The Korean Journal of Applied Statistics
2022, Vol. 35, No. 3, 421–434

DOI: https://doi.org/10.5351/KJAS.2022.35.3.421
Print ISSN1225-066X / Online ISSN 2383-5818

Exploratory data analysis for Chatterjee's ξ coefficient

Dae-Heung Jang[1,a]

[a]Department of Statistics and Data Science, Pukyong National University

Abstract

Chatterjee (2021) proposed a new correlation coefficient ξ. Focusing on two questions (1. Is ξ coefficient distinguishable for Anscombe's quartet data set?, 2. How does the ξ coefficient value change according to the number of data for various kinds of scatterplots?), an exploratory data analysis is attempted for ξ coefficient. We can compare three measures (ξ coefficient, Pearson's correlation coefficient and mutual information).

Keywords: coefficient of correlation, ξ coefficient, mutual information

1. 서론

확률변수 Y가 상수가 아니라는 가정 하에 이변량 확률벡터 (X, Y)에 대응하는 이변량 데이터 쌍 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$이 얻어졌다고 하자. $\mathbf{x}$를 대상으로 정렬 $(x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)})$을 하면 데이터쌍 $(x_{(1)}, y_{(1)}), (x_{(2)}, y_{(2)}), \ldots, (x_{(n)}, y_{(n)})$이 얻어진다. X와 Y 사이의 선형성을 평가하는 측도로 피어슨 상관계수를 주로 사용한다. 또한 이상점에 대한 강건한 통계량으로서 스피어만 상관계수를 사용한다. 피어슨 상관계수의 단점을 극복하기 위한 대안으로서 많은 측도들이 제안되었다 (예로, 최대상관계수, 결합누적분포함수나 순위 기반 계수, 커널기반 계수, 정보이론 기반 계수, 코퓰라 기반 계수, 쌍별거리 기반 계수 등). Chatterjee (2021)는 이러한 측도들의 문제점들을 두 가지로 대략 지적하였다. 첫째, 대부분의 방법들이 X와 Y 사이의 관계성(relationship)의 강도(strength)를 측정하기 위한 측도로 개발되었다기보다는 독립성 검정에 촛점을 맞추어 개발되었다. 둘째, 대부분의 방법들이 귀무가설 하에서 p-값을 쉽게 계산할 수 있는 간단한 점근적 이론이 없다.

$\mathbf{x}$값이 모두 다른 경우 Chatterjee (2021)는 기존의 피어슨상관계수의 단점을 극복하기 위한 또 다른 대안으로서 새로운 상관계수 ξ를 다음과 같이 제시하였다.

$$\xi_n(X, Y) = 1 - \frac{3 \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1}, \tag{1.1}$$

여기서 r_i는 $y_{(i)}$에 대한 순위이다. Chatterjee (2021)는 $\mathbf{x}$값에 같은 값이 있는 경우의 ξ 계수 공식을 동시에 제시하였다.

Chatterjee (2021)는 새로운 상관계수 ξ에 대하여 다음과 같이 정리하였다.

1. 공식이 아주 간단하다.
2. 간단한 공식 때문에 개념을 이해하기 쉽고 측도 계산 속도가 아주 빠르다.
3. 순위를 기반으로 하기 때문에 이상점에 강건하다.

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. 2020R1I1A3A04037799).

[1] Department of Statistics and Data Science, Pukyong National University, 45, Yongso-ro, Nam-Gu, Busan 48513, Repulic of Korea. E-mail: dhjang@pknu.ac.kr

Published 30 June 2022 / journal homepage: http://kjas.or.kr
© 2022 The Korean Statistical Society. All rights reserved.

4. X와 Y가 독립이면 ξ 계수 값이 0이고 Y가 X의 가측함수(measurable function)가 되면 ξ 계수 값이 1이 된다. 그러므로 ξ 계수를 X와 Y 사이의 관계성의 강도를 측정하기 위한 측도로 사용할 수 있다.
5. 독립성 검정에서 아주 간단한 근사 이론을 가지고 있고 모든 대립가설에 대하여 일치성을 가지고 있다.
6. Y가 상수가 아니라는 가정 외에 X와 Y 사이의 다른 가정이 필요없다.
7. $\xi_n(X, Y)$는 대칭적이지 않기 때문에 X가 Y의 가측함수(measurable function)인지를 확인하고자 한다면 $\xi_n(Y, X)$를 사용하여야 한다.
8. 진동신호를 탐지하는 시뮬레이션이나 실제 데이터 분석에서 다른 방법들보다 좋다.
9. 단점으로서 부드럽고 비진동인 신호 대상 독립성 검정에서 표본크기가 작은 경우 다른 방법들보다 검정력이 떨어진다. 그러나 표본의 크기가 큰 경우 문제가 되지 않고 계산 속도가 무척 빠르다.

본 논문의 2절에서는 Chatterjee (2021)가 제시한 새로운 상관계수 ξ에 대하여 다음과 같은 두 가지 질문을 중심으로 탐색적자료분석을 시도하였다.

1. Anscombe's quartet 데이터셋에 대하여 ξ 계수는 구별이 가능한가?
2. 다양한 종류의 산점도에서 데이터의 개수에 따라 ξ 계수 값은 어떻게 변하는가?

탐색적자료분석 시 Chatterjee (2021)가 제시한 새로운 상관계수 ξ 와의 비교를 위하여 피어슨상관계수와 상호정보(mutual information; MI)를 사용하였다. 상호정보는 정보이론 기반 통계량으로서 X와 Y 사이의 선형성을 포함한 X와 Y 사이의 관계성의 강도를 측정하기 위한 측도로 쓰일 수 있다 (Yi 등, 2015).

3절에서 결론을 내렸다.

2. ξ 계수에 대한 탐색적자료분석

Chatterjee (2021)가 제시한 새로운 상관계수 ξ에 대하여 서론에서 언급한 두 가지 질문을 각 소절로 나누어 ξ 계수에 대한 탐색적자료분석을 시도하였다.

2.1. Anscombe의 quartet 데이터셋에 대한 ξ 계수의 구별

Anscombe의 quartet 데이터셋 (Anscombe, 1973)은 거의 동일한 기술 통계량을 갖지만 매우 다른 분포를 가지며 그래프로 표시할 때 매우 다르게 나타나는 4개의 데이터 세트로 구성된다. 각 데이터셋은 각각 11개의 (x, y) 점으로 구성되어 있다. 데이터를 분석할 때 그래프의 중요성을 강조할 때 많이 언급되는 데이터셋이다. 각 데이터셋을 대상으로 **x**, **y** 각각의 산술평균, 분산, **x**와 **y** 사이의 피어슨상관계수, 직선회귀식을 구하면 모두 같은 값이다. 그러나 산점도를 그려보면 다음 Figure 1처럼 다른 모습을 나타낸다.

각 데이터셋을 대상으로 피어슨상관계수, 스피어만상관계수, ξ 계수, 상호정보 값을 비교하면 Table 1과 같다. 피어슨상관계수는 네 개의 데이터셋을 구별해 내지 못한다. X와 Y 사이의 선형 관계의 강도를 나타내는 측도로서는 피어슨상관계수가 좋은 통계량이나 일반적인 X와 Y 사이의 관계의 강도를 나타내는 측도로서는 미진한 부분이 있다. 스피어만상관계수, ξ 계수, 상호정보는 네 개의 데이터셋을 구별하고 있음을 알 수 있다. ξ 계수와 상호정보는 데이터셋 IV에서 X와 Y 사이의 관계성의 강도가 제일 약하고 데이터셋 I보다는 데이터셋 III에서 X와 Y 사이의 관계성의 강도가 더 세다고 평가한다. 데이터셋 II에 대해서는 ξ 계수는 X와 Y 사이의 관계성의 강도가 데이터셋 III보다 조금 약하다고 평가한 반면 상호정보는 데이터셋 I과 같이 약하다고 평가한다. 흥미로운 사실은 데이터셋 I에 대해 피어슨상관계수와 스피어만상관계수는 X와 Y 사이의 선형성이 강하다고 판단한 반면 ξ 계수와 상호정보는 X와 Y 사이의 관계성의 강도가 약하다고 판단하고 있다는 사실이다. 데이터셋 IV에서 ξ 계수가 -0.075가 나오나 독립성 검정 시 p-값이 0.65가 나오므로 ξ 계수가 0

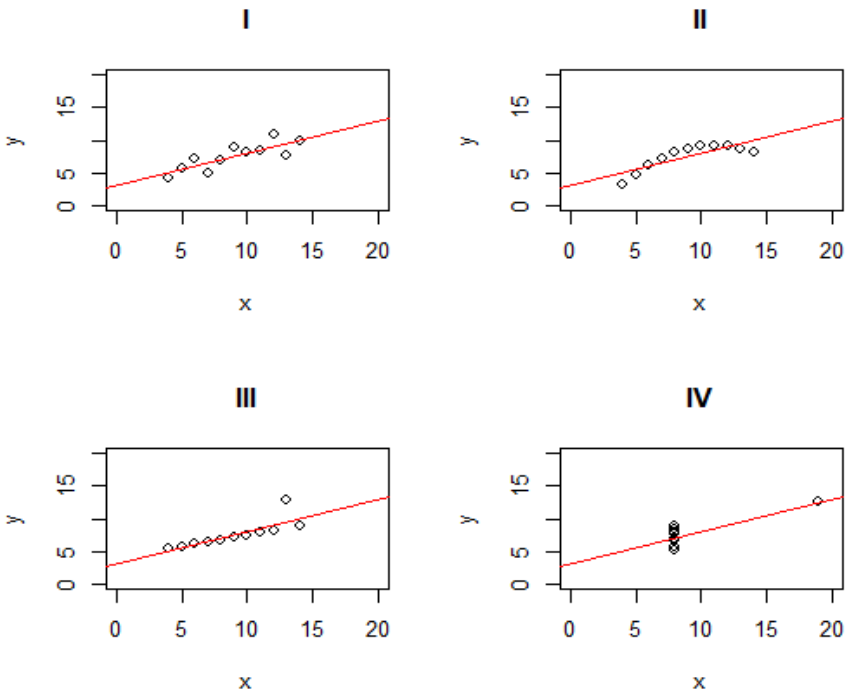

Figure 1: *Scatterplot and estimated linear regression line for Anscombe's quartet.*

Table 1: Measures for the strength of the relationship in Anscombe's quartet.

Measures	Dataset I	Dataset II	Dataset III	Dataset IV
Pearson's r	0.816	0.816	0.816	0.816
Spearman's ρ	0.818	0.691	0.991	0.500
ξ coefficient	0.275	0.600	0.725	−0.075
MI	0.216	0.216	0.689	0.077

이라고 할수 있다. 스피어만상관계수와 ξ 계수는 모두 순위를 기반으로 하는 계수이나 4개의 데이터셋에 대한 계수 값이 많이 다름을 알 수 있다. 특히 데이터셋 I과 데이터셋 IV에서 차이가 많이 난다. 이런 차이가 나는 것은 스피어만상관계수가 피어슨상관계수와 같이 X와 Y 사이의 선형 관계의 강도를 나타내는 측도인 반면 ξ 계수는 X와 Y 사이의 관계성의 강도를 측정하기 위한 측도라는 차이 때문에 기인한다.

2.2. 다양한 종류의 산점도에서 데이터 개수에 따른 ξ 계수의 변화

산점도에서 나타나는 X와 Y 사이의 구조를 네 가지 경우로 나누어 살펴보고자 한다.

1. (X, Y) **관계가 직선 관계인 경우**

X와 Y 사이에 다음과 같이 완전한 직선 관계가 있는 경우를 먼저 고려해 보자.

$$Y = X(-5 \le X \le 5).$$

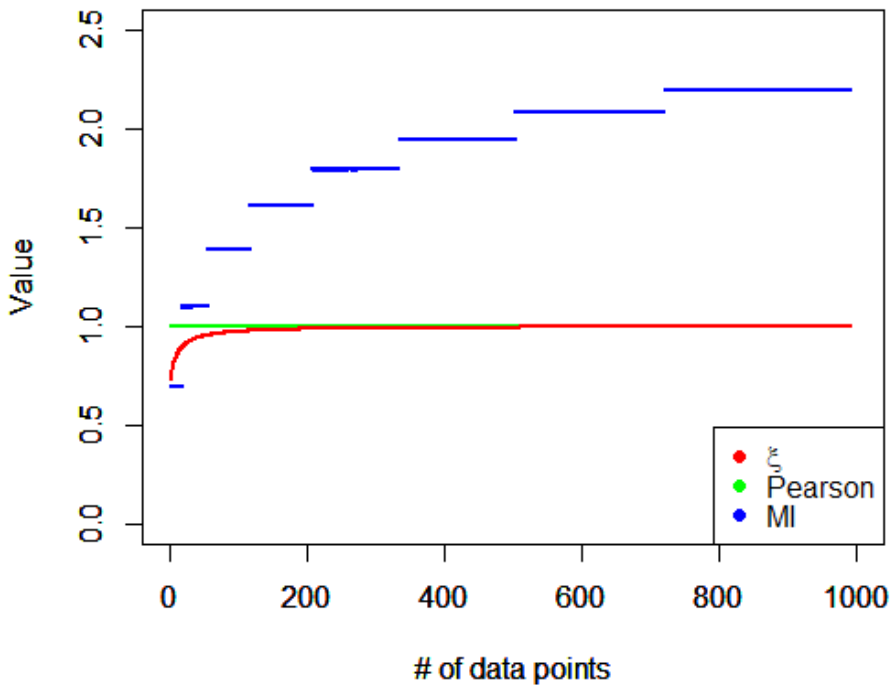

Figure 2: *Changes in measure values as the sample size increases with no noise model $Y = X$.*

X의 범위에서 균등난수 개수를 10에서 1,000까지 증가시키면서 피어슨상관계수, ξ 계수, 상호정보 값들을 구하는 작업을 100번 반복한 후 세 측도들 값의 변화를 보면 다음 Figure 2와 같다. 직선의 속성상 각 표본크기에서의 피어슨상관계수, ξ 계수, 상호정보 값들의 변동은 0이 됨을 확인할 수 있다. 피어슨상관계수는 표본의 크기에 관계없이 항상 1을 유지한다. 반면 ξ 계수는 표본의 크기가 증가할 때 ξ 계수 값은 증가하며 1로 단조수렴한다. 표본의 크기가 29일 때 0.900, 표본의 크기가 100일 때 0.970, 표본의 크기가 299일 때 0.990이 된다. 모든 표본 크기에서 **x** 데이터의 위치에 따른 ξ 계수 값의 표준편차가 0이 돼서 ξ 계수 값은 **x** 데이터의 위치에 관련이 없고 오직 데이터의 개수에 의해 좌우된다. 상호정보는 표본의 크기가 증가할 때 특이하게 계단 함수 패턴을 가지며 증가한다. 상호정보의 상한값이 없음이 상호정보의 단점임을 알 수 있다.

다음으로 X와 Y 사이에 다음과 같이 완전한 직선 관계에 약한 잡음 (평균이 0이고 표준편차가 0.5인 정규난수)을 첨가하는 경우를 살펴보자.

$$Y = X + \epsilon(-5 \le X \le 5), \quad \epsilon \sim N\left(0, 0.5^2\right)$$

X의 범위에서 균등난수 개수를 10에서 1,000까지 증가시키면서 피어슨상관계수, ξ 계수, 상호정보 값들을 구하는 작업을 100번 반복한 후 세 측도들 값의 변화를 보면 다음 Figure 3과 같다. 피어슨상관계수는 표본의 크기가 증가할 때 평균이 0.985로 수렴하는 반면 ξ 계수는 표본의 크기가 증가할 때 값은 증가하나 평균이 0.834로 수렴한다. 완전한 직선 관계에 약한 잡음이 있는 경우 피어슨상관계수는 영향을 받기는 하지만 X와 Y 사이에 매우 강한 직선 관계를 유지한다. 반면 ξ 계수는 X와 Y 사이에 직선 관계가 상대적으로 더 약화됨을 알 수 있다. 피어슨상관계수가 약한 잡음에 강건하다고 해석할 수도 있으나 다른 측면에서 X와 Y 사이의 관계성의 강도를 측정하는 입장에서는 오히려 약한 잡음을 더 민감하게 감지해야 하는 것이 중요할 수 있다. 피어슨상관계수에 비해 ξ 계수는 값의 더 큰 변화를 통하여 이러한 약한 잡음의 영향을 민감하게 잘 반영하고

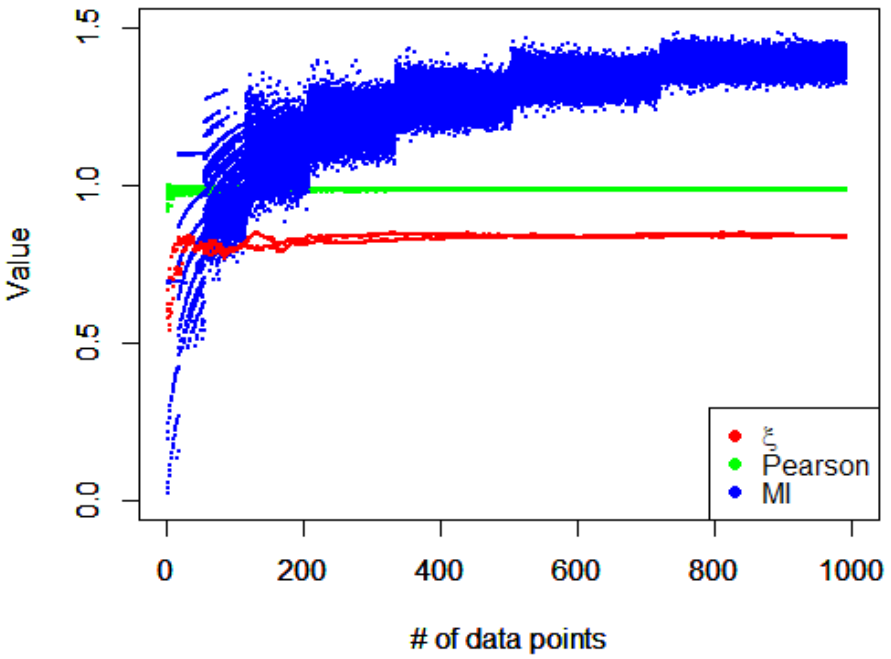

Figure 3: *Changes in measure values as the sample size increases with weak noise model* $Y = X + \epsilon, \epsilon \sim N(0, 0.5^2)$.

있다고 볼 수 있다. 모든 표본 크기에서 **x** 데이터의 위치에 따른 ξ 계수 값의 표준편차가 거의 0이 돼서 ξ 계수 값은 **x** 데이터의 위치에 거의 관련이 없고 오직 데이터의 개수에 의해 좌우된다. 상호정보는 표본의 크기가 증가할 때 변동이 계단식으로 감소하는 계단함수 형태의 특이한 패턴을 가지며 증가하나, 잡음이 없는 완전한 직선 관계일 때와 비교하면 그 값이 상대적으로 작다.

다음으로 X와 Y 사이에 다음과 같이 완전한 직선 관계에 강한 잡음 (평균이 0이고 표준편차가 2인 정규난수)을 첨가하는 경우를 살펴보자.

$$Y = X + \epsilon(-5 \le X \le 5), \quad \epsilon \sim N\left(0, 2^2\right)$$

X의 범위에서 균등난수 개수를 10에서 1,000까지 증가시키면서 피어슨상관계수, ξ 계수, 상호정보 값들을 구하는 작업을 100번 반복한 후 세 측도들 값의 변화를 보면 다음 Figure 4와 같다. 피어슨상관계수는 표본의 크기가 증가할 때 평균이 0.821로 수렴하는 반면 ξ 계수는 표본의 크기가 증가할 때 값은 증가하나 평균이 0.482로 수렴한다. 약한 잡음이 있는 경우보다 변동폭이 매우 크나 데이터 개수가 증가하며 변동폭이 점차 줄어든다. 완전한 직선 관계에 강한 잡음이 있는 경우 피어슨상관계수는 영향을 받기는 하지만 X와 Y 사이에 강한 직선 관계를 유지하나 ξ 계수는 X와 Y 사이에 직선 관계가 상대적으로 매우 약화됨을 알 수 있다. 피어슨상관계수가 강한 잡음에도 불구하고 강건하다고 해석할 수도 있으나 다른 측면에서 X와 Y 사이의 관계성의 강도를 측정하는 입장에서는 오히려 약한 잡음 뿐만 아니라 강한 잡음을 감지해야 하는 것도 중요할 수 있다. 피어슨상관계수에 비해 ξ 계수는 이러한 강한 잡음의 영향을 ξ 계수 값의 더 큰 변화를 통하여 매우 민감하게 잘 반영하고 있다고 볼 수 있다. 완전한 직선 관계가 있는 경우 ξ 계수 수렴값이 1, 완전한 직선 관계에 약한 잡음이 있는 경우는 ξ 계수 수렴값이 0.834, 완전한 직선 관계에 강한 잡음이 있는 경우는 ξ 계수 수렴값이 0.482로 민감하게 반응하는 현상은 다음에 다룰 (X, Y) 관계가 직선 관계가 아닌 곡선관계, 삼각함수 관계 등

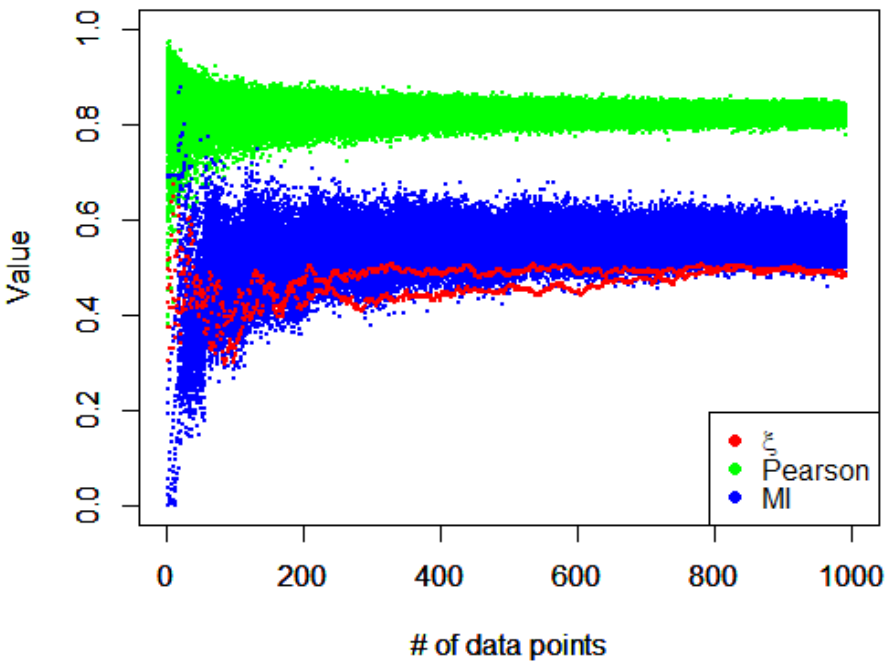

Figure 4: *Changes in measure values as the sample size increases with strong noise model* $Y = X + \epsilon, \epsilon \sim N(0, 2^2)$.

다양한 함수 관계에도 동일하게 적용된다. 이러한 성질을 갖는 계수를 Reshef 등 (2011)은 공정한(equitable) 계수라 정의하였다. X와 Y 사이에 함수 관계가 존재한다면 우리는 모의실험을 통하여 ξ 계수가 이러한 공정성질을 가지고 있음을 확인할 수 있다. 모든 표본 크기에서 **x** 데이터의 위치에 따른 ξ 계수 값의 표준편차가 거의 0이 돼서 ξ 계수 값은 **x** 데이터의 위치에 거의 관련이 없고 오직 데이터의 개수에 의해 좌우된다. 상호정보는 표본의 크기가 증가할 때 증가하며 평균이 0.551로 수렴한다. 약한 잡음이 있는 경우에 비해 급격하게 수렴값이 작아진다.

2. (X, Y) **관계가 곡선 관계인 경우**

X와 Y 사이에 다음과 같이 완전한 곡선 관계가 있는 경우를 먼저 고려해 보자.

$$Y = X^2(-5 \le X \le 5)$$

X의 범위에서 균등난수 개수를 10에서 1,000까지 증가시키면서 피어슨상관계수, ξ 계수, 상호정보 값들을 구하는 작업을 100번 반복한 후 세 측도들 값의 변화를 보면 다음 Figure 5와 같다. 피어슨상관계수는 표본의 크기가 증가할 때 평균이 0에 수렴한다. 그러므로 X와 Y 사이에 직선 관계성이 전혀 없음을 알 수 있으나 피어슨상관계수는 X와 Y 사이의 곡선 관계를 전혀 탐지하지 못한다. 표본의 크기가 증가하면 **x** 데이터의 위치에 따른 피어슨상관계수 값의 변동이 점점 작아지기는 하나 꽤 크게 유지된다. 반면 ξ 계수는 표본의 크기가 증가할 때 값이 증가하며 1로 단조수렴함으로 X와 Y 사이의 곡선 관계를 탐지할 수 있게 된다. 모든 표본 크기에서 **x** 데이터의 위치에 따른 ξ 계수 값의 표준편차가 거의 0이거나 0이 돼서 ξ 계수 값은 **x** 데이터의 위치에 거의 관련이 없고 오직 데이터의 개수에 의해 좌우된다. 표본의 크기가 57일 때 ξ 계수 값이 0.901, 표본의 크기가 100일 때 ξ 계수 값이 0.941, 표본의 크기가 599일 때 ξ 계수 값이 0.990이 된다. 상호정보는 표본의 크기가 증

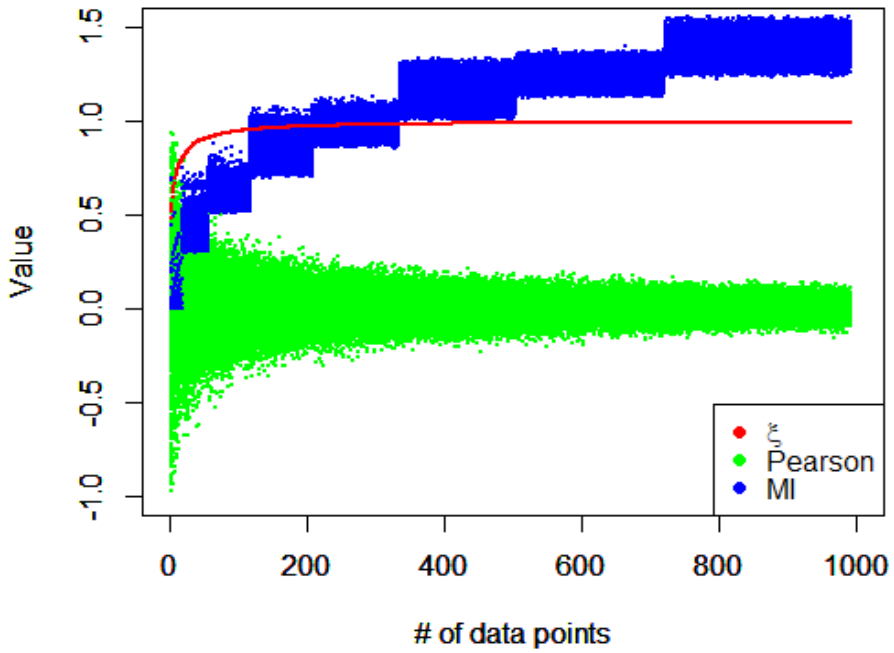

Figure 5: *Changes in measure values as the sample size increases with no noise model* $Y = X^2$.

가할 때 변동이 계단식으로 변화하는 계단 함수 형태의 특이한 패턴을 가지며 증가한다. 상호정보를 통해서도 X와 Y 사이의 곡선 관계를 탐지할 수 있게 된다.

다음으로 X와 Y 사이에 다음과 같이 완전한 곡선 관계에 약한 잡음 (평균이 0이고 표준편차가 0.5인 정규 난수)을 첨가하는 경우를 살펴보자.

$$Y = X^2 + \epsilon(-5 \le X \le 5), \quad \epsilon \sim N\left(0, 0.5^2\right)$$

X의 범위에서 균등난수 개수를 10에서 1,000까지 증가시키면서 피어슨상관계수, ξ 계수, 상호정보 값들을 구하는 작업을 100번 반복한 후 세 측도들 값의 변화를 보면 다음 Figure 6과 같다. 피어슨상관계수는 표본의 크기가 증가할 때 평균이 0에 수렴한다. 그러므로 X와 Y 사이에 직선 관계성이 전혀 없음을 알 수 있으나 피어슨상관계수는 X와 Y 사이의 곡선 관계를 전혀 탐지하지 못한다. 반면 ξ 계수는 표본의 크기가 증가할 때 값은 증가하며 평균이 0.906으로 수렴한다. 완전한 곡선 관계에 약한 잡음이 있는 경우 ξ 계수의 수렴값이 평균이 1이되지 못하고 평균이 0.906으로 작아짐으로 X와 Y 사이에 곡선 관계가 약화됨을 알 수 있다. 이러한 약한 잡음의 영향을 ξ 계수 값의 작은 변화를 통하여 민감하게 잘 반영하고 있다고 볼 수 있다. 모든 표본 크기에서 **x** 데이터의 위치에 따른 ξ 계수 값의 표준편차가 거의 0이 돼서 ξ 계수 값은 **x** 데이터의 위치에 거의 관련이 없고 오직 데이터의 개수에 의해 좌우된다. 상호정보는 표본의 크기가 증가할 때 변동이 계단식으로 감소하는 계단함수 형태의 특이한 패턴을 가지며 증가한다. 잡음이 없는 완전한 곡선 관계일 때와 비교하면 약한 잡음의 영향으로 상호 정보의 값이 작아진다.

다음으로 X와 Y 사이에 다음과 같이 완전한 곡선 관계에 강한 잡음 (평균이 0이고 표준편차가 3인 정규 난수)을 첨가하는 경우를 살펴보자.

$$Y = X^2 + \epsilon(-5 \le X \le 5), \quad \epsilon \sim N\left(0, 3^2\right)$$

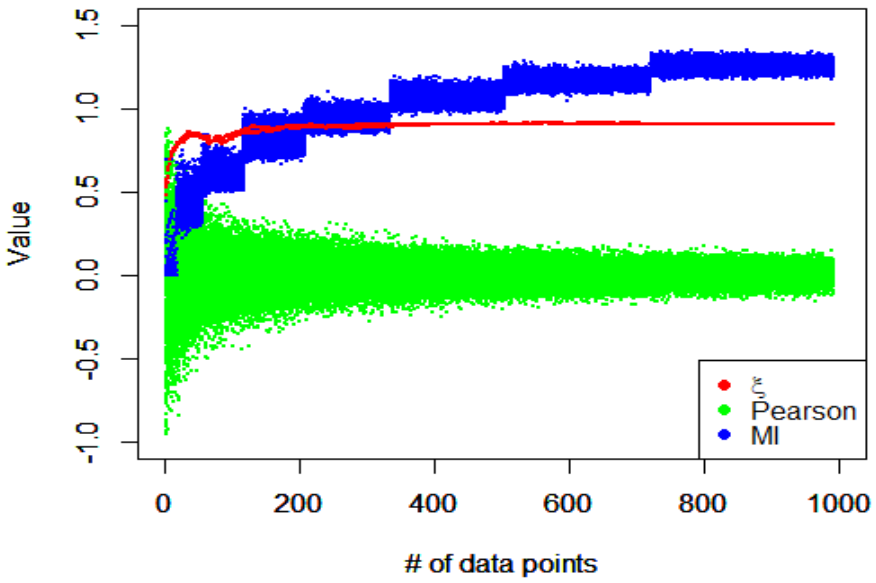

Figure 6: *Changes in measure values as the sample size increases with weak noise model* $Y = X^2 + \epsilon,\ \epsilon \sim N(0, 0.5^2)$.

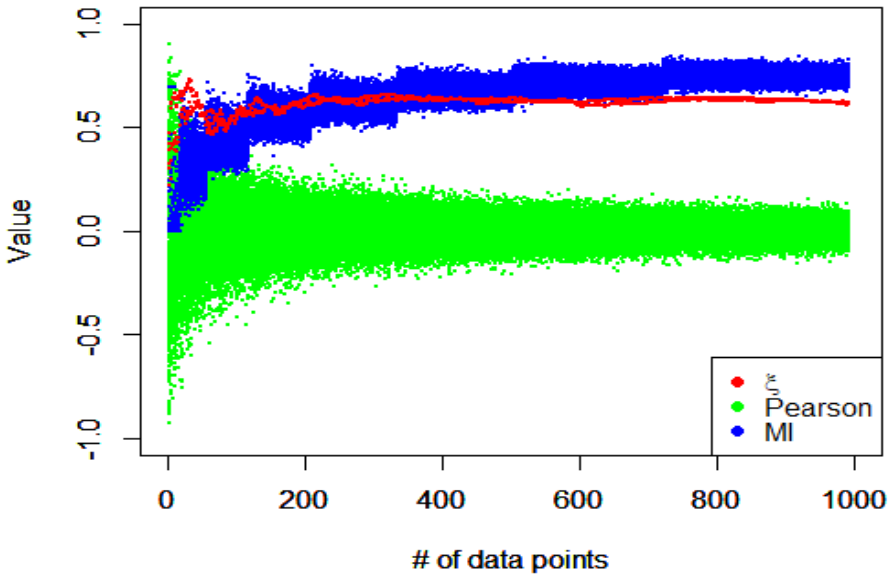

Figure 7: *Changes in measure values as the sample size increases with strong noise model* $Y = X^2 + \epsilon,\ \epsilon \sim N(0, 3^2)$.

X의 범위에서 균등난수 개수를 4에서 1,000까지 증가시키면서 피어슨상관계수, ξ 계수, 상호정보 값들을 구하는 작업을 100번 반복한 후 세 측도들 값의 변화를 보면 다음 Figure 7과 같다. 피어슨상관계수는 표본의

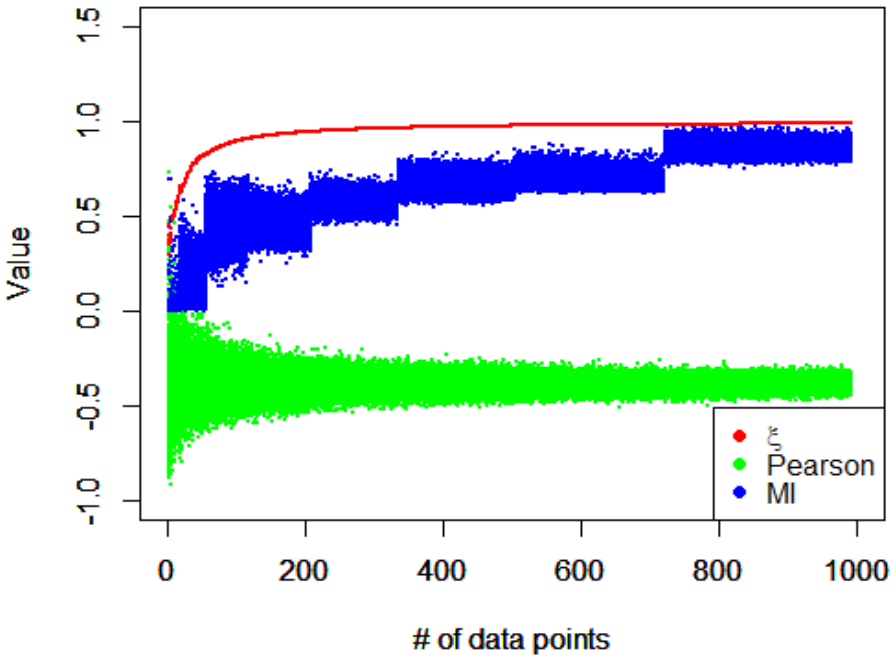

Figure 8: *Changes in measure values as the sample size increases with no noise model* $Y = \sin(X)$.

크기가 증가할 때 평균이 0에 수렴한다. 그러므로 X와 Y 사이에 직선 관계성이 전혀 없음을 알 수 있으나 피어슨상관계수는 X와 Y 사이의 곡선 관계를 전혀 탐지하지 못한다. 반면 ξ 계수는 표본의 크기가 증가할 때 값은 증가하며 평균이 0.614로 수렴한다. 완전한 곡선 관계에 강한 잡음이 있는 경우 ξ 계수의 수렴값이 평균이 1이되지 못하고 평균이 0.614로 매우 작아짐으로 X와 Y 사이에 곡선 관계가 매우 약화됨을 알 수 있다. 이러한 강한 잡음의 영향을 ξ 계수 값의 큰 변화를 통하여 민감하게 잘 반영하고 있다고 볼 수 있다. 완전한 곡선 관계가 있는 경우 ξ 계수 수렴값이 1, 완전한 곡선 관계에 약한 잡음이 있는 경우는 ξ 계수 수렴값이 0.906, 완전한 곡선 관계에 강한 잡음이 있는 경우는 ξ 계수 수렴값이 0.614로 민감하게 반응하는 현상을 통하여 ξ 계수가 공정 성질을 가지고 있음을 확인할 수 있다. 모든 표본 크기에서 **x** 데이터의 위치에 따른 ξ 계수 값의 표준편차가 거의 0이 돼서 ξ 계수 값은 **x** 데이터의 위치에 거의 관련이 없고 오직 데이터의 개수에 의해 좌우된다. 상호정보는 표본의 크기가 증가할 때 변동이 계단식으로 감소하는 계단함수 형태의 특이한 패턴을 가지며 증가한다. 약한 잡음이 있는 경우와 비교하면 강한 잡음의 영향으로 상호 정보의 값이 더 작아진다.

3. (X, Y) **관계가 주기함수 관계인 경우**

X와 Y 사이에 다음과 같이 완전한 주기함수 관계가 있는 경우를 먼저 고려해 보자.

$$Y = \sin(X)(0 \le X \le 4\pi)$$

X의 범위에서 균등난수 개수를 10에서 1,000까지 증가시키면서 피어슨상관계수, ξ 계수, 상호정보 값들을 구하는 작업을 100번 반복한 후 세 측도들 값의 변화를 보면 다음 Figure 8과 같다. 피어슨상관계수는 표본의 크기가 증가할 때 평균이 −0.390에 수렴한다. 그러므로 X와 Y 사이에 직선 관계성이 크지 않음을 알 수 있으나 피어슨상관계수는 X와 Y 사이의 주기함수 관계를 전혀 탐지하지 못한다. 표본의 크기가 증가하면

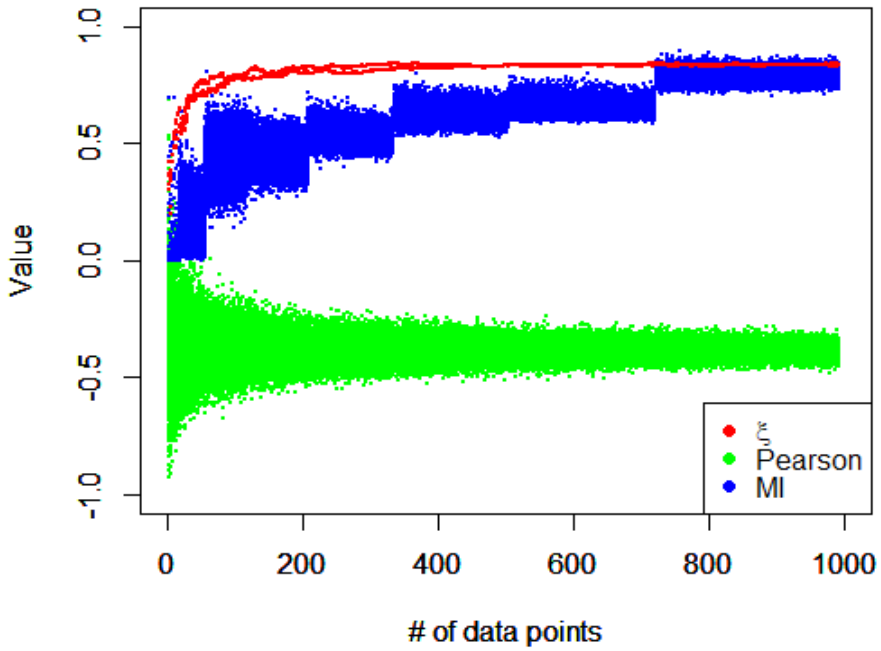

Figure 9: *Changes in measure values as the sample size increases with weak noise model* $Y = sin(X) + \epsilon,\ \epsilon \sim N(0, 0.1^2)$.

x 데이터의 위치에 따른 피어슨상관계수 값의 변동이 점점 작아지기는 하나 꽤 크게 유지된다. 반면 ξ 계수는 표본의 크기가 증가할 때 값이 증가하며 1로 단조수렴함으로 X와 Y 사이의 주기함수 관계를 탐지할 수 있게 된다. 모든 표본 크기에서 **x** 데이터의 위치에 따른 ξ 계수 값의 표준편차가 거의 0이거나 0이 돼서 ξ 계수 값은 **x** 데이터의 위치에 거의 관련이 없고 오직 데이터의 개수에 의해 좌우된다. 표본의 크기가 100일 때 ξ 계수 값이 0.884, 표본의 크기가 238일 때 ξ 계수 값이 0.950이 된다. 상호정보는 표본의 크기가 증가할 때 변동이 계단식으로 감소하는 계단 함수 형태의 특이한 패턴을 가지며 증가한다. 상호정보를 통해서도 X와 Y 사이의 주기함수 관계를 탐지할 수 있게 된다.

다음으로 X와 Y 사이에 다음과 같이 완전한 주기함수 관계에 약한 잡음 (평균이 0이고 표준편차가 0.1인 정규 난수)을 첨가하는 경우를 살펴보자.

$$Y = \sin(X) + \epsilon(0 \le X \le 4\pi), \quad \epsilon \sim N\left(0, 0.1^2\right)$$

X의 범위에서 균등난수 개수를 10에서 1,000까지 증가시키면서 피어슨상관계수, ξ 계수, 상호정보 값들을 구하는 작업을 100번 반복한 후 세 측도들 값의 변화를 보면 다음 Figure 9와 같다. 피어슨상관계수는 표본의 크기가 증가할 때 평균이 −0.389에 수렴한다. 그러므로 X와 Y 사이에 직선 관계성이 크지 않음을 알 수 있으나 피어슨상관계수는 X와 Y 사이의 주기함수 관계를 전혀 탐지하지 못한다. 반면 ξ 계수는 표본의 크기가 증가할 때 값은 증가하며 평균이 0.834로 수렴한다. 완전한 주기함수 관계에 약한 잡음이 있는 경우 ξ 계수의 수렴값이 평균이 1이되지 못하고 평균이 0.834로 작아짐으로 X와 Y 사이에 곡선 관계가 약화됨을 알 수 있다. 이러한 약한 잡음의 영향을 ξ 계수 값의 변화를 통하여 민감하게 잘 반영하고 있다고 볼 수 있다. 모든 표본 크기에서 **x** 데이터의 위치에 따른 ξ 계수 값의 표준편차가 거의 0이 돼서 ξ 계수 값은 **x** 데이터의 위치에 거의 관련이 없고 오직 데이터의 개수에 의해 좌우된다. 상호정보는 표본의 크기가 증가할 때 변동이 계단식으로 감소하는

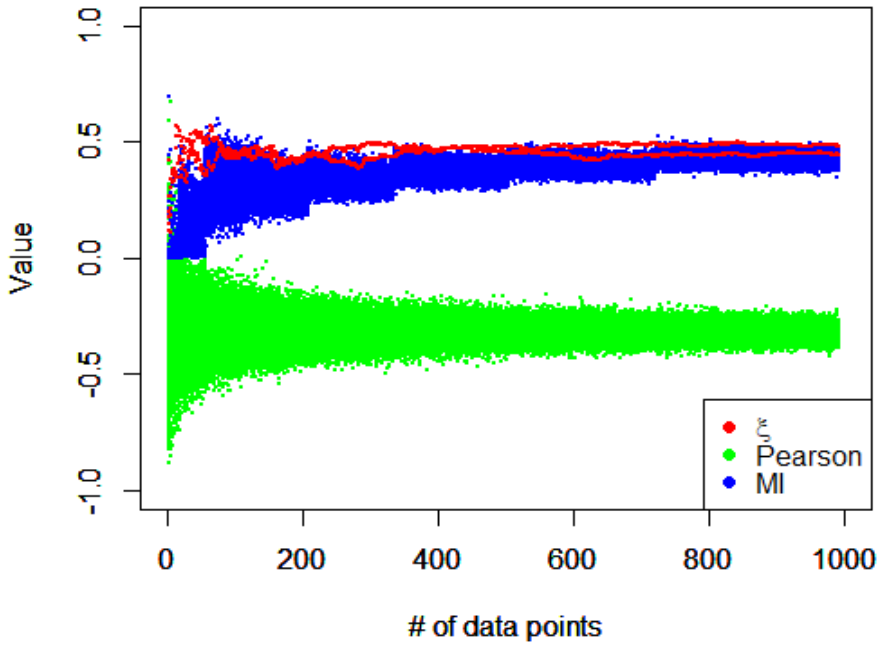

Figure 10: *Changes in measure values as the sample size increases with strong noise model* $Y = \sin(X) + \epsilon, \epsilon \sim N(0, 0.5^2)$.

계단함수 형태의 특이한 패턴을 가지며 증가한다. 잡음이 없는 완전한 주기함수 관계일 때와 비교하면 약한 잡음의 영향으로 상호 정보의 값이 작아진다.

다음으로 X와 Y 사이에 다음과 같이 완전한 주기함수 관계에 강한 잡음 (평균이 0이고 표준편차가 0.5인 정규 난수)을 첨가하는 경우를 살펴보자.

$$Y = \sin(X) + \epsilon(0 \le X \le 4\pi), \quad \epsilon \sim N\left(0, 0.5^2\right)$$

X의 범위에서 균등난수 개수를 10에서 1,000까지 증가시키면서 피어슨상관계수, ξ 계수, 상호정보 값들을 구하는 작업을 100번 반복한 후 세 측도들 값의 변화를 보면 다음 Figure 10과 같다. 피어슨상관계수는 표본의 크기가 증가할 때 평균이 −0.322에 수렴한다. 그러므로 X와 Y 사이에 직선 관계성이 크지 않음을 알 수 있으나 피어슨상관계수는 X와 Y 사이의 주기함수 관계를 전혀 탐지하지 못한다. 반면 ξ 계수는 표본의 크기가 증가할 때 값은 증가하며 평균이 0.447로 수렴한다. 완전한 주기함수 관계에 강한 잡음이 있는 경우 ξ 계수의 수렴값이 평균이 1이되지 못하고 평균이 0.447로 매우 작아짐으로 X와 Y 사이에 곡선 관계가 매우 약화됨을 알 수 있다. 이러한 강한 잡음의 영향을 ξ 계수 값의 큰 변화를 통하여 민감하게 잘 반영하고 있다고 볼 수 있다. 완전한 주기함수 관계가 있는 경우 ξ 계수 수렴값이 1, 완전한 주기함수 관계에 약한 잡음이 있는 경우는 ξ 계수 수렴값이 0.834, 완전한 주기함수 관계에 강한 잡음이 있는 경우는 ξ 계수 수렴값이 0.447로 민감하게 반응하는 현상을 통하여 ξ 계수가 공정 성질을 가지고 있음을 확인할 수 있다. 모든 표본 크기에서 $\mathbf{x}$ 데이터의 위치에 따른 ξ 계수 값의 표준편차가 거의 0이 돼서 ξ 계수 값은 $\mathbf{x}$ 데이터의 위치에 거의 관련이 없고 오직 데이터의 개수에 의해 좌우된다. 상호정보는 표본의 크기가 증가할 때 변동이 계단식으로 감소하는 계단함수 형태의 특이한 패턴을 가지며 증가한다. 약한 잡음이 있는 경우와 비교하면 강한 잡음의 영향으로 상호

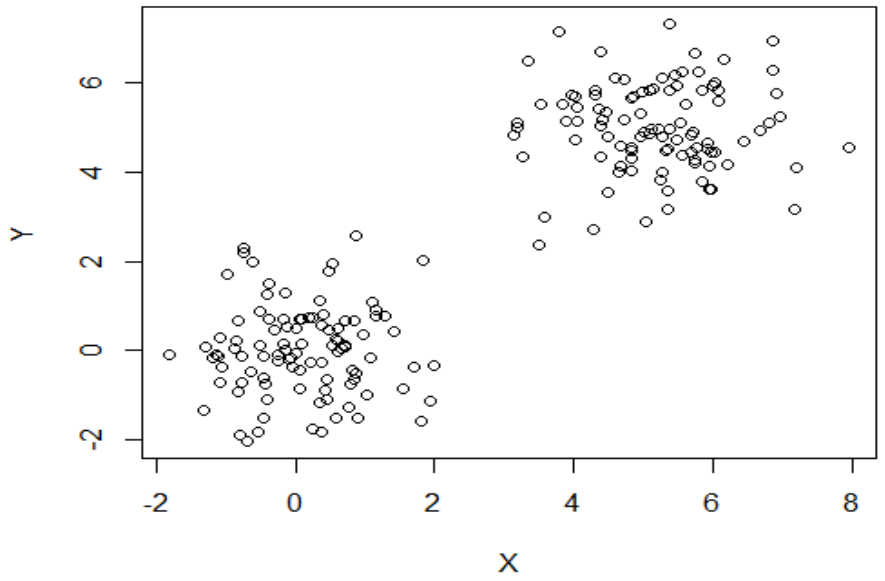

Figure 11: *Mixture of bivariate normals.*

정보의 값이 더 작아진다. 상호정보는 표본의 크기가 증가할 때 평균이 0.425로 수렴하여 ξ 계수의 수렴값의 평균 0.447과 비슷한 값을 가지게 된다.

4. (X, Y) **관계가 방울(blob) 구조를 갖는 경우**

(X, Y)가 Figure 11과 같이 방울 (blob) 구조를 갖는 경우 데이터 개수가 증가하면서 ξ 계수 값의 변화는 어떠한지를 살펴보자. 다음과 같은 이변량 정규혼합모형을 고려해 보자.

$$0.5\text{BN}(\mu_1, I) + 0.5\text{BN}(\mu_2, I),$$

여기서, $\mu_1' = (0, 0)$, $\mu_2' = (5, 5)$, I는 2×2 단위행렬이다.

데이터의 개수를 10에서 1,000까지 증가시키면서 피어슨상관계수, ξ 계수, 상호정보 값들을 구하는 작업을 100번 반복한 후 값들의 변화를 보면 다음 Figure 12와 같다. 피어슨상관계수는 표본의 크기가 증가할 때 평균이 0.862에 수렴한다. 그러므로 X와 Y 사이에 직선 관계성이 크다고 계산되어지나 피어슨상관계수는 X와 Y 사이의 방울 구조를 전혀 탐지하지 못한다. 반면 ξ 계수는 표본의 크기가 작을 때는 변동폭이 커서 불안정하나 표본의 크기가 증가할수록 변동폭이 감소하며 평균이 0.488로 수렴한다. 상호정보는 표본의 크기가 증가할 때 대략 0.6 값을 중심으로 위아래로 평균값이 이동하며 변동이 계단식으로 감소하는 계단함수 형태의 아주 특이한 패턴을 갖는다.

3. 결론

Chatterjee의 ξ 계수에 대하여 두 가지 질문을 통하여 이 계수가 기존의 피어슨상관계수의 단점을 극복하기 위한 좋은 대안이 될 수 있는지 탐색적으로 살펴보았다. 그 결과로서 Chatterjee의 ξ 계수가 기존의 피어슨상관계수의 단점을 극복하기 위한 좋은 대안이 될 수 있음을 확인하였다. 우리는 머신러닝 및 예측분석 시

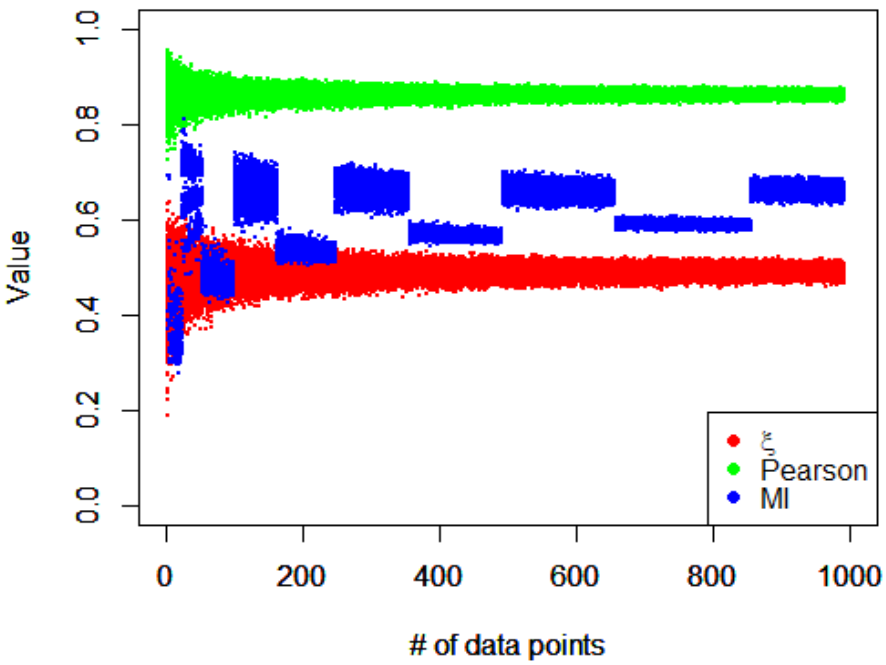

Figure 12: *Changes in measure values as the sample size increases with normal mixture.*

특징[변수] 선택의 판정기준 (X와 Y 사이의 관계성의 강도를 측정하기 위한 측도)으로서 ξ 계수를 다양하게 사용할 수 있을 것이다.

모의실험을 통하여 X와 Y 사이에 함수 관계가 존재할 때 ξ 계수가 공정 성질을 가지고 있음을 확인하였다. 이러한 성질이 ξ 계수의 정의(definition)로부터 이론적으로 유도될 수 있는지를 살피는 작업이 추후의 연구과제가 될 수 있다.

References

Anscombe FJ (1973). Graphs in statistical analysis, *American Statistician*, **27**, 17–21.

Chatterjee S (2021). A new coefficient of correlation, *Journal of American Statistical Association*, **116**, 2009–2022.

Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, and Sabeti P (2011), Detecting novel associations in large datasets, *Science*, **334**, 1518–1524.

Yi S and Jang DH (2015). An application of mutual information in mathematical statistics education, *Journal of the Korean Data & Information Science Society*, **26**, 1017–1025.

Received February 24, 2022; Revised March 19, 2022; Accepted April 28, 2022

Chatterjee의 ξ 계수에 대한 탐색적자료분석

장대흥[1,a]

[a]부경대학교 통계·데이터사이언스학과

요 약

Chatterjee (2021)는 새로운 상관계수 ξ를 제안하였다. 두 가지 질문 (1. Anscombe's quartet 데이터셋에 대하여 ξ 계수는 구별이 가능한가?, 2. 다양한 종류의 산점도에서 데이터의 개수에 따라 ξ 계수 값의 변화는 어떠한가?)을 중심으로 ξ 계수에 대한 탐색적자료분석을 시도하였다. 세 가지 측도 (ξ 계수, 피어슨상관계수, 상호정보)를 서로 비교하였다.

주요용어: 상관계수, ξ 계수, 상호정보

이 논문은 한국연구재단(NRF)의 지원을 받아 연구되었음(No. 2020R1I1A3A04037799).

[1](48513) 부산광역시 남구 용소로 45, 부경대학교 통계·데이터사이언스학과. E-mail: dhjang@pknu.ac.kr