Optimal Scaling of Interaction Effects in Generalized Linear Models*

Joost van Rosmalen, Alex J. Koning, and Patrick J.F. Groenen
Econometric Institute
Erasmus University Rotterdam

October 2007

Econometric Institute Report EI 2007-44

Abstract

Multiplicative interaction models, such as Goodman's RC(M) association models, can be a useful tool for analyzing the content of interaction effects. However, most models for interaction effects are only suitable for data sets with two or three predictor variables. Here, we discuss an optimal scaling model for analyzing the content of interaction effects in generalized linear models with any number of categorical predictor variables. This model, which we call the optimal scaling of interactions (OSI) model, is a parsimonious, one-dimensional multiplicative interaction model. We discuss how the model can be used to visually interpret the interaction effects. Two empirical data sets are used to show how the results of the model can be applied and interpreted. Finally, several multidimensional extensions of the one-dimensional model are explored.

1 Introduction

The analysis of data sets with categorical variables often requires studying interaction effects between these variables. If the relationship between the response variable and the predictor variables is linear, the interaction effects can be studied using analysis of variance (ANOVA). If this relationship is nonlinear, models from the class of generalized

*Correspondence concerning this article should be addressed to Joost van Rosmalen, Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. E-mail: vanrosmalen@few.eur.nl

linear models (GLMs) are often used. Generalized linear models, which are extensions of the general linear model, can be used to model various kinds of relationships between variables and can account for nonnormality and nonlinearity. GLMs have been thoroughly described in Nelder and Wedderburn (1972) and McCullagh and Nelder (1989). Including all two-way interactions in a GLM may often require the estimation of a large number of parameters, especially if there are many categorical variables and if they have many levels. Because of the large number of parameters, the estimated individual interaction effects are often not interpreted, and only their combined effect is tested for significance.

Models for representing interaction effects parsimoniously have been proposed before, especially for the case of two categorical predictor variables. For example, Goodman (1981) proposed row-column (RC(M)) association models for the analysis of two-way contingency tables. RC(M) association models can be considered as a special case of generalized additive main effects and multiplicative interaction (GAMMI) models, which are mainly used in agricultural science (see, for example, Van Eeuwijk, 1995, 1996). Similar models were proposed by Gabriel (1998). Algorithmic approaches for these kinds of models were discussed by De Falguerolles and Francis (1992). These types of models often use *biplots* (see, for example, Gower & Hand, 1996) to represent interaction effects between two variables, by plotting the categories of both variables in a two-dimensional space. Specialized models for the case of three categorical predictor variables also exist (see, for example, Anderson, 1996; Choulakian, 1996; Siciliano & Mooijaart, 1997; Wong, 2001).

For the case of more than three predictor variables, Groenen and Koning (2006) proposed the interaction decomposition model. They sketched an outline of an algorithm for parameter estimation in this model and gave graphical representations of their results. For log-linear analysis (a special case of generalized linear modeling) with more than three variables, a variety of models were proposed by Anderson and Vermunt (2000). In their article, the interaction effects are parsimoniously modeled by assuming the presence of latent variables.

In this article, we use the methodology of *optimal scaling* for modeling interaction effects parsimoniously. Optimal scaling (see, for example, Young, 1981; Gifi, 1990; Linting, Meulman, Groenen, & Van der Kooij, 2007) is a methodology originating from psychometrics that assigns numeric values to categorical variables in an optimal way. Gifi (1990) discusses a host of multivariate analysis techniques (multiple correspondence analysis, nonlinear principal components analysis, generalized nonlinear canonical correlation analysis, etc.) all having in common that the variables are categorical and that some optimal recoding is being done. That is, the categories of the original categorical variables are replaced by their so-called category quantifications, and from then on the variables are considered to be quantitative variables. The word optimal refers to the fact that these category quantifications are chosen in such a way that they help optimize the criterion. Optimal scaling has also been applied in a regression context, with techniques such as

MONANOVA (Kruskal, 1965), ADDALS (De Leeuw, Young, & Takane, 1976), MORALS (Young, De Leeuw, & Takane, 1976), ACE (Breiman & Friedman, 1985), and generalized additive models (Hastie & Tibshirani, 1990).

We describe a model for representing interaction effects in generalized linear models with any number of categorical predictor variables in a clear and parsimonious way. It is assumed that only main effects and two-way interaction effects are empirically relevant, and we do not account for higher-way interaction effects, as they are often not required in empirical applications. The main assumption of our model is that interactions between categorical predictor variables can be modeled using continuous predictor variables on which we have partial knowledge. This assumption leads to a model in which the estimated parameters may be interpreted in terms of an optimal scaling of the categorical predictor variables. Because of this assumption, we refer to our model as the *optimal scaling of interactions* (OSI) model. The OSI model requires a number of parameters that is only linear in the total number of categories of the categorical variables and quadratic in the number of variables. By contrast, a standard two-way interaction model requires a number of parameters that is quadratic in the total number of categories of the variables. With our model, we construct one-dimensional graphical representations of the interaction effects, which can help interpret these effects. As a one-dimensional model may be restrictive in some cases, multidimensional extensions of our model are also explored.

The outline of this article is as follows. In the next section, we introduce some notation and our optimal scaling of interactions model. Section 3 describes the application of our model to two empirical data sets. In Section 4, we discuss multidimensional generalizations of our model. The final section summarizes our findings.

2 Optimal Scaling of Interactions Model

The model we propose is based on generalized linear modeling (see, for example, Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989). The observations $y_i, i = 1 \ldots n$ are assumed to be independently distributed with $\mathrm{E}(y_i) = \mu_i$. Each y_i has a distribution in the exponential family, with probability density function given by

$$f_Y(y; \theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}, \tag{1}$$

where $a(.)$, $b(.)$, and $c(.)$ are given functions. The exponential family includes the normal, Poisson, binomial, gamma, and inverse Gaussian distributions. The systematic part of a generalized linear model consists of a predictor η_i, which typically is a linear function of the predictor variables and the parameters. A *link function* $h(.)$ relates the linear predictor η_i to the response variable according to

$$\eta_i = h(\mu_i). \tag{2}$$

Common link functions are the identity, inverse, logarithm, and logit functions. In practice, one often uses *canonical links*, such as a logarithm link in combination with a Poisson error distribution. The canonical links can be derived from the theory of sufficient statistics.

In this paper, we aim to model two-way interaction effects using a generalized linear modeling framework. Suppose continuous predictor variables $\mathbf{x}_j$, $j = 1, \ldots, m$ are known. Then, the main effects and the interaction effects of these variables can be modeled according to

$$\eta_i = c + \sum_{j=1}^{m} b_j x_{ij} + \sum_{j=1}^{m-1} \sum_{l=j+1}^{m} w_{jl} s_{jl} x_{ij} x_{il}, \tag{3}$$

where c is a constant term, b_j is the main effect of variable $\mathbf{x}_j$, and s_{jl} is the interaction effect of variables $\mathbf{x}_j$ and $\mathbf{x}_l$. The $m \times m$ upper-triangular matrix $\mathbf{W} = (w_{jl})$ specifies which interaction effects are to be estimated in the GLM, with $w_{jl} = 1$ if the interaction between predictor variables j and l is taken into account and $w_{jl} = 0$ otherwise. The diagonal elements of $\mathbf{W}$ are not used, as these elements refer to main effects that are already modeled by the second term in (3).

Here, we restrict ourselves to cases in which all predictor variables are categorical instead of continuous, so that (3) cannot be used directly. The central assumption of our model is that interaction effects between the categorical predictor variables can be modeled in approximately the same way as interaction effects between continuous predictor variables. To do so, we apply the idea of *optimal scaling* (see, for example, Gifi, 1990) to the categorical predictor variables for modeling their interaction effects, hence the name optimal scaling of interactions (OSI) model.

To be able to introduce optimal scaling in model (3), we need some notation. Let there be m categorical predictor variables with each variable having k_j categories. To code the categorical predictor variables we use indicator matrices $\mathbf{G}_j$ with rows $\mathbf{g}_{ij}$ of length k_j; element l of $\mathbf{g}_{ij}$ has value 1 if observation y_i belongs to category l of predictor variable j and 0 otherwise. In the OSI model, we use separate optimally scaled variables for the main effects and the interaction effects, so that

$$\eta_i = c + \sum_{j=1}^{m} b_j r_{ij} + \sum_{j=1}^{m-1} \sum_{l=j+1}^{m} w_{jl} s_{jl} q_{ij} q_{il}, \tag{4}$$

where $\mathbf{r}_j$ is the optimally scaled variable for the main effect of variable j, and $\mathbf{q}_j$ is the optimally scaled variable that is used for the interaction effects of variable j. In principle, one could also use the same optimally scaled variables for both the main effects and the interaction effects, so that $\mathbf{r}_j = \mathbf{q}_j$. However, we find this approach too restrictive, and we therefore do not explore it here.

The values of the continuous, optimally scaled predictor variables $\mathbf{r}_j$ and $\mathbf{q}_j$ are not known in our model and need to be estimated. The $\mathbf{r}_j$s are related to the categorical

predictor variables according to $\mathbf{r}_j = \mathbf{G}_j\mathbf{a}_j$, where $\mathbf{a}_j$ is a $k_j \times 1$ parameter vector that contains the category quantifications for the main effects of variable j. The $\mathbf{q}_j$s are constructed similarly as $\mathbf{q}_j = \mathbf{G}_j\mathbf{y}_j$, with $\mathbf{y}_j$ a $k_j \times 1$ parameter vector that contains the category quantifications for the interaction effects of variable j. Instead of (4), the OSI model may also be described as

$$\eta_i = c + \sum_{j=1}^{m} b_j \mathbf{g}'_{ij}\mathbf{a}_j + \sum_{j=1}^{m-1}\sum_{l=j+1}^{m} w_{jl}s_{jl}\mathbf{g}'_{ij}\mathbf{y}_j\mathbf{y}'_l\mathbf{g}_{il}. \tag{5}$$

In this way, the main effects appear in the same manner as in an ordinary GLM with categorical predictor variables. For the interaction effects, the OSI model uses a multiplicative specification that is relatively parsimonious. The parameter vector $\mathbf{y}_j$ reflects the content of the interaction effects of variable j. The goal of the parameter s_{jl} is to estimate the size of the interaction effect between variables j and l. Therefore, we refer to s_{jl} as a *scaling factor*.

Once the category quantifications $\mathbf{a}_j$ and $\mathbf{y}_j$ are estimated and hence are known, the optimally scaled variables can be treated as ordinary continuous variables. Then, the parameters b_j and s_{jl} can be computed using the ordinary GLM in (3). Because the $\mathbf{q}_j$s are not known, and the way the interactions appear in (4), the predictor η_i is a nonlinear function of the model parameters. Therefore, the OSI model is not an ordinary GLM. As the $\mathbf{q}_j$s are restricted by $\mathbf{q}_j = \mathbf{G}_j\mathbf{y}_j$, and $\mathbf{y}_j$ can be estimated from the data, the OSI model can be seen as a GLM with optimal scaling of the categorical predictor variables.

Several parameter constraints, including location and scale constraints, are required for model identification. We use the following parameter constraints, which originate from the optimal scaling methodology and differ from the constraints typically used in multiplicative interaction models. We impose that the optimally scaled variables $\mathbf{r}_j$ and $\mathbf{q}_j$ have mean zero and variance one, which is a customary constraint in optimal scaling. For the interaction effects, this results in the location constraints $\sum_{i=1}^{n} \mathbf{g}'_{ij}\mathbf{y}_j = 0$ and the scale constraints $\sum_{i=1}^{n} \mathbf{g}'_{ij}\mathbf{y}_j\mathbf{y}'_j\mathbf{g}_{ij} = n$. For the main effects, we impose that $\sum_{i=1}^{n} \mathbf{g}'_{ij}\mathbf{a}_j = 0$ and $\sum_{i=1}^{n} \mathbf{g}'_{ij}\mathbf{a}_j\mathbf{a}'_j\mathbf{g}_{ij} = n$. In addition, the value of the scaling factor s_{jl} cannot be estimated if $w_{jl} = 0$; therefore, we set $s_{jl} = 0$ whenever $w_{jl} = 0$. Finally, simultaneously changing the signs of the elements of $\mathbf{y}_j$ and s_{jl} for all l does not affect the predictor η_i. To improve the interpretability of the model parameters, we simultaneously reflect the $\mathbf{y}_j$s and the scaling factors s_{jl} in such a way, that the sum of the estimated scaling factors is maximized. To do so, each of the 2^m possible combinations of reflections of the $\mathbf{y}_j$s is considered, and the combination that maximizes $\sum_{j=1}^{m-1}\sum_{l=j+1}^{m} s_{jl}$ is used to interpret the model's results.

Additional parameter constraints may be required if few observations are available, or if $w_{jl} = 0$ for many values of j and l. Whether such additional constraints are necessary can be determined empirically, for example, by checking whether the estimated parameters are unique maximizers of the log-likelihood function. This can be done by estimating the

model parameters multiple times using randomly chosen starting values; if no additional parameter constraints are necessary, the estimated parameters must be the same in every instance.

The categorical predictor variables can have either a nominal or an ordinal measurement level. For ordinal predictor variables, it is possible to impose their ordering on $\mathbf{y}_j$. However, imposing such ordinality constraints may not be appropriate, as the interaction effects can reflect nonmonotonic relations between the predictor variables and the response variable. Therefore, we do not impose the ordering of ordinal predictor variables on the model parameters. The OSI model can also be extended to include continuous predictor variables, for example, by modeling the $\mathbf{y}_j$s as (spline) transformations of these continuous variables. In that case, one again needs to consider whether such transformations need to be monotonic. More information on splines and other nonlinear transformations is given in Gifi (1990).

As the OSI model is not an ordinary GLM, a special algorithm for parameter estimation is needed. In our implementation, the parameters are estimated by maximizing the log-likelihood function using the BFGS quasi-Newton optimization routine in the MATLAB Optimization Toolbox (version 3.0.4). Standard errors of the estimated parameters are computed using the negative inverse of Hessian (the matrix of second-order partial derivatives of the log-likelihood function), evaluated at the final parameter estimates.

The OSI model has several relationships with existing models for interaction effects. A standard GLM with two-way interaction effects can be described as

$$\eta_i = c + \sum_{j=1}^{m} b_j \mathbf{g}'_{ij} \mathbf{a}_j + \sum_{j=1}^{m-1} \sum_{l=j+1}^{m} w_{jl} \mathbf{g}'_{ij} \mathbf{B}_{jl} \mathbf{g}_{il}, \tag{6}$$

where $\mathbf{B}_{jl}$ is a $k_j \times k_l$ parameter matrix of interaction effects between variables j and l. The OSI model can be obtained from the full two-way interaction GLM by imposing that each interaction matrix $\mathbf{B}_{jl}$ equals a matrix $\bar{\mathbf{B}}_{jl}$ with

$$\bar{\mathbf{B}}_{jl} = s_{jl} \mathbf{y}_j \mathbf{y}'_l. \tag{7}$$

Thus, the OSI model implicitly approximates each matrix of interaction effects $\mathbf{B}_{jl}$ by a matrix of rank one.

The OSI model also resembles a few multiplicative interaction models that have been proposed previously. If there are only two categorical predictor variables, the OSI model is equivalent with the generalized additive main effects and multiplicative interaction models discussed by Van Eeuwijk (1995, 1996), which are a generalization of the RC association models discussed by Goodman (1981). For the case of log-linear analysis (that is, generalized linear modeling with link function $\eta_i = \log(\mu_i)$ and a Poisson probability distribution), the OSI model is equivalent with equation (20) of Anderson and Vermunt (2000); they interpreted the $\mathbf{y}_j$s as latent variables.

3 Empirical Applications

To determine how the OSI model performs in practice, we apply it to two empirical data sets. We also compare its usefulness with other models and show how the interaction effects can be visually represented and interpreted.

3.1 STAR Data Set

The first data set we use is based on the STAR data set, which can be found in the "Ecdat" package in the R programming language. This data set contains the results of 5,748 Tennessee primary school students on tests of math and reading skills. The data were collected as a part of the Student/Teacher Achievement Ratio (STAR) project (see `http://www.heros-inc.org/star.htm` for additional information). This project investigates the effects of class size on the performance of primary school students. Each student was assigned to either a small class (13 to 17 students per teacher), a regular size class (22 to 25 students per teacher), or a regular-with-aide class (22 to 25 students with a full-time teacher's aide). The data set also contains personal background characteristics of the students, the level of experience of the teacher, and the school at which the test was taken.

The aim is to explain the results of the math skills test using six categorical predictor variables that are also in the STAR data set. We will focus on the two-way interaction effects of these predictor variables.

- Class size: A categorical predictor with levels "small", "regular", and "regular with aide"
- Teaching experience: A categorical predictor with levels "$<$ 5 years", "5-9 years", "10-14 years", "15-19 years", and "$>$ 19 years"
- Sex: A categorical predictor with levels "boy" and "girl"
- Race: A categorical predictor with levels "white" and "black"
- Free lunch: A categorical predictor with levels "Free lunch" and "No free lunch"
- School id: A categorical predictor with 79 levels, which identifies the school at which the test was taken.

The effects of the variables Sex and Race are combined, so that a new predictor variable (denoted by "Sex, race") with four levels is obtained. We do not take the interaction effects of "School id" into account in our analysis, as the levels of this factor have no meaning to the reader; only the main effects of "School id" are modeled. In addition, "School ID" is modeled as a random factor (the schools used in the study are a sample of the

Table 1: Residual degrees of freedom, log-likelihood values, and values of the Akaike information criterion of various models for the STAR data set

Model	Residual df	Log-likelihood	AIC
GLM with main effects only	5659	-29,483.34	59,144.68
OSI model	5647	-29,450.89	59,103.78
GLM with all two-way categorical interactions	5624	-29,433.69	59,115.38

Table 2: ANOVA table with all two-way interactions for STAR data set

Source	Type III sum sq.	d.f.	Mean sq.	F	p-value	Partial η^2
School ID	2,230,156	78	28,591.7	17.04	0.000	0.191
Class size	50,189	2	25,094.3	14.96	0.000	0.005
Teaching exp	16,425	4	4,106.3	2.45	0.044	0.002
Sex, race	80,821	3	26,940.3	16.06	0.000	0.008
Free lunch	158,946	1	158,945.8	11.53	0.001	0.017
Class size × Teaching exp	71,125	8	8,890.6	5.30	0.000	0.007
Class size × Sex, race	16,631	6	2,771.7	1.65	0.129	0.002
Class size × Free lunch	209	2	104.6	0.06	0.940	0.000
Teaching exp × Sex, race	47,845	12	3,987.1	2.38	0.005	0.005
Teaching exp × Free lunch	1,737	4	434.3	0.26	0.904	0.000
Sex, race × Free lunch	23,268	3	7,755.9	4.62	0.003	0.002
Error	9,436,513	5,624	1,677.9			
Total	13,115,339	5,747				

population of schools in Tennessee), whereas all other predictor variables are fixed factors. The math score is a continuous variable and ranges from 320 (worst performance) to 626 (best performance), with an average score of 486. As the response variable is continuous and approximately normally distributed, generalized linear modeling with an identity link and a normal error distribution (which is analysis of variance) seems most appropriate.

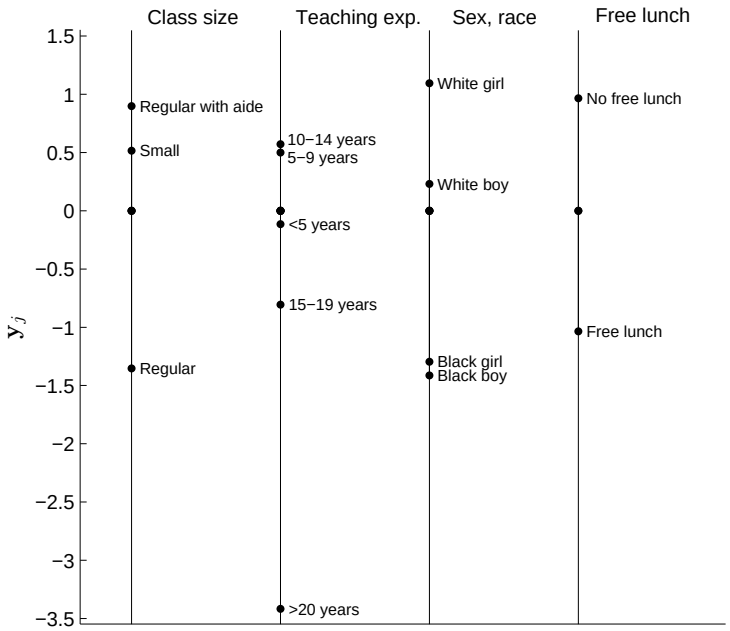
Table 1 contains the results for the one-dimensional OSI model, a standard GLM with only main effects, and a standard GLM with full two-way interaction effects. From this table, we can observe that the OSI model accounts for most of the interaction effects, as the difference in log-likelihood between the OSI model and a full two-way interaction model is relatively small. Therefore, a full two-way interaction model does not seem necessary. In addition, the OSI model has a lower value for the Akaike information criterion (AIC) than either a model with only main effects or a full two-way interaction model. Table 2 gives the ANOVA table for the model with full two-way interactions. This table shows that the interactions of "Teaching experience" with "Class size" and "Sex, race" are relatively large and statistically significant at the 5% level.

An important objective of the OSI model is interpreting the two-way interactions. Therefore, we focus on the interaction effects in this paper and do not report the estimated main effects $\mathbf{a}_j$ and b_j. To interpret the interactions, we can construct visualizations of the estimated model parameters, which should provide an understanding of the interaction effects. The $\mathbf{y}_j$s may be graphically represented using an *interaction plot* that shows the elements of these parameter vectors. In such a figure, each level of each categorical variable is represented by a single parameter. Such an interaction plot is similar to a one-dimensional version of a biplot (see Gower & Hand, 1996) used in, for example, principal components analysis and correspondence analysis. The scaling factors s_{jl} constitute a matrix, which can be shown in a simple table.

Figure 1 shows the interaction plot and the estimated scaling factors of the OSI model for the STAR data set. To improve the interpretability of this figure, the estimated category quantifications are shown in a separate axis for each predictor variable. The estimated standard errors of the scaling factors are shown in parentheses in Figure 1. Using these results, we can interpret the content of the interaction effects as follows. The values of the scaling factors s_{jl} determine the relative importance of the interaction effects; large absolute values of these scaling factors correspond to large interaction effects. The scaling factors in Figure 1 show that the interaction effects between "Teaching experience" and "Class size" ($s_{12} = 3.861$) and between "Teaching experience" and "Sex, race" ($s_{23} = 2.265$) are relatively large and statistically significant. The content of relevant or statistically significant interaction terms can be determined using the $\mathbf{y}_j$s. If the corresponding scaling factor is positive, pairs of categories of different variables with quantifications $\mathbf{y}_j$ of the same sign have positive estimated interaction effects. For the interaction between "Class size" and "Teaching experience", $\mathbf{y}_1$ and $\mathbf{y}_2$ show that there are fairly high positive estimated interaction effects between a high level of teaching experience and regular size classes; this is also true for a high level of teaching experience in combination with black students. Therefore, we can conclude that teachers with more than 15 years of experience appear more capable of handling regular size classes and classes with black students than other teachers. It seems best to assign small classes and classes with few black students to less experienced teachers. There is also a strong interaction between "Sex, race" and "Free lunch". The variable "Free lunch" is mainly determined by the household income of the student. There appear to be more severe negative effects of having a low household income on math performance for white students than for black students.

3.2 General Social Survey Data

The second data set used in this paper is based on the 1994 General Social Survey (Davis & Smith, 1996). This data set contains the responses of 899 respondents on four questions on attitudes of the labor roles of women and was also used in Anderson and Vermunt



	Class size	Teaching exp.	Sex, race	Free lunch
Class size	-	3.861 (0.711)	-0.702 (0.761)	0.124 (0.624)
Teaching exp.	-	-	2.265 (0.835)	-0.262 (0.647)
Sex, race	-	-	-	2.593 (0.765)
Free lunch	-	-	-	-

Figure 1: Interaction plot and corresponding scaling factors s_{jl} of the OSI model for the STAR data set. The estimated standard errors of the scaling factors are shown in parentheses.

Table 3: Residual degrees of freedom, deviance values, p-values, and values of the AIC of various models for the General Social Survey data set

Model	Residual df	Deviance	p-value	AIC
GLM with main effects only	187	1,063.25	0.00	1,089,25
OSI model	175	277.85	0.00	327.85
GLM with all two-way categorical interactions	136	117.93	0.87	245.93

(2000). The four questions were as follows:

1. Woman earning money: "Do you approve or disapprove of a married woman earning money in business or industry if she has a husband capable of supporting her?" (approve, disapprove).

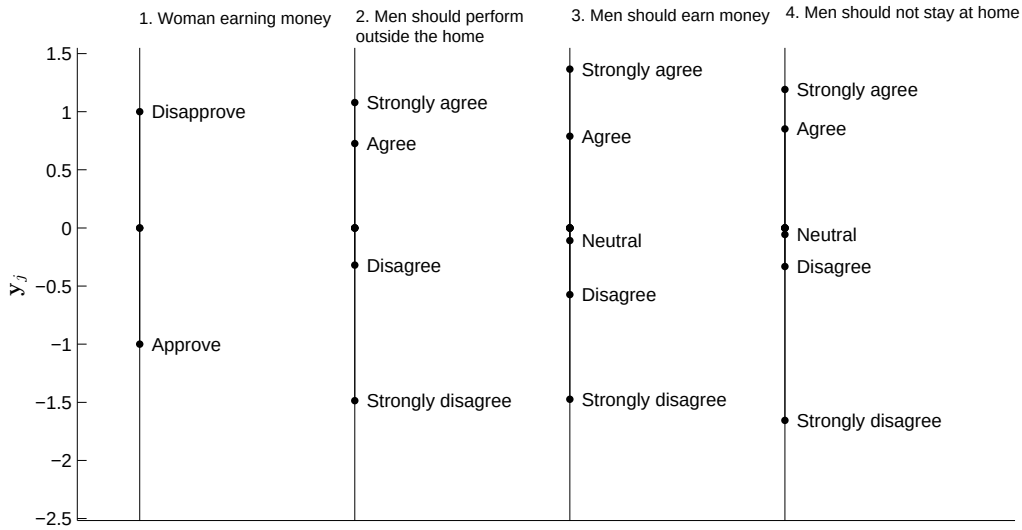
2. Men should perform outside the home: "It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family." (strongly agree, agree, disagree, strongly disagree).

3. Men should earn money: "A man's job is to earn money; a woman's job is to look after the home and family." (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree).

4. Men should not stay at home: "It is not good if the man stays at home and cares for the children and the woman goes out to work." (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree).

From this data set, we construct a contingency table, so that we can apply log-linear analysis (generalized linear modeling with a log link and a Poisson error distribution) with the four questions as predictor variables.

Table 3 gives results of a GLM with only main effects, the OSI model, and a full two-way interaction model for this data set. Figure 2 shows the interaction plot and the estimated scaling factors of the OSI model. The deviance values in Table 3 show that the OSI model does not fit the data, as it is too restrictive. In addition, the values of the Akaike information criterion indicate that a model with full two-way interactions is preferable to the OSI model. Nevertheless, the results of the OSI model can still help us interpret the interaction effects. As all estimated scaling factors are positive, the interaction plot shows that the respondents tend to have similar opinions for items 2, 3, and 4. The similarities are largest between items 2 and 3 ($s_{23} = 1.288$) and between items 3 and 4 ($s_{34} = 0.817$). The categories of item 1 appear inverted compared to the other three variables. This was to be expected, as a negative response to item 1 implies a conservative attitude towards gender roles; for the other three items, a positive response



Item	1	2	3	4
1. Woman earning money	-	0.161 (0.080)	0.224 (0.073)	0.171 (0.070)
2. Men should perform outside the home	-	-	1.288 (0.132)	0.383 (0.081)
3. Men should earn money	-	-	-	0.817 (0.094)
4. Men should not stay at home	-	-	-	-

Figure 2: Interaction plot and scaling factors s_{jl} of the OSI model for the General Social Survey data set. The estimated standard errors of the scaling factors are shown in parentheses.

indicates such a conservative attitude. The reported scaling factors are all statistically significant different from zero and correspond to fairly large effects. For example, for a respondent who responds "strongly disagree" to items 2 and 3, the additive interaction effect between these two items on η_i is $1.288 \times -1.485 \times -1.473 = 2.82$. For a log-linear model, the expected frequency μ_i is calculated as $\mu_i = \exp(\eta_i)$. Therefore, the estimated interaction effect between items 2 and 3 increases the probability of responding "strongly disagree" to both items by a factor of $\exp(2.82) = 16.8$.

4 Multidimensional Extensions

The one-dimensional OSI model (5) is relatively straightforward to interpret. However, as can be seen in the analysis of the GSS data set, this model may yield an inadequate fit for some data sets. In that case, a less restrictive model may be considered. Here, we discuss several ways to generalize the one-dimensional OSI model to a multidimensional model, which should provide a better fit. We also discuss what identification constraints are required for such models and how they are related to previously proposed models.

The most natural generalization of the one-dimensional OSI model consists of allowing for multiple optimally scaled variables per categorical predictor variable. We use this approach for our general multidimensional model, so that it is given by

$$\eta_i = c + \sum_{j=1}^{m} b_j r_{ij} + \sum_{j=1}^{m-1} \sum_{l=j+1}^{m} w_{jl} \sum_{p=1}^{P} s_{jlp} q_{ijp} q_{ilp}, \tag{8}$$

where q_{ijp} is the score of person i on the p-th optimally scaled variable for categorical variable j, and s_{jlp} is the coefficient of the p-th optimally scaled variable for the interaction between categorical variables j and l. By writing this model in terms of the categorical predictor variables, it can be also described as

$$\eta_i = c + \sum_{j=1}^{m} b_j \mathbf{g}'_{ij} \mathbf{a}_j + \sum_{j=1}^{m-1} \sum_{l=j+1}^{m} w_{jl} \mathbf{g}'_{ij} \mathbf{Y}_j \mathbf{S}_{jl} \mathbf{Y}'_l \mathbf{g}_{il}, \tag{9}$$

where $\mathbf{Y}_j$ and $\mathbf{S}_{jl}$ are matrices of sizes $k_j \times P$ and $P \times P$ respectively, with P the dimensionality of the model. The matrix $\mathbf{S}_{jl}$ is constrained to be diagonal. For $P = 1$, (9) simplifies to the one-dimensional OSI model.

For the general multidimensional model (9), we impose similar location constraints as for the one-dimensional model, so that $\sum_{i=1}^{n} \mathbf{g}'_{ij} \mathbf{a}_j = 0$ and $\sum_{i=1}^{n} \mathbf{g}'_{ij} \mathbf{Y}_j = \mathbf{0}'$. For the main effects, the scale constraints are $\sum_{i=1}^{n} \mathbf{g}'_{ij} \mathbf{a}_j \mathbf{a}'_j \mathbf{g}_{ij} = n$; for the interaction effects, we require that $\sum_{i=1}^{n} \mathbf{g}'_{ij} \mathbf{y}_{jp} \mathbf{y}'_{jp} \mathbf{g}_{ij} = n$, where $\mathbf{y}_{jp}$ denotes the p-th column of $\mathbf{Y}_j$. In addition, we must set $s_{jlp} = 0$ for every $w_{jl} = 0$. Furthermore, just as in the one-dimensional OSI model, we change the signs of $\mathbf{y}_{jp}$ and, correspondingly, s_{jlp} in such a

way that the elements of $\sum_{j=1}^{m-1}\sum_{l=j+1}^{m}\mathbf{S}_{jl}$ are maximized. Finally, in a multidimensional model, it is often convenient to ensure that the amount of explained variation decreases with the dimension, so that the first dimension is the most important one. For the general multidimensional model (9), we accomplish this by requiring that the diagonal elements of $\sum_{j=1}^{m}\sum_{l=1}^{m}|\mathbf{S}_{jl}|$ are decreasing.

Determining the number of degrees of freedom in model (9) may be somewhat difficult, as the degrees of freedom are influenced by characteristics of both the model and the research design (that is, the values of the predictor variables in the data set). However, if all interaction terms are present in the model (that is $w_{jl} = 1$ for all $j < l$), and the number of observations is sufficiently large, the number of parameters in the general multidimensional model equals $1 + Pm(m-1)/2 + (1+P)\sum_{j=1}^{m} k_j$, and the total number of parameter restrictions is $(P+2)m$. In that case, the number of degrees of freedom required by this model is given by

$$df = 1 + (1+P)\sum_{j=1}^{m} k_j - (1+2P)m + \frac{1}{2}Pm(m-1). \tag{10}$$

Graphically representing the results of (9) in a way similar to Figures 1 and 2 would require several plots; we believe such a representation would be hard to interpret. Instead, one can construct a biplot for each interaction term separately, which should be straightforward to interpret. To do so, one may calculate a compact singular value decomposition of $\bar{\mathbf{B}}_{jl} = w_{jl}\mathbf{Y}_j\mathbf{S}_{jl}\mathbf{Y}_l'$, so that $\mathbf{U}\mathbf{\Sigma}\mathbf{V}' = w_{jl}\mathbf{Y}_j\mathbf{S}_{jl}\mathbf{Y}_l'$, where $\mathbf{\Sigma}$ is $P \times P$ diagonal matrix, and $\mathbf{U}$ and $\mathbf{V}$ are orthogonal (so that $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$). Matrices $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V}$ that meet these requirements must exist, as the rank of $\bar{\mathbf{B}}_{jl}$ cannot be greater than P. A biplot can then be constructed by plotting $\mathbf{U}\mathbf{\Sigma}^{1/2}$ and $\mathbf{V}\mathbf{\Sigma}^{1/2}$ simultaneously in one figure.

Restricted Multidimensional Models

For some data sets, the general multidimensional model (9) may require prohibitively many parameters, leading to instability of the estimated parameters. In addition, interpreting the estimated interaction effects using graphical representations may be difficult if there are many predictor variables. In such cases, alternative generalizations of the one-dimensional OSI model with fewer parameters can be considered. Here, we discuss three such generalizations, which consist of restricting the parameters in $\mathbf{S}_{jl}$ to be equal for each interaction term or for each dimension.

First, we can restrict the elements of $\mathbf{S}_{jl}$ to be equal across dimensions, which leads to the model

$$\eta_i = c + \sum_{j=1}^{m} b_j \mathbf{g}_{ij}'\mathbf{a}_j + \sum_{j=1}^{m-1}\sum_{l=j+1}^{m} w_{jl}s_{jl}\mathbf{g}_{ij}'\mathbf{Y}_j\mathbf{Y}_l'\mathbf{g}_{il}. \tag{11}$$

The location and scale constraints of the general multidimensional model are also used for this model. As for any orthogonal rotation matrix $\mathbf{T}$, $\mathbf{Y}_j\mathbf{Y}_l' = s\mathbf{Y}_j\mathbf{T}\mathbf{T}'\mathbf{Y}_l'$, simultaneously rotating the matrices $\mathbf{Y}_j$ does not alter the values of η_i. Therefore, *rotation restrictions* are also required for this model. We require that $\sum_{j=1}^{m}\mathbf{Y}_j'\mathbf{G}_j'\mathbf{G}_j\mathbf{Y}_j$ is diagonal, thereby imposing $P(P-1)/2$ parameter restrictions on the $\mathbf{Y}_j$s.

A second type of restricted model can be obtained by imposing that $\mathbf{S}_{jl} = \mathbf{S}$ for every interaction term, so that

$$\eta_i = c + \sum_{j=1}^{m} b_j\mathbf{g}_{ij}'\mathbf{a}_j + \sum_{j=1}^{m-1}\sum_{l=j+1}^{m} w_{jl}\mathbf{g}_{ij}'\mathbf{Y}_j\mathbf{S}\mathbf{Y}_l'\mathbf{g}_{il}, \tag{12}$$

where $\mathbf{S}$ is diagonal. Here, the scale constraint $\sum_{i=1}^{n}\mathbf{g'}_{ij}\mathbf{y}_{jp}\mathbf{y'}_{jp}\mathbf{g}_{ij} = n$ cannot be imposed without loss of generality. Instead, we impose the constraint $\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbf{g}_{ij}'\mathbf{y}_{jp}\mathbf{y}_{jp}'\mathbf{g}_{ij} = mn$. For log-linear modeling, this model coincides with equation (24) of Anderson and Vermunt (2000), though the parameter restrictions used in their article are different.

Finally, one may consider restricting the parameter matrices $\mathbf{S}_{jl}$ to be equal for every interaction term and for every dimension (so that $\mathbf{S}_{jl} = \mathbf{I}$), essentially removing these parameters from the model. In that case, the model is

$$\eta_i = c + \sum_{j=1}^{m} b_j\mathbf{g}_{ij}'\mathbf{a}_j + \sum_{j=1}^{m-1}\sum_{l=j+1}^{m} w_{jl}\mathbf{g}_{ij}'\mathbf{Y}_j\mathbf{Y}_l'\mathbf{g}_{il}, \tag{13}$$

and we obtain the interaction decomposition model proposed by Groenen and Koning (2006). Here, the sizes of the interaction effects are determined by the parameter matrices $\mathbf{Y}_j$. The same location and rotation constraints as in (11) can be imposed; however, no scale constraints can be imposed on the $\mathbf{Y}_j$s without loss of generality. If the estimated interactions effects are of a similar size, the results of (13) can be conveniently visualized using a biplot in which the $\mathbf{Y}_j$s are simultaneously plotted in a P-dimensional space. However, if various interaction interaction effects differ in size significantly, the visualization may break down, and the model may fit poorly.

Model (13) is a special case of (12), which can be obtained by setting $s_p = 1$ for all p; the results of these two models may appear to be almost identical, although they are not equivalent. Model (13) can be obtained from (12) by multiplying the elements of each $\mathbf{y}_{jp}$ with $\sqrt{s_p}$, but this is only possible if all s_p are nonnegative, so that (12) is more flexible than (13).

Experimentation with these multidimensional models suggests that unique parameter estimates that maximize the log-likelihood function may not always exist for models (11) and (13). In that case, the parameter estimates that optimization algorithms produce may fail to converge to finite values and could approach infinity instead. This effect does not appear to occur in the one-dimensional OSI model or in model (12).

Table 4: Maximum degrees of freedom associated with various models

Model	Degrees of freedom
Main effects only GLM	$1 + \sum_{j=1}^{m} k_j - m$
One-dimensional OSI model (5)	$1 + 2\sum_{j=1}^{m} k_j - 3m + m(m-1)/2$
General multidimensional model (9)	$1 + (1+P)\sum_{j=1}^{m} k_j - (1+2P)m + Pm(m-1)/2$
Restricted multidimensional model (11)	$1 + (1+P)\sum_{j=1}^{m} k_j - (2+P)m + m(m-1)/2 - P(P-1)/2$
Restricted multidimensional model (12)	$1 + (1+P)\sum_{j=1}^{m} k_j - (1+P)m - P(P-1)/2$
Restricted multidimensional model (13)	$1 + (1+P)\sum_{j=1}^{m} k_j - (1+P)m - P(P-1)/2$
Full two-way interaction GLM	$1 + \sum_{j=1}^{m} k_j - m + \sum_{j=1}^{m-1}\sum_{l=j+1}^{m}(k_j-1)(k_l-1)$

Table 4 gives an overview of the maximum numbers of degrees of freedom for a number of models, based on the location, scale, and rotation constraints that were described previously. The values in this table are upper bounds on the actual degrees of freedom; they can only be attained if all interaction terms are taken into account, and both the number of observations and the number of variables are large enough. Whether the model parameters are identified can often be determined empirically. This can, for example, be done by calculating the matrix of second-order partial derivatives of the log-likelihood function at the final parameter estimates and then checking whether this matrix is positive definite, which is a necessary condition for parameter identification.

5 Discussion

Optimal scaling is a useful methodology for modeling the effects of categorical predictor variables (see, for example Gifi, 1990). In this article, we have applied this methodology to modeling two-way interactions effects in generalized linear models. The resulting optimal scaling of interactions (OSI) model is a multiplicative interaction model that can help interpret the content of interaction effects. This model has the additional advantages that it requires fewer parameters than a full two-way interaction model and that it can be used to construct (graphical) representations of the interaction effects. The OSI model can be seen as an extension of several models for parsimoniously representing interaction effects, including Goodman's RC(M) association models and models that were proposed by Anderson and Vermunt (2000) and Groenen and Koning (2006).

Using two empirical data sets, we have shown how the OSI model can be applied in practice and we have compared its usefulness with other models. Based on the results, the one-dimensional OSI model appears to be most useful, as it is easy to apply and appears to give good results. Multidimensional models may lead to representations that are not so easy to interpret. Based on our experience with these models, we recommend using

the one-dimensional OSI model. We believe that this model can be useful for interpreting interaction effects in an applied setting.

An advantage of the one-dimensional OSI model is that it uses different sets of parameters to model the strength of an interaction term (using the scaling factors s_{jl}) and the content of an interaction term (using $\mathbf{y}_j$). This separation helps to understand the results. A limitation of the one-dimensional OSI model is that it may have an inadequate fit for some data sets. This problem can be solved by applying one of the multidimensional extensions in Section 4. In some cases, however, degenerate solutions may occur that are avoided in the one-dimensional OSI model. Clearly, the models presented in this paper are not capable of taking three-way or higher interaction effects into account. Multiplicative interaction models can also be constructed for modeling higher-way terms, as was done in Van Eeuwijk and Kroonenberg (1998) for three-way interactions effects. However, we believe that two-way interactions are the most important ones to explore in practice.

References

Anderson, C. J. (1996). The analysis of three-way contingency tables by three-mode association models. *Psychometrika*, *61*, 465-483.

Anderson, C. J., & Vermunt, J. K. (2000). Log-multiplicative association models as latent variable models for nominal and/or ordinal data. *Sociological Methodology*, *30*, 81-122.

Breiman, L., & Friedman, J. H. (1985). Estimating optimal transformation for multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, *80*, 580-619.

Choulakian, V. (1996). Generalized bilinear models. *Psychometrika*, *61*, 271-283.

Davis, J. A., & Smith, T. W. (1996). *General Social Surveys 1972–1996: Cumulative codebook*. Chicago, IL: National Opinion Research Center.

De Leeuw, J., Young, F. W., & Takane, Y. (1976). Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, *41*(4), 471-503.

De Falguerolles, A., & Francis, B. (1992). Algorithmic approaches for fitting bilinear models. In Y. Dodge & J. Whittaker (Eds.), *Compstat 1992* (Vol. 1, p. 77-82). Physica-Verlag Heiderberg.

Gabriel, K. R. (1998). Generalised bilinear regression. *Biometrika*, *85*(3), 689-700.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: Wiley.

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.

Gower, J. C., & Hand, D. J. (1996). *Biplots* (No. 54). London: Chapman & Hall.

Groenen, P. J. F., & Koning, A. J. (2006). A new model for visualizing interactions in analysis of variance. In M. Greenacre & J. Blasius (Eds.), *Multiple correspondence analysis and related methods* (p. 487-502). Chapman & Hall.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. New York: Chapman and Hall.

Kruskal, J. B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society. Series B (Methodological)*, *27*, 251-263.

Linting, M., Meulman, J. J., Groenen, P. J. F., & Van der Kooij, A. J. (2007). Nonlinear principal components analysis: Introduction and application. *Psychological Methods*, *12*, 336-358.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (Second ed.). Chapman & Hall.

Nelder, J., & Wedderburn, R. (1972). Generalized linear models. *Journal of The Royal Statistical Society. Series A (General)*, *135*(3), 370-384.

Siciliano, R., & Mooijaart, A. (1997). Three-factor association models for three-way contingency tables. *Computational Statistics and Data Analysis*, *24*, 337-356.

Van Eeuwijk, F. A. (1995). Multiplicative interaction in generalized linear models. *Biometrics*, *51*, 1017-1032.

Van Eeuwijk, F. A. (1996). *Between and beyond additivity and non-additivity: The statistical modelling of genotype by environment interaction in plant breeding*. Unpublished doctoral dissertation, Wageningen Agricultural University, Wageningen, the Netherlands.

Van Eeuwijk, F. A., & Kroonenberg, P. M. (1998). Multiplicative models for interaction in three-way anova, with applications to plant breeding. *Biometrics*, *54*(4), 1315-1333.

Wong, R.-K. (2001). Multidimensional association models: A multilinear approach. *Sociological Methods and Research*, *30*(2), 197-24.

Young, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika*, *46*, 357–388.

Young, F. W., De Leeuw, J., & Takane, Y. (1976). Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, *41*(4), 505-529.