Visualization of Object Relations for Outlier Finding

Takeo Okazaki

Faculty of Engineering, University of the Ryukyus, Okinawa, 903-0213 Japan

Summary
The visualization of the relationship between objects such as distance and influence is one of the useful technique for outlier finding. Asymmetric data are particularly complicated. In this study, I proposed a method of two-dimensional coordinate arrangement according to dissimilarity between objects by decomposing the distortion and the symmetrical part. For the symmetrical part, I proposed an optimization method of deployment as a whole, which was an extension of the optimization techniques of conventional. For the distortion part, I defined the two components of the direction and magnitude to the skewness. The results of application experiments with two cases of real data showed the effectiveness of the outlier detection by the two-dimensional display of the distortion section in particular.

Key words:
2-dimensional visualization, Asymmetric relationship, Outlier

1. Introduction

In order to make the visual perception about the specificity into a readily, geometrical dummies, such as a facial and a body, and a coordinate set may express the poly dimensional data of the sociometric data etc. which are collected in the territory of social science and a behavioral science. Sunagawa [1] defined whenever dissimilarity between objects and expressed the relative relationship. However, by a real data, also when observed as a data whenever asymmetrical dissimilarity, for a certain reason, it is necessary to express not only a far and near relationship but a tensile strain. It was difficult to recognize a far and near relationship and the magnitude of a tensile strain to a directly from the set obtained by the so-far technique, although the inner product and outer product after a set expressed the relative relationship. Then, I proposed the technique from the data in which a direct perception of the relationship between objects is possible whenever unsymmetrical dissimilarity. Thereby, I made the singularity detection in a data into a possible.

2. Dissimilarity

When two or more objects are observed as the same poly dimensional variables, X^q denotes the (m,n)-dimensional data.

$$X^q = \begin{pmatrix} x_1^q \\ x_2^q \\ \vdots \\ x_m^q \end{pmatrix} \quad (q = 1,2,\cdots,n) \tag{1}$$

x_i^q in actual data can be assumed as an ordinal scale with the exception that real number. However, the real number is suitable for the definition of dissimilarity, rankit transformation is applied. Rankit transformation formula is given by Formula (2) in n<50 case.

$$\left(x_i^q\right)' = \frac{n!}{(x_i^q - 1)!(n-1)!} \times \int_{-\infty}^{\infty} \left\{\Phi(x_i^q)\right\}^{x_i^q - 1} \left\{1 - \Phi(x_i^q)\right\}^{n-1} \phi(x_i^q) x_i^q dx_i^q \tag{2}$$

$$\phi(x_i^q) = \Phi'(x_i^q) = \int_z^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x_i^q)^2}{2}\right\} dx_i^q \tag{3}$$

Then dissimilarity between objects i , j from (m,n) -dimensional data X^q are defined as Formula (4).

$$e_{ij} = \sqrt{\sum_{p=1}^{m} \left(X_p^{(i)} - X_p^{(j)}\right)^2} \tag{4}$$

Now we can get the dissimilarity matrix $E = \{e_{ij}\}$ $(1 \le i, j \le n)$.

In this dissimilarity matrix, $e_{ij} = e_{ji}, e_{ii} = 0$. On the other hand, there are some cases that observations are gotten as dissimilarity matrix, that means asymmetric matrix. Then I expanded to the dissimilarity matrix $S = \{s_{ij}\}$ $(1 \le i, j \le n)$ including asymmetric type.

When the expanded dissimilarity matrix is placed on the 2-dimensional area, it is necessary to express the skewness of dissimilarity besides the distance between objects.

Manuscript received September 5, 2014
Manuscript revised September 20, 2014

Chino[2] gave the placement in ASYMSCAL by setting the objective function that evaluated the consistency of inner and cross product between the dissimilarity and the placement. This procedure can express the direction of skewness, but it is hard to get the straightforward recognition from the resultant placement, because the procedure shows only the largeness of inner or cross product for the distance and skewness.

In this study, I decomposed S uniquely to symmetric part $S^{\#}$ and skewness part S^{*} utilizing that dissimilarity matrix S was square matrix.

$$S^{\#} = \frac{1}{2}\left(S + S^{T}\right) \tag{5}$$

$$S^{*} = \frac{1}{2}\left(S - S^{T}\right) \tag{6}$$

Where S^{T} is the transposed matrix of S. Symmetric part expresses the distance and skewness part expresses the direction or the largeness of skewness, we can get the placement to enable the direct recognition.

3. Placement of symmetrical part

3.1 Placement procedure

In the placement based on the symmetric matrix $S^{\#} = \left\{s_{ij}^{\#}\right\}$ $(1 \le i, j \le n)$, the distance d_{ij} between arbitrary two points $i(x_i^{\#}, y_i^{\#})$, $j(x_j^{\#}, y_j^{\#})$ can be gotten by the Formula (7).

$$d_{ij} = \sqrt{\left(x_i^{\#} - x_i^{\#}\right)^2 + \left(y_i^{\#} - y_i^{\#}\right)^2} \ge 0 \tag{7}$$

R is the correlation coefficient between $s_{ij}^{\#}$ and d_{ij} to evaluate the consistency between the distance d_{ij} and the element of the symmetric matrix $s_{ij}^{\#}$.

$$R = \frac{\frac{1}{N}\sum_{i<j}\left(s_{ij}^{\#} - \bar{s}^{\#}\right)\left(d_{ij} - \bar{d}\right)}{\sqrt{\frac{1}{N}\sum_{i<j}\left(s_{ij}^{\#} - \bar{s}^{\#}\right)^2 \frac{1}{N}\sum_{i<j}\left(d_{ij} - \bar{d}\right)^2}} \tag{8}$$

$$N = \frac{n(n-1)}{2} \tag{9}$$

$$\bar{s}^{\#} = \frac{1}{N}\sum_{i<j} s_{ij}^{\#} \tag{10}$$

$$\bar{d} = \frac{1}{N}\sum_{i<j} d_{ij} \tag{11}$$

The correlation coefficient R has a real number from -1 to 1, and the value 1 means the highest consistency. Then it is enough to find the placement that maximizes the correlation coefficient.

Especially, when all $s_{ij}^{\#}$ values are equal, the variance is zero, then the correlation coefficient cannot be derived. However, it is trivial that the placement of all objects has equally-spaced intervals.

3.2 Placement normalization

Though the placement derived from the dissimilarity is relative, there are an infinite number of placements those give a same correlation coefficient. So we need a standardization of the placement. A proposal standardization procedure by the transformation of homogeneous coordinate expression is as follows.

Step 1. Set all means of *x* and *y* coordinate to $\bar{x}$ and $\bar{y}$, translate $A(\bar{x}, \bar{y})$ to be the position of the origin.

$$S = \begin{bmatrix} 0 & 0 & -\bar{x} \\ 0 & 0 & -\bar{y} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_m^i \\ y_m^i \\ 1 \end{bmatrix} \tag{12}$$

Step 2. Set the variances of *x* and *y* coordinate to V_x and V_y respectively, change scale as the sum of V_x and V_y to be 1.

$$Z = \begin{bmatrix} \frac{1}{V_x + V_y} & 0 & 0 \\ 0 & \frac{1}{V_x + V_y} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_m^i \\ y_m^i \\ 1 \end{bmatrix} \tag{13}$$

Step 3. Calculate the norm of all points, rotate them so that the point with biggest norm to be on the *x* coordinate and *x*>0.

$$R = \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_m^i \\ y_m^i \\ 1 \end{bmatrix} \quad (0 \le \alpha \le 2\pi) \tag{14}$$

Step 4. Turn around so that the point with second bigger norm to be *y*>0.

$$M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\beta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_m^i \\ y_m^i \\ 1 \end{bmatrix} \quad (\beta = 0, \pi) \tag{15}$$

Note that if there are more than 2 points with biggest norm at Step 3 and Step 4, the point with biggest *x* coordinate is chosen. All transformation at each step don't change the correlation coefficient, arbitrary combination transformation of them have a same correlation coefficient.

3.3 Optimization

Though it is not necessarily obtained algebraic solutions at the maximum of the correlation coefficient, we need some optimization procedure. In this study, BFGS method as quasi-Newton's method and Powell method as the conjugate gradient method were improved, and compared.

The traditional BFGS method [3] updates the one point until the derivative of the point converges to zero. To optimize the whole placement, the proposal calculates derivative of all points at each step, updates the point with maximum derivative. The proposal optimization algorithm is as follows. Note that $\nabla R(x)$ is a partial differential of the correlation coefficient by *x*.

Step 1. Prepare the positive definite matrix $H^k = H_0 \in R^2$ that is symmetrical to the starting point $x^k \in R^2$ $(k = 1, \cdots, n)$ at *n* points.

Step 2. Calculate $\nabla R(x)$ for all points, set the point with a maximum $\nabla R(x)$ as update point.

Step 3. Calculate $d^k = -H^k \nabla^T R(X_0)$.

Step 4. If $\nabla f(X_0) d^k \geq 0$, then set $H^k = H_0$ and back to Step 3.

Step 5. Calculate α with Wolfe linear search method [4], and calculate $X_1 = X_0 + \alpha d^k$.

Step 6. Set $\delta = X_1 - X_0$, $\gamma = \nabla^T R(X_1) - \nabla^T R(X_0)$

and calculate.

$$H^k = \left\{ I - \frac{\delta(\gamma)^T}{(\delta)^T \gamma} \right\} H^k \left\{ I - \frac{\gamma(\delta)^T}{(\delta)^T \gamma} \right\} + \frac{\delta(\delta)^T}{(\delta)^T \gamma} \tag{16}$$

Step 7. If $\nabla R(X_1) = 0$ then exit. Otherwise back to Step 2.

Zangwill Powell method [5] concentrates the one point and optimizes in a similar way of BFGS method. The proposal optimization algorithm applies the Powell method to all points, and selects the update point due to the magnitude of the direction vector as follows.

Step 1. Prepare the linearly independent vector $d_0^k, d_1^k \in R^2$ $(k = 1, \cdots, n)$ and the starting point $x_0^k \in R^2$ at each *n* point.

Step 2. Calculate α_i $(i = 1,2)$ by the secondary completion method [6], and $x_{i+1}^k = x_i^k + \alpha_i d_i^k$.

Step 3. Set $d_2^k = x_2^k - x_0^k$, and calculate α_i by the secondary completion method, and $x_3^k = x_2^k + \alpha_2 d_2^k$.

Step 4. If $\left\| x_3^k - x_0^k \right\| = 0$ then stop.

Step 5. Select *p* according to $\left\| x_{p+1}^k - x_p^k \right\| = \max_{0 \leq i \leq 1} \left\| x_{i+1}^k - x_i^k \right\|$.

If $\dfrac{\left\| x_{p+1}^k - x_p^k \right\|}{\left\| d_n^k \right\|} < 1$, then keep the value d_i^k, otherwise set

$$d_i^k = \begin{cases} d_i^k, & i \neq p \\ d_2^k, & i = p \end{cases}.$$

Step 6. Execute the operation one time from Step 1 to Step 5 for all coordinate points, update the coordinate $x_0^k = x_3^k$ at the point with max $\left\| d_2^k \right\|$.

3.4 Verification experiments

To confirm the validity of the proposal, 2 kinds of simulated data were prepared. Initial placement was resolved by the standardization of the simulated data using a random search method. The average of replicate experiments 10 times were used as the optimization results.

Two conditions to generate simulated data were considered as follows.

(1) Case includes dissimilarity 0

(2) Case includes equivalent elements in the dissimilarity matrix

Figure 1 shows the simulation placement considering the first condition.

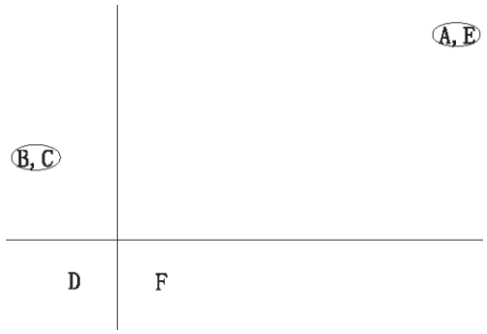


Fig.1: Simulation placement 1

Distance matrix obtained by the standardization of this placement was assumed as dissimilarity matrix, and set as simulated data 1. That is, simulated data 1 has a solution with correlation coefficient 1.

Table 1: Dissimilarity matrix (Simulation data 1)

	B	C	D	E	F
A	10.4403	10.4403	10.8162	0.0000	9.2195
B		0.0000	3.1623	10.4403	4.2426
C			3.1623	10.4403	4.2426
D				10.8167	2.0000
E					9.2195

Figure 2 shows the simulation placement considering the second condition.

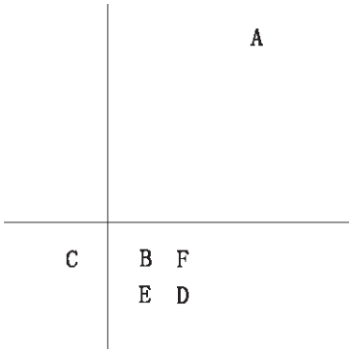


Fig.2: Simulation placement 2

Distance matrix obtained by the standardization of the placement Figure 2 was set as simulated data 2. Also simulated data 2 has a solution with correlation coefficient 1.

Table 2: Dissimilarity matrix (Simulation data 2)

	B	C	D	E	F
A	6.7082	7.8103	7.2801	7.6158	6.3246
B		2.0000	1.4142	1.0000	1.0000
C			3.1623	2.2361	3.0000
D				1.0000	1.0000
E					1.4142

The verification results of the simulated data 1 by proposal modified BFGS method was shown in Table 3.

Table 3: Optimization results (Data 1, BFGS)

Initial correlation coefficient	Mean	0.9314
	Variance	0.0085
Convergence correlation coefficient	Mean	1.0000
	Variance	0.0000
Average number of iterations		40.9
Number of convergence to 1		10

The convergence correlation coefficient was 1 for all initial patterns. Figure 3 shows the average placement in this case. The marks A, B, C, D, E, F denote the ideal placement and the marks a, b, c, d, e, f denote the resultant placement.

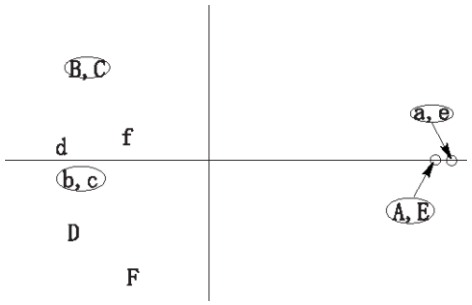


Fig.3: Optimal placement (Data 1, BFGS)

The verification results of the simulated data 1 by proposal modified Powell method was shown in Table 4.

Table 4: Optimization results (Data 1, Powell)

Initial correlation coefficient	Mean	0.9314
	Variance	0.0085
Convergence correlation coefficient	Mean	0.9991
	Variance	0.0005
Average number of iterations		19.8
Number of convergence to 1		2

The number of iterations was less than modified BFGS method case. The convergence value was stable, but the number of convergence to 1 was few. Figure 4 shows the average placement in this case.

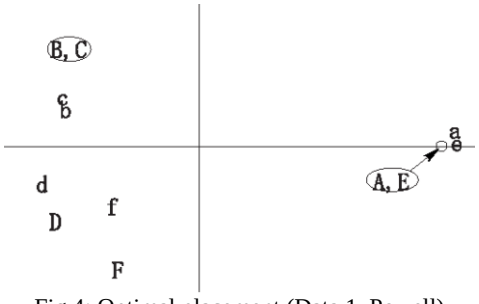


Fig.4: Optimal placement (Data 1, Powell)

The verification results of the simulated data 2 by modified BFGS method was shown in Table 5.

Table 5: Optimization results (Data 2, BFGS)

Initial correlation coefficient	Mean	0.9328
	Variance	0.0058
Convergence correlation coefficient	Mean	1.0000
	Variance	0.0000
Average number of iterations		45.2
Number of convergence to 1		10

The convergence correlation coefficient was 1 for all initial patterns. Figure 5 shows the average placement in this case.

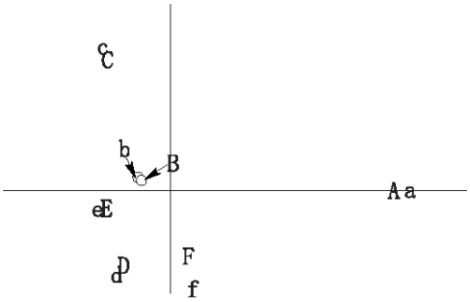


Fig.5: Optimal placement (Data 2, BFGS)

Table 6 shows the result of the modified Powell method.

Table 6: Optimization results (Data 2, Powell)

Initial correlation coefficient	Mean	0.9328
	Variance	0.0058
Convergence correlation coefficient	Mean	0.9999
	Variance	0.0000
Average number of iterations		66.1
Number of convergence to 1		7

The convergence correlation coefficient was 1 for most of initial patterns, though the variance of the iteration was large. Figure 6 shows the average placement in this case.

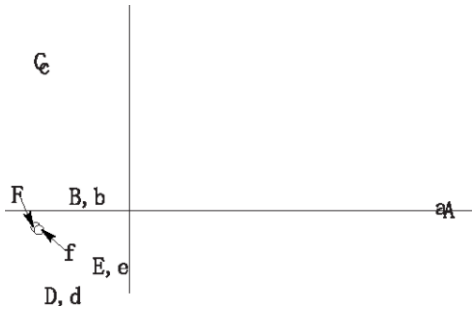


Fig.6: Optimal placement (Data 2, Powell)

From the above results and differentiability of correlation coefficient R, modified BFGS method was well suited for this optimization problem.

4. Placement of distortional part

In the distortion matrix $S^* = \{s^*_{ij}\}(1 \le i, j \le n)$, $s^*_{ij} = -s^*_{ji}$ is realized.

I defined the representation of the distortion relationship by allocating this matrix to 2 coordinates those are skewness magnitude and direction respectively.

For the skewness magnitude, in order to express which of the positive trend value and the negative trend value was larger, I defined as the magnitude of the distortion of each object the sum of the distortion of up to *n*-1 that each object had. The magnitude of the distortion of each object was as follows.

$$x^*_j = \sum_{i=1}^{n} s^*_{ij} \quad (j = 1, \cdots, n) \tag{17}$$

For the skewness direction, in order to express which of the positive trend number and the negative trend number was superior, the code of each element in the distortion matrix was converted to numeric, and the sum of the value was defined as skewness direction. The skewness direction of each object was as follows.

$$W(x)_i = \begin{cases} 1 & (s^*_{ij} > 0) \\ 0 & (s^*_{ij} = 0) \\ -1 & (s^*_{ij} < 0) \end{cases} \tag{18}$$

$$y^*_j = \sum_{i=1}^{n} W(x)_i \quad (j = 1, \cdots, n) \tag{19}$$

Formula (17), (18) and (19) gave the 2-dimensional placement of the distortion matrix S^*. Figure 7 shows the interpretation of the placement.

Fig.7: Interpretation of distortional part

5. Application to actual case

5.1 Trade volume of bearing

Table 7 shows that trade volume of the bearing among six countries (Japan, the United States, the United Kingdom, France, Germany, Italy) in 2002 by the Japan Bearing Industrial Association [7].

Table 7: Trade volume of bearing in 2002 (unit:1000yen)

	JPN	USA	GBR	FRA	DEU	ITA
JPN		1,091,551	170,232	45,286	227,318	66,380
USA	3,117,595		272,679	402,374	1,058,599	217,799
GBR	320,476	564,432		495,764	928,800	331,788
FRA	301,531	794,984	423,636		1,771,934	626,326
DEU	1,077,315	491,433	681,613	1,353,802		2,093,474
ITA	49,154	256,429	445,920	1,200,122	2,001,529	

Assuming this data as dissimilarity matrix, Table 8 shows the symmetrical part and Table 9 shows the distortional part.

Table 8: Symmetrical part of trade volume

	USA	GBR	FRA	DEU	ITA
JPN	2,105,573	245,354	173,408.5	652,316.5	57,767
USA		418,555.5	598,679	775,016	237,114
GBR			459,700	805,206.5	388,854
FRA				1,562,868	913,224
DEU					2,047,501.5

Table 9: Distortional part of trade volume

	JPN	USA	GBR	FRA	DEU	ITA
JPN		-1,013,022	-75,122	-128,122.5	-424,998.5	8613
USA	1,013,022		-145,876.5	-196,305	283,583	-19,315
GBR	75,122	145,876.5		36,064	123,593.5	-57,066
FRA	128,122.5	196,305	-36,064		209,066	-286,898
DEU	424,998.5	-283,583	-123,593.5	-209,066		45,972.5
ITA	-8,613	19,315	57,066	286,898	-45,972.5	

Figure 8 shows the symmetrical part placement and Figure 9 shows the distortional part placement. The correlation coefficient of this symmetrical part was 0.767939.

Fig.8: 2-dimensional placement of symmetrical part

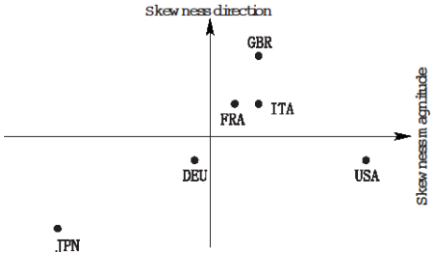


Fig.9: 2-dimensional placement of distortional part

5.2 Tourists from abroad

Table 10 shows the data of foreign tourists of nine countries (Japan, South Korea, China, Taiwan, Hong Kong, Malaysia, Singapore, Thailand, Australia) across in the Asia-Pacific Ocean region of 2000 due to the Singapore Tourism Board: Country Report: A Comparative Study of Visitor Arrivals to Selected Asia Destinations 2000.

Table 10: The number of tourists from foreign countries in the Asia-Pacific Ocean region of 2000 (unit:1000person)

	JPN	KOR	CHN	TWN	HKG	MYS	SGP	THA	AUS
JPN		2,472	2,202	916	1,382	456	930	1,207	720
KOR	1,064		1,345	84	373	72	354	448	160
CHN	352	443		0	3,786	425	434	704	124
TWN	913	127	0		2,386	213	291	712	135
HKG	49	201	0	361		76	286	495	0
MYS	62	60	441	58	315		565	1,056	153
SGP	73	83	399	95	451	5,420		660	276
THA	65	88	241	133	229	940	247		74
AUS	147	40	234	32	352	237	510	326	

Assuming this data as dissimilarity matrix, Table 11 shows the symmetrical part and Table 12 shows the distortional part.

Table 11: Symmetrical part of the number of tourists

	KOR	CHN	TWN	HKG	MYS	SGP	THA	AUS
JPN	1,768	1,277	914.5	715.5	259	501.5	636	433.5
KOR		894	105.5	287	66	218.5	268	100
CHN			0	1,893	433	416.5	472.5	179
TWN				1,373.5	135.5	193	422.5	83.5
HKG					195.5	368.5	362	176
MYS						2,992.5	998	195
SGP							453.5	393
THA								200

Table 12: Distortional part of the number of tourists

	JPN	KOR	CHN	TWN	HKG	MYS	SGP	THA	AUS
JPN		704	925	1.5	666.5	197	428.5	571	286.5
KOR	-704		451	-21.5	86	6	135.5	180	60
CHN	-925	-451		0	1,893	-8	17.5	231.5	-55
TWN	-1.5	21.5	0		1,012.5	77.5	98	289.5	51.5
HKG	-666.5	-86	-1,893	-1,012.5		-119.5	-82.5	133	-176
MYS	-197	-6	8	-77.5	119.5		-2,427.5	58	-42
SGP	-428.5	-135.5	-17.5	-98	82.5	2,424.5		206.5	-117
THA	-571	-180	-231.5	-289.5	-133	-58	-206.5		-126
AUS	-286.5	-60	55	-51.5	176	42	117	126	

Figure 10 shows the symmetrical part placement and Figure 11 shows the distortional part placement. The correlation coefficient of this symmetrical part was 0.680421.

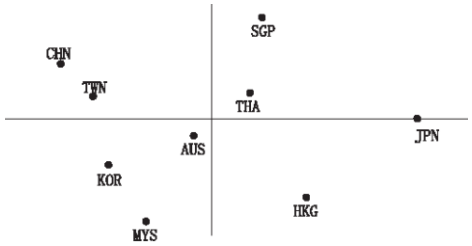


Fig.10: 2-dimensional placement of symmetrical part

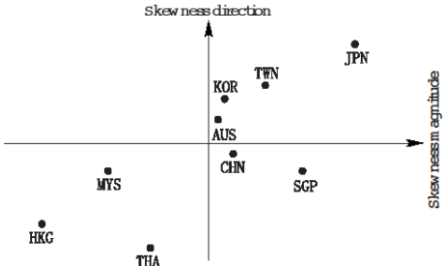


Fig.11: 2-dimensional placement of distortional part

6. Conclusion

I proposed a method of two-dimensional coordinate arrangement according to dissimilarity between objects that are characterized by multi-dimensional data. By decomposing the distortion part and the symmetrical part, a two-dimensional arrangement of each could be expressed as can be recognized directly from the arrangement direction of the distortion and perspective between objects.

For the symmetrical part, I defined an evaluation function using the correlation coefficient of the distance and placement, proposed an optimization method of deployment as a whole, which was an extension of the optimization techniques of conventional, showed the effectiveness by numerical simulation. For the distortion part, I defined the two components of the direction and magnitude to the skewness. The results of application experiments with two cases of real data showed the effectiveness of the outlier detection by the two-dimensional display of the distortion section in particular.

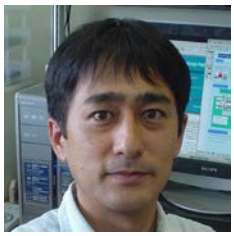
It is expected that the recognition of the coordinates disposed becomes difficult when object number is increased. It can be accommodated by the pre-processing such as reducing the number of objects like cluster analysis.

Acknowledgment

The authors would like to express their cordial thanks to Mr. Yoshichika Sunagawa for his experimental support.

References

[1] Y. Sunagawa, “A coordinates placement by dissimilarity between objects”, Proceedings of Hinokuni Information Symposium 2002, pp.311-317, 2002.
[2] N. Chino, “Asymmetric Multidimensional Scaling”. Gendai-Sugakusha, 1997.
[3] D. F. Shanno,“Conditioning of quasi-newton methods for function minimization”, Mathematics of Computation, Vol.24, pp.647-656, 1970.
[4] P. Wolfe, “A method of convergence conditions for ascent methods”, SIAM Review, Vol.11, pp.226-235, 1969.
[5] W. T. Zangwill, “Minimizing a function without calculating derivatives”, Computer Journal, Vol.10, pp.293-296, 1967.
[6] M. J. D. Powell, "On the convergence of variable metric algorithms”, Journal of Institute Mathematical Application, Vol.7, pp.21-36, 1970.
[7] Japan Bearing Industry Association, “Trade Report 2002”, http://www.jbia.or.jp/stat/main.html



Takeo Okazaki received the B.Sci. and M.Sci. degrees from Kyushu University in 1987 and 1989, and D.Eng. degree from University of the Ryukyus in 2014, respectively. He had been a research assistant at Kyushu University from 1989 to 1995. He has been an assistant professor at University of the Ryukyus since 1995. His research interests are statistical data normalization for analysis, statistical causal relationship analysis. He is a member of JSCS, IEICE, JSS, GISA, and BSJ Japan.