On collapsing categories in two–way contingency tables

Maria Kateri[1] and George Iliopoulos[2]

[1]Department of Philosophy - Education - Psychology, University of Ioannina, 45110 Ioannina, Greece, e-mail: *me00126@cc.uoi.gr*

[2]Department of Mathematics, University of the Aegean, 83200 Karlovasi, Samos, Greece, e-mail: *geh@aegean.gr*

Abstract

The issue of collapsing categories of a contingency table's classification variables is well–known and has been dealt in the framework of classical models such as models of independence and association, canonical correlation and logistic regression. The most often used criterion is based on the homogeneity of the corresponding categories which was connected to association and correlation models by Goodman (1981a, b) and Gilula (1986) respectively. In this paper we relate homogeneity to a class of generalized association models, based on the f–divergence. The main issue raised in this paper is that the homogeneity and the structural criteria can not be contradictory. It is proved that collapsing among homogeneous categories does not affect the underlying structure of the table.

AMS 2000 subject classifications: 62H17

Key words: Contingency tables, collapsing categories, generalized association models, f–divergence.

1 Introduction

The subject of grouping rows or/and columns (usually successive) of a contingency table is as old as contingency tables analysis itself and one can find related references even in the very early literature on contingency tables (Yates, 1948). The issue of collapsibility remains active and each time is connected with the new developments on the analysis of contingency tables. The main motivations for collapsing classification categories are to easily face the problem of small counts in a particular row or column (it is collapsed to the most relevant), to detect possible overdispersions of a classification scale and to simplify the analysis of association between the two classification variables (usually a simpler model fits well to the reduced table). The basic criteria for collapsing (or not) are those of homogeneity (Benzécri, 1973; Hirotsu, 1983; Gilula, 1986; Gilula and Krieger, 1989; Weller and Romney, 1990; Beh, 1997,

1998; Wermuth and Cox, 1998) and structure (Goodman, 1981b, 1985). Also, when a classification variable is ordinal, the violation of the ordering of certain estimated scores is a reason to collapse the corresponding categories to ensure the known order (Goodman, 1985, 1986; Agresti et al., 1987; Ritov and Gilula, 1991, 1993).

Let $\mathbf{\Pi} = (\pi_{ij})$ be the $I \times J$ probability table corresponding to the cross–classification of two categorical variables with I and J categories respectively. Denoting their marginal distributions by $\pi_{i\cdot} = \sum_{j=1}^{J} \pi_{ij}$ and $\pi_{\cdot j} = \sum_{i=1}^{I} \pi_{ij}$, two rows labeled s and t are said to be homogeneous, if

$$\frac{\pi_{sj}}{\pi_{s\cdot}} = \frac{\pi_{tj}}{\pi_{t\cdot}} \ , \quad \forall \ j = 1, \ldots, J \ , \tag{1.1}$$

that is, the corresponding conditional column probabilities are equal. Note that $\pi_{sj}/\pi_{s\cdot}$, $j = 1, \ldots, J$, has been named by Benzécri (1973) as the s-th row profile. Thus homogeneity is expressed as equality of the corresponding row profiles. This definition can obviously be extended to more than two rows while homogeneous columns are defined similarly.

Homogeneity was initially related to the basic model of independence, due to its basic property: "Independence holds for every subtable formed by homogeneous rows or columns." As a consequence, if we denote by $\mathbf{I}$ and $\tilde{\mathbf{I}}$ the models of independence for the initial $I \times J$ and the collapsed $\tilde{I} \times \tilde{J}$ tables respectively ($\tilde{I} \leqslant I$, $\tilde{J} \leqslant J$), then the difference of the likelihood ratio statistics for the fit of models $\mathbf{I}$ and $\tilde{\mathbf{I}}$, $G^2(\mathbf{I}) - G^2(\tilde{\mathbf{I}})$, should not be statistically significant, provided the collapsing has been done among homogeneous rows and columns (see Williams, 1952 and Goodman, 1985). A similar idea has been also developped in correspondence analysis framework by Benzécri (1973), who introduced the principle of distributional equivalence.

Goodman (1981b) noted that when independence is rejected for the initial table, collapsing homogeneous categories can affect the underlying association structure, although the fit of independence remains very bad. This led him to the introduction of the structural criterion, according to which, two (or more) homogeneous categories can be collapsed only if the association structure remains unchanged.

The main issue raised in this paper is that the predominant criteria of homogeneity and structure for which so far was supported that they can sometimes be contradictory (Goodman, 1981b; Gilula, 1986), are in agreement. That is, collapsing between homogeneous classification categories ensures the preservation of the underlying structure of the probability table $\mathbf{\Pi}$. As a consequence, no simpler model should be appropriate for the collapsed table. Nevertheless, some simple association structures should naturally be excluded from this statement, since they can not coexist with certain homogeneities (see Section 5). In our context, the structure is described in terms of a generalized association model based on an information theoretic setup,

which includes the models used by Goodman and Gilula as special cases.

Next we outline the layout of our paper. In Section 2 we describe the generalized association models used throughout this paper. Section 3 contains the theoretical results that support our assertions about the homogeneity and structural criteria explained above. An illustrative example is provided in Section 4, while Section 5 contains comments and conclusions.

2 Generalized association models

In the context of contingency tables analysis, the association and correlation models are well–known (cf. Goodman, 1985, 1986). For an $I \times J$ contingency table $\mathbf{\Pi} = (\pi_{ij})$ and for $1 \leqslant K \leqslant M = \min(I, J) - 1$, the association model of K-th order, denoted by RC(K), is defined by

$$\pi_{ij} = a_i b_j \exp\left(\sum_{k=1}^{K} \phi_k \mu_{ik} \nu_{jk}\right), \quad i = 1, \ldots, I \ , \ j = 1, \ldots, J \ . \tag{2.1}$$

In particular, for $K = 1$, the model is the multiplicative row–column association model, simply noted by RC, whereas when $K = M$, RC(M) is the saturated model. The parameters a_i, $i = 1, \ldots, I$, and b_j, $j = 1, \ldots, J$, are the row and column main effects respectively, while the vectors $\boldsymbol{\mu}_k = (\mu_{1k}, \ldots, \mu_{Ik})'$ and $\boldsymbol{\nu}_k = (\nu_{1k}, \ldots, \nu_{Jk})'$ are the row and column scores corresponding to the k-th term of the interaction sum, $k = 1, \ldots, K$. In the related literature, the k-th term is referred as the k-th axis due to the graphical displays of the row and column scores, often used for visualization purposes. The ϕ_k's are known as the intrinsic association parameters. On the row and column scores are imposed the constraints

$$\sum_{i=1}^{I} w_{1i} \mu_{ik} = \sum_{j=1}^{J} w_{2j} \nu_{jk} = 0 \ , \quad k = 1, \ldots, K \ , \tag{2.2}$$

and

$$\sum_{i=1}^{I} w_{1i} \mu_{ik} \mu_{i\ell} = \sum_{j=1}^{J} w_{2j} \nu_{jk} \nu_{j\ell} = \delta_{k\ell} \ , \quad k, \ell = 1, \ldots, K \ , \tag{2.3}$$

where $\delta_{k\ell}$ is the Kronecker's delta, while w_{1i} $(i = 1, \ldots, I)$ and w_{2j} $(j = 1, \ldots, J)$ are row and column positive weights respectively. In the literature, the common used weights are the uniform ($w_{1i} = w_{2j} = 1$, for all i, j) and the marginal ($w_{1i} = \pi_{i\cdot}$, $w_{2j} = \pi_{\cdot j}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$). For a detailed related justification see Goodman (1985) and Becker and Clogg (1989).

In an analogue manner, the correlation model of K-th order is defined as

$$\pi_{ij} = \pi_{i.}\pi_{.j}\left(1 + \sum_{k=1}^{K} \phi_k \mu_{ik}\nu_{jk}\right), \quad i = 1, \ldots, I \ , \ j = 1, \ldots, J \ , \tag{2.4}$$

and denoted by CA(K). The simplest model is obtained for $K = 1$ and is denoted by CA, while CA(M) is the saturated model. The row and column scores $\boldsymbol{\mu}_k$ and $\boldsymbol{\nu}_k$ of CA(K) satisfy also the constraints (2.2) and (2.3) but with the marginal weights.

The main qualitative difference between these two classes of models is that although both of them are models of dependence, the association models are (under certain conditions) the closest to independence in terms of the Kullback–Leibler distance, while correlation models in terms of the Pearsonian distance (Gilula et al., 1988). Rom and Sarkar (1992), Kateri and Papaioannou (1994) and Goodman (1996) introduced general classes of dependence models which express the departure from independence in terms of generalized measures and include association and correlation models as special cases.

The generalized measure used by Kateri and Papaioannou (1994) was the f–divergence. If $\mathbf{P} = (p_{ij})$ and $\mathbf{Q} = (q_{ij})$ are two discrete finite bivariate probability distributions, then the f–divergence between $\mathbf{P}$ and $\mathbf{Q}$ (or Csiszar's measure of information in $\mathbf{Q}$ about $\mathbf{P}$) is defined by

$$\mathrm{I}^{\mathrm{C}}(\mathbf{P}, \mathbf{Q}) = \sum_{i,j} q_{ij} f(p_{ij}/q_{ij}), \tag{2.5}$$

where f is a real–valued convex function on $[0, \infty)$ with $f(1) = f'(1) = 0$, $0f(0/0) = 0$, $0f(y/0) = \lim_{x\to\infty} f(x)/x$.

Let $F(x) = f'(x)$. Kateri and Papaioannou (1994) introduced the generalized association model of order K, which in the sequel will be denoted by RC[f](K), setting

$$\pi_{ij} = \pi_{i.}\pi_{.j} F^{-1}\left(\alpha_i + \beta_j + \sum_{k=1}^{K} \phi_k \mu_{ik}\nu_{jk}\right) \ , \ i = 1, \ldots, I, \ j = 1, \ldots, J \ , \tag{2.6}$$

where F^{-1} denotes the inverse function of F and $\boldsymbol{\mu}_k$ and $\boldsymbol{\nu}_k$ satisfy (2.2) and (2.3). Model (2.6) is equivalent to the generalized linear model

$$\begin{aligned} F\left(\frac{\pi_{ij}}{\pi_{i.}\pi_{.j}}\right) &= \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_{ij}^{(12)} \ , \\ &= \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \sum_{k=1}^{K} \phi_k \mu_{ik}\nu_{jk}, \ i = 1, \ldots, I, \ j = 1, \ldots, J \ , \end{aligned} \tag{2.7}$$

where

$$\sum_{i=1}^{I} w_{1i}\lambda_i^{(1)} = \sum_{j=1}^{J} w_{2j}\lambda_j^{(2)} = \sum_{i=1}^{I} w_{1i}\lambda_{ij}^{(12)} = \sum_{j=1}^{J} w_{2j}\lambda_{ij}^{(12)} = 0 \ , \tag{2.8}$$

and the matrix of interactions $\mathbf{\Lambda} = (\lambda_{ij}^{(12)})$ is of rank K. Via the Generalized Singular Value Decomposition (GSVD), $\mathbf{\Lambda}$ is expressed as $\mathbf{\Lambda} = \mathbf{M\Phi N}'$, where $\mathbf{M} = (\mu_{ik})$ $(I \times K)$ and $\mathbf{N} = (\nu_{jk})$ $(J \times K)$, the left and right singular vectors respectively, are orthonormalized with respect to $\mathbf{W}_1 = \mathrm{diag}(w_{11}, \ldots, w_{1I})$ and $\mathbf{W}_2 = \mathrm{diag}(w_{21}, \ldots, w_{2J})$, e.g. they satisfy $\mathbf{M}'\mathbf{W}_1\mathbf{M} = \mathbf{N}'\mathbf{W}_2\mathbf{N} = \mathbf{I}_K$, the K-th order identity matrix, and $\mathbf{\Phi} = \mathrm{diag}(\phi_1, \ldots, \phi_K)$ with $\phi_1 \geqslant \ldots \geqslant \phi_K > 0$. It is important to highlight that the order K of the generalized association model coincides with the rank of the generalized linear model interaction parameters matrix and is not affected by the choice of the weights. In particular, the following lemma holds.

Lemma 2.1. *If* $\mathbf{W}_1^* = \mathrm{diag}(w_{11}^*, \ldots, w_{1I}^*)$ *and* $\mathbf{W}_2^* = \mathrm{diag}(w_{21}^*, \ldots, w_{2J}^*)$ *are matrices of weights, then the corresponding interaction parameters matrix* $\mathbf{\Lambda}^*$ *is expressed in terms of* $\mathbf{\Lambda}$ *as*

$$\mathbf{\Lambda}^* = [\mathbf{I}_I - \mathbf{1}_I\mathbf{W}_1^*/\mathrm{trace}(\mathbf{W}_1^*)]\ \mathbf{\Lambda}\ [\mathbf{I}_J - \mathbf{1}_J\mathbf{W}_2^*/\mathrm{trace}(\mathbf{W}_2^*)]' \ , \tag{2.9}$$

where $\mathbf{1}_I$, $\mathbf{1}_J$ *are the* $I \times I$, $J \times J$ *matrices of ones. Moreover,* $\mathrm{rank}(\mathbf{\Lambda}^*) = \mathrm{rank}(\mathbf{\Lambda}) = K$.

Notice that the equality of ranks stated in Lemma 2.1 is not immediate since the matrices multiplying $\mathbf{\Lambda}$ in (2.9) are idempotent with ranks $I - 1$ and $J - 1$ respectively. The row and column scores of the corresponding generalized association model RC$[f](K)$ are the generalized singular vectors of $\mathbf{\Lambda}^*$ orthonomalized with respect to $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$.

Remark 2.1. The parameters ϕ_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\nu}_k$, $k = 1, \ldots, K$, in (2.1), (2.4) and (2.6) are not the same. We adopt unified notation for these parameters of any model since we make use only of their qualitative identity and not their magnitude. Note also that the rank K of $\mathbf{\Lambda}$ in (2.6) in general varies for different choices of f.

Examples. 1) Let $f_0(x) = x\log x + 1 - x$, $F(x) = f_0'(x) = \log x$. Then, model (2.7) is equivalent to the well–known log–linear model

$$\log \pi_{ij} = u + u_i^{(1)} + u_j^{(2)} + u_{ij}^{(12)},\ i = 1, \ldots, I,\ j = 1, \ldots, J \ , \tag{2.10}$$

with

$$u = \lambda + \frac{\sum_i w_{1i} \log \pi_{i\cdot}}{\sum_i w_{1i}} + \frac{\sum_j w_{2j} \log \pi_{\cdot j}}{\sum_j w_{2j}} \ ,$$

$$u_i^{(1)} = \lambda_i^{(1)} + \log \pi_{i\cdot} - \frac{\sum_i w_{1i} \log \pi_{i\cdot}}{\sum_i w_{1i}} \ ,$$

$$u_j^{(2)} = \lambda_j^{(2)} + \log \pi_{\cdot j} - \frac{\sum_j w_{2j} \log \pi_{\cdot j}}{\sum_j w_{2j}} \ , \quad \text{and}$$

$$u_{ij}^{(12)} = \lambda_{ij}^{(12)} \ .$$

The parameters of model (2.10) satisfy also the constraints (2.8). Considering uniform weights, (2.8) are reduced to the traditional constraints used in the log–linear models framework. Since $\text{rank}(\mathbf{\Lambda}) = K$, model (2.10) is equivalent to the standard association model RC(K) in (2.1).

2) Consider the power divergence loss function $f_r(x) = [x^{r+1} - x + r(1-x)]/[r(r+1)]$, $r \neq -1, 0$ (see Read and Cressie, 1988, p.128), for which $F(x) = f_r'(x) = (x^r - 1)/r$, i.e. the Box and Cox (1964) power transformation. Then, model (2.7) becomes

$$\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \left[1 + r \left(\alpha_i + \beta_j + \sum_{k=1}^{K} \phi_k \mu_{ik} \nu_{jk} \right) \right]^{1/r} \ , \tag{2.11}$$

which is essentially equivalent to the power model of Baccini et al. (1993). For $K = 1$ the model is first introduced by Rom and Sarkar (1992). Notice also that in the special case $r = 1$ and for marginal weights, model (2.11) reduces to the canonical correlation model CA(K) in (2.4), while for $r \to 0$ it coincides with the association model RC(K), since $\lim_{r \to 0} f_r(x) = f_0(x)$.

3 Main results

As already mentioned in the Introduction, the predominant criterion for collapsibility is that of homogeneity. Goodman (1981b, 1986) connected homogeneity to association models by stating that the equality of the scores in the saturated association model, RC(M), implies homogeneity of the corresponding categories. Later, Gilula (1986) proved the equivalence of these two issues for the saturated canonical correlation model CA(M). In Theorem 3.2 below, we extend this result for the generalized association model RC[f](K), for any choice of the divergence measure f and for any $K \leqslant M$.

The following lemma provides two useful equalities connecting the interaction parameters of the generalized linear model (2.7) with corresponding row or column scores. They are originally derived by Goodman (1996, p.421).

Lemma 3.1. *Let* $\mathbf{\Pi} = (\pi_{ij})$ *be an* $I \times J$ *contingency table with structure* RC$[f](K)$ *given by (2.6),* $K \leqslant \min(I, J) - 1$. *Then, for* $1 \leqslant s, t \leqslant I$, $1 \leqslant p, q \leqslant J$, *the following equalities hold.*

$$\sum_{j=1}^{J} w_{2j}(\lambda_{sj}^{(12)} - \lambda_{tj}^{(12)})^2 = \sum_{k=1}^{K} \phi_k^2(\mu_{sk} - \mu_{tk})^2 \ , \tag{3.1}$$

$$\sum_{i=1}^{I} w_{1i}(\lambda_{ip}^{(12)} - \lambda_{iq}^{(12)})^2 = \sum_{k=1}^{K} \phi_k^2(\nu_{pk} - \nu_{qk})^2 \ . \tag{3.2}$$

Since w_{2j}, ϕ_k^2 are strictly positive for all j, k, equality (3.1) implies that $\mu_{sk} = \mu_{tk}$, $k = 1, \ldots, K$, if and only if $\lambda_{sj}^{(12)} = \lambda_{tj}^{(12)}$, $j = 1, \ldots, J$, that is, $\mathbf{\Lambda}$ has its s-th and t-th rows identical. The analogous conclusion follows from (3.2) for the p-th and q-th columns of $\mathbf{\Lambda}$.

Theorem 3.2. *Let* $\mathbf{\Pi} = (\pi_{ij})$ *be an* $I \times J$ *contingency table with structure* RC$[f](K)$ *given by (2.6),* $K \leqslant \min(I - 2, J - 1)$. *A necessary and sufficient condition for two distinct rows* s *and* t *of* $\mathbf{\Pi}$ *to be homogeneous is that* $\mu_{sk} = \mu_{tk}$, $k = 1, \ldots, K$, *where* μ_{sk} *and* μ_{tk} *are the* s*-th and* t*-th row scores of the underlying generalized association model* RC$[f](K)$.

Proof. Set $F_{ij} = F(\pi_{ij}/\pi_{i\cdot}\pi_{\cdot j})$. Then, by (1.1), homogeneity of the rows s and t is equivalent to

$$F_{sj} = F_{tj} \ , \ j = 1, \ldots, J \ , \tag{3.3}$$

since F is a strictly monotone function. Furthermore,

$$F_{sj} - F_{tj} = (\lambda_s^{(1)} - \lambda_t^{(1)}) + (\lambda_{sj}^{(12)} - \lambda_{tj}^{(12)}) \ , \ j = 1, \ldots, J \ . \tag{3.4}$$

Sufficiency. Let $\mu_{sk} = \mu_{tk}$, $k = 1, \ldots, K$, or equivalently, $\lambda_{sj}^{(12)} = \lambda_{tj}^{(12)}$, $j = 1, \ldots, J$. Then, (3.4) becomes $F_{sj} - F_{tj} = \lambda_s^{(1)} - \lambda_t^{(1)}$, $j = 1, \ldots, J$. Assume that $\lambda_s^{(1)} - \lambda_t^{(1)}$ is positive (resp., negative). Since F is a strictly increasing function, it holds $\pi_{sj}/\pi_{s\cdot} >$ (resp., $<$) $\pi_{tj}/\pi_{t\cdot}$, $j = 1, \ldots, J$. The last equality leads to a contradiction since both sides add up to 1. Hence, $\lambda_s^{(1)} - \lambda_t^{(1)} = 0$ and $F_{sj} = F_{tj}$ for all j, i.e. rows s and t are homogeneous.

Necessity. Multiplying both sides of (3.4) by w_{2j} and adding over j yields

$$\lambda_s^{(1)} - \lambda_t^{(1)} = \frac{\sum_{j=1}^{J} w_{2j}(F_{sj} - F_{tj})}{\sum_{j=1}^{J} w_{2j}} \ . \tag{3.5}$$

Using (3.3), (3.4) and (3.5) it follows that $\lambda_{sj}^{(12)} = \lambda_{tj}^{(12)}$ for all j and thus $\mu_{sk} = \mu_{tk}$, $k = 1, \ldots, K$. □

Remark 3.1. Notice that in Theorem 3.2, the order K of the generalized association model is taken at most $\min(I-2, J-1)$ rather than $\min(I-1, J-1)$ which is its usual upper bound. This will be clarified after the presentation of Theorem 3.3 below.

Remark 3.2. From the proof of Theorem 3.2 arises that homogeneity of two rows is equivalent to equality of the corresponding interaction *and* main effect parameters in the generalized linear model (2.7). Moreover, the weighted euclidean distance

$$r_{st} = \sum_{j=1}^{J} w_{2j}(F_{sj} - F_{tj})^2 \tag{3.6}$$

can measure the inhomogeneity of rows s and t, with a value of zero indicating homogeneity. For the special case of canonical correlation model, r_{st} becomes the chi–squared distance between s-th and t-th row profiles (Benzécri, 1973).

The formulation of Theorem 3.2 for two homogeneous columns is obvious as well as its extension to the general case of multiple collapses of sets of homogeneous rows or/and columns. An interesting issue is that when performing collapses over homogeneous categories, the structure of the association of the reduced table remains unchanged, as the following theorem states. For simplicity, this theorem is also expressed for the case of collapsing two homogeneous rows while the more general result is provided by Corollary 3.4 below.

Theorem 3.3. *Let* $\mathbf{\Pi} = (\pi_{ij})$ *be an* $I \times J$ *contingency table with structure* $\mathrm{RC}[f](K)$, $K \leqslant \min(I-2, J-1)$, *having homogeneous rows* s *and* t. *Let also* $\tilde{\mathbf{\Pi}} = (\tilde{\pi}_{ij})$ *be the* $(I-1) \times J$ *table obtained by collapsing these homogeneous rows. Then, the structure of* $\tilde{\mathbf{\Pi}}$ *is the same as that of* $\mathbf{\Pi}$, *i.e.* $\mathrm{RC}[f](K)$.

Proof. Without loss of generality consider $s < t$ and place the sum of the homogeneous rows s and t at row s. Then,

$$\tilde{\pi}_{ij} = \begin{cases} \pi_{ij} & , \quad i < s, \ s < i < t \ , \\ \pi_{sj} + \pi_{tj} & , \quad i = s \ , \\ \pi_{i+1,j} & , \quad i \geqslant t \ , \end{cases} \tag{3.7}$$

and since $\tilde{\pi}_{sj}/\tilde{\pi}_{s\cdot} = \pi_{sj}/\pi_{s\cdot}$, $\tilde{\pi}_{\cdot j} = \pi_{\cdot j}$, one has

$$\tilde{F}_{ij} = F(\tilde{\pi}_{sj}/\tilde{\pi}_{s\cdot}\tilde{\pi}_{\cdot j}) = \begin{cases} F_{ij} & , \quad i < t \ , \\ F_{i+1,j} & , \quad i \geqslant t \ . \end{cases} \tag{3.8}$$

Let the generalized linear model expression for $\tilde{\mathbf{F}} = (\tilde{F}_{ij})$ be

$$\tilde{F}_{ij} = \tilde{\lambda} + \tilde{\lambda}_i^{(1)} + \tilde{\lambda}_j^{(2)} + \tilde{\lambda}_{ij}^{(12)} \ , \ i = 1, \ldots, I-1, \ j = 1, \ldots, J \ , \tag{3.9}$$

with $\sum_{i=1}^{I-1} \tilde{w}_{1i}\tilde{\lambda}_i^{(1)} = \sum_{j=1}^{J} \tilde{w}_{2j}\tilde{\lambda}_j^{(2)} = \sum_{i=1}^{I-1} \tilde{w}_{1i}\tilde{\lambda}_{ij}^{(12)} = \sum_{j=1}^{J} \tilde{w}_{2j}\tilde{\lambda}_{ij}^{(12)} = 0$, where $\tilde{w}_{1i}$'s are related to w_{1i}'s of the initial model (2.7) by the analog of (3.7), while $\tilde{w}_{2j} = w_{2j}$, $j = 1, \ldots, J$. Since RC[f](K) is the underlying model for $\mathbf{\Pi}$, then rank($\mathbf{\Lambda}$) = K, where $\mathbf{\Lambda} = (\lambda_{ij}^{(12)})$ is the corresponding interaction parameters matrix in (2.7). Due to the homogeneity of rows s and t, it holds by Theorem 3.2 that $\mathbf{\Lambda}$ has its s-th and t-th rows equal. It can be seen that the matrix $\tilde{\mathbf{\Lambda}} = (\tilde{\lambda}_{ij}^{(12)})$ arises from $\mathbf{\Lambda}$ by deleting its t-th row. Obviously, rank($\tilde{\mathbf{\Lambda}}$) = rank($\mathbf{\Lambda}$) = K and thus the underlying model for $\tilde{\mathbf{\Pi}}$ is also RC[f](K). By Lemma 2.1 the choice of weights does not affect the order of the model. □

Remark 3.3. It is clear now the demand of $K \leqslant \min(I-2, J-1)$ in Theorems 3.2 and 3.3. Since the order of the generalized association model remains the same for the collapsed table $\tilde{\mathbf{\Pi}}$, K has to be consistent also with its size. As a consequence, if the initial's table structure is RC[f](M), i.e. saturated, then it is not possible to exist any homogeneities in the smallest dimension. In the special case of a square contingency table with saturated structure, there are not any homogeneities at all.

Remark 3.4. Defining weighted euclidean column distances in analogy to r_{st} in (3.6), it can be seen that when collapsing homogeneous rows these column distances do not change. This generalizes the principle of distributional equivalence of Benzécri (1973).

Corollary 3.4. *Let $\tilde{\mathbf{\Pi}} = (\tilde{\pi}_{ij})$ be the $\tilde{I}\times\tilde{J}$ table obtained by collapsing all homogeneous rows and columns of $\mathbf{\Pi}$ ($\tilde{I} \leqslant I$, $\tilde{J} \leqslant J$). Then, the collapsed table $\tilde{\mathbf{\Pi}}$ will have the same structure,* RC[f](K), $K \leqslant \min(\tilde{I}, \tilde{J}) - 1$, *as the initial table $\mathbf{\Pi}$.*

The parameters of the RC[f](K) model for the collapsed table $\tilde{\mathbf{\Pi}}$ are connected to the parameters of the corresponding model for the initial table $\mathbf{\Pi}$ as stated below.

Corollary 3.5. *Let $A_1, \ldots, A_{\tilde{I}}$ (resp., $B_1, \ldots, B_{\tilde{J}}$) be the partition of $A = \{1, \ldots, I\}$ (resp., $B = \{1, \ldots, J\}$) formed by homogeneous rows (resp., columns) of $\mathbf{\Pi}$ with structure* RC[f](K), $K \leqslant \min(\tilde{I}, \tilde{J}) - 1$. *Then, the cell probabilities of $\tilde{\mathbf{\Pi}}$ are given by*

$$\tilde{\pi}_{qr} = \tilde{\pi}_{q\cdot}\tilde{\pi}_{\cdot r} F^{-1}\left(\lambda + \tilde{\lambda}_q^{(1)} + \tilde{\lambda}_r^{(2)} + \sum_{k=1}^{K} \phi_k \tilde{\mu}_{qk}\tilde{\nu}_{rk}\right) \ , \quad q = 1, \ldots, \tilde{I} \ , \ r = 1, \ldots, \tilde{J} \ , \tag{3.10}$$

with

$$\tilde{\lambda}_q^{(1)} = \lambda_i^{(1)} \ , \quad i \in A_q \ , \quad q = 1, \ldots, \tilde{I} \ , \tag{3.11}$$

$$\tilde{\lambda}_r^{(2)} = \lambda_j^{(2)} \ , \quad j \in B_r \ , \quad r = 1, \ldots, \tilde{J} \ , \tag{3.12}$$

$$\tilde{\mu}_{qk} = \mu_{ik} \ , \quad i \in A_q \ , \quad q = 1, \ldots, \tilde{I} \ , \quad k = 1, \ldots, K \ , \tag{3.13}$$

$$\tilde{\nu}_{rk} = \nu_{jk} \ , \quad j \in B_r \ , \quad r = 1, \ldots, \tilde{J} \ , \quad k = 1, \ldots, K \ , \tag{3.14}$$

provided that the weights for the collapsed table are

$$\tilde{w}_{1q} = \sum_{i \in A_q} w_{1i} \ , \quad q = 1, \ldots, \tilde{I} \ , \qquad \tilde{w}_{2r} = \sum_{j \in B_r} w_{2j} \ , \quad r = 1, \ldots, \tilde{J} \ . \tag{3.15}$$

Proof. Let $\mathbf{F} = (F_{ij})$, $\tilde{\mathbf{F}} = (\tilde{F}_{ij})$, where $F_{ij} = F(\pi_{ij}/\pi_{i\cdot}\pi_{\cdot j})$, $\tilde{F}_{ij} = F(\tilde{\pi}_{ij}/\tilde{\pi}_{i\cdot}\tilde{\pi}_{\cdot j})$. Under (3.15), the interaction parameters matrix $\tilde{\mathbf{\Lambda}}$ for $\tilde{\mathbf{F}}$ arises from the corresponding matrix $\mathbf{\Lambda}$ for $\mathbf{F}$ by deleting appropriate rows and columns (see also the proof of Theorem 3.3). Let $\tilde{\mathbf{M}} = (\tilde{\mu}_{qk})$ and $\tilde{\mathbf{N}} = (\tilde{\nu}_{rk})$, where $\tilde{\mu}_{qk}$ and $\tilde{\nu}_{rk}$ are as in (3.13) and (3.14). It can be seen that $\tilde{\mathbf{\Lambda}} = \tilde{\mathbf{M}}\mathbf{\Phi}\tilde{\mathbf{N}}'$ and, due to (3.15), $\tilde{\mathbf{M}}'\tilde{\mathbf{W}}_1\tilde{\mathbf{M}} = \tilde{\mathbf{N}}'\tilde{\mathbf{W}}_2\tilde{\mathbf{N}} = \mathbf{I}_K$, where $\tilde{\mathbf{W}}_1 = \mathrm{diag}(\tilde{w}_{11}, \ldots, \tilde{w}_{1\tilde{I}})$ and $\tilde{\mathbf{W}}_2 = \mathrm{diag}(\tilde{w}_{21}, \ldots, \tilde{w}_{2\tilde{J}})$. Thus, $\tilde{\mathbf{M}}$ and $\tilde{\mathbf{N}}$ are the matrices containing the left and right singular vectors of $\tilde{\mathbf{\Lambda}}$ orthonormalized with respect to $\tilde{\mathbf{W}}_1$ and $\tilde{\mathbf{W}}_2$. The uniqueness of the GSVD ensures that the $\tilde{\mu}_{qk}$'s and $\tilde{\nu}_{rk}$'s are the row and column scores of the RC$[f](K)$ model for $\tilde{\mathbf{\Pi}}$. Moreover, relations (3.11) and (3.12) are justified by Remark 3.2. □

Condition (3.15) is satisfied by the marginal weights. Hence, if marginal weights are used, the RC$[f](K)$ model's parameters for table $\tilde{\mathbf{\Pi}}$ are immediate provided from the corresponding parameters for $\mathbf{\Pi}$ by (3.11) – (3.14). On the other hand, (3.15) is not satisfied by the uniform weights. Thus, marginal weights are preferable over the uniform ones since they preserve the invariance of the parameters under collapsing.

4 An illustrative example

Although, as mentioned in the Introduction, the detection of homogeneous categories may be done by testing independence in corresponding subtables, Theorem 3.2 allows for an alternative approach. Since homogeneity of, say, rows s and t is equivalent to the equality of the corresponding row scores, one may consider the hypothesis

$$H_0^{s,t} : \mu_{sk} = \mu_{tk} \ , \ k = 1, \ldots, K \ ,$$

a significance test of which can be based on any asymptotically normally distributed estimators of μ's. In the literature, there are algorithms calculating maximum likelihood estimators for the parameters of correlation and association models as well as their covariance matrix (see Gilula and Haberman, 1986 and Haberman, 1995) as well as results about generalized least squared estimation in power models (see Baccini et al., 2000). Having obtained the estimators one can compute appropriate Mahalanobis distances and checking their significance by comparing with an upper quantile of the chi–square distribution with $K - 1$ degrees of freedom. A non–significant distance indicates homogeneity of the corresponding categories. Note that so far in the related

literature collapsing of categories based on the equality of the corresponding row or column scores was illustrated only for the case $K = 1$ and equality was decided upon simple observation of closeness without performing any test of significance.

We materialized the above described procedure for the classical association model RC(K), which is the most well–known member of the class of generalized association models, using Haberman's (1995) algorithm. A drawback of this algorithm is that there is no option of selecting weights being restricted to the case of uniform ones. We modified it appropriately, in order to control the use of weights.

The procedure is applied on the data provided in Table 1 which origins from Wermuth and Cox (1998) and classifies adults in Germany according to age and type of their education. Model RC(2) is appropriate for the table (see Table 2) and applying the modified Haberman's algorithm we confirm that only the last two columns (4 and 5) are homogeneous and we collapse them. As expected (by Theorem 3.3), the RC(2) model is the most parsimonious describing the data also for the collapsed table (see Table 2). For the RC(2) model the scores' estimates for the initial table as well as its collapsed version are provided in Table 3 for the cases of uniform and marginal weights.

Wermuth and Cox (1998) suggested the grouping of columns 4 and 5 as well, but they reduced further the table by collapsing rows 1 and 2. This is a decision based on the acceptable fit (p–value $= 0.078$) of the independence model of the 2×4 subtable formed by the first two rows and after collapsing the last two columns. It is also consistent with the natural motivation to group the first row, which has relatively small frequencies. However, the corresponding Mahalanobis distances using either uniform or marginal weights are both significant indicating that these rows are not homogeneous. This can be observed clearer in Figure 1, based on the RC(2) model, where the row scores of the first two rows are close for the first axis but apart for the second. Notice also in Table 3, that when using marginal weights, the scores' change for the first reduced table is negligible with respect to the initial table, whereas the corresponding change for the second one is more substantial, especially for the second axis. This is in agreement with Corollary 3.5.

5 Discussion

Summarizing, our major point to make is that there exist no contradiction between homogeneity and structural criteria. Whenever we collapse homogeneous categories, the underlying association structure is not affected. Nevertheless, if in practice happens after collapsing categories for which we have the indication that they are homogeneous, a simpler model to be appropriate for the reduced table, we have to be

cautious. The assumption of either the homogeneity or the association structure's order for one of the two tables is false. It can not be the case that all these assumptions are correct but not in agreement.

In the association models framework, there are special association structures, which assume the row or/and column scores in (2.1) as known. In particular, for $K = 1$, if the column (resp., row) scores are pre–specified (equidistant) we are led to the Row (resp., Column) effect model, denoted by R (resp., C). If further all scores are considered as known, the Uniform (U) association model is achieved. Extensions of this type of models for $K > 1$ have been considered by Goodman (1981a) and Kateri et al. (1998). We would like to emphasize that Theorem 3.3 refers only to RC–type models and can not be applied to the special association models, mentioned above. For example, the U model, by its definition, can never express the structure of an initial table in presence of homogeneities, i.e. some of the row or/and column scores being equal. However it can be the case the underlying association structure to be of the U–type, the U model to be consistent with the collapsed table but not with the initial one due to the homogeneity noise. Analogous observations can also be done for other models of this type. So far this has been faced as a structural contradiction and collapsings have been rejected (Goodman, 1981b).

It is important to highlight that the existence of homogeneities among classification categories is transferred to equalities of the corresponding scores and vice versa, for any choice of the link function F of the RC$[f](K)$ model, with K not necessarily remaining constant for different choices of f. However, in case of independence then $K = 0$ for all possible choices of f.

From a different point of view, it is sometimes preferable not to combine indistinguishable categories but only appreciate their similarity (Anderson, 1984 and Goodman 1985, 1986). For example, in the association models framework, when the scores' ordering of an ordinal classification variable is violated or two scores are close, Goodman equates them but does not collapse the corresponding categories (see also Gilula and Haberman 1986, 1988). Hence, collapsibility is a matter of policy. When equating the scores of the homogeneous categories without collapsing them, we are led to a model of better performance, since the fit remains approximately the same while the degrees of freedom are augmented (less parameters due to scores' equality). On the other hand when we collapse the categories, the performance of the model seems worse since the fit remains the same while the degrees of freedom become less due to the reduction of the table.

In the case of commensurable classification variables (usually occurring in the framework of panel data) the grouping of categories has to be applied simultaneously to rows and columns (Goodman, 1981b). It is straightforward to adapt Theorems

3.2 and 3.3 for this special case. In this framework, it is sometimes meaningful to impose the additional constraint $\mu_{ik} = \nu_{ik}$, $i = 1, \ldots, I$, $k = 1, \ldots, K \leqslant M$, i.e. assume symmetric interaction. When $K = M$, it will be equivalent to the generalized model of Quasi Symmetry (QS[f]), based on f–divergence, introduced by Kateri and Papaioannou (1997). For $K < M$, it will be a special case of the QS[f] model. In the panel data framework it is often the case that large frequencies occur on the main and secondary diagonals and the table is sparse at the corners. Also in some cases, from the nature of the data, there exists no diagonal (athletic data: games results, for example). These tables need special care and research could be developed towards these directions.

Acknowledgement

The authors wish to thank the anonymous referee for his suggestions, especially for calling their attention to Benzécri's principle of distributional equivalence.

References

Agresti, A., Chuang, C. and Kezouh, A. (1987). Order–restricted score parameters in association models for contingency tables. *J. Amer. Statist. Association*, **82**, 619–623.

Anderson, J.A. (1984). Regression and ordered categorical variables. *J. R. Statist. Soc.* B, **46**, 1, 1–30.

Baccini, A., Caussinus, H. and de Falguerolles, A. (1993). Analyzing dependence in large contingency tables: Dimensionality and patterns in scatter plots. In: *Multivariate Analysis: Future Directions* (Cuadras, C. M. and Rao, C. R. Eds.), **2**, 245–263. North Holland, Amsterdam.

Baccini, A., Fekri, M. and Fine, J. (2000). Generalized least squares estimation in contingency tables analysis: Asymptotic properties and applications. *Statistics*, **34**, 267–300.

Becker, M.P. and Clogg, C.C. (1989). Analysis of sets of two–way contingency tables using association models. *J. Amer. Statist. Association*, **84**, 142–151.

Beh, E.J. (1997). Simple correspondence analysis of ordinal cross–classifications using orthogonal polynomials. *Biom. J.*, **39**, 589–613.

Beh, E.J. (1998). A comparative study of scores for correspondence analysis with ordered categories. *Biom. J.*, **40**, 413–429.

Benzécri, J.P. (1973). *L' analyse des données, vol. 2 (L' analyse des correspondance).* Dunod, Paris.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *J. Roy. Statist. Soc.* B, **26**, 211–252.

Gilula, Z. (1986). Grouping and association in contingency tables: An exploratory canonical correlation approach. *J. Amer. Statist. Assoc.*, **81**, 773–779.

Gilula, Z. and Haberman, S.J. (1986). Canonical analysis of contingency tables by maximum likelihood. *J. Amer. Statist. Assoc.*, **81**, 780–788.

Gilula, Z. and Haberman, S.J. (1988). The analysis of multivariate contingency tables by restricted canonical and restricted association models. *J. Amer. Statist. Assoc.*, **83**, 760–771.

Gilula, Z. and Krieger, A.M. (1989). Collapsed two–way contingency tables and the chi–square reduction principle. *J. Roy. Statist. Soc.* B, **51**, 425–433.

Gilula, Z., Krieger, A.M. and Ritov, Y. (1988). Ordinal association in contingency tables: some interpretive aspects. *J. Amer. Statist. Assoc.*, **83**, 540–545.

Goodman, L.A. (1979). Simple models for the analysis of association in cross–classifications having ordered categories. *J. Amer. Statist. Assoc.*, **74**, 537–552.

Goodman, L.A. (1981a). Association models and canonical correlation in the analysis of cross–classifications having ordered categories. *J. Amer. Statist. Assoc.*, **76**, 320–334.

Goodman, L.A. (1981b). Criteria for determining whether certain categories in a cross–classification table should be combined, with special reference to occupational categories in an occupational mobility table. *American Journal of Sociology*, **87**, 3, 612–650.

Goodman, L.A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models and asymmetry models for contingency tables with or without missing entries. *Ann. Statist.*, **13**, 10–69.

Goodman, L.A. (1986). Some useful extensions of the usual correspondence analysis and the usual log-linear models approach in the analysis of contingency tables with or without missing entries. *Int. Stat. Rev.*, **54**, 243–309.

Goodman, L.A. (1996). A single general method for the analysis of cross-classified data: reconciliation and synthesis of some methods of Pearson, Yule, and Fisher, and also some methods of correspondence analysis and association analysis. *J. Amer. Statist. Assoc.*, **91**, 408–428.

Haberman, S.J. (1995). Computation of maximum likelihood estimates in association models. *J. Amer. Statist. Assoc.*, **90**, 1438–1446.

Hirotsu, S. (1983). Defining the pattern of association in two-way contingency tables. *Biometrika*, **70**, 579–589.

Kateri, M. and Papaioannou, T. (1994). f–divergence association models. *Int. J. Math. Stat. Sci.*, **3**, 179–203.

Kateri, M. and Papaioannou, T. (1997). Asymmetry models for contingency tables. *J. Amer. Statist. Assoc.*, **92**, 1124–1131.

Kateri, M., Ahmad, R. and Papaioannou, T. (1998). New features in the class of association models. *Appl. Stochastic Models Data Anal.*, **14**, 125–136.

Read, T.R.C. and Cressie, N.A.C. (1988). *Goodness-of-fit statistics for discrete multivariate data.* Springer-Verlag, New York.

Ritov, Y. and Gilula, Z. (1991). The ordered restricted RC model for ordered contingency tables: Estimation and testing for fit. *Ann. Statist.*, **19**, 2090–2101.

Ritov and Gilula, Z. (1993). Analysis of contingency tables by correspondence models subject to order constraints. *J. Amer. Statist. Assoc.*, **88**, 1380–1387.

Rom, D. and Sarkar, S.K. (1992). A generalized model for the analysis of association in ordinal contingency tables. *J. Statist. Plan. Infer.*, **33**, 205–212.

Weller, S. and Romney, A.K. (1990). *Metric Scaling Correspondence Analysis (Quantitative Applications in the Social Sciences).* Sage University.

Wermuth, N. and Cox, D.R. (1998). On the application of conditional independence to ordinal data. *Int. Stat. Rev.*, **66**, 2, 181–199.

Williams, E.J. (1952). Use of scores for the analysis of association in contingency tables. *Biometrika*, **39**, 274–289.

Yates, F. (1948). The analysis of contingency tables with groupings based on quantitative characters. *Biometrika*, **35**, 176–183.

	Age Group				
Type of Schooling	18-29	30-44	45-59	60-74	>74
basic, incomplete	12	13	12	20	7
basic, complete	215	507	493	460	137
medium	277	300	192	126	38
upper medium	52	91	47	15	6
intensive	233	225	102	74	19

Table 1: Classification of adults according to age and type of their education.

Model	G^2	d.f.	p–value	Change of G^2(I) from the initial table (p-val. for the change)
I	357.146	16	.000	Initial Table
RC	24.275	9	.039	
RC(2)	2.599	4	.627	
I	356.310	12	.000	
RC	23.487	6	.001	Collapsed columns: 4, 5
RC(2)	1.809	2	.405	.835 (.934)
I	349.487	9	.000	
RC	16.677	4	.002	Collapsed rows: 1, 2
RC(2)	1.800	1	.178	6.823 (.078)

Table 2: Models' fit for Table 1, the table with collapsed columns 4 and 5 and the table with additionally collapsed rows 1 and 2.

Uniform weights				**Marginal weights**			
Initial Table				Initial Table			
$\phi_1 = 1.783$		$\phi_2 = .690$		$\phi_1 = .315$		$\phi_2 = .082$	
μ_{i1}-scores	ν_{j1}-scores	μ_{i2}-scores	ν_{j2}-scores	μ_{i1}-scores	ν_{j1}-scores	μ_{i2}-scores	ν_{j2}-scores
-.575	.529	-.431	-.684	-.701	1.522	3.952	1.025
-.484	.428	.605	.400	-.972	.365	-.158	-.941
.168	.073	-.209	.573	.761	-.573	.434	-.898
.517	-.520	.467	-.181	1.192	-1.243	-3.295	1.165
.374	-.511	-.432	-.107	1.293	-1.286	.496	1.014
Collapsed Table (columns 4, 5)				Collapsed Table (columns 4, 5)			
$\phi_1 = 1.462$		$\phi_2 = 0.680$		$\phi_1 = .315$		$\phi_2 = .082$	
μ_{i1}-scores	ν_{j1}-scores	μ_{i2}-scores	ν_{j2}-scores	μ_{i1}-scores	ν_{j1}-scores	μ_{i2}-scores	ν_{j2}-scores
-.549	.506	-.468	-.666	-.698	1.522	3.987	1.027
-.518	.356	.579	.421	-.972	.365	-.159	-.943
.180	-.081	-.197	.541	.761	-.573	.436	-.896
.489	-.781	.492	-.296	1.191	-1.253	-3.284	1.130
.399		-.406		1.293		.489	
Collapsed Table (columns 4, 5 and rows 1, 2)				Collapsed Table (columns 4, 5 and rows 1, 2)			
$\phi_1 = 1.189$		$\phi_2 = .492$		$\phi_1 = .314$		$\phi_2 = .069$	
μ_{i1}-scores	ν_{j1}-scores	μ_{i2}-scores	ν_{j2}-scores	μ_{i1}-scores	ν_{j1}-scores	μ_{i2}-scores	ν_{j2}-scores
-.843	.662	.072	-.509	-.963	1.515	-.044	1.032
-.096	.234	-.374	.495	.760	.371	.532	-.933
.361	-.219	.787	.505	1.198	-.569	-3.868	-.910
.387	-.677	-.485	-.491	1.293	-1.259	.617	1.126

Table 3: Intrinsic association parameters and scores estimates for the RC(2) model fitted on Table 1 and on the reduced tables using uniform and marginal weights.

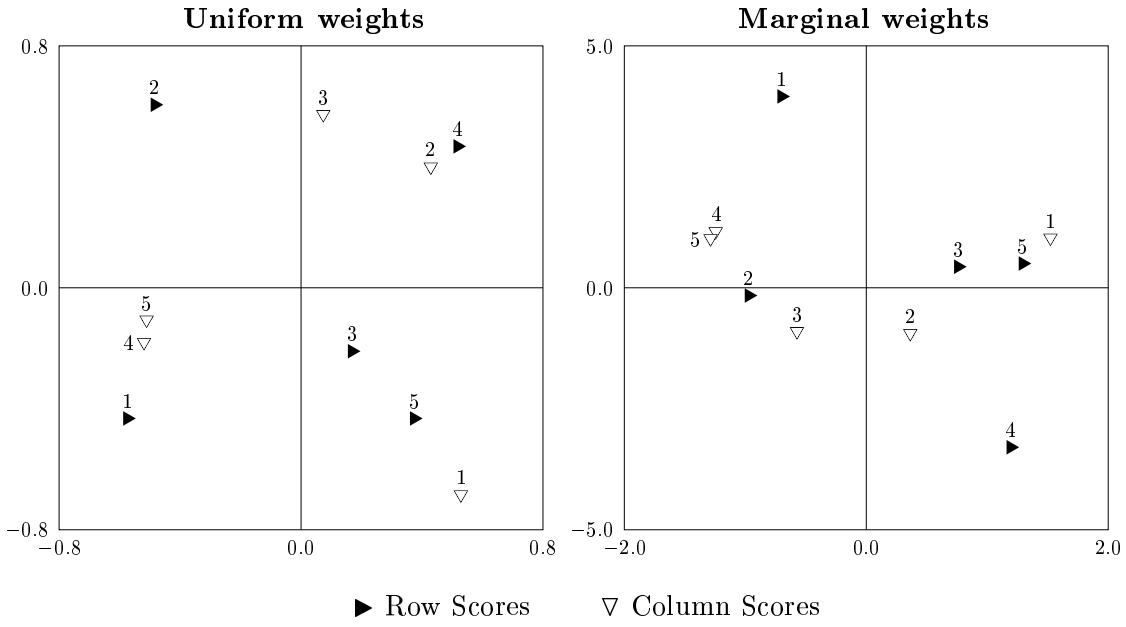

Figure 1: Estimated scores for the RC(2) model fitted on Table 1 using uniform and marginal weights.