Appl. Statist. (1994)
43, *No.* 3, *pp.* 527–530

Letters to the Editors

End-member Analysis and Latent Budget Analysis

The work described by Renner (1993) is similar to but different from work performed in the social sciences. For example, what Renner calls 'an end-member' is called 'a latent budget' in our work (for an overview see van der Heijden *et al.* (1992)). This term was adopted because the model was originally applied to time–budget data, and in this context a latent (time) budget was interpreted as a typical way of allocating time. The observed time budgets of people are then approximated by a convex combination of a number of latent time budgets.

The so-called 'latent budget model' has the same parametric form as that of the model used by Renner, namely

$$\mathbf{X} = \mathbf{LB} + \mathbf{E}.$$

However, in the latent budget model the estimation of the parameters is currently made by maximum likelihood under the assumption that the frequencies in each of the rows of $\mathbf{X}$ follow a multinomial distribution. The current estimation procedure of the latent budget model therefore differs from the estimation procedure used in end-member analysis.

Since Renner does not discuss the identification issue, I would like to point out here that the model parameters are unidentified. We could write

$$\mathbf{X} = \mathbf{LB} + \mathbf{E} = (\mathbf{LT})(\mathbf{T}^{-1}\mathbf{B}) + \mathbf{E} = \mathbf{L}^{+}\mathbf{B}^{+} + \mathbf{E},$$

where $\mathbf{T}$ is a $k \times k$ transformation matrix. The matrix $\mathbf{T}$ should be chosen such that the two restrictions of the original parameters $\mathbf{L}$ and $\mathbf{B}$ are also fulfilled for the new parameters $\mathbf{L}^{+}$ and $\mathbf{B}^{+}$. The first restriction is that $\mathbf{L}^{+}\mathbf{u} = \mathbf{u}$ and $\mathbf{B}^{+}\mathbf{u} = \mathbf{u}$, where $\mathbf{u}$ is a column vector of 1s, the length of which depends on the context. This requirement is fulfilled by choosing $\mathbf{T}$ so that $\mathbf{Tu} = \mathbf{u}$, i.e. the elements of $\mathbf{T}$ add up to 1, rowwise. The second requirement is that both $\mathbf{L}^{+} \geqslant \mathbf{0}$ and $\mathbf{B}^{+} \geqslant \mathbf{0}$. In general this still allows an infinite number of choices for $\mathbf{T}$. However, as pointed out by van der Heijden *et al.* (1992), the range from which a possible matrix $\mathbf{T}$ can be picked is sometimes so limited that some specific admissible choice for $\mathbf{T}$ will barely influence the interpretation of the parameter estimates. However, this identification problem leads to the conclusion that, in the general case, what Renner calls 'the estimated end-members' cannot be seen as estimates of '*true* end-members'. We usually solve the identification problem by choosing $\mathbf{T}$ such that elements of $\mathbf{L}^{+}$ or of $\mathbf{B}^{+}$ are restricted to 0. If possible, choices for a specific $\mathbf{T}$ should be guided by substantive theory. For more details see de Leeuw *et al.* (1990), de Leeuw and van der Heijden (1991) and van der Heijden *et al.* (1992).

Finally, the model described above is closely related to other contingency table models, namely correspondence analysis (see, for example, Greenacre (1984)) and latent class analysis of a two-way contingency table (see, for example, Clogg (1981)). In fact, de Leeuw and van der Heijden (1991) showed that when $k = 2$ the latent budget model (end-member model with two estimated end-members) is equivalent to correspondence

 0035–9254/94/43527

analysis with $k-1$ dimensions. For $k>2$, the latent budget model implies correspondence analysis, but the reverse does not hold. So if both models are estimated by using the same criterion (e.g. maximum likelihood), then the estimates of expected frequencies of both models are the same when $k=2$, and they can be the same (but do not have to be the same) when $k>2$.

Since correspondence analysis is usually estimated by using a generalized singular value decomposition (see, for example, Greenacre (1984)), it is worthwhile to study the implications of the choices for the metrics made in this estimation procedure compared with the choices made by Renner (1993), who used a simple singular value decomposition. This brings me to my final remark: the criterion with which the estimate space is determined is unclear. The reason is that a least squares criterion is used to approximate $\mathbf{X}$ by a matrix of lower rank $\mathbf{X}^*$ only in the first step, by means of the singular value decomposition. However, Renner subsequently makes adjustments to $\mathbf{X}^*$ to ensure that the elements of $\mathbf{X}^*$ are non-negative and sum to 1, rowwise. Thus the final result no longer approximates the matrix $\mathbf{X}$ in a least squares sense.

References

Clogg, C. C. (1981) Latent structure models of mobility. *Am. J. Sociol.*, **86**, 836–868.

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

van der Heijden, P. G. M., Mooijaart, A. and de Leeuw, J. (1992) Constrained latent budget analysis. *Sociol. Methodol.*, **22**, 279–320.

de Leeuw, J. and van der Heijden, P. G. M. (1991) Reduced rank models for contingency tables. *Biometrika*, **78**, 229–232.

de Leeuw, J., van der Heijden, P. G. M. and Verboon, P. (1990) A latent time budget model. *Statist. Neerland.*, **44**, 1–22.

Renner, R. M. (1993) The resolution of a compositional data set into mixtures of fixed source compositions. *Appl. Statist.*, **42**, 615–631.

Peter G. M. van der Heijden
Department of Methodology and Statistics
Faculty of Social Sciences
Utrecht University
Postbus 80.140
3508 TC Utrecht
The Netherlands

[Received October 1993]