Hacettepe Journal of Mathematics and Statistics
Volume 31 (2002), 37–43

HIERARCHY ANALYSIS OF THREE–WAY TABLES

İ. Akhisar* and A. Bener†

Received 21. 03. 2000

Abstract

In this work a hierarchical method is given for analyzing three-way data tables using classical factorial methods.

Key Words: Hierarchical methods, indices of hierarchy, quality of splitting, logic tables, metric classification, inertia moments.

1. Introduction

Classical factorial analysis methods as applied to two-way data tables are insufficient for three-way data tables. Therefore different techniques should be used. The first formal work on this subject was carried out in 1982 by J. P. Benzecri [4]. The main principle of this method is based on transforming three-way comparison tables to logic tables. As it is known, the most important problem in three dimensional data analysis is to express the data in terms of principle components, namely the factors. The problem was solved in 1982 by a formula that completely defines the three dimensional data in terms of principle components, including third order interactions [6]. A sufficient approach can be obtained even if only second order interactions are taken [2]. The expression of this formula, known as the reconstruction formula, in terms of relative frequencies is as follows

$$f_{ijt} = f_i \cdot f_j \cdot f_t \left(1 + \sum\left\{\varphi_\alpha^i \cdot \varphi_\alpha^j \cdot a_\alpha \mid \alpha \in A\right\} + \sum\left\{\varphi_\beta^j \cdot \varphi_\beta^t \cdot b_\beta \mid \beta \in B\right\} + \sum\left\{\varphi_\gamma^i \cdot \varphi_\gamma^t \cdot c_\gamma \mid \gamma \in C\right\} + \sum\left\{\psi_\delta^i \cdot \psi_\delta^j \cdot \psi_\delta^t \cdot d_\delta \mid \delta \in D\right\}\right),$$

where the first three sums are second order and the last sum a third order interaction, besides φ_α, φ_β, φ_γ and ψ_δ are functions with zero mean and unit variance defined on the sets I, J and T respectively [1].

We cannot examine three-way data tables directly using classical factorial analysis methods or obtain a good classification of these tables by applying hierarchical methods directly. The grouping between individuals as well as variables in plain

*Istanbul Technical University, Department of Mathematics, Istanbul, Turkey
†Yıldız Technical University, Department of Statistics, Istanbul, Turkey

representations obtained by an application of suitable methods will usually be different from the grouping obtained by classical methods. Therefore the hierarchical classifications available from such tables should be appropriate to these groupings [3].

As is well known hierarchical methods, usually factorial analysis, can either be applied to the distances between individuals obtained from correspondence tables, or directly to logic tables. So a good classification can be obtained by the application of known hierarchical methods to logic tables obtained from such tables.

On the other hand, since the expression (reconstruction formula) connecting three dimensional data to principle components is known, a classification can be obtained by applying known hierarchical methods to principle components instead of distance tables.

One may ask which of these two methods should be used to classify three-way tables. In the first case, the classification can be done using hierarchical methods after transforming to a logic table, without using factorial analysis. However, this method may be expensive since logic tables are very high dimensional. In the second case, factorial analysis methods must be used to obtain the principle components [5].

In fact, for the selection of the method, a sufficient criteria will be to consider the purpose of the study. To apply a method a metric must be chosen. This will be either the Euclid or the χ^2 metric.

2. Metric Classification

In hierarchy analysis, the data is usually given as a two-way table, as in the case of factorial analysis techniques:

$$x_i^j,\ i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, p.$$

Such a table can be considered as a system of values taken on the set of individuals using p variables. If each individual x_i of the set is assigned the mass $m_i > 0$, then the total mass becomes $m = \sum\{m_i \mid i \in [1, n]\}$. In addition, defining the relative mass for each individual as $p_i = m_i/m$, the total mass becomes $\sum p_i = 1$. For each individual, the corresponding point in $\mathbb{R}^p$ is given as follows:

$$\forall\, i \in [1, n],\ x_i = [x_i^1, x_i^2, \ldots, x_i^p] \in \mathbb{R}^p.$$

Define the cloud N of individuals by $N = \{(x_i, p_i) \mid i \in [1, n],\ p_i > 0\}$. Assume that a symmetric and positive definite bilinear form M and scalar product with respect to M is defined by the following.

$$\forall\, i, j,\ \langle x_i, x_j \rangle_M = M(x_i, x_j) = x_i' M x_j = \parallel x_i - x_j \parallel_M^2 .$$

The center of mass (or gravity) of the cloud, denoted by g_N (or briefly by g) is given by $g = \sum p_i x_i$, where

$$g = [g_1, g_2, \ldots, g_p] \in \mathbb{R}^p,\ g_j = \sum p_i x_i^j.$$

The total mass, total gravity and the center of gravity of a part a of the cloud are denoted by m_a, p_a and g_a, respectively, where:

$$\begin{aligned} m_a &= \sum\Big\{m_i \mid x_i \in a\Big\},\ g_a = \sum\left\{\frac{m_i}{m_a}x_i \mid x_i \in a\right\} \\ p_a &= \frac{m_a}{m} = \sum\Big\{p_i \mid x_i \in a\Big\} \quad \text{and} \quad g_a = \frac{1}{p_a}\sum\Big\{p_i x_i \mid x_i \in a\Big\}. \end{aligned}$$

3. Moments of Inertia

3.1. Definition. The second central moments, variance and inertia according to any point y in the space $\mathbb{R}^p$ of a cloud N with center of gravity g, are denoted by $M^2(N)$, $V(N)$ and $I_y(N)$, and defined by

$$\begin{aligned} M^2(N) &= \sum\Big\{m_i\|x_i - g\|_M^2 \mid x_i \in N\Big\}, \\ V(N) &= \sum\Big\{p_i\|x_i - g\|_M^2 \mid x_i \in N\Big\}, \\ I_y(N) &= \sum\Big\{p_i\|x_i - y\|_M^2 \mid x_i \in N\Big\}, \end{aligned}$$

respectively. It is obvious that the moment of the cloud according to its center of gravity is equal to its variance, the value of $I_g = V(N)$ expresses the prevalence around y of the cloud N.

3.2. Definition. The second moment, variance and inertia with respect to any point y in $\mathbb{R}^p$ of a part a in $P(N)$, with center of gravity at g_a, are given by

$$\begin{aligned} M^2(a) &= \sum\Big\{m_i\|x_i - g\|_M^2 \mid x_i \in a\Big\}, \\ V(a) &= \sum\Big\{\frac{m_i}{m}\|x_i - g\|_M^2 \mid x_i \in a\Big\}, \\ I_y(a) &= \frac{1}{p_a}\sum\Big\{p_i\|x_i - g\|_M^2 \mid x_i \in a\Big\}, \end{aligned}$$

respectively.

It is clear that one may obtain the relations

$$M^2(a) = m_a I_g(a) = mV(a)$$

among $M^2(a)$, $V(a)$ and $I_g(a)$.

Now consider a partition Q of N, that is $N = \bigcup\{a \mid a \in Q\}$ where $a, a' \in Q$, $a \neq a' \Longrightarrow a \cap a' = \emptyset$. For each class a define the inertia of the class to be its moment of inertia $I_{g_a}(a)$ with respect to the center of gravity g_a. That is,

$$I_{g_a}(a) = \frac{1}{p_a}\sum\{p_i\|x_i - g\|_M^2 \mid x_i \in a\}.$$

Let G denote the set of centers of gravity of the classes, i.e. $G = \{g_a \mid a \in Q\}$. If a gravity p_a is assigned to each g_a then the moment of inertia of the set G with

respect to the center of gravity g of the cloud is said to be the inertia within the class and is given by

$$I_g(G) = \sum \Big\{ p_a \|g_a - g\|_M^2 \mid a \in Q \Big\}$$

4. The Method of Metric Classification

We need to give criteria measuring the proximity among parts for use in algorithms giving an application of methods related to metric classification. Here we will give two such criteria. Given a metric in the space $\mathbb{R}^p$ including the cloud N, a distance function may be defined as

$$\forall\, x_i, x_j \in \mathbb{R}^p,\ d(x_i, x_j) = d_{ij} = [M(x_i - x_j, x_i - x_j)]^{1/2} = \ \|x_i - x_j\|_M.$$

The main purpose of a metric classification is to separate a cloud N consisting of n individuals into k classes so that individuals in the same class have similar properties, while it is possible to distinguish among classes. Such a grouping is only possible by giving a criterion that measures the proximity of individuals belonging to the same class, so giving the quality of splitting. In fact, if such a criterion is given it is possible to choose the best splitting by examining all possible splittings. But to carry out this takes too much time. As an example, the number of 4-class splittings of a class consisting of 14 individuals, is more than 10 million.

Because of this, instead of trying to obtain the best solution, we have to set up algorithms that give approximate solutions.

1) Each class has to have maximum homogeneity in terms of the quality p that the individuals have,

2) The classes have to be as different from each other as possible.

Achieving this end depends on the difference between the classes as well as on the best measurement of homogeneity in the same class. Hence it is necessary to define a suitable proximity criterion in the cloud N, and we try to set up algorithms which ensure that the metric classification has the two proximity criteria given above. Efficiency of the algorithms depends on the transition formula, which give proximity criteria while passing from the hierarchy P_{k-1} to the hierarchy P_k [7]. Given classes a, b and c, the main step is to relate $\|g_a - g_{b\cup c}\|$ with $\|g_a - g_b\|$, $\|g_a - g_c\|$ and $\|g_c - g_b\|$. Clearly

$$p_b g_b + p_c g_c = (p_b + p_c) g_{b\cup c}$$

and so

$$(p_b + p_c)(g_a - g_{b\cup c}) = p_b(g_a - g_b) + p_c(g_a - g_c).$$

Hence we have

$$\begin{aligned}(p_b + p_c)^2 \|g_a - g_{b\cup c}\|^2 \ = \ & p_b^2 \|g_a - g_b\|^2 + p_c^2 \|g_a - g_c\|^2 \\ & + 2 p_b p_c M(g_a - g_b, g_a - g - c).\end{aligned}$$

On the other hand, since $g_c - g_b = (g_a - g_b) - (g_a - g_c)$, we have

$$\|g_c - g_b\|^2 = \|g_a - g_b\|^2 - 2M(g_a - g_b, g_a - g_c) + \|g_a - g_c\|^2$$

and so

$$\begin{aligned}(p_b + p_c)^2\|g_a - g_{b\cup c}\|^2 &= p_b(p_b + p_c)\|g_a - g_b\|^2 \\ &\quad + p_c(p_b + p_c)\|g_a - g_c\|^2 - p_b p_c\|g_c - g_b\|^2.\end{aligned}$$

Finally we obtain

$$\begin{aligned}(p_b + p_c)\|g_a - g_{b\cup c}\|^2 &= p_b\|g_a - g_b\|^2 + p_c\|g_a - g_c\|^2 \\ &\quad - \frac{p_b p_c}{p_b + p_c}\|g_b - g_c\|^2, \qquad (1)\end{aligned}$$

as required.

5. Indexing the Hierarchy

To index a hierarchy i.e. to compute the values of the indexing function Δ, we will use an appropriate distance δ between the parts.

One advantage of indexing a hierarchy is that the proximity of splittings on the classification tree can be measured in terms of the proximity index between previously given parts [8].

If for $c = a \cup b$ we define $\Delta(c) = \max\{\Delta(a), \Delta(b), \delta(a, b)\}$ then we obtain a hierarchy index consistent with these criteria.

6. An Algorithm related to the Inertia

The prevalence of a class is measured using the inertia of that class. The smaller the inertia of a class the higher the homogeneity. Now consider the following formula that relates the inertia between classes to the inertia within classes.

$$I_g(N) = I_g(G) + \sum\Big\{Ig_a(a) \mid a \in Q\Big\}$$

Since the total inertia $I_g(N)$ is constant, to minimize the total inertia within classes is equivalent to maximizing the inertia between classes. So, if we take inertia within a class as a measure of homogeneity and inertia between classes as a measure of difference, investigating different classes will be equivalent to investigating homogeneous classes.

This criterion will choose the best splitting having a fixed number of classes in the cloud N, but is not useful for comparing splittings having different numbers of classes because it favours small classes.

In fact, the total within-class inertia of the best k-class splitting is always greater than the total within-class inertia of the best $k + 1$ class splitting. So the $k+1$ class splitting will be better. Therefore, the best possible splitting according to this view will occur when each class has one element. In this case, we will have $\sum\{Ig_a(a) \mid a \in Q\} = 0$ since each class coincides with its center of gravity.

We therefore see that, while investigating the best splitting of the set, we must examine how to determine the number k of classes. We can express the quality of a splitting using between-class and within-class inertia. A good splitting occurs when between-class inertia is large and, consequently, within-class inertia is small. When one passes from a $k+1$ class splitting to a k class splitting by combining two classes, between-class inertia will decrease. Given that this is so (see Theorem 6.1 below), our criterion should therefore be to group two classes for which the loss of between-class inertia will be least. This means that if the distance between two classes is chosen to represent the loss of inertia after they are combined, then we should combine classes which are close together. In other words, a pair of classes for which the loss of inertia will be least will be the closest together, and should therefore be combined. The following theorem is important in this respect:

6.1. Theorem: *Let Q_k be a k-class splitting of the individuals of the cloud N, $N = \bigcup\{a \mid a \in Q_k\}$, where $a \cap a' = \emptyset$ for $a \neq a'$. The within-class inertia $I_g(G_k) = \sum\{p_a \|g_a - g\|_M^2 \mid a \in Q_k\}$ corresponding to this splitting is a non-increasing function of k. Here p_a, g_a and G_k are as before*

Now let a and b belong to Q_k, and define the $k-1$ class splitting Q_{k-1} by combining the classes a and b. According to Theorem 6.1 the number $I_g(G_k) - I_g(G_{k-1})$ is non-negative, and as explained above it is appropriate to take this as a measure of the distance between the classes a and b. Hence we define

$$\delta(a,b) = I_g(G_k) - I_g(G_{k-1})$$

and call this the *distance between the classes a and b*. Using the equality (1) it may be proved that

$$\delta(a,b) = \frac{p_a p_b}{p_a + p_b}\|g_a - g_b\|_M^2.$$

Now let us try to give a transition formula. Suppose that the hierarchy H_k is obtained by grouping any two elements b and c of the hierarchy H_{k-1}. Then if a is any part in H_k the distance $\delta(a, b \cup c)$ is given by

$$\begin{aligned}
\delta(a,b\cup c) &= \frac{p_a(p_b+p_c)}{p_a+p_b+p_c}\|g_a - g_{b\cup c}\|^2 \\
&= \frac{p_a(p_b+p_c)}{p_a+p_b+p_c}\left[\frac{p_b}{p_b+p_c}\|g_a-g_b\|^2 + \frac{p_c}{p_b+p_c}\|g_a-g_c\|^2 \right. \\
&\qquad \left. - \frac{p_b p_c}{p_b+p_c}\|g_c-g_b\|^2\right] \\
&= \frac{p_a+p_b}{p_a+p_b+p_c}\delta(a,b) + \frac{p_a+p_c}{p_a+p_b+p_c}\delta(a,c) - \frac{p_a(p_b+p_c)}{p_a+p_b+p_c}\delta(c,b).
\end{aligned}$$

7. The selection of the number of classes and the quality of splitting

Let us consider again the formula that relates between-class inertia to within-class inertia, namely $I_g(N) = I_g(G) + \sum\left\{Ig_a(a) \mid a \in Q\right\}$. This formula may be written

briefly as $I_T = I_1 + I_2$. Hence, we have $I_1/I_T + I_2/I_T = 1$. The within-class inertia I_1 belonging to the splitting can always be computed whatever method is used to obtain a k class hierarchy. So given any k the ratio $\tau = I_2/I_T$, called the *quality of splitting*, can be computed. It is clear that $0 \leq \tau \leq 1$. When $k = 1$ then $\tau = 0$ and when $k = n$ then $t = 1$. We can determine the most convenient value of k by obtaining several solutions for different values of k between the smallest and the largest and examining the corresponding τ values.

References

[1] Bener, A. Etude par L'Analyse des Correspondences des Interactions dans un Tableau Ternarie Applications a des donnees Linguistiques Universite P. et M. Curie (These), 1981.

[2] Bener, A. Decomposition des interaction dans une correspondence multiple chaier de l'analyse de donnes, Vol I, (Donut), 1982.

[3] Benzecri, J. P. Theory and methods of Scaling, John Wiley, New York, 1958.

[4] Benzechri, J. P. Analyse des Donees Correspondances, Donud, 1976.

[5] Leeuw, J. D. and Heijden, P. G. M. Correspondence analysis of incomplete contingency tables, Psychometrika, 27, 223–233, 1988.

[6] Little, R. J. A. and Beale, E. M. L. Missing values in multivariate analysis, Jour. of Royal Statis. Soc. 41, B37, 129–146, 1979.

[7] Little, R. J. A. Maximum likelihood inference for multiple regression, Journal of Royal Statisitcal Society 41, B1, 76–87, 1979.

[8] Wilkinson, J. H. The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, 1965.