PSYCHOMETRIKA—VOL. 53, NO. 2, 287–291
JUNE 1988
NOTES AND COMMENTS

COMMENT ON "CORRESPONDENCE ANALYSIS USED COMPLEMENTARY TO LOGLINEAR ANALYSIS"

PETER G. M. VAN DER HEIJDEN

DEPARTMENT OF PSYCHOMETRICS
UNIVERSITY OF LEIDEN

KEITH J. WORSLEY

DEPARTMENT OF MATHEMATICS AND STATISTICS
MCGILL UNIVERSITY

In van der Heijden and de Leeuw (1985) it was proposed to use loglinear analysis to detect interactions in a multiway contingency table, and to explore the form of these interactions with correspondence analysis. After performing the exploratory phase of the analysis, we will show here how the results found in this phase can be used for confirmation.

Key words: loglinear models, association models, correspondence analysis, data analysis, exploration, confirmation.

Introduction

Van der Heijden and de Leeuw (1985) show that correspondence analysis (CA) of a so-called multiple table can be interpreted as providing a decomposition of the difference between the observed frequencies and expected frequencies for a restrictive loglinear model. They propose to use this result to circumvent the "cumbersome job" of interpreting a large number of loglinear interaction parameters: It is shown that loglinear analysis can be used to detect interactions, whereas CA can represent these interactions graphically. In de Leeuw and van der Heijden (1988) the same approach is used for incomplete contingency tables.

Another way to facilitate the interpretation of interaction parameters is to restrict the interaction in some form or another, for example, to have a product form. In this way the number of parameters to be interpreted can be reduced considerably. An example of such a model is the RC association model, which is closely related to CA (Goodman, 1981, 1985, 1986). In this paper we use CA as a tool for exploring the interaction, in order to obtain simpler models in which interaction parameters are restricted.

Relation between CA and the RC association model

In the two-variable case CA and the RC association model are related in the following way. CA representations in k dimensions can be written as

$$m_{ij} = p_{i\cdot}p_{\cdot j}\left(1 + \sum_{\alpha=1}^{k} \lambda_\alpha r_{i\alpha} c_{j\alpha}\right), \tag{1}$$

This research was conducted while the authors were visiting the Laboratoire de Statistique et Probabilité, Universite Paul Sabatier, Toulouse. This visit was partly made possible by a joint grant of the Netherlands Organisation for the Advancement of Pure Research (Z.W.O.) and the French National Center for Scientific Research (C.N.R.S.). For helpful comments, the authors are indebted to H. Caussinus, J. de Leeuw, and two anonymous reviewers.

Requests for reprints should be sent to P. G. M. van der Heijden, Department of Psychometrics, University of Leiden, Hooigracht 15, 2312KM Leiden, THE NETHERLANDS.

0033-3123/88/0600-9006$00.50/0

where p_{ij} is the observed proportion, and m_{ij} is the reconstituted proportion for cell (i, j); λ_α is the singular value for dimension α; $r_{i\alpha}$ and $c_{j\alpha}$ are scores for row i and column j on dimension α normalized so that $\sum_i p_{i\cdot} r_{i\alpha} = 0 = \sum_j p_{\cdot j} c_{j\alpha}$ and $\sum_i p_{i\cdot} r_{i\alpha}^2 = 1 = \sum_j p_{\cdot j} c_{j\alpha}^2$. Escoufier (1982) noted that if $x = \sum_{\alpha=1}^{k} \lambda_\alpha r_{i\alpha} c_{j\alpha}$ is small compared to 1, so that $\log(1 + x)$ is approximately equal to x, we can rewrite (1) as

$$\log m_{ij} \approx u + u_{1(i)} + u_{2(j)} + \sum_{\alpha=1}^{k} \lambda_\alpha r_{i\alpha} c_{j\alpha}, \tag{2}$$

where $u = 0$, $u_{1(i)} = \log(p_{i\cdot})$, $u_{2(j)} = \log(p_{\cdot j})$. Goodman (1981) showed that, if $k = 1$, and the p_{ij} stem from a discretized bivariate normal distribution (or a distribution that is bivariate normal after a proper transformation of the rows and columns), no matter how large λ_1, (1) is closely related to the RC association model

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \phi u_{1(i)}^* u_{2(j)}^* \tag{3}$$

in the following way: $\phi = \lambda_1$, and $r_{i1} = u_{1(i)}^*$ and $r_{j1} = u_{2(j)}^*$, where $u_{1(i)}^*$ and $u_{2(j)}^*$ are normalized in the same way as $r_{i\alpha}$ and $c_{j\alpha}$. Note that (2) is an approximation in k components, whereas (3) has only one component. For more details and related material, see Goodman (1985, 1986).

Model (3) can be viewed as a loglinear model allowing for interaction, where the interaction $u_{12(ij)}$ is restricted to be $u_{12(ij)} = \phi u_{1(i)}^* u_{2(j)}^*$. For restrictions of the RC association model we refer to Goodman (1985, 1986). Because of the similarity between (1) and (3) CA provides us with indications for restricting the loglinear interaction parameters. In this paper we illustrate how this result can be used in practical situations.

Example

We further analyze the three-way table in van der Heijden and de Leeuw (1985), having variables "method of suicide" (M), "age" (A) and "sex" (S). We will use chi-squares in a descriptive way: firstly, because of the large sample size (in fact, we deal with a German population), all models are significant. Secondly, due to usage of CA solutions and repeated testing, we are not able to interpret the significance levels in the usual way, and, strictly speaking, the chi-squares cannot be used for confirmation (we come back to this in the discussion). Generally, likelihood-ratio statistics are preferable to test restrictive models against each other, because they can be partitioned. However, here we will use Pearson chi-squares χ^2, because the relation between the eigenvalues of CA and the χ^2 for independence is explicitly defined.

In order to perform CA, van der Heijden and de Leeuw (1985) constructed a two-way table in which A and S were coded interactively, thus creating a new variable with $17 \times 2 = 34$ categories. The first two CA dimensions were shown in their Figure 1. These display $\tau_1 = .519$ and $\tau_2 = .381$ of the association as measured by $\chi^2 = 9995$ for model $[M][SA]$. CA decomposes the residuals from this model, being the terms $u_{12(ij)} + u_{13(ik)} + u_{123(ijk)}$. Here the variables M, S and A are denoted as 1, 2 and 3, indexed with i, j and k respectively. We shall now pick out which particular parts of the three terms are shown by the first two dimensions of CA.

Dimension 1 reveals a difference between men and women relative to their use of the methods, and does not involve age. We conclude that we can write the interaction between S and M, $u_{12(ij)}$, as a fixed term different for men and women, $u_{2(j)}^*$, multiplied by the parameters for the method categories, $u_{1(i)}^*$: $u_{12(ij)} = u_{1(i)}^* u_{2(j)}^*$. This is not restrictive, since S has two categories, and the main effect term $u_{1(i)}$ is already present in the model. If $u_{1(i)}^* u_{2(j)}^*$ is added to $[M][SA]$, then the model becomes $[MS][SA]$. Fitting $[MS][SA]$ we find $\chi^2 = 4519$ (df is 256). The proportion of χ^2 that is "explained" by including $u_{1(i)}^* u_{2(j)}^*$, is

(9995 − 4519)/9995 = .548, which is near $\tau_1 = .519$. Thus our interpretation of dimension 1 as showing $S - M$ interaction appears to be adequate.

Dimension 2 of CA shows that age scores are approximately the same for both sexes. Hence the second dimension reveals the interaction between A and M only. It also reveals that the interaction between A and M might be linear in A, when we exclude 10+ and 15+. Hence we write:

$$\begin{aligned} &u_{13(ik)} \text{ unrestricted for } k = 1, 2 \\ &u_{13(ik)} = u^{**}_{1(i)}k, \text{ for } k \geq 3. \end{aligned} \tag{4}$$

When we fit model [MS][MA][SA] with Restriction (4), we find $\chi^2 = 999$ (df is 232), providing a reduction of (4519–999)/9995 = .352; that is close to $\tau_2 = .381$. This tends to confirm that the interaction between A and M as displayed on Dimension 2 is linear (from 15+ onward). In Figure 1, $u^{*}_{1(i)}$ and $u^{**}_{1(i)}$ are plotted. The resemblance with the CA

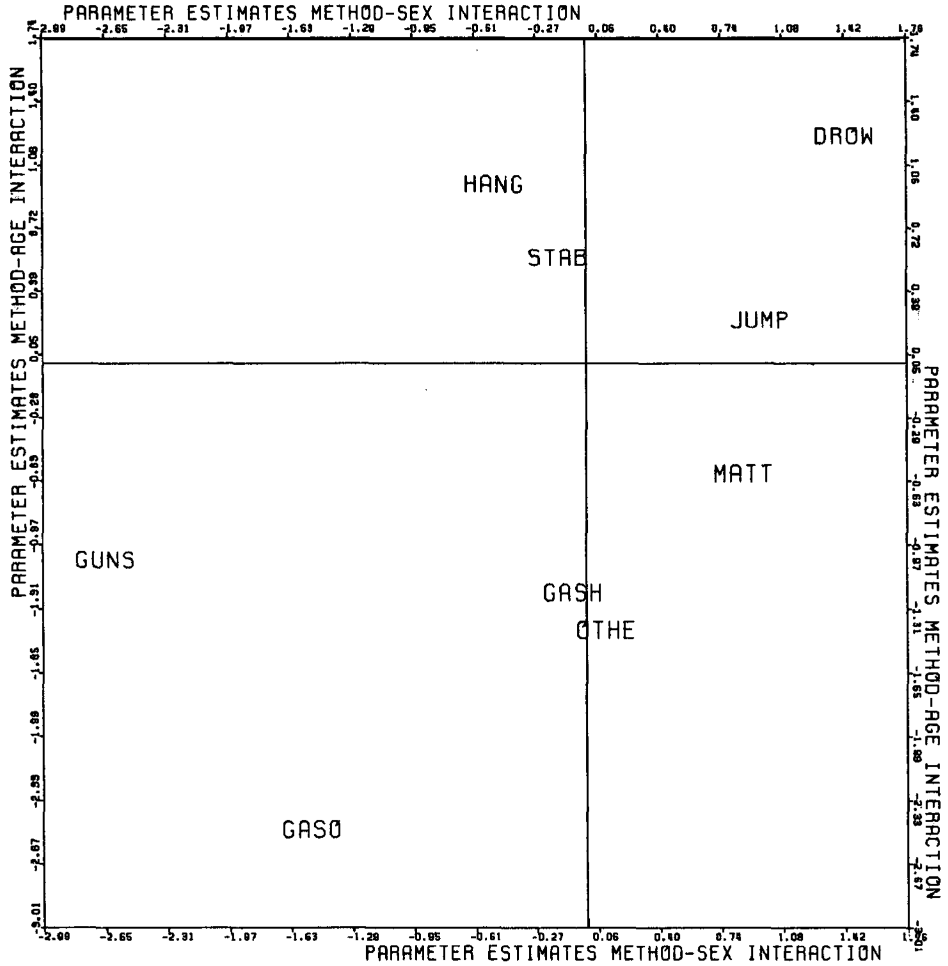

FIGURE 1
Plot of method scores $u^{*}_{1(i)}$ and $u^{**}_{1(i)}$. The scores are normalized so that $\sum_i p_{i\cdot} r_{i\alpha} = 0 = \sum_j p_{\cdot j} c_{j\alpha}$ and $\sum_i p_{i\cdot} r^2_{i\alpha} = 1 = \sum_j p_{\cdot j} c^2_{j\alpha}$.

solution is striking. However, for model [MS][MA][SA] in which the interactions are unrestricted, $\chi^2 = 436$ (df is 128). This χ^2 is much lower than the χ^2 for model [MS][MA][SA] with (4), so (4) is clearly not sufficient to model the complete A–M interaction. Therefore we studied the third CA dimension. It reveals roughly a quadratic effect in A. When we model this by replacing the second part of (4) by $u_{13(ik)} = u^{**}_{1(i)}k + u^{***}_{1(i)}k^2$ for $k \geq 3$, we find $\chi^2 = 646$, with 224 df. This becomes quite near the value given above for the unrestricted $A - M$ interaction. 96 d.f.'s are won by restricting $u_{13(ik)}$ in this way.

To get ideas for a restrictive modeling of the second order interaction, we use Figure 4 of van der Heijden and de Leeuw (1985). Dimension 1 of this figure indicates that a large part of this interaction can be written in product form as $u_{123(ijk)} = u^{****}_{1(i)}u^{*}_{2(j)}k^2$, where $u^{*}_{2(j)}$ for S is fixed as before, A is taken to be quadratic, and only parameters $u^{****}_{1(i)}$ for M have to be estimated. Model [MSA] with $u_{123(ijk)} = u^{****}_{1(i)}u^{*}_{2(j)}k^2$ has $\chi^2 = 263$, with 120 df. Compared to [MS][MA][SA], (436 – 263)/436 = .400 of the second-order interation is modeled restrictively (using only 8 df). It departs somewhat from $\tau_1 = .491$ for Figure 4, which is partly because the sum of the eigenvalues of generalized CA do not correspond precisely to χ^2.

Discussion and Conclusion

We think that CA can be useful as an exploratory method that helps finding a model with restrictions on the interaction parameters. For example, (some of) the categories of a variable can play a linear or quadratic role in the interaction. CA can also show that over different interactions the parameter estimates for some variable could be the same without a great loss of fit. So CA guides one directly to the proper restrictions to be chosen. As a reviewer indicated, a different approach is to plot estimated u-terms directly, for example, by setting out category numbers horizontally, and estimates vertically, and connecting estimates by lines. However, this does not always guide easily to the proper restrictions, because the lines of u-terms are unrestricted; in CA the categories are scaled optimally in some sense, so that one line is obtained. Secondly, if the categories of the variable do not have an a priori order, the plotting approach is more difficult to apply. Another alternative is not to stick with conventional CA, but, as in (2), to take off row and column effects and to apply PCA to the residual. We have no practical experience with this.

If an ordinary CA solution does not provide indications as clear as in our example, it might be useful to study generalized CA decomposing residuals from less restricted models. Thus some interactions are not shown, making the remaining interactions clearly visible (compare our use of Figure 4). Another option is to use another multiple table. In general, if one is interested in the role of a particular variable in some interaction, this role is more easily studied when this variable is not merged with another one. Thus the variable receives only one set of scores for each dimension.

A serious danger of our exploratory search for a model is that of overfitting the data. This was one of the reasons we used chi-squares in a descriptive way. However, the problem of overfitting can be solved by using cross-validation procedures (compare Bonett & Bentler, 1983).

References

Bonett, D. G., & Bentler, P. M. (1983). Goodness-of-fit procedure for the evaluation and selection of log-linear models. *Psychological Bulletin, 93*, 149–166.

de Leeuw, J., & van der Heijden, P. G. M. (1988). Correspondence analysis of incomplete tables. *Psychometrika, 53*, 223–233.

Escoufier, Y. (1982). L'analyse des tableaux de contingence simples et multiples [The analysis of simple and multiple contingency tables]. In R. Coppi (Ed.), Proceedings of the international meeting on the analysis of multidimensional contingency tables. *Metron, 40*, 53–77. (Rome, 1981)

Goodman, L. A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *Journal of the American Statistical Association*, *76*, 320–334.

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. *Annals of Statistics*, *13*, 10–69.

Goodman, L. A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear model approach in the analysis of contingency tables. *International Statistical Review*, *54*, 243–309.

van der Heijden, P. G. M., & de Leeuw, J. (1985). Correspondence analysis used complementary to loglinear analysis. *Psychometrika*, *50*, 429–447.

Manuscript received 7/25/86
Final version received 5/19/87