MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
SUMY STATE UNIVERSITY
UKRAINIAN FEDERATION OF INFORMATICS

PROCEEDINGS

OF THE V INTERNATIONAL SCIENTIFIC CONFERENCE

ADVANCED INFORMATION SYSTEMS AND TECHNOLOGIES

AIST-2017

(Sumy, May 17–19, 2017)

SUMY
SUMY STATE UNIVERSITY
2017

Detecting bivariate outliers on the basis of normalizing transformations for non-Gaussian data

S. Prykhodko, N. Prykhodko, L. Makarova, O. Kudin, T. Smykodub, A. Prykhodko.
Admiral Makarov National University of Shipbuilding, sergiy.prykhodko@nuos.edu.ua, http://ikitn.nuos.edu.ua/

***Abstract* – The statistical technique for detecting outliers in bivariate non-Gaussian data on the basis of normalizing transformations, prediction ellipse and a test statistic (TS) for the Mahalanobis squared distance (MSD), which has an approximate *F* distribution, is proposed. Application of the technique is considered for detecting outliers in two bivariate non-Gaussian data sets: the first, actual effort (hours) and size (adjusted function points) from 145 maintenance and development projects, the second, effort (hours) and mass (tonnes) of designed the section of the ship from 188 designs of sections.**

***Keywords* – outlier; normalizing transformation; bivariate non-Gaussian data; Mahalanobis squared distance; F distribution; prediction ellipse.**

I. INTRODUCTION

An important step in data processing is the outlier detection. Today the problem of outlier detection in a bivariate data set is solved with different methods including statistical [1, 2]. However, well-known statistical methods (for example, bivariate outlier detection based on a prediction ellipse or a test statistic (TS) for the Mahalanobis squared distance (MSD), which has an approximate the *F* distribution) are used to detect outliers in a data set under the assumption that the data is generated by a bivariate Gaussian distribution. And this assumption is valid only in particular cases. In [3] and [4] statistical outlier detection techniques for multivariate non-Gaussian data on the basis of normalizing transformations and MSD, which has an approximate the Chi-Square distribution and the *F* distribution respectively, were proposed. We propose a statistical outlier detection technique for bivariate non-Gaussian data on the basis of normalizing transformations, prediction ellipse and TS for MSD, which has an approximate *F* distribution. The technique consists of two steps. In the first step, bivariate non-Gaussian data is normalized using a bivariate normalizing transformation. In the second step, MSD, prediction ellipse and TS for MSD are calculated and compared with a quantile of the *F* distribution. The data values for which a value of TS 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 TS for MSD will be less than or equal to the quantile of the *F* distribution.

II. THE STATISTICAL TECHNIQUE

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

$$\mathbf{Z} = \psi(\mathbf{X}). \qquad (\square\square\square)$$

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

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

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

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

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

A test statistic for d_i^2 can be created as follows [5]

$$N(N-2) d_i^2 / 2(N^2 - 1), \qquad \square(4)$$

which has an approximate *F* distribution with 2 and $N-2$ degrees of freedom.

The equation for the prediction ellipse is defined by [6].

$$(\mathbf{Z} - \overline{\mathbf{Z}})^T S^{-1} (\mathbf{Z} - \overline{\mathbf{Z}}) = \frac{2(N^2 - 1)}{N(N-2)} F_{2,N-2,\alpha}, \qquad (5)$$

where $F_{2,N-2,\alpha}$ is a quantile of the *F* distribution; α is significance level. We take α as 0.05.

A test statistic for MSD (4) is compared with $F_{2,N-2,\alpha}$. 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 bivariate 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. BIVARIATE 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 [7]

$$\mathbf{Z} = \boldsymbol{\gamma} + \boldsymbol{\eta}\mathbf{h}\left[\boldsymbol{\lambda}^{-1}(\mathbf{X} - \varphi)\right] \sim N_m(0_m, \Sigma), \quad (6)$$

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

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

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

IV. EXAMPLES

We consider the examples of detecting outliers in two bivariate non-Gaussian data sets: the first, actual effort (hours) and size (adjusted function points) from 145 maintenance and development projects [8], the second, effort (hours) and mass (tonnes) of designed the section of the ship from 188 designs of sections.

Table I contains the data from 145 maintenance and development projects [8], MSD and TS for MSD for standardized data sample units, which are [2]

$$Z_{ki} = \left(X_{ki} - \bar{X}_k\right)/S_{X_k} \quad k = 1,2 \; ; \; i = 1,2,\ldots,145. \quad (7)$$

The last column in Table I reveals that projects 3, 9, 38, 51, 101 and 102 are bivariate outliers, since $F_{2,143;0.05}$=3.06.

TABLE III. TS FOR MSD FOR THE STANDARDIZED DATA

Pro - ject	Size (adjusted function points)	Actual effort (hours)	Z_{1i}	Z_{2i}	d_i^2	TS for MS D
1	101.65	485	-0.28088	-0.74421	0.57	0.28
2	57.12	990	-0.31024	-0.46733	0.22	0.11
3	1010.88	13635	0.31859	2.97504	11.17	5.51
4	45.6	1576	-0.31783	0.14631	0.24	0.12
…	…	…	…	…	…	…
9	144.72	584	-0.25248	2.82538	12.41	6.12
…	…	…	…	…	…	…
17	609.7	186	0.05408	-0.15302	0.05	0.02
…	…	…	…	…	…	…
38	172.96	497	-0.23386	2.17432	7.50	3.70
…	…	…	…	…	…	…
51	15.36	462	-0.33777	1.91240	6.29	3.10
…	…	…	…	…	…	…
101	1285.7	548	0.49978	2.55597	7.61	3.75
102	18137.48	946	11.6103	5.53437	135.5	66.83
…	…	…	…	…	…	…
138	698.54	308	0.11266	0.75995	0.70	0.35
139	752.64	217	0.14833	0.07896	0.02	0.01
140	809.25	40	0.18565	-1.24560	2.58	1.27
141	178.1	253	-0.23047	0.34837	0.37	0.18
142	81.48	405	-0.29418	1.48585	3.89	1.92
143	1093.86	241	0.37330	0.25856	0.14	0.07
144	1002.76	156	0.31323	-0.37753	0.52	0.26
145	551.88	92	0.01596	-0.85646	1.05	0.52

Table II contains the normalized data from 145 projects, MSD and TS for MSD for normalized data. These data is normalized by S_U family of the transformation (6). In these case the parameters are such: $\gamma_1 = -1{,}448408$, $\gamma_2 = -0{,}489606$, $\eta_1 = 0{,}717501$, $\eta_2 = 0{,}655549$, $\varphi_1 = 71{,}11167$, $\varphi_2 = 1178{,}5237$, $\lambda_1 = 46{,}09214$ and $\lambda_2 = 513{,}9309$. The sample correlation matrix (3) of the **Z** is used as the approximate moment-matching estimator of correlation matrix Σ

$$S_N = \begin{pmatrix} 0.993109 & 0.716010 \\ 0.716010 & 0.993119 \end{pmatrix}.$$

In Table II the last column reveals that projects 4, 17, 101, 102, 138, 140 and 144 are bivariate outliers, since $F_{2,143;0.05}$=3.06. We note, only for two projects 101 and 102 the results are the same in both cases. For other projects, the results of bivariate outliers do not match. First of all, this is due to poor normalization (or normality) of standardized data by formula (7). It is known that Mardia's multivariate kurtosis [9] β_2 equals 8 under bivariate normality. The values of β_2 equal respectively 131.20 and 8.21 for the data from Table I and Table II. These values indicate that the necessary condition for bivariate normality is practically performed for the normalized data from Table II and does not hold for standardized data from Table I by the formula (7).

The prediction ellipses (Fig. 1 and Fig. 2) indicate on the same results. On Fig. 1 and Fig. 2 the standardized and normalized data set for 145 projects and the prediction ellipses are presented. On Fig. 2 the prediction ellipse (5) also reveals that seven data points (projects 4, 17, 101, 102, 138, 140 and 144) are bivariate outliers as in Table II.

TABLE IV. TS FOR MSD FOR THE NORMALIZED DATA

Project	Normalized size	Normalized actual effort	d_i^2	TS for MSD
1	-1.002326	-1.216116	1.52	0.75
2	-1.662999	-0.724990	3.26	1.61
3	1.212616	2.054895	4.40	2.17
4	-1.827636	-0.023010	6.88	3.39
…	…	…	…	…
9	-0.553410	-0.496653	0.33	0.16
…	…	…	…	…
17	0.814075	-1.140640	6.93	3.42
…	…	…	…	…
38	-0.348010	0.648245	1.82	0.90
…	…	…	…	…
51	-2.181756	-1.010318	5.46	2.69
…	…	…	…	…
101	-0.417388	0.556097	9.29	4.57
102	1.350987	2.154155	13.70	6.76
…	…	…	…	…
138	0.923278	-1.354447	9.42	4.64
139	0.982474	-0.178135	2.62	1.69
140	1.039605	-1.331836	10.17	5.02
141	-0.315695	-1.129770	1.81	0.89
142	-1.288339	-1.344194	2.03	1.00
143	1.273260	0.666751	1.76	0.87
144	1.206397	-0.664230	6.40	3.16
145	0.732924	-0.843953	4.49	2.21

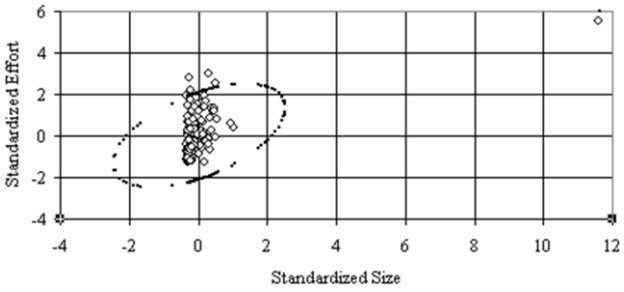

Figure 1. Standardized data set for 145 projects

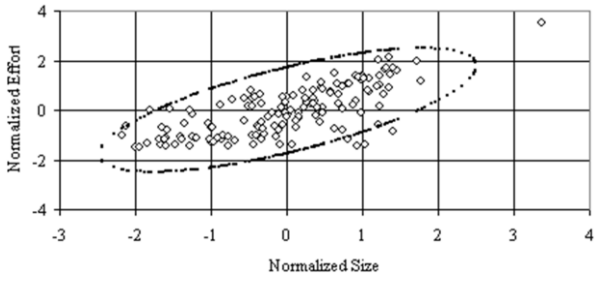

Figure 2. Normalized data set for 145 projects

On Fig. 3 the data set for 145 projects and the transformed prediction ellipse are presented.

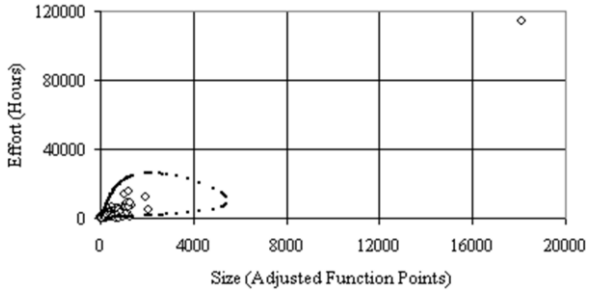

Figure 3. Data set for 145 projects

On Fig. 3 the transformed prediction ellipse also reveals that seven data points (projects 4, 17, 101, 102, 138, 140 and 144) are bivariate outliers. We note, if the anomaly detection technique [10] based on the Grubb test applies for detecting outliers in the normalized data for 145 projects then 144 data sample units do not appear to be an outlier in each of the univariate distributions.

CONCLUSIONS

From the examples we conclude that the proposed technique is promising. For other bivariate non-Gaussian data set of effort and mass of designed the section of the ship from 188 designs of sections the results are similar.

REFERENCES:

[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] S. Prykhodko, N. Prykhodko, L. Makarova and K. Pugachenko, "Multivariate outlier detection technique based on normalizing transformations for non-Gaussian data", unpublished.

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

[6] V. Chew "Confidence, prediction and tolerance regions for the multivariate normal distribution" *Journal of the American Statistical Association*, Vol. 61, Issue 315, pp.605-617, 1966.

[7] 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.

[8] B. Kitchenham, S.L. Pfleeger, B. McColl, and S. Eagan, "An empirical study of maintenance and development estimation accuracy", *The Journal of Systems and Software*, 64, pp.57-77, 2002.

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

[10] 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.