Analyzing the fine structure of distributions 요약

2022.04.06. 예지혜

[Abstract]
데이터의 구조를 이해할 때, 데이터가 하나의 과정에서 발생되었는지, 여러 다른 과정을 거친 데이터가 합쳐진 것인지 판단하는 것은 중요한 관심사이다. 단변량 분포를 시각화하는 방법으로 히스토그램, ridgeline plot, bean plot, violin plot 등이 있지만 이러한 방법은 multimodal, skewed 분포 등에서 문제가 발생할 수 있다. 따라서 이 논문은 새로운 시각화 툴인 MD plot을 제안한다. 이 방법은 분포 추정을 위한 파라미터가 필요없어 비전문가도 쓰기 좋다.

[Introduction]
기존의 방법
- univariate : QQ plot, 히스토그램, cdf, pdf
- many features : box plot, violin plot, bean plot, ridgeline plot

exploratory statistics에서 분포 추정은 꽤 어려운 문제이고, 다양한 파라미터를 설정해줘야 하기 때문에 많은 비전문가들은 디폴트 옵션을 사용한다. 반면, MD plot은 PDE(파레토 분포 추정) 방법에 기반하기 때문에 파라미터가 필요하지 않다.

[Methods]

<Performance comparison (시각화의 성능 비교 방법)>
1. 샘플링을 통해 어떠한 특징을 갖는 인위적인 데이터를 발생시킨다.
이때, 특징마다 샘플의 사이즈가 다른데, skewness와 bimodality를 만족하는 데이터를 위해서는 해당 통계적 검정을 위한 maximum size가 사용되고, uniform 분포를 위해서는 minimum size가 사용된다. (여기서 샘플 사이즈의 범위는 269부터 31000이다.) 이 범위 내에서는 다양한 방법을 비교할 때 그 결과가 바뀌지 않기 때문이다.
2. 각 특징에 대한 통계적 검정을 진행한다.
multimodality를 위한 sensitivity는 Hartigan's dip statistic을 사용한다. skewness를 위해서는 D' Agostino test of skewness를 사용한다.
3. 특징을 알지 못하는 새로운 데이터셋에 대해 다양한 요소들을 탐색한다. 이 과정을 통해 각 방법의 문제점들을 이해한다. 잘 알지 못하는 데이터를 탐색할 때 적절한 파라미터를 설정하는 것은 굉장히 중요하므로, 디폴트 파라미터로도 각 방법이 특징을 잘 파악해내는지 시험한다.

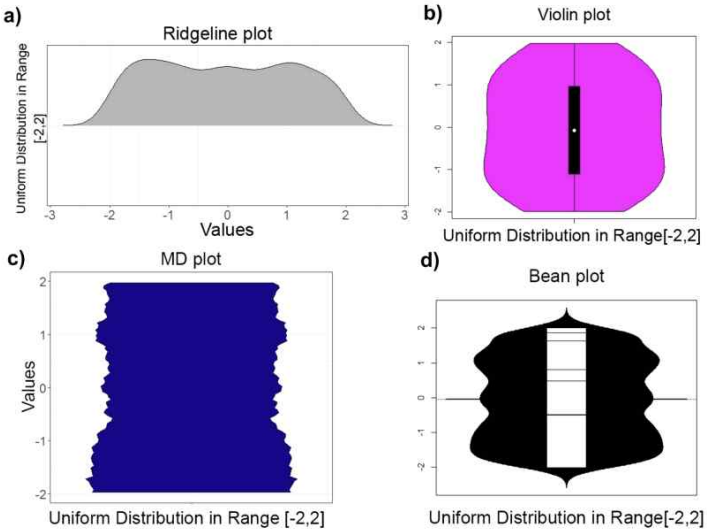
Comparing visualizations is challenging because they have the same problems as the estimation of quantiles or clustering algorithms such as k-means or Ward: they depend on the specific implementation. [18-21]

〈Visualization tools〉
1. Finite mixture models
가우시안과 같은 파라미터 함수의 중첩을 찾아낸다.
2. Variable kernel methods
커널 방법은 로컬 근사를 통해 분포를 추정하는데, variable 커널 방법은 로컬을 결정할 radius를 조절할 수 있다. (전체 범위에서 서로 다른 반경을 사용)
3. Uniform kernel algorithm
이 커널 방법은 고정된 global radius를 사용한다. 예를 들어 히스토그램은 전 범위에서 동일한 binwidth를 사용한다.

vaseplot, boxplot, bean plot 등..
violin plot은 고정된 반경을 이용하는 smooth 커널 함수로, nonparametric 분포 추정을 사용한다.
ridgeline plot은 부분적으로 중첩된 line plot이다.

〈Mirrored Density plot - MD plot〉
Pareto density estimation (PDE)에서는, hypersphere의 반경을 information theoretic ideas에 따라 적절히 선택한다. 즉, 최소한의 부피를 이용해 최대한의 정보를 담을 수 있는 반경을 선택한다. (만약 hypersphere가 데이터의 20%를 담고 있으면 80%의 정보를 담고있는 셈이다.)
이러한 PDE 분포 추정 방법을 PDF 미러링과 결합시킨 것이 MD plot이다.
MD plot은 로그 변형, 큰 데이터 자동 샘플링, 정규성 검정 등 분포에 대한 다양한 탐색을 가능하게 한다. 또한, 만약 통계적 검정 결과 가우시안 분포가 맞다면 그에 상응하는 평균과 분산을 가진 정규분포 plot을 함께 보여줌으로써 통계적 가설 검정이 잡아내지 못한 non-Gaussian의 가능성을 확인하게 해준다. (If all tests agree that a feature is Gaussian distributed, then the plot of the feature is automatically overlaid with a normal distribution of robustly estimated mean and variance equal to the data. This step allows the marking of possible non-Gaussian distributions of single feature investigations with a quantile-quantile plot in cases where statistical testing may be insensitive.)
MD plot은 두 가지 threshold를 적용하는데, 먼저 unique data의 최소량(minimal amount of unique data)을 정의해 그 이하의 데이터는 분포 추정을 하지 않고 산점도로 시각화한다. 또 하나는 값의 최소값(minimal amount of values in data)인데, 이보다 작은 데이터는 산점도로 표현한다. 이 두 가지 한계점은 missing data나 알고리즘에 의한 quantized error를 잡아내는데 도움이 된다.

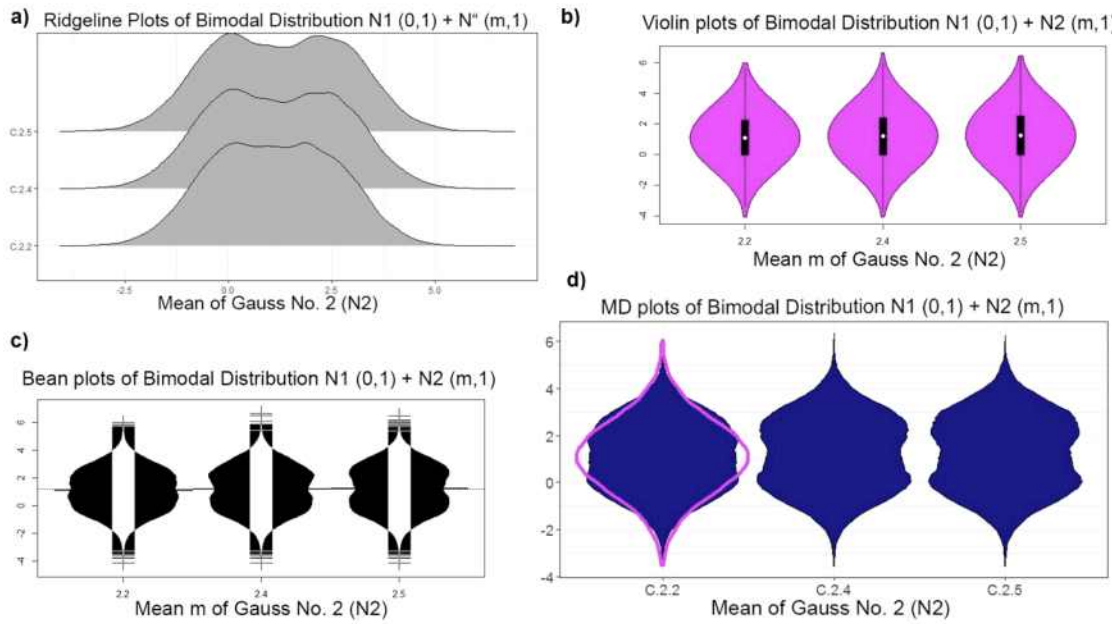
[Results]



먼저 유니폼 분포를 다양한 방법으로 시각화한 결과, ridgeline plot, histogram, bean plot은 multimodality를 보였고, ridgeline, bean, violin plot은 양쪽 끝 부분이 좁아지는 형태를 보였다. 반면 MD plot은 구간 내에서 평평한 형태를 띠며 분포를 잘 나타내었다. 실제로, 가설 검정 결과 unimodal에 not skewed로 샘플 자체에는 문제가 없었다.

실험 1. Multimodality vs. Unimodality

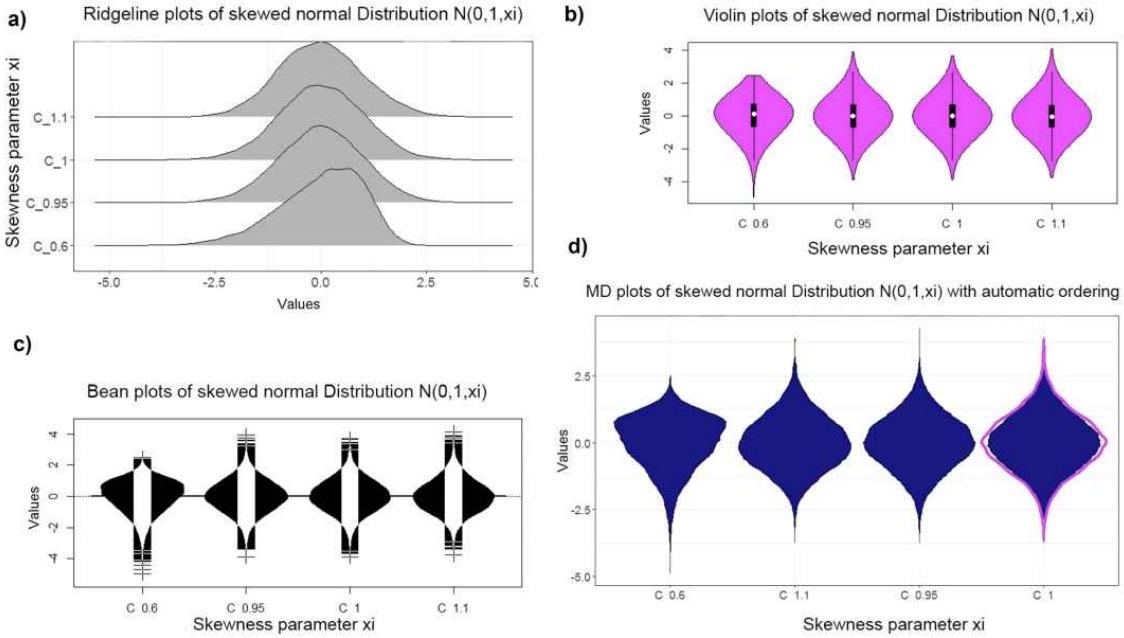
가우시안 분포 2개로 실험을 진행한다. 하나는 mean = 0 으로 고정하고, 하나는 mean을 계속 바꿔가며 실험하였다. Hartigans' Dip test에서는 2.4부터 통계적으로 차이가 있다는 결과를 보였다.



violin plot은 식별해내지 못했고, 나머지는 2.4부터 multimodality를 보였다. 다만 MD plot에

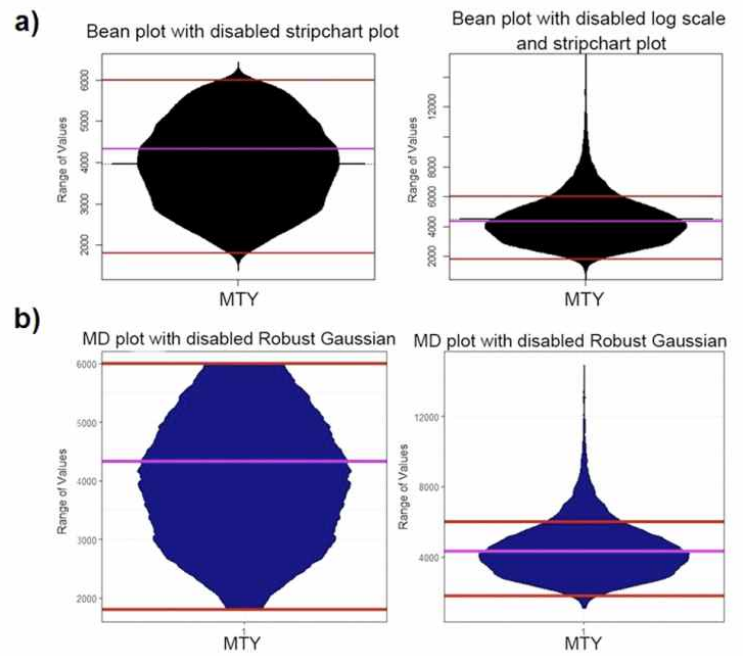
서는 2.2일 때 통계적으로 unimodal이라 그에 해당하는 정규분포를 함께 시각화해주었는데, 둘을 비교해보면 실제로는 unimodal이라 보기 어려움을 확인할 수 있다.

실험 2. Skewness vs. normality



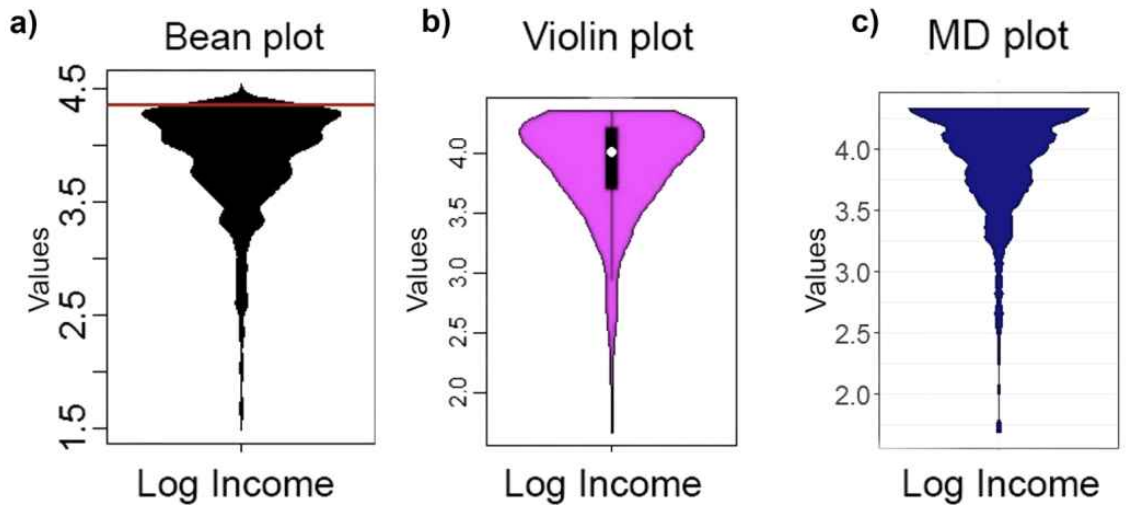
skewed 정규분포를 샘플링하여 각 방법들을 비교해보았다. D' Agustino 검정에 따르면 [0.95, 1.05] 밖에서의 skewness가 통계적으로 유의하다. 시각화 결과 violin plot을 제외하곤 해당 구간에서 skewness가 확인된다.

실험 3. Data clipping vs. heavy-tailedness



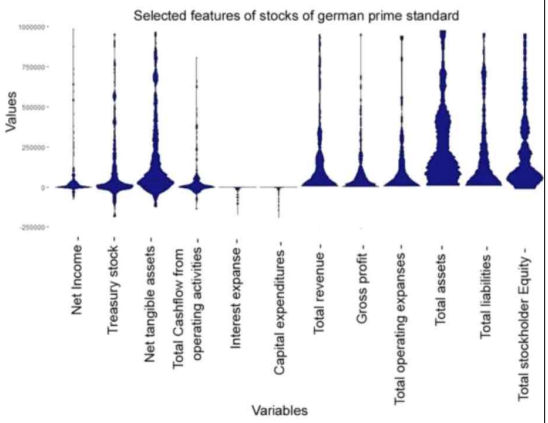
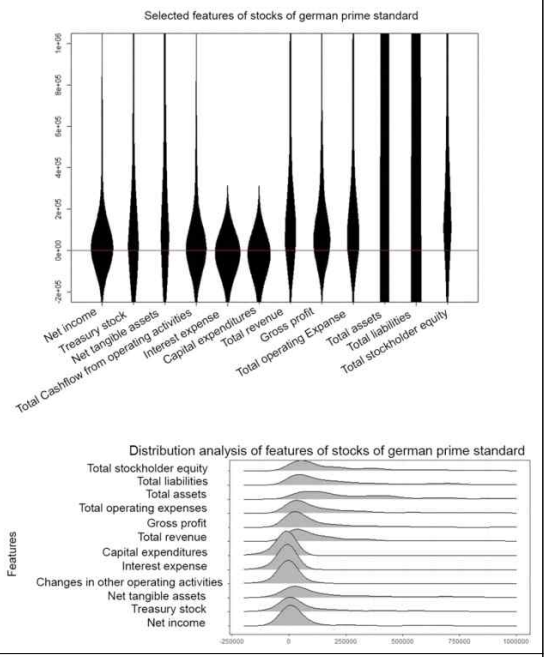
오른쪽 데이터를 빨간 선으로 clipping한 데이터를 시각화한 것이 왼쪽 plot이다. bean plot은 잘린 범위 밖으로도 존재하지 않는 데이터를 시각화하며 범위 내부는 낮게 추정하였다. 반면, MD plot은 정확하게 시각화하였다.

실험 4. multimodality, skewness, data clipping 결합

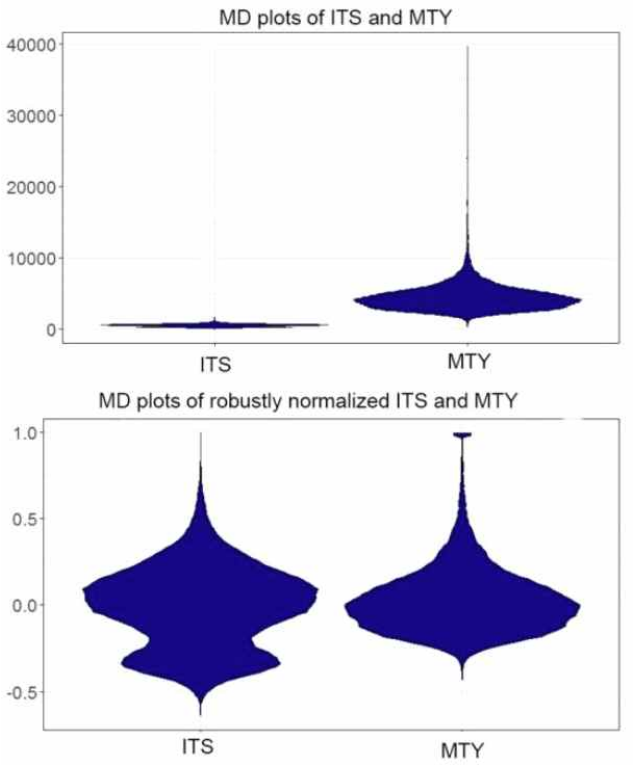


- violin plot은 md plot보다 덜 skewed된 형태
- violin, ridgeline, bean plot은 4~4.5에서 최빈값이 나타나는 반면 md plot과 히스토그램을 보면 4.35가 최댓값이다. 즉, 나머지 방법들은 최댓값을 넘어서서 시각화한 것이다.

실험 5. 분포의 시각적 탐색

MD plot	bean plot & ridgeline plot
	
- 0 근처 hard cut - highly skewed - multimodal 요약통계량과 발견된 특성들이 일치한다. concave ordering이 가능하다.	- unimodal이나 uniform으로 바꿔서 시각화하였다. (bean plot) - 0 근처에서 뚝 끊기는 부분이 없다. (bean, ridgeline plot)

실험 6. feature 간 범위가 매우 상이할 때



그냥 시각화하면 위 쪽의 그래프처럼 좁은 범위의 변수는 확인하기 어렵다. md plot에서는 robust normalization 옵션을 이용하여 주요 특성들을 변화시키지 않고 시각화시킬 수 있다.

[Discussion]

MD plot은 모든 문제 상황을 해결할 수 있는 유일한 시각화 도구이다.

- 데이터가 클 때는 다른 방법들도 bimodality와 skewness를 잘 잡아내지만 데이터가 적을 때는 MD plot이 훨씬 우수하다. (실험 5)
- robust estimated Gaussian을 겹쳐 시각화하는 것은 통계적 검정보다도 더 민감하게 데이터를 해석할 수 있게 한다. (실험 1)
- feature들의 automatic ordering은 skewness를 더 잘 보이게 한다. (실험 5)

등등

MD plot은 10^5 정도의 데이터 크기까지는 충분히 진행할 수 있다. 분포 추정과 반경 계산을 각각 진행해야 하므로 컴퓨팅 계산량도 늘어난다. 10^5 이상의 큰 데이터셋에서는 subsampling을 통해 사이즈를 줄인다. 50 이하의 작은 데이터셋에서는 분포 추정을 하지 않고 산점도로 나타낸다.

MD plot은 multimodality 탐지를 위해 고안된 방법이기에 예민하게 잡아내는 것을 고려해야 한다. 간혹 overestimate 할 수 있다.