An Overview of Statistical Methods for the Analysis of Compositional Data

Jan A. van den Brakel *

July, 1995

Abstract

Compositional data can be regarded as vectors representing the distribution of proportions over D categories of some unit. Because elements of a compositions denote proportions, they are non-negative and they add up to one. Therefore compositions can be regarded as vectors in the simplex, a constrained sub-space of the real space.

Aitchison (1986) showed that there are very specific problems with the statistical analysis of compositions observed on the simplex. In the first place, compositional vectors have a degenerate covariance structure. This leads to several problems in the interpretation of the covariance matrix of a composition. In the second place the assumption of a multivariate normal distribution for compositional vectors is doubtful. Therefore standard multivariate statistical methods can lead to distorted inferences. In this article, four statistical methods, more appropriate for the analysis of compositional data, are reviewed and discussed.

Keywords & phrases: biplots, canonical correspondence analysis, latent budget analysis, logratio analysis, seemingly unrelated regression equations models, simplex.

*Department of Statistical Methods, Division Research and Development, Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, E-mail: JBRL@CBS.NL.

This research was performed at the Interuniversity Center of the Social Science Theory and Methodology (ICS), Utrecht University. The author wish to thank prof. dr. T.A.B. Snijders and prof. dr. P.G.M. van der Heijden for reading and commenting former drafts of this manuscript.

1 Introduction

Compositional data can be regarded as vectors representing proportions of some unit. Let vector $\mathbf{p_D}$ denote a D part composition: $\mathbf{p_D} = (p_{1|i}, \ldots, p_{j|i}, \ldots, p_{D|i})^t$. Vector $\mathbf{p_D}$ denotes for instance the time allocation of person i over $j = 1, \ldots, D$ activities in proportions (conditional on person i). Other applications are for instance time budgets, financial budgets or the chemical compositions of rocks expressed as proportions. Because the elements of $\mathbf{p_D}$ denote proportions, they are non-negative ($p_{j|i} \geq 0 \; \forall j$) and they add up to one ($\sum_{j=1}^{D} p_{j|i} = 1$), the "unit-sum constraint". Aitchison (1986) emphasized that due to the unit-sum constraint, standard multivariate methods are not applicable for the analysis of compositional data. In the first place, there are problems with the interpretation of the covariances between the elements of the compositional vector $\mathbf{p_D}$ because the covariance matrix of $\mathbf{p_D}$ is degenerate. Aitchison showed that the elements of the correlation matrix of $\mathbf{p_D}$ are subject to restrictions. From $\mathrm{cov}(p_{1|i}, \sum_{j=1}^{D} p_{j|i}) = 0$ it can be deduced that $-\mathrm{var}(p_{1|i}) = \sum_{j=2}^{D} \mathrm{cov}(p_{1|i}, p_{j|i})$. Due to this, zero covariance (or zero correlation) between two components cannot be interpreted as absence of association. Other problems with the interpretation of the covariance matrix of $\mathbf{p_D}$ are extensively discussed in Aitchison (1986), Ch.3. In the second place the assumption of a multivariate normal distribution for $\mathbf{p_D}$ is doubtful, because the elements of composition $\mathbf{p_D}$ can only take values between zero and one. Standard multivariate methods like multivariate regression, MANOVA, principal components analysis or factor analysis, based on the covariance matrix of $\mathbf{p_D}$ and on the assumption of multivariate normality of $\mathbf{p_D}$, can lead to distorted inferences (Aitchison, 1986).

In this article, logratio analysis (section 2), seemingly unrelated regression equation models (section 3), latent budget analysis (section 4) and canonical correspondence analysis (section 5) for the analysis of compositional data are reviewed and discussed. In de Leeuw, van der Heijden and Verboon (1990) and in van der Heijden and van den Brakel (1993), latent budget analysis, correspondence analysis and logcontrast principal component analysis are compared and applied on time budgets.

2 Logratio analysis

Aitchison (1986) emphasized that compositions $\mathbf{p_D}$ are, essentially $d = D - 1$ dimensional vectors because the last element of a composition is fixed, due to the unit-sum constraint. He emphasized that compositional data can be regarded as vectors observed on the simplex, defined by:

$$\mathcal{S}^d = \{(p_{1|i}, p_{2|i}, \ldots, p_{D|i}) : p_{1|} \geq 0, \ldots, p_{D|i} \geq 0; p_{1|i} + p_{2|i} + \ldots + p_{D|i} = 1\}.$$

Vectors observed on a constrained space such as the simplex do not have necessarily the same interpretation as vectors that can assume any values in the real space $\mathcal{R}^d$ (Aitchison, 1986). To get rid of these constraints, Aitchison proposed to transform the composition $\mathbf{p_D}$ with a logratio transformation to a vector $\mathbf{y_d}$:

$$\mathbf{y_d} = [y_{j|i}] = [\ln\left(\frac{p_{j|i}}{p_{D|i}}\right) : j = 1, 2, \ldots, d]. \tag{1}$$

The logratio transformation transforms the vector $\mathbf{p_D}$, observed on the constrained space of the simplex $\mathcal{S}^d$ to the space of real numbers $\mathcal{R}^d$. The inverse of the logratio transformation is the logistic transformation:

$$\begin{aligned} p_{j|i} &= \frac{\exp(y_{j|i})}{\exp(y_{1|i}) + \exp(y_{2|i}) + \ldots + \exp(y_{d|i}) + 1} \quad (j = 1, 2, \ldots, d) \\ p_{D|i} &= \frac{1}{\exp(y_{1|i}) + \exp(y_{2|i}) + \ldots + \exp(y_{d|i}) + 1}. \end{aligned}$$

The logratio transformed composition $\mathbf{y_d}$ can assume any value in $\mathcal{R}^d$. If $\mathbf{y_d}$ follows a d-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the composition $\mathbf{p_D}$ is said to follow a logistic normal distribution with the same parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (Aitchison, 1986):

$$\mathbf{y_d} \simeq \mathcal{N}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Leftrightarrow \mathbf{p_D} \simeq \mathcal{L}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

with

$$\boldsymbol{\mu} = [\mu_{j|i}] = [\mathrm{E}(y_{j|i})] = [\mathrm{E}\left\{\ln\left(\frac{p_{j|i}}{p_{D|i}}\right)\right\} : j = 1, 2, \ldots, d]$$

and

$$\boldsymbol{\Sigma} = [\sigma_{jj'}] = [\mathrm{cov}\left\{\ln\left(\frac{p_{j|i}}{p_{D|i}}\right), \ln\left(\frac{p_{j'|i}}{p_{D|i}}\right)\right\} : j, j' = 1, 2, \ldots, d].$$

Through the close relationship of the logistic normal distribution with the normal distribution, all the powerful multivariate analysis methods (like multivariate regression analysis, MANOVA, logcontrast principal component analysis and factor analysis) based on the assumption of normality become applicable for compositional data.

These methods are appropriate if each composition can be regarded as an independent replication. For example time budgets or financial budgets where individual persons specify how they spent their time or income over different categories. The logistic normal distribution, together with the logratio covariance structure ($\boldsymbol{\Sigma}$) define a powerful parametric class of distributions on the simplex, flexible enough to describe or model all possible dependency structures that can occur on the constraint space of the simplex. This class of probability distributions makes it possible to test hypotheses concerning model parameters, in order to investigate different kinds of research questions.

Aitchison (1986, Ch.5, Ch.9 and Ch.10) showed how different forms of statistical independency between the components of a composition can be investigated by testing hypotheses about the structure of $\boldsymbol{\Sigma}$. In order to investigate how compositions depend on other explanatory variables, the logratio transformed composition $\mathbf{y}$ can be modeled with covariables in a multivariate linear regression model. Under the assumption that $\mathbf{y}$ is $\mathcal{N}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distributed, hypotheses concerning the significance of regression coefficients can be tested (Aitchison, 1986, Ch.7).

Under the assumption that $\mathbf{y}$ is $\mathcal{N}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distributed, it is also possible to test the hypothesis that $\boldsymbol{\mu}$ and/or $\boldsymbol{\Sigma}$ for two or more groups are different. If groups are significant different, a classification rule for new observed compositions can be constructed with logistic discriminant analysis (Aitchison, 1986, Ch.7 and Ch.12).

To describe the observed variation in a composition with a smaller number of components, logcontrast principal component analysis can be used as data dimension reduction technique (Aitchison, 186, Ch.8). Only the most important methods proposed by Aitchison are mentioned here. For an extensive overview the reader is referred to Aitchison (1986).

3 Seemingly unrelated regression equations models

The idea to use seemingly unrelated regression equations (SURE) models for the analysis of budget data was raised by Pyndyck and Rubinfeld (1981). The analysis of SURE models was extensively discussed by Srivastava and Giles (1987). In SURE models, each element of the composition $\mathbf{p_D}$ is modeled with the same explanatory variables in a linear regression model. The elements of the composition are the dependent variables. This leads to a system of regression equations:

$$
\begin{aligned}
p_{1|i} &= \alpha_1 + \beta_{11} z_{i1} + \beta_{12} z_{i2} + \ldots + \beta_{1k} z_{ik} + \ldots + \beta_{1K} z_{iK} + \epsilon_{i1} \\
p_{2|i} &= \alpha_2 + \beta_{21} z_{i1} + \beta_{22} z_{i2} + \ldots + \beta_{2k} z_{ik} + \ldots + \beta_{2K} z_{iK} + \epsilon_{i2} \\
&\vdots \\
p_{j|i} &= \alpha_j + \beta_{j1} z_{i1} + \beta_{j2} z_{i2} + \ldots + \beta_{jk} z_{ik} + \ldots + \beta_{jK} z_{iK} + \epsilon_{ij} \\
&\vdots \\
p_{D|i} &= \alpha_D + \beta_{D1} z_{i1} + \beta_{D2} z_{i2} + \ldots + \beta_{Dk} z_{ik} + \ldots + \beta_{DK} z_{iK} + \epsilon_{iD}
\end{aligned}
$$

were z_{ik} denotes the k'th explanatory variable for each category j of the composition $\mathbf{p_D}$; α_j the intercept of the j'th regression equation; β_{jk} the regression coefficient for z_{ik} and ϵ_{ij} the i'th value of the error term of the j'th regression equation. The unit-sum constraint implies the following additional restrictions for the regression parameters:

$$\sum_{j=1}^{D} \alpha_j = 1 \tag{2}$$

$$\sum_{j=1}^{D} \beta_{jk} = 0 \quad (k = 1, 2, \ldots, K) \tag{3}$$

$$\sum_{j=1}^{D} \epsilon_{ij} = 0 \quad (i = 1, 2, \ldots, N). \tag{4}$$

Due to these constraints, the error terms ϵ_{ij} are mutually correlated. Applying ordinary least squares estimation to this system of equations leads to inefficient parameter estimates because the correlation between the error terms of the regression equations are ignored. An efficient estimation procedure, which takes into account the correlation between the equations of the system, was devised by Zellner (1962) and is called Zellner Estimation. Zellner suggests that efficiency in estimation can be gained by application of generalised least squares (GLS) estimation to a group of seemingly unrelated regression equations.

Suppose that N compositions $\mathbf{p_D}$ are observed. For each composition K explanatory variables are observed. The aim is to explain the relationship between the composition and the explanatory variables in a linear regression model. Every category of the composition forms a linear regression equation with the same K explanatory variables. This D regression equations can be expressed compactly in matrix notation:

$$\begin{pmatrix} \mathbf{p^1} \\ \mathbf{p^2} \\ \vdots \\ \mathbf{p^j} \\ \vdots \\ \mathbf{p^D} \end{pmatrix} = \begin{pmatrix} \mathbf{Z} & \mathbf{0} & \ldots & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z} & \ldots & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & \vdots & & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{Z} & \ldots & \mathbf{0} \\ \vdots & \vdots & & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \ldots & \mathbf{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^1 \\ \boldsymbol{\beta}^2 \\ \vdots \\ \boldsymbol{\beta}^{\mathbf{j}} \\ \vdots \\ \boldsymbol{\beta}^{\mathbf{D}} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}^1 \\ \boldsymbol{\epsilon}^2 \\ \vdots \\ \boldsymbol{\epsilon}^{\mathbf{j}} \\ \vdots \\ \boldsymbol{\epsilon}^{\mathbf{D}} \end{pmatrix}, \tag{5}$$

or, equivalently,

$$\mathbf{P} = \mathbf{I_D} \otimes \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{6}$$

where $\mathbf{p^j}$ is the $N \times 1$ vector which contains the N observations of category j from the composition $\mathbf{p_D}$; $\mathbf{Z}$ is the $N \times (K+1)$ matrix with the explanatory variables for category j (the first column of $\mathbf{Z}$ contains ones for the intercept); $\boldsymbol{\beta}^{\mathbf{j}}$ is the $(K+1) \times 1$ vector with the regression parameters for category j. The first element of $\boldsymbol{\beta}^{\mathbf{j}}$ corresponds to the intercept parameter α_j; $\boldsymbol{\epsilon}^{\mathbf{j}}$ is a $N \times 1$ vector with the error terms for the observations of category j; $\mathbf{P}$ is a $(ND \times 1)$ vector consisting of the D vectors $\mathbf{p^j}$; $\mathbf{I_D}$ is the $D \times D$ identity matrix; $\boldsymbol{\beta}$ is a $(K+1)D \times 1$ vector consisting of the D regression coefficient vectors $\boldsymbol{\beta}^{\mathbf{j}}$; $\boldsymbol{\epsilon}$ is a $ND \times 1$ vector consisting of the D error vectors $\boldsymbol{\epsilon}^{\mathbf{j}}$ and $\otimes$ denotes the Kronecker product.

Restriction (4), causes interdependence between the error terms of the D equations: $\text{cov}(\boldsymbol{\epsilon}^{\mathbf{j}}, \boldsymbol{\epsilon}^{\mathbf{j}'}) = \text{E}(\boldsymbol{\epsilon}^{\mathbf{j}}\boldsymbol{\epsilon}^{\mathbf{j}'\mathbf{t}}) = \sigma_{jj'}\mathbf{I_N}$. Estimating (6) as a SURE model, the following assumption is made:

$$\text{E}(\boldsymbol{\epsilon}) = \mathbf{0},$$

$$\text{cov}(\boldsymbol{\epsilon}) = \text{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathbf{t}}) = \begin{pmatrix} \sigma_{11}\mathbf{I_N} & \sigma_{12}\mathbf{I_N} & \cdots & \sigma_{1D}\mathbf{I_N} \\ \sigma_{21}\mathbf{I_N} & \sigma_{22}\mathbf{I_N} & \cdots & \sigma_{2D}\mathbf{I_N} \\ \vdots & \vdots & & \vdots \\ \sigma_{D1}\mathbf{I_N} & \sigma_{D2}\mathbf{I_N} & \cdots & \sigma_{DD}\mathbf{I_N} \end{pmatrix} = \boldsymbol{\Sigma} \otimes \mathbf{I_N} = \boldsymbol{\Psi},$$

where $\boldsymbol{\Sigma}$ is the $D \times D$ matrix with elements $\sigma_{jj'}$. Note that the variance of ϵ_{ij} and the covariance between ϵ_{ij} and $\epsilon_{ij'}$ are assumed to be constant for all i and that the covariance between ϵ_{ij} and $\epsilon_{i'j'}$ are zero for all j and for all j'.

For a given $\boldsymbol{\Sigma}$ the best linear unbiased estimator (BLUE) $\hat{\boldsymbol{\beta}}_{\mathbf{GLS}}$ for the regression coefficients $\boldsymbol{\beta}$ is obtained by applying GLS estimation:

$$\hat{\boldsymbol{\beta}}_{\mathbf{GLS}} = [(\mathbf{I_D} \otimes \mathbf{Z})^{\mathbf{t}}(\boldsymbol{\Sigma} \otimes \mathbf{I_N})^{-1}(\mathbf{I_D} \otimes \mathbf{Z})]^{-1}(\mathbf{I_D} \otimes \mathbf{Z})^{\mathbf{t}}(\boldsymbol{\Sigma} \otimes \mathbf{I_N})^{-1}\mathbf{P}. \tag{7}$$

The variance covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathbf{GLS}}$ can be estimated as

$$\text{V}(\hat{\boldsymbol{\beta}}_{\mathbf{GLS}}) = [(\mathbf{I_D} \otimes \mathbf{Z})^{\mathbf{t}}(\boldsymbol{\Sigma} \otimes \mathbf{I_N})^{-1}(\mathbf{I_D} \otimes \mathbf{Z})]^{-1}. \tag{8}$$

In the general case, BLUE for the regression coefficients are provided by Zellner estimation (Zellner 1962). Because the dependent variables are all modeled with the same explanatory variables, the GLS estimates turn out to be equivalent to the OLS estimates for the regression coefficients (Srivastava and Giles 1987, Ch.2). When the Kronecker products in (7) and (8) are worked out it follows that:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\mathbf{GLS}} &= \mathbf{I_D} \otimes (\mathbf{Z}^{\mathbf{t}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathbf{t}}\mathbf{P}, \\ \text{V}(\hat{\boldsymbol{\beta}}_{\mathbf{GLS}}) &= \boldsymbol{\Sigma} \otimes (\mathbf{Z}^{\mathbf{t}}\mathbf{Z})^{-1}. \end{aligned}$$

Note that the SURE model reduces to a multivariate regression model (Mardia, Kent and Bibby, 1979, Ch.6).

The elements of $\boldsymbol{\Sigma}$ can be estimated by

$$\begin{aligned} \hat{\boldsymbol{\Sigma}} &= [\hat{\sigma}_{jj'}] = \left[\frac{\hat{\boldsymbol{\epsilon}}^{\mathbf{j}}\hat{\boldsymbol{\epsilon}}^{\mathbf{j}'\mathbf{t}}}{(N-K)} : j = 1, 2, \ldots, D \right], \\ \hat{\boldsymbol{\epsilon}}^{\mathbf{j}} &= \mathbf{p}^{\mathbf{j}} - \mathbf{Z}\hat{\boldsymbol{\beta}}^{\mathbf{j}}_{\mathbf{OLS}}. \end{aligned}$$

Because the dependent variables are restricted to the hyperplane $\sum_{j=1}^{D} p_{j|i} = 1$, the parameter restrictions (2), (3) and (4) are automatically fulfilled. Due to the restriction $\sum_{j=1}^{D} p_{j|i} = 1$ it follows that $\boldsymbol{\Sigma}\mathbf{j_D} = \mathbf{0}$ (with $\mathbf{j_D}$ a D dimensional vector with each element 1). Therefore, the covariance matrix $\boldsymbol{\Sigma} \otimes \mathbf{I_n}$ will be always singular. In fact the system

of regression equations is completely determined by $D-1$ regression equations. The last regression equation is known by the restriction that the dependent variables are on the hyperplane $\sum_{j=1}^{D} p_{j|i} = 1$ and the parameter restrictions (2), (3) and (4).

SURE models seem to be appropriate to relate the components of a D part composition to K explanatory variables in a linear regression model. However, it should be emphasized here that only the restriction of the unit-sum constraint is incorporated in the SURE model and that the restriction that each component is non-negative is not used. SURE does not model compositional data observed on the simplex $\mathcal{S}^d$, but data observed in a hyperplane in the space of $\mathcal{R}^D$, defined by the unit-sum constraint. A consequence of this is that after having estimated the regression parameters, it is possible to predict compositions outside the range $[0,1]$. This is not problematic if $\hat{\mathbf{\Sigma}}$ is small enough and the elements of $\hat{\mathbf{p}}_\mathbf{D}$ are not to close to the boundaries of the simplex. However, if there are many compositions that take values close to the boundaries of the simplex, it is better to transform the compositions $\mathbf{p_D}$ with the logratio transformation (1) and to model these transformed compositions in a multivariate regression model (section 2). The advantage of the use of SURE models for the analysis of compositional data instead of logratio analysis is that the interpretation of the estimation results is less complicated because no non-linear transformation is applied.

4 Latent budget analysis

The methods described in section 2 and 3 are appropriate if from each object a composition is obtained which can be regarded as an independent replication observed on the simplex. However, compositional data can also arise in a completely different way from contingency tables. Let the $I \times D$ matrix $\mathbf{N}$ denote a contingency table. The rows correspond with I objects and the columns correspond with D different categories of a composition. For example by time budget analysis, the rows correspond with I persons or groups of persons and the columns correspond with D different activities of time spending. Counting at random points in time which of the D activities, the different objects are doing, gives information how these objects spend their time over these activities. If these counts can be regarded as independent observations, a $I \times D$ contingency table $\mathbf{N}$ arise and can be assumed to be generated by a product multinomial distribution. From contingency table $\mathbf{N}$, for each object (row) a budget that specifies the proportions of time spent at each of the D activities can be obtained by dividing each element of $\mathbf{N}$ by its row total. Note that the elements of these budgets are conditional probabilities which can be regarded as compositions because they are non-negative and add up to one.

The Latent Budget Model (LBM) was originally proposed by Clogg (1981) for the analysis of square social mobility tables. The (LBM) was also proposed by de Leeuw

and van der Heijden (1988) and by de Leeuw, van der Heijden and Verboon (1990) for the analysis of time budgets. The basic data are a contingency table $\mathbf{N}$ with I rows (corresponding with I objects) and D columns (corresponding to the D categories of the composition). Element n_{ij} denotes the number of times that object i is observed in category j. For each object i there is a (unknown) budget $\boldsymbol{\pi}_{\mathbf{D}} = (\pi_{1|i}, \ldots, \pi_{j|i}, \ldots, \pi_{D|i})^t$. This budget can be estimated by $\mathbf{p}_{\mathbf{D}} = (p_{1|i}, \ldots, p_{j|i}, \ldots, p_{D|i})^t$ with $p_{j|i} = n_{ij}/n_{i+}$ ($n_{i+} = \sum_{j=1}^{D} n_{ij}$). The budgets (or compositions) $\boldsymbol{\pi}_{\mathbf{D}}$ are the probabilities over the D categories conditional on object i.

The LBM tries to explain the budgets $\boldsymbol{\pi}_{\mathbf{D}}$ (estimated by $\mathbf{p}_{\mathbf{D}}$) with a mixture of K latent or typical budgets:

$$\pi_{j|i} = \sum_{k=1}^{K} \alpha_{k|i}\beta_{j|k} = \sum_{k=1}^{K} \pi_{jk|i} \quad (i = 1, 2, \ldots, I; j = 1, 2, \ldots, J; k = 1, 2, \ldots, K).$$

with the constraints:

$$\begin{aligned} \sum_{k=1}^{K} \alpha_{k|i} &= 1 \quad \alpha_{k|i} \geq 0 \quad (i = 1, 2, \ldots I; k = 1, 2, \ldots, K), \\ \sum_{j=1}^{J} \beta_{j|k} &= 1 \quad \beta_{j|k} \geq 0 \quad (j = 1, 2, \ldots J; k = 1, 2, \ldots, K). \end{aligned}$$

The latent budgets are defined by parameters $(\beta_{1|k}, \beta_{2|k}, \ldots, \beta_{D|k})^t$. These are also conditional probabilities, specifying latent or typical distributions over the D categories conditional on latent budget k. For each object there are K parameters $\alpha_{k|i}$. These are the mixture parameters for the K latent budgets. Parameter $\alpha_{k|i}$ is a probability conditional on object i and specifies to what extent latent budget k explains the budget of object i.

Under the assumption of a product multinomial distribution as a sampling model for data matrix $\mathbf{N}$, maximum likelihood estimates of the parameters can be obtained with the EM algorithm (Dempster, Laird and Rubin, 1977). See de Leeuw, van der Heijden and Verboon (1990) for expressions of the maximum likelihood estimates for $\alpha_{k|i}$ and $\beta_{j|k}$.

On the α and β parameters of the LBM, different types of parameter constraints can be posed; e.g. equality constraints, fixed value constraints and multinomial logit constraints. These constraints are important for several reasons. First, imposing constraints simplifies the model because it reduces the number of parameters to be interpreted. Second, substantive research questions can some times be formulated in terms of constraints on the parameters of the LBM. Testing the constraints then provides answers to these research questions. Third, constraining the parameters reduces the standard errors of the unconstrained parameter estimates. These parameter constraints are extensively described in van der Heijden, Mooijaart and de Leeuw (1992).

Fixed value constraints specify that specific parameters are equal to certain values. These constraints can be used to test whether an estimate of a parameter is different from

values that are of theoretical interest. In many circumstances such values will be zero or one. The constraint $\alpha_{k|i} = 0$ (or 1), tests the hypothesis that the budget for object i is not (or completely) explained with latent budget k. The constraint $\beta_{j|k} = 0$, tests the hypothesis that latent budget k does not participate in category j.

Equality constraints specify that certain parameters are unknown but equal to each other. These constraints can be used to test if two or more parameters are different. An important application is the test for identity of two budgets $\boldsymbol{\pi}_{\mathbf{D}}$ $(\alpha_{k|i} = \alpha_{k|i'}, \forall k)$.

Additional information about the objects (rows) or categories (columns) can be used to constrain the α or β parameters. Because the parameters $\alpha_{k|i}$ and $\beta_{j|k}$ are conditional probabilities, the multinomial logit model is an appropriate model. This idea is proposed and worked out in van der Heijden, Mooijaart and de Leeuw (1992).

Let v_{im} denote the variables containing additional information for the objects i $(m = 1, 2, \ldots, M)$. The α parameters can be modeled using these variables by the multinomial logit model:

$$\alpha_{k|i} = \frac{\exp\left(\sum_{m=1}^{M} v_{im}\gamma_{mk}\right)}{\sum_{k=1}^{K} \exp\left(\sum_{m=1}^{M} v_{im}\gamma_{mk}\right)},$$

were the γ_{mk} are parameters. This model can be identified by constraining $\gamma_{m1} = 0, \forall m$.

In a similar way the β parameters can be modeled using additional information in a multinomial logit model. Let variables w_{jh} containing additional information about the categories $(h = 1, 2, \ldots H)$. The multinomial logit model for the β parameters reads as:

$$\beta_{j|k} = \frac{\exp\left(\sum_{h=1}^{H} w_{jh}\psi_{hk}\right)}{\sum_{j=1}^{J} \exp\left(\sum_{h=1}^{H} w_{jh}\psi_{hk}\right)},$$

were the ψ_{hk} are parameters. This model can be identified by constraining $\psi_{h1} = 0, \forall h$.

Because the α and β parameters are conditional probabilities, the interpretation is very easy, also for non-statisticians. A disadvantage is that the model assumes a product multinomial distribution as sampling model. In many applications this requirement is not met, which limits the application of the model.

Related appropriate models are latent class models and loglinear models with latent variables because both assume a (product) multinomial distribution as a sampling model (Haberman 1979, Ch.10 or Hagenaars 1990).

5 Canonical correspondence analysis

Correspondence Analysis (CA) is popular as an explorative data analysis technique and may be regarded as a tool to obtain a low-rank approximation of a data matrix. This

CA application is used to study the relationships between rows and columns of two-way contingency tables graphically. Although CA is mainly intended for two-way contingency tables, it can be used for the analysis of any two-way matrix with non-negative entries. For this reason CA seems to be an appropriate analysis technique for compositional data.

The low-rank approximation can be obtained with least squares or with maximum likelihood. A least squares approximation has the advantage that no assumptions about the form of a sampling model of the matrix (or contingency table) are required. Up to now, maximum likelihood requires a Poisson or (product)multinomial distribution as sampling model. One of the advantages of a maximum likelihood approach is that hypotheses about parameter constraints of the model can be tested if the assumptions are met (Goodman 1985 and 1986, Gilula and Haberman 1986). Only the least squares approximation is discussed here.

In CA the relationship between rows and between columns of the data matrix is studied in an explorative way. For discussion of correspondence analysis, the notation of Greenacre (1984) is used. Consider the $I \times D$ data matrix $\mathbf{N}$ with non-negative elements n_{ij}. Data matrix $\mathbf{N}$ can be scaled to the so called $I \times D$ correspondence matrix $\mathbf{P}$ with proportions in each of its elements p_{ij}. Thus $p_{ij} = n_{ij}/n_{++}$ with $n_{++} = \sum_{i=1}^{I} \sum_{j=1}^{D} n_{ij}$ and $\sum_{i=1}^{I} \sum_{j=1}^{D} p_{ij} = 1$. Let $\mathbf{r}$ be the vector of row sums of $\mathbf{P}$ and let $\mathbf{c}$ be the vector of column sums of $\mathbf{P}$. Let $\mathbf{D_r}$ be the $I \times I$ diagonal matrix with the elements of $\mathbf{r}$ and let $\mathbf{D_c}$ be the $D \times D$ diagonal matrix with the elements of $\mathbf{c}$.

In CA differences between the rows of $\mathbf{P}$ are measured by the chi-squared distance between the so called row profiles, defined as the rows of the $I \times D$ matrix $\mathbf{R} = \mathbf{D_r^{-1}P}$. Scaling the row vectors of $\mathbf{P}$ to row profiles $\mathbf{R}$ makes comparison between the rows easier. The chi-squared distance between the row profiles i and i' (in the metric $\mathbf{D_c^{-1}}$) is defined as:

$$\delta_{ii'} = \sum_{j=1}^{J} \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2. \tag{9}$$

Note that the elements r_{ij} of $\mathbf{R}$ are non-negative and that the elements of $\mathbf{R}$ add up to one row wise. Thus $\mathbf{R}$ corresponds with the characteristics for compositional data mentioned in section 1. Because the row profiles correspond with the compositional vectors, CA seems to be well suited for compositional data analysis.

In a similar way differences between the columns of $\mathbf{P}$ are measured by the chi-squared distances between the column profiles, which are the rows of the $D \times I$ matrix $\mathbf{C} = \mathbf{D_c^{-1}P^t}$. The chi-squared distance between the column profiles j and j' (in the metric $\mathbf{D_r^{-1}}$) is defined as:

$$\delta_{jj'} = \sum_{i=1}^{I} \frac{1}{p_{i+}} \left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}} \right)^2. \tag{10}$$

Let M denote the rank of the centered correspondence matrix $\mathbf{P} - \mathbf{rc}^t$ ($M \leq \min(I - 1, D - 1)$). Principal axes that maximize the chi-squared distances between the row profiles (9) and the chi-squared distances between the column profiles (10) are obtained by performing a Generalised Singular Value Decomposition (GSVD) on the centered [1] correspondence matrix $\mathbf{P} - \mathbf{rc}^t$ in the metrics $\mathbf{D}_\mathbf{r}^{-1}$ and $\mathbf{D}_\mathbf{c}^{-1}$ (Greenacre, 1984, Ch.4):

$$\mathbf{P} - \mathbf{rc}^t = \mathbf{A}\mathbf{D}_\mathbf{u}\mathbf{B}^t, \quad \mathbf{A}^t\mathbf{D}_\mathbf{r}^{-1}\mathbf{A} = \mathbf{B}^t\mathbf{D}_\mathbf{c}^{-1}\mathbf{B} = \mathbf{I}.$$

The columns of the $I \times M$ matrix $\mathbf{A}$ correspond to the M left singular vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M$ and form an orthonormal basis for the columns of $\mathbf{P} - \mathbf{rc}^t$. The columns of the $D \times M$ matrix $\mathbf{B}$ consists of the M right singular vectors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_M$ and form an orthonormal basis for the rows of $\mathbf{P} - \mathbf{rc}^t$. The $M \times M$ diagonal matrix $\mathbf{D}_\mathbf{u}$ contains the singular values $u_1 \geq u_2 \geq \ldots \geq u_{M-1} \geq u_M > 0$.

Let $\mathbf{F}$ be the $I \times M$ matrix that contains the principal coordinates of the row profiles with respect to the principal axes of $\mathbf{B}$. The right singular vectors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_M$ define the principal axes for principal coordinates of the row profiles. Now the principal coordinates of the row profiles in $\mathbf{F}$ with respect to the principal axes of $\mathbf{B}$ in the chi-squared metric $\mathbf{D}_\mathbf{c}^{-1}$ can be shown to be (Greenacre, 1984, Ch.4):

$$\mathbf{F} = \mathbf{D}_\mathbf{r}^{-1}\mathbf{A}\mathbf{D}_\mathbf{u}. \tag{11}$$

Because $\mathbf{B}$ is orthonormal in the metric $\mathbf{D}_\mathbf{c}^{-1}$ the euclidean distance between the row points of $\mathbf{F}$ is equivalent to the chi-squared distance between the row points of $\mathbf{R}$. The beauty of the GSVD is that an optimal m^*-dimensional sub-space (in a least squares sense) for the row profiles is provided by the first m^* columns of $\mathbf{F}$ (Greenacre 1984, Ch.3). Let the $I \times 2$ matrix $\mathbf{F}_{[2]}$ contain the first two columns of $\mathbf{F}$. A plot of the row points of $\mathbf{F}_{[2]}$ provides a graphical display of the approximate chi-squared distances between the row profiles (in a least squares sense). Thus a plot of $\mathbf{F}_{[2]}$ provides a graphical representation of the differences between the row profiles $\mathbf{R}$ expressed as chi-squared distances (9) and thus express the differences between the rows of $\mathbf{N}$ or $\mathbf{P}$. The greater the chi-square differences between the row profiles, the greater the euclidean distance between the row points in the plot of $\mathbf{F}_{[2]}$.

The left singular vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M$ define the principal axes for the principal coordinates of the column profiles. The principal coordinates of the column profiles in $\mathbf{G}$ with respect to the principal axes of $\mathbf{A}$ in the chi-squared metric $\mathbf{D}_\mathbf{r}^{-1}$ can be shown to be (Greenacre, 1984, Ch.4):

[1] The corresponding matrix $\mathbf{P}$ is centered by subtracting $\mathbf{rc}^t$ (the expected proportions of the independence matrix). If $\mathbf{P}$ is not centered, the first singular value and the first left and right singular vectors in the GSVD are always the trivial solutions $u_1 = 1$, $\mathbf{a}_1 = \mathbf{r}$ and $\mathbf{b}_1 = \mathbf{c}$. For this reason, the trivial solution is removed by centering of $\mathbf{P}$.

$$\mathbf{G} = \mathbf{D_c^{-1}BD_u}. \tag{12}$$

An optimal two-dimensional sub-space (in a least squares sense) for the column profiles is provided by the first two columns of $\mathbf{G}$. Let the $D \times 2$ matrix $\mathbf{G_{[2]}}$ contain the first two columns of $\mathbf{G}$. A plot of the row points of $\mathbf{G_{[2]}}$ provides a graphical display of the approximate chi-squared distances between the column profiles (in a least squares sense). Thus a plot of the row points of $\mathbf{G_{[2]}}$ provides a graphical display of the differences of the column profiles $\mathbf{C}$ expressed as chi-squared distances (10) and thus express the differences between the columns of $\mathbf{N}$ or $\mathbf{P}$. Column points that are close together are more alike than column points that are far apart.

Additional information about the row profiles can be incorporated with so called canonical correspondence analysis (CCA), developed by ter Braak (1986). Here the scores for the row profiles are restricted to be a linear combination of covariables. Let $\mathbf{Q}$ be an $I \times M$ matrix of covariables variables pertaining to the row structure of the $I \times D$ matrix $\mathbf{P}$. The matrix $\mathbf{Q}$ may contain continuous and/or discrete variables. In CCA representations of the rows and columns of $\mathbf{P}$ are sought under the restriction that the row representation is a linear combination of $\mathbf{Q}$. The canonical correspondence analysis of the row and column profiles of the correspondence matrix $\mathbf{P}$ under the restriction that the row profiles are a linear combination of the additional variables in $\mathbf{Q}$ is obtained by performing a GSVD on the centered correspondence matrix $\mathbf{P} - \mathbf{rc^t}$ in the metrics $(\mathbf{Q^tD_rQ})^-$ and $\mathbf{D_c^{-1}}$:

$$\mathbf{P - rc^t = AD_uB^t}, \quad \mathbf{A^t(Q^tD_rQ)^-A = B^tD_c^{-1}B = I},$$

where $(\mathbf{Q^tD_rQ})^-$ is the Moore Penrose pseudo-inverse of $\mathbf{Q^tD_rQ}$ (in many cases rank$(\mathbf{Q}) < M$). Note that CA is the special case of CCA obtained for $\mathbf{Q} = \mathbf{I}$.

Principal coordinates of the row profiles are obtained from (11) where $\mathbf{D_r^{-1}}$ is replaced by $(\mathbf{Q^tD_rQ})^-$. Principal coordinates of the column profiles are obtained from (12).

To study the relationships between rows and columns of $\mathbf{P}$, the plots of the principal coordinates of the row profiles and column profiles from $\mathbf{F}$ and $\mathbf{G}$ respectively should not be merged into one plane. Distances between points from the row profiles or between points from the column profiles are explicitly defined in terms of weighted chi-squared distances and can be visualized graphically in plots of the principal coordinates of $\mathbf{F}$ and $\mathbf{G}$ respectively. Distances between points from row profiles and column profiles are not defined. Relationships between rows and columns of a matrix can be studied graphically in one plot with the so-called biplots (Gabriel (1971) and Gabriel (1981)). A biplot is a graphical display of an $I \times D$ matrix $\mathbf{N}$ of rank M by means of $m \times 1$ vectors $\mathbf{k_1}, \mathbf{k_2}, \ldots, \mathbf{k_I}$ for its rows and $m \times 1$ vectors $\mathbf{l_1}, \mathbf{l_2}, \ldots, \mathbf{l_D}$ for its columns. These vectors are chosen in such a way that the inner products $\mathbf{k_i^t l_j}$ represents the i,jth element n_{ij} of $\mathbf{N}$ (Gabriel 1971). Any $I \times D$ matrix $\mathbf{N}$ of rank M can be factorised as:

$$\mathbf{N} = \mathbf{K}\mathbf{L}^{\mathbf{t}}, \tag{13}$$

where $\mathbf{K}$ is an $I \times M$ matrix, which rows correspond with the vectors $\mathbf{k_1}, \mathbf{k_2}, \ldots, \mathbf{k_I}$ representing the rows of $\mathbf{N}$ and $\mathbf{L}$ is an $D \times M$ matrix, which rows correspond with the vectors $\mathbf{l_1}, \mathbf{l_2}, \ldots, \mathbf{l_D}$ representing the columns of $\mathbf{N}$. The matrices $\mathbf{K}$ and $\mathbf{L}$ are both of rank M. The factorisation (13) assigns vectors $\mathbf{k_1}, \mathbf{k_2}, \ldots, \mathbf{k_I}$, one to each of the rows of $\mathbf{N}$ and vectors $\mathbf{l_1}, \mathbf{l_2}, \ldots, \mathbf{l_J}$, one to each of the columns of $\mathbf{N}$. These $I + J$ vectors of order M provide a representation of $\mathbf{N}$ in a M dimensional space. The vectors $\mathbf{k_1}, \mathbf{k_2}, \ldots, \mathbf{k_I}$ may be regarded as row effects of $\mathbf{N}$ and the vectors $\mathbf{l_1}, \mathbf{l_2}, \ldots, \mathbf{l_J}$ as column effects of $\mathbf{N}$.

If matrix $\mathbf{N}$ is of rank 2, the vectors $\mathbf{k_1}, \mathbf{k_2}, \ldots, \mathbf{k_I}$ and $\mathbf{l_1}, \mathbf{l_2}, \ldots, \mathbf{l_J}$ are vectors of order 2. A plot of these $I + J$ vectors provides in an exact representation of the IJ elements of $\mathbf{N}$ by means of the inner products of the corresponding row effect and column effect vectors. Element n_{ij} of $\mathbf{N}$ is represented as the inner product of vectors $\mathbf{k_i}$ and $\mathbf{l_j}$. The inner product of two vectors $\mathbf{k_i}$ and $\mathbf{l_j}$ may be interpreted visually as the product of the length of one of these vectors times the length of the other vectors projection onto it (Gabriel 1971).

Matrices of ranks higher than two cannot be represented exactly by a biplot. With a GSVD a two rank least squares approximation $\mathbf{N_{[2]}}$ of $\mathbf{N}$ can be obtained. A biplot of $\mathbf{N_{[2]}}$ provides a least squares approximation biplot of the original matrix $\mathbf{N}$ and makes it easy to study the main relationships between the rows and columns of matrix $\mathbf{N}$.

In CCA the correspondence matrix $\mathbf{P}$ is decomposed as: $\mathbf{P} = \mathbf{Q^t D_r Q(J + F D_u^{-1} G^t) D_c}$ ($\mathbf{J}$ is an $I \times D$ matrix with each element 1). This is called the reconstitution formula and is the result of the decomposition of the centered correspondence matrix $\mathbf{P - rc^t}$, obtained with a GSVD (Greenacre 1984, Ch. 4). With the biplot a joint representation of the row and column points of $\mathbf{N}$ can be obtained. The reconstruction formula can be rewritten as: $\mathbf{(Q^t D_r Q)^-(P - rc^t) D_c^{-1} = F D_u^{-1} G^t}$. Three natural candidates for the factorisation matrices $\mathbf{K}$ and $\mathbf{L}$ of (13) are possible:

1. An asymmetric correspondence analysis display where the rows of $\mathbf{N}$ are displayed in principal coordinates $\mathbf{F}$ and the columns of $\mathbf{N}$ in standard coordinates $\mathbf{G D_u^{-1}}$:

   $$\mathbf{K = F_{[2]}}, \quad \mathbf{L = G_{[2]} D_{u[2]}^{-1}}.$$

   This biplot can be interpreted as a graphical display of the rows of $\mathbf{F}$ in a weighted average of the columns of $\mathbf{G D_u^{-1}}$.

2. An asymmetric correspondence analysis display where the rows of $\mathbf{N}$ are displayed in standard coordinates $\mathbf{F D_u^{-1}}$ and the columns of $\mathbf{N}$ in principal coordinates $\mathbf{G}$:

   $$\mathbf{K = F_{[2]} D_{u[2]}^{-1}}, \quad \mathbf{L = G_{[2]}}.$$

   This biplot can be interpreted as a graphical display of the rows of $\mathbf{G}$ in a weighted average of the columns of $\mathbf{F D_u^{-1}}$.

3. Or a symmetric correspondence analysis display:

$$\mathbf{K} = \mathbf{F}_{[2]}\mathbf{D}_{\mathbf{u}[2]}^{-\frac{1}{2}}, \quad \mathbf{L} = \mathbf{G}_{[2]}\mathbf{D}_{\mathbf{u}[2]}^{-\frac{1}{2}}.$$

In this symmetric biplot display there is not such a straight forward interpretation as by the two former asymmetric biplot displays.

6 Discussion and conclusions

In this article four different methods for the analysis of compositional data were reviewed. The statistical inferences with the methods proposed by Aitchison are very powerful. Because Aitchison starts with the simplex $\mathcal{S}^d$ as the sampling space and the logistic normal distribution $\mathcal{L}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as an appropriate probability density to model compositions, the whole range of multivariate statistical methods based on the multivariate normal distribution becomes applicable for compositional data analysis. Aitchison emphasized that D part compositions are essentially d-dimensional vectors. In order to analyze compositions with these standard statistical methods, they must be transformed with the logratio transformation from the constrained space of $\mathcal{S}^d$ to $\mathcal{R}^d$. The disadvantage of this methods is that due to the application of this non-linear transformation the interpretation of the estimation results becomes more complicated.

The advantage of SURE models is that no non-linear transformations are applied. This simplifies the interpretation of the estimation results. Theoretically, SURE models are not a completely appropriate analysis methods for compositional data, because the restriction that the compositions are elements of the simplex $\mathcal{S}^d$ is replaced by the restriction that compositions are elements of the hyperplane in $\mathcal{R}^D$ defined by the unit-sum constraint. The restriction that the components of the composition only take values in the range between zero and one and the dimensionality of compositional vectors is ignored. A consequence of this is that after having estimated the regression coefficients, it is possible to predict proportions that are negative or greater than one. This method is applicable if the values of the components of the composition do not take values close to the boundaries of the simplex and if the variance of the error terms is not to large and if the residuals are tested for normality.

If it is reasonable to assume that the observed data follow a product multinomial distribution, the Latent Budget Model is an appropriate analysis method. Because the parameters of the LBM can be interpreted as conditional probabilities, the interpretation of the results of this model is very simple. Many substantive research questions can be answered by testing restrictions, that can be posed on the parameters of the LBM. A disadvantage of the LBM is the assumption of a product multinomial distribution as a sampling model. Because this assumption often is unfounded, the applicability of the

LBM is limited. If it can be assumed that the data are generated by a product multinomial distribution, a second advantage of the LBM in comparison with the methods proposed by Aitchison is that differences between the objects can be studied. The reason is that in the LBM there are for each object (the rows of the data matrix) specific parameters (α), explaining the observed budget as a mixture of K latent budgets (β). This is possible because for each object a budget is estimated on the basis of n_{i+} independent observations. By Aitchisons methods there are no such row-specific parameters because each observed budget is regarded as one independent observed replication drawn from a logistic normal distribution. Using Aitchisons methods for the analysis of budget data, generated by a (product)multinomial sampling model, has the disadvantage that the number of independent observations reduces from n_{++} to I and it is no longer possible to investigate differences between the individual objects.

In Correspondence Analysis differences between the rows and columns of the data matrix are expressed as chi-squared distances between row profiles and between column profiles. Because the row profiles are also the compositional vectors, CA seems to be very well suited for compositional data analysis. Because there are no model assumption, CA is always applicable as an explorative analysis method, to investigate relationships between rows and columns of a data matrix. However, the inferential side of CA is not very well developed.

References

[1] Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman and Hall, London.

[2] Clogg, C.C. (1981). Latent Structure Models of Mobility. American Journal of Sociology, 86, 836-868.

[3] de Leeuw, J. and P.G.M. van der Heijden (1988). The Analysis of Time-Budgets with a Latent Time-Budget Model. In: E. Diday e.a. (eds), *Data Analysis and Informatics*, Amsterdam: North Holland, 5, 159-166.

[4] de Leeuw, J., P.G.M. van der Heijden and P. Verboon (1990). A Latent Budget Model. *Statistica Neerlandica*, 44, 1-22.

[5] Dempster, A.P., N.M. Laird and D.B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, ser. B, 39, 1-38.

[6] Gabriel, K.R. (1971). The Biplot Graphic Display of Matrices with application to Principal Component Analysis. *Biometrica*, 58, 453-467.

[7] Gabriel, K.R. (1981). Biplot Display of Multivariate Matrices for Inspection of Data and Diagnoses. In: Barnett, V. (eds), *Interpreting Multivariate Data*, pp. 147-174. Wiley, Chichester, UK.

[8] Goodman, L.A. (1985). The Analysis of Cross-Classified Data having Ordered and/or Unordered Categories: Association Models, Correlation Models, and Asymmetric Models for Contingency Tables with or without Missing Entries. The Annals of Statistics, 13(1), 10-69.

[9] Goodman, L.A. (1986). Some Useful Extensions of the Correspondence Analysis Approach and the Usual Log-Linear Models Approach in the Analysis of Contingency Tables. International Statistical Revieuw, 54(3), 243-309.

[10] Gilula, Z. and S.J. Haberman. (1986). Canonical Analysis of Contingency Tables by Maximum Likelihood. Journal of the American Statistical Association, 81, 780-788.

[11] Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis.* Academic Press, New York.

[12] Haberman, S.J. (1979). *Analysis of Qualitative Data*, vol 2. New York: Academic Press.

[13] Hagenaars, J.A. (1990). *Categorical Longitudinal Data.* Sage, London.

[14] Mardia, K.V., J.T. Kent and J.M. Bibby (1979). *Multivariate Analysis.* Academic Press, London.

[15] Pindyck, R.S. and D.L. Rubinfeld (1981). *Econometric Models and Economic Forecasts.* McGraw-Hill, New York.

[16] Srivastava, V.K. and D.E.A. Giles (1987). *Seemingly Unrelated Regression Equations Models.* Marcel Dekker, INC., New York.

[17] ter Braak, C.J.F. (1986). Canonical Correspondence Analysis: A new Eigenvector Technique for Multivariate Gradient Analysis. *Ecology*, 67, 97-120.

[18] van der Heijden, P.G.M. and J.A. van den Brakel (1993). Three Data Reduction Methods for the Analysis of Time Budgets. *Time use methodology: toward concensus*, Rome, Istat, ed. 1993, no. 3, 151-160.

[19] van der Heijden, P.G.M., A. Mooijaart and J. de Leeuw (1992). Constrained Latent Budget Analysis. In: Marsden, P. (eds), *Sociological Methodology*, Cambridge: Blackwell Publishers, volume 20, 279-320.

[20] Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association*, 57, 348-368.

Ontvangen: 24-4-1995
Geaccepteerd: 12-10-1995