Generalized Bi-additive Modelling for Categorical Data

Patrick J.F. Groenen[1] and Alex J. Koning[1]

Econometric Institute,
Erasmus University Rotterdam, The Netherlands
groenen@few.eur.nl, koning@few.eur.nl

Econometric Institute Report EI 2004-05

Abstract. Generalized linear modelling (GLM) is a versatile statistical technique, which may be viewed as a generalization of well-known techniques such as least squares regression, analysis of variance, loglinear modelling, and logistic regression. In many applications, low-order interaction (such as bivariate interaction) terms are included in the model. However, as the number of categorical variables increases, the total number of low-order interactions also increases dramatically. In this paper, we propose to constrain bivariate interactions by a bi-additive model which allows a simple graphical representation in which each category of every variable is represented by a vector.

1 Introduction

Generalized linear modelling (GLM) is a versatile statistical technique, which may be viewed as a generalization of well-known techniques such as least squares regression, analysis of variance, loglinear modelling, logistic regression (Nelder and Wedderburn (1972); McCullagh and Nelder(1989)). In this paper, we limit ourselves to categorical predictor variables. Then, GLMs may consist of main effects, bivariate, and higher-order interactions. Since higher order interactions are generally difficult to interpret, we consider only bivariate interactions here. Note that as the number of categorical variables increases, the total number of bivariate interactions may also increase dramatically. In addition, if there are many categories per variable, it becomes increasingly difficult to interpret the estimated interaction parameters because there are so many of them. Our aim here is to provide a simple graphical representation to facilitate the interpretation of all bivariate interactions. To reach this goal, we impose rank restrictions on the bivariate interactions, thus leading to a bi-additive model.

For two categorical variables, van Eeuwijk (1995), De Falguerolles and Francis (1992) and Gabriel (1996) have provided algorithms for a bi-additive model within GLM. Here we propose a bi-additive model for more than two categorical predictors. To some extent, the proposed model can be seen as a generalization of multiple correspondence analysis to GLM.

2 Generalized Bi-Additive Modelling

Let us introduce some notation needed for generalized linear modelling (GLM). Let $\mathbf{y}$ be the dependent vector of n objects that needs to be predicted by m categorical variables. Also, let the categorical variable j be represented by the indicator matrix $\mathbf{G}_j$ with a zero-one variable for each category with $g_{ijk} = 1$ if observation i falls in category k of variable j and $g_{ijk} = 0$ otherwise. Let the number of categories of variable j be denoted by K_j and $\mathbf{g}'_{ij}$ be row i of $\mathbf{G}_j$.

The central idea behind GLM is that the distribution of the dependent variable belongs to a given family of distributions (popular choices are the Normal, Poisson, binomial, gamma, and inverse Gaussian families). This leaves some freedom, which allows the distribution to vary from object to object. Especially, the expectation μ_i of the dependent variable may differ from object to object, and is assumed to depend on the values taken by the predictor variables through the linear predictor η_i (examples are given below). Finally, the inverse of the link function $g(\mu_i) = \eta_i$ relates the linear predictor to μ_i. Some standard link functions are the logarithm, power, logistic, identity, and probit (McCullagh and Nelder(1989)).

A simple linear predictor may be specified by $\eta_i = c + \sum_{j=1}^{m} \mathbf{g}'_{ij}\mathbf{a}_j$, where c is an overall mean and $\mathbf{a}_j$ is a vector of main effect for variable j. However, we are interested in bivariate interactions as well, so that we need the linear predictor

$$\eta_i = c + \sum_{j=1}^{m} \mathbf{g}'_{ij}\mathbf{a}_j + \sum_{j=1}^{m}\sum_{l=j+1}^{m} \mathbf{g}'_{ij}\mathbf{B}_{jl}\mathbf{g}_{il},$$

where $\mathbf{B}_{jl}$ is the $K_j \times K_l$ matrix of bivariate interactions between variables j and l. It is easily verified that $\mathbf{g}'_{ij}\mathbf{B}_{jl}\mathbf{g}_{il}$ selects the appropriate row and column element that corresponds to the categories of the variables j and l of object i. Note that summation over $j > l$ or $j < l$ gives the same results, because one can always choose $\mathbf{B}_{jl} = \mathbf{B}'_{lj}$ so that $\mathbf{g}'_{ij}\mathbf{B}_{jl}\mathbf{g}_{il} = \mathbf{g}'_{il}\mathbf{B}_{lj}\mathbf{g}_{ij}$.

We can obtain more insight and compact notation by joining the effects of all variables. Let $\mathbf{G}$ be the *super* indicator matrix with all m variables next to each other, that is, $\mathbf{G} = [\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_m]$, $\mathbf{g}'_i$ be row i of $\mathbf{G}$, and $\mathbf{a}$ be the vector of all main effects. Finally, all bivariate interaction effects are joined into the symmetric partitioned block matrix

$$\mathbf{B} = \begin{bmatrix} \mathbf{0} & \mathbf{B}'_{12} & \ldots & \mathbf{B}'_{1m} \\ \mathbf{B}_{12} & \mathbf{0} & \ldots & \mathbf{B}'_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_{1m} & \mathbf{B}_{2m} & \ldots & \mathbf{0} \end{bmatrix}.$$

The diagonal blocks are zero because $j \neq l$. Then, we may write

$$\eta_i = c + \sum_{j=1}^{m} \mathbf{g}'_{ij}\mathbf{a}_j + \sum_{j=1}^{m}\sum_{l=j+1}^{m} \mathbf{g}'_{ij}\mathbf{B}_{jl}\mathbf{g}_{il} = c + \mathbf{g}'_i\mathbf{a} + \tfrac{1}{2}\mathbf{g}'_i\mathbf{B}\mathbf{g}_i. \qquad (1)$$

The basic idea of this paper is to impose constraints on the interaction terms $\mathbf{B}_{jl}$. The type of constraint that we consider is the one of common rank-reduction, that is, to require that

$$\mathbf{B}_{jl} = \mathbf{Y}_j \mathbf{Y}_l' \tag{2}$$

with $\mathbf{Y}_j$ a $K_j \times p$ matrix and $\mathbf{Y}_l$ a $K_l \times p$ matrix being of rank not higher than $p > 0$. Such rank constraints are similar to the ones used in multiple correspondence analysis, joint correspondence analysis, or homogeneity analysis. This rank constrained bi-additive model can be expressed as

$$\eta_i = c + \sum_{j=1}^{m} \mathbf{g}_{ij}' \mathbf{a}_j + \sum_{j=1}^{m} \sum_{l=j+1}^{m} \mathbf{g}_{ij}' \mathbf{Y}_j \mathbf{Y}_l' \mathbf{g}_{il}. \tag{3}$$

To avoid that $\mathbf{Y}_j$ also estimates main effects, we impose the restriction that $\mathbf{Y}_j$ has column mean zero. This restriction also implies that $\mathbf{B}_{jl} = \mathbf{Y}_j \mathbf{Y}_l'$ has zero row and column mean, which is a restriction that is usually imposed on bivariate interactions to ensure uniqueness. We shall refer to the matrix $\mathbf{Y}$ as the *interaction generating* matrix, and to the k^{th} column of $\mathbf{Y}_j$ as the k^{th} dimension of interaction generators belonging to the categorical variable j. To fit this model, we have developed a prototype in MatLab that optimizes the likelihood by iterated weighted least squares and iterative majorization.

Note that standard likelihood theory applies to model (3), and hence we may employ the likelihood ratio test to determine the rank p. From this perspective, it is relevant to know the degrees of freedom associated to the rank p model. Observe that the parameters in the rank p model are the constant term c, the main effect vectors $\mathbf{a}_j$ and the elements of the interaction generating matrix $\mathbf{Y}$. Hence, the number of parameters in this model equals

$$1 + \sum_{j=1}^{m} K_j + p \sum_{j=1}^{m} K_j.$$

However, we have also imposed several restrictions on these parameters. Each of the m main effect vectors and each of the p dimensions of the m interaction generators should add up to zero. Moreover, the interaction generating matrix $\mathbf{Y}$ can be rotated by any orthonormal rotation matrix $\mathbf{T}$ (with $\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}$) without affecting $\mathbf{B}_{jl}$ since $\mathbf{B}_{jl} = \mathbf{Y}_j \mathbf{Y}_l' = \mathbf{Y}_j \mathbf{T}\mathbf{T}'\mathbf{Y}_l'$ for all j and l. Therefore, without loss of generality, we impose the restriction that $\sum_j \mathbf{Y}_j$ is orthogonal thereby making the $\mathbf{Y}_j$'s unique. This restriction implies fixing $p(p-1)/2$ of the elements of the the $\mathbf{Y}_j$'s. Summarizing, the number of restrictions in the rank p model is equal to

$$m + mp + \frac{p(p-1)}{2}$$

Table 1. Frequencies of occurrence of lung cancer for (non) smokers in eight Chinese cities (taken from Liu (1992), see also Tabel 3.3 in Agresti (1996), p. 60).

		City								
Lung cancer	Smoker	Beijing	Shanghai	Shenyang	Nanjing	Harbin	Zengzhou	Taiyuan	Nanchang	Total
yes	yes	126	908	913	235	402	182	60	104	2930
yes	no	35	497	336	58	121	72	11	21	1121
no	yes	100	688	747	172	308	156	99	89	2359
no	no	61	807	598	121	215	98	43	36	1979
	Total	322	2900	2594	586	1046	508	213	250	8419

The difference in number of parameters and number of restrictions

$$df_{\text{model}} = \left[1 + \sum_{j=1}^{m} K_j + p \sum_{j=1}^{m} K_j\right] - \left[m + mp + \frac{p(p-1)}{2}\right]$$
$$= 1 + (1+p)\sum_{j=1}^{m}(K_j - 1) - \frac{p(p-1)}{2}.$$

is the degrees of freedom associated to the rank p model. The residual degrees of freedom df_{res} is obtained by subtracting df_{model} from the degrees of freedom n (the number of objects) associated to the so-called saturated model.

3 Application: lung cancer in China

Lung cancer is one of the leading causes of death in the People's Republic of China. It has been estimated that in the year 2025, the number of new cancer cases in China will reach three million, of which two million are associated with smoking, and the remaining one million attributable to other causes (Peto (1994); Peto, Chen and Boreham (1996)).

A number of epidemiological studies have investigated the association between lung cancer and smoking in China. In Liu (1992), a meta-analysis is presented of eight case-control studies conducted in Beijing, Shanghai, Shenyang, Nanjing, Harbin, Zhengzhou, Taiyuan, and Nanchang. Table 1 cross-classifies a total of 4081 lung cancer cases and 4338 controls according to smoking behaviour and city. In this paragraph, we investigate the relation between smoking and lung cancer in China by applying the generalized biadditive model introduced in the previous section to the data in Table 1. Note that $m = 3$, $K_1 = 2 - 1$, $K_2 = 2 - 1$, and $K_3 - 1$, and hence $\sum_j^m (K_j - 1) = 1 + 1 + 7 = 9$.

Table 2. Summary of fit for different models, using the Poisson family of distributions and the log link function on the data of Table 1.

Model	Deviance	df_{res}	p
Main effects	457.1	22	.0000
Bi-additive interaction model, rank 1	35.7	13	.0007
Bi-additive interaction model, rank 2	5.4	5	.3690

Table 2 summarizes the fit of three models when using the Poisson family of distributions and a log link function. The main effects model may be regarded as a special case of model (3) with rank $p = 0$. Next, adding a first dimension of bilinear interactions to each variable yields model (3) with rank $p = 1$. Finally, adding a second dimension of bilinear interactions to each variable yields model (3) with rank $p = 2$. Adding further dimensions would lead to models with degrees of freedom exceeding 32, the degrees of freedom of the saturated model; such models are unidentified, and have no statistical use.

Note that the three models in Table 2 are nested, and thus may be compared as usual by means of the likelihood ratio test (that is, by relating differences in deviance to the chi-square distribution corresponding to the difference in degrees of freedom). Obviously, the rank 2 model is to be preferred, as it is the only model that fits the data.

Table 3. Estimation results obtained by applying the rank 2 bi-additive decomposition model to the Chinese lung cancer-smoking data.

		Main	Interaction generators	
Variable	Category	effects	Dim 1	Dim 2
Overall mean	c	5.005		
Lung cancer	Yes	0.410	-0.494	-0.169
	No	-0.410	0.494	0.169
Smoking	Yes	-0.124	-0.055	-0.413
	No	0.124	0.055	0.413
City	Beijing	-0.727	-0.015	-0.052
	Shanghai	1.557	0.348	-0.133
	Shenyang	1.406	0.124	-0.053
	Nanjing	-0.121	0.004	-0.057
	Harbin	0.472	0.047	-0.071
	Zengzhou	-0.239	0.081	-0.080
	Taiyuan	-1.270	-0.342	0.421
	Nanchang	-1.078	-0.248	0.025

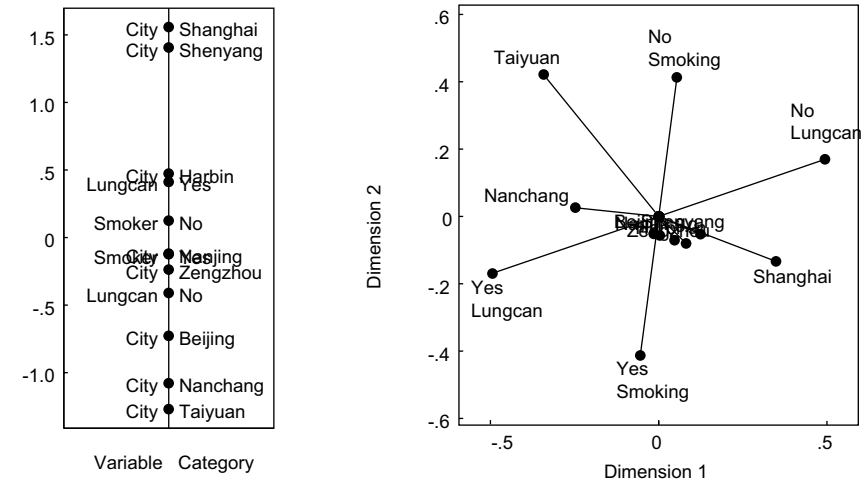

Fig. 1. Representation of the bi-additive decomposition model of the Chinese lung cancer-smoking data. Panel (a) shows the main effects and Panel (b) the decomposition of the interactions in two dimensions.

Applying the rank 2 bi-additive decomposition model to the Chinese lung cancer-smoking data yields the estimation results listed in Table 3. These results are visualized in Figure 1, where Panel (a) shows the main effects and Panel (b) gives the vectors for the interaction generators $\mathbf{Y}_j$.

Figure 1 may be interpreted as follows. To see the size of the interaction effect, project the vector of one category onto a category of a different variable. For example, Taiyuan is characterized by having more nonsmokers than smokers, when corrected for the main effects of smoking. The reason is the nonsmoking vector projects highly on the vector of Taiyuan. Long vectors lead to longer projections and thus to a stronger interaction effect. Conversely, short vectors have short projections, indicating a small interaction effect. Therefore, the cities Beijing, Harbin, Nanjing, Zengzhou, and Shenyang will only have small interaction effects with the other variables.

In Figure 1, three Chinese cities relatively stand out: Nanchang, Taiyuan and Shanghai. Nanchang was badly battered after the Communist takeover, but reinvented itself as a centre of modern steel and chemical industry (Leffman, Lewis and Atiyah (2000), p. 52). Taiyuan's extensive coal mines were constructed by the Japanese in 1940; serious industrialization began after the Communist takeover and today it is the factories that dominate, relentlessly processing the region's coal and mineral deposits (Leffman, et al. (2000), p. 225). After forty years of stagnation, Shanghai seems certain to recapture its position as East Asia's leading business city, a status it last held before World War II (Leffman, et al. (2000), p. 337).

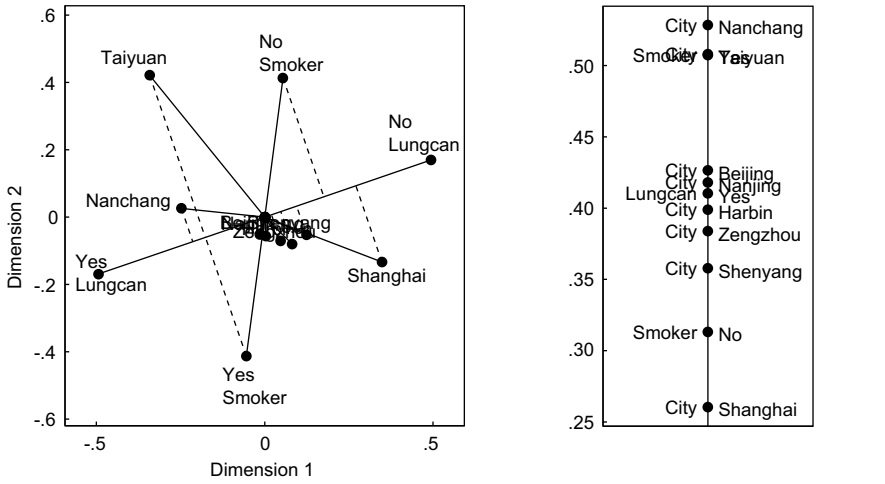

Fig. 2. Interaction terms conditioned on category 'yes' of the variable 'lung cancer'. Panel (a) shows the projections of all other catgories on category 'yes' of the variable 'lung cancer' and Panel (b) the values of the interactions.

Recall that Table 1 was compiled by Liu (1992) from the results of eight different case-control studies investigating lung cancer in China. In this respect, the interactions between the variable Lung Cancer on one hand and the variables Smoking and City on the other hand are of primary interest. Thus, to gain insight into the causes of lung cancer in China, we should project each category of the variables Smoking and City onto the category Yes of the variable Lung Cancer, as indicated in Panel (a) of Figure 2. Panel (b) visualizes the interaction values according to the bi-additive interaction model for the categories of Smoking and City conditioned on having lung cancer.

As to be expected, smoking is clearly a risk factor for obtaining lung cancer. However, Panel (b) of Figure 2 seems to suggest that there is also an environmental risk factor, as the interaction between the industrial cities of Nanchang and Taiyuan on one hand and the presence of lung cancer on the other hand is much higher than the interaction between the leading business city Shanghai and the presence of lung cancer.

4 Conclusions

We have proposed a new model for representing two-way interactions for a GLM with more than two categorical predictor variables, where we constrain

the two-way interactions to have reduced rank. Each category is represented by a vector in a plot. The interaction effect between two categories of different predictor variables is obtained by projecting the vector of one category onto the vector of another. Categories of the same variable should not be compared within the plot, but only by looking at the main effects. The main advantage of the bi-additive interaction model is that interactions between several variables can be visualized in a relatively simple display, even when the total number of categories is large.

References

AGRESTI, A. (1996): *An introduction to categorical data analysis*. New York, Wiley.

DE FALGUEROLLES, A. and FRANCIS, B. (1992): Algorithmic approaches for fitting bilinear models. In: Y. Dodge and J. Whittaker (Eds.): *Compstat 1992*. Heidelberg, Physica-Verlag, 77–82.

EEUWIJK, F. A. (1995): Multiplicative interaction in generalized linear models. *Biometrics, 85, 1017-1032*.

GABRIEL, K. R. (1996): Generalised bilinear regression. *Biometrika, 85, 689-700*.

LEFFMAN, D., LEWIS, S. and ATIYAH, J. (2000): *China*. London, Rough Guides Ltd.

LIU, Z. (1992): Smoking and lung cancer in China. *Internal Journal of Epidemiology, 21, 197-201*.

MCCULLAGH, P. and NELDER, J. A. (1989): *Generalized linear models*. London, Chapman and Hall.

NELDER, J. A. and WEDDERBURN, R. W. M. (1972): Generalized linear models. *Journal of the Royal Statistical Society A, 135, 370-384*.

PETO R (1994): Smoking and death - the past 40 years and the next 40. *British Medical Journal 309, 937–939*.

PETO, R., CHEN, Z. M. and BOREHAM. J. (1996): Tobacco - The growing epidemic in China. *Journal of the American Medical Association, 275, 1683–1684*.