The Korean Journal of Applied Statistics
2023, Vol. 36, No. 3, 261–269

DOI: https://doi.org/10.5351/KJAS.2023.36.3.261
Print ISSN1225-066X / Online ISSN 2383-5818

Applications of Chatterjee correlation coefficient

Sojin Ahn[a], Dae-Heung Jang[1,a]

[a]Division of Data and Information Sciences, Pukyong National University

Abstract

Chatterjee (2021) proposed a new correlation coefficient ξ as an alternative to overcome the disadvantages of the existing Pearson's correlation coefficient. Since this correlation coefficient is rank-based, it is robust to outliers, and the simple formula makes the concept easy to understand and the measure calculation speed is very fast. Through this paper, the application of this correlation coefficient was examined in two aspects (1. Bivariate distribution, 2. High-dimensional data).

Keywords: Chatterjee correlation coefficient, bivariate distribution, high-dimensional data

1. 서론

확률 변수 Y가 상수가 아니라는 가정 하에 이변량 확률 벡터(X, Y)에 대응하는 이변량 데이터 쌍 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$이 얻어졌을 때 $\mathbf{x}$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 상관계수가 주로 사용된다. $\mathbf{x}$값이 모두 다른 경우 Chatterjee (2021)는 기존의 피어슨 상관계수의 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)가 제시한 상관계수 ξ(앞으로의 논의에서 CCC (Chatterjee correlation coefficient)라는 약어를 사용하고자 한다.)는 다음과 같은 특징들이 있다.

1. 순위를 기반으로 하기 때문에 이상점에 강건하다.
2. 공식이 아주 간단하다. 이러한 간단한 공식 때문에 개념을 이해하기 쉽고 측도 계산 속도가 아주 빠르다. 또한 Y가 상수가 아니라는 가정 외에 X와 Y 사이의 다른 가정이 필요없다.
3. X와 Y가 독립이면 ξ 계수 값이 0이고 Y가 X의 가측 함수(measurable function)가 되면 ξ 계수 값이 1이 된다. 그러므로 ξ 계수를 X와 Y 사이의 관계성의 강도를 측정하기 위한 측도로 사용할 수 있다. $\xi_n(X, Y)$는 대칭적이지 않기 때문에 X가 Y의 가측 함수(measurable function)인지를 확인하고자 한다면 $\xi_n(Y, X)$를 사용하여야 한다.

This work was supported by a Research Grant of Pukyong National University (2021).

[1] Corresponding author: Division of Data and Information Sciences, Pukyong National University, 45 Yongso-ro, Nam-Gu, Busan 48513, Korea. E-mail: dhjang@pknu.ac.kr

Published 30 June 2023 / journal homepage: http://kjas.or.kr

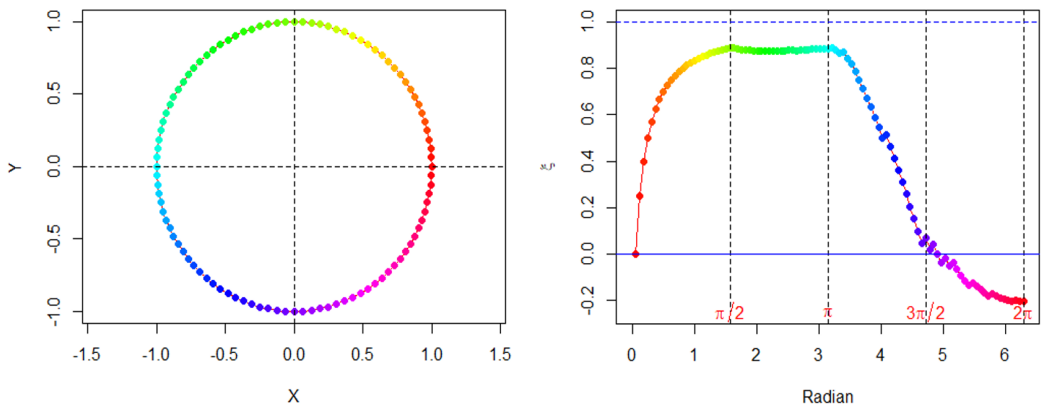

Figure 1: *A scatterplot of points on a circle and the corresponding CCC plot.*

4. 독립성 검정에서 아주 간단한 근사 이론을 가지고 있고 모든 대립가설에 대하여 일치성을 가지고 있다.
5. 진동신호를 탐지하는 시뮬레이션이나 실제 데이터 분석에서 다른 방법들보다 좋다. 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.

X와 Y 사이의 관계성의 강도를 측정하기 위한 측도로서의 CCC의 특징을 살펴보기 위하여 원 $X^2+Y^2=1$을 대상으로 $(1,0)$에서 시작하여 원 상의 점들을 일정한 각도를 유지하며 증가시키면서 CCC를 계산하면 Figure 1과 같이 나타난다. $(1,0)$에서 시작하여 원 상의 점들을 일정한 각도를 유지하며 증가시켜 극좌표에서의 각도가 $\pi/2$ 라디안에 이르게 되면 CCC가 0에서 비선형적으로 증가하며 0.9 가까이 도달하게 되고, 이 값은 π 라디안까지 유지되다가 $3\pi/2$ 라디안까지 거의 선형적으로 0 가까이 감소한다.

Chatterjee (2021) 논문 발표 이후 CCC에 대한 관심이 높아지고 있다. Sadeghi (2022)는 지구화학연구(geochemical study)에 CCC를 사용할 것을 제안하였다. Shi 등 (2022)는 독립성 검정에서 CCC의 검정력에 대하여 세 가지 검정방법(Hoeffding's D, Blum–Kiefer–Rosenblatt's R, Bergsma–Dassios–Yanagimoto's τ)과 비교를 행하였다. Auddy 등 (2021)은 독립성 검정에서 Chatterjee 검정 통계량에 대한 정확한 탐지 임계값에 대하여 논의하였고 Lin과 Han (2021)은 독립성 검정에서 CCC의 검정력을 강화하는 방안에 대하여 연구하였다.

본 논문의 2절에서는 CCC의 응용을 두 가지 측면(1. 이변량 분포, 2. HDLSS (high dimension low sample size) 자료)에서 살펴보았다. CCC와의 비교를 위하여 피어슨 상관계수(Pearson correlation coefficient; PCC), 스피어만 상관계수(Spearman correlation coefficient; SCC), 상호정보(mutual information; MI)를 사용하였다.

3절에서 결론을 내렸다.

2. Chatterjee 상관계수의 응용

CCC의 응용을 두 가지 측면(1. 이변량 분포, 2. HDLSS 자료)에서 살펴보고자 한다.

2.1. 이변량 분포 하에서의 Chatterjee 상관계수

첫 번째 이변량 분포로서 모평균 벡터가 $\mu'=(0,0)$이고 분산-공분산 행렬이 $\Sigma=\begin{pmatrix}1&\rho\\ \rho&1\end{pmatrix}$인 이변량 정규 분포 $\mathrm{BN}(\mu,\Sigma)$에 대하여 생각해 보자. $\rho=-0.99$에서 $\rho=0.99$까지 0.01씩 증가시키면서 이변량 정규 분포 $\mathrm{BN}(\mu,\Sigma)$에서 난수 100개를 추출하여 PCC를 구하는 작업을 100번 반복하면 PCC plot를 그릴 수 있다. 같은 작업을

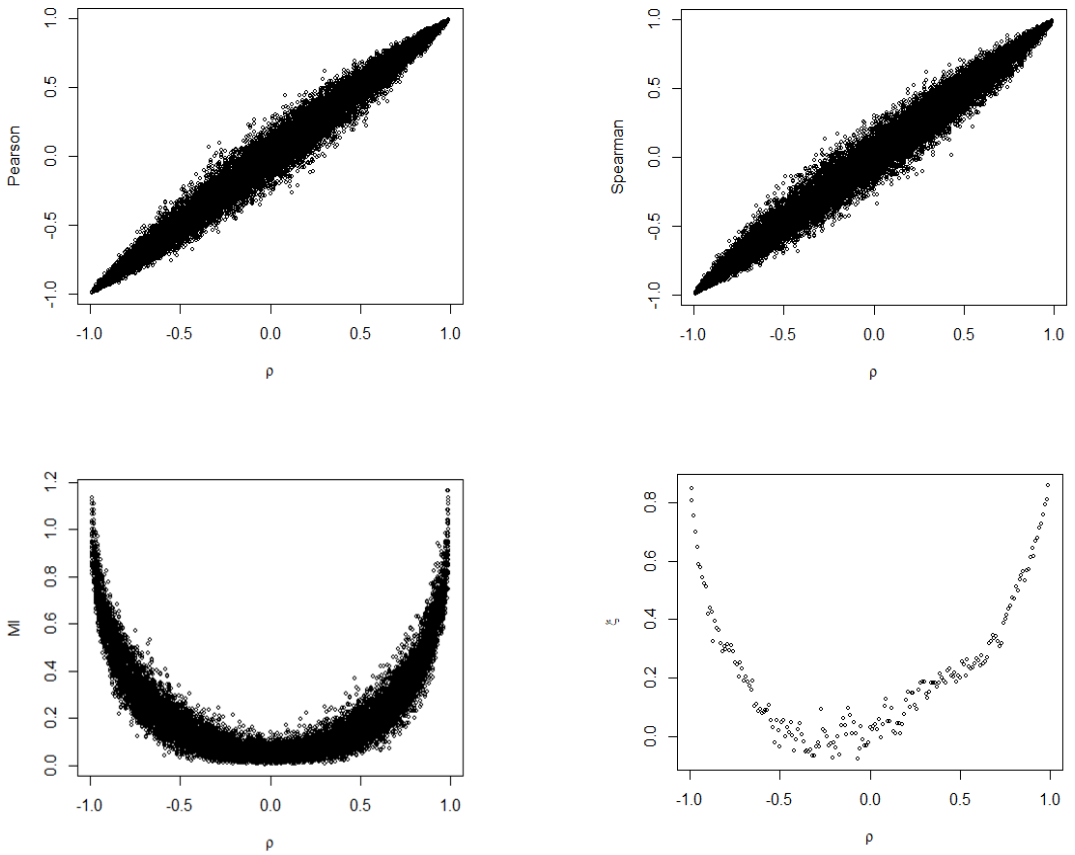

Figure 2: *PCC plot, SCC plot, MI plot, and CCC plot with bi-variate normal distribution.*

SCC, MI, CCC에 대하여 각각 시행하면 SCC plot, MI plot, CCC plot를 그릴 수 있다. Figure 2는 이 4개의 그림을 나타낸다.

PCC plot이나 SCC plot에서는 ρ의 절대값이 1에 가까우면 PCC나 SCC 값의 변동이 거의 없으나 ρ의 절대값이 1에서 점점 작아지면 PCC나 SCC 값의 변동이 점점 커지다가 0에 가까우면 PCC나 SCC 값의 변동이 제일 큰 것을 알 수 있다. MI plot에서는 ρ의 전 범위에 걸쳐 MI 값의 변동이 심한 것을 알 수 있다. 반면 CCC plot에서는 ρ의 전 범위에 걸쳐 변동이 거의 없음을 알 수 있다.

난수의 개수가 PCC, MI, CCC 값에 미치는 영향을 관찰하기 위하여 $\rho = 0, 0.5, 0.95$인 이변량 정규 분포 BN(μ, Σ)에서 난수의 개수를 10에서 1,000까지 증가시키면서 PCC, MI, CCC 값을 구하는 작업을 100번 반복하면 Figures 3–5를 그릴 수 있다. Figure 3에서 보는 것처럼 $\rho = 0$일 때 난수의 개수가 증가하면 PCC, MI, CCC 값 변동이 모두 작아지나 PCC > MI > CCC 순으로 변동이 작음을 알 수 있다. 난수의 개수가 600 이상이면 MI 값 변동과 CCC 값 변동이 거의 같아짐을 알 수 있다.

Figure 4에서 보는 것처럼 $\rho = 0.5$일 때 난수의 개수가 증가하면 PCC, MI, CCC 값 변동이 모두 작아지나 PCC > MI > CCC 순으로 변동이 작음을 알 수 있다.

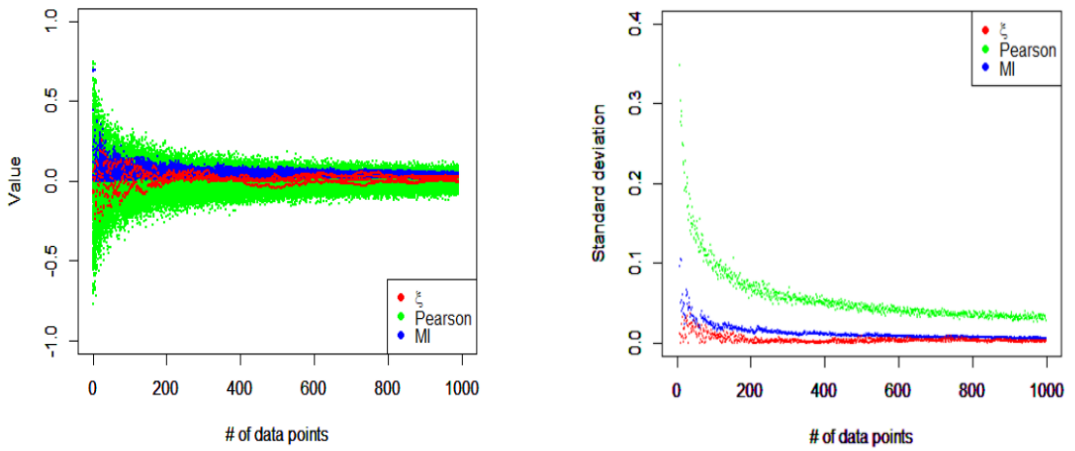

Figure 3: *Changes in measure values as the sample number size increases with bi-variate normal distribution ($\rho = 0$).*

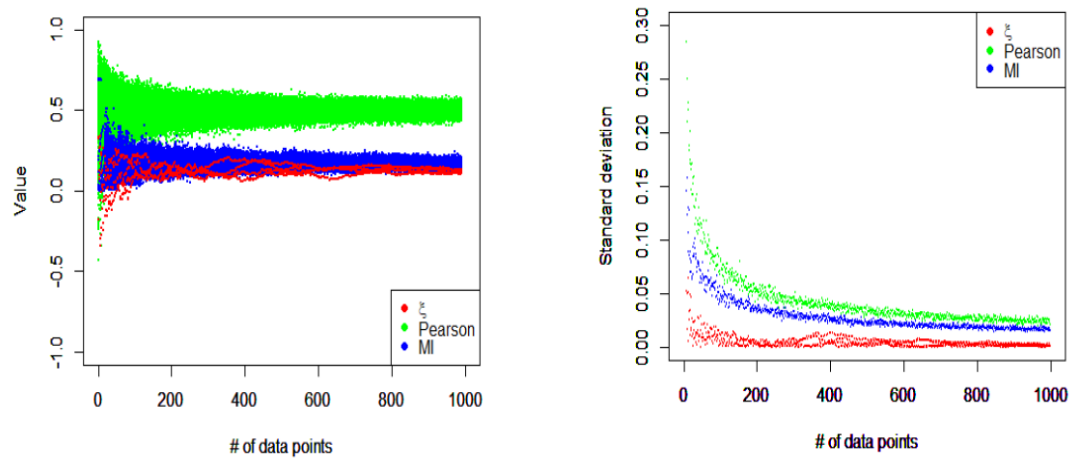

Figure 4: *Changes in measure values as the sample number size increases with bi-variate normal distribution ($\rho = 0.5$).*

Figure 5에서 보는 것처럼 $\rho = 0.95$일 때 난수의 개수가 증가하면 PCC, MI, CCC 값 변동이 모두 작아지나 MI > PCC > CCC 순으로 변동이 작음을 알 수 있다.

두 번째 이변량 분포로서 이변량 코시 분포를 고려해 보자. 이변량 코시 분포에서 난수를 추출하면 이상값이 자주 나타난다. 이변량 코시 분포에서 난수를 추출하기 위해 R library (fMultivar)를 사용하였다. Figure 6은 $\rho = 0.5$인 이변량 코시분포 하에서 난수 100개를 뽑아 그린 산점도이다. 이상값이 3개 나타남을 알 수 있다.

$\rho = -0.99$에서 $\rho = 0.99$까지 0.01씩 증가시키면서 이변량 코시 분포에서 난수 100개를 추출하여 PCC를 구하는 작업을 100번 반복하면 PCC plot를 그릴 수 있다. 같은 작업을 SCC, MI, CCC에 대하여 각각 시행하면

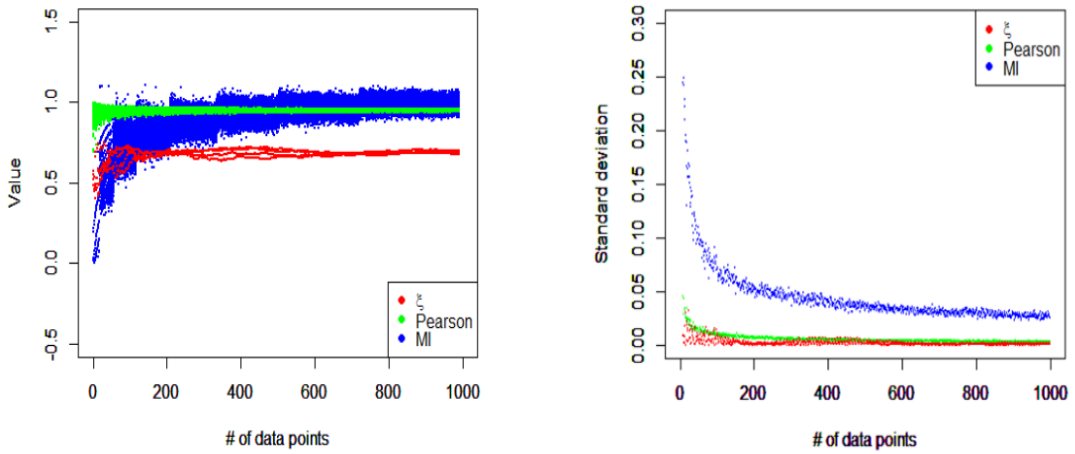

Figure 5: *Changes in measure values as the sample number size increases with bi-variate normal distribution ($\rho = 0.95$).*

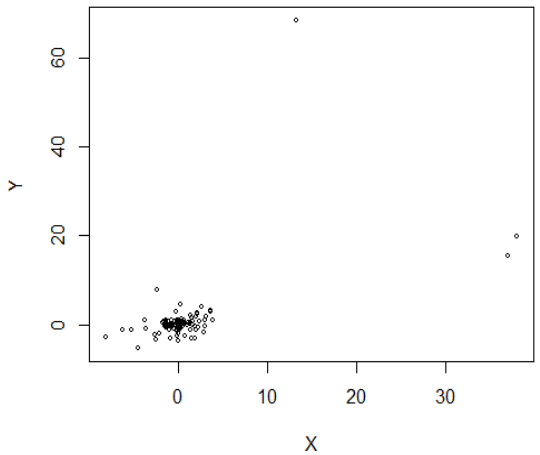

Figure 6: *Scatterplot with bi-variate Cauchy distribution ($\rho = 0.5$).*

SCC plot, MI plot, CCC plot를 그릴 수 있다. Figure 7은 이 4개의 그림을 나타낸다.

PCC plot에서는 ρ의 전 범위에 걸쳐 PCC값의 변동이 매우 심한 것을 알 수 있다. 소수의 이상값이 PCC값에 불안정한 영향을 주어 변동을 매우 크게 만든다. SCC plot에서는 ρ의 절대값이 1에 가까우면 SCC값의 변동이 거의 없으나 ρ의 절대값이 1에서 점점 작아지면 SCC값의 변동이 점점 커지다가 0에 가까우면 SCC값의 변동이 제일 큰 것을 알 수 있다. MI plot에서는 ρ의 전 범위에 걸쳐 MI값의 변동이 심한 것을 알 수 있다. 반면 CCC plot에서는 ρ의 전 범위에 걸쳐 CCC값의 변동이 거의 없음을 알 수 있다. 이변량 정규 분포 뿐만이 아니라 이변량 코시 분포에서도 ρ의 전 범위에 걸쳐 CCC값의 변동이 거의 없음을 알 수 있다.

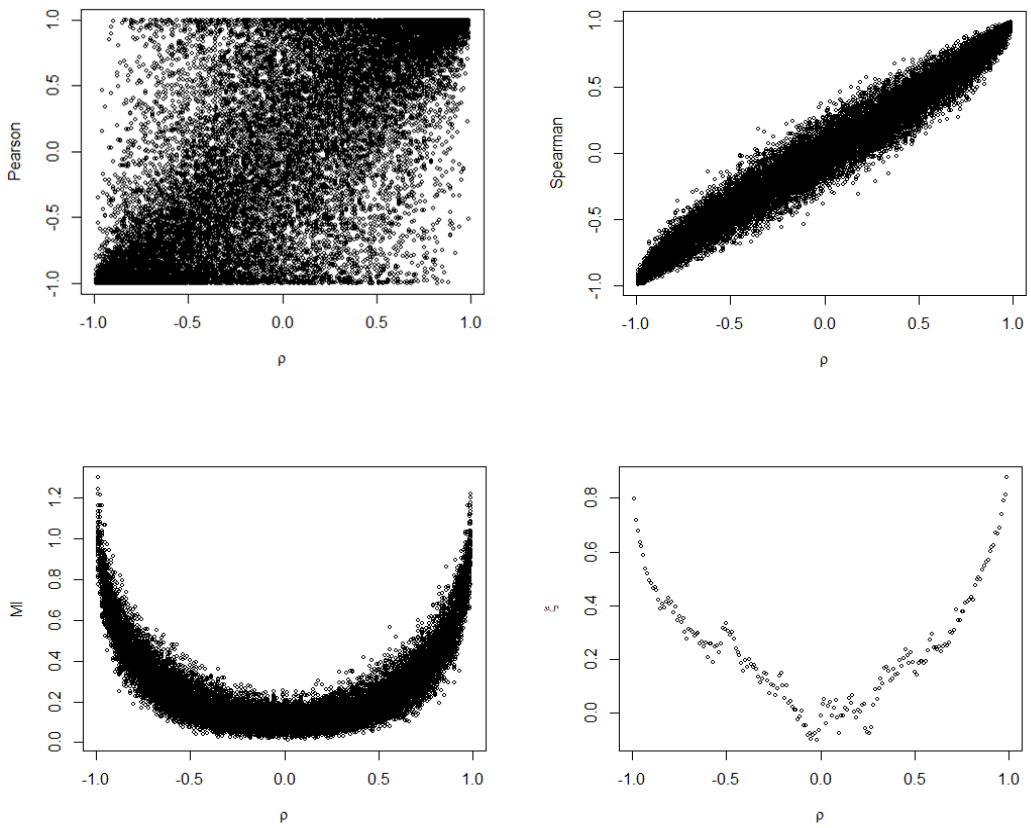

Figure 7: *PCC plot, SCC plot, MI plot, and CCC plot with bi-variate Cauchy distribution.*

Table 1: HDLSS datasets

Datasets	No. of features	No. of observations
Leukemia	7,129	72
Prostate	12,600	136
Colon	2,000	62
Breast cancer	22,215	118

2.2. HDLSS 자료에서의 Chatterjee 상관계수

HDLSS 자료는 고차원이지만 자료가 적은 자료를 말하며 HDLSS 자료는 유전자 데이터에서 많이 나타난다. 바이오마커 특성(features)의 수를 p라고 하고 관측값(환자)의 수를 n이라고 하면, 암 진단을 위한 자료는 통상 n보다 p가 큰 ($n < p$) 고차원자료 또는 n보다 p가 매우 큰 ($n << p$) 초고차원 자료이다.

PCC와 CCC의 관계를 연구하기 위하여 Table 1과 같은 4 종류의 HDLSS 자료를 선정하였다.

Dataset ‘leukemia’는 Golub 등 (1999)이 소개한 백혈병 자료이며 급성 골수성 백혈병(acute myeloid leukem

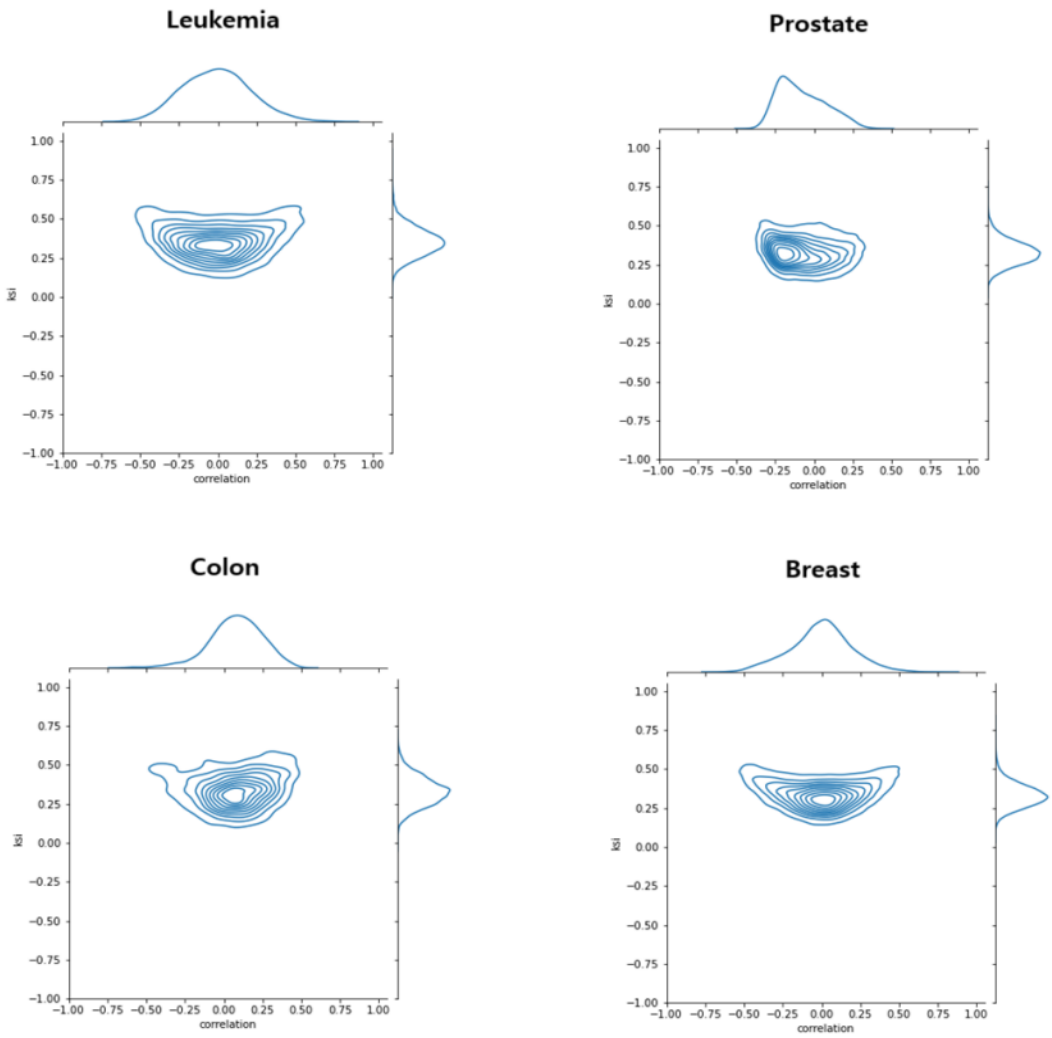

Figure 8: *Joint plots for 4 datasets.*

ia; AML) 및 급성 림프절성 백혈병(acute lymphoblastic leukemia; ALL) 환자들을 분류하는데 사용되었다. R 패키지 'SIS'에 내장되어 있다. 학습 자료와 테스트 자료가 각각 leukemia.train과 leukemia.test로 제공되며, leukemia train에는 27명의 급성 림프절성 백혈병과 11명의 급성 골수성 백혈병, leukemia.train에는 20명의 급성 림프절성 백혈병과 14명의 급성 골수성 백혈병으로 구성되어 있다. 설명 변수는 7,129개 유전자이고 반응 변수는 급성 골수성 백혈병을 1, 급성 림프절성 백혈병을 0으로 나타낸다.

Dataset 'prostate'는 Singh 등 (2002)이 소개한 전립선암 자료(prostate cancer data)이며, R 패키지 'SIS'에 내장되어 있다. 학습 자료와 테스트 자료가 각각 prostate.train과 prostate.test로 제공되어 있다. prostate.train에는 52명의 전립선 종양환자와 50명의 전립선 비 종양환자가 있으며, prostate.test에는 25명의 전립선 종양환자와 9명의 전립선 비 종양환자로 구성되어 있다. 설명 변수는 12,600개 유전자이며, 반응 변수는 전립선 비

종양을 1, 전립선 종양을 0으로 나타낸다.

Dataset ‘colon’은 Alon 등 (1999)이 제공한 대장암 자료(colon cancer data)이다. R 패키지 ‘datamicroarray’에 alon 자료로 내장되어 있다. 종양 조직 40명, 정상 조직 22명이고 설명 변수는 2,000개 유전자이다. 반응 변수는 대장암 환자를 1, 정상 환자를 0으로 나타낸다.

Dataset ‘breast cancer’는 Chin 등 (2006)이 제공한 유방암 자료(breast cancer data)이다. 유방암 환자 75명, 정상 환자 43명이고 설명 변수는 22,215개 유전자이다. 반응 변수는 유방암 환자를 1, 정상 환자를 0으로 나타낸다.

4개의 dataset 각각 반응 변수와 설명 변수들 간의 PCC와 CCC값을 구한 후 PCC와 CCC값 사이의 관계를 보기 위하여 이차원 KDE를 사용하여 Figure 8과 같은 결합도(joint plot)를 그려 보았다.

4개의 dataset 모두 원형이 아닌 좌우로 조금 길쭉하고 아래로 볼록한 형태를 이루고 있음을 알 수 있다.

3. 결론

Chatterjee의 ξ 계수가 기존의 피어슨상관계수의 단점을 극복하기 위한 좋은 대안이 될 수 있는지, 어떤 특징이 있는 지를 파악하기 위하여 CCC의 두 가지 응용(1. 이변량 분포, 2. HDLSS 자료)을 고려하여 보았다.

CCC는 다른 상관계수들에 비하여 상대적으로 이상값에 강건하고 변동이 적은 측도임을 탐색적으로 알 수 있었다.

References

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, and Levine AJ (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proceedings of the National Academy of Sciences USA*, **96**, 6745–6750.

Auddy A, Deb N, and Nandy S (2021). Exact detection thresholds for Chatterjee’s correlation, *Unpublished paper*, Available from: https://arxiv.org/abs/2104.15140(https://arxiv.org/pdf/2104.15140.pdf)

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

Chin K, DeVries S, Fridlyand J *et al*. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies, *Cancer Cell*, **10**, 529–541.

Golub TR, Slonim DK, Tamayo P *et al*. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, **286**, 531–537.

Lin Z and Han F (2021). On boosting the power of Chatterjee’s rank correlation, *Unpublished paper*, Available from: https://arxiv.org/abs/2108.06828(https://arxiv.org/pdf/2108.06828.pdf)

Sadeghi B (2022). Chatterjee correlation coefficient: A robust alternative for classic correlation methods in geochemical studies- (including “TripleCpy” Python package), *Ore Geology Reviews*, **146**, 104954.

Shi H, Drton M, and Han F (2022). On the power of Chatterjee’s rank correlation, *Biometrika*, **109**, 317–333.

Singh D, Febbo PG, Ross K *et al*. (2002). Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, **1**, 203–209.

Received February 15, 2023; Revised February 24, 2023; Accepted February 24, 2023

Chatterjee 상관계수의 응용

안소진[a], 장대흥 [1,a]

[a]부경대학교 데이터정보과학부 통계·데이터사이언스전공

요 약

Chatterjee (2021)는 기존의 피어슨 상관계수의 단점을 극복하기 위한 하나의 대안으로서 새로운 상관계수 ξ를 제안하였다. 이 상관계수는 순위를 기반으로 하기 때문에 이상점에 강건하고, 간단한 공식 때문에 개념을 이해하기 쉽고 측도 계산 속도가 아주 빠르다. 본 논문을 통하여 이 상관계수의 응용을 두 가지 측면(1.이변량 분포, 2. 고차원자료)에서 살펴보았다.

주요용어: Chatterjee 상관계수, 이변량 분포, 고차원자료

이 논문은 부경대학교 자율창의학술연구비(2021년)에 의하여 연구되었음.

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