Some useful multiple comparison procedures in the analysis of the two-way contingency table

C. Hirotsu*

Meisei University, 2-1-1 Hodokubo, Hino-City, Tokyo 191-8506, Japan

S. Yamamoto

RPM Co., Ltd., 5F Shinwa-building, 3-2-4, Nishi-shinjuku, Shinjuku-ku, Tokyo 160-0023, Japan

Abstract

A clustering procedure for the rows of a two-way contingency table has been proposed in Hirotsu (1983a) and verified to be useful on several occasions as compared with other multiple comparison approaches, see Greenacre (1988) and Hirotsu (1993). It is essentially Scheffé type multiple comparisons and Hirotsu (2009) raised its power by introducing a generalized squared distance among any number of clusters. It is usually easy to obtain and interpret those significant clusters when the number of rows is small, say, up to 10. However, if it is more than 10, we need some stopping rule working automatically for obtaining a significant clustering with the reasonable number of clusters. One of the purposes of the present paper is therefore to propose such a stopping rule. When there is a natural ordering in the columns the procedure is essentially unchanged excepting the definition of the squared distance among clusters reflecting the natural ordering and the related distribution theory. The related distributions are those of the largest eigen root of the orthogonal and non-orthogonal Wishart matrices for the nominal and ordinal columns, respectively. When the columns are nominal the rows and columns are symmetrically dealt with and Scheffé type multiple comparisons can be applied simultaneously to rows and columns. For the ordinal columns, however, we are not interested in all the permutations of them and apply the change-point

*Corresponding author

Email address: `hirotsu@ge.meisei-u.ac.jp` (C. Hirotsu)

URL: `http://corec.meisei-u.ac.jp/labs/hirotsu/` (C. Hirotsu)

Preprint submitted to Computational Statistics & Data Analysis *November 22, 2012*

type contrasts to columns. Then the related asymptotic distribution is that of the max accumulated chi-square, which is the maximum of the correlated chi-squares (Hirotsu et al., 1992).

Keywords: Exact algorithm, Generalized squared distance, Ordinal categories, Row-wise multiple comparisons, Stopping rule, Wishart distribution

1. Introduction

An overall goodness of fit chi-squared test for independence is a well known approach to a contingency table. It cannot, however, give any detailed information on the association between the rows and columns. On the other hand the multiple comparison approach based on one degree of freedom chi-squared variable is less powerful and the result of the analysis is often unclear since the total degrees of freedom for interaction is usually so large. Therefore the row-wise multiple comparisons have been proposed in Hirotsu (1983a) and verified to be useful in several occasions as compared with other multiple comparison approaches, see Greenacre (1988) and Hirotsu (1993). The idea has been shown to be useful also for the two-way ANOVA model in Hirotsu (1983b, 1991) and Hirotsu et al. (2003). They are nothing but the multiple comparisons of the treatment effects if the data are taken as the one-way layout with categorical responses instead of the usual normal variables. The row-wise multiple comparisons proposed are essentially Scheffé type and naturally lead to clustering of rows so that the rows within a cluster are homogeneous and a large deviation from independence exists only among clusters. This gives a simple structure of the association between the rows and columns. In particular the generalized squared distance among clusters introduced in Hirotsu (2009) raised the power of the multiple comparisons. It is usually easy to obtain and interpret those significant clusters when the number of rows is not large. However, if it is large it is impossible to search all the possible classifications and obtain an optimal clustering in any sense. So in Section 6 of the present paper we propose a stopping rule working automatically for obtaining significant classification into the reasonable number of clusters. In Section 2 we explain the notation and the overall goodness of fit chi-square is described according to the notation. In Section 3 several definitions of the squared distance are given including the generalized squared distance among any number of clusters. They are extended in Section 4 to the case where there is a natural ordering in columns. Section 5 is for the algorithm to obtain the reference value. Section 7 is for clustering columns and Section 8 is for real examples. Finally in Section 9 the concluding remarks are stated.

2. An overall goodness of fit chi-square

2.1. The notation and an overall goodness of fit chi-square

Following the notation in Hirotsu (2009) let a two-way contingency table be denoted by $\{y_{ij}\}_{a\times b}$ and the row, column and the grand totals by $R_i = y_{i\cdot}$ $(i = 1, \cdots, a)$, $C_j = y_{\cdot j}$ $(j = 1, \cdots, b)$ and $N = y_{\cdot\cdot}$, respectively, where we employ the usual dot notation to express the summation with respect to the suffix replaced by dot. We assume a multinomial distribution for the cell probabilities $\{p_{ij} \,|\, p_{\cdot\cdot} = 1\}$. The null hypothesis of interest is then

$$H : p_{ij} = p_{i\cdot}\, p_{\cdot j} \quad \text{for all } i \text{ and } j,$$

and the statistical inference is based on the conditional distribution given $\{R_i\}$ and $\{C_j\}$. For the row-wise multiple comparisons define

$$\boldsymbol{r} = N^{-1/2}\left(\sqrt{R_1}, \cdots, \sqrt{R_a}\right)', \quad \boldsymbol{c} = N^{-1/2}\left(\sqrt{C_1}, \cdots, \sqrt{C_b}\right)'$$

and then define $R'_{a-1\times a}$ and $C'_{b-1\times b}$ so that $\binom{\boldsymbol{r}'}{R'}$ and $\binom{\boldsymbol{c}'}{C'}$ are the a- and b-dimensional orthogonal matrices, respectively, where the prime denotes a transpose of a matrix. Define a column vector $\boldsymbol{z}$ with the elements $z_{ij} = y_{ij}/\sqrt{R_iC_j/N}$ arranged in the dictionary order. Then under the null hypothesis H the conditional expectation and variance of $(R' \otimes C')\,\boldsymbol{z}$ given R_i and C_j are

$$E\{(R' \otimes C')\ \boldsymbol{z}\} = \boldsymbol{O}_{(a-1)\,(b-1)},$$
$$V\{(R' \otimes C')\ \boldsymbol{z}\} = (N/(N-1))\,\boldsymbol{I}_{(a-1)\,(b-1)}$$

where $\boldsymbol{O}_n$ and $\boldsymbol{I}_n$ are n-dimensional zero vector and the identity matrix, respectively and $\otimes$ denotes a Kronecker product. It should be noted that in $(R' \otimes C')\ \boldsymbol{z}$ every row of R' is constructing the orthogonal contrast in rows and similarly for the row of C'. In the following we ignore the coefficient $(N/(N-1))$ in the variance since our example of the contingency table is usually large. Then

$$\chi^2 = \|\,(R' \otimes C')\ \boldsymbol{z}\,\|^2$$

is nothing but the goodness of fit χ^2 for H and every element of $(R' \otimes C')\ \boldsymbol{z}$ gives the partition of χ^2 into one degree of freedom, where $\|\ \boldsymbol{.}\ \|^2$ denotes the squared norm of a vector.

2.2. The largest root of a Wishart matrix

Another important overall test is given by the largest root

$$W_1 = \max_{\boldsymbol{\gamma}'\boldsymbol{r}=0,\, \|\boldsymbol{\gamma}\|=1} \| (\boldsymbol{\gamma}' \otimes C')\ \boldsymbol{z} \|^2 \tag{1}$$

whose asymptotic null distribution is shown to be that of the largest root of the Wishart matrix $W\left(\max(a-1,\, b-1),\, \boldsymbol{I}_{\min(a-1,\, b-1)}\right)$. This reference distribution has been introduced in Hirotsu (1983a) and employed by the other authors including Greenacre (1988). The squared distances among clusters introduced in the next Section are bounded above by W_1 so that we can use the upper tail probability of the largest root of the Wishart matrix as the reference value.

3. The generalized squared distance among any number of clusters of rows

Without any loss of generality we assume a partition of rows into m clusters: $G_1 = \{1, \cdots, q_1\}$, $G_2 = \{q_1+1, \cdots, q_1+q_2\}, \cdots,$
$G_m = \{q_1 + \cdots + q_{m-1} + 1, \cdots, q_1 + \cdots + q_m\}$.
Then the generalized squared distance among them is defined by

$$\chi^2\,(G_1; \cdots; G_m) = \max \| (\boldsymbol{\gamma}' \otimes C')\ \boldsymbol{z} \|^2, \tag{2}$$

where the maximization is taken with respect to $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_a)'$ under the restriction

$$\boldsymbol{\gamma}'\boldsymbol{r} = 0, \quad \|\boldsymbol{\gamma}\| = 1,$$
$$\gamma_i \equiv \lambda_k\,(R_i/T_k)^{1/2}, \quad i \in G_k,\ \ T_k = \sum_{i \in G_k} R_i, \quad k = 1, \cdots, m.$$

It is actually the maximization by $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_m)'$ under the restriction

$$\sum_{k=1}^{m} \sqrt{T_k}\lambda_k = 0, \quad \sum_{k=1}^{m} \lambda_k^2 = 1 \tag{3}$$

The basic idea is to give a constant coefficient for the rows within a cluster so that it cannot contribute to the maximization.

Let $Y_{ij} = \sum_{i \in G_k} y_{ij}, \quad k = 1, \cdots, m$, denote the frequency of the kth clusters at the jth column so that $\{Y_{kj}\}$ gives the $m \times b$ table with the row total T_k collapsing those pooled rows. Then eq.(2) reduces to

$$\chi^2 (G_1; \cdots ; G_m) = \max \boldsymbol{\lambda}' \begin{pmatrix} \boldsymbol{w}_1' \\ \vdots \\ \boldsymbol{w}_m' \end{pmatrix} (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m) \ \boldsymbol{\lambda} \tag{4}$$

with

$$\boldsymbol{w}_k = (T_k/N)^{-1/2} C' \left(C_1^{-1/2} Y_{k1}, \cdots, C_b^{-1/2} Y_{kb} \right) \tag{5}$$

In particular we have

$$W \left(\sqrt{T_1}, \cdots, \sqrt{T_m} \right)' = \sum_k \sqrt{T_k} \boldsymbol{w}_k = NC' \boldsymbol{c} = \boldsymbol{0}$$

suggesting $\left(\sqrt{T_1}, \cdots, \sqrt{T_m} \right)'$ to be the eigen vector of $W'W$ corresponding to a zero root, where $W = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m)$. Then the maximization reduces to the problem of the largest root of $W'W$ and the condition (3) is automatically satisfied. The statistic (4) is the same type of statistic for the pooled $m \times b$ table as (1) from the original $a \times b$ table and obviously coincides with it when $m = a$ and all the clusters are composed of a single row.

By restricting the vector $\boldsymbol{\gamma}$ in appropriate ways we can introduce the squared distances between two rows or between two clusters.

The squared distance between the two rows i and i' :

$$\chi^2(i; i') = \| \left(\boldsymbol{r}'(i; i') \otimes C' \right) \ \boldsymbol{z} \|^2,$$
$$\boldsymbol{r}'(i; i') = \left(\frac{1}{R_i} + \frac{1}{R_i'} \right)^{-1/2} \left(0 \cdots 0 \, R_i^{-1/2} \, 0 \cdots 0 \ - R_{i'}^{-1/2} \, 0 \cdots 0 \right),$$
$$i, \ i' = 1, \cdots, a.$$

The squared distance between the two clusters $G_1 = \{1, \cdots, q_1\}$ and

$G_2 = \{q_1 + 1, \cdots, q_1 + q_2\}$:

$$\chi^2(G_1; G_2) = \| (\boldsymbol{r}'(G_1; G_2) \otimes C')\ \boldsymbol{z}\|^2,$$

$$\boldsymbol{r}'(G_1; G_2) = \left(\frac{1}{T_1} + \frac{1}{T_2}\right)^{-1/2} \times \left(\frac{\sqrt{R_1}}{T_1} \cdots \frac{\sqrt{R_{q_1}}}{T_1} \; - \frac{\sqrt{R_{q_1+1}}}{T_2} \cdots - \frac{\sqrt{R_{q_1+q_2}}}{T_2} \, 0 \cdots 0\right),$$

$$T_1 = \sum_{i \in G_1} R_i, \quad T_2 = \sum_{i \in G_2} R_i.$$

Those squared distances are obviously bounded above by (1).

4. An extension to the natural ordering in columns

In the case where there is a natural ordering in columns we are interested in distinguishing the up- or down-ward tendency along with the columns. For the purpose an order sensitive squared distance has been proposed in Hirotsu (2009). It is based on the cumulative chi-squared statistic which is originated from the complete class lemma for testing the ordered alternatives (Hirotsu, 1982). The method simply replaces the matrix C' by $C^{*'}$ in the definition of the squared distances in Section 3, where the jth row of $C^{*'}$ is

$$\boldsymbol{c}^{*'}(j; j') = \left(\frac{1}{U_j} + \frac{1}{\overline{U}_j}\right)^{-1/2} \left(\frac{\sqrt{C_1}}{U_j} \cdots \frac{\sqrt{C_j}}{U_j} \; - \frac{\sqrt{C_{j+1}}}{\overline{U}_j} \cdots - \frac{\sqrt{C_b}}{\overline{U}_j}\right) \tag{6}$$

with

$$U_j = \sum_{k=1}^{j} C_k, \quad \overline{U}_j = \sum_{k=j+1}^{b} C_k, \quad \text{for } j = 1, \cdots, b-1.$$

We call the vector $\boldsymbol{c}^{*'}(j; j')$ a change-point type contrast. Then in case of $a \geqslant b$ the reference distribution is obtained as the largest root W_1^* of the non-orthogonal Wishart matrix $W\left(a-1, C^{*'}C^*\right)$, where $C^{*'}C^* = \{\rho_{jk}\}$ is made from

$$\rho_{jk} = \sqrt{\delta_j/\delta_k}, \quad 1 \leqslant j \leqslant k \leqslant b-1 \tag{7}$$

$$\text{with } \delta j = \frac{U_j}{\overline{U}_j} = \frac{C_1 + \cdots + C_j}{C_{j+1} \cdots + C_b}.$$

5. Algorithm for the reference value

5.1. The largest root of the Wishart matrix $W(\nu, \boldsymbol{I}_q)$

We employ the formula based on the tube method in Kuriki and Takemura (2001). Our function computes

$$\Pr(W_1 \geqslant \lambda_0) = \sum_{e:even,\, e=0}^{\min(q,\nu)-1} \delta_{q+\nu-1-e} \overline{G}_{q+\nu-1-e}(\lambda_0),$$

$$\delta_{q+\nu-1-e} = \frac{\sqrt{\pi}\Gamma(q)\Gamma(\nu)}{\Gamma(q/2)\Gamma(\nu/2)} \left(-\frac{1}{2}\right)^{e/2} \times \frac{\Gamma\left(\frac{1}{2}(q+\nu-1) - \frac{1}{2}e\right)}{\Gamma(q-e/2)\Gamma(\nu-e/2)(e/2)!},$$

where $\overline{G}_l$ is the tail probability of the chi-squared distribution with the degrees of freedom l.

5.2. The largest root of the non-orthogonal Wishart matrix $W(a-1, C^{*'}C^{*})$

We employ the chi-squared approximation $d\chi^2_\nu$ proposed in Hirotsu (2009), where d and ν are determined by the equation

$$d\nu = E(W_1^*) = n\rho_1 + \left(1 - \frac{2}{n}\right) \sum_2^p \frac{\rho_1\rho_j}{\rho_1 - \rho_j} + \frac{2}{n} \sum\sum_{2\leqslant j<k\leqslant p} \frac{\rho_1\rho_j\rho_k}{(\rho_1-\rho_j)(\rho_1-\rho_k)},$$

$$2d^2\nu = V(W_1^*) = 2n\rho_1^2 + \frac{-2n^3+32n+144}{n^3} \sum_2^p \left(\frac{\rho_1\rho_j}{\rho_1-\rho_j}\right)^2 + \frac{2\left(n^2+6n-4\right)}{n^2} \sum\sum_{2\leqslant j<k\leqslant p} \frac{\rho_1^2\rho_j\rho_k}{(\rho_1-\rho_j)(\rho_1-\rho_k)}.$$

with $p = b-1$, $n = a-1$ and ρ_j the jth largest root of $C^{*'}C^{*}$. Therefore the tail probability is approximately given by

$$\Pr(W_1^* \geqslant \lambda_0^*) \approx \overline{G}_\nu(\lambda_0^*/d)$$

The approximation improves Anderson (2003) and Hirotsu (1991).

6. Stopping rule

6.1. Clustering algorithm into the prespecified number K of clusters

We employ the algorithm proposed in Hirotsu (2009) for obtaining a classification such that the generalized squared distance among clusters is large achieving simultaneously the homogeneity within each cluster.

(i) Specify K the number of clusters.

(ii) Start from a clusters each of which is composed of one row.

(iii) Let $G_1, \cdots, G_{a-k+1}$ be the cluster at the kth stage. Find two clusters G_i and $G_{i'}$ that give the smallest squared distance $\chi^2(G_i; G_{i'})$ based on (3) among all the possible combinations of two clusters from $G_1, \cdots, G_{a-k+1}$. Then, combine those two clusters to form $(a-k)$ clusters for the next $(k+1)$th stage.

(iv) Continue the procedure (iii) until the number of clusters becomes the prespecified number K. The resulting partition is denoted by $G_1, \cdots, G_K$. Then make an adjustment by the next algorithm.

(v) First, calculate the squared distance between the row 1 and the clusters $G_1(1), \cdots, G_K(1)$, where $G_k(1)$ denotes that the row 1 is eliminated from G_k. Then, classify the row 1 into the cluster that gives the smallest squared distance $\chi^2(1; G_k(1))$ among $k = 1, \cdots, K$. Do the same thing between the row 2 and the renewed clusters with row 2 eliminated. Continue the process repeatedly until no reduction in the generalized squared distance $\chi^2(G_1; \cdots; G_K)$ is obtained.

6.2. Stopping rule

We begin with $K = 2$ and continue the process until the generalized squared distance among the clusters $(G_1, \cdots, G_K)$ becomes significant for the first time at the pre-specified level α_1. Then we evaluate the variation within each cluster by the maximum eigen root at level α_2. If all the K clusters show the non-significant within variation we stop here concluding there are K clusters and give the interpretation of them. Example 2 at $K = 2$ is typical for this case. If all the K clusters show the significant within variation we proceed to the $K+1$ clusters and continue the process. As an intermediate case let the within variation be significant for n clusters and non-significant for $m(= K - n)$ clusters. Then we fix those m clusters and apply the clustering procedure anew to the rows in n clusters. As an example we obtain a significant classification at $K = 2$ in Example 1 with two clusters $G_1 = (1, \cdots, 7)$ and $G_8 = (8)$. The within variation in G_1 is significant so that $n = m = 1$ in this case. We fix and eliminate G_8 and apply the clustering procedure anew to G_1. In this particular case we need not adjust the significance level α_1 since it follows the closed testing procedure by Marcus, Peritz and Gabriel (1976). It suggests that it is reasonable to take $\alpha_2 = \alpha_1$ since otherwise we apply different α's to G_1 in testing within variation and in clustering procedure, respectively. We apply this generally in the following since there is no reason to choose any other particular value for α_2. The subgroup G_1 was not analyzed separately but analyzed as a part of the original table in Hirotsu (2009). It leads to putting the coefficient γ_j to zero for the eliminated rows in calculating the generalized squared distance among the clusters from G_1. Then the maximization does not reduce to the maximal eigen root problem and require a very complicated optimization procedure. The difference is only in the treatment of the column totals of the two-way table and there is only a slight difference in the outcome. Therefore we deal with G_1 independently from the eliminated rows in this article. In the general case of $m > 1$ we also apply the same level α_1 for the generalized squared distance in the sub-table composed of the n clusters after eliminating the m clusters.

7. Clustering columns

7.1. Both of the row and column categories are nominal

If both of the row and column categories are nominal without ordering then we can deal with the rows and columns symmetrically. Therefore we can apply the common Wishart distribution $W(\max(a-1, b-1), I_{(a-1,b-1)})$ for the columns as for rows, see Example 1.

7.2. Only the column categories are ordinal

Because of the natural ordering all the permutations of columns do not make sense and we are interested in the $(b-1)$ change-point type contrasts defined by the rows of $C^{*'}$. Then the reference distribution is that of

$$\max \chi^2(C^{*'}) = \max_j \left\| \left\{ R' \otimes \boldsymbol{c}^{*'}(j; j') \right\} \boldsymbol{z} \right\|^2,$$

where $\boldsymbol{c}^{*'}(j, j')$ is the jth row of $C^{*'}$ as defined in (6). It is asymptotically that of the maximum of the correlated chi-squared variables and a very efficient algorithm for the p-value is given in Hirotsu et al. (1992) based on the Markov property of the successive components. The explicit form of the cumulative distribution function is given below and we realized it up to $b = 5$ as an R system, where ρ_{jk} is given in (7) and $G_l(x)$ is the cumulative distribution function of the chi-square distribution with the degrees of freedom l.

$$b = 3 \qquad \Pr(\max \chi^2(C^{*'}) \leqslant c) = \left(1 - \rho_{12}^2\right)^{\frac{\nu}{2}} \sum_{k=0}^{\infty} \rho_{12}^{2k} \frac{\Gamma\left(\frac{\nu}{2} + k\right)}{\Gamma\left(\frac{\nu}{2}\right) k!} G_{\nu+2k} \left(\frac{c}{1 - \rho_{12}^2}\right)^2$$

$$\begin{aligned} b = 4 \qquad & \Pr(\max \chi^2(C^{*'}) \leqslant c) \\ & = \left\{\frac{(1-\rho_{12}^2)(1-\rho_{23}^2)}{1-\rho_{13}^2}\right\}^{\frac{\nu}{2}} \\ & \times \sum_{k=0}^{\infty} \left\{\frac{\rho_{12}^2(1-\rho_{23}^2)}{1-\rho_{13}^2}\right\}^k G_{\nu+2k}\left(\frac{c}{1-\rho_{12}^2}\right) \\ & \times \sum_{m=0}^{\infty} \left\{\frac{(1-\rho_{12}^2)\rho_{34}^2}{1-\rho_{12}^2}\right\}^m G_{\nu+2m}\left(\frac{c}{1-\rho_{23}^2}\right) \\ & \times \frac{\Gamma\left(\frac{\nu}{2}+k+m\right)}{\Gamma\left(\frac{\nu}{2}\right) k!\, m!} G_{\nu+2k+2m}\left(\frac{(1-\rho_{13}^2)\, c}{(1-\rho_{12}^2)(1-\rho_{23}^2)}\right) \end{aligned}$$

$$\begin{aligned} b = 5 \qquad & \Pr(\max \chi^2(C^{*'}) \leqslant c) \\ & = \left\{\frac{(1-\rho_{12}^2)(1-\rho_{23}^2)(1-\rho_{34}^2)}{(1-\rho_{13}^2)(1-\rho_{24}^2)}\right\}^{\frac{\nu}{2}} \\ & \times \sum_{k=0}^{\infty} \left\{\frac{(1-\rho_{12}^2)\rho_{23}^2(1-\rho_{34}^2)}{(1-\rho_{13}^2)(1-\rho_{24}^2)}\right\}^k \sum_{m=0}^{\infty} \left\{\frac{(1-\rho_{23}^2)\rho_{34}^2}{1-\rho_{24}^2}\right\}^m \\ & \times \frac{\Gamma\left(\frac{\nu}{2}+k+m\right)}{\Gamma\left(\frac{\nu}{2}\right) k!\, m!} G_{\nu+2k+2m}\left(\frac{(1-\rho_{24}^2)\, c}{(1-\rho_{23}^2)(1-\rho_{34}^2)}\right) G_{\nu+2m}\left(\frac{c}{1-\rho_{34}^2}\right) \\ & \times \sum_{j=0}^{\infty} \left\{\frac{\rho_{12}^2(1-\rho_{23}^2)}{1-\rho_{13}^2}\right\}^j \frac{\Gamma\left(\frac{\nu}{2}+k+j\right)}{\Gamma\left(\frac{\nu}{2}+k\right) j!} \\ & \times G_{\nu+2k+2j}\left(\frac{(1-\rho_{13}^2)\, c}{(1-\rho_{12}^2)(1-\rho_{23}^2)}\right) G_{\nu+2j}\left(\frac{c}{1-\rho_{12}^2}\right) \end{aligned}$$

If the $\max\chi^2(C^{*'})$ is significant between j and $j+1$ we apply anew the procedure to each of the subgroups $G_1 = (1, \cdots, j)$ and $G_2 = (j+1, \cdots, b)$. If the $\max\chi^2$ is not significant at any stage we stop there concluding the subgroups are homogeneous.

7.3. Both of the row and column categories are ordinal

We apply the change-point type contrasts to both of rows and columns. There are $(a-1)$ cut points in rows and $(b-1)$ cut points in columns so that we have $(a-1)(b-1)$ sub-tables. We take the maximum goodness of fit χ^2 of those 2×2 tables and call it $\max\max\chi^2(R^{*'}\times C^{*'})$,

$$\max\max\chi^2(R^{*'}\times C^{*'}) = \max_{1\leqslant i\leqslant a-1}\max_{1\leqslant j\leqslant b-1}\left\|\left(\boldsymbol{r}^{*'}(i;i')\otimes\boldsymbol{c}^{*'}(j;j')\right)\boldsymbol{z}\right\|^2$$

where $R^{*'}$ and $\boldsymbol{r}^{*'}(i;i')$ are similarly defined as $C^{*'}$ and $\boldsymbol{c}^{*'}(j;j')$. The exact algorithm for the distribution function of $\max\max\chi^2$ is given in Hirotsu (1997). Let

$$Y_{IJ} = \sum_{i\leqslant I}\sum_{j\leqslant J} y_{ij}, \quad R_I = \sum_{i\leqslant I} R_i, \quad C_J = \sum_{j\leqslant J} C_j$$

and

$$Y_{IJ}^* = \left(Y_{IJ} - \frac{R_I C_J}{N}\right)\left(\frac{R_I(N-R_I)C_J(N-C_J)}{N^3}\right)^{-1/2}$$

be the standardized version of Y_{ij}. Define a conditional probability

$$F_k(\boldsymbol{Y}_k) = \Pr\left\{\boldsymbol{Y}_1^* \le c_{\boldsymbol{j}}, \cdots, \boldsymbol{Y}_k^* \le c_{\boldsymbol{j}} \mid \boldsymbol{Y}_k\right\},$$

where $\boldsymbol{Y}_k = (Y_{1k}, \cdots, Y_{ak})'$ and $\boldsymbol{Y}_k^* \le c_{\boldsymbol{j}}$ means $Y_{Ik}^* < c$ for $I = 1, \cdots, a-1$. It should be noted here that Yak is a fixed marginal total. Then we have a recurrence formula

$$F_{k+1}(\boldsymbol{Y}_{k+1}) = \sum_{\boldsymbol{Y}_k} F_k(\boldsymbol{Y}_k) f(\boldsymbol{Y}_k \mid \boldsymbol{Y}_{k+1})$$

where $f(\boldsymbol{Y}_k \mid \boldsymbol{Y}_{k+1})$ is a conditional probability of $\boldsymbol{Y}_k$ given $\boldsymbol{Y}_{k+1}$. To be exact define a matrix

$$D = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}_{a\times a}$$

then $D\boldsymbol{Y}_k$ is distributed as a multinomial given the row total $D\boldsymbol{Y}_{k+1}$ and the column total Y_{ak}. The p-value is obtained finally by $1 - F_K(\boldsymbol{Y}_K)$. The algorithm is realized as the R system.

8. Examples

8.1. Israeli adults cross classified by principal worries and residence places

Greenacre (1988) applied the method of Hirotsu (1983a) for clustering the rows and columns of the 8×5 contingency table reported by Guttman (1971). It cross tabulates 1554 Israeli adults according to the row categories of principal worries and the column categories depending on their place of residence and that of their respective fathers. The row categories are as follows:

(i) OTH-other worries
(ii) POL-political situation
(iii) MIL-military situation
(iv) ECO-economic situation
(v) ENR-enlisted relative
(vi) SAB-sabotage
(vii) MTO-more than one worry
(viii) PER-personal economics.

First applying the clustering procedure to rows we obtain a highly significant squared distance 77.90 between the clusters $G_1(1, 2, 3, 4, 5, 6, 7)$ and $G_8(8)$ at $K = 2$ for the reference value 23.55 at $\alpha = 0.05$ from Wishart distribution $W(7, I_4)$. The within variation 24.66 of the cluster G_1 evaluated as the largest root of $W(6, I_4)$ is significant with the p-value 0.018. Therefore we separate G_8 and apply the clustering procedure anew to G_1. The generalized chi-squared distances $\chi^2(1; 2, 3, 4, 5, 6, 7) = 20.77$ at $K = 2$ and $\chi^2(1; 2; 3, 4, 5, 6, 7) = 20.99$ at $K = 3$ are non-significant as the largest root of $W(6, I_4)$ and we obtain a significant result first at $K = 4$ with the chi-squared distance $\chi^2(1; 2; 3, 4; 5, 6, 7) = 23.195$ and the related p-value 0.030. The generalized squared distance among five clusters $G_1(1), G_2(2), G_3(3, 4), G_5(5, 6, 7)$, $G_8(8)$ is 91.81 and its relative contribution to the largest root 92.73 of the original table is 0.99.

Now, the software first present the original data, calculate the Wishart matrix $W'W$ via the vector $\boldsymbol{w}_k$ (5) with $m = a$ and its largest root as 92.73

which is evaluated as highly significant by the Wishart $W(7, I_4)$ distribution. Then the clustering algorithm starts for the prespecified number of clusters $K = 2, \cdots, 8$. The search for significant clustering at prespecified $\alpha = 0.05$ goes like this. First try $K = 2$ to find clustering into $G_1(1, \cdots, 7)$ and $G_8(8)$ is highly significant with p-value 0.13×10^{-10}. Then check the within variation of sub-clusters to find G_1 is inhomogeneous with $p = 0.018$ by the Wishart $W(6, I_4)$. Therefore re-clustering of G_1 starts to obtain a significant clustering first at $K = 4$ with the generalized squared distance 23.20 and the related p-value 0.030. The within variations of sub-clusters $G_3(3, 4)$ and $G_5(5, 6, 7)$ are evaluated non-significant with the p-values 0.949 and 0.718 by the Wishart $W(4, I_1)$ and Wishart $W(4, I_2)$, respectively. Therefore the algorithm stops here and gives the summary of classification, the generalized squared distance among sub-clusters 91.84 and its contribution 0.99 to the original largest root 92.73.

In this case, however, the number of clusters 5 is too large for $a = 8$ and the relative contribution looks excessively high. Therefore we may try other significance level $\alpha = 0.10$, say. Then we can separate the row 1 from G_1 with p-value 0.081. The within variation in the counterpart sub-cluster $G_2(2, \cdots, 7)$ is 15.17 with the p-value 0.19 by the Wishart $W(5, I_4)$ and the algorithm stops here. The relative contribution of $G_1(1)$, $G_2(2, \cdots, 7)$ and $G_8(8)$ is still 0.88 and reasonably high.

Since the column categories are also nominal we can apply the same procedure as rows and obtain a significant clustering $G_1(1, 2)$ and $G_3(3, 4, 5)$ at $K = 2$. Applying the largest root test to each of G_1 and G_3 the latter is found to be homogeneous. On the other hand the largest root 24.55 of G_1 is inhomogeneous at $\alpha = 0.05$ and it immediately suggests that G_1 should be separated into $G_1(1)$ and $G_2(2)$ since this is only one possible classification. The generalized squared distance 90.54 among G_1, G_2, G_3 explains 98% of the largest root 92.73 of the original table. If we employ $\alpha = 0.10$, we have the same result.

The collapsed data and the simple departure measures from independence $y_{ij}/(R_i C_j/N)$ are given in Table 1. It is seen that the row 8 is strongly associated with the cluster $G_3(3, 4, 5)$ of columns. Row 1 is strongly associated with column 2 and the cluster $G_2(2, \cdots, 7)$ of rows is associated with column 1.

Table 1 about here

8.2. Cancer patients cross classified by their occupation and initial condition of illness

The data of Table 2 were first analyzed in Hirotsu (1977), where a highly significant classification into $G_1(1,2,3,6,7,9)$ and $G_4(4,5,6,10)$ was obtained based on the cumulative chi-squared statistics reflecting on the ordinal nature of column categories. Then to evaluate the within variation of G_1 and G_4 the largest roots $W_1^*(1,2,3,6,7,9) = 3.37$ and $W_1^*(4,5,8,10) = 2.83$ are calculated. Then the respective p-values of W_1^* are obtained as 0.86 and 0.34 by the chi-squared approximation for the largest root of $W(5 : C_1^{*'}C_1^*)$ and $W(3 : C_4^{*'}C_4^*)$, where the matrices $C_i^{*'}C_i^*$, $i = 1,4$ are calculated by the equation (6). Therefore the procedure stops here concluding there are two clusters.

Since the column categories are ordinal we apply the change-point-type contrasts for clustering columns and find the max $\chi^2(C^{*'})$ to be 91.25 and highly significant for the partition $F_1(1)$ and $F_2(2,3)$ by the method of Section 7.2 with $b = 3$. The max $\chi^2(C^{*'})$ for the cluster $F_2(2,3)$ is 5.23 and non-significant as the chi-squared with 9 degrees of freedom. Therefore we conclude there are two clusters F_1 and F_2. The collapsed data in two-ways and their simple departure measures $y_{ij}/(R_iC_j/N)$ are given in Table 3. The cluster G_1 is clearly characterized by the high proportion of the slight condition F_1 relatively to G_4.

Table 2 about here

Table 3 about here

The result of clustering suggests a simple block interaction model

$$p_{ij} = p_{i.}\, p_{.j}\, \theta_{\ell m}, \qquad \begin{pmatrix} \ell = 1 \ \text{ for } \ i = 1,2,3,6,7,9; \ \ell = 2 \ \text{ for } \ i = 4,5,8,10 \\ m = 1 \ \text{ for } \ j = 1; \ \ m = 2 \ \text{ for } \ j = 2,3 \end{pmatrix}$$

with the parameter θ for interaction with only one degree of freedom. The goodness of fit chi-square reduces from 95.75 of the independence model to 8.04 by adding only one parameter for interaction. The fitted values are in the right side of the original data with upper column from independence model and the lower column from block interaction model. It is seen that the improvement of fit is remarkable.

9. Concluding remarks

The row- and column-wise multiple comparison procedures have been proposed for a two-way contingency table. In particular the change-point type contrasts were employed for reflecting the up- and down- ward tendency along with the ordinal categories. The p-value calculations were implemented for the largest root of an orthogonal and non-orthogonal Wishart matrices. For evaluating the maximum of the correlated chi-squared statistics a recurrence formula was also implemented. A stopping rule working automatically for obtaining a clustering of rows and columns into a reasonable number of clusters was introduced for dealing with a large table. The real examples in Section 8 show that the 8 by 5 and 10 by 3 tables can be reduced to 3 by 3 and 2 by 2 tables, respectively, giving a simple and clear interpretation of the data.

References

[1] Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley Intersciences, New York.

[2] Greenacre, M.J. (1988). Clustering the rows and columns of a contingency table. *J. Classification*, 5, 39-51.

[3] Guttman, L. (1971). Measurement as structural theory. *Psychometrika*, 36, 329-347.

[4] Hirotsu, C. (1977). Multiple comparisons and clustering rows in a contingency table. *Quality* 7, 27-33 (in Japanese).

[5] Hirotsu, C. (1982), Use of cumulative efficient scores for testing ordered alternatives in discrete models, *Biometrika*, 69, 567-577.

[6] Hirotsu, C. (1983a). Defining the pattern of association in two-way contingency tables. *Biometrika*, 70, 579-589.

[7] Hirotsu, C. (1983b). An Approach to defining the pattern of interaction effects in a two-way layout. *Annals Inst. Statist. Math. A*, 35, 77-90.

[8] Hirotsu, C. (1991). An approach to comparing treatments based on repeated measures. *Biometrika*, 78, 583-594.

[9] Hirotsu, C., Kuriki, S., Hayter, A.J. (1992), Multiple comparison procedures based on the maximal component of the cumulative chisquared statistic, *Biometrika*, 79, 381-392.

[10] Hirotsu, C. (1993). Beyond analysis of variance techniques: Some applications in clinical trials. *Int. Statist.* Rev. 61, 183-201.

[11] Hirotsu, C., Ohta, E., Hirose, N., Shimizu, K. (2003). Profile analysis of 24-h measurements of blood pressure. *Biometrics*, 59, 907-915.

[12] Hirotsu, C. (2009), Clustering rows and/or columns of a two-way contingency table and a related distribution theory. *Comp. Stat. and Data Analysis*, 53, 4508-4515.

[13] Kuriki, S., Takemura, A. (2001), Tail Probabilities of the Maxima of Multilinear Forms and Their Applications *The Annals of Statistics*, Vol.29, No.2, 328-371

[14] Marcus, R., Peritz, E., Gabriel, K.R. (1976), On closed testing procedures with special reference to ordered analysis of variance, *Biometrika*, 63, 655-660.

Table 1: Collapsed data

	Collapsed table			Simple departure measure		
Cluster	$G_1(1)$	$G_2(2)$	$G_3(3,4,5)$	$G_1(1)$	$G_2(2)$	$G_3(3,4,5)$
$G_1(1)$	128	52	107	0.88	1.58	0.98
$G_2(2,\cdots,7)$	610	110	356	1.12	0.89	0.87
$G_8(8)$	48	16	127	0.50	0.73	1.75

Table 2: Number of cancer patients cross classified by occupation and severity of illness with fitted values on the right side

Severity Occupation	Slight		Medium		Serious		Total
1. Professional & technical workers	148	123.3 148.5	444	473.9 452.4	86	80.8 77.1	678
2. Manager and officials	111	93.1 112.1	352	357.9 341.6	49	60.0 58.2	512
3. Clerical and related workers	645	524.6 631.4	1911	2015.7 1924.4	328	343.7 328.1	2884
4. Sales workers	165	191.9 160.7	771	737.4 764.0	119	125.7 130.3	1055
5. Formers, lumbermen, fisherman, quarrymen & etc.	383	458.9 384.3	1829	1763.4 1827.2	311	300.6 311.5	2523
6. Workers in transport and communication systems	96	79.3 95.5	293	304.7 290.9	47	52.0 49.6	436
7. Craftsmen	98	88.4 106.4	330	339.7 324.3	58	57.9 55.3	486
8. Production process workers	199	223.4 187.1	874	858.3 889.3	155	146.3 151.6	1228
9. Service workers	59	52.4 63.1	199	201.3 192.2	30	34.3 32.8	288
10. Persons without regular occupations	262	330.7 276.9	1320	1270.7 1316.6	236	216.6 224.5	1818
Total	2166		8323		1419		11908

Table 3: Collapsed data

	Collapsed table		Simple departure measure	
Cluster	Slight	Medium & Serious	Slight	Medium & Serious
$G_1(1,2,3,4,7,9)$	1157	4127	1.20	0.95
$G_2(4,5,8,10)$	1009	5615	0.84	1.04