Control and Cybernetics

vol. **29** (2000) No. 4

Detecting rows and columns of contingency table, which outlie from a total positivity pattern

by

Wiesław Szczęsny

Department of Econometrics and Computer Science
Warsaw Agricultural University
Nowoursynowska 166, 02-787 Warsaw, Poland

Institute of Computer Science, Polish Academy of Sciences
Ordona 21, 01-237 Warsaw, Poland

Abstract: It is known that the procedure called *Grade Correspondence Analysis* (GCA) transforms any bivariate contingency table into an approximation of table with a very regular positive dependence, called *total positivity of order two* (TP_2). This fact is reminded in Sections 2 and 3, illustrated there by the GCA transformation of an artificial contingency table $T_{8\times6}$. A search for rows and/or columns of table $T_{8\times6}$, which most strongly outlie from the TP_2 pattern, is described in Section 4. Section 5 presents the outliers found in three large contingency tables, containing the occupational mobility data from Britain and Poland and the parliamentary election data from Poland.

Keywords: computer-intensive methods, contingency table, graphical display, occupational mobility, outliers, scatterplot, total positive dependence.

1. Introduction

Commonly, the term *outlier* designates such an element of a considered set, which is far from "the main body of elements". Data analysts have been especially interested in univariate and multivariate outliers occurring in data matrices (see, e.g. Bartkowiakowa and Szustalewicz, 1997). In case of contingency tables, gross errors are being traced as well as any non-robust behaviour of the contents of particular cells; moreover, statisticians used to test whether a table as a whole can be treated as a random sample from a particular model of bivariate distributions, etc.

In the present paper we propose a procedure which finds out which rows and/or columns of a bivariate contingency table most strongly outlie from the

It is shown that the first step should rearrange rows and columns in order to maximise the value of Spearman's rho. This transformation is called the Grade Correspondence Analysis (GCA), introduced in Ciok et al. (1995). GCA and its link with the TP_2 pattern is described in Sections 2 and 3, referring to facts established by Kowalczyk (2000).

Exclusion of outliers among rows and columns is very important in exploratory data analysis. Here we will only mention that it is a necessary preliminary procedure preceding clustering of rows and of columns based on GCA. Generally, we believe that it will be an important tool of recognising the structure of a contingency table, and this is the direction of the author's further research.

2. Grade Correspondence Analysis

2.1. Contingency tables $\mathrm{T}_{m\times 2}$

In this section we consider bivariate contingency tables with two columns, denoted $\mathrm{T}_{m\times 2} = (N_{ij};\ i = 1, \ldots, m,\ j = 1, 2)$. In an artificial example given in Table 2.1a, rows correspond to school regions and row total N_i for $i = 1, \ldots, m$ denotes the number of pupils who finished school in region i during the last three years. Each total N_i splits into the numbers of those who failed to become a student (N_{i1}) and those who became students (N_{i2}). The regions are presumed

Table 2.1a. Numbers of pupils' failures and successes.

region$_i$	N_{i1} (failure)	N_{i2} (success)	**Total (N_i)**
1	2470	618	**3088**
2	1600	1530	**3130**
3	150	100	**250**
4	400	170	**570**
5	1650	70	**1720**
6	330	120	**450**
7	200	194	**394**
8	200	198	**398**
Total	**7000**	**3000**	**10000**

Table 2.1b. Probability table and its column distributions.

region$_i$	p_{i1} (failure)	p_{i2} (success)	Total ($p_{i\bullet}$)	$P_{\bullet 1}$	$P_{\bullet 2}$
1	0.2470	0.0618	0.3088	0.35286	0.08829
2	0.1600	0.1530	0.3130	0.22857	0.21857
3	0.0150	0.0100	0.0250	0.02143	0.01429
4	0.0400	0.0170	0.0570	0.05714	0.02429
5	0.1650	0.0070	0.1720	0.23571	0.01000
6	0.0330	0.0120	0.0450	0.04714	0.01714
7	0.0200	0.0194	0.0394	0.02857	0.02771
8	0.0200	0.0198	0.0398	0.02857	0.02829
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Table 2.1c. Indices of overrepresentation for Table 2.1a.

region$_i$	h_{i1} (failure)	h_{i2} (success)
1	1.1427	0.6671
2	0.7303	1.6294
3	0.8571	1.3333
4	1.0025	0.9942
5	1.3704	0.1357
6	1.0476	0.8889
7	0.7252	1.6413
8	0.7179	1.6583

Table 2.1d. Permuted table of indices of overrepresentation when regions are ordered according to increasing likelihood ratio (last column).

region$_i$	h_{i1} (failure)	h_{i2} (success)	$p_{i\|2}/p_{i\|1}$
5	1.3704	0.1357	0.0990
1	1.1427	0.6671	0.5838
6	1.0476	0.8889	0.8485
4	1.0025	0.9942	0.9917
3	0.8571	1.3333	1.5556
2	0.7303	1.6294	2.2313
7	0.7252	1.6413	2.2633
8	0.7179	1.6583	2.3100

to be preliminarily somehow ordered, e.g. according to summarised results of final school exams. Denote

$$p_{ij} = N_{ij}/\sum_{i=1}^{m} N_i,\ p_{\bullet j} = \sum_{i=1}^{m} p_{ij},\ p_{i\bullet} = \sum_{i=1}^{2} p_{ij},\ p_{i|j} = p_{ij}/p_{\bullet j},$$
$$P_{\bullet j} = (p_{1|j}, \dots, p_{m|j}),\ i = 1, \dots, m,\ j = 1, 2.$$

The ratio $p_{i|2}/p_{i|1}$, called the likelihood ratio and calculated in Table 2.1d, is the ratio of odds of an alumnus in region i to become and to not become a student. It is seen that initially the odds are not ordered increasingly (i.e. they are not matched with the results of final school exams). So we have two orderings of regions: the initial one and that corresponding to increasing odds as in Table 2.1d. The second ordering ensures *maximal separation* between the conditional column distributions $P_{\bullet 2}$ and $P_{\bullet 1}$, calculated on the basis of the so-called *concentration curve* of $P_{\bullet 2}$ w.r.t. $P_{\bullet 1}$. The curve is shown in Fig. 2.1 as curve C. It consists of eight segments joining the following points

$$(0,0),\ (p_{1|1}, p_{1|2}),\ (p_{1|1} + p_{2|1}, p_{1|2} + p_{2|2}),$$
$$(p_{1|1} + p_{2|1} + p_{3|1}, p_{1|2} + p_{2|2} + p_{3|2}), \dots, (1,1).$$

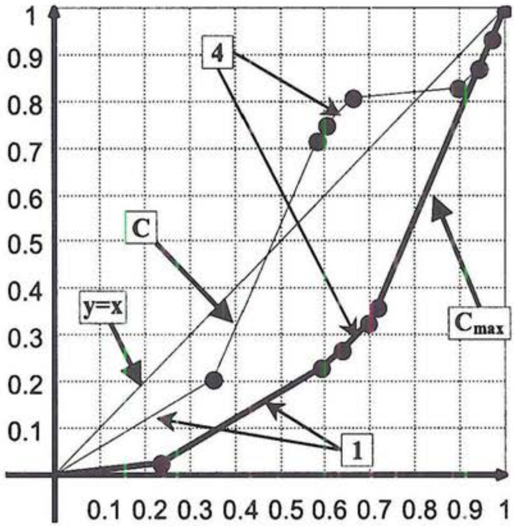

Figure 2.1. Concentration curves C and $C_{\max}$ (below C) for column distrbutions in Table 2.1b. For $C_{\max}$, regions are ordered 5, 1, 6, 4, 3, 2, 7, 8.

Under curve C there lies the concentration curve for column distributions *permuted to make the likelihood ratios increasing* as in Table 2.1d; this curve is called the *maximal concentration curve* and is denoted $C_{\max}$. The integral

$$2\int_0^t (t - C(t))\,dt, \tag{2.1}$$

called *the concentration index* and denoted $ar(P_{\bullet 2} : P_{\bullet 1})$, is a numerical measure of separation between these two column distributions. The concentration index for $C_{\max}$ is denoted $ar_{\max}$. The indices ar and $ar_{\max}$ for C and $C_{\max}$ shown in Fig. 2.1, which are easily expressed geometrically by means of the areas between the diagonal and the curves, are equal to 0.0240 and to 0.4455.

The probability table can be transformed into a continuous distribution defined on the unit square with the density function which is constant on rectangles

$$R_{ij} = \Big\{(u,v) : \sum_{s=1}^{i-1} p_{s\bullet} \le u \le \sum_{s=1}^{i} p_{s\bullet},\ \sum_{t=1}^{j-1} p_{\bullet j} \le v \le \sum_{t=1}^{j} p_{s\bullet}\Big\},$$
$$i = 1, \dots, m,\ j = 1, 2,$$

and is equal on R_{ij} to

This ratio h_{ij} (see Table 2.1c) will be called the *overrepresentation index for cell* (i, j), since it shows overrepresentation of the contents of cell (i, j) as related to its fair representation emerging from the marginals. The density h_{ij} on R_{ij} is called *grade* density of Table $\mathrm{T}_{m\times 2}$. The corresponding correlation coefficient, called the *grade correlation* of table $\mathrm{T}_{m\times 2}$ and denoted ρ^* (and also named Spearman's rho), is related to $ar(P_{\bullet 2} : P_{\bullet 1})$ by the formula

$$\rho^*(\mathrm{T}_{m\times 2}) = 3p_{\bullet 2}p_{\bullet 1}ar(P_{\bullet 2} : P_{\bullet 1}).$$

Note that another well-known dependence measure called Kendall's tau and denoted τ is defined by the formula

$$\tau(\mathrm{T}_{m\times 2}) = 2p_{\bullet 2}p_{\bullet 1}ar(P_{\bullet 2} : P_{\bullet 1})$$

so that for any table with two columns (or two rows) $\rho^*(\mathrm{T}_{m\times 2}) = (3/2)\tau(\mathrm{T}_{m\times 2})$.

We see from Table 2.1d that the rearrangement of regions according to increasing likelihood ratio results in a rather strong overrepresentation of failures in case of the initial region No 5 and of successes in case of the last regions No 3, 2, 7, 8, while rather strong underrepresentation appears for successes in regions No 5 and 1.

The related concentration index can be expressed as (Kowalczyk, 2000):

$$ar(P_{\bullet 2} : P_{\bullet 1}) = \frac{1}{p_{\bullet 1}p_{\bullet 2}} \sum_{i=1}^{m} \sum_{j=i+1}^{m} (p_{i1}p_{j2} = p_{j1}p_{i2}).$$

Its value for column distributions in Table 2.1b is 0.4455.

It is immediately seen that the likelihood ratio of the column distributions is increasing if and only if for all pairs (i, j), $i = 1, \dots, m$, $j = i+1, \dots, m$, the following inequality is satisfied

$$p_{i1}p_{j2} - p_{j1}p_{i2} \geq 0,$$

which means that in all the subtables 2×2 formed by rows i, j $(i < j)$ the likelihood ratios are increasing. This property of a table $\mathrm{T}_{m\times 2}$ is also known as its *total positivity of order two*.

2.2. Contingency tables with m rows and k columns

The notion of the grade density can be easily extended to $m \times k$ tables, with the overrepresentation indices h_{ij} defined as $p_{ij}/(p_{i\bullet}p_{\bullet j})$ for $i = 1, \dots, m$, $j = 1, \dots, k$. Similarly, the definitions of Spearman's rho and Kendall's tau are extended as:

$$\rho^*(\mathrm{T}_{m\times k}) = 3 \sum^{k} \sum^{t-1} \Big[(S_t + S_{t-1} - S_s S_{s-1}) \sum^{m} \sum^{m} (p_{is}p_{it} - p_{is}p_{it})\Big],$$

where $S_u = \sum_{i=1}^{u} p_{\bullet i}$ for $u = 1, \ldots, k$, and

$$\tau(\mathrm{T}_{m\times k}) = 2\sum_{t=2}^{k}\sum_{s=1}^{t-1}\sum_{i=1}^{m}\sum_{j=i+1}^{m}(p_{is}p_{jt} - p_{js}p_{it}).$$

By a suitable permutation of rows and columns one gets a pair (possibly more than one) of permutations which maximise ρ^* (an algorithm was proposed in Ciok et al., 1995). Usually, there is just one pair of optimal permutations and usually the same pair maximises the value of ρ^* and of τ; but whenever the optimal pairs of permutations for ρ^* and of τ are different, they usually differ only slightly. The operation of permuting rows and columns of $\mathrm{T}_{m\times k}$ in order to maximise ρ^* is called the *Grade Correspondence Analysis* (GCA) *of* $\mathrm{T}_{m\times k}$. The analogous procedure maximising τ is called the *Grade Correspondence Analysis based on* τ (denoted GCA$|\tau$). Both procedures will be applied here to the 8×6 contingency table given in Table 2.2a, which contains data related to Table 2.1a: three first columns of Table 2.2a sum up to the first column of Table 2.1a and denote, respectively, the numbers of failures in three consecutive years, while three last columns of Table 2.2a sum up to the second column of Table 2.1a

Table 2.2a. Numbers of pupils' failures and successes in three consecutive years.

	failures			successes			**Total**
region$_i$	year 1	year 2	year 3	year 1	year 2	year 3	(N_i)
1	850	820	800	230	210	178	**3088**
2	500	540	560	500	540	490	**3130**
3	50	50	50	32	33	35	**250**
4	90	130	180	140	30	0	**570**
5	570	550	530	22	24	24	**1720**
6	120	110	100	43	40	37	**450**
7	66	67	68	61	66	67	**394**
8	67	66	67	65	66	67	**398**
Total	**2313**	**2333**	**2355**	**1093**	**1009**	**898**	**10000**

Table 2.2b. Indices of overrepresentation: rows (regions) and columns ordered according to GCA.

	failures			successes		
region$_i$	year 1	year 2	year 3	year 1	year 2	year 3
5	1.433	1.371	1.309	0.117	0.138	0.155
1	1.190	1.138	1.100	0.681	0.674	0.642
6	1.153	1.048	0.944	0.874	0.881	0.916
4	0.683	0.978	1.341	2.247	0.522	0.000
3	0.865	0.857	0.849	1.171	1.308	1.559
2	0.691	0.739	0.760	1.462	1.710	1.743
7	0.719	0.729	0.728	1.416	1.660	1.894

and denote, respectively, the numbers of successes in three consecutive years. The values of ρ^* and τ for Table 2.2a are 0.0162 and 0.0092. After GCA, which provides here the same results as GCA$|\tau$, the overrepresentation indices are as shown in Table 2.2b, and the values of ρ^* and τ increase to their maximal values of 0.2971 and 0.1998. We see that a rather strong overrepresentation occurs in case of region No 5 for columns 1, 2, 3, in case of regions No 3, 2, 7, 8 for columns 3, 4, 5, and also in case of region No 4 for columns 3 and 4; a rather strong underrepresentation appears in case of regions No 5 and 1 for columns 4, 5 and 6, and also in case of region No 4 for columns 1, 5 and 6.

Table 2.2b provides a good insight into the chances of failures and of successes. GCA does not lead to the interchange of columns concerning failures (numbered 1, 2, 3) and columns concerning successes (numbered 4, 5, 6), which means that differences, which occurred in consecutive years, were negligible as compared to those between successes and failures. The optimal ordering of regions in Table 2.2b remains the same as in Table 2.1d, in which failures and successes are aggregated over years.

3. Total positivity of order two

Procedures GCA and GCA$|\tau$ provide patterns of positive dependence between the row variable and the column variable such that the strength of positive dependence is maximised. Then, we can ask how *regular* is this dependence. Looking backward to tables with only two columns, discussed in Sec. 2.2, we become aware that in this case GCA ensures an ordering of rows according to the *increasing likelihood ratio* for the conditional distributions corresponding to the two columns. Now we ask: does this condition hold *for any pair* (i, j) *of columns* of a $\mathrm{T}_{m \times k}$ table when $i < j$? The answer is that *generally it does not hold*, although such requirement would certainly be desirable. Whenever it holds, we deal with a very regular pattern of positive dependence between row and column variables. It is easy to check (Kowalczyk, 2000) that this requirement holds if and only if

$$p_{is}p_{jt} - p_{js}p_{it} \geq 0 \tag{3.1}$$

for any 2×2 subtable of $\mathrm{T}_{m \times k}$ with cells in rows i, j and columns s, t such that $1 \leq i < j \leq m$, $1 \leq s < t \leq k$. Formula (3.1) entails that such model of positive dependence is called *total positivity of order two* (TP_2).

The aforesaid condition imposed on all pairs of columns is equivalent to such condition imposed on all pairs of rows. Moreover (Kowalczyk, 2000), if $\mathrm{T}_{m \times k}$ is TP_2, then it remains unchanged under GCA as well as under GCA$|\tau$.

A useful characterization of TP_2, based on the expression

$$\tau_{abs} = \sum^{k} \sum^{t-1} \sum^{m} \sum^{m} |p_{is}p_{jt} - p_{js}p_{it}|, \tag{3.2}$$

states that a table $T_{m\times k}$ is TP_2 if and only if $\tau(T_{m\times k}) = \tau_{abs}$. It has been therefore suggested in Kowalczyk (2000) to use $1-\tau/\tau_{abs}$ as a measure of departure of $T_{m\times k}$ from the family of TP_2 tables. This measure is nonnegative, equal to zero if and only if $T_{m\times k}$ is TP_2, and it attains its minimal value in the set of all tables obtained from $T_{m\times k}$ by permutations of rows and/or columns when $T_{m\times k}$ is transformed according to GCA$|\tau$. So, we say that GCA$|\tau$ applied to $T_{m\times k}$ provides the best approximation of the TP_2 property with respect to $1-\tau/\tau_{abs}$.

For Table 2.2a, τ_{abs} is equal to 0.2181, $\tau_{max} = 0.2010$ and hence $1 - \tau_{\max}/\tau abs = 0.0785$. This implies that Table 2.2a transformed by GCA is almost a TP_2 table.

In practice, however, we are less interested in how distant from TP_2 a table is, than in detecting which rows and/or columns are particularly responsible for this departure. Then, we could throw these rows and/or columns out and deal with a more regular positive trend between the row variable and the column variable. The row variable is well represented by *the grade regression function defined on rows*, the column variable — by *the grade regression function defined on columns*, where by definition the grade regression function is the regression function of the grade distribution. It should be noted that in a TP_2 table the first regression is increasing w.r.t. the likelihood ratio for any pair (s,t) of columns $(s < t)$, and the second regression is increasing w.r.t. the likelihood ratio for any pair (i,j) of rows. This is why *we are often inclined to represent the whole vector of columns of a* TP_2 *table (or of a table close to* TP_2*) solely by the first regression*; this possibility can be exploited in further exploratory analysis of that table (when it is confronted with other tables or when the data are additional explanatory variables).

4. Search for rows and/or columns, which most strongly outlie from TP_2

The requirement put on pairs of columns in the definition of TP_2 is equivalent (Kowalczyk, 2000) to the statement: table $T_{m\times k}$ is TP_2 iff, for each pair of columns, distributions $(P_{\bullet s}, P_{\bullet t})$ satisfy

$$ar(P_{\bullet s} : P_{\bullet t}) = ar_{\max}(P_{\bullet s} : P_{\bullet t}); \tag{4.1}$$

the analogous statement can be also formulated for all pairs of row distributions $P_{i\bullet}$ and $P_{j\bullet}$. Therefore we will consider two sets: the scatterplot

$$S^{\text{GCA}}_{\text{columns}} = \{(ar(P^{\text{GCA}}_{\bullet t} : P^{\text{GCA}}_{\bullet s}, ar_{\max}(P^{\text{GCA}}_{\bullet t} : P^{\text{GCA}}_{\bullet s})) : \\ s = 1,\ldots,k,\ t = s+1,\ldots,k\}$$

(when we are interested in outliers from TP_2 among colums) and the scatterplot

$$S^{\text{GCA}}_{\text{rows}} = \{(ar(P^{\text{GCA}}_{j\bullet} : P^{\text{GCA}}_{i\bullet}, ar_{\max}(P^{\text{GCA}}_{j\bullet} : P^{\text{GCA}}_{i\bullet})) : \\ i = 1,\ldots,m,\ j = 1,\ldots,m\}$$

The indices ar and $ar_{\max}$ for rows of Table 2.2a transformed by GCA are given in Table 4.1, and the resulting set $S^{\mathrm{GCA}}_{\mathrm{rows}}$ is shown in Fig. 4.1. Since in this table the equality (4.1) holds or nearly holds for the majority of pairs of row distributions, almost all points in Fig. 4.1 are close to the diagonal $y = x$ (called in the sequel the TP_2 line); however, there are a few exceptions (marked grey), which refer to the following pairs of regions: $(4,6)$, $(4,3)$, $(4,1)$, $(4,2)$, $(4,7)$, $(4,5)$. Clearly, any row, say i, of the table is described by the subset of S_{GCA} consisting of points $(ar(i,j), ar_{\max}(i,j))$, $j = 1,\dots,m$. The position of this subset among all points in S_{GCA} indicates whether points corresponding to row i tend to be more distant from the TP_2 line than in the case of remaining rows. This is a visual suggestion that row i is an outlier. According to that, Fig. 4.1 suggests that region No 4 is an outlier from TP_2 in the set of regions. After removing this region from the data set we get a new scatterplot $S^{\mathrm{GCA}}_{\mathrm{rows}}$ presented in Fig. 4.2, which practically lies on the diagonal. We note that according to Table 2.1a the size of region No 4 slightly exceeds four other regions which do not outlie from TP_2, so we have no reason to think that region No 4 outlies because of having small size (i.e. it is not a *make believe* outlier from TP_2 in the set of regions).

Table 4.1. Indices ar (below the diagonal) and $ar_{\max}$ (above the diagonal).

	region 5	**region 1**	**region 6**	**region 4**	**region 3**	**region 2**	**region 7**	**region 8**
region 5	0	0.1621	0.2436	0.3958	0.3688	0.4688	0.4629	0.4649
region 1	0.1608	0	0.0835	0.3194	0.2124	0.3096	0.3080	0.3090
region 6	0.2083	0.0520	0	0.3337	0.1568	0.2511	0.2504	0.2510
region 4	0.3624	0.1598	0.0881	0	0.3474	0.3562	0.3747	0.3724
region 3	0.3686	0.2124	0.1568	0.1060	0	0.1033	0.0965	0.1011
region 2	0.4679	0.3096	0.2506	0.2262	0.0917	0	0.0237	0.0279
region 7	0.4628	0.3080	0.2504	0.2267	0.0945	0.0049	0	0.0132
region 8	0.4643	0.3090	0.2510	0.2288	0.0948	0.0050	0.0000	0

Turning to columns, we see from the scaterplot $S^{\mathrm{GCA}}_{\mathrm{columns}}$ in Fig. 4.3 that none of the columns of Table 4.1a transformed by GCA ought to be treated as an outlier even when region No 4 is not excluded. After exclusion of this region, the scatterplot of columns in Fig. 4.4 transmits the same visual message as that obtained from Fig. 4.2: the GCA transform of Table 2.2a with region No 4 excluded is almost a TP_2 table.

According to the remark at the end of Section 3, the sequence of regions 5, 1, 6, 3, 7, 2, 8 (with region No 4 excluded) can be well represented by the respective grade regression function defined on rows. This function could be next compared with various explanatory variables, which describe the regions (for example, in order to find out which factors influence the more successful regions). However, in practice it is rarely so that there is just one definite outlier, and the points in $S^{\mathrm{GCA}}_{\mathrm{rows}}$ and $S^{\mathrm{GCA}}_{\mathrm{columns}}$ are usually much more distant from the

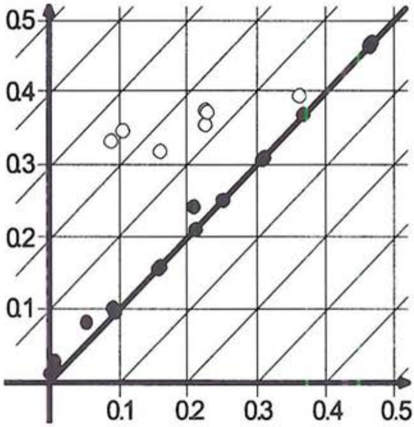

Figure 4.1. Scatterplot S_{GCA} for Table 2.2a (rows)

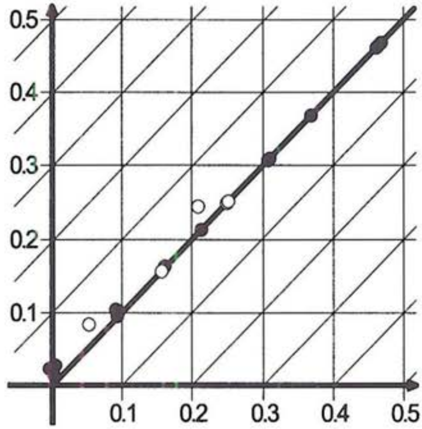

Figure 4.2. Scatterplot S_{GCA} for Table 2.2a (rows) when region No 4 is excluded

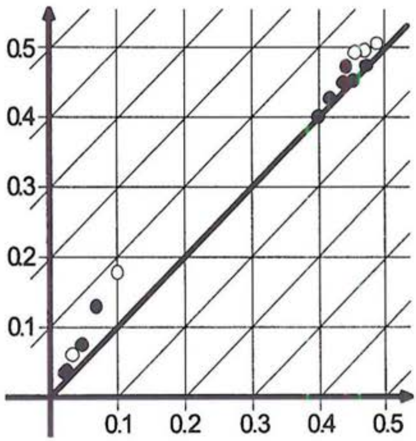

Figure 4.3. Scatterplot S_{GCA} for Table 2.2a (columns)

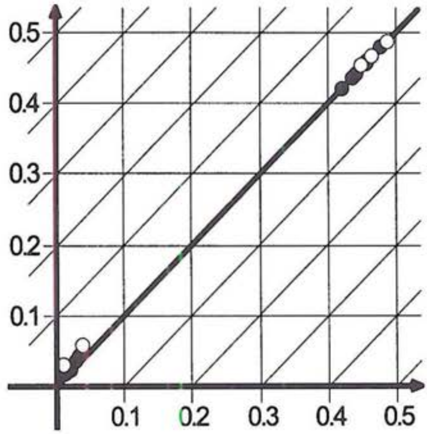

Figure 4.4. Scatterplot s_{GCA} for Table 2.2a (columns) when region No 4 is excluded

Apart from visual suggestions, we will introduce numerical measures describing how distant is a row or a column from the TP_2 line. This is simply done by of points. So we introduce *the mean distance from* TP_2 *line of row* (i) as

$$\overline{d}_{\text{rows}}(i) = \sum_{\{j: j\leq m,\, ar(i,j)\geq 0\}} d[(ar(i,j), ar_{\max}(i,j)),\ \text{line } y = x] + \sum_{\{j: j\leq m,\, ar(i,j)< 0\}} d[(ar(i,j), ar_{\max}(i,j)),\ \text{line } y = -x]$$

for $i = 1, \ldots, m$, and let *the mean distance from the* TP_2 *line of column* (i), denoted $\overline{d}_{\text{columns}}(i)$, be defined analogously. It follows that those rows and columns can be ordered, according to their mean distances, from those most to those least likely to be treated as an outlier. There are many possibilities of further decisions and actions to be undertaken by a data analyst but this exceeds the scope of this paper. Some possibilities will be discussed in a next paper being currently prepared by the present author. Now, we only suggest that a contingency table can be described by the following real-valued statistics:

$$\frac{1}{m}\sum_{i=1}^{m}\overline{d}_{\text{rows}}(i)$$

(called *total mean distance from* TP_2 *line in case of rows*);

$$\frac{1}{k}\sum_{i=1}^{k}\overline{d}_{\text{columns}}(i)$$

(called *total mean distance from* TP_2 *line in case of columns*);

$$\max\{\overline{d}_{\text{rows}}(i);\ i = 1, \ldots, m\}$$

(called *maximal distance from* TP_2 *line among rows*);

$$\max\{\overline{d}_{\text{columns}}(i);\ i = 1, \ldots, k\}$$

(called *maximal distance from* TP_2 *line among columns*),

and by a vector $(q_1, q_2, \ldots)$, where q_s for $s = 1, 2, \ldots$ is a fraction of points in S_{GCA} satisfying

$$|ar(i,j)| + 0.1(s-1) \leq ar_{\max}(i,j) < |ar(i,j)| + 0.1s$$

(i.e. q_1 is the fraction of points which are distant from the TP_2 line or the line $y = -x$ by no more than $0.1\sqrt{2}$, etc.). In a TP_2 table, or in a table very close to it, $(q_1, q_2, \ldots) = (1, 0, 0 \ldots)$.

In case of Table 2.2a, the total mean distance from TP_2 line is 0.0315 in case of rows and 0.0182 in case of columns, maximal distance from the TP_2 line is 0.1737 among rows and 0.0559 among columns, and $(q_1, q_2, q_3, q_4, \ldots) = (0.786, 0.143, 0.071, 0, \ldots)$ for rows and $(1, 0, \ldots)$ for columns. When region No 4 is excluded, these statistics take the values: 0.0051, 0.0038, 0.0249, 0.0148,

5. Examples of graphical and numerical analysis of outliers in three large data sets

Three contingency tables will be analyzed:

(i) Table $\mathrm{BRIT}_{7\times 7}$ containing frequencies of father/son pairs such that father's occupation belongs to category i and son's occupation belongs to category j $(i, j = 1, 2, \ldots, 7)$. The table, which summarizes the results of a study made in Britain, was published in many statistical papers on data analysis, e.g. Gifi (1990), Kowalczyk (1999),

(ii) Table $\mathrm{POH}_{12\times 12}$ which also deals with father/son occupational mobility data for 12 categories, summarizing the results of a study performed in Poland (Pohoski, 1983, Kowalczyk, 1999),

(iii) Table $\mathrm{ELECT}_{52\times 25}$ summarizing the results of two elections to the Polish parliament, in 1993 and 1997, with vote numbers $\{n_{ij}\}$ obtained in 52 election regions by altogether 25 political parties (15 in 1993, 10 in 1997). This data table was analyzed in Szczęsny et al. (1998).

The scatterplots S_{GCA} for fathers (rows) and sons (columns) in case of $\mathrm{BRIT}_{7\times 7}$ are presented in Figs. 5.1 and 5.2; the respective scatterplots for fathers and for sons in case of $\mathrm{POH}_{12\times 12}$ are presented in Figs. 5.3 and 5.4; S_{GCA} for political parties (columns) in $\mathrm{ELECT}_{52\times 25}$ is presented in Fig. 5.5. The points corresponding to the row or column, which is the most distant from TP_2, are distinguished on every figure.

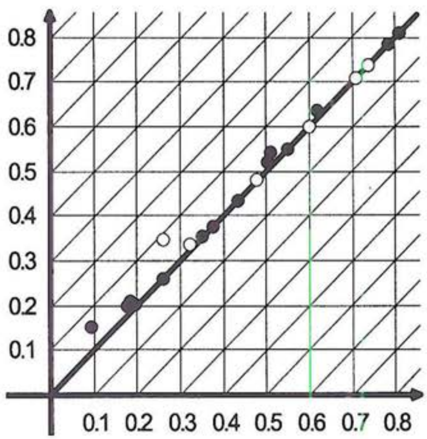

Figure 5.1. Scatterplot S_{GCA} for

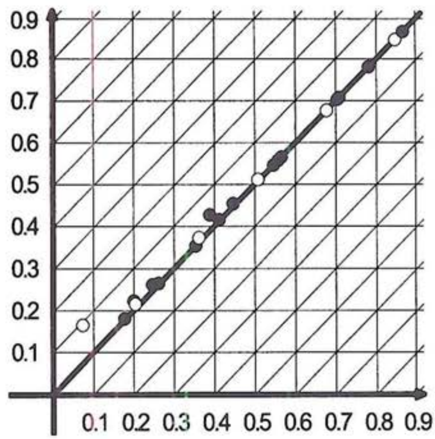

Figure 5.2. Scatterplot S_{GCA} for

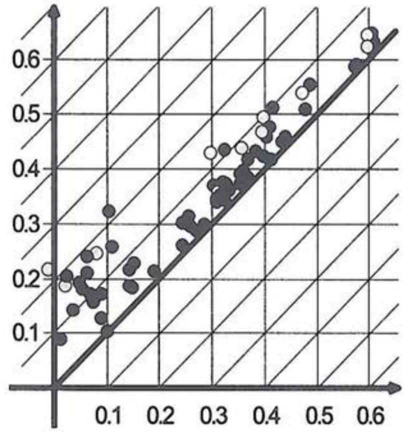

Figure 5.3. Scatterplot S_{GCA} for $\mathrm{POH}_{12\times 12}$ (rows).

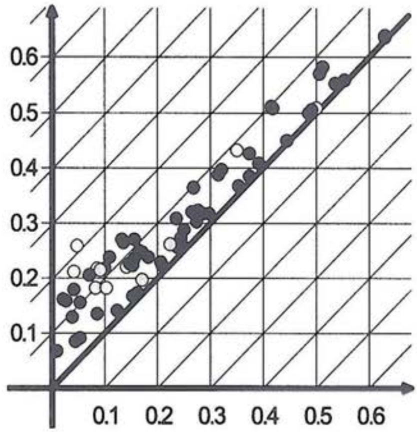

Figure 5.4. Scatterplot S_{GCA} for $\mathrm{POH}_{12\times 12}$ (columns).

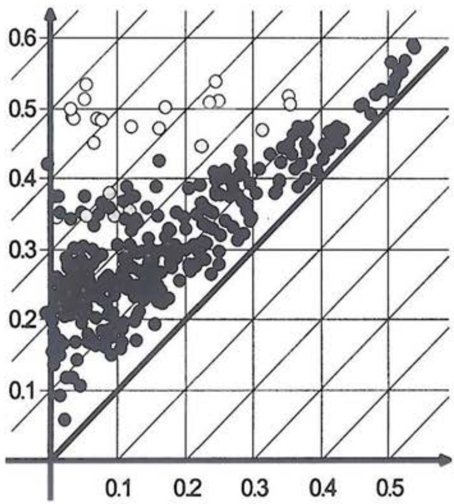

Figure 5.5. Scatterplot S_{GCA} for $\mathrm{ELECT}_{52\times 25}$ (columns).

The table $\mathrm{BRIT}_{7\times 7}$ is equal to its GCA transform, and it is immediately seen that it is very close to TP_2, with no outliers among its rows and columns. The GCA transform of table $\mathrm{POH}_{12\times 12}$ is visually less close to TP_2 than $\mathrm{BRIT}_{7\times 7}$

On the other hand, the GCA transform of table $\mathrm{ELECT}_{52\times 16}$ is irregular and contains at least one definite outlier among political parties ("SAMOOBRONA", Engl. "SELF-DEFENCE").

The results for examples (i), (ii), (iii) appearing in Figs. 5.1–5.5 and Table 5.1 imply convincingly that there are no outliers in (i) and (ii) but there is one obvious outlier among columns in (iii). Although these conclusions are certainly true, we have to stress once more that neither scatterplots S_{GCA} nor values of the statistics used in Table 5.1 are directly comparable from one study to another. They depend on the total $N = \sum\sum N_{ij}$, on the numbers of categories m and k and on the extent of diversification of probabilities in marginal distributions, and also on the strength of maximal positive dependence. In examples (i)–(iii), the totals N are very large (3497 in (i), 8767 in (ii), over 20,000,000 in (iii)) and the quotients $N/(mk)$ are rather similar, but probabilities in marginal distributions are diversified in different ways. It is evident that a row with very small $p_{i\bullet}$ or column with very small $p_{\bullet j}$ could induce a very large value of mean distance from TP_2 as compared with those for other rows and columns. Therefore, we checked the marginal probability for "SAMOOBRONA" in $\mathrm{ELECT}_{52\times 25}$ and found it equal to 0.014, which is not exceptionally small (seven other parties in $\mathrm{ELECT}_{52\times 25}$ had smaller probabilities, while all of them had the mean distance from TP_2 much smaller than "SAMOOBRONA").

Table 5.1. Numerical description of departure from TP_2 for examples (i), (ii), (iii).

	GCA transforms of tables		
	(i) $\mathrm{BRIT}_{7\times 7}$	(ii) $\mathrm{POH}_{12\times 12}$	(iii) $\mathrm{ELECT}_{52\times 25}$
$q_1,\ldots,q_5$ in case of rows	1, 0, 0, 0, 0	.803, .167, .030, 0, 0	not calculated
$q_1,\ldots,q_5$ in case of columns	1, 0, 0, 0, 0	.788, .197, .015, 0, 0	.343, .417, .177, .037, .027
Total mean distance from TP_2 line in case of rows	.009	.048	not calculated
Total mean distance from TP_2 line in case of columns	.007	.064	.105
Maximal distance from TP_2 line among rows	.012	.075	not calculated
Maximal distance from TP_2 line among columns	.017	.067	.231

Yet, inference from outliers is usually more obscure and a general method of standardization is needed. When m, k and N are rather small, checking could be based on simulation from the discretized binormal distribution with correlation coefficient and marginal distributions such as in the observed table. By drawing N times, we build a random contingency table, form S_{GCA}, and calculate values of all real-valued statistics, which are of interest. Then, from a sufficiently large number of random tables, we find thresholds for those statistics.

References

BARTKOWIAK, A. and SZUSTALEWICZ, A. (1997) Detecting multivariate outliers by a grant tour, *Machine Graphics & Vision*, **6.4**, 487–505.

CIOK, A., KOWALCZYK, T., PLESZCZYŃSKA, E. and SZCZĘSNY, W. (1995) Algorithms of grade correspondence-cluster analysis. *Archiwum Informatyki Teoretycznej i Stosowanej*, **7**, 1-4, 5–22.

CIOK, A., KOWALCZYK, T. and PLESZCZYŃSKA, E. (1998) How a New Statistical Infrastructure Induced a New Computing Trend in Data Analysis. In: *Rough Sets and Current Trends in Computing*, *Lecture Notes in Artificial Intelligence* 1424, *Proceedings of the First International Conference*, *RSCTC'98*, *Warsaw*, *Poland*, *June 22–26*, *1998*, Polkowski, L. and Skowron, A., eds., Springer, 75–82.

CIOK, A., KOWALCZYK, T., PLESZCZYŃSKA, E. and SZCZESNY, W. (1995) Algorithms of grade correspondence-cluster analysis. *Archiwum Informatyki Teoretycznej i Stosowanej*, **7**, 1-4, 5–22.

GIFI, A. (1990) *Nonlinear multivariate analysis*. J. Wiley & Sons, New York.

KOWALCZYK, T. (1999) Decomposition of $m \times m$ tables on father/son occupational status. In: *Intelligent Information Systems VIII*, *Procedings of the Workshop held in Ustroń*, *14–18 June 1999*, Kłopotek, M. and Michalewicz, M., eds., 60–64.

KOWALCZYK, T. (2000) Link between grade measures of dependence and of separability in pairs of conditional distributions. *Statistics and Probability Letters*, **46**, No 4, 371–379.

POHOSKI, M. (1983) Social mobility and social inequalities (in Polish). *Kultura i Społeczeństwo*, **27** (1983), 135–164.

SZCZĘSNY, W., CIOK, A., KOWALCZYK, T., PLESZCZYŃSKA, E. and WYSOCKI, W. (1998) Grade correspondence analysis in contingency tables. Applications to data of the 1993 and 1997 Elections to the Polish Parliament (in Polish). *Studia Socjologiczne*, No 3 (150), 49–68.