Statistical Modelling using Euclidean Distances

Mark de Rooij
Leiden University, Faculty of Social Sciences (FSW)
Department of Psychometrics and Research Methodology
Wassenaarseweg 52
2333 AK, Leiden, The Netherlands
rooijm@fsw.leidenuniv.nl

1. Introduction

De Rooij and Heiser (2002a; 2002b) show how to use Euclidean distances as model terms in log-linear models for two-way contingency tables. The advantages of such an approach are that distance plots are easily interpretable, and instead of having a bunch of numbers all effects can be shown in a single graph, which is highly attractive. Interaction effects are represented by common dimensions while the main effects are represented by unique dimensions. Since squared distances are used, the interpretation is easy since effects of different dimensions are additive; a proof of which is obtained through the theorem of Pythagoras. Log-linear models are just one instance of generalized linear models (McCullagh and Nelder, 1989) and here we will present the distance representation in that broader context. We will only consider categorical covariates, like in the log-linear model. First we will discuss generalized linear models, then make a translation to distance models for two covariates. We briefly discuss generalizations to multiple predictors.

2. Generalized linear models

Let Y be a vector of observations with expectation μ assuming an error distribution from the exponential family. Often μ or a function of μ is modelled by a linear combination of covariates given in a matrix X. Generalized linear models can then be written as

$$g(\mu) = \eta = X\beta, \tag{1}$$

where the *link* function $g(\cdot)$ is a monotone differentiable function, for example the identity function, the log function, or the logit function, and β is the vector with model parameters. In the current paper we will only deal with categorical covariates, often named factors. With two factors generalized linear models can be written as

$$g(\mu) = \eta = m + a_i + b_j + ab_{ij} \tag{2}$$

where the m denotes a constant, a_i $(i = 1, \ldots, I)$ and b_j $(j = 1, \ldots, J)$ are main effects for the variables A with I categories, and B with J categories, respectively, and ab_{ij} represent the interaction effect between variables A and B.

Estimation of generalized linear models is performed using a *Iteratively Re-weighted Least Squares* algorithm, which converges to the maximum likelihood estimates. In each iteration the weights and the dependent variable are updated. Afterwards the parameter vector is re-estimated. For a detailed discussion of the models and estimation procedure see McCullagh and Nelder (1989).

A problem in generalized linear models with factors is that for the interaction effect ab_{ij} the number of parameters is often rather large, i.e. $(I-1)(J-1)$, which in designs with multi-category variables can seriously effect the power. An example can be found in Hays, (1981, p 374) where the anxiety level of respondent in rooms of different size and color were measured (see Table 1).

Table 1: Anxiety measure dependent on the relation between room size and wall color.

	Room color	Red	Yellow	Green	Blue
Room size	Small	160	134	104	86
		155	139	175	71
		170	144	96	112
	Medium	175	150	83	110
		152	156	89	87
		167	159	79	100
	Large	180	170	84	105
		154	133	86	93
		141	128	83	85

An analysis of variance gives an insignificant interaction effect ($F(6, 24) = 1.87$; $p > .05$) but visualization of the data shows that within the green rooms there might be an interaction (Figure 1). The fact that we have an insignificant interaction is possibly due to the fact that 6 parameters are needed to model this interaction. If a reduction of this number of parameters is possible without changing the decomposition of sums of squares too much a significant interaction can be expected.

Reducing the number of parameters can be performed by using (generalized) biadditive models (see, for example, Denis and Gower, 1994) that decompose the interaction parameters by a singular value decomposition. Interpretation of these models is like biplot models (Gabriel, 1971; Gower and Hand, 1996)

In the current paper we will translate the generalized linear model in distance terms, and through dimension reduction we obtain models with less parameters. We think the distance parametrization has an intuitively clearer interpretation compared to the biplot interpretation.

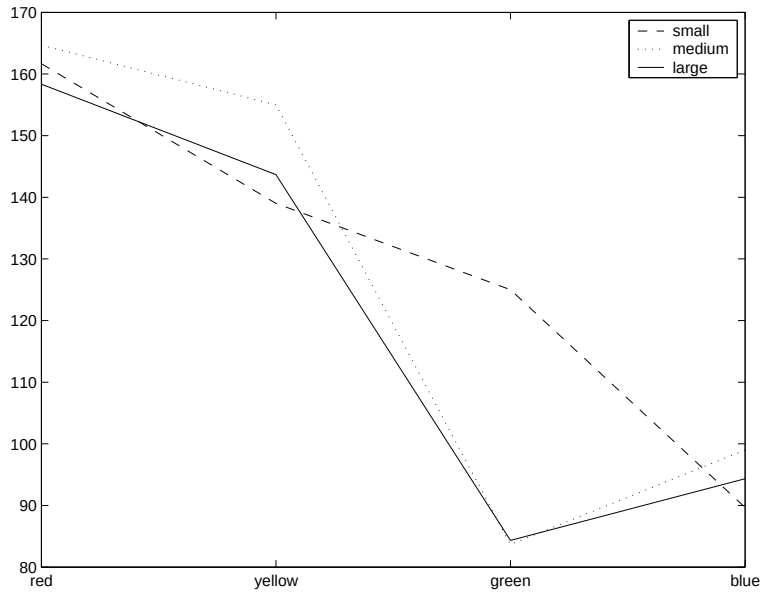

Figure 1: Visualization of anxiety data.

3. Translation to distance models

The interaction parameters can be translated into a squared two-mode distance function, i.e.

$$ab_{ij} = -d^2_{ij}(Z^A; Z^B) = -\sum_{p=1}^{P}(z^A_{ip} - z^B_{jp})^2, \tag{3}$$

where Z^A is a matrix with coordinates (z^A_{ip}) for points that represent the categories of variable A on dimension p, and Z^B is a matrix with coordinates of points that represent the categories of variable B. The dimensionality P is maximally equal to $\min(I-1,J-1)$, in which case the representation is exact, that is the predicted values of the distance model are equal to the predicted values of GLMs with interaction terms. The closer two points of different sets are in Euclidean space the higher the expected value of the variable of interest (μ).

The new model can be estimated again using *Iteratively Re-weighted Least Squares* to obtain maximum likelihood estimates. The two-dimensional solution for the anxiety data is shown in Figure 2, where uppercase letters (S, M, L) are used for the different sizes of the rooms, and lower case (r, y, g, b) for different colors. It is clear that the horizontal dimension determines the major part of the distances, and that some reduction of parameters might be possible by fitting only the first dimension.

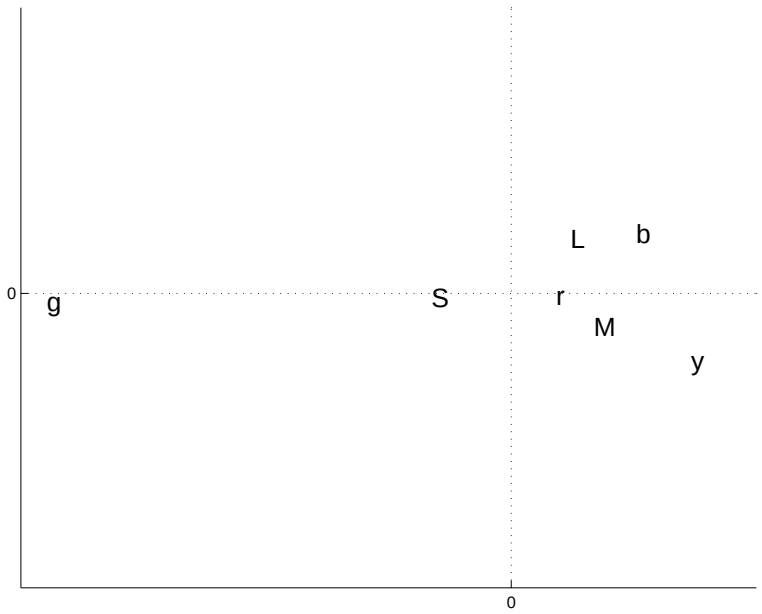

Figure 2: A distance representation of the interaction.

4. Extension to more factors

Distances represent the relationship between two categories. In a generalization to three factors, where the generalized linear model is

$$g(\mu) = \eta = m + a_i + b_j + c_k + ab_{ij} + ac_{ik} + bc_{jk}, \tag{4}$$

each of the three first order interactions could be modelled by a distance, and (4) can be rewritten as

$$g(\mu) = \eta = m + a_i + b_j + c_k - d^2_{ij}(Z^A_{AB}; Z^B_{AB}) - d^2_{ik}(Z^A_{AC}; Z^C_{AC}) - d^2_{jk}(Z^B_{BC}; Z^C_{BC}). \tag{5}$$

Coordinate matrices Z_{AB}^{A} denote the coordinates of variable A in the interaction with B (AB), and likewise for other coordinate matrices. In (5) we did not model the second order interaction. By using equality constraints such that the coordinate matrix for variable A is the same in the interaction AB as in the interaction AC we obtain a *triadic distance* (De Rooij and Heiser, 2000). Again this shows that triadic distance models do not model the second order interaction (see also De Rooij, 2002), but that it is a nice way of a simultaneous representation of all first order interactions.

REFERENCES

Denis, J.-B and Gower, J.C. (1994). Biadditive models. Biometrics, 50, 310-311.

De Rooij, M. (2002). Distance models for three-way tables and three-way association. Journal of Classification, 19, 161-178.

De Rooij M. and Heiser W.J. (2000). Triadic distance models for the analysis of asymmetric three-way proximity data, British Journal of Mathematical and Statistical Psychology, 53, 99-119

De Rooij, M and Heiser W.J. (2002, submitted) Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data.

De Rooij, M and Heiser W.J. (2002). A distance representation of the quasi-symmetry model and related distance models. In: New developments on psychometrics: proceedings of the international meeting of the psychometric society (Eds H. Yanai, A. Okada, K. Shigemasu, Y. Kano, and J. J. Meulman). (pp. 487-494) Tokyo: Springer-Verlag

Gabriel, K.R. (1971). The biplot graphical display of matrices with application to principal component analysis. Biometrika, 58, 453-467.

Gower, J.C. and Hand, D.J. (1996). Biplots. London: Chapman and Hall.

Hays, W.L. (1981) Statistics. New York: Holt, Rinehart, and Winston.

McCullagh, P. and Nelder, J.A. (1989). Generalized linear models. London: Chapman and Hall.

RÉSUMÉ

Les modèles linéaires généraux offrent un cadre unificateur commode pour l'analyse des données multivariables. Cependant, quand les covariables sont qualitatives, le nombre de paramètres nécessaires à modeler les effets d'interaction est considérable. Par conséquent, dans le cas de variables à multiples catégories, la puissance du test est souvent trop diminuée pour pouvoir trouver un effet d'interaction. Nous montrons comment les termes de distance peuvent être utilisés dans les modèles linéaires généraux afin de réduire le nombre de paramètres. Cela a pour effet non seulement une augmentation de la puissance, mais aussi une amélioration de l'interprétation de la solution. Des extensions pour le cas de plus de deux covariables sont brièvement traitées.