Multivariate Outlier Detection Technique Based on Normalizing Transformations for Non-Gaussian Data

Sergiy Prykhodko
Department of Software of Automated Systems
Admiral Makarov National University of Shipbuilding
Mykolaiv, Ukraine
E-mail: sergiy.prykhodko@nuos.edu.ua

Natalia Prykhodko
Finance Department
Admiral Makarov National University of Shipbuilding
Mykolaiv, Ukraine
E-mail: natalia.prykhodko@nuos.edu.ua

Lidiia Makarova
Department of Software of Automated Systems
Admiral Makarov National University of Shipbuilding
Mykolaiv, Ukraine
E-mail: lidiia.makarova@nuos.edu.ua

Kateryna Pugachenko
Department of Information Systems and Technologies
Admiral Makarov National University of Shipbuilding
Mykolaiv, Ukraine
E-mail: kateryna.puhachenko@gmail.com

Метод Визначення Багатовимірних Викидів, що Базується на Нормалізуючих Перетвореннях для Негаусівських Даних

Сергій Приходько
кафедра програмного забезпечення автоматизованих систем
Національний університет кораблебудування імені адмірала Макарова
Миколаїв, Україна
sergiy.prykhodko@nuos.edu.ua

Наталія Приходько
кафедра фінансів
Національний університет кораблебудування імені адмірала Макарова
Миколаїв, Україна
natalia.prykhodko@nuos.edu.ua

Лідія Макарова
кафедра програмного забезпечення автоматизованих систем
Національний університет кораблебудування імені адмірала Макарова
Миколаїв, Україна
lidiia.makarova@nuos.edu.ua

Катерина Пугаченко
кафедра інформаційних управляючих систем та технологій
Національний університет кораблебудування імені адмірала Макарова
Миколаїв, Україна
kateryna.puhachenko@gmail.com

***Abstract*—The statistical technique for detecting outliers in multivariate non-Gaussian data on the basis of normalizing transformations and a test statistic for the Mahalanobis squared distance (MSD), which has an approximate *F* distribution, is proposed. Application of the technique is considered for detecting outliers in the data of four-variate measurements.**

***Анотація*—Пропонується статистичний метод для виявлення викидів в багатовимірних негаусівських даних на основі нормалізуючих перетворень і тестової статистики для квадрата відстані Махаланобісу (MSD), який має наближений *F* розподіл. Розглянуто застосування цього методу для виявлення викидів за даними вимірювань чотирьох величин.**

Keywords—outlier; normalizing transformation; multivariate non-Gaussian data; Mahalanobis squared distance; F distribution

Ключові слова—викид; нормалізуюче перетворення; багатовимірні негауссівськи дані; квадрат відстані Махаланобісу; F-розподіл

I. Introduction

A very important step in data analysis and processing is the outlier detection. Today the problem of outlier detection in a multivariate data set is solved with different methods including statistical [1, 2]. However, well-known statistical methods (for example, multivariate outlier detection based on a test statistics for MSD, which have an approximate the Chi-Square distribution or F distribution) are used to detect outliers in a multivariate data set under the assumption that the data is generated by a Gaussian distribution. And this assumption is valid only in particular cases.

In [3] a statistical outlier detection technique for multivariate non-Gaussian data on the basis of normalizing transformations and MSD, which has an approximate the Chi-Square distribution, was proposed. However, the quantile of the Chi-Square distribution does not depend from the number of data. We propose a statistical outlier detection technique for multivariate non-Gaussian data on the basis of normalizing transformations and a test statistic for MSD, which has an approximate F distribution. At that the quantile of the F distribution depends from the number of data. The technique consists of two steps. In the first step, multivariate non-Gaussian data is normalized using a multivariate normalizing transformation. In the second step, MSD and a test statistic for MSD are calculated and compared with a quantile of the F distribution. The data values for which a value of test statistic for MSD is greater than the quantile of the F distribution are considered as outliers and these values are cut off. Two steps should be repeated for the data after outlier cutoff until all values of test statistic for MSD will be less than or equal to the quantile of the F distribution.

II. The statistical technique

The outlier detection technique for multivariate non-Gaussian data is based on normalizing transformations and a test statistic for MSD, which has an approximate F distribution. Consider bijective multivariate normalizing transformation of non-Gaussian random vector $\mathbf{X} = \{X_1, X_2, \ldots, X_m\}^T$ to Gaussian random vector $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_m\}^T$ is given by

$$\mathbf{Z} = \psi(\mathbf{X}). \tag{1}$$

The values of the sample observations or multivariate data points $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ are normalized using the transformation (1).

The Mahalanobis squared distance for each multivariate data point i, $i = 1,2,\ldots,N$, is denoted by d_i^2 and given by

$$d_i^2 = (\mathbf{Z}_i - \overline{\mathbf{Z}})^T S_N^{-1} (\mathbf{Z}_i - \overline{\mathbf{Z}}), \tag{2}$$

where $\overline{\mathbf{Z}}$ is the sample mean vector and S_N is the sample correlation matrix

$$S_N = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{Z}_i - \overline{\mathbf{Z}})(\mathbf{Z}_i - \overline{\mathbf{Z}})^T. \tag{3}$$

A test statistic (TS) for d_i^2 can be created as follows [4]

$$(N-m) N d_i^2 / (N^2 - 1) m, \tag{4}$$

which has an approximate F distribution with m and $N - m$ degrees of freedom.

A test statistic for MSD (4) is compared with a quantile of the F distribution, which is noted as $F_{m,N-m,\alpha}$. Here α is significance level. We take α as 0.05. The data values for which a value of TS (4) is greater than the quantile of the F distribution are considered as outliers and these values are cut off. After outlier cutoff the reduced number of multivariate data points are normalized using the transformation (1) again until all values of TS (4) will be less than or equal to the quantile of the F distribution.

III. Multivariate Normalizing Transformations

Some transformations have been proposed for normalizing multivariate non-Gaussian data, such as, transformation on the basis of the decimal logarithm, the Box-Cox transformation, the Johnson translation system and others. However, only a few normalizing transformations are bijective. Such bijective transformation is the transformation of S_U family of the Johnson translation system. The Johnson normalizing translation is given by [5]

$$\mathbf{Z} = \boldsymbol{\gamma} + \boldsymbol{\eta} \mathbf{h}\left[\boldsymbol{\lambda}^{-1}(\mathbf{X} - \varphi)\right] \sim N_m(0_m, \Sigma), \tag{5}$$

where Σ is the correlation matrix; $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$, φ and $\boldsymbol{\lambda}$ are parameters of the Johnson normalizing translation; $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_m)^T$; $\boldsymbol{\eta} = diag(\eta_1, \eta_2, \ldots, \eta_m)$; $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_m)^T$; $\boldsymbol{\lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_m)$; $\mathbf{h}[(y_1, y_2, \ldots, y_m)] = \{h_1(y_1), h_2(y_2), \ldots, h_m(y_m)\}^T$; $h_i(\cdot)$ is one of the translation functions

$$h = \begin{cases} \ln(y), & \text{for } S_L \text{ (log normal) family;} \\ \ln[y/(1-y)], & \text{for } S_B \text{ (bounded) family;} \\ \mathrm{Arsh}(y), & \text{for } S_U \text{ (unbounded) family;} \\ y & \text{for } S_N \text{ (normal) family.} \end{cases}$$

Here $y = (x - \varphi)/\lambda$; $\mathrm{Arsh}(y) = \ln\left(y + \sqrt{y^2 + 1}\right)$.

IV. Examples

We consider the examples of detecting outliers in multivariate non-Gaussian data for two cases: the first, under the assumption that the data set is generated by a Gaussian distribution and, the second, this assumption is not valid. In [2]

example of detecting outliers in the lumber stiffness data (example 4.15, p.190) is presented. Table I contains the data on lumber, MSD and a test statistic for MSD for standardized measurements, which are

$$Z_{ki} = \left(X_{ki} - \overline{X}_k\right) \Big/ S_{X_k}\ ,\ k = 1,2,3,4\ ;\ i = 1,2,\ldots,30. \quad (6)$$

TABLE I. MSD for the standardized data

i	X_1	X_2	X_3	X_4	d_i^2	TS for MSD
1	1889	1651	1561	1778	0.61	0.13
2	2493	2048	2087	2197	7.48	1.62
3	2119	1700	1815	2222	7.85	1.70
4	1645	1627	1110	1533	5.23	1.13
5	1976	1916	1614	1883	1.44	0.31
6	1712	1712	1439	1546	2.27	0.49
7	1943	1685	1271	1671	5.17	1.12
8	2104	1820	1717	1874	1.34	0.29
9	2983	2794	2412	2581	12.37	2.68
10	1745	1600	1384	1508	0.78	0.17
11	1710	1591	1518	1667	2.05	0.44
12	2046	1907	1627	1898	0.5	0.11
13	1840	1841	1595	1741	2.78	0.60
14	1867	1685	1493	1678	0.12	0.03
15	1859	1649	1389	1714	1.11	0.24
16	1954	2149	1180	1281	17.53	3.80
17	1325	1170	1002	1176	3.63	0.79
18	1419	1371	1252	1308	4.08	0.89
19	1828	1634	1602	1755	1.48	0.32
20	1725	1594	1313	1646	1.48	0.32
21	2276	2189	1547	2111	10.25	2.22
22	1899	1614	1422	1477	4.8	1.04
23	1633	1513	1290	1516	0.8	0.17
24	2061	1867	1646	2037	2.64	0.57
25	1856	1493	1356	1533	4.33	0.94
26	1727	1412	1238	1469	3.37	0.73
27	2168	1896	1701	1834	2.13	0.46
28	1655	1675	1414	1597	2.93	0.64
29	2326	2301	2065	2234	6.76	1.47
30	1490	1382	1214	1284	2.67	0.58

These data consist of four different measures of stiffness X_1, X_2, X_3 and X_4, on each $N = 30$ of boards. For significance level which equals to 0.05 the last column in Table I reveals that specimen 16 is multivariate outlier, since $F_{m,N-m,\alpha}$ =2.74.

Table II contains the normalized data on lumber, MSD and a test statistic for MSD for normalized data. These data is normalized by S_U family of the transformation (5).

TABLE II. MSD FOR THE NORMOLIZED DATA

i	Z_1	Z_2	Z_3	Z_4	d_i^2	TS for MSD
1	0.1984	-0.1910	0.3966	0.3539	0.77	0.17
2	1.4562	1.0898	1.5843	1.2985	2.88	0.63
3	0.8508	0.0237	1.0712	1.3430	4.23	0.92
4	-0.9978	-0.3027	-1.7482	-0.4821	9.07	1.97
5	0.4882	0.7632	0.5598	0.6343	0.78	0.17
6	-0.6296	0.0734	-0.0412	-0.4297	2.31	0.50
7	0.3859	-0.0401	-0.8259	0.0249	4.54	0.98
8	0.8177	0.4724	0.8404	0.6117	1.13	0.25
9	1.9326	2.1752	2.0416	1.8902	5.26	1.14
10	-0.4518	-0.4328	-0.2728	-0.5859	0.59	0.13
11	-0.6406	-0.4770	0.2530	0.0116	1.99	0.43
12	0.6799	0.7381	0.5977	0.6715	0.58	0.13
13	0.0031	0.5407	0.5029	0.2456	1.54	0.33
14	0.1139	-0.0401	0.1644	0.0480	0.14	0.03
15	0.0819	-0.2002	-0.2507	0.1630	0.92	0.20
16	0.4209	1.2978	-1.3383	-1.7120	18.95	4.11
17	-2.2232	-2.2878	-2.3454	-2.3094	6.33	1.37
18	-1.9595	-1.5575	-0.9289	-1.5624	5.40	1.17
19	-0.0488	-0.2697	0.5241	0.2872	1.17	0.25
20	-0.5587	-0.4622	-0.6080	-0.0594	1.68	0.36
21	1.1474	1.3723	0.3511	1.1373	4.25	0.92
22	0.2352	-0.3648	-0.1103	-0.7201	3.97	0.86
23	-1.0620	-0.8721	-0.7256	-0.5522	1.50	0.32
24	0.7172	0.6214	0.6518	0.9874	1.16	0.25
25	0.0697	-0.9741	-0.4000	-0.4821	4.34	0.94
26	-0.5479	-1.3717	-1.0063	-0.7558	3.78	0.82
27	0.9523	0.7069	0.7996	0.5081	1.42	0.31
28	-0.9435	-0.0836	-0.1435	-0.2339	3.52	0.76
29	1.2270	1.5616	1.5482	1.3640	3.14	0.68
30	-1.7150	-1.5089	-1.1418	-1.6952	3.96	0.86

In these case the parameters are such: $\gamma_1 = -0.76817$, $\gamma_2 = -0.81796$, $\gamma_3 = -1.52027$, $\gamma_4 = -2.11559$, $\eta_1 = 1.01985$, $\eta_2 = 1.29026$, $\eta_3 = 1.58452$, $\eta_4 = 1.87154$, $\varphi_1 = 1686.87$,

$\varphi_2 = 1523.59$, $\varphi_3 = 1148.94$, $\varphi_4 = 1209.93$, $\lambda_1 = 184.403$, $\lambda_2 = 252.185$, $\lambda_3 = 269.810$ and $\lambda_4 = 327.027$.

The sample correlation matrix (3) of the $\mathbf{Z}$ is used as the approximate moment-matching estimator of correlation matrix Σ

$$S_N = \begin{pmatrix} 1.00000 & 0.85039 & 0.79001 & 0.82173 \\ 0.85039 & 1.00000 & 0.71725 & 0.73854 \\ 0.79001 & 0.71725 & 1.00000 & 0.87227 \\ 0.82173 & 0.73854 & 0.87227 & 1.00000 \end{pmatrix}.$$

In Table II the last column reveals that only specimen 16 is multivariate outlier, since $F_{m,N-m,\alpha}$=2.74. We note the value of TS for MSD for the normalized data from Table II is greater than the same value for the data from Table I in this case.

Table III contains the normalized data on lumber and MSD for normalized data after outlier cutoff of specimen 16 (according Table II). These data is normalized by S_U family of the transformation (5). After outlier cutoff the parameters are such: $\gamma_1 = -0.70625$, $\gamma_2 = -0.62410$, $\gamma_3 = -0.70133$, $\gamma_4 = -21.2264$, $\eta_1 = 0.33081$, $\eta_2 = 0.23699$, $\eta_3 = 0.21906$, $\eta_4 = 6.20266$, $\varphi_1 = 1700.55$, $\varphi_2 = 1587.45$, $\varphi_3 = 1304.12$, $\varphi_4 = -181.883$, $\lambda_1 = 10.202$, $\lambda_2 = 2.123$, $\lambda_3 = 1.495$ and $\lambda_4 = 124.038$.

The sample correlation matrix (3) of the $\mathbf{Z}$ is such

$$S_N = \begin{pmatrix} 1.00000 & 0.70699 & 0.74789 & 0.77134 \\ 0.70699 & 1.00000 & 0.72856 & 0.75745 \\ 0.74789 & 0.72856 & 1.00000 & 0.70002 \\ 0.77134 & 0.75745 & 0.70002 & 1.00000 \end{pmatrix}.$$

The last column in Table III reveals there are no multivariate outliers in the data, since $F_{m,N-m,\alpha}$=2.76. We note, if the parametric anomaly detection technique [6] based on the Grubb test applies for detecting outliers in the normalized data from Table II and Table III then all data sample units do not appear to be an outlier in each of the univariate distributions.

Following [7] Mardia's multivariate kurtosis β_2 is calculated for the data from Table I, Table II and Table III as

$$\beta_2 = \frac{1}{N}\sum_{i-1}^{N}\left\{\left(\mathbf{Z}_i - \overline{\mathbf{Z}}\right)^T S_N^{-1}\left(\mathbf{Z}_i - \overline{\mathbf{Z}}\right)\right\}^2.$$

It is known that $\beta_2 = m(m+2)$ holds under multivariate normality. The given equality is a necessary condition for multivariate normality. In our case $\beta_2 = 24$. The values of multivariate kurtosis equal, respectively, 30.71, 23.84 and 23.78 for the data from Table I, Table II and Table III. These values indicate that the necessary condition for multivariate normality is practically performed for the normalized data from Table II and Table III and does not hold for standardized data from Table I by the formula (6).

TABLE III. MSD FOR THE NORMOLIZED DATA AFTER OUTLIER CUTOFF

i	Z_1	Z_2	Z_3	Z_4	d_i^2	TS for MSD
1	0.4880	0.3457	0.5780	0.1988	0.49	0.11
2	0.9629	0.8150	0.8221	1.3986	2.10	0.45
3	0.7517	0.4811	0.7286	1.4633	3.18	0.69
4	-1.4990	0.2334	-1.9193	-0.6276	10.43	2.25
5	0.6135	0.7350	0.6191	0.5219	0.58	0.13
6	-0.3870	0.5051	0.4369	-0.5809	3.23	0.70
7	0.5713	0.4472	-1.5320	-0.1487	11.66	2.52
8	0.7397	0.6531	0.6819	0.4949	0.66	0.14
9	1.1222	1.0433	0.8982	2.3257	7.32	1.58
10	0.0142	-0.0371	0.3221	-0.7184	2.09	0.45
11	-0.4323	-0.3193	0.5379	-0.1621	2.29	0.49
12	0.6883	0.7284	0.6281	0.5667	0.61	0.13
13	0.3886	0.6736	0.6052	0.0808	1.06	0.23
14	0.4470	0.4472	0.5106	-0.1253	1.21	0.26
15	0.4308	0.3381	0.3354	-0.0067	0.59	0.13
16	-2.1285	-2.0400	-2.0162	-2.0706	5.33	1.15
17	-2.0332	-1.8843	-1.6313	-1.4974	4.75	1.03
18	0.3589	0.2720	0.6104	0.1257	0.56	0.12
19	-0.1743	-0.1869	-0.1577	-0.2327	0.05	0.01
20	0.8571	0.8783	0.5657	1.1705	1.53	0.33
21	0.5051	0.1392	0.4073	-0.8330	4.59	0.99
22	-1.5628	-1.6314	-1.3458	-0.6892	4.74	1.02
23	0.7024	0.6967	0.6406	0.9673	0.94	0.20
24	0.4244	-1.6878	0.2276	-0.6276	9.57	2.06
25	-0.1502	-1.8345	-1.6834	-0.8629	8.16	1.76
26	0.7883	0.7201	0.6733	0.3732	0.98	0.21
27	-1.4346	0.4216	0.3920	-0.4009	8.38	1.81
28	0.8847	0.9188	0.8159	1.4942	2.46	0.53
29	-1.9372	-1.8719	-1.7512	-1.5977	4.33	0.93

The F plots are computed for a test statistic for MSD from Table I, Table II and Table III. A F plot is similar to a Q-Q plot. The F plots for the test statistic for MSD (4) from Table I, Table II and Table III are shown on Fig. 1, Fig. 2 and Fig. 3.

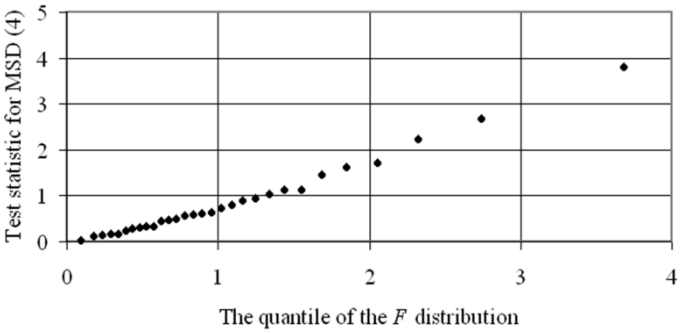

Fig. 1. *F* plot for a test statistic for MSD (4) from Table I

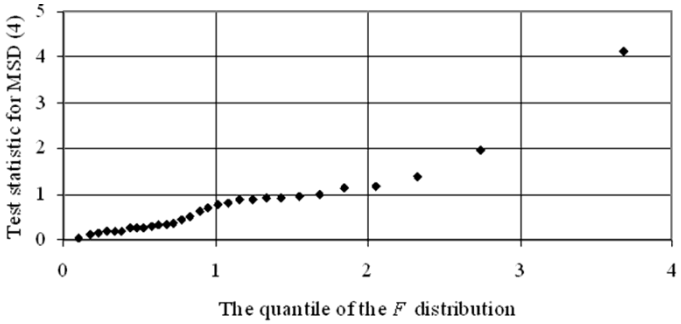

Fig. 2. *F* plot for a test statistic for MSD (4) from Table II

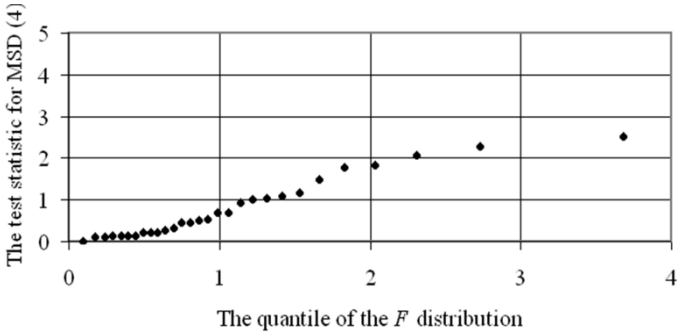

Fig. 3. *F* plot for a test statistic for MSD (4) from Table III

On Fig. 1 and Fig. 2 the quantiles of the *F* distribution are calculated for $m = 4$ and $N - m$ =26 degrees of freedom and the 30 probability levels, which are computed as $\frac{(1-0.5)}{30} = 0.0167$, $\frac{(2-0.5)}{30} = 0.05, \ldots,$ $\frac{(30-0.5)}{30} = 0.9833$.

On Fig. 3 the quantiles of the *F* distribution are calculated for $m = 4$ and $N - m$ =25 degrees of freedom and the 29 probability levels, which are computed as $\frac{(1-0.5)}{29} = 0.0172$, $\frac{(2-0.5)}{29} = 0.0517, \ldots,$ $\frac{(29-0.5)}{30} = 0.9828$.

It is necessary to check if points of the *F* plot approximately fall on a straight line. On Fig. 1 and Fig. 2 the one specimen with largest MSD is clearly removed from straight line pattern. We note on Fig. 1 and Fig. 2 the point with largest MSD is situated upper from an approximate straight line. On Fig. 3 all points do not appear to be an outlier. The same results were obtained by proposed technique for detecting outliers in multivariate non-Gaussian data.

V. Conclusions

From the examples we conclude that the proposed technique is promising, since, firstly, for multivariate outlier the value of test statistic for the Mahalanobis squared distance for the normalized data is greater than the same value for the standardized data and, secondly, the multivariate data point is not flagged as an outlier when we only consider the univariate data. However, some refinements may be necessary. We intend to try other normalizing transformations and test statistics for the Mahalanobis squared distance.

References Литература

[1] D.M. Hawkins. Identification of Outliers, London; New York: Chapman and Hall, 1980, 188 p.

[2] R.A. Johnson and D.W. Wichern. Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007, 800 p.

[3] S. Prykhodko, N. Prykhodko, L. Makarova and K. Pugachenko, "Detecting outliers in multivariate non-Gaussian data on the basis of normalizing transformations", unpublished.

[4] P.M. Stanfield, J.R. Wilson, G.A. Mirka, N.F. Glasscock, J.P. Psihogios, J.R. Davis. "Multivariate input modeling with Johnson distributions", in *Proceedings of the 28th Winter simulation conference WSC'96*, December 8-11, 1996, Coronado, CA, USA, ed. S.Andradyttir, K.J.Healy, D.H.Withers, and B.L.Nelson, IEEE Computer Society Washington, DC, USA, 1996, pp. 1457-1464.

[5] A.A. Afifi and S.P. Azen. Statistical analysis: a computer oriented approach, New York; London: Academic Press, 1972, 366 p.

[6] S.B. Prykhodko. "Statistical anomaly detection techniques based on normalizing transformations for non-Gaussian data", in *"Computational Intelligence (Results, Problems and Perspectives)", Proceedings of the International Conference*, Kyiv-Cherkasy, Ukraine, May 12-15, 2015, pp. 286-287.

[7] K.V. Mardia. "Measures of multivariate skewness and kurtosis with applications", *Biometrika*, 57, pp. 519–530, 1970.