The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.2.355
29(2), 355–368

Firework plot as a graphical exploratory data analysis tool for evaluating the impact of outliers in skewness and kurtosis of univariate data

Sungho Moon[a,1]

[a]Department of Data Management, Busan University of Foreign Studies

(Received January 18, 2016; Revised January 26, 2016; Accepted January 28, 2016)

Abstract

Outliers and influential data points distort many data analysis measures. Jang and Anderson-Cook (2014) proposed a graphical method called a firework plot for exploratory analysis purpose so that there could be a possible visualization of the trace of the impact of the possible outlying and/or influential data points on the univariate/bivariate data analysis and regression. They developed 3-D plot as well as pairwise plot for the appropriate measures of interest. This paper further extends their approach to identify its strength. We can use firework plots as a graphical exploratory data analysis tool to evaluate the impact of outliers in skewness and kurtosis of univariate data.

Keywords: outliers, influential data point, skewness, kurtosis, (3-D) firework plot, firework plot matrix

1. 서론

특이값(특이점, 이상점이라고도 함) 및 영향점(영향력이 큰 관측값이라고도 함)은 자료분석을 하는 데 사용되는 계량적이고 기술적인 많은 측도들(measures)을 왜곡한다. 우리는 민감도 곡선(sensitivity curve)이나 영향력 함수(influence function)를 사용하여 이러한 특이값을 탐지할 수 있다 (Hampel, 1974; Maronna 등, 2006). 특이점 및 영향점의 평가는 주로 회귀분석에서 다루어졌다 (Chatterjee와 Hadi, 2012). 회귀진단에 관련된 많은 참고문헌 중에서 Beckman과 Cook (1983)의 논문은 이상점에 관한 전반적인 정보를 담고 있다. 영향점과 관련해서는 Cook의 거리통계량(Cook's Distance measure; Cool, 1977, 1979)을 위시해서 Belsley 등 (1980)의 문헌들이 대표적으로 존재한다. 그리고 요약된 숫자 통계량 뿐 아니라 이를 활용한 다양한 도시적인 방법이 존재 하는데 레버리지에 대한 인덱스 그림(index plot), Cook의 거리 통계량에 대한 인덱스 그림 뿐 아니라 Emerson과 Strenio (1983)의 산포-수준 그림(spread-level plot), 그리고 Fox (2008)의 회귀 영향그림(regression influence plot) 등이 문헌에 존재한다.

Jang과 Anderson-Cook (2014)은 불꽃그림이란 이름을 붙인 그림도구를 발표하였는데 이상점이나 영향점이 일변량/이변량 자료분석 및 회귀분석에 어떠한 영향을 미치는지 알기 위하여 3-D 불꽃그림 및

This work was supported by the research grant of Busan University of Foreign Studies in 2015.
[1]Department of Data Management, Busan University of Foreign Studies, Busan 46234, Korea.
E-mail: shmoon@bufs.ac.kr

짝진 불꽃그림 행렬을 제시하였다. 관측값에 부여된 가중치를 1에서 0으로 변화함에 따라 이상점이 일변량/이변량 자료분석시의 여러 수치적 측도들에 어떠한 영향을 미치는 지 3-D 불꽃그림 및 불꽃그림 행렬을 통하여 살펴보았다. 또한, 이상점이나 영향점이 회귀계수 및 잔차제곱합(SSE)에 어떠한 영향을 미치는지 3차원 그림에 추적곡선을 그려 보았을 뿐 아니라 쌍으로 대비시켜 봄으로써 분석의 시각적인 효과를 증대시켰다. 본 연구에서는 이러한 불꽃그림이 일변량 자료의 왜도와 첨도에서 특이점의 영향을 평가하기 위한 탐색적 자료분석 그림도구로 사용될 수 있음을 보이고자 한다. 제 2절은 일변량 자료의 왜도와 첨도에서 특이점의 영향을 평가하기 위한 탐색적 자료분석 그림도구로서의 불꽃그림을 제시하고 컴퓨터 시뮬레이션을 시행하여 불꽃그림의 유용성을 살펴보았다. 제 3절은 실제 사례를 통하여 불꽃그림의 유용성을 살펴보았다. 제 4절에서 결론으로 마무리하였다.

2. 특이값의 영향을 평가하기 위한 탐색적 자료분석 그림도구로서의 불꽃그림

일변량 데이터가 주어졌을 때 이 데이터를 통하여 분포의 대칭성 및 꼬리의 두꺼운 정도를 측정하기 위하여 우리는 왜도와 첨도를 사용한다. 일변량 데이터 $x_1, x_2, \ldots, x_n$이 주어졌을 때 이 데이터에 대한 표본왜도(sample skewness)와 표본첨도(sample kurtosis)를 다음과 같이 각각 정의한다.

$$b_1 = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^3}{n}, \qquad g_2 = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^4}{n} - 3,$$

여기서 $\bar{x}$는 표본평균이고, s는 표본표준편차이다. 우리는 이러한 표본왜도 및 표본첨도를 통하여 분포의 대칭성 및 꼬리의 두꺼운 정도를 측정할 수 있다.

특이값은 자료분석을 하는 데 사용되는 계량적이고 기술적인 많은 측도들을 왜곡한다. 표본왜도나 표본첨도는 데이터의 표본적률임으로 특이값에 매우 민감하다. 즉 특이값은 분포의 대칭성을 심하게 왜곡할 수 있고 분포의 꼬리를 두껍게 한다. 이러한 특이값이 표본왜도나 표본첨도에 어떤 영향을 주는지를 평가하기 위하여 우리는 불꽃그림을 사용할 수 있다. i번째 자료에 부여된 가중치의 값 $w_i\,(i = 1, 2, \ldots, n)$을 1에서 0으로 변화하게 하고, 나머지 가중치인 경우 $w_j = 1\,(j \neq i, j = 1, 2, \ldots, n)$로 고정한 다음 표본왜도 및 표본첨도에 어떠한 영향이 있는지 연속적으로 그 변화를 추적하고 이를 모든 관측값에 적용, 상응하는 그림을 시도할 수 있다. 이를 위하여 우리는 다음과 같은 가중표본왜도(weighted sample skewness)와 가중표본첨도(weighted sample kurtosis)를 다음과 같이 각각 정의한다.

$$b_{1w} = \frac{\sum_{i=1}^{n} w_i \left(\frac{x_i - \bar{x}_w}{s_w}\right)^3}{\sum_{i=1}^{n} w_i}, \qquad g_{w2} = \frac{\sum_{i=1}^{n} w_i \left(\frac{x_i - \bar{x}_w}{s_w}\right)^4}{\sum_{i=1}^{n} w_i} - 3,$$

여기서 $\bar{x} = \left[\sum_{i=1}^{n} w_i x_i\right] / \sum_{i=1}^{n} w_i$는 가중표본평균이고, $s_w = \sqrt{\left[\sum_{i=1}^{n} w_i \left(x_i - \bar{x}_w\right)^2\right] / \sum_{i=1}^{n} w_i}$는 가중표본표준편차이다.

가중표본왜도-가중표본첨도 불꽃그림에서는 i번째 데이터에 부여된 가중치의 값 $w_i\,(i = 1, 2, \ldots, n)$을 1에서 0으로 연속적으로 변화하게 하고, 나머지 가중치인 경우 $w_j = 1\,(j \neq i, j = 1, 2, \ldots, n)$로 고정한 다음 가중표본왜도 및 가중표본첨도를 계산하여 i번째 데이터에 대응하는 그림을 그린다. 이를 모든 관측값에 적용하여 그림을 완성한다. 예로, 첫 번째 데이터에 대한 불꽃그림을 그리기 위해서는 데이터셋에서 두 번째 데이터부터 n번째 데이터는 그대로 두고 첫 번째 데이터만 가중치의 값을 1에서 0으로 연속적으로 변화시켜 가며 가중표본왜도 및 가중표본첨도를 계산한 후 이렇게 계산된 가중표본왜도 및 가중표본첨도를 산점도에 나타내면 첫 번째 데이터에 대한 불꽃그림이 완성된다.

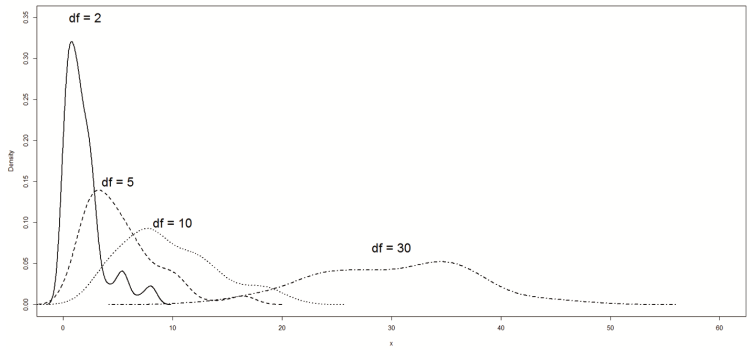

Figure 2.1. Kernel density estimators of 100 chi-squared random variates in case of 2, 5, 10 and 30 degree of freedoms.

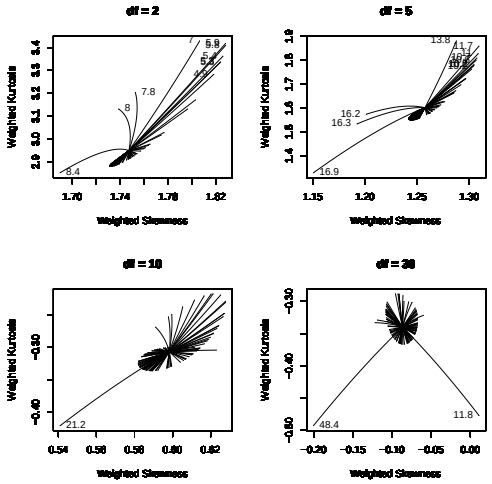

Figure 2.2. Skewness-kurtosis firework plots of 100 chi-squared random variates in case of 2, 5, 10 and 30 degree of freedoms, respectively.

이러한 불꽃그림의 유용성을 알아보기 위하여 먼저 카이제곱분포와 정규분포를 비교하여 보자. 카이제곱분포는 자유도가 작을 때는 양의 방향으로 치우친 분포이어서 왜도가 0보다 크게 되고 자유도가 커짐에 따라 점점 대칭분포가 된다. 자유도가 2, 5, 10, 30인 경우 각각 카이제곱난수 100개를 생성하여 커널밀도추정량을 함께 그려보면 다음 Figure 2.1과 같다.

Figure 2.2는 자유도가 2, 5, 10, 30에 대응되는 카이제곱난수 100개에 대한 각각의 왜도-첨도 불꽃그림이다. 카이제곱분포에서는 분포의 평균이 자유도가 된다. 각 불꽃그림에서 숫자들은 데이터 중 분포의 평균보다 상대적으로 크거나 작은 수치값들을 나타낸다. 각각의 불꽃그림을 보면 데이터가 평균보다 매우 큰 값이나 매우 작은값이 왜도나 첨도에 상대적으로 더 큰 영향을 줌을 알 수 있다. 그러나, 이러한 값들도 정규분포의 왜도와 첨도값 0과 상대적으로 비교하면 왜도와 첨도에 매우 큰 변화를 일으킨다고는 할 수 없다.

Figure 2.3은 Figure 2.2에 나타난 불꽃그림 4개를 하나의 그림으로 모아 나타낸 그림이다. 자유도가 2,

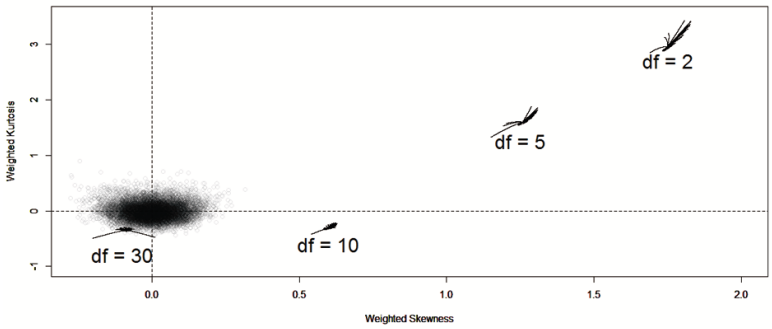

Figure 2.3. Combined skewness-kurtosis firework plots of 100 chi-squared random variates in case of 2, 5, 10 and 30 degree of freedoms, respectively.

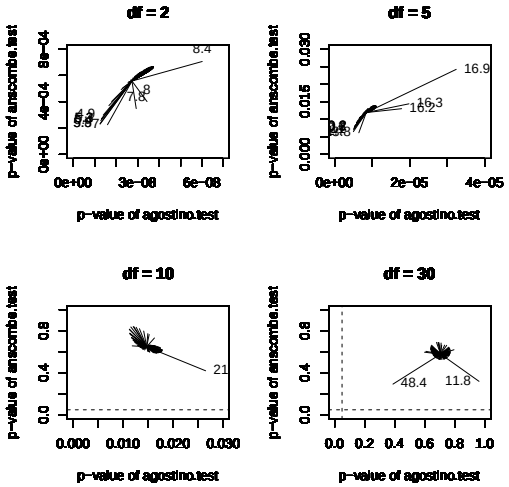

Figure 2.4. Firework plots for skewness-kurtosis test of 100 chi-squared random variates in case of 2, 5, 10 and 30 degree of freedoms, respectively.

5, 10, 30에 대응되는 각각의 불꽃그림을 보면 데이터가 평균보다 매우 큰 값이나 매우 작은값이 왜도나 첨도에 상대적으로 더 큰 영향을 주나, 이러한 값들도 정규분포의 왜도와 첨도값 0과 상대적으로 비교하면 왜도와 첨도에 매우 큰 변화를 일으킨다고는 할 수 없음을 알 수 있다. 자유도가 커지면서 점점 왜도와 첨도가 0에 가까워짐을 볼 수 있다. 즉, 자유도가 커지면서 카이제곱분포의 정규근사화를 확인할 수 있다. 원점 주위의 10,000개의 점들은 각각 정규난수 1,000개를 이용하여 구한 왜도와 첨도값이다. 정규분포는 특이값이 매우 드물게 나타남을 알 수 있다. Figure 2.3에서 자유도가 30일 때 카이제곱난수의 왜도와 첨도값이 이 10,000개의 점들 집합 변두리에 인접해 있음을 확인할 수 있다. 이 시뮬레이션 결과 10,000개의 왜도와 첨도에 대한 하위 2.5%와 상위 2.5%를 제외시켰을 때의 범위가 왜도는 $[-0.15, 0.15]$, 첨도는 $[-0.27, 0.33]$이었다. 탐색적 자료분석 입장에서 어떤 데이터셋이 정규분포와 같은 왜도 및 첨도를 가지고 있는지를 판단하기 위한 임계값으로 대략 왜도는 절대값으로 0.15, 첨도는 절대값으로 0.3을 사용하여 볼 수 있을 것이다.

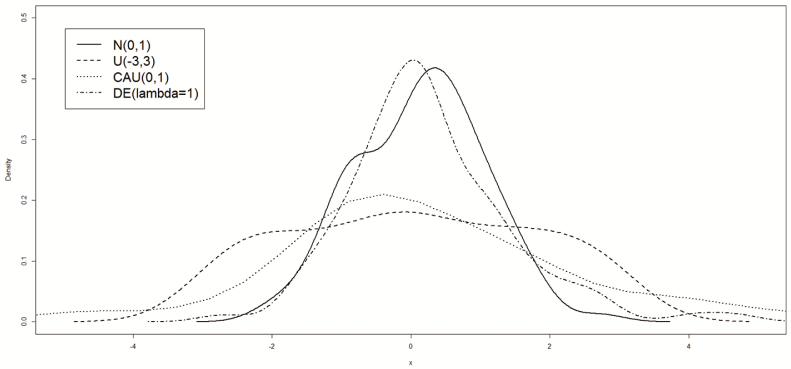

Figure 2.5. Kernel density estimators of 100 random variates in case of normal, uniform, Cauchy and double exponential distribution.

자유도가 2, 5, 10, 30에 대응되는 카이제곱난수 100개에 대하여 D'Agostino 왜도 검정과 Anscombe-Glynn 첨도 검정 p-값에 대한 영향을 파악하기 위하여 불꽃그림을 그려보면 Figure 2.4와 같다. 이 불꽃그림에서는 i번째 데이터에 부여된 가중치의 값 $w_i\,(i = 1, 2, \ldots, n)$을 1($i$번째 데이터 삽입)과 0($i$번째 데이터 제거) 두 가지 경우만 고려하고, 나머지 가중치인 경우 $w_j = 1\,(j \neq i, j = 1, 2, \ldots, n)$로 고정한 다음 D'Agostino 왜도 검정과 Anscombe-Glynn 첨도 검정 p-값을 계산하여 i번째 데이터에 대응하는 그림을 그린다. 이를 모든 관측값에 적용하여 그림을 완성한다. 이 불꽃그림에서 우리가 알 수 있는 사실은 Figure 2.2나 Figure 2.3에서처럼 자유도가 2, 5, 10, 30에 대응되는 각각의 불꽃그림을 보면 데이터가 평균보다 매우 큰 값이나 매우 작은값이 왜도나 첨도에 대한 검정의 p-값에 상대적으로 더 큰 영향을 주나, 이러한 값들도 정규성 검정의 입장에서 보면 왜도와 첨도에 대한 검정 결과를 바꿀 정도로 왜도와 첨도에 매우 큰 변화를 일으킨다고는 할 수 없음을 알 수 있다. 자유도가 커지면서 점점 왜도와 첨도가 0에 가까워짐에 따라 왜도와 첨도에 대한 검정 결과 p-값도 0.05보다 커짐을 볼 수 있다. 즉, 자유도가 커지면서 카이제곱분포의 정규근사화를 확인할 수 있다.

코시분포나 이중지수분포는 꼬리가 정규분포보다 두꺼워 특이값이 자주 발생하는 분포이어서 첨도가 0보다 크게 된다. 반면 연속적 균등분포는 꼬리가 없어 특이값이 나타나지 않는 분포이어서 첨도가 0보다 작게 된다. 다음 Figure 2.5는 표준정규난수 $N(0, 1)$, 연속적 균등난수 $U(-3, 3)$, 코시난수 CAU$(0, 1)$, 이중지수난수 DE$(\lambda = 1)$를 각각 100개씩 생성시켜 구한 커널밀도추정량들이다.

Figure 2.6은 4개의 난수 데이터셋에 대응되는 각각의 왜도-첨도 불꽃그림이다. 각 불꽃그림에서 숫자들은 데이터 중 절대값이 상대적으로 큰 수치값들을 나타낸다. 정규분포나 균등난수에서는 두드러진 특이값이 없어 왜도나 첨도에 미치는 영향이 미미하나 코시난수에서는 3개의 난수가 왜도나 첨도에 미치는 영향이 크고 특히 수치값 221.6은 왜도나 첨도에 큰 영향을 미치고 가중값에 따라 특이한 패턴을 이룸을 알 수 있다. 이중지수난수에서도 2∼3개의 데이터가 왜도나 첨도에 미치는 영향이 큼을 알 수 있다.

4개의 난수 데이터셋에 대하여 D'Agostino 왜도 검정과 Anscombe-Glynn 첨도 검정 p-값에 대한 영향을 파악하기 위하여 불꽃그림을 그려보면 Figure 2.7과 같다. 이 불꽃그림에서 우리가 알 수 있는 사실은 Figure 2.6에서처럼 코시난수와 이중지수 난수에 대응되는 각각의 불꽃그림을 보면 데이터의 절대값이 매우 큰 값이 왜도나 첨도에 대한 검정의 p-값에 상대적으로 더 큰 영향을 주나, 이러한 값들도 정규성 검정의 입장에서 보면 왜도와 첨도에 대한 검정 결과를 바꿀 정도로 왜도와 첨도에 매우 큰 변화를

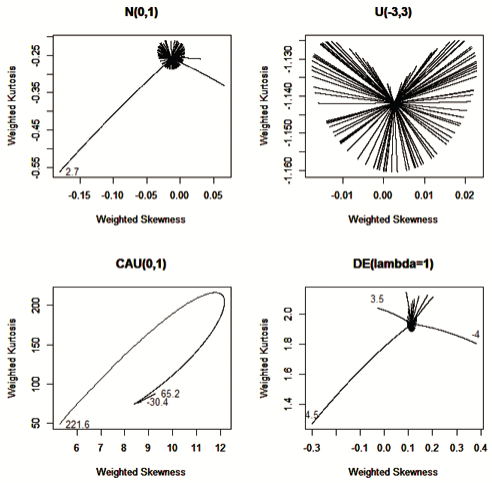

Figure 2.6. Skewness-kurtosis firework plots of 100 random variates in case of 4 distributions, respectively.

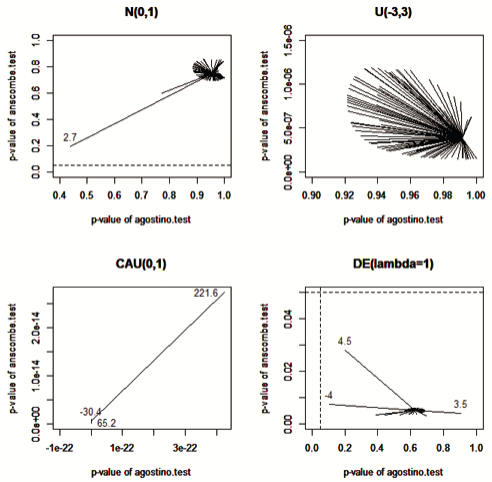

Figure 2.7. Firework plots for skewness-kurtosis test of 100 random variates in case of 4 distributions, respectively.

일으킨다고는 할 수 없음을 알 수 있다.

우리는 왜도 및 첨도를 평균, 분산, 표준편차 같은 다른 측도들과 같이 고려하여 3차원 불꽃 그림이나 불꽃그림 행렬을 작성할 수 있다. 100개의 또 다른 코시난수를 생성시켜 이 데이터셋에 대하여 상자그림 및 커널밀도함수추정량을 그려 보니 다음 Figure 2.8과 같았다. 다수의 특이값들이 존재함을 알 수 있다.

Figure 2.9는 100개의 코시난수를 이용하여 구한 3-D 표준편차-왜도-첨도 불꽃그림이다. 이 3-D 불꽃그림을 통하여 각 데이터에 대하여 표준편차, 왜도 및 첨도에 어떠한 영향을 주는지를 파악할 수 있다. 5개의 데이터 $(-54.8, -25.4, 31.1, 32.1, 45.6)$가 표준편차, 왜도 및 첨도에 영향을 줌을 알 수 있다. 특

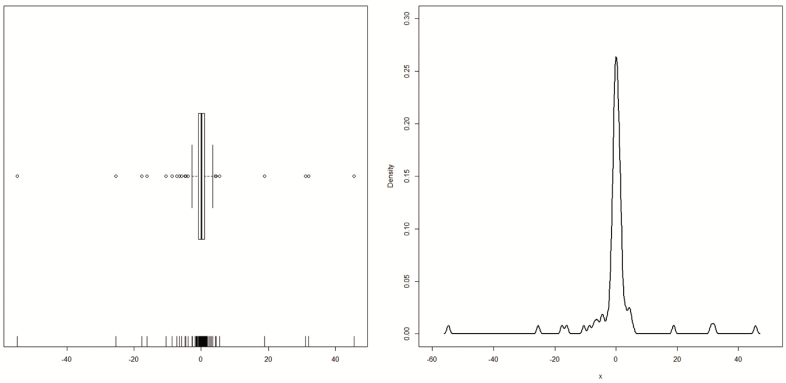

Figure 2.8. Box plot and Kernel density estimator for Cauchy random variates.

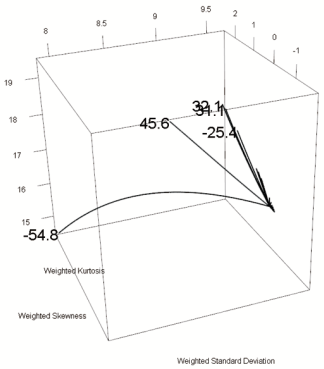

Figure 2.9. 3-D standard deviation-skewness-kurtosis firework plot for Cauchy random variates.

히 2개의 데이터 −54.8과 45.6은 표준편차에 대한 영향 패턴은 비슷한 경향 (감소시키는 패턴)이 있는 반면 데이터 −54.8의 삭제는 왜도의 증가, 첨도의 감소를 일으키나 데이터 45.6의 삭제는 왜도의 감소, 첨도의 증가를 일으켜 영향을 주는 패턴이 데이터 −54.8와 다르다.

다음 Figure 2.10은 코시난수를 이용하여 구한 불꽃그림 행렬이다. 이 불꽃그림 행렬을 통하여 각 데이터에 대하여 평균, 분산, 왜도 및 첨도에 어떠한 영향을 주는지를 파악할 수 있다. 전체적으로 2개의 데이터 $(-54.8, 45.6)$가 평균, 분산, 왜도 및 첨도에 영향을 줌을 알 수 있다. 데이터 −54.8과 45.6은 표준편차에 대한 영향 패턴은 비슷한 경향 (감소시키는 패턴)이 있는 반면 데이터 −54.8의 삭제는 왜도의 증가, 첨도의 감소를 일으키나 데이터 45.6의 삭제는 왜도의 감소, 첨도의 증가를 일으켜 영향을 주는 패턴이 데이터 −54.8와 다르다.

코시분포와 비교하기 위하여 표준정규분포에서 난수 100개를 생성시켜 보았다. 다음 Figure 2.11은 표준정규난수를 이용하여 구한 3-D 표준편차-왜도-첨도 불꽃그림이다. 이 3-D 불꽃그림을 통하여 각 데이터에 대하여 표준편차, 왜도 및 첨도에 어떠한 영향을 주는지를 파악할 수 있다. 어떤 데이터도 표준편차, 왜도 및 첨도에 큰 영향을 주지 못함을 알 수 있다. 모든 난수값들을 대상으로 각각 가중값을 1에서 0으로 서서히 바꾸어 가며 표준편차, 왜도 및 첨도가 어떻게 달라지는지를 살펴보면 거의 이들 측도값들에 변화가 거의 없음을 알 수 있다.

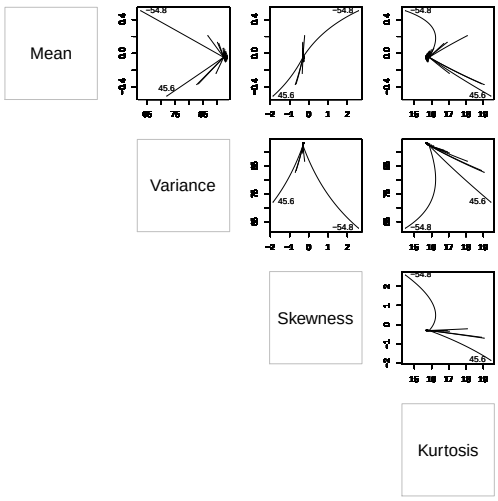

Figure 2.10. Firework plot matrix for Cauchy random variates.

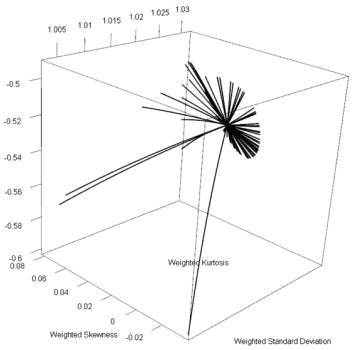

Figure 2.11. 3-D standard deviation-skewness-kurtosis firework plot for normal random variates.

다음 Figure 2.12는 표준정규난수를 이용하여 구한 불꽃그림 행렬이다. 이 불꽃그림 행렬을 통하여 각 데이터에 대하여 평균, 분산, 왜도 및 첨도에 어떠한 영향을 주는지를 파악할 수 있다. 어떤 데이터도 평균, 분산, 왜도 및 첨도에 큰 영향을 주지 못 함을 알 수 있다. 모든 난수값들을 대상으로 각각 가중값을 1에서 0으로 서서히 바꾸어 가며 표준편차, 왜도 및 첨도가 어떻게 달라지는지를 살펴보면 거의 이들 측도값들에 변화가 거의 없음을 알 수 있다.

3. 실제 사례

실제 사례를 통하여 불꽃그림의 유용성을 살펴보자. Forbes지(www.forbes.com)가 평가한 2014년 미국 MLB 30개 프로야구팀들의 경영평가 데이터는 7개의 변수로 구성되어 있다 (1. 순위(rank based on current value), 2. 팀이름(team name), 3. 현재구단가치(current value ($mil)), 4. 1년구단가치변화율(one-year value change (%)), 5. 가치대비부채비율(debt/value (%)), 6. 수입(revenue ($mil)), 7. 운영수익(operating income ($mil))).

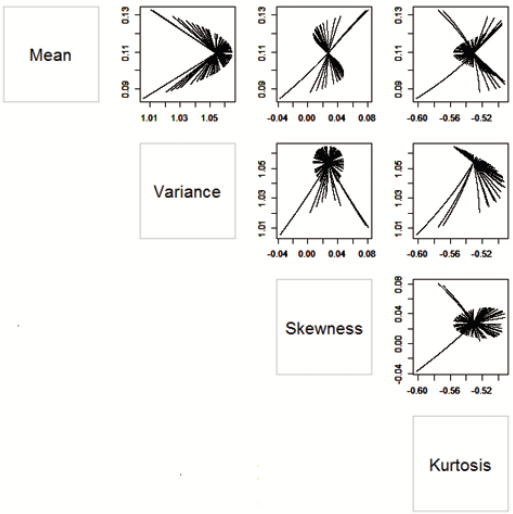

Figure 2.12. Firework plot matrix for normal random variates.

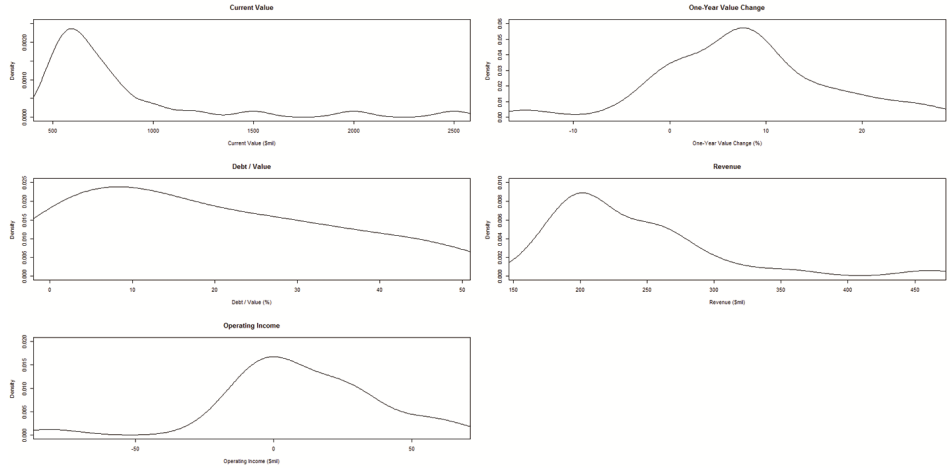

Figure 3.1. Kernel density estimators for 5 variables of 2014 MLB team values.

5개의 변수(현재구단가치, 1년구단가치변화율, 가치대비부채비율, 수입, 운영수익) 각각에 대하여 구한 커널밀도추정량을 그려보면 다음 Figure 3.1과 같다. '1년구단가치변화율' 변수값이 좌우대칭형에 가까움을 알 수 있다.

다음 Figure 3.2는 5개의 변수(현재구단가치, 1년구단가치변화율, 가치대비부채비율, 수입, 운영수익) 각각에 대하여 구한 가중표본왜도-가중표본첨도 불꽃그림들이다. 특이값이 왜도나 첨도에 영향을 주는 변수는 '현재구단가치', '1년구단가치변화율', '수입', '운영수익'인 반면 '가치대비부채비율' 변수에서는 왜도나 첨도에 영향을 주는 데이터가 없다. 특이값은 순위 1인 뉴욕 양키스, 순위 2인 LA 다저스, 순위 3인 보스톤 레드삭스, 순위 26인 휴스턴 애스트로스 네 팀임을 알 수 있다. 세 팀은 순위 1, 2, 3위인 팀인 데 반하여 휴스턴 애스트로스는 순위가 매우 낮은 26위인 팀이라 이채롭다.

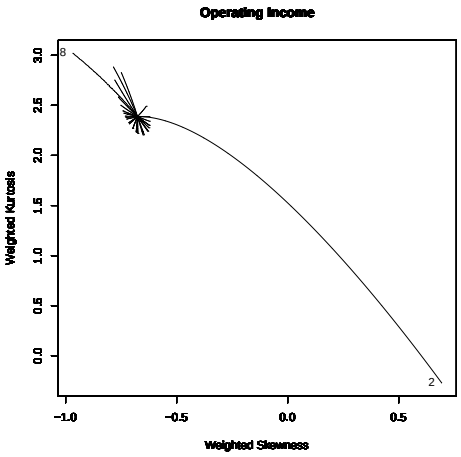

Figure 3.2. Skewness-kurtosis firework plots for 5 variables of 2014 MLB team values (The numbers identify the MLB team ranking).

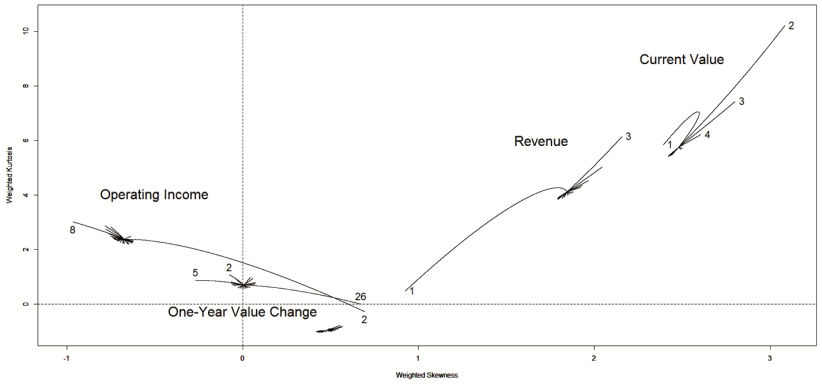

Figure 3.3. Combined skewness-kurtosis firework plots for 5 variables in 2014 MLB team values (The numbers identify the MLB team ranking).

다음 Figure 3.3은 Figure 3.1에 나타난 불꽃그림 5개를 하나의 그림으로 모아 나타낸 그림이다. 특이값이 왜도나 첨도에 영향을 주는 변수는 '현재구단가치', '1년구단가치변화율', '수입', '운영수익'인 반면 '가치대비부채비율' 변수에서는 왜도나 첨도에 영향을 주는 데이터가 없음을 한 눈에 알 수 있다. Figure 3.3 하단의 아주 작은 왜도-첨도 불꽃그림이 '가치대비부채비율' 변수에 대응되는 불꽃그림이다. 순위 1인 뉴욕 양키스 삭제 시 '수입' 변수에서 왜도와 첨도가 동시에 작아진다. 순위 2인 LA 다저스 삭제 시 '현재구단가치' 변수에서 왜도와 첨도가 동시에 커지는 반면 '운영수익' 변수에서는 왜도는 커지나 첨도는 작아진다, 순위 3인 보스톤 레드삭스 삭제 시 '현재구단가치' 변수와 '운영수익' 변수에서 모두 왜도와 첨도가 동시에 커진다. 순위 26인 휴스턴 애스트로스 삭제 시 '1년구단가치변화율' 변수에서 왜도가 커진다. Table 3.1을 보면 휴스턴 애스트로스가 1년구단가치변화율 −15%로 가장 작다.

다음 Figure 3.4는 '현재구단가치' 변수를 대상으로 구한 3-D 표준편차-왜도-첨도 불꽃그림이다. 이 3-D 불꽃그림을 통하여 각 팀에 대하여 표준편차, 왜도 및 첨도에 어떠한 영향을 주는지를 파악할 수 있

Table 3.1. MLB management evaluation data of top 30 teams in 2014.

Rank	Team	Current Value ($mil)	1-Yr Value Change (%)	Debt Value (%)	Revenue ($mil)	Operating Income ($mil)
1	New York Yankees	2,500	9	1	461	−9.1
2	Los Angeles Dodgers	2,000	24	20	293	−80.9
3	Boston Red Sox	1,500	14	0	357	25.3
4	Chicago Cubs	1,200	20	35	266	27.3
5	San Francisco Giants	1,000	27	9	316	53.3
6	Philadelphia Phillies	975	9	10	265	−20.9
7	Texas Rangers	825	8	20	257	−4.9
8	St Louis Cardinals	820	15	35	283	65.2
9	New York Mets	800	−1	44	238	1.6
10	Los Angeles Angels of Anaheim	775	8	4	253	5.8
11	Atlanta Braves	730	16	0	253	38.4
12	Seattle Mariners	710	10	2	210	5.3
13	Washington Nationals	700	11	49	244	22.4
14	Chicago White Sox	695	0	1	210	−2.7
15	Detroit Tigers	680	6	26	262	7.5
16	Baltimore Orioles	620	0	24	198	1.6
17	San Diego Padres	615	2	31	207	33.0
18	Toronto Blue Jays	610	7	0	218	−14.9
19	Minnesota Twins	605	5	37	221	30.2
20	Cincinnati Reds	600	10	7	209	−11.6
21	Arizona Diamondbacks	585	0	25	192	−5.8
22	Colorado Rockies	575	7	11	197	13.7
23	Pittsburgh Pirates	572	19	16	204	21.8
24	Cleveland Indians	570	2	15	196	−1.9
25	Milwaukee Brewers	565	1	9	197	6.8
26	Houston Astros	530	−15	49	186	55.9
27	Miami Marlins	500	−4	44	159	−8.0
28	Oakland Athletics	495	6	13	187	27.4
29	Kansas City Royals	490	7	11	178	−6.5
30	Tampa Bay Rays	485	8	28	181	15.3

다. 특이값은 순위 1인 뉴욕 양키스, 순위 2인 LA 다저스임을 알 수 있고 영향 패턴이 상이함을 알 수 있다. 즉 표준편차에 대한 영향 패턴은 비슷한 경향(감소시키는 패턴)이 있으나 순위 1인 뉴욕 양키스의 삭제는 왜도는 약간 작게 되나 첨도는 거의 변화가 없다. 반면 순위 2인 LA 다저스 삭제는 왜도와 첨도가 오히려 커지는 영향을 준다.

다음 Figure 3.5는 '현재구단가치' 변수를 대상으로 구한 불꽃그림 행렬이다. 이 불꽃그림 행렬을 통하여 각 팀에 대하여 평균, 분산, 왜도 및 첨도에 어떠한 영향을 주는지를 파악할 수 있다. 특이값은 순위 1인 뉴욕 양키스, 순위 2인 LA 다저스임을 알 수 있고 평균과 분산에 대한 영향 패턴은 비슷한 경향 (감소시키는 패턴)이 있으나 왜도와 첨도에 대한 영향패턴은 상이함을 알 수 있다. 즉 순위 1인 뉴욕 양키스의 삭제는 왜도는 약간 작게 되나 첨도는 거의 변화가 없다. 반면 순위 2인 LA 다저스 삭제는 왜도와 첨도가 오히려 커지는 영향을 준다.

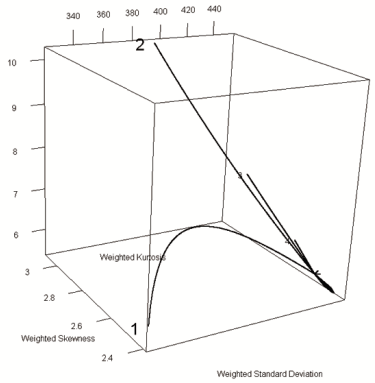

Figure 3.4. 3-D standard deviation-skewness-kurtosis firework plot for 'current value' of 2014 MLB team values (The numbers identify the MLB team ranking).

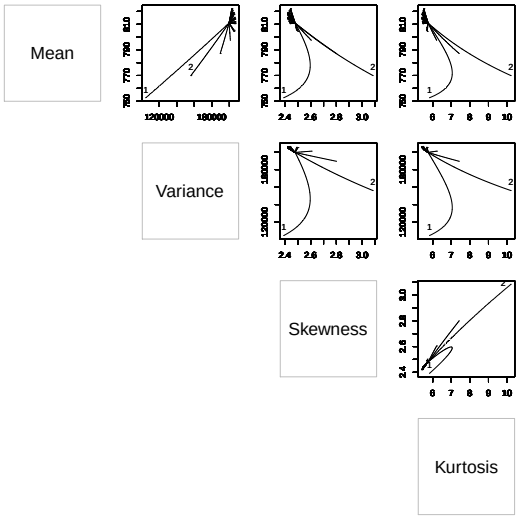

Figure 3.5. Firework plot matrix for 'current value' of 2014 MLB team values (The numbers identify the MLB team ranking).

4. 결론

왜도나 첨도는 분포의 대칭성 및 꼬리의 두꺼운 정도를 파악하는 데 유용한 측도들이다. 컴퓨터 시뮬레이션을 통하여 비대칭 분포인 카이제곱분포를 대상으로 자유도가 커져감에 따라 왜도 및 첨도에 어떤 변화가 있는지, 난수값들이 왜도 및 첨도에 어떤 영향을 주는지를 왜도-첨도 불꽃그림을 통하여 살펴보았고 이 왜도-첨도 불꽃그림을 사용하여 꼬리가 없는 균등분포와 특이값이 자주 나타나는 코시분포 및 이중지수분포를 정규분포와 비교하여 보았다. 미국프로야구 30개 팀들에 대한 경영평가자료를 통하여 불꽃그림이 일변량 자료의 왜도와 첨도에서 특이점의 영향을 평가하기 위한 탐색적 자료분석 그림도구로 사용될 수 있음을 살펴보았다. 이 그래픽 방법은 다양한 자료를 대상으로 각 관측값이 왜도 및 첨도에

어떤 영향을 주는지를 평가할 수 있는 탐색적 자료분석 그림도구로서 유용하게 쓰일 수 있다. 왜도-첨도 불꽃그림 뿐만이 아니라 다른 수치적 측도, 예를 들어 평균, 분산, 표준편차 등을 동시에 사용하여 삼차원 불꽃그림 및 불꽃그림행렬을 작성할 수 있다.

References

Beckman, R. J. and Cook, R. D. (1983). Outlier....s, *Technometrics*, **25**, 119–147.

Belsley, D. A., Kuh, E., Welch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Source of Collinearity*, Wiley, New York.

Chatterjee, S. and Hadi, A. S. (2012). *Regression Analysis by Example*, 5th ed, Wiley, Hoboken.

Cook, R. D. (1977). Detection of influential observation in linear regression, *Technometrics*, **19**, 15–18.

Cook, R. D. (1979). Influential observation in linear regression, *Journal of American Statistical Association*, **74**, 169–174.

Emerson, J. D. and Strenio, J. (1983). The Spread-versus-Level plot in Hoaglin, D. C., Mosteller, F., and Tukey, J. W.(Eds.) (1983). *Understanding Robust and Exploratory Data Analysis*, Wiley, New York.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*, 2nd ed., Sage, New York.

Hampel, F. R. (1974). The influence curve and its role in robustness, *The Annal of Statistics*, **45**, 383–393.

Jang, D. H. and Anderson-Cook, C. M. (2014). Firework plot as a graphical exploratory data analysis tool for evaluating the impact of outliers in data exploration and regression, *Quality and Reliability Engineering International*, **30**, 1409–1425.

Maronna, R. A., Martin, D., and Yohai, V. J. (2006). *Robust Statistics*, John Wiley & Sons, New York.

일변량 자료의 왜도와 첨도에서 특이점의 영향을 평가하기 위한 탐색적 자료분석 그림도구로서의 불꽃그림

문승호[a,1]

[a]부산외국어대학교 데이터경영학과

(2016년 1월 18일 접수, 2016년 1월 26일 수정, 2016년 1월 28일 채택)

요 약

특이점 및 영향점은 자료분석을 하는 데 사용되는 계량적이고 기술적인 많은 측도들을 왜곡한다. 각종 자료분석에 있어서의 특이점 검색을 위한 검정 통계량이나 그림도구에 관한 연구는 꾸준히 전개되어 왔다. Jang과 Anderson-Cook (2014)은 불꽃그림이란 이름을 붙인 그림도구를 발표하였는데 이상점이나 영향점이 일변량/이변량 자료분석 및 회귀분석에 어떠한 영향을 미치는지 알기 위하여 3-D 불꽃그림 및 불꽃그림 행렬을 제시하였다. 본 연구에서는 이러한 불꽃그림이 일변량 자료의 왜도와 첨도에서 특이점의 영향을 평가하기 위한 탐색적 자료분석 그림도구로서 사용될 수 있음을 보였다.

주요용어: 특이점, 영향점, 왜도, 첨도, (3-D) 불꽃그림, 불꽃그림 행렬

이 논문은 2015학년도 부산외국어대학교 학술연구조성비에 의해 연구되었음.

[1](46234) 부산시 금정구 금샘로 485번길, 부산외국어대학교 데이터경영학과. E-mail: shmoon@bufs.ac.kr