일변량 및 이변량 자료에 대하여 특이값의 영향을 평가하기 위한 그래픽 방법

장대홍

부경대학교 수리과학부 통계학전공

Graphical Methods for Evaluating the Effect of Outliers in Univariate and Bivariate Data

Jang, Dae-Heung

Division of Mathematical Sciences, Pukyong National University

Key Words: Dandelion Seed Plot, Influence Graph, Cumulative Deletion Plot

Abstract

We usually use two techniques(influence function and local influence) for detecting outliers. But, we cannot use these difficult techniques in elementary industrial statistics course for college students. We can use some simple graphical methods(box plot, dandelion seed plot, influence graph and cumulative deletion plot) for univariate and bivariate outlier detection and outlier effect in elementary industrial statistics course for college students.

1. 서론

통계자료분석 시 데이터에 특이값이 존재하면 이 특이값이 자료분석을 위한 여러 가지 측도들을 크게 왜곡시킨다. 특이값의 영향을 측정하기 위한 도구로서 우리는 크게 두 가지 방법을 사용한다. 이 두 가지 방법은 Hampel(1974)이 제안한 'influence function' 방법과 Cook(1986)이 제안한 'local influence' 방법이다. 그러나 우리는 품질경영 관련 학부생들을 위한 공업통계학 교수 시 특이값의 영향을 측정하기 위한 도구로서 이 두 가지 방법들을 사용하는 것은 대단히 어려운 일이다. 그래서 특이값의 영향을 측정하기 위한 도구로서 간단한 그림 도구들을 제안할 수 있다. 이러한 방법들은 품질경영 관련 학부생들이 이해하기에 충분히 쉬운 도구들이다.

2. 일변량 자료에 대하여 특이값의 영향을 평가하기 위한 그래픽 방법

일변량 자료에 특이값이 존재하게 되면 대표값과 산포도에 관계되는 수치적 측도들이 영향을 받게 된다. 대표값에서는 산술평균이 대표적으로 특이값에 민감한(sensitive) 수치적 측도이고 산포도에서는 분산, 표준편차, 범위 등이 대표적으로 특이값에 민감한 수치적 측도들이다. 이러한 수치적 측도들이 특이값에 얼마나 민감한지를 나타내는 방법으로서 그래픽 방법을 사용한다면 품질경영 관련 학부생들을 위한 공업통계학 교수 시 많은 도움을 얻을 수 있을 것이다.

일변량 자료에 대하여 특이값의 영향을 평가하기 위하여 우리는 상자그림을 이용할 수 있다. 전체 n개의 자료에 대한 상자그림과 전체 n개의 자료에서 하나씩의 자료를 차례로 뺀 후 남은 $(n-1)$개의 자료들에 대한 상자그림들을 병렬상자그림(side-by-side box plot)으로 그려 보면 일변량 자료에 대하여 특이값의 영향을 평가하여 볼 수 있다. 이 때 상자그림들은 산술평균을 같이 나타내는 상자그림이어야 한다.

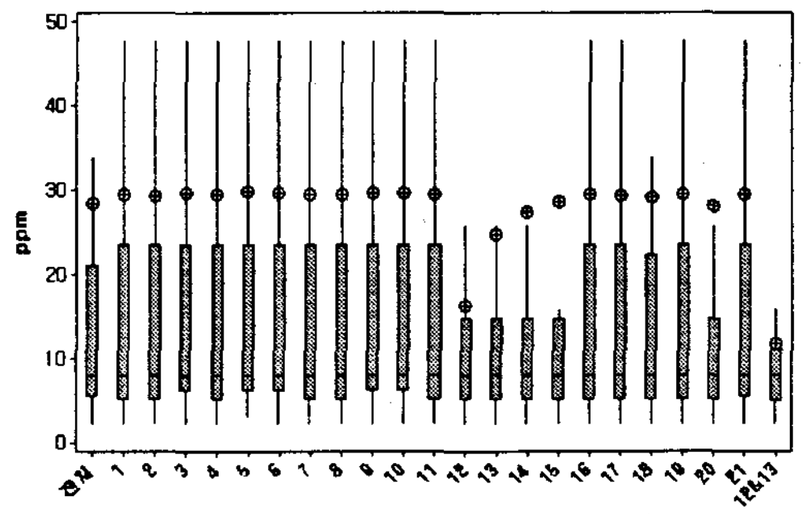
다음의 자료는 어떤 특정 타입의 플라스틱 제품 21개에 포함된 알루미늄 오염수치(ppm)를 나타낸 표이다.

<표 1> 어떤 특정 타입의 플라스틱 제품 21개에 포함된 알루미늄 오염수치(ppm)

NO.	1	2	3	4	5	6	7	8	9	10	11
ppm	8	11	5	7	2	3	8	8	4	3	8

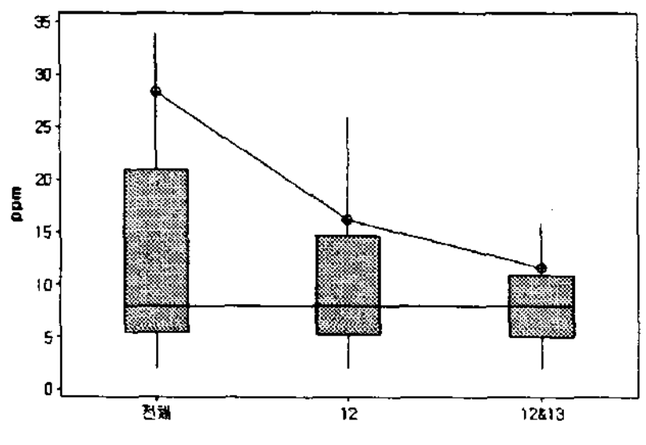
NO.	12	13	14	15	16	17	18	19	20	21
ppm	272	103	48	26	8	10	16	7	34	6

이 자료에 대하여 특이값의 영향을 평가하기 위하여 전체 21개의 자료에 대한 상자그림과 전체 21개의 자료에서 하나씩의 자료를 차례로 뺀 후 남은 20개의 자료들에 대한 상자그림들을 병렬상자그림으로 그려 보면 다음 <그림 1>과 같다. 그림에서 십자표시가 산술평균을 나타내고 가장 오른쪽 상자그림은 12번째(제일 큰 특이값)와 13번째(두 번째 큰 특이값) 두 개의 자료를 동시에 뺀 후 남은 19개 자료에 대한 상자그림이다. 산술평균의 변화를 통하여 12번째 자료가 가장 큰 영향을 주는 특이값임을 알 수 있다.



<그림 1> 병렬상자그림 1

전체 21개의 자료에 대한 상자그림, 전체 21개의 자료에서 12번째의 자료를 뺀 후 남은 20개 자료에 대한 상자그림, 그리고 12번째와 13번째 두 개의 자료를 동시에 뺀 후 남은 19개 자료에 대한 상자그림을 병렬 상자그림으로 그려 보면 다음 <그림 2>와 같다. 중앙값은 변화가 없으나 산술평균에는 많은 변화가 있음을 알 수 있다.

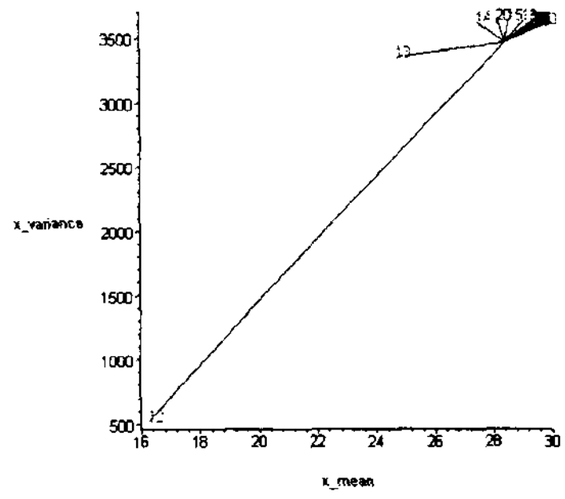


<그림 2> 병렬상자그림 2

자료를 $x_1, x_2, \ldots, x_n$이라 할 때 가중산술평균과 가중분산을 우리는 다음과 같이 정의할 수 있다.

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \quad s_{x_w}^2 = \frac{\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2}{\sum_{i=1}^{n} w_i}$$

n개의 자료 각각에 대하여 가중치 $w_i\,(i=1,2,\ldots,n)$를 1에서 0으로 변화시켜 가며 가중산술평균과 가중분산을 구한 후 이렇게 구한 가중산술평균과 가중분산들을 연결하면 하나의 그림이 완성되는 데 이러한 그림을 평균-분산 민들레씨그림(dandelion seed plot)이라 명하자. <그림 3>은 <표 1>의 자료에 대하여 그린 평균-분산 민들레씨그림이다.

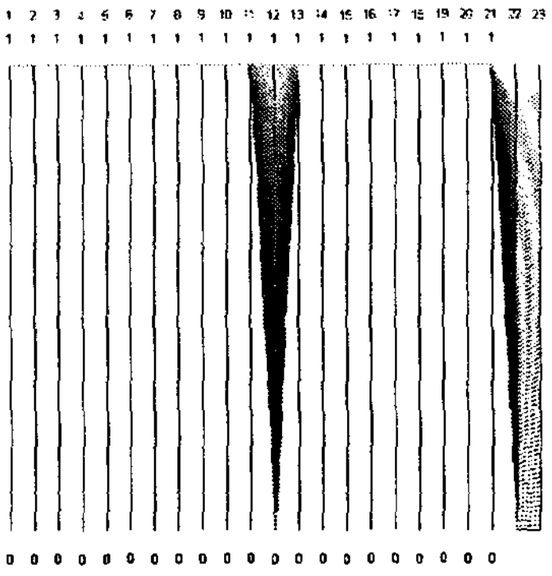


<그림 3> 평균-분산 민들레씨그림

12번째 자료에서 가중산술평균과 가중분산이 극적인 변화를 겪고 있음을 알 수 있다. 그러므로 12번째 자료가 가장 큰 영향을 주는 특이값임을 알 수 있다. 그림에서의 가중산술평균과 가중분산의 변화량은 가중산술평균과 가중분산에 대한 표준화민감도곡선(standardized sensitivity curve, Maronna 외 2인(2006) 참조.)값에 비례한다.

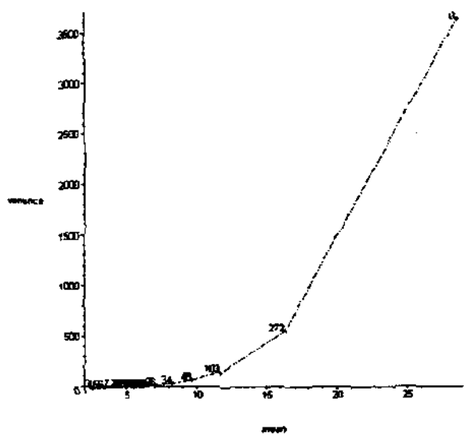
평행좌표그림(parallel coordinate plot)을 변형하면 일변량 자료에 대하여 특이값의 영향을 평가할 수 있다. 평행좌표그림은 Inselberg(1985)의 제안 이후 최근 탐색적 그래픽자료분석 및 시각 데이터마이닝에서 중요한 수단으로 쓰이고 있다. n개의 평행좌표축에 n개의 자료를 각각 대응시키고 다음 두 개의 평행좌표축에 가중산술평균과 가중분산을 대응시킨다. $i\,(i=1,2,\ldots,n)$-번째 자료에 대응되는 i-번째 평행좌표축 상에서 위에서 아래로 내려가며 가중치 w_i를 1에서 0으로 변화시킨다. 각 가중치에 대응되는 가중산술평균과 가중분산의 값을 가중산술평균과 가중분산 평행좌표축에 표

시한다. 이러한 평행좌표그림의 변형을 영향력그림(influence graph)이라고 칭하자. <그림 4>는 12번째 자료에 대한 영향력그림이다. 12번째 자료의 가중치 w_i를 1에서 0으로 변화시켰을 때 가중산술평균과 가중분산이 극적인 변화를 겪고 있음을 알 수 있다. 그러므로 12번째 자료가 가장 큰 영향을 주는 특이값임을 알 수 있다.



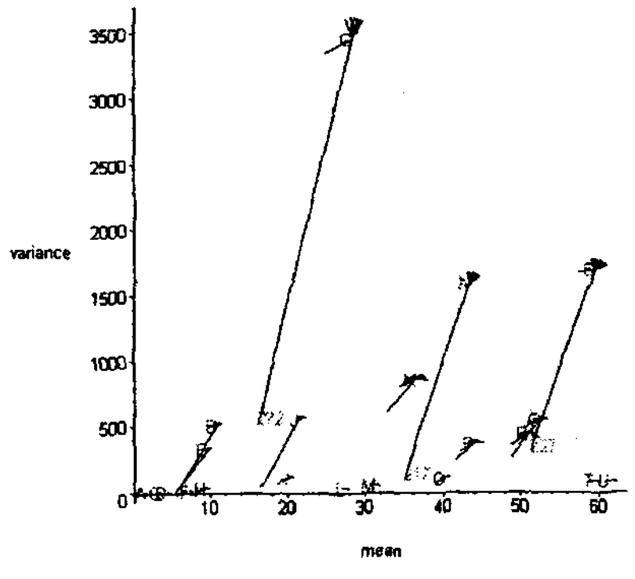
<그림 4> 영향력그림

누적제거그림(cumulative deletion plot)은 자료의 오름차순이나 내림차순의 순서대로 하나씩 제거시키며 가중산술평균과 가중분산의 값을 계산 한 후 연결시킨 그림이다. <그림 5>는 12번째 자료에 대한 누적제거그림이다.

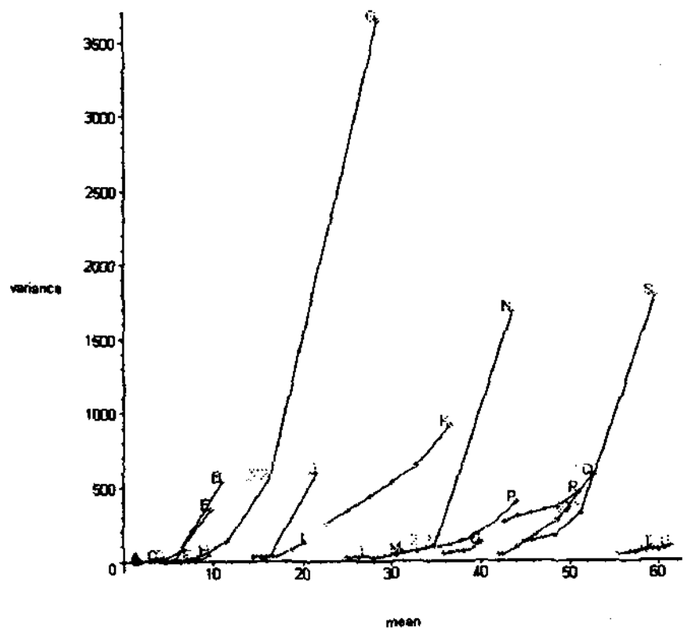


<그림 5> 누적제거그림

다음 페이지에 있는 표는 21개 회사(A-U) 각각에 대하여 플라스틱 제품 21개에 포함된 알루미늄 오염수치(ppm)를 나타낸 표이다. 이 자료에 대한 민들레씨그림과 누적제거그림은 <그림 6>과 <그림 7>과 같다. <그림 7>의 누적제거그림은 내림차순의 순서대로 5개까지 하나씩 제거시키며 그린 누적제거그림이다.



<그림 6> 평균-분산 민들레씨그림



<그림 7> 누적제거그림

G, N, S 세 회사가 영향이 큰 특이값들을 갖고 있음을 알 수 있다.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
3	2	5	4	9	7	8	14	16	17	20	33	48	39	69	51	72	71	72	61	70
1	2	4	3	11	10	11	12	22	17	17	31	43	52	43	52	52	74	54	48	63
3	6	5	1	10	14	5	10	14	15	10	15	26	24	30	40	45	77	45	43	41
1	3	2	4	6	9	7	7	15	24	10	26	33	43	38	23	40	67	51	65	69
2	12	3	4	5	9	2	11	9	10	2	39	21	35	50	30	40	71	37	68	70
2	106	8	1	3	15	3	15	12	6	37	25	20	26	39	30	33	52	31	64	69
1	36	2	0	4	8	8	12	16	7	10	24	26	24	48	29	50	67	33	56	64
6	14	10	10	11	7	8	33	47	124	76	27	29	217	44	95	93	71	58	92	59
1	14	1	6	4	15	4	9	23	24	82	22	32	44	26	37	66	50	105	54	78
1	3	2	10	7	5	3	8	26	19	172	22	23	46	31	99	32	21	51	52	53
0	4	3	6	8	9	8	11	19	23	65	28	34	36	39	51	38	32	28	61	57
3	4	4	3	4	6	272	8	17	22	72	35	41	38	40	44	44	20	28	74	60
3	1	5	4	4	7	108	11	52	24	48	31	41	36	52	58	130	37	227	61	61
1	2	6	6	3	6	48	8	28	21	81	31	32	36	50	59	49	32	42	66	68
1	1	5	9	90	2	26	4	20	15	18	27	32	32	49	45	49	25	60	53	56
0	0	1	5	8	4	6	5	8	12	18	26	26	22	29	39	36	33	45	54	56
1	2	0	4	4	2	10	4	13	9	37	24	30	19	32	21	38	97	42	67	49
6	7	4	2	6	7	16	9	25	26	21	38	38	54	45	42	49	44	66	69	72
3	10	5	6	5	3	7	5	15	12	13	16	22	25	29	30	50	39	51	57	54
2	2	2	3	1	6	34	2	10	8	25	38	31	26	32	24	43	54	63	53	73
0	1	2	5	1	9	6	3	16	18	13	29	26	40	37	30	35	37	61	47	47

3. 이변량 자료에 대하여 특이값의 영향을 평가하기 위한 그래픽 방법

이변량 자료를 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$라 할 때 가중산술평균, 가중분산, 가중공분산, 가중상관계수들을 우리는 다음과 같이 정의할 수 있다.

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \quad s_{x_w}^2 = \frac{\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2}{\sum_{i=1}^{n} w_i}$$

$$\bar{y}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}, \quad s_{y_w}^2 = \frac{\sum_{i=1}^{n} w_i (y_i - \bar{y}_w)^2}{\sum_{i=1}^{n} w_i}$$

$$cov_w(x, y) = \frac{\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum_{i=1}^{n} w_i}$$

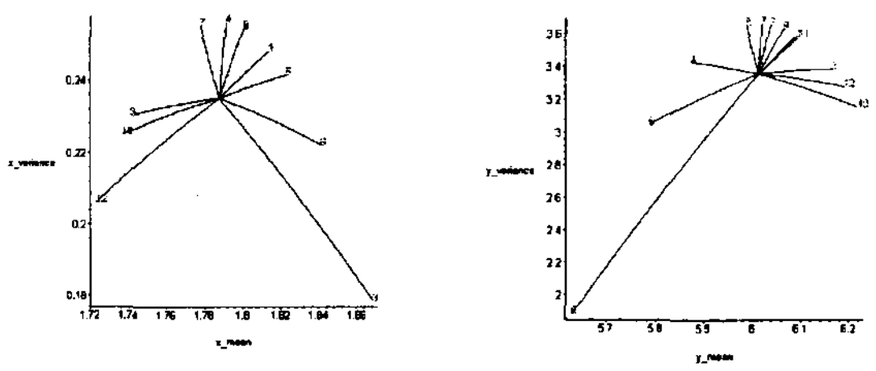
$$corr_w(x, y) = \frac{cov_w(x, y)}{\sqrt{s_{x_w}^2}\sqrt{s_{y_w}^2}}$$

n개의 이변량 자료 각각에 대하여 가중치 $w_i\,(i = 1, 2, \ldots, n)$를 1에서 0으로 변화시켜가며 가중산술평균, 가중분산, 가중공분산, 가중상관계수들을 구한 후 이렇게 구한 가중산술평균과 가중분산들을 연결하여 각각 변수 X와 Y에 대한 평균-분산 민들레씨그림을 작성하고 가중공분산과 가중상관계수들을 연결하여 공분산-상관계수 민들레씨그림을 그릴 수 있다. 다음 <표 2>는 생화학 자료이다.

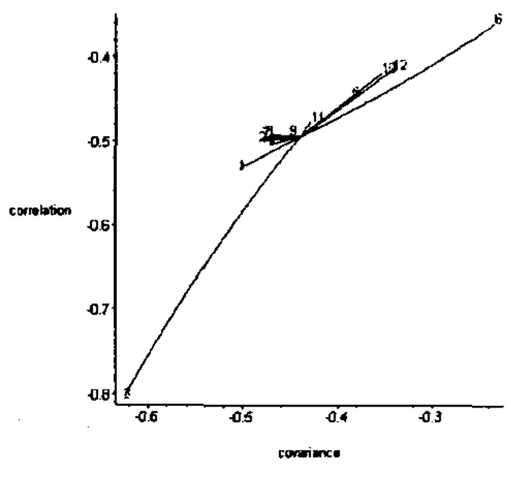
<표 2> 생화학 자료

관측치	인삼염	염화물
1	1.50	5.15
2	1.65	5.75
3	0.90	4.35
4	1.75	7.55
5	1.40	8.50
6	1.20	10.25
7	1.90	5.95
8	1.65	6.30
9	2.30	5.45
10	2.35	3.75
11	2.35	5.10
12	2.50	4.05

이 자료에 대하여 평균-분산 민들레씨그림과 공분산-상관계수 민들레씨그림을 그리면 다음 <그림 8>과 <그림 9>와 같다.



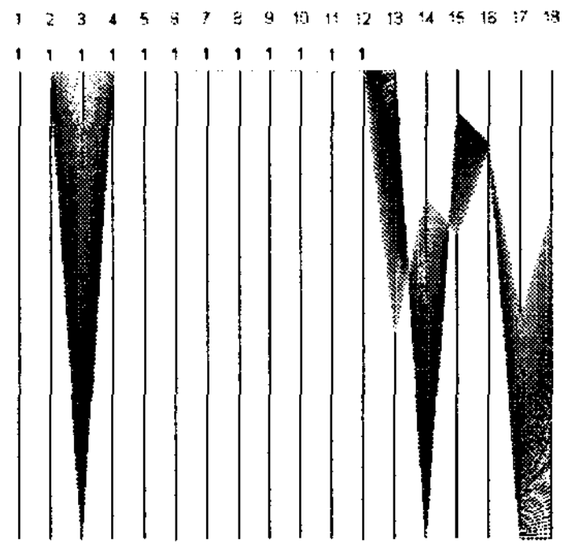
<그림 8> 변수 X와 Y에 대한
평균-분산 민들레씨그림



<그림 9> 공분산-상관계수 민들레씨그림

3번째 자료가 특이값임을 알 수 있다. 변수 X에 대한 가중산술평균과 가중분산, 가중공분산과 가중상관계수에서 큰 변화가 있음을 알 수 있다.

다음 <그림 10>은 3번째 자료에 대한 영향력그림이다. 12개의 평행좌표축에 12개의 자료를 각각 대응시키고 다음 여섯 개의 평행좌표축에 변수 X와 Y에 대한 가중산술평균과 가중분산, 가중공분산과 가중상관계수를 대응시킨다. 변수 X에 대한 가중산술평균과 가중분산, 가중공분산과 가중상관계수에서 큰 변화가 있음을 알 수 있다.



<그림 10> 영향력그림

4. 결론

본 논문에서는 특이값의 영향을 측정하기 위한 도구로서 간단한 그림도구들(민들레씨그림, 영향력그림, 누적제거그림)을 제안하였다. 이러한 방법들은 품질경영 관련 학부생들을 위한 공업통계학 교수 시 매우 유용하게 쓰일 수 있다.

참고문헌

[1] Cook, R. D.(1986), "Assessment of Local Influence", *Journal of Royal Statistical Society*, *B* 48, 133-169.

[2] Hampel, F. R.(1974), "The Influence Curve and Its Role in Robustness, *The Annal of Statistics*, 45, 383-393.

[3] Inselberg, A.(1985), "The Plane with Parallel Coordinates", *The Visual Computer*, 1, 69-97.

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