \mu_1, \Sigma_1) + 0.5\Phi(X_1, X_2, X_3; \mu_2, \Sigma_2), \tag{2.2}$$

여기서 $\mu_1 = (0,0,0)^t$, $\mu_2 = (2,2,2)^t$ 그리고 Σ_1과 Σ_2는 2.1절에서 논의한 삼변량 표준정규분포의 분

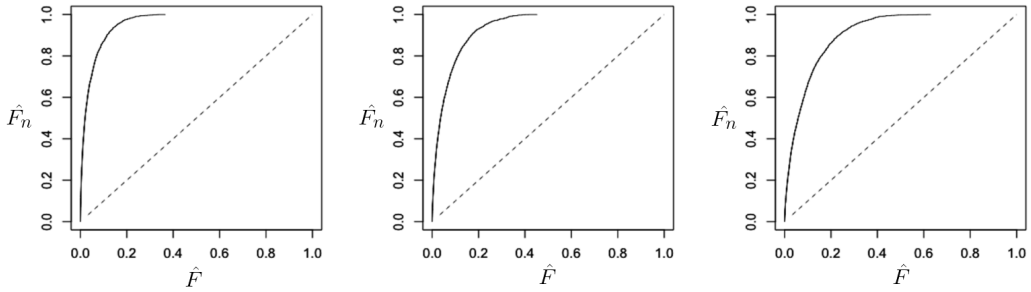

Figure 2.3. Tri-variate empirical distribution plots ($\rho = -0.6, 0, 0.6$).

산공분산행렬로 설정한다. 상관계수 ρ는 -0.6, 0.0, 0.6인 경우에 크기 10,000의 표본을 추출하여 2.1에서 정의한 경험분포함수 $\hat{F}_n$와 추정한 정규혼합분포 $\hat{F}_i$를 구한 후 삼변량 경험분포그림을 Figure 2.3에 구현하였다.

Figure 2.3를 살펴보면, 이변량과 마찬가지로 상관계수가 음수일 경우에 $\hat{F}_i$값이 작게 나타나 경험분포함수 $\hat{F}_n$와의 차이가 크며, 상관계수가 음수에서 양수로 증가할수록 차이가 줄어든다. 이변량보다 $\hat{F}_i$값들이 작게 나타나 상관계수가 음수일 때 왼쪽으로 더 치우쳐 있으며 삼변량 혼합 경험분포그림의 $(\hat{F}_i, \hat{F}_n)$곡선과 참조선과의 차이는 삼변량정규분포보다 차이가 나고 이변량정규분포보다는 더욱 큰 차이가 나타난다.

Figure 2.2와 Figure 2.3을 통해 탐색하였듯이 이변량 정규분포에서 상관계수의 크기에 따라 경험분포그림이 다르게 표현되었으며, 삼변량 정규분포에 대한 경험분포그림은 이변량 경험분포그림과 차이가 있으며 심지어 동일한 상관계수를 가질 때에도 차이를 발견할 수 있다. 본 논문에서는 이변량 정규분포와 삼변량 정규혼합분포만을 비교 분석하였지만, 일반적으로 변량의 수가 증가하거니 분산공분산행렬의 형태가 복잡할수록 경험분포그림의 형태는 다르게 표현되며 나아가 다변량 분포함수의 종류에 따라 경험분포그림의 형태는 다르다는 사실을 탐색할 수 있다. 경험분포그림은 다변량 경험분포함수를 이차원 평면에 표현할 수 있는 장점이 있으며, 다변량 경험분포함수를 구하기 위해 가정한 다변량 분포함수의 변량수와 공분산행렬 형태 그리고 분포함수의 종류에 따라 경험분포그림의 형태는 변하는 것을 탐색할 수 있다.

3. 다변량 경험분포그림을 이용한 적합도 검정

3.1. 적합도 검정방법

경험분포함수를 구하기 위해 추정한 다변량 분포함수에 따라 경험분포그림의 형태는 다르게 표현되는 경험분포그림의 특징을 이용하여, 경험분포함수를 구하기 위한 다변량 분포함수의 적합도검정에 대해 살펴본다.

표본크기 n 그리고 k변량의 확률표본 $X^k = (X_1^k, \ldots, X_n^k)$의 누적분포함수 $F(\cdot, \ldots, \cdot)$를 특정한 분포함수 $F^0(\cdot, \ldots, \cdot)$이라고 설정한 귀무가설과 대립가설은 다음과 같다.

$$H_0 : F(x_1, \ldots, x_k) = F^0(x_1, \ldots, x_k) \quad \text{vs.} \quad H_1 : F(x_1, \ldots, x_k) \neq F^0(x_1, \ldots, x_k).$$

귀무가설 H_0하에서 다변량 경험분포함수에서의 분포함수는 $F^0(\cdot, \ldots, \cdot)$이므로 대표본크기 m의 확률

표본으로부터 다음과 같은 경험분포함수를 구한다.

$$F_m^0 = F_m^0(x_1, \ldots, x_k) = \frac{1}{m}\sum_{j=1}^{m} I(F_j^0 \le f), \tag{3.1}$$

여기서 $F_j^0 = F^0(X_{1j}, \ldots, X_{kj})$이며, f는 0과 1 사이의 값으로 간주한다.

2장에서 논의한 크기 n의 확률표본에 대한 경험분포그림의 좌표는 $(\hat{F}_i, \hat{F}_n)$이며 귀무가설 하에서 대표본의 확률표본에 대한 경험분포그림의 좌표는 (F_j^0, F_m^0)이므로 수평축의 좌표가 일치할 때 즉, $\hat{F}_i = F_j^0$일 때 $\hat{F}_n$과 F_m^0의 차이에 대응하는 적합도 검정통계량을 고려한다. 본 연구에서는 일변량 분포함수의 적합도 검정에 많이 사용하는 다섯 종류의 대표적인 검정통계량인 Kolmogorov-Smirnov K-S통계량 (Kolmogorov, 1933; Smirnov, 1933), Kuiper V 통계량 (Kuiper, 1960; Watson, 1961), Cramer-Von Mises W^2통계량 (Anderson, 1962), Watson U^2 통계량 (Watson, 1961; Stephens, 1965), 그리고 Anderson-Daring A^2 통계량 (Anderson과 Darling, 1952, 1954)을 다변량 분포함수의 적합도 검정통계량으로 다음과 같이 변환하여 제안한다.

$$\begin{aligned}
K\text{-}S &= \text{Max}\left|\hat{F}_n - F_m^0\right|, \quad \text{when } \hat{F}_i = F_j^0,\\
V &= \text{Max}\left|\hat{F}_n - F_m^0\right| + \text{Max}\left|F_m^0 - \hat{F}_n\right|, \quad \text{when } \hat{F}_i = F_j^0,\\
W^2 &= n\int_{-\infty}^{\infty}\left(\hat{F}_n - F_m^0\right)^2 dF^0(x), \quad \text{when } \hat{F}_i = F_j^0,\\
U^2 &= n\int_{-\infty}^{\infty}\left[\left(\hat{F}_n - F_m^0\right) - \int_{-\infty}^{\infty}\left(\hat{F}_n - F_m^0\right)dF^0(x)\right]^2 dF^0(x), \quad \text{when } \hat{F}_i = F_j^0,\\
A^2 &= n\int_{-\infty}^{\infty}\frac{\left(\hat{F}_n - F_m^0\right)^2}{\hat{F}_n\left(1-\hat{F}_n\right)}dF^0(x), \quad \text{when } \hat{F}_i = F_j^0.
\end{aligned}$$

귀무가설 H_0하에서의 대표본크기 $m = 10{,}000$개를 추출해 F_j^0와 F_m^0를 설정하고, 소표본 크기 $n(= 10, 20, 30, 40$ 등)의 확률표본에서의 $\hat{F}_i$과 $\hat{F}_n$를 구한다. 그리고 $\hat{F}_i = F_j^0$일 때 경험분포함수들의 차이인 $\hat{F}_n - F_m^0$에 대응하는 다섯 종류의 적합도 검정통계량을 계산한다. 이 과정을 10,000번 반복하여 상위 $\alpha(= 0.01, 0.05, 0.10, 0.15)$를 분위수를 구하면 검정 통계량의 기각역으로 설정할 수 있다.

3.2. 이변량 정규분포에서의 검정 방법

상관계수 ρ를 포함하고 있는 이변량 표준정규분포에서 다양한 표본크기 n의 확률표본을 추출해 각각의 통계량들을 구하고 이를 반복해 α수준에 따른 기각역을 정리하였다. $n = 30$이고 $\rho = -0.3, 0, 0.3$인 경우에서의 각 검정통계량의 상위 $\alpha(= 0.01, 0.05, 0.10, 0.15)$를 분위수를 기각역으로 설정하여 Table 3.1에 정리하였다. 각 검정통계량의 첫 번째 행인 critical value는 문헌에서 찾을 수 있는 일변량 적합도 검정통계량의 기각역이다 (D'Agostino와 Stephens, 1986).

경험분포함수를 활용해 구한 다섯 종류의 검정통계량의 기각역은 ρ의 크기와 부호에 따라 큰 차이가 나타나지 않으며, 문헌에서 제시한 일변량 검정통계량의 기각역(각 통계량의 첫 번째 행인 critical values)과 3–5% 정도의 차이를 나타낸다. 이 결과를 바탕으로 3.1절에서 제안한 적합도 검정방법을 위한 다섯 종류의 검정통계량의 기각역을 별도로 구하여 설정하지 않고 어렵지 않게 구할 수 있는 문헌에서 제시한 기각역을 사용해도 큰 무리 없이 검정할 수 있음을 발견하였다. Table 3.1에서는 $n = 30$인 경우

Table 3.1. Critical value for bivariate normal distribution

Test statistic			α			
			0.01	0.05	0.10	0.15
K-S	Critical values*		0.2700	0.2180	0.1900	-
	ρ	-0.3	0.2827	0.2290	0.2033	0.1880
		0.0	0.2782	0.2287	0.2038	0.1886
		0.3	0.2780	0.2280	0.2045	0.1879
V	Critical value*		0.3520	0.3070	0.2850	0.2690
	ρ	-0.3	0.3511	0.3057	0.2838	0.2691
		0.0	0.3475	0.3059	0.2851	0.2711
		0.3	0.3519	0.3084	0.2825	0.2667
W^2	Critical value*		0.7430	0.4610	0.3470	0.2840
	ρ	-0.3	0.7196	0.4444	0.3395	0.2789
		0.0	0.7299	0.4611	0.3559	0.2842
		0.3	0.7358	0.4627	0.3436	0.2793
U^2	Critical value*		0.2680	0.1870	0.1520	0.1310
	ρ	-0.3	0.2663	0.1854	0.1488	0.1307
		0.0	0.2534	0.1858	0.1538	0.1323
		0.3	0.2603	0.1883	0.1479	0.1255
A^2	Critical value*		3.8800	2.4920	1.9330	1.6100
	ρ	-0.3	3.7564	2.4226	1.9307	1.6083
		0.0	3.8342	2.5464	1.9626	1.6172
		0.3	3.7746	2.4709	1.9279	1.5909

* D'Agostino and Stephens (1986) p.105.

만을 제시하였지만, 그 외 $n = 10, 20, 40$ 등에서도 매우 유사한 결과가 발생하였기 때문에 추가하지 않았다.

제 2.1절 마지막에서 설명한 삼변량 표준정규분포에 대하여 다양한 ρ의 변화에 따라 각 통계량의 기각역을 구한 결과 이변량 표준정규분포의 결과와 유사함을 발견하였다. 본 연구에서는 삼변량 정규분포에 대한 설명을 추가하지 않고 일반적인 정규분포가 아닌 특정한 분포함수에 대해 연구를 한다.

3.3. 특정한 다변량분포에서의 검정 방법

제 2.2절에서 정규혼합분포를 가정하고 다변량 경험분포그림에 대하여 토론하였다. 여기에서는 추정한 정규혼합분포의 적합도 검정하기 위하여 귀무가설의 분포함수를 2.2절에서의 $F^0(X_1, X_2, X_3) = 0.5\Phi(x_1, x_2, x_3; \hat{\mu}_1, \hat{\Sigma}_1) + 0.5\Phi(x_1, x_2, x_3; \hat{\mu}_2, \hat{\Sigma}_2)$로 설정하고 $\rho = -0.3, 0, 0.3$인 경우에서의 $n = 30$인 확률표본을 추출해 각 검정통계량의 상위 α분위수를 유의수준α에 대응하는 기각역으로 설정하여 Table 3.2에 정리하였다.

Table 3.2에서도 Table 3.1과 유사하게 다섯 종류의 검정통계량의 기각역은 ρ의 크기와 부호에 따라 큰 차이가 나타나지 않으며, 문헌에서 제시한 일변량 검정통계량의 기각역(Table 3.1에서 각 통계량의 첫 번째 행인 critical values)과 5% 미만의 차이를 나타난다. Table 3.2에서는 $n = 30$인 경우만을 제시하였지만, 그 외 $n = 10, 20, 40$ 등에서도 매우 유사한 결과가 발생하였기 때문에 추가하지 않았다. 3.2와 3.3절의 결과를 바탕으로 3.1절에서 제안한 적합도 검정방법을 위한 다섯 종류의 검정통계량의 기각역을 다변량 분포함수에 따라 별도로 구할 필요없이 기존 문헌에서 제시한 기각역을 사용해도 검정할 수

Table 3.2. Critical value for a certain normal mixture distribution

Test statistic	ρ	α			
		0.01	0.05	0.10	0.15
K-S	−0.3	0.2723	0.2328	0.2058	0.1910
	0.0	0.2765	0.2309	0.2090	0.1944
	0.3	0.2710	0.2242	0.2036	0.1900
V	−0.3	0.3596	0.3113	0.2856	0.2713
	0.0	0.3560	0.3120	0.2880	0.2726
	0.3	0.3510	0.3066	0.2820	0.2690
W^2	−0.3	0.7470	0.4401	0.3466	0.2890
	0.0	0.7588	0.4767	0.3577	0.2930
	0.3	0.6808	0.4407	0.3307	0.2811
U^2	−0.3	0.2702	0.1924	0.1583	0.1318
	0.0	0.2769	0.1979	0.1570	0.1338
	0.3	0.2626	0.1857	0.1523	0.1306
A^2	−0.3	3.9574	2.4427	1.9882	1.6787
	0.0	3.9691	2.5154	1.9965	1.6731
	0.3	3.6936	2.4049	1.8745	1.6159

있음을 발견하였다.

4. 실증예제

4.1. 삼변량 정규분포

2016년 8월 1일부터 9월 29일까지의 40일간의 대신증권, 한화생명, 미래에셋생명의 주식종가 자료를 단순수익률로 변환한 삼변량 자료를 사용하였다. 표준화한 삼변량 자료의 정규성 검정을 실시한 결과, Royston의 정규성 검정 통계량값은 $H = 7.7771$이고 이에 대응하는 p-값은 0.0528로서 정규분포에 적합하며, Mardia의 왜도와 첨도의 통계량값들에 대응하는 p-값들은 모두 0.05보다 훨씬 크므로 정규분포에 적합하다고 판단할 수 있다. 그러나 Henze-Zirkler의 정규성 검정 통계량 값은 HZ = 0.9326이고 이에 대응하는 p-값은 0.0375로서 정규분포에 적합하지 않다. 세 검정 통계량 중에서 두 종류(Royston, Mardia)가 정규분포에 적합하고 나머지 하나는 5%보다 조금 작기 때문에 전체적으로 삼변량 실증예제 자료가 정규분포에 적합하다고 판단할 수 있다. 다음과 같이 표준화된 삼변량 자료의 분산공분산 행렬을 이용하여,

$$\hat{\Sigma} = \begin{pmatrix} 1 & 0.51 & 0.49 \\ 0.51 & 1 & 0.35 \\ 0.49 & 0.35 & 1 \end{pmatrix}.$$

경험분포함수를 구하고 이를 바탕으로 삼변량 경험분포그림을 Figure 4.1에 표현한다. 그리고 동일한 분산공분산 행렬을 갖는 삼변량 표준정규분포에서 수집한 매우 큰 표본에 대한 경험분포그림을 Figure 4.1 왼쪽에 곡선으로 표현한다.

다음으로 자료가 정규분포인지를 검정하는 귀무가설과 대립가설은 다음과 같다.

$$H_0 : F(x_1, x_2, x_3) = \Phi\left(x_1, x_2, x_3; \hat{\Sigma}\right) \quad \text{vs.} \quad H_1 : F(x_1, x_2, x_3) \neq \Phi\left(x_1, x_2, x_3; \hat{\Sigma}\right).$$

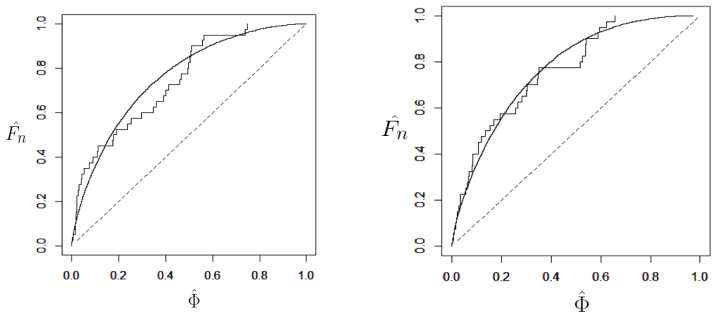

Figure 4.1. Empirical distribution plot.

추정한 정규혼합분포를 이용하여 3절에서 제안한 검정 방법을 이용하여 대신증권, 한화생명, 미래에셋생명의 삼변량 주식종가의 단순수익률 자료에 대한 삼변량 정규분포의 적합도 검정통계량은 각각 다음과 같다.

$$K\text{-}S = 0.1351, \quad V = 0.2489, \quad W^2 = 0.1652, \quad U^2 = 0.1622, \quad A^2 = 0.9847.$$

대신증권, 한화생명, 미래에셋생명의 삼변량 주식종가의 단순수익률 자료에서 계산한 다섯 종류의 적합도 검정통계량은 Table 3.1 critical value에서의 5%유의수준 기각역보다 모두 작기 때문에 귀무가설을 기각할 수 없다. 따라서 주식종가의 단순수익률 자료의 분포함수가 삼변량 정규분포를 따른다고 판단할 수 있다. 그러므로 경험분포함수를 구하기 위하여 실증예제 자료가 정규분포를 따른다고 예측한 분포함수를 바탕으로 경험분포그림을 작성하였고, 이를 바탕으로 3절에서 논의한 적합도 검정을 실시한 결과 예측한 삼변량 정규분포가 실증예제 자료를 잘 설명하는 것으로 판단할 수 있다.

4.2. 특정한 다변량 정규분포

2016년 8월 1일부터 9월 29일까지의 40일 간의 삼성화재, 키움증권의 주식종가 자료를 단순수익률로 변환한 이변량 자료를 사용하였다. 표준화한 이변량 자료의 정규성 검정을 실시한 결과, Henze-Zirkler의 정규성 검정통계량값은 HZ = 1.8262이고 이에 대응하는 p-값은 0.0001이며 그리고 Royston의 정규성 검정 통계량값은 $H = 10.6969$이고 이에 대응하는 p-값은 0.0048으로 정규분포에 적합하지 않다. 또한 Mardia의 왜도 통계량값에 대응하는 p-값은 0.05보다 훨씬 크지만, 첨도 통계량값에 대응하는 p-값은 0.05보다 훨씬 작으므로 정규분포에 적합하지 않기 때문에 세 종류의 적합도 검정방법 모두 삼성화재, 키움증권의 실증예제 자료는 정규분포를 따르지 않는다고 판단할 수 있다.

자료에 적합한 분포를 두 정규분포의 혼합모형으로 가정할 수 있으며, R Package “Mixtool”을 이용하여 두 정규분포의 모수 $\lambda, \mu_1, \mu_2, \Sigma_1, \Sigma_2$를 추정하면 각각 다음과 같다:

$$\hat{\lambda} = 0.50, \quad \hat{\mu}_1 = (0.77, 0.54), \quad \hat{\mu}_2 = (-0.78, 0.54), \quad \hat{\Sigma}_1 = \begin{pmatrix} 0.17 & 0.04 \\ 0.04 & 0.60 \end{pmatrix}, \quad \hat{\Sigma}_2 = \begin{pmatrix} 0.56 & 0.54 \\ 0.54 & 0.77 \end{pmatrix}.$$

모수 추정량을 포함한 혼합 정규분포에 대한 확률표본 $n = 40$의 경험분포함수를 구하고 이를 바탕으로 경험분포그림과 혼합 정규분포에서 수집한 매우 큰 표본 $n = 10{,}000$에 대한 경험분포그림을 Figure 4.1 오른쪽에 계단형태와 곡선으로 각각 표현한다. 삼성화재, 키움증권의 이변량 주식종가의 단순수익률 자료가 두 정규분포의 혼합모형에 적합한 지에 대한 검정을 위한 귀무가설과 대립가설은 다음과 같이 설정

할 수 있다:

$$H_0 : F(x_1, x_2) = F^0(x_1, x_2) \quad \text{vs.} \quad H_1 : F(x_1, x_2) \neq F^0(x_1, x_2),$$

여기서 $F^0(x_1, x_2) = \hat{\lambda}\Phi(x_1, x_2; \hat{\mu}_1, \hat{\Sigma}_1) + (1 - \hat{\lambda})\Phi(x_1, x_2; \hat{\mu}_2, \hat{\Sigma}_2)$로 간주한다.

추정한 정규혼합분포를 이용하여 계산한 적합도 검정통계량 값은 각각 다음과 같다.

$$K\text{-}S = 0.1218, \quad V = 0.2251, \quad W^2 = 0.1069, \quad U^2 = 0.1068, \quad A^2 = 0.5744.$$

실증자료에서 계산한 다섯 종류의 검정통계량이 유의수준이 5%의 기각역보다 모두 작아 귀무가설을 기각할 수 없다. 따라서 삼성화재, 키움증권의 이변량 주식종가의 단순수익률 자료의 분포함수가 두 정규분포의 혼합분포를 따른다고 판단할 수 있다. 그러므로 경험분포함수를 구하기 위하여 실증예제 자료가 혼합 정규분포를 따른다고 예측한 분포함수를 바탕으로 경험분포그림을 작성하였고, 이를 바탕으로 적합도 검정을 실시한 결과 예측한 이변량 혼합 정규분포가 실증예제 자료를 잘 설명하는 것으로 판단할 수 있다.

5. 결론

다변량 자료에서 각 표본의 크기 순서는 일변량과 다르게 쉽게 결정할 수 없다. 그러나 다변량 자료의 공분산행렬을 포함한 자료에 적합한 분포함수를 알고 있거나 추정할 수 있으면 가능하다. 다변량 확률표본의 특정한 분포함수를 추정하면서 Hong 등 (2017)은 다변량 경험분포함수를 정의하고, 특히 이변량 경험분포함수를 시각화하는 계단그림과 분위그림에 대하여 설명하였다.

본 연구에서는 다변량 경험분포함수를 이차원 평면 중에서 단위 1의 정사각형에 표현할 수 있는 다변량 경험분포그림을 제안하였다. 다양한 공분산행렬을 포함하는 이변량 정규분포에 대하여 경험분포그림을 구현하여 살펴본 결과 공분산행렬을 포함된 상관계수에 따라 그림 모양이 다르다는 것을 발견하였다. 정규분포 이외의 특정한 분포에 대하여 살펴보니 다양한 정규혼합분포에 대한 경험분포그림도 분포함수에 따라서도 민감하게 반응하는 것을 탐색하였다.

다변량 분포함수에 따라 경험분포그림의 형태가 다르게 표현되기 때문에 본 연구에서 제안한 다변량 경험분포그림을 통하여 추정한 다변량 분포함수가 자료를 잘 설명하는지 결정할 수 있는 적합도 검정방법을 제안하였다. 잘 알려진 대표적인 다섯 종류의 적합도 검정방법을 사용하고, 이변량 이상 그리고 다양한 분산공분산행렬의 형태를 가진 분포함수들에 대하여 다섯 종류의 검정통계량의 기각역을 각각 구하였다. 다섯 종류의 적합도 검정통계량의 기각역은 문헌에서 구할 수 있는 각각의 기각역과 큰 차이가 없음을 탐색하였다. 그러므로 본 연구에서 제안한 적합도 검정방법은 잘 알려진 다섯 종류의 적합도 검정통계량과 그 기각역을 사용할 수 있는 장점이 있음을 발견하였다.

이변량과 삼변량의 정규분포를 따르는 모의실험과 실증예제에 대하여 살펴보았고 정규분포함수 이외의 특정한 분포함수로서는 다양한 정규혼합분포를 고려하여 정규혼합분포를 따르는 실증예제에 대하여도 경험분포그림을 작성하고 그림의 형태와 특징에 대하여 설명하였다. 그리고 분포에 대응하는 경험분포그림의 특징을 바탕으로 분포함수의 적합도 검정을 할 수 있음을 토론하였다.

References

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes, *The Annals of Mathematical Statistics*, **23**, 193–212.

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit, *Journal of the American Statistical Association*, **49**, 765–769.

Anderson, T. W. (1962). On the distribution of the two-sample Cramér-von Mises criterion, *The Annals of Mathematical Statistics*, **33**, 1148–1159.

D'Agostino, R. B. and Stephens, M. A. (1986). Goodness-of-fit techniques, *Statistics, a Series of Textbooks and Monographs*, **68**, Marcel Dekker Inc., New York.

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data, *Biometrics*, **28**, 81–124.

Gnanadesikan, R., Kettenring, J. R., and Landwehr, J. M. (1977). Interpreting and assessing the results of cluster analyses, *Bulletin of the International Statistical Institute*, **47**, 451–463.

Hong, C. S., Park, J., and Park, Y. H. (2017). Multivariate empirical distribution functions and descriptive methods, *Journal of the Korean Data & Information Science Society*, **28**, 87–98.

Justel, A., Peña, D., and Zamar, R. (1997). A multivariate Kolmogorov-Smirnov test of goodness of fit, *Statistics & Probability Letters*, **35**, 251–259.

Kim, N. H. (2004). An approximate Shapiro-Wilk statistic for testing multivariate normality, *The Korean Journal of Applied Statistics*, **17**, 35–47.

Kim, N. H. (2005). The limit distribution of an invariant test statistic for multivariate normality, *Communications for Statistical Applications and Methods*, **12**, 71–86.

Kim, N. H. (2006). Testing multivariate normality based on EDF statistics, *The Korean Journal of Applied Statistics*, **19**, 241–256.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione, *Giornale dell'Instuto Italiano degli Attuari*, **4**, 83–91.

Koziol, J. A. (1982). A class of invariant procedures for assessing multivariate normality, *Biometrika*, **69**, 423–427.

Kuiper, N. H. (1960). Tests concerning random points on a circle. In *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Series A*, **63**, 38–47.

Malkovich, J. F. and Afifi, A. A. (1973). On tests for multivariate normality, *Journal of the American Statistical Association*, **68**, 176–179.

Meintanis, S. G. and Hlávka, Z. (2010). Goodness-of-fit tests for bivariate and multivariate skew-normal distributions, *Scandinavian Journal of Statistics*, **37**, 701–714.

Moore, D. S. and Stubblebine, J. B. (1981). Chi-square tests for multivariate normality with application to common stock prices, *Communications in Statistics-Theory and Methods*, **10**, 713–738.

Rosenblatt, M. (1952). Remarks on a multivariate transformation, *The Annals of Mathematical Statistics*, **23**, 470–472.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis, *The Annals of Mathematical Statistics*, **24**, 220–238.

Royston, J. P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk W, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **32**, 121–133.

Singh, A. (1993). *Omnibus robust procedures for assessment of multivariate normality and detection of multivariate outliers*, Multivariate environmental statistics, North-Holland, Amsterdam, 445–488.

Smirnov, N. V. (1933). Estimate of deviation between empirical distribution functions in two independent samples, *Bulletin Moscow University*, **2**, 3–16.

Stephens, M. A. (1965). The goodness-of-fit statistic V_n: Distribution and significance points, *Biometrika*, **52**, 309–321.

Thode, H. C. (2002). *Testing for Normality*, Marcel Dekker Inc., New York, **164**.

Watson, G. S. (1961). Goodness-of-fit tests on a circle, *Biometrika*, **48**, 109–114.

Zhu, L. X., Fang, K. T., and Bhatti, M. I. (1997). On estimated projection pursuit-type cramer-von mises statistics, *Journal of Multivariate Analysis*, **63**, 1–14.

다변량 경험분포그림과 적합도 검정

홍종선[a,1] · 박용호[a] · 박준[a]

[a]성균관대학교 통계학과

(2017년 6월 14일 접수, 2017년 7월 12일 수정, 2017년 7월 26일 채택)

요 약

다변량 자료의 분포함수를 알고 있거나 추정할 수 있으면 다변량 경험분포함수를 정의할 수 있다. 이변량인 경우에는 계단그림과 분위그림을 사용하여 경험분포함수를 시각화할 수 있는데, 본 연구에서는 다변량인 경우에 경험분포함수를 정사각형에 표현할 수 있는 다변량 경험분포그림을 제안하였다. 여러 종류의 다변량 정규분포와 특정한 분포에 대하여 경험분포그림을 작성하고 특징을 살펴보니, 다양한 분산공분산행렬을 포함된 분포함수에 따라 경험분포그림이 민감하게 반응하는 것을 탐색하였다. 이를 바탕으로 경험분포함수를 구할 때 가정한 다변량 분포함수의 적합도 검정방법을 제안하였다. 대표적인 다섯 종류의 적합도 검정방법을 사용하고, 다양한 분포함수들에 대하여 각각의 검정통계량 기각역을 구하였다. 본 연구에서 얻은 기각역은 문헌에서 구할 수 있는 기각역과 큰 차이가 없음을 발견하였다. 그러므로 본 연구에서 제안한 적합도 검정방법을 문헌에서 제시한 기각역으로 쉽게 사용할 수 있는 장점이 있다.

주요용어: 경험분포그림, 분위벡터, 정규혼합

[1]교신저자: (03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: cshong@skku.edu