Journal of the Korean
Statistical Society
Vol. 21, No. 1, 1992

GAUSS DISCREPANCY TYPE MEASURE OF DEGREE OF RESIDUALS FROM SYMMETRY FOR SQUARE CONTINGENCY TABLES

Sadao Tomizawa[1] and Mariko Murata[2]

ABSTRACT

A measure is proposed to represent the degree of residuals from the symmetry model for square contingency tables with nominal categories. The measure is derived by modifying the sum of squared singular values for a skew symmetric matrix of the residuals from the symmetry model. The proposed measure would be useful for comparing the degree of residuals from the symmetry model in several tables.

1. INTRODUCTION

For two-way contingency tables, van der Heijden *et al.* (1989) uses correspondence analysis (CA) and one of its generalization as tools for the analysis of residuals from various log-linear models. As described in van der Heijden *et al.* (1989), the

[1]Department of Information Sciences, Faculty of Science & Technology, Science University of Tokyo, Noda City, Chiba, 278, Japan
[2]Department of Statistical Science, School of Mathematical and Physical Science, The Graduate University for Advanced Studies, The Institute of Statistical Mathematics 4-6-7 Minami-Azabu, Minato-ku, Tokyo, 106, Japan

singular value decomposition (SVD) of a matrix of residuals that are left after fitting a specific restricted log-linear model, could help to find the structure in residuals. In particular, for square contingency tables, van der Heijden *et al.* (1989) considers the SVD of a skew symmetric matrix (denoted by N) of which elements are residuals from the symmetry model (see Section 2), and analyzes the structure in residuals from the model with CA approach.

For square contingency tables, Tomizawa (1989) proposed two kinds of measures to represent the degree of departure from the symmetry model. In it, measures are expressed by using the Kullback-Leibler information and the Pearson's chi-squared type discrepancy (see Appendix). We are now interested in representing the degree of residuals with a single summary measure for the matrix of residuals from the symmetry model, *i.e.*, for the skew symmetric matrix N described above.

The purpose of this paper is to propose a measure to represent the degree of residuals from the symmetry model. The measure is obtained by modifying the sum of squared singular values (SSSV) of matrix N. It would be useful for comparing the degree of residuals from the symmetry model in several tables.

2. SINGULAR VALUE DECOMPOSITION OF RESIDUALS FROM SYMMETRY

This section describes briefly the SVD of a matrix of residuals from the symmetry model described in van der Heijden *et al.* (1989). Consider an $R \times R$ contingency table P of nominal categories with observed proportions p_{ij} $(i = 1, 2, \dots, R; j = 1, 2, \dots, R)$. The square matrix P is decomposed as

$$P = M + N, \tag{2.1}$$

where M is a symmetric matrix with elements $m_{ij} = m_{ji} = (p_{ij} + p_{ji})/2$, and N is a skew symmetric matrix with elements $n_{ij} = -n_{ji} = (p_{ij} - p_{ji})/2$. Let π_{ij} denote the probability for cell (i, j) of the $R \times R$ table. The symmetry model is defined by

$$\pi_{ij} = \pi_{ji} \quad \text{for } i = 1, 2, \dots, R; j = 1, 2, \dots, R,$$

(see Bishop *et al.*, 1975, p.282). Assume that the observations have a multinomial distribution. The maximum likelihood estimates (MLEs) of π_{ij} under this model are equal to m_{ij}. Therefore, N is a matrix of residuals that are left after fitting the symmetry model. The SVD of N has the form

$$N = U\Lambda JU',$$

where U is orthogonal and J is a block diagonal matrix with blocks $\begin{matrix} 0 & 1 \\ -1 & 0 \end{matrix}$ and the singular values are ordered in pairs $\lambda_1, \lambda_1, \lambda_2, \lambda_2, \ldots$. When R is odd, $\lambda_R = 0$ and $J_R = 0$. The SSSV of N is expressed as

$$\text{trace}\,\Lambda^2 = \text{trace}\,NN' = \text{trace}\,(P - M)(P - M)'.$$

[Note that trace Λ^2 is often called the 'total inertia' in CA (see van der Heijden *et al.*, 1989, p.268).]

3. A MEASURE OF DEGREE OF RESIDUALS FROM SYMMETRY

Let $\delta = \sum\sum_{i \neq j} p_{ij}$. Assuming that $\delta > 0$, consider a measure defined as in the sample version,

$$\psi = \frac{2}{\delta^2}\text{trace}\,\Lambda^2.$$

This measure may be also expressed as

$$\psi = \frac{1}{\delta^2}\sum\sum_{i<j}(p_{ij} - p_{ji})^2.$$

Put $q_{ij} = p_{ij}/\delta$. Then ψ may be further expressed as

$$\psi = 2\Delta_G(q; q^S),$$

where

$$\Delta_G(q; q^S) = \sum\sum_{i \neq j}(q_{ij} - q_{ij}^S)^2,$$

$$q_{ij}^S = (q_{ij} + q_{ji})/2.$$

Note that $\Delta_G(q; q^S)$ is the Gauss discrepancy metric (see Linhart and Zucchini, 1986, p.18) between $\{q_{ij}\}_{i \neq j}$ and $\{q_{ij}^S\}_{i \neq j}$ with $\sum\sum_{i \neq j} q_{ij} = \sum\sum_{i \neq j} q_{ij}^S = 1$. It is easily seen that ψ must lie between 0 and 1. And (i) there is a structure of symmetry in table P [that is, trace $\Lambda^2 = 0$] if and only if $\psi = 0$, and (ii) there is such a structure that the degree of residuals from symmetry is largest in a sense [that is , trace Λ^2 is maximum on condition that the observations fall in the off-diagonal cells of square table (then trace $\Lambda^2 = \delta^2/2$)] if and only if $\psi = 1$. According to the Gauss discrepancy metric, the degree of residuals from symmetry increases as the value of

ψ increases. The population version of ψ, *i.e.*, ψ^*, is given by ψ with $\{p_{ij}\}$ replaced by $\{\pi_{ij}\}$. Thus,

$$\psi^* = \frac{1}{(\delta^*)^2} \sum\sum_{i<j} (\pi_{ij} - \pi_{ji})^2,$$

where $\delta^* = \sum\sum_{i \neq j} \pi_{ij}$.

Let f_{ij} denote the observed frequency in cell (i, j) of the $R \times R$ table, and let $n = \sum\sum f_{ij}$ (note that $p_{ij} = f_{ij}/n$). We shall consider an approximate standard error and large-sample confidence interval of ψ^*, using the *delta method* of which descriptions are given by Bishop *et al.* (1975, Sec.14.6). Using the delta method, $\sqrt{n}(\psi - \psi^*)$ has asymptotically (as $n \to \infty$) a normal distribution with mean zero and variance

$$(\sigma^*)^2 = \frac{4}{(\delta^*)^4} [\sum\sum_{i<j} (\pi_{ij} - \pi_{ji})^2 (\pi_{ij} + \pi_{ji}) - (\delta^*)^3 (\psi^*)^2].$$

For details, see Appendix. Note that this asymptotic distribution is not applicable whenever $\psi^* = 0$ and 1 because then $(\sigma^*)^2 = 0$. Let σ^2 denote $(\sigma^*)^2$ with $\{\pi_{ij}\}$ replaced by $\{p_{ij}\}$. The term $\sigma/\sqrt{n}$ is an estimated approximate standard error of ψ, and $\psi \pm z_{p/2}\sigma/\sqrt{n}$ is an approximate $100(1-p)$ percent confidence interval for ψ^*, where $z_{p/2}$ is the percentage point from the standard normal distribution corresponding to a two-tail probability equal to p.

4. EXAMPLES

Table 1 taken from Fienberg (1980, p.24) is the repeat-victimization data. Table 2 taken from Escofier (1984) is home-work traffic data of which the frequency in each cell gives the number of persons which live in one south-eastern suburb of Paris and work in another. Table 3 taken from Harshman *et al.* (1982), deals with car changing data which were surveyed in 1979 for new car buyers to collect information on their old and new car.

When the degrees of residuals from the symmetry model in Tables 1, 2, and 3 are compared using the confidence interval for ψ^* (see Table 4), it would be easily seen that (i) it in Table 3 is the strongest of these tables, (ii) it in Table 2 is the second strongest, and (iii) it in Table 1 is the weakest of them.

Note: For the data in Tables 1, 2, and 3, the values of the likelihood ratio statistic for testing the goodness-of-fit of the symmetry model are 31.1, 12805.7, and 229357, respectively, with 28, 105, and 120, degrees of freedom, respectively. Therefore the symmetry model fits the data in Table 1 well, but fits the other data very poorly.

5. REMARKS

The followings are remarked:

(I) The structure of symmetry based on the proportions $\{p_{ij}\}$, *i.e.*, $p_{ij} = p_{ji}$ for $i \neq j$, may be also expressed as $q_{ij} = q_{ji}$ for $i \neq j$, using the conditional proportions $\{q_{ij}\}$. So, in parallel to (2.1), we could get

$$\widetilde{P} = \widetilde{M} + \widetilde{N}, \tag{5.1}$$

where $\widetilde{P}$ is a matrix with elements $\tilde{p}_{ij} = q_{ij}$ for $i \neq j$ and $\tilde{p}_{ii} = 0$; $\widetilde{M}$ is a symmetric matrix with elements $\widetilde{m}_{ij} = \widetilde{m}_{ji} = (q_{ij}+q_{ji})/2$ for $i \neq j$ and $\widetilde{m}_{ii} = 0$, and $\widetilde{N}$ is a skew symmetric matrix with elements $\tilde{n}_{ij} = -\tilde{n}_{ji} = (q_{ij}-q_{ji})/2$ for $i \neq j$ and $\tilde{n}_{ii} = 0$. Then we can see easily that ψ is equal to twice the SSSV of $\widetilde{N}$, *i.e.*, $2\text{trace}\widetilde{\Lambda}^2 = 2\Delta_G(q; q^S)$.

(II) The symmetry model imposes no restriction on the diagonal cell probabilities $\{\pi_{ii}\}$. Therefore, the MLEs of π_{ii} under this model are equal to p_{ii}, namely, the residuals for these cells are always theoretically zero without depending on the diagonal proportions $\{p_{ii}\}$. Thus, the degree of residuals also might be considered for only the off-diagonal cells of table P in (2.1) [*i.e.*, for table $\widetilde{P}$ in (5.1)].

(III) The range of SSSV of N depends on the diagonal proportions $\{p_{ii}\}$ because of $0 \leq 2\text{trace}\,\Lambda^2 \leq \delta^2$ $[= (1 - \sum p_{ii})^2]$. On the other hand, the range of SSSV of $\widetilde{N}$ does not depend on the diagonal proportions $\{p_{ii}\}$ because of $0 \leq 2\text{trace}\,\widetilde{\Lambda}^2 \leq 1$. Therefore, as a measure, twice the SSSV of $\widetilde{N}$ (*i.e.*, ψ) rather than twice the SSSV of N would be useful for comparing the degree of residuals from the symmetry model in several tables.

(IV) Measure ψ would be preferable to measures $\hat{\phi}_s$ and $\hat{\psi}_s$ considered by Tomizawa (1989; also see Appendix) when one wants to see the degree of residuals for the skew symmetric matrix N with a single summary measure, *i.e.*, when one wants to see with a single summary measure how far the observed proportions matrix P is distant from the symmetric matrix M [though $\hat{\phi}_s(\hat{\psi}_s)$ may be useful when one wants to see with a minimum Kullback-Leibler distance measure (Pearson's chi-squared type discrepancy measure) how far the sample cell probability distribution is distant from the estimated cell probability distribution with symmetry].

(V) If ψ from table A is smaller than that from table B but the confidence interval from table A is included in that from table B, then the comparison between the degrees of the residuals in tables A and B would be impossible and the suitable interpretation could not be obtained.

(VI) Even if ψ is very close to either 0 or 1, the estimated standard error of ψ is not always close to 0 because that of ψ depends on the sample size n.

ACKNOWLEDGEMENTS

The authors would like to express their sincere thanks to Professor K. Kunisawa, Associate Professor K. Shimizu, and Mr. M. Sagae, Science University of Tokyo, and also to Mr. O. Kikuchi, for their useful advices, stimulating discussions and encouragement. They are also grateful to two anonymous referees for their valuable suggestions on improving this paper.

APPENDIX

(I) Measures $\hat{\phi}_s$ and $\hat{\psi}_s$ of departure from symmetry, which were considered by Tomizawa(1989), are defined as follows: In the sample version,

$$\begin{aligned}\hat{\phi}_s &= \frac{1}{\delta \log 2} \sum \sum_{i \neq j} p_{ij} \log \frac{2p_{ij}}{p_{ij} + p_{ji}}, \\ \hat{\psi}_s &= \frac{1}{\delta} \sum \sum_{i<j} \frac{(p_{ij} - p_{ji})^2}{p_{ij} + p_{ji}},\end{aligned}$$

where $\delta = \sum \sum_{i \neq j} p_{ij}$ and $0 \log 0 = 0$.

(II) The asymptotic distribution of $\sqrt{n}(\psi - \psi^*)$ is derived as follows: Let

$$p = (p_{11}, p_{12}, \ldots, p_{1R}, \ldots, p_{R1}, p_{R2}, \ldots, p_{RR})'$$

and let us define the vector π in terms of π_{ij}'s in the same way as p. Then $\sqrt{n}(p - \pi)$ is asymptotically distributed as normal $N(0, \Lambda(\pi))$, where $\Lambda(\pi) = \mathrm{diag}(\pi) - \pi\pi'$ and $\mathrm{diag}(\pi)$ denotes a diagonal matrix with the i-th element of π as the i-th diagonal element. We also get

$$\psi = \psi^* + d(\pi)(p - \pi) + o(\|p - \pi\|),$$

where $d(\pi) = \{\partial \psi^* / \partial \pi'\}$. Using the delta method (see Bishop *et al.*, 1975, Sec.14.6), $\sqrt{n}(\psi - \psi^*)$ is asymptotically distributed as normal $N(0, (\sigma^*)^2)$, where

$$\begin{aligned}(\sigma^*)^2 &= d(\pi)\Lambda(\pi)d(\pi)' \\ &= \frac{4}{(\delta^*)^4} [\sum \sum_{i<j} (\pi_{ij} - \pi_{ji})^2 (\pi_{ij} + \pi_{ji}) - (\delta^*)^3 (\psi^*)^2],\end{aligned}$$

and $\delta^* = \sum \sum_{i \neq j} \pi_{ij}$.

REFERENCES

(1) Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis : Theory and Practice.* MIT Press, Cambridge, Massachusetts.

(2) Escofier, B. (1984). Analyse factorielle en référence à un modèle: application à l'analyse de tableaux d'échanges. *Revue de Statistique Appliquée*, 32, 25-36.

(3) Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data* (2nd ed.). MIT Press, Cambridge, Massachusetts.

(4) Harshman, R.A., Green, P.E., Wind, Y. and Lundy, M.E. (1982). A model for the analysis of asymmetric data in marketing research. *Marketing Science*, 1, 205-242.

(5) van der Heijden, P.G.M., de Falguerolles, A. and de Leeuw, J. (1989). A combined approach to contingency table analysis using correspondence analysis and log-linear analysis. *Applied Statistics*, 38, 249-292.

(6) Linhart, H. and Zucchini, W. (1986). *Model Selection.* John Wiley & Sons, New York.

(7) Tomizawa, S. (1989). Two kinds of measures of departure from symmetry in square contingency tables having nominal categories. Manuscript.

Table 1 : Repeat-victimization information, *i.e.*, successive pairs of victimizations for households in the National Crime Survey.

First Victimization in Pair	Second Victimization in Pair								
	Ra	A	Ro	PP/PS	PL	B	HL	MV	Total
Ra	26	50	11	6	82	39	48	11	273
A	65	2997	238	85	2553	1083	1349	216	8586
Ro	12	279	197	36	459	197	221	47	1448
PP/PS	3	102	40	61	243	115	101	38	703
PL	75	2628	413	229	12137	2658	3689	687	22516
B	52	1117	191	102	2649	3210	1973	301	9595
HL	42	1251	206	117	3757	1962	4646	391	12372
MV	3	221	51	24	678	301	367	269	1914
Total	278	8645	1347	660	22558	9565	12394	1960	57407

A is assault; B, burglary; HL, household larceny; MV, motor vehicle theft; PL, personal larceny; PP/PS, pocket picking and purse snatching; Ra, rape; Ro, robbery.

Table 2: Migration in the suburbs of Paris. Rows are destinations, and columns are origins.

	CHA	IVR	KRE	GEN	VIT	ALF	CHO	BON	VAL	ORL	RUN	FRE	THI	JOI	SUC	Total
Charenton	6238	269	45	14	204	1824	57	250	70	76	16	36	0	403	189	9691
Ivry	270	11268	1113	257	2483	1450	530	708	166	878	166	205	281	457	174	20406
Kremlin	34	585	11353	1001	1493	32	143	62	133	207	327	549	226	133	0	16278
Gentilly	0	106	1389	10695	425	100	99	220	27	111	215	1037	26	152	117	14719
Vitry	186	667	894	281	11263	1009	1577	148	123	1021	154	265	860	314	90	18852
Alfort	713	258	134	75	632	16420	595	1675	563	250	29	0	118	507	297	22266
Choisy	0	181	78	41	763	148	5590	24	396	964	104	38	745	25	87	9184
Bonneuil	51	81	68	0	133	1094	109	9235	107	92	0	28	39	1831	491	13359
Valenton	31	34	34	28	34	316	271	148	6161	628	0	0	59	83	228	8055
Orly	14	108	492	177	353	104	528	209	568	6461	315	408	551	191	130	10609
Rungis	0	21	160	83	81	33	23	20	64	248	1455	110	106	21	0	2425
Fresnes	0	53	310	260	156	0	0	0	0	82	481	3889	131	0	0	5362
Thiais	0	66	21	0	151	40	421	24	43	248	26	0	1498	25	0	2563
Joinville	327	43	0	63	206	801	42	1362	0	40	54	90	35	17045	774	20882
Sucy	0	0	0	26	26	20	28	159	591	102	0	0	0	403	5624	6979
Total	7864	13740	16091	13001	18403	23391	10013	14244	9012	11408	3342	6655	4675	21590	8201	181630

Note: Taken directly from Escofier (1984).

Table 3: 1979 car changing data. Rows denote cars disposed of, and columns denote new cars.

	SUBD	SUBC	SUBI	SMAD	SMAC	SMAI	COML	COMM	COMI	MIDD	MIDI	MIDS	STDL	STDM	LUXD	LUXI	Total
SUBD	23272	1487	10501	18994	49	2319	12349	4061	545	12622	481	16329	4253	2370	949	127	110708
SUBC	3254	1114	3014	2656	23	551	959	894	223	1672	223	2012	926	540	246	37	18344
SUBI	11344	1214	25986	9803	47	5400	3262	1353	2257	5195	1307	8347	2308	1611	1071	288	80793
SMAD	11740	1192	11149	38434	69	4880	6047	2335	931	8503	1177	23898	3238	4422	4114	410	122539
SMAC	47	6	0	117	4	0	0	49	0	110	0	10	0	0	0	0	343
SMAI	1772	217	3622	3453	16	5249	1113	313	738	1631	1070	4937	338	901	1310	459	27139
COML	18441	1866	12154	15237	65	1626	27137	6182	835	20909	566	15342	9728	3610	910	170	134778
COMM	10359	693	5841	6368	40	610	6223	7469	564	9620	435	9731	3601	5498	764	85	67901
COMI	2613	481	6981	1853	10	1023	1305	632	1536	2738	1005	990	454	991	543	127	23282
MIDD	33012	2323	22029	29623	110	4193	20997	12155	2533	53002	2140	61350	28006	33913	9808	706	315900
MIDI	1293	114	2844	1242	5	772	1507	452	565	3820	3059	2357	589	1052	871	595	21137
MIDS	12981	981	8271	18908	97	3444	3693	1748	935	11551	1314	56025	10959	18688	12541	578	162714
STDL	27816	1890	12980	15993	34	1323	18928	5836	1182	28324	938	37380	67964	28881	6585	300	256354
STDM	17293	1291	11243	11457	41	1862	7731	6178	1288	20942	1048	30189	15318	81808	21974	548	230211
LUXD	3733	430	4647	5913	6	622	1652	1044	476	3068	829	8571	2964	9187	63509	1585	108236
LUXI	105	40	997	603	0	341	75	55	176	151	589	758	158	756	1234	3124	9162
Total	179075	15339	142259	180654	616	34215	112978	50756	14784	183858	16181	278226	150804	194228	126429	9139	1689511

SUBD is Subcompact / domestic; SUBC, Subcompact / captive imports; SUBI, Subcompact / imports; SMAD, Small speciality / domestic; SMAC, Small speciality / captive imports; SMAI, Small speciality / imports; COML, Low price compacts; COMM, Medium price compacts; COMI, Import compact; MIDD, Midsize domestic; MIDI, Midsize imports; MIDS, Midsize speciality; STDL, Low price standard; STDM, Medium price standard; LUXD, Luxury domestic; LUXI, Luxury import. A more elaborate description can be found in Harshman *et al.*(1982).

Table 4 : Estimate of ψ^*, estimated approximate standard error for ψ, and approximate 95% confidence interval for ψ^*, applied to Tables 1, 2, and 3

Applied data	Estimated measure ψ	Standard error $\sigma/\sqrt{n}$	Confidence interval $\psi \pm z_{0.025}\, \sigma/\sqrt{n}$
Table 1	0.00002	0.00002	(−0.00001, 0.00006)
Table 2	0.00400	0.00009	(0.00382, 0.00417)
Table 3	0.00463	0.00002	(0.00459, 0.00467)