The Korean Journal of Applied Statistics (2018) DOI: http://dx.doi.org/10.5351/KJAS.2018.31.4.517
31(4), 517–528

Detecting outliers in multivariate data and visualization-R scripts

Sung-Soo Kim[a,1]

[a]Department of Information Statistics, Korea National Open University

(Received June 28, 2018; Revised August 1, 2018; Accepted August 1, 2018)

Abstract

We provide R scripts to detect outliers in multivariate data and visualization. Detecting outliers is provided using three approaches 1) Robust Mahalanobis distance, 2) High Dimensional data, 3) density-based approach methods. We use the following techniques to visualize detected potential outliers 1) multidimensional scaling (MDS) and minimal spanning tree (MST) with k-means clustering, 2) MDS with fviz_cluster, 3) principal component analysis (PCA) with fviz_cluster. For real data sets, we use MLB pitching data including Ryu, Hyun-jin in 2013 and 2014. The developed R scripts can be downloaded at "http://www.knou.ac.kr/∼sskim/ddpoutlier.html" (R scripts and also R package can be downloaded here).

Keywords: potential outliers, visualization, Mahalanobis distance, multidimensional scaling, minimal spanning tree, principal component analysis

1. 서론 및 개발 배경

다변량 자료의 분석에서 가장 먼저 접하는 문제 중의 하나는 특이점(outlier)의 검출이라고 할 수 있다. 특이점은 그 자체를 밝히는 것도 중요한 의미를 지니고 있을 뿐만 아니라, 자료 자체의 문제점 파악, 데이터 모델링에 미치는 영향 등을 고려할 때 특이점을 검출하는 것은 데이터 분석의 기본이라고 할 수 있다. 여기서 특이점은 분석의 방향에 따라 여러 가지로 정의 될 수 있지만, 일반적으로 인용되는 정의로서는 "an outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism (Hawkins, 1980)", 또는 "an outlying observation, or outlier, is one that appears to deviates markedly from other members of the sample in which it occurs (Barnett와 Lewis, 1994)"과 같이 대부분의 다른 관찰점들과는 다른 형태를 보이는 관찰점을 특이점으로 정의할 수 있다.

변수가 하나인 단변량 자료나 변수가 둘인 이변량 자료에 있어서의 특이점의 검출은 상자그림, 히스토그램, 산점도, 이변량 상자그림(bivariate boxplot) (Rousseeuw 등, 1999) 등을 이용하여 쉽게 밝혀낼 수 있지만, 다변량 자료의 경우에는 데이터의 구조를 그래프로 시각화하는 것이 쉽지 않기 때문에 상대적으로 어려운 것이 사실이다. 다변량 자료에서 특이점을 검출하는 방법을 크게 분류하면, 1) Mahalanobis

This research was supported by Korea National Open University Research Fund in 2015.
[1]Department of Information Statistics, Korea National Open University, Daehakno 86, Jongnogu, Seoul 03087, Korea. E-mail: sskim@knou.ac.kr

거리를 이용하여 검출하는 통계적인 방법 (Mahalanobis, 1936), 2) 로버스트 PCA(robust principal component analysis) 기반 방법, 3) 비모수적 방법으로 거리 기반 검출 방법(distance-based methods), 밀도-근거 방법(density-based method), 군집분석 접근 방법(clustering-based methods) 등으로 나눌 수 있다 (Kriegel 등, 2010; Pamula 등, 2011; Jayakumar와 Thomas, 2013). 이러한 방법들은 데이터의 분석, 데이터의 구조 등에 따라 효율성이 다르므로, 실무에 있어서는 어느 한 방법만을 쓰는 것보다는 여러 방법을 이용하여 잠재적 특이점(potential outliers)을 검출하고, 검출된 특이점이 데이터의 구조에서 어떠한 위치에 있는 지, 어떠한 영향을 미칠 수 있는 지 등을 시각적으로 검토할 필요가 있다.

개발된 R 스트립트는 R 시스템에서 제공되고 있는 다변량 검출 방법을 활용하여 여러 관점에서 잠재적 특이점을 밝히고, 이를 시각화를 이용하여 살펴보는 시스템이다. 제공되는 R 스크립트는 크게 다변량 특이점 검출 함수, 다변량 특이점 시각화 함수로 구분된다. 특이점 검출 함수로는 1) Mahalanobis 거리 이용방법, 2) 다차원자료에서 PCA 기반 검출방법, 3) 밀도-근거 접근 방법으로 구분하여 제공하고, 특이점 시각화와 연결하여 데이터 구조를 나타내기 위한 방법으로 K-means 군집방법을 이용하였다. 시각화 방법으로는 multidimensional scaling (MDS)를 이용하거나, PCA를 이용한 시각화를 제공하였다. 특히 MDS를 이용한 방법으로는 minimal spanning tree (MST)와 연결한 방법, ggplot2 (Wickham, 2010) 접근 방식을 이용한 fviz_cluster (R 패키지 factoextra의 시각화 기능을 수행하는 함수) 시각화와 연결하여 제공한다. 또한 다이내믹 시각화 관점에서 특이점을 grand tour에 연결하여 보여주는 자바 프로그램도 4장 결론 및 제언에서 소개한다. 여기서 이용한 자료는 미국 Major League Baseball (MLB) 투수자료 중에서 류현진 선수가 적극적으로 활동했던 2013년, 2014년 자료를 이용하여 살펴보도록 한다.

2. 다변량 자료에서 특이점 검출

앞에서 소개한 것처럼 다변량 자료에서 특이점을 검출하는 다양한 방법들은 데이터의 구조, 속성 등에 따라 특이점의 검출 관점이 다르고, 어느 하나의 방법이 다른 방법보다 더 효율적이라고 할 수 없다 (Penny와 Jolliffe, 2001; Jayakumar와 Thomas, 2013). 따라서 다양한 관점에서 여러 특이점을 검출하고, 이를 비교, 검토할 필요가 있다. 이를 위해서는 특이점 검출을 효율적으로 하기 위한 종합 분석시스템이 요구된다.

개발된 R 스크립트는 다변량에서 특이점을 검출하기 위한 함수를 효율적으로 활용할 수 있도록 하였다. 또한 검출된 특이점을 시각화와 연결하여 살펴볼 수 있도록 하였다. 특이점 검출을 위한 사례로서 미국 MLB 자료(http://mlb.mlb.com/stats에서 Pitching을 선택) 중에서 류현진이 적극적으로 활동했던 2013년, 2014년 투수자료를 예로 들어 살펴보도록 한다. 이용한 자료는 방어율(earned run average; ERA) 상위 50명을 추출한 자료로서 Table 2.1, Table 2.2와 같다 (http://www.knou.ac.kr/∼sskim/ddpoutlier.html에서 다운).

자료를 읽은 후, 특이점 검출을 위한 R 스크립트를 실행해보도록 하자.

```
> pitch2013 <- read.table("c:/data/pitch2013.txt", header=T)
> name2013 <- pitch2013[, "name"]            # pitcher's name
> pitch2013 <- pitch2013[, -c(1,2)]          # raw data
> source("c:/pgm/DDPoutlier(Ver1-1).r")      # Run R scripts
> tpitch2013 <- standardfn(pitch2013)        # standardization: 0-1 or Zscore
> out2013 <- mvpoutliers(tpitch2013)
```

Table 2.1. MLB pitcher data in year 2013 (part)

Num	Name	W	L	ERA	GS	IP	H	R	ER	HR	BB	SO	AVG	WHIP
1	Kershaw	16	9	1.83	33	236.0	164	55	48	11	52	232	0.195	0.92
2	Fernandez	12	6	2.19	28	172.2	111	47	42	10	58	187	0.182	0.98
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
13	Ryu	14	8	3.00	30	192.0	182	67	64	15	49	154	0.252	1.20
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	Tillman	16	7	3.71	33	206.1	184	87	85	33	68	179	0.241	1.22

Table 2.2. MLB pitcher data in year 2014 (part)

Num	Name	W	L	ERA	GS	IP	H	R	ER	HR	BB	SO	AVG	WHIP
1	Kershaw	21	3	1.77	27	198.1	139	42	39	9	31	239	0.196	0.86
2	Hernandez	15	6	2.14	34	236.0	170	68	56	16	46	248	0.200	0.92
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
33	Ryu	14	7	3.38	26	152.0	152	60	57	8	29	139	0.257	1.19
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	Harang	12	12	3.57	33	204.1	215	88	81	15	71	161	0.273	1.40

먼저 자료의 표준화를 시행하고(본 사례분석에서는 0-1변환(min-max scaling 또는 min-max normalization; $\{x - \min(x)\}/\{\max(x) - \min(x)\}$으로 계산)을 선택함), 특이점 검출을 위한 함수 mvpoutliers()를 실행한다. 특이점 검출 선택 메뉴에서 mvoutliers::aq.plot 함수를 선택한 결과는 다음과 같다.

```
-------------------------------------------------------------------------
 <A> Detecting Outliers using (Robust)Mahal. distance.
    ---------------------------------------------------------------------
     1. mvoutliers::aq.plot         2. mvoutliers::dd.plot
     3. chemometrics::Moutlier      4. rrcovHD::OutlierMahdist
    ---------------------------------------------------------------------
 <B> Detecting Outliers in High Dimensions
     5. mvoutliers::pcout           6. rrcovHD::OutlierPCOut
    ---------------------------------------------------------------------
 <C> Detecting Outliers usig density-based approach
     7. DMwR::lofactor              8. fpc::dbscan
    ---------------------------------------------------------------------
     99. Exit(Result of pre-selected method is returned)
 -------------------------------------------------------------------------
  Select : The number is ---- : 1

Potential Outliers :  1 2 41 3 43 16 4 28 5 6 50 21 11
Robust Mahal. distance of potential outliers
[1] 15.703 15.029 11.624 11.560 11.206 9.317 8.829 8.474 7.828 6.598 6.302 6.237
```

```
    5.457
> poutlier2013 <- out2013$poutlier
> poutlier2013
 [1]  1  2  3  4  5  6 11 16 21 28 41 43 50
```

실행 결과에서는 잠재적 특이점을 거리가 큰 순서로 보여주고, 이에 대응되는 로버스트 Mahalanobis 거리(aq.plot을 수행하면 Mahalanobis distance를 제공하지 않는다. 특이점들의 거리를 보면 상대적인 강도를 파악하는데 도움이 되기에 aq.plot을 약간 수정하였다.)를 보여주고 있다. aq.plot에서는 디폴트로 평균과 공분산의 추정을 minimum covariance determinant (MCD) 방법(기타 추정방법으로는 minimum volume ellipsoid (MVE) 방법이 있다. 일반적으로 MCD 방법이 더 좋은 것으로 알려져 있다 (Butler 등, 1993; Rousseeuw와 Van Drissen, 1999))을 이용하고, 특이점의 결정을 위한 Cutoff 값은 $\max(\sqrt{\text{Adjusted quantile}}, \sqrt{\chi^2(0.975, \text{df})})$을 이용한다. Adjusted Quantile은 Filzmoser (2004)에 자세히 설명되어 있다. 검출되는 특이값은 로버스트 거리를 구할 때 랜덤표본을 사용하기 때문에 시행시마다 결과가 다르게 된다.

다음 분석에 들어가기 전에 개발된 R 스크립트에서 사용하는 특이점 검출 함수들을 간략히 살펴보자.

- mvoutliers::dd.plot : 특이점, Mahalanobis 거리 및 로버스트 Mahalanobis 거리를 반환해준다. Cutoff 값은 $\min(\sqrt{\text{Adjusted quantile}}, \sqrt{\chi^2(0.975, \text{df})})$을 이용한다.
- chemometrics::Moutlier : 특이점, Mahalanobis 거리 및 로버스트 Mahalanobis 거리를 반환해준다. Cutoff 값은 $\sqrt{\chi^2(0.975, \text{df})}$을 이용하고 있다.
- rrcovHD::OutlierMahdist : 로버스트 Mahalanobis 거리를 반환한다. 반환 객체는 가상 클래스 (virtual class) Outlier의 부클래스(subclass)로서 특이점 및 거리를 구하는 방법은 다음과 같다.

```
out.rrd <- OutlierMahdist(kdata)
poutlier <- getOutliers(out.rrd)
pout.dist <- getDistance(out.rrd)
```

- mvoutliers::pcout : 데이터가 방대하고 변수가 많은 고차원 데이터에서 주성분분석 방법을 이용하여 특이점을 검출한다 (Filzmoser 등, 2008).
- rrcovHD::OutlierPCOut : 데이터가 방대하고 변수가 많은 고차원 데이터에서 주성분분석 방법을 이용하여 특이점을 검출한다 (Filzmoser 등, 2008). 특이점 및 거리를 구하는 함수는 rrcovHD::Outlier Mahdist와 같다.
- DMwR::lofactor : 각 관찰점에 대한 local outlier factor (LOF) 값을 구한다 (Breunig 등, 2000). LOF값을 순서대로 정렬하여 상위 케이스를 잠재적 특이점으로 정한다(디폴트로 10개 케이스를 선택함).
- fpc::dbscan : Density-based clustering (DBSCAN) (Ester 등, 1996)을 실시한다. 이웃한 케이스 수가 작은 케이스를 특이점으로 분류한다. DBSCAN의 모수에서 최적 eps값을 정하기 위한 kNN 그림도 제공한다.

3. 특이점 시각화

앞 장에서 검출된 잠재적 특이점이 다른 관찰점에 비해 어떠한 위치에 있는 가를 파악하기 위해서는 주성분분석 그림에 특이점을 표시하거나, 또는 다차원 척도(multidimensional scaling)에 특이점을 나타

내는 방법이 이용될 수 있다. 여기에 데이터의 구조를 파악하기 위한 탐색적 방법의 하나로서 K-평균 군집분석과 연결한다면 시각화의 효과가 가중된다.

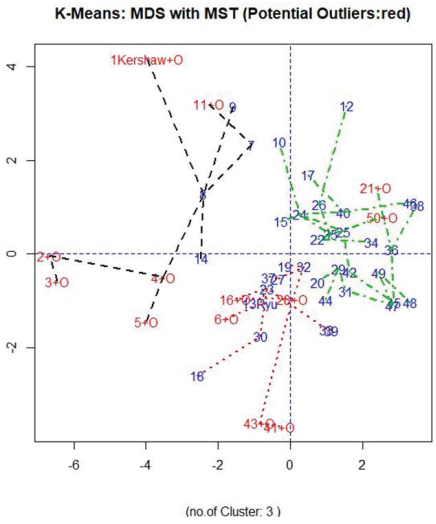
R 스크립트에서 제공하는 시각화 함수 displayoutliers()의 실행 절차와 결과는 다음과 같다.

```
  > id2013 <- seq(1, nrow(tpitch2013))
  > id2013[1] <- paste(id2013[1], name2013[1], sep="")      # Kershaw
  > id2013[13] <- paste(id2013[13], name2013[13], sep="")   # Ryu
  > displayoutliers(tpitch2013, poutlier2013, id2013)
  Type the number of clusters
  # of Clst(2-49)(To stop, only RETURN): 3
===  K-means result ==
   Cluster Size =  10 15 25
   Cluster num =  3
   SSB/SST    =  0.3891
 (G= 1 ) 1Kershaw 2 3 4 5 7 8 9 11 14
 (G= 2 ) 6 13Ryu 16 18 19 23 27 28 30 32 33 37 39 41 43
 (G= 3 ) 10 12 15 17 20 21 22 24 25 26 29 31 34 35 36 38 40 42 44 45 46 47 48 49 50
 Potential outliers :  1 2 3 4 5 6 11 16 21 28 41 43 50
```

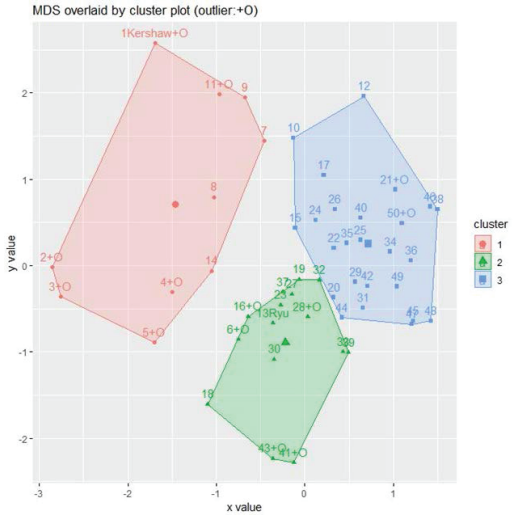
처음 세 줄은 우리의 관심이 류현진 투수이고, 또한 류현진과 더불어 잘 아는 투수가 Kershaw이므로 케이스 표시에 이를 추가하였다. 디폴트는 케이스 번호이다. 실행 결과는 먼저 K-평균 군집분석 결과를 보여주고, 차례대로 Figure 3.1의 세 개의 그래프를 보여준다. 여기서 K-평균 군집분석은 Hierarchical K-군집 방법 (Kassambara, 2017)으로, Ward 방법을 이용하여 초기군집 중심을 구한 후, K-평균 군집을 할 수 있도록 하였다. 함수 displayoutliers()에서는 초기 군집 중심을 구하기 위하여 랜덤표본을 추출하는 비율을 정할 수 있다. 디폴트는 100으로 전체 자료를 이용한다는 의미이다.

Figure 3.1은 2013년도 자료에서 군집 수가 3인 경우의 특이점을 나타낸 시각화 그림이다. 첫 번째는 MDS 그림에 K-means 군집 결과를 연결한 것이다. 특이점은 빨간색으로 번호 뒤에 +O 표시를 하여 다른 케이스와 구분하였다. 연결된 군집은 K-means 군집을 수행한 후, 같은 군집에 속한 케이스를 MST (Prim, 1957; Kim 등, 2000)로 연결하였다. MST는 각 케이스와 거리가 가까운 케이스를 연결하면서 동시에 속성이 유사한 케이스를 찾는데 도움이 된다. 여기서 군집 1은 Kershaw가 소속된 군집으로 $(1, 2, 3, 4, 5, 11)$ 케이스가 특이점으로 검출된 것을 알 수 있고, 특별히 이 군은 뛰어난 투수군이라는 것을 알 수 있다. 우리가 관심이 있는 투수 류현진은 군집2에 속하면서 군집2의 중심에 있는 것으로 보아서 안정된 역할을 했다는 것으로 짐작할 수 있다. 여기서 류현진과 MST로 연결된 케이스를 보면 6번과 23번 투수이고, 따라서 이들이 류현진이 비슷한 속성을 지니고 있다고 할 수 있다. 군집 2의 경우에는 $(6, 16, 28, 41, 43)$ 케이스가 특이점으로 표시되어 있고, 특히 $(41, 43)$번째 케이스는 다른 케이스와 다른 속성을 지니고 있다는 것을 보여준다. 군집 3의 경우는 많은 투수가 소속된 군으로 $(21, 50)$번째 케이스가 특이점으로 표시되어 있지만 별다른 특성을 보이지 않는 것을 알 수 있다. 구체적인 사항은 야구 전문가와 더불어 검토할 사항이다.

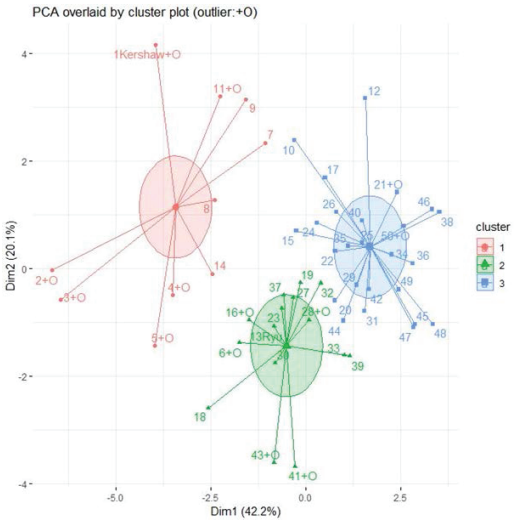
Figure 3.1에서 두 번째는 MDS 결과와 K-means 군집 결과를 ggplot2 속성을 지니고 개발된 factoextra::fviz_cluster를 이용하여 시각화한 그림이다. 객체 지향언어의 속성인 상속성을 이용하여 개발된 것으로 MDS 결과를 fviz_cluster 함수를 이용하여 시각화하기 위해서는상속성(inheritance)을 갖도록 하기 위하여 structure 함수(R 스크립트에서 fv.km <– structure(list(cluster=km.cluster), class =



(a) MDS with MST



(b) MDS with cluster

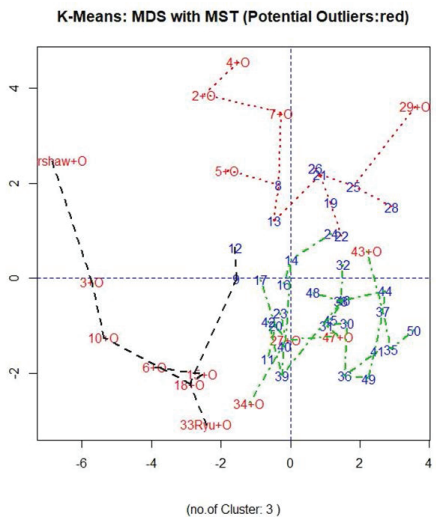


(c) PCA with cluster

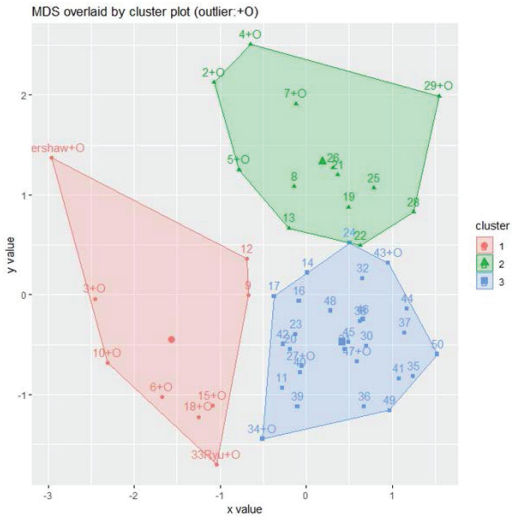
Figure 3.1. Outlier visualization with cluster size 3 (year 2013). MDS = multidimensional scaling; MST = minimal spanning tree, PCA = principal component analysis.

"kmeans") 참조)를 활용하면 된다. 각 군집을 MDS 결과와 연결하여 표현하면 군집의 타당성, 군집 중심의 위치, 각 특이점의 위치 등을 통하여 다른 케이스와의 관계, 특이점의 판단, 예를 들어 군집2의 경우 28번째 케이스 등의 특이점 판단 여부 등에 참고할 수 있을 것이다.

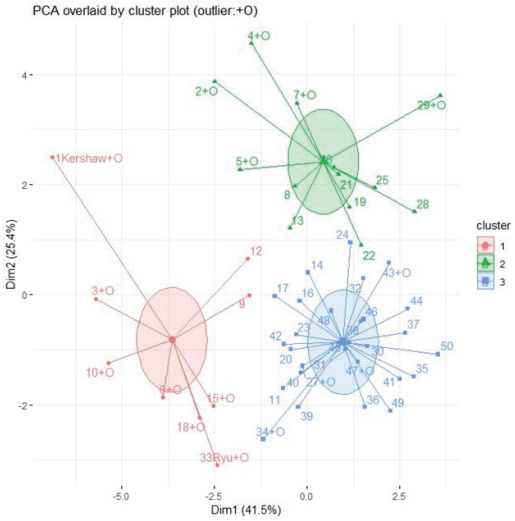
Figure 3.1에서 세 번째는 fviz_cluster에서 제공하는 기능을 이용한 것이다. 기본적으로 변수가 2개인 경우는 두 변수의 산점도를 이용하고, 세 변수 이상인 경우는 주성분 분석(principal component analysis; PCA)에서 (제1주성분, 제2주성분)을 산점도로 표시한다. 각 중심의 위치와 케이스를 연결하여 군



(a) MDS with MST



(b) MDS with cluster

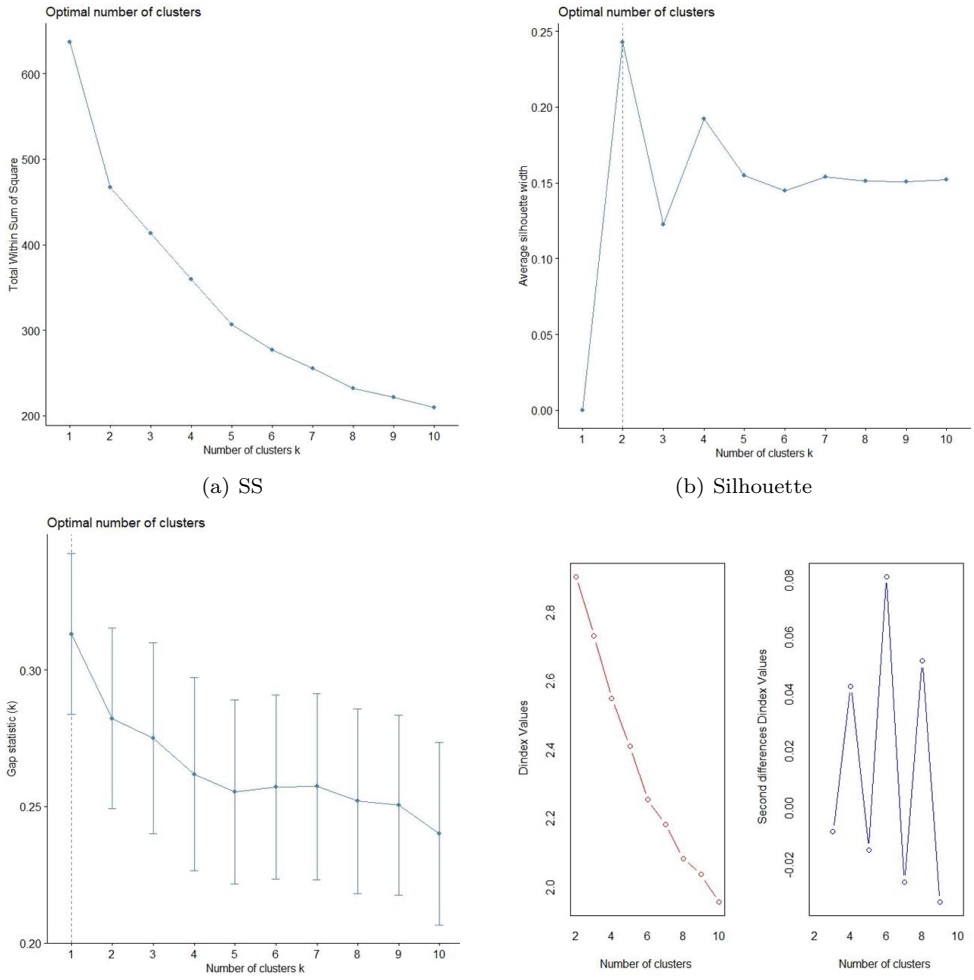


(c) PCA with cluster

Figure 3.2. Outlier visualization with cluster size 3 (year 2014). MDS = multidimensional scaling; MST = minimal spanning tree, PCA = principal component analysis

집의 타당성, 상대적 위치 등을 파악할 수 있다. 군집1의 Kershaw, 군집2의 Ryu의 위치와 연결하여 살펴보면 투수 류현진의 활동 위치, 위상 등을 짐작할 수 있다.

군집 수를 4개, 5개, 6개로 변화시켜가면서 살펴보자(결과는 지면 관계로 생략). 적정 군집 수의 결정, 군집의 타당성, 특이점의 상대적 속성 등을 파악하는데 도움이 될 것이다. 참고로 Figure 3.2는 2014년도 MLB자료를 이용하여 잠재적 특이점을 검출하고, 시각화한 그림이다. 2013년도 비해 2014년도 류현진 투수의 위상이 어느 방향으로 바뀌고 있는 지도 짐작할 수 있다. 군집 수를 바꾸어 가면서 2013년



(a) SS (b) Silhouette

(c) Gap statistic (d) NbClust

Figure 3.3. Plots of Total of within SS (a), Silhouette (b), Gap statistic (c), NbClust (d).

도와 비교해보는 것도 재미있는 작업이다.

K-평균 군집은 군집 수가 미리 주어지고, 초기 군집 중심에 따라 결과가 달라질 수 있기 때문에 시각화에 앞서 적정 군집 수를 미리 파악하면 많은 도움이 된다. 이를 위하여 K-평균 군집분석에서 적정 군집 수를 파악하기 위한 함수, KmeansClustering()를 제공하였다. 여기서는 Mojena의 규칙 (Mojena, 1977; Mojena와 Wishart, 1980), Total with SS (Kassambara, 2017), Silhouette method (Rousseeuw, 1987), Gap statistics (Tibshirani 등, 2001), Nbclust (Charrad 등, 2014)의 결과를 이용하여 적정 군집 수의 범위를 분석자가 파악할 수 있도록 하였다 (Figure 3.3). KmeansClustering()의 예는 다음과 같다.

```
> KmeansClustering(tpitch2013)
 < Process to determine the optimal number of clusters >
```

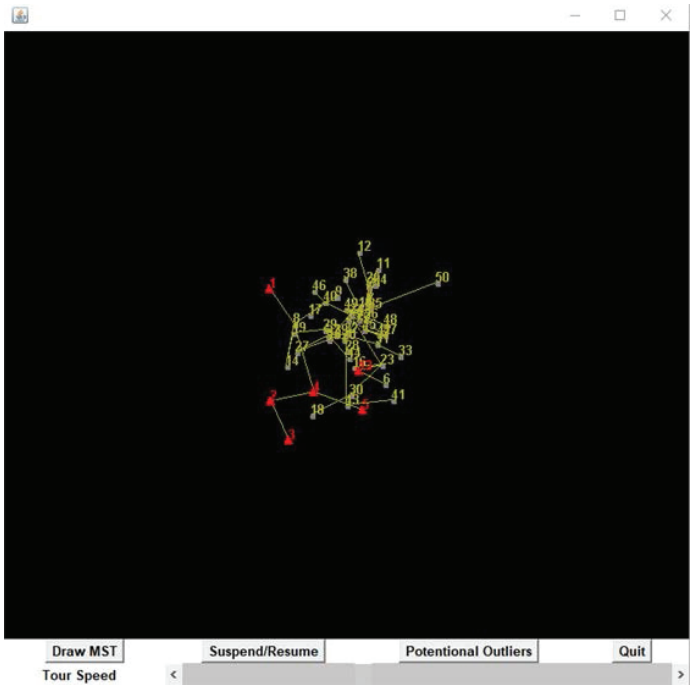


Figure 4.1. Grand tour combined with outlier visualization.

```
 1: HK-Means Clustering
 Number of clusters by Mojena rule (k=1.25,1.5,2.0,2.5)
             k=1.25 1.5 2.0 2.5
# of cluster      4   3   3   2
 2: Total within SS
 3: Silhouette method
 4: Gap statistic method
 5: NbClust::NbClust
```

4. 결론 및 제언

다변량 자료에서 특이점을 검출하기 위한 방법은 매우 많다. 더구나 분석 데이터의 크기가 커질수록, 응용 범위가 높아질수록 각 분야에 유용한 특이점 검출 방법은 더 많아질 것이다. 특이점의 검출에 있어서는 데이터의 분포를 고려하여 Mahalanobis 거리를 이용하는 방법이 기본적인 출발이라고 할 수 있다. Mahalanobis 거리를 구하는데 있어서 특이점의 영향을 덜 받는 로버스트 Mahalanobis 거리를 구하는 방법은 랜덤표본을 이용하여 작업이 이루어지기 때문에 시행 때마다 검출되는 특이점이 변하게 된다. 따라서 같은 방법이라도 여러 번 반복 시행하여 살펴보는 것이 좋다. 또한 특이점을 구하는 여러 함수들을 이용하여 다양한 관점에서 특이점을 검출할 수 있도록 하였다. 같은 방법을 여러 번 반복하여 검출하고, 다양한 관점에서 검출되는 특이점을 동시에 고려한다면 특이점을 종합하여 분석할 수 있을 것이다. 개발된 R스크립트에서 사용한 특이점 검출함수 외에도 여러 함수가 있으니 활용하기 바란다(참

고로 MVN::mvn, dprep::robout은 2018년 7월 현재 R 버전 3.5.1에서는 안정되지 않아 포함하지 않았다). 또한 각 방법들을 시각화와 연결하여 장단점을 비교하는 것도 좋은 연구 주제가 될 것이다. 참고로 다변량 특이점 검출방법을 비교, 연구한 결과로서는 Penny와 Jolliffe (2001)를 들 수 있다.

검출되는 특이점은 실제적으론 특이점이 아닐 수도 있고, 또 다른 목적에 유용한 관찰점이 될 수 있다. 이러한 의미에서 잠재적 특이점으로 명명하고, 이를 데이터 구조의 시각화와 연결하였다. 시각화 방법으로는 데이터의 구조를 파악하기 위한 탐색적 방법으로 K-평균 군집분석을 시행하고, 이를 MDS와 PCA와 연결하여 잠재적 특이점을 표시하였다. 이러한 시각화를 통해서 다양한 군집화, 특이점의 속성 등을 파악할 수 있을 것이다.

개발된 R 스크립트는 "http://www.knou.ac.kr/∼sskim/ddpoutlier.html"에서 다운 받으면 된다. 참고로 데이터의 시각화에 있어서 다이내믹 그래픽 기능인 Grand Tour를 이용한 특이점 연결(자바소스 프로그램은 http://www.public.iastate.edu/∼dicook/papers/Metrika/paper.html (Kim 등, 2000)에 있다. 저자가 기능을 일부 축소하고, 특이점은 붉은 색(삼각형)으로 표시하였다.)은 또 다른 재미를 줄 수 있다. Figure 4.1은 이러한 시도 중의 하나이다.

References

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data* (3rd ed), John Wiley & Sons, Chichester.

Breunig, M., Kriegel, H., Ng, R., and Sander, J. (2000). LOF: identifying density-based local outliers. In *SIGMOD '00 Proceedings of the 2000 ACM SIGMOD International Conference on Management of data*, Texas, 93–104.

Butler, R. W., Davies, P. L., and Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator, *The Annals of Statistics*, **21**, 1385–1400.

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). NbClust: an R package for determining the relevant number of clusters in a data set, *Journal of Statistical Software*, **61**, 1–36.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Oregon, 226–231.

Filzmoser, P. (2004). A multivariate outlier detection method, from: http://file.statistik.tuwien.ac.at/filz/papers/minsk04.pdf

Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions, *Computational Statistics & Data Analysis*, **52**, 1694–1711.

Hawkins, D. M. (1980). *Identification of Outliers*, Chapman & Hall, London.

Jayakumar, D. S. and Thomas, B. J. (2013). A new procedure of clustering based on multivariate outlier detection, *Journal of Data Science*, **11**, 69–84.

Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*, STHDA.

Kim, S., Kwon, S., and Cook, D. (2000). Interactive visualization of hierarchical clusters using MDS and MST, *Metrika*, **51**, 39–51.

Kriegel, H.-P., Kröger, P., and Zimek, A. (2010). Outlier detection techniques, *The 2010 SIAM International Conference on Data Mining*.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. In *Proceedings of the National Institute of Sciences (Calcutta)*, India, **2**, 49–55.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: an evaluation, *The Computer Journal*, **20**, 359–363.

Mojena, R. and Wishart, D. (1980). Stopping rules for Ward's clustering method. In *COMPSTAT 1980 Proceedings*, Physica-Verlag, 426–432.

Pamula, R., Deka, J. K., and Nandi, S. (2011). An outlier detection method based on clustering, *2011 Second International Conference on Emerging Applications of Information Technology*, 253–256.

Penny, K. I. and Jolliffe, I. T. (2001). A comparison of multivariate outlier detection methods for clinical laboratory safety data, *Journal of the Royal Statistical Society. Series D (The Statistician)*, **50**, 295–308.

Prim, R. C. (1957). Shortest connection networks and some generalizations, *Bell System Technical Journal*, **36**, 1389–1401.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Rousseeuw, P. J., Ruts, I., and Tukey, J. W. (1999). The Bagplot: a bivariate boxplot, *The American Statistician*, **53**, 382–387.

Rousseeuw, P. J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, **41**, 212–223.

Tibshirani, R., Walther, G., and Hastie, T. (2001), Estimating the number of clusters in a data set via the gap statistic, *Journal of Royal Statistical Society: Series B (Statistical Methodology)*, **63**, 411–423.

Wickham, H. (2010). ggplot2: Elegant Graphics for Data Analysis, *Journal of Statistical Software*, **35**, Book Review 1.

다변량 자료에서 특이점 검출 및 시각화 - R 스크립트

김성수[a,1]

[a]한국방송통신대학교 정보통계학과

(2018년 6월 28일 접수, 2018년 8월 1일 수정, 2018년 8월 1일 채택)

요 약

다변량 자료에서 특이점을 검출하고, 검출된 특이점을 시각화와 연결한 R 스크립트를 제공한다. 개발된 R 스크립트는 특이점을 검출하는 방법으로서 1) Robust Mahalanobis distance, 2) High Dimensional data, 3) Density-based approach 방법을 이용하였다. 특이점을 연결하면서 데이터 구조를 파악하기 위한 시각화 방법으로는 1) multidimensional scaling (MDS)와 minimal spanning tree (MST)를 K-means 군집분석과 연결하여 표시하는 방법, 2) MDS를 fviz_cluster와 연결하는 방법, 3) principal component analysis (PCA)를 fviz_cluster와 연결한 방법을 이용하였다. 사례분석의 예로서는 Major League Baseball (MLB) 자료에서 류현진이 적극적으로 활동하던 2013년, 2014년 투수자료를 이용하였다. 개발된 R 스트립트는 "http://www.knou.ac.kr/∼sskim/ddpoutlier.html (R 스크립트와 R 패키지도 다운로드 받을 수 있다. 실행방법도 설명되어 있다.)"에서 다운받으면 된다.

주요용0어: 잠재적 특0점, 시각화, Mahalanobis 거리, MDS, MST, PCA

이 논문은 2015년도 한국방송통신대학교 학술연구비 지원을 받아 작성된 것임.

[1](03087) 서울시 종로구 대학로 86, 한국방송통신대학교 정보통계학과. E-mail: sskim@knou.ac.kr