CONSTRUCTING THE TRANSFORMED PREDICTION ELLIPSES ON THE BASIS OF NORMALIZING TRANSFORMATIONS FOR BIVARIATE NON-GAUSSIAN DATA

UDC 004.412:519.237

ПРИХОДЬКО Сергей Борисович
д.т.н., профессор, заведующий кафедрой программного обеспечения автоматизированных систем, Национальный университет кораблестроения имени адмирала Макарова.
Научные интересы: математическое моделирование случайных величин и процессов в информационных технологиях.

ПРИХОДЬКО Наталья Васильевна
к.э.н., доцент, доцент кафедры финансов, Национальный университет кораблестроения имени адмирала Макарова, natalia.prykhodko@nuos.edu.ua
Научные интересы: математическое моделирование случайных величин и процессов в информационных технологиях.

КУДИН Олег Алексеевич
старший преподаватель кафедры морского приборостроения, Национальный университет кораблестроения имени адмирала Макарова
Научные интересы: математическое моделирование случайных величин и процессов в информационных технологиях.

СМЫКОДУБ Татьяна Георгиевна
старший преподаватель кафедры программного обеспечения автоматизированных систем, Национальный университет кораблестроения имени адмирала Макарова.
Научные интересы: математическое моделирование случайных величин и процессов в информационных технологиях.

INTRODUCTION

In the bases of statistical methods of bivariate data analysis, there is the ellipse [1-3]. However, well-known statistical methods (for example, bivariate outlier detection based on a prediction ellipse) are used under the assumption that the data is generated by a bivariate Gaussian distribution. But this assumption is valid in particular cases only. This leads to the need to transform the prediction ellipse for bivariate non-Gaussian data.

We propose a technique for constructing the transformed prediction ellipses on the basis of normalizing transformations for bivariate non-Gaussian data. As and in [4] the technique consists of three steps. In the first step, bivariate non-Gaussian data is normalized using a bijective bivariate normalizing transformation and linear regression is built on the basis of the normalized data. In the second step, the prediction ellipse for the normalized data is built. In the third step, the transformed prediction ellipse for bivariate non-Gaussian data is constructed on the basis of the prediction ellipse for the normalized data and the normalizing transformation.

THE TECHNIQUE

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

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

and the inverse transformation for (1)

$$\mathbf{X} = \psi^{-1}(\mathbf{Z}). \tag{2}$$

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 equation for the prediction ellipse is defined by [3]

$$(\mathbf{Z} - \mathrm{m}_Z)^T \mathbf{S}^{-1} (\mathbf{Z} - \mathrm{m}_Z) = \frac{2(N^2 - 1)}{N(N-2)} F_{2,N-2,\alpha}, \tag{3}$$

where $F_{2,N-2,\alpha}$ is a quantile of the F distribution; m_Z is the mean vector, $\mathrm{m}_Z = (\mathrm{m}_{Z_1}, \mathrm{m}_{Z_2})^T$; α is significance level (we take α as 0.05); **S** is the covariance matrix

$$\mathbf{S} = \begin{pmatrix} S^2_{Z_1} & S_{Z_1Z_2} \\ S_{Z_1Z_2} & S^2_{Z_2} \end{pmatrix}.$$

The left side of the equation (3) can be also merged into

$$\frac{S^2_{Z_1} S^2_{Z_2}}{S^2_{Z_1} S^2_{Z_2} - S^2_{Z_1Z_2}} \left[\frac{(Z_1 - \mathrm{m}_{Z_1})^2}{S^2_{Z_1}} + \frac{(Z_2 - \mathrm{m}_{Z_2})^2}{S^2_{Z_2}} - \frac{2S_{Z_1Z_2}(Z_1 - \mathrm{m}_{Z_1})(Z_2 - \mathrm{m}_{Z_2})}{S^2_{Z_1} S^2_{Z_2}} \right] = \frac{2(N^2-1)}{N(N-2)} F_{2,N-2,\alpha}. \tag{4}$$

The equation for the transformed prediction ellipse for bivariate non-Gaussian data is constructed on the basis of the prediction ellipse (3) for the normalized data and the transformation (1)

$$\frac{[\psi_1(X_1) - \mathrm{m}_{Z_1}]^2}{S^2_{Z_1}} + \frac{[\psi_2(X_2) - \mathrm{m}_{Z_2}]^2}{S^2_{Z_2}} - \frac{2S_{Z_1Z_2}[\psi_1(X_1) - \mathrm{m}_{Z_1}][\psi_2(X_2) - \mathrm{m}_{Z_2}]}{S^2_{Z_1} S^2_{Z_2}} = \frac{2(N^2-1)(S^2_{Z_1} S^2_{Z_2} - S^2_{Z_1Z_2})}{N(N-2) S^2_{Z_1} S^2_{Z_2}} F_{2,N-2,\alpha}. \tag{5}$$

The equation (5) is used for constructing the transformed prediction ellipse for bivariate non-Gaussian data. A same ellipse can be built by the inverse transformation (2) of the values of variables Z_1 and Z_2 from equation (3).

BIVARIATE NORMALIZING TRANSFORMATIONS

Some bivariate transformations have been proposed for normalizing bivariate non-Gaussian data, such as, transformation on the basis of 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]

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

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

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

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

EXAMPLES

We consider the examples of constructing the transformed prediction ellipses for two bivariate non-Gaussian data sets: the first, actual effort (hours) and size (adjusted function points) from 145 maintenance and development projects [6], the second, effort (hours) and mass (tonnes) from 144 designs of ship units. On Fig. 1 the normalized data set for 145 projects and the prediction ellipse for $F_{2,143,0.05} = 3.059$ are presented. The prediction ellipse (dotted line on Fig. 1) reveals that seven data points (projects 4, 17, 101, 102, 138, 140 and 144) are bivariate outliers.

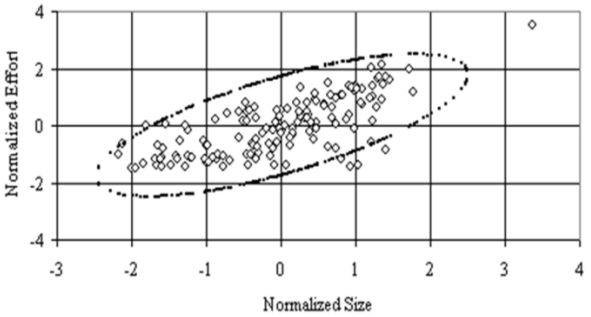

Figure 1. Normalized data set for 145 projects

These data is normalized by S_U family of the transformation (6). In these case the point estimates of 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 covariance matrix of the $\mathbf{Z}$ is used as the approximate moment-matching estimator of covariance matrix **S**

$$\mathbf{S}_N = \begin{pmatrix} 0.993109 & 0.716010 \\ 0.716010 & 0.993119 \end{pmatrix}.$$

On Fig. 2 the data set for 145 projects and the transformed prediction ellipse are presented. The transformed prediction ellipse (dotted line on Fig. 2) indicates on the same results.

Table I contains the mass (tonnes) X_1 and effort (hours) X_2 for 144 designs of ship units.

On Fig. 3 the normalized data set for 144 designs of ship units and the prediction ellipse for $F_{2,142,0.05} = 3.060$ are presented. The prediction ellipse (dotted line on Fig. 3) reveals that eight data points (designs 71, 82, 83, 103, 107, 108, 110 and 142) are bivariate outliers.

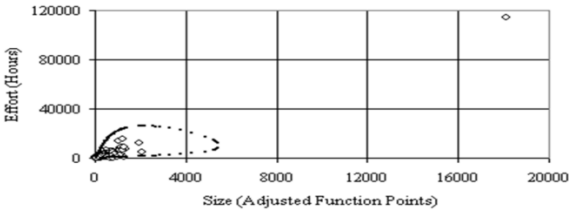

Figure 2. Data set for 145 projects

These data from Table I is normalized by S_B family of the transformation (6). In these case the point estimates of parameters are such: $\gamma_1 = 0.152798$, $\gamma_2 = 0.977402$, $\eta_1 = 0.702754$, $\eta_2 = 0.775194$, $\varphi_1 = -1.83139$, $\varphi_2 = 6.15860$, $\lambda_1 = 104.3435$ and $\lambda_2 = 429.5826$. The sample covariance matrix of the $\mathbf{Z}$ is used as the approximate moment-matching estimator of covariance matrix **S**

$$\mathbf{S}_N = \begin{pmatrix} 1.00000 & 0.38045 \\ 0.38045 & 0.99999 \end{pmatrix}.$$

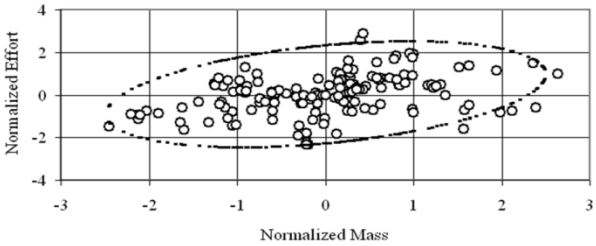

Figure 3. Normalized data set for 144 designs

On Fig. 4 the data set for 144 designs and the transformed prediction ellipse are presented. The transformed prediction ellipse (dotted line on Fig. 4) indicates on the same results: eight data points (designs 71, 82, 83, 103, 107, 108, 110 and 142) are bivariate outliers.

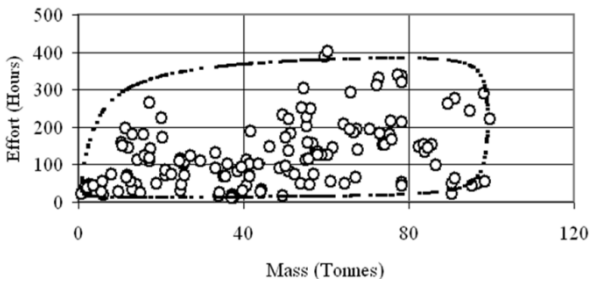

Figure 4. Data set for 144 designs

We note, Mardia's multivariate kurtosis [7] β_2 equals 8 under bivariate normality for our cases. The values of point estimate of kurtosis β_2 equal respectively 8.21 and 7.98 for the normalized data on Fig. 1 and Fig. 3. These values indicate that the necessary condition for bivariate normality is practically performed for the normalized data.

Table I

The mass X_1 and effort X_2 for 144 designs of ship units

No	X_1	X_2	*No*	X_1	X_2	*No*	X_1	X_2	*No*	X_1	X_2
1	40.62	111	37	41.57	68	73	1.69	41	109	64.58	49
2	40.62	45	38	38.53	82	74	2.25	39	110	95.31	44
3	49.51	233	39	43.6	101	75	3.56	43	111	86.64	98
4	49.51	17	40	35.7	67	76	78.19	328	112	61.13	54
5	33.25	132.5	41	11.89	72	77	78.26	51	113	67.24	65
6	33.25	90.5	42	5.66	28	78	78.39	337	114	52.39	74
7	25.54	101	43	17.14	110	79	78.41	44	115	70.5	193
8	24.63	114	44	2.1	39	80	50.93	221	116	55.4	158
9	41.81	188	45	2.1	32	81	50.93	83	117	50.3	173
10	25.67	106	46	27.06	124	82	59.54	388	118	72.78	331
11	21.10	70	47	24.91	47	83	60.36	402	119	94.9	242
12	14.02	46	48	14.94	25	84	77.23	340	120	84.7	144
13	14.02	24	49	18.82	49	85	78.24	319	121	67.68	138
14	5.92	56	50	39.42	93	86	85.62	154	122	59.02	125
15	9.64	28	51	46.29	148	87	82.50	148	123	64.16	208
16	37.07	17	52	12.13	145	88	73.74	154	124	66.04	292
17	37.14	17	53	10.5	159	89	75.40	179	125	56.19	250
18	37.22	18	54	17.46	143	90	67.30	194	126	56.95	74
19	33.97	16	55	10.67	151	91	83.57	154	127	75.81	167
20	44.36	34	56	11.35	197	92	55.31	202	128	66.77	187
21	37.65	12	57	13.16	179	93	61.49	144	129	65.79	193
22	37.25	13	58	12.61	56	94	75.58	215	130	57.27	127
23	37.15	12	59	14.34	113	95	78.18	212	131	54.95	112
24	44.15	26	60	21.04	84	96	60.44	127	132	57.89	134
25	39.81	31	61	25.63	70	97	91.15	275	133	50.95	136
26	24.83	31	62	24.55	109	98	51.36	183	134	50.3	96
27	6.15	20	63	15.88	124	99	55.22	226	135	56.12	46
28	13.08	34	64	15.83	180	100	56.68	156	136	55.86	116
29	35.07	69	65	20.15	224	101	74.16	152	137	57.91	126
30	34.15	24	66	20.38	172	102	83.90	134	138	98.17	289
31	8.01	73	67	17.15	264	103	96.54	48	139	89.51	263
32	41.48	71	68	17.23	118	104	53.91	48	140	54.56	302
33	41.34	101	69	22.64	74	105	90.43	50	141	54.15	252
34	29.69	108	70	2.64	47	106	91.25	64	142	99.56	222
35	48.66	90	71	0.69	23	107	98.35	55	143	72.13	311
36	36.06	100	72	11.53	66	108	90.32	21	144	73.07	183

CONCLUSIONS

From the examples we conclude that the proposed technique for constructing the transformed prediction ellipses is promising. The equations for the transformed prediction ellipses for two bivariate non-Gaussian data sets are constructed on the basis of the Johnson normalizing

translation for S_U and S_B families. Application of these equations for bivariate outlier detection in the bivariate non-Gaussian data sets is demonstrated. The results are similar for the bivariate non-Gaussian data sets of the examples.

REFERENCES

1. Michael Friendly, Georges Monette and John Fox "Elliptical Insights: Understanding Statistical Methods Through Elliptical Geometry" Statistical Science, Vol. 28, No. 1, pp. 1–39, 2013.
2. R.A. Johnson and D.W. Wichern, Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007, 800 p.
3. 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.
4. 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.
5. 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.
6. 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.
7. K.V. Mardia, "Measures of multivariate skewness and kurtosis with applications", Biometrika, 57, pp. 519–530, 1970.

Рецензент: *Dr.Sc., Petro Guchek*
Associate Professor at the Department of information technologies,
Kherson National Technical University,
Institute of biocybernetics and biomedical engineering of M. Nalecha
of the Polish academy of Sciences.