LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR STATISTIK
SONDERFORSCHUNGSBEREICH 386

Tutz, Scholz:

Semiparametric Modelling of Multicategorical Data

Sonderforschungsbereich 386, Paper 209 (2000)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

gsf

MAX-PLANCK-GESELLSCHAFT

TUM

Semiparametric modelling of multicategorical data

G. Tutz, T. Scholz
University of Munich, Institute of Statistics
Akademiestrasse 1, 80799 München
tutz@stat.uni-muenchen.de scholz@stat.uni-muenchen.de

Abstract

Parametric multicategorical models are an established tool in statistical data analysis. Alternative semi-parametric models are introduced where part of the explanatory variables is still linearly parametrized and the rest is given as a sum of unspecified functions of the explanatory variables. The modelling approach distinguishes between global and category specific variables, in contrast to global variables the latter may have different values for differing categories of the response. Estimation procedures are derived which make use of an expansion in basis functions which are localized on a grid of values of the explanatory variables. Regularization of the estimates is obtained by penalization.

1 Introduction

In generalized linear models (Nelder & Wedderburn, 1972) the mean of a response variable y is related to a vector of covariates by

$$g(\mu) = x^T\beta$$

where g is a specified link function. Various extensions of the model have been given in order to avoid the restrictive parametric form of the predictor. An extension which still has additive structure is the generalized additive model (Hastie & Tibshirani, 1990) where the linear predictor is replaced by an additive term yielding

$$g(\mu) = \beta_0 + \gamma_{(1)}(x_1) + \ldots + \gamma_{(p)}(x_p)$$

with $x_1, \ldots, x_p$ being the components in $x^T = (x_1, \ldots, x_p)$ and $\gamma_{(j)}$ are unspecified unknown functions. Partially linear models may be treated as a special case where part of the covariates, say $x_1, \ldots, x_{p_1}$ have linear form. These types of models are given by

$$g(\mu) = \beta_0 + \sum_{j=1}^{p_1} x_j \beta_j + \sum_{j=p_1+1}^{p} \gamma_{(j)}(x_j)\,.$$

While these models are in wide use there have been relatively few extensions to the multivariate case. Yee & Wild (1996) considered a multivariate additive model using smoothing spline methodology with estimation based on the backfitting algorithm. In the neural network literature multivariate models are treated extensively, however, without constraints on the form of the predictor (e.g. Bishop, 1995). Thus in neural networks one assumes a multi-dimensional unspecified function of the covariates. However, the estimated functions do not allow to determine the influence of specific covariates which is a central issue in statistical analysis. New developments are found in marketing where multicategorical models are an established tool to model the choice of brands. Abe (1999) develops a multicategorical model which is based on penalized likelihood whereas Hruschka (1998), Hruschka et al. (1999) base their brand choice models on artificial networks.

The models considered here are differing from these approaches in several respects. While Yee & Wild (1996) consider only global variables i.e. variables which do not vary across response categories, in the marketing literature only the other type of covariates, namely brand-specific variables which do vary across categories are used. Here both types of variables are included. In addition we use a simple way for estimating smooth functions which allows to embed the estimation into the framework of multivariate generalized linear models.

In the following multivariate response means multicategorical response. In section 2 first we consider the parametric multinomial model as a special case of a multivariate generalized model. The underlying maximization of random utilities will be considered as a basic tool for the construction of semi-parametric extensions. In section 3 estimation of the general model is derived. The method used assumes that the unspecified functions may be approximated by a finite number of basis functions which in the estimation procedure are penalized with reference to the localization of the basis functions. In section 4 the methodology is applied to the preferences for political parties in Germany.

2 Semiparametric models

2.1 The multinomial logit model and latent utilities

Let Y denote the response variable with possible values $1, \ldots, k$ with the numbers representing only labels for categories on a nominal scale. In probabilistic choice theory it is often assumed that the response categories are linked to unobserved utilities. Let more generally U_r be a latent variable which has the form

$$U_r = u_r + \varepsilon_r$$

where u_r is a fixed value associated with the rth response category and $\varepsilon_1, \ldots, \varepsilon_k$ are iid random variables with distribution function F which represent the noise. The *principle of maximum random utility* links the observable response Y to the latent variables $U_1, \ldots, U_k$ in the form

$$Y = r \Leftrightarrow U_r = \max_{j=1,\ldots,k} U_j.$$

It is well known (McFadden, 1973, McFadden, 1981) that the assumption of the extreme (maximal) value distribution

$$F(x) = \exp(-\exp(-x))$$

yields the multinominal logit model

$$P(Y = r) = \frac{\exp(u_r)}{\sum_{s=1}^{k} \exp(u_s)}.$$

However, only differences between the latent utilities $u_1, \ldots, u_k$ are identifiable. By choosing a reference category, which in the following is k one obtains

$$P(Y = r) = \frac{\exp(u_r - u_k)}{1 + \sum_{s=1}^{q} \exp(u_s - u_k)} = \frac{\exp(\eta_r)}{1 + \sum_{s=1}^{q} \exp(\eta_s)}$$

where $\eta_r = u_r - u_k$ is the identifiable difference between the rth fixed utility and the utility of the reference category k and $q = k - 1$.

In parametric models the utilities u_r and η_r are specified as known functions of the explanatory variables, most often they have linear form. It is useful to distinguish between two types of variables namely global variables which only characterize the respondent, e.g. age in the choice of transportation mode or

the choice of brands, and category-specific variables which are specific for the alternatives $1, \ldots, k$, e.g. the price of the alternatives. Let the data at hand be given by $(Y_i, x_i, w_{i1}, \ldots, w_{ik})$, $i = 1, \ldots, n$, where $Y_i \in \{1, \ldots, k\}$ is the response variable, x_i is a vector of global variables and $w_{i1}, \ldots, w_{ik}$ is a set of vector-valued variables which are connected to the alternatives $1, \ldots, k$. An often used linear specification is given by

$$u_{ir} = \bar{\gamma}_{0r} + x_i^T \bar{\gamma}_r + w_{ir}^T \alpha, \quad r = 1, \ldots, k$$

which yields the differences

$$\begin{aligned} \eta_{ir} &= u_{ir} - u_{ik} \\ &= \gamma_{0r} + x_i^T \gamma_r + (w_{ir}^T - w_{ik}^T)\alpha, \quad r = 1, \ldots k-1, \end{aligned}$$

where $\gamma_{0r} = \bar{\gamma}_{0r} - \bar{\gamma}_{0k}$, $\gamma_r = \bar{\gamma}_r - \bar{\gamma}_k$. The resulting multinomial logistic model has the form

$$P(Y_i = r|x_i, \{w_{ij}\}) = \frac{\exp(\gamma_{0r} + x_i^T \gamma_r + (w_{ir} - w_{ik})^T \alpha)}{1 + \sum_{s=1}^{q} \exp(\gamma_{0s} + x_i^T \gamma_s + (w_{is} - w_{ik})^T \alpha)}$$

or

$$\log\left(\frac{P(Y_i = r|x_i, \{w_{ij}\})}{P(Y_i = k|x_i, \{w_{ij}\})}\right) = \eta_{ir} = \gamma_{0r} + x_i^T \gamma_r + (w_{ir} - w_{ik})^T \alpha.$$

The model may be written in the form of a multivariate generalized model

$$g(\pi_i) = Z_i \beta \quad \text{or} \quad \pi_i = h(Z_i \beta)$$

where $\pi_i^T = (\pi_{i1}, \ldots, \pi_{iq})$, $q = k-1$, is the vector of the response probabilities with components $\pi_{ir} = P(Y_i = r|x_i, \{w_{ij}\})$, g is a link function, the inverse function $h = g^{-1}$ is the response function, $\beta^T = (\gamma_{01}, \ldots, \gamma_{0q}, \gamma_1^T, \ldots, \gamma_q^T, \alpha^T)$ is the vector of parameters and Z_i is a design matrix composed from $x_i, \{w_{ij}\}$, $i = 1, \ldots, n$ which has the form

$$Z_i = \left(\begin{array}{ccc|ccc|c} 1 & & & x_i^T & & & w_{i1} - w_{ik} \\ & 1 & & & x_i^T & & w_{i2} - w_{ik} \\ & & \ddots & & & \ddots & \vdots \\ & & 1 & & & x_i^T & w_{iq} - w_{ik} \end{array}\right).$$

Thus the vector valued predictor $\eta_i^T = (\eta_{i1}, \ldots, \eta_{iq})$ has the linear form $\eta_i = Z_i \beta$. For details see Fahrmeir & Tutz (2000).

2.2 The semiparametric multinomial logit model

Instead of assuming a known parametrized function for u_{ir} as in the common multinomial model here the form is an unspecified additive function. The utility is determined by the variables $x_i^T = (x_{i1}, \ldots, x_{ip})$ and $w_{ir}^T = (w_{i1r}, \ldots, w_{imr})$ in the form

$$u_{ir} = \bar{\gamma}_{0r} + \sum_{j=1}^{p} \bar{\gamma}_{(j),r}(x_{ij}) + \sum_{j=1}^{m} \alpha_{(j),r}(w_{ijr})$$

yielding the predictor for observation i and response category r

$$\eta_{ir} = u_{ir} - u_{ik} = \gamma_{0r} + \sum_{j=1}^{p} \gamma_{(j),r}(x_{ij}) + \sum_{j=1}^{m} \alpha_{(j),r}(w_{ijr}) - \alpha_{(j),k}(w_{ijk}) \qquad (1)$$

where $\gamma_{0r} = \bar{\gamma}_{0r} - \bar{\gamma}_{0k}$ and $\gamma_{(j),r}(x_{ij}) = \bar{\gamma}_{(j),r}(x_{ij}) - \bar{\gamma}_{(j),k}(x_{ij})$. Thus the logits $\log(\pi_{ir}/\pi_{ik})$, $r = 1, \ldots, q$ have an additive form with $\gamma_{(j),r}(x_{ij})$ being the effect of the jth global variable and $\alpha_{(j),r}(w_{ijr}) - \alpha_{(j),k}(w_{ijk})$ being the effect of category specific covariates. A simpler version of (1) follows from the assumption that the response-specific variables have identical influence upon the response categories. Thus one assumes $\alpha_{(j),r}(w) = \alpha_{(j)}(w)$ for $r = 1, \ldots, q$. This type of model has been considered for brand choice by Abe (1999).

The parametric multinomial model results as the special case where

$$\begin{aligned} \gamma_{(j),r}(x_{ij}) &= x_{ij}\gamma_{jr}, \\ \alpha_{(j),r}(w_{ijr}) &= w_{ijr}\alpha_j. \end{aligned}$$

If the predictor has the form (1) each explanatory variable is given as a smooth function. However, if some of the explanatory variables are categorical a smooth function is not appropriate. Thus a slightly more general semiparametric model is considered in which some of the explanatory variables have a linear form and some are given as unknown functions. In the general model the predictor has the form

$$\begin{aligned} \eta_{ir} &= \gamma_{0r} + \sum_{j=1}^{p_1} x_{ij}\gamma_{jr} + \sum_{j=p_1+1}^{p} \gamma_{(j),r}(x_{ij}) \\ &\quad + \sum_{j=1}^{m_1} w_{ijr}\alpha_{jr} - w_{ijk}\alpha_{jk} + \sum_{j=m_1+1}^{m} \alpha_{(j),r}(w_{ijr}) - \alpha_{(j),k}(w_{ijk}). \qquad (2) \end{aligned}$$

3 Estimation by penalized basis functions

3.1 Basis functions

In the following the smooth functions are specified as a finite number of basis functions which are characterized by knots or anchor points. A well known approach of this type is based on smoothing splines where the basis functions are for example B-splines and the knots are given by the observations. Alternative functions have been used in the neural network community in the form of so-called radial basis functions. A radial basis function has the form

$$G(x \mid x_0) = \Phi\left(||x - x_0||\right)$$

where Φ is some unidimensional function. Functions which have been used are the Gaussian kernel, thin-plate spline functions $\Phi(x) = x^2 \log(x)$, as well as the linear function $\Phi(x) = x$ (see e.g. Bishop, 1995). The basis function G is centered around the anchor point x_0 which determines the location of the basis function. In contrast to the thin-plate spline function the Gaussian kernel is local in the sense that $\Phi(y) \longrightarrow 0$ as $|y| \longrightarrow \infty$.

In the following it is assumed that the smooth functions are given by

$$\gamma_{(j),r}(x) = \sum_{s=1}^{P_j} \gamma_{jrs} G_{js}(x), \tag{3}$$

$$\alpha_{(j),r}(w) = \sum_{s=1}^{M_j} \alpha_{jrs} A_{js}(w), \tag{4}$$

where $G_{js}(x) = G(x|x_{j(s)})$, $A_{js}(w) = A(w|w_{j(s)})$ are basis functions which are linked to knots $x_{j(s)}, w_{j(s)}$ which characterize the location of the basis function. The knots $x_{j(s)}$ are chosen on a grid from the range of the jth global variable, the knots $w_{j(s)}$ are from the range of the jth category-specific variable. In the following it is assumed that the knots are ordered, i.e. $x_{j(1)} < \cdots < x_{j(P_j)}$ and $w_{j(1)} < \cdots < w_{j(M_j)}$. Since basis functions are linked to these knots one obtains an ordering of the basis functions. Each basis function is considered as linked to just one knot, for kernel functions this may be the mode of the kernel, for B-splines which have several knots the knot which is linked to the basis function can be chosen for example as the leftmost knot at which the B-spline starts to become nonzero.

Although usually the basis functions will be of the same type different functions could be used for different components. The semiparametric model (2) follows

simply by specifying for the variables which have linear form the identity function as basis function and the number of basis functions as one. With $G_{j1}(x) = x$ and $P_j = 1$, $j = 1, \ldots, p_1$, one obtains

$$\gamma_{(j),r}(x_{ij}) = \gamma_{jr1} G_{j1}(x_{ij}) = \gamma_{jr} x_{ij}\,, \; j = 1, \ldots, p_1,$$

where $\gamma_{jr} = \gamma_{jr1}$.

The essential advantage in using basis functions is that the resulting model may be embedded into the framework of parametric generalized linear models. Use of (3) and (4) yields the linear predictor

$$\begin{aligned}
\eta_{ir} &= \gamma_{0r} + \sum_{j=1}^{p} \sum_{s=1}^{P_j} \gamma_{jrs} G_{js}(x_{ij}) + \sum_{j=1}^{m} \sum_{s=1}^{M_j} \alpha_{jrs} A_{js}(w_{ijr}) - \sum_{s=1}^{M_j} \alpha_{jks} A_{js}(w_{ijk}) \\
&= \gamma_{0r} + \sum_{j=1}^{p} \gamma_{jr}^T c_{ij} + \sum_{j=1}^{m} \alpha_{jr}^T a_{ijr} - \alpha_{jk}^T a_{ijk}
\end{aligned}$$

where

$$\begin{aligned}
\gamma_{jr}^T &= (\gamma_{jr1}, \ldots, \gamma_{jrP_j}), \\
c_{ij}^T &= (G_{j1}(x_{ij}), \ldots, G_{jP_j}(x_{ij})), \\
\alpha_{jr}^T &= (\alpha_{jr1}, \ldots, \alpha_{jrM_j}), \\
a_{ijr}^T &= (A_{j1}(w_{ijr}), \ldots, A_{jM_j}(w_{ijr})).
\end{aligned}$$

Thus the vector valued predictor of the multivariate model $\eta_i^T = (\eta_{i1}, \ldots, \eta_{iq})$ is given by

$$\eta_i = \gamma_0 + \sum_{j=1}^{p} C_{ij} \gamma_j + \sum_{j=1}^{m} A_{ij} \alpha_j - A_{ijk} \alpha_{jk} = Z_i \beta$$

where

$$\begin{aligned}
\gamma_0^T &= (\gamma_{01}, \ldots, \gamma_{0q}), \\
C_{ij} &= \mathrm{Diag}(c_{ij}^T, \ldots, c_{ij}^T), \quad \gamma_j^T = (\gamma_{j1}^T, \ldots, \gamma_{jq}^T), \\
A_{ij} &= \mathrm{Diag}(a_{ij1}^T, \ldots, a_{ijq}^T), \quad \alpha_j^T = (\alpha_{j1}^T, \ldots, \alpha_{jq}^T), \\
A_{ijk} &= (a_{ijk}, \ldots, a_{ijk})^T.
\end{aligned}$$

The corresponding design matrix Z_i is a collection of these parts, i.e. with I denoting the unit matrix one has

$$\begin{aligned}
Z_i &= (I, C_{i1}, \ldots, C_{ip}, A_{i1}, A_{i1k}, \ldots, A_{im}, A_{imk}) \\
\beta^T &= (\gamma_0^T, \gamma_1^T, \ldots, \gamma_p^T, \alpha_1^T, \alpha_{1k}^T, \ldots, \alpha_m^T, \alpha_{mk}^T).
\end{aligned}$$

The resulting model $\mu_i = h(Z_i \beta)$ can be estimated as a generalized linear model.

3.2 Regularization by penalization

The expansion of the unspecified functions of explanatory variables may be based on quite different basis functions. One might use the orthonormal Fourier basis or the truncated power series basis which is used within the regression spline approach. Here the basis functions are always localized meaning that they are linked to an anchor point or knot and have essentially local support. For example the Gaussian kernel has no local support in a strict sense but the kernel is decreasing with increasing distance from its mode, thus it is localized around its mode. For basis functions of this type it is still a question which knots to select. If few knots are selected the number of parameters to be estimated remains low but the range of functions which may be approximated is severely limited. An alternative approach which we will pursue here has been outlined by Eilers & Marx (1996) for the B-spline basis within the framework of univariate generalized linear models. The basic idea is to use many basis functions linked to an equidistant grid of knots. Thus the number of parameters is increased but the estimates may adapt very flexible to a whole range of functions. The high number of parameters to be estimated is restricted by penalizing the variation of weights which correspond to adjacent basis functions. However, since many parameters have to be estimated some regularization is useful in order to obtain smooth estimates. Therefore we consider penalized likelihood estimation where the penalty is given as finite differences of the coefficients of adjacent basis functions.

The penalized log-likelihood which is to be maximized is given by

$$\begin{aligned} \mathrm{pl}(\beta) &= \sum_{i=1}^{n} l_i(y_i, \eta_i) \\ &\quad - \frac{1}{2} \sum_{j=1}^{p} \sum_{r=1}^{q} \lambda_{jr} \sum_{s=d_{jr}+1}^{P_j} (\Delta^{d_{jr}} \gamma_{jrs})^2 - \frac{1}{2} \sum_{j=1}^{m} \sum_{r=1}^{k} \lambda^c_{jr} \sum_{s=d^c_{jr}+1}^{M_j} (\Delta^{d^c_{jr}} \alpha_{jrs})^2 \end{aligned}$$

where $l_i(y_i, \eta_i)$ is the log-likelihood contribution of the model with predictor η_i and Δ is the difference operator operating on adjacent coefficients of the basis functions, i.e. $\Delta\gamma_{jrs} = \gamma_{jrs} - \gamma_{jr,s-1}$, $\Delta^2\gamma_{jrs} = \Delta(\gamma_{jrs} - \gamma_{jr,s-1})$ etc., and $\lambda_{jr}, \lambda^c_{jr}$ are smoothing parameters for the global and category-specific variables which determine the smoothness of the estimated functions.

The difference operator may be given as a matrix D^d where D is a contrast matrix containing $-1, 1$ in each row. With $K_d = (D^d)^T(D^d)$ one obtains in matrix notation

$$\mathrm{pl}(\beta) = \sum_{i=1}^{n} l_i(y_i, \eta_i) - \frac{1}{2} \sum_{j=1}^{p} \sum_{r=1}^{q} \lambda_{jr} \gamma_{jr}^T K_{d_{jr}} \gamma_{jr} - \frac{1}{2} \sum_{j=1}^{m} \sum_{r=1}^{k} \lambda^c_{jr} \alpha_{jr}^T K_{d^c_{jr}} \alpha_{jr} \qquad (5)$$

where d_{jr}, d^c_{jr} are the degrees of differences for the effect of the jth variable on category r which may vary across variables and categories.

Maximization of the penalized log-likelihood is obtained by solving the generalized estimation equation $\mathrm{ps}(\beta) = \partial l(\beta)/\partial\beta = 0$. It may be performed by a version of Fisher scoring which has the form

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \tilde{F}(\hat{\beta}^{(k)})^{-1}\,\mathrm{ps}(\hat{\beta}^{(k)})\,,$$

where $\tilde{F}(\beta) = F(\beta) + \Lambda K$ is the Pseudo-Fisher matrix and $\mathrm{ps}(\beta)$ is the pseudo-score function given by $\mathrm{ps}(\beta) = Z^T D^T(\beta)\,\Sigma^{-1}(\beta)\,(y-\mu) - \Lambda K\beta$ (see Appendix).

If the underlying smooth functions may be represented by sums of basis functions with the specified number of knots $P_1, \ldots, P_p, M_1, \ldots, M_m$ for fixed smoothing parameters $\lambda_{jr}, \lambda^c_{jr}$ the usual asymptotic results hold under standard assumptions if $n \to \infty$, $\lambda_{jr}/n, \lambda^c_{jr}/n \to 0$. In this case estimates of variances may be computed by $\widehat{\mathrm{cov}}(\hat{\beta}) = \tilde{F}^{-1}(\hat{\beta})$ or $\widehat{\mathrm{cov}}(\hat{\beta}) = F^{-1}(\hat{\beta})$ with the same asymptotic behaviour. The multivariate analog to the sandwich estimator which has been used by Marx & Eilers (1998) in the univariate case is given by

$$\widehat{\mathrm{cov}}(\hat{\beta}) = \tilde{F}^{-1}(\hat{\beta})\,F(\hat{\beta})\,\tilde{F}^{-1}(\hat{\beta})\,. \tag{6}$$

The sandwich estimator may be motivated by considering the case of an univariate normally distributed response with the identity link. In this case the estimator becomes $\hat{\beta} = (F(\beta) + \Lambda K)^{-1}\,Z^T\,\Sigma^{-1}(\beta)\,y$ with $F(\beta) = Z^T\,\Sigma^{-1}(\beta)\,Z$ and one obtains (6) as the exact formula for the covariance. Thus it may be a better approximation if the conditions for asymptotics are not fulfilled. The essential condition for asymptotic considerations is that the number of knots and the penalties as represented by the smoothing parameters are fixed. In practice a moderate number of knots which is a fraction of the sample size works pretty well.

4 Example

We consider data on preferences for political parties which have been collected in the German socio-economic panel. The data set comprises 1913 individuals who were asked if they prefer the christian-democratic party (CDU/CSU, $Y = 1$), the liberal party (FDP, $Y = 2$) or the social-democratic party (SPD, $Y = 3$). The covariates were gender and age and an indicator for the area with AREA= 1 if the individual lives in the former western part of Germany and AREA= 0 otherwise. The highest degree of education for all individuals was secondary school (Hauptschule).

We consider the parametric model

$$\begin{aligned}\log(\pi_r/\pi_k) &= \gamma_{0r} + \text{GENDER} * \gamma_{G,r} + \text{AREA} * \gamma_{AR,r} + \text{AGE} * \gamma_{A,r} \\ &\quad + \text{AGE}^2 * \gamma_{A^2,r} + \text{AGE}^3 * \gamma_{A^3,r} \end{aligned} \tag{7}$$

Table 1 shows the parameter estimates for model (7). Among the categorical variables only area seems to have a significant fact upon the preference of category 1 over category 3.

	$\hat{\gamma}_0$	$s(\hat{\gamma}_0)$	$\hat{\gamma}_G$	$s(\hat{\gamma}_G)$	$\hat{\gamma}_{AR}$	$s(\hat{\gamma}_{AR})$
category 1 (CDU/CSU)	0.02	0.13	-0.14	0.10	-0.29	0.13
category 2 (FDP)	-3.16	0.44	-0.14	0.36	-0.38	0.44
	$\hat{\gamma}_A$	$s(\hat{\gamma}_A)$	$\hat{\gamma}_{A^2}$	$s(\hat{\gamma}_{A^2})$	$\hat{\gamma}_{A^3}$	$s(\hat{\gamma}_{A^3})$
category 1 (CDU/CSU)	0.02	0.11	$5 \cdot 10^{-5}$	$2 \cdot 10^{-3}$	$-8 \cdot 10^{-7}$	$1 \cdot 10^{-5}$
category 2 (FDP)	0.28	0.63	$-3 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	$8 \cdot 10^{-6}$	$8 \cdot 10^{-5}$

Table 1: Effects of gender and area preference for parties for the parametric model

Alternatively the smooth model

$$\log(\pi_r/\pi_k) = \gamma_{0r} + \text{GENDER} * \gamma_{G,r} + \text{AREA} * \gamma_{AR,r} + \gamma_r(\text{AGE})$$

is considered where the form of age is unspecified. In a first step the smoothing parameters have been chosen in order to give good visualization.

λ		$\hat{\gamma}_0$	$s(\hat{\gamma}_0)$	$\hat{\gamma}_G$	$s(\hat{\gamma}_G)$	$\hat{\gamma}_{AR}$	$s(\hat{\gamma}_{AR})$
0.3	category 1 (CDU/CSU)	0.025	0.128	-0.142	0.096	-0.295	0.129
	category 2 (FDP)	-3.259	0.446	-0.136	0.357	-0.376	0.442
1.0	category 1 (CDU/CSU)	0.025	0.128	-0.140	0.095	-0.293	0.129
	category 2 (FDP)	-3.199	0.441	-0.140	0.356	-0.364	0.441
12.2	category 1 (CDU/CSU)	0.028	0.127	-0.139	0.095	-0.295	0.128
	category 2 (FDP)	-3.086	0.431	-0.158	0.355	-0.401	0.439
54.6	category 1 (CDU/CSU)	0.034	0.127	-0.144	0.095	-0.298	0.127
	category 2 (FDP)	-3.022	0.427	-0.174	0.355	-0.448	0.437

Table 2: Effects of gender and area upon preference for parties for the semiparametric model

For comparison the estimated effects of categorical variables are shown in Table 2 for a wide range of smoothing parameters and B-spline basis functions for 20

equidistant knots. It is seen that for category 1 which is the only one which shows significance, the estimates are quite stable, varying smoothing parameters yield rather similar estimates. In Figure 1 the estimated probabilities are given for smoothing parameter $\lambda = 54.6$. It is seen that the preference for the christian-democratic party is increasing across age with the effect that for eastern countries the overall preference for this party is stronger than for the social-democratic party if people are above 60 years of age. The preference for the liberal party is very low for all subpopulation and age groups.

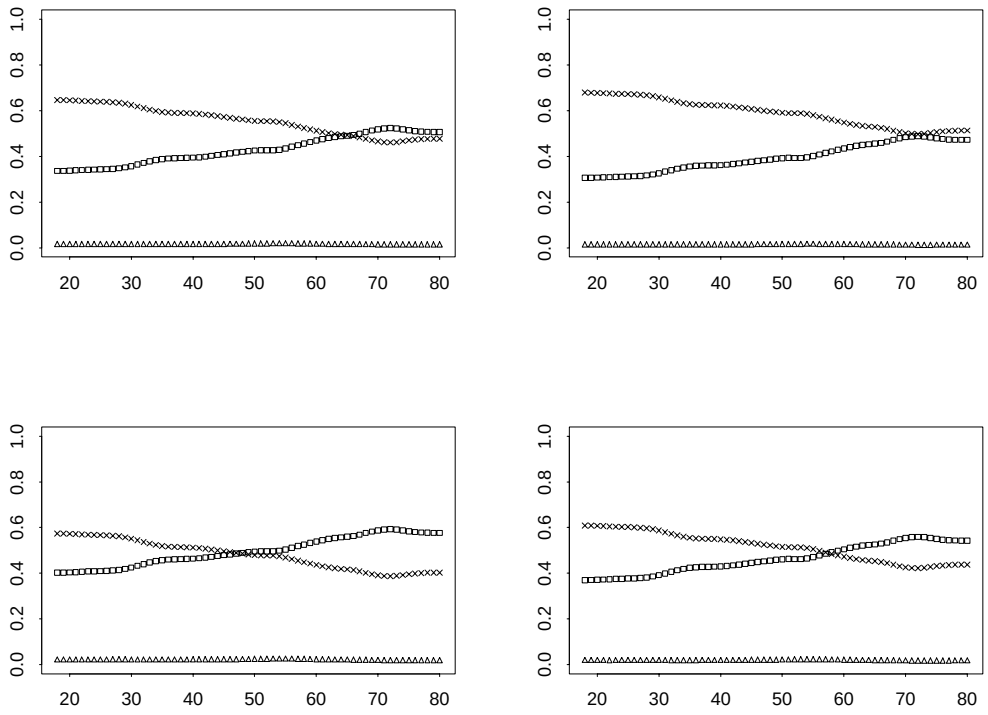

Figure 1: Estimated probabilities plotted against age, squares show preference for CDU, crosses for SPD and triangles for FDP. The subpopulations are western countries, women (top left), western countries, men (top right), eastern countries, women (bottom left), eastern countries, men (bottom right).

Appendix

The penalized log-likelihood has the form

$$\mathrm{pl}(\beta) = \sum_{i=1}^{n} l_i(y_i, \eta_i) - \frac{1}{2}\sum_{j=1}^{p}\sum_{r=1}^{q} \lambda_{jr}\gamma_{jr}^T K_{d_{jr}}\gamma_{jr} - \frac{1}{2}\sum_{j=1}^{m}\sum_{r=1}^{k} \lambda_{jr}^c \alpha_{jr}^T K_{d_{jr}^c}\alpha_{jr}$$

or equivalently

$$\mathrm{pl}(\beta) = \sum_{i=1}^{n} l_i(y_i, \eta_i) - \frac{1}{2}\sum_{j=1}^{p} \gamma_j^T \Lambda_j K_j \gamma_j - \frac{1}{2}\sum_{j=1}^{m} \alpha_j^T \Lambda_j^c K_j^c \alpha_j + \lambda_{jk}^c \alpha_{jk}^T K_{d_{jk}^c}\alpha_{jk}$$

where

$$\begin{aligned} K_j &= \mathrm{Diag}(K_{d_{j1}}, \ldots, K_{d_{jq}}), \\ \Lambda_j &= \mathrm{diag}(\lambda_{j1}1_{G_j}, \ldots, \lambda_{jq}1_{G_j}), \\ K_j^c &= \mathrm{Diag}(K_{d_{j1}^c}, \ldots, K_{d_{jq}^c}), \\ \Lambda_j^c &= \mathrm{diag}(\lambda_{j1}^c 1_{M_j}, \ldots, \lambda_{jm}^c 1_{M_j}). \end{aligned}$$

One obtains with $D_i = \frac{\partial h(\eta_i)}{\partial \eta}, \quad \mu_i = h(Z_i\beta)$

$$\begin{aligned} \partial \mathrm{pl}(\beta)/\partial\gamma_0 &= \sum_{i=1}^{n} D_i^T \Sigma_i^{-1}(y_i - \mu_i) \\ \partial \mathrm{pl}(\beta)/\partial\gamma_j &= \sum_{i=1}^{n} C_{ij}^T D_i^T \Sigma_i^{-1}(y_i - \mu_i) - \Lambda_j K_j \gamma_j \\ \partial \mathrm{pl}(\beta)/\partial\alpha_j &= \sum_{i=1}^{n} A_{ij}^T D_i^T \Sigma_i^{-1}(y_i - \mu_i) - \Lambda_j^c K_j^c \alpha_j \\ \partial \mathrm{pl}(\beta)/\partial\alpha_{jk} &= \sum_{i=1}^{n} A_{ijk}^T D_i^T \Sigma_i^{-1}(y_i - \mu_i) - \lambda_{jk}^c K_{d_{jk}^c}\alpha_{jk} \end{aligned}$$

The Pseudo-Fisher matrix is given by

$$\tilde{F} = E\left(-\frac{\partial^2 \mathrm{pl}(\beta)}{\partial\beta\partial\beta^T}\right) = (\tilde{F}_{rs})_{\substack{r=0,\ldots,p+2m \\ s=0,\ldots,p+2m}}$$

with

$$\tilde{F}_{00} = \sum_{i=1}^{n} D_i^T \Sigma_i^{-1} D_i$$

$$\tilde{F}_{r0} = \begin{cases} \sum_{i=1}^{n} C_{ir}^T D_i^T \Sigma_i^{-1} D_i, & r = 1, \ldots, p \\ \sum_{i=1}^{n} A_{il}^T D_i^T \Sigma_i^{-1} D_i, & r = p + 2l - 1,\ l = 1, \ldots, m\,, \\ \sum_{i=1}^{n} A_{ilk}^T D_i^T \Sigma_i^{-1} D_i, & r = p + 2l,\ l = 1, \ldots, m \end{cases} \qquad \tilde{F}_{0r} = \tilde{F}_{r0}^T$$

$$\tilde{F}_{ss} = \begin{cases} \sum_{i=1}^{n} C_{is}^T D_i^T \Sigma_i^{-1} D_i C_{is} + \Lambda_s K_s, & s = 1, \ldots, p \\ \sum_{i=1}^{n} A_{il}^T D_i^T \Sigma_i^{-1} D_i A_{il} + \Lambda_l^c K_l^c, & s = p + 2l - 1,\ l = 1, \ldots, m \\ \sum_{i=1}^{n} A_{ilk}^T D_i^T \Sigma_i^{-1} D_i A_{ilk} + \lambda_{lk}^c K_{d_{lk}^c}, & s = p + 2l,\ l = 1, \ldots, m \end{cases}$$

$$\tilde{F}_{rs} = \begin{cases} \sum_{i=1}^{n} C_{ir}^T D_i^T \Sigma_i^{-1} D_i C_{is}, & r = 2, \ldots, p,\ 0 < s < r \\ \sum_{i=1}^{n} A_{il}^T D_i^T \Sigma_i^{-1} D_i C_{is}, & r = p + 2l - 1,\ l = 1, \ldots, m, \\ & 0 < s \le p \\ \sum_{i=1}^{n} A_{ilk}^T D_i^T \Sigma_i^{-1} D_i C_{is}, & r = p + 2l,\ l = 1, \ldots, m, \\ & 0 < s \le p\,, \\ \sum_{i=1}^{n} A_{il}^T D_i^T \Sigma_i^{-1} D_i A_{it}, & r = p + 2l - 1,\ l = 2, \ldots, m, \\ & s = p + 2t - 1,\ t = 1, \ldots, l - 1 \\ \sum_{i=1}^{n} A_{ilk}^T D_i^T \Sigma_i^{-1} D_i A_{itk}, & r = p + 2l,\ l = 2, \ldots, m, \\ & s = p + 2t,\ t = 1, \ldots, l - 1 \\ \sum_{i=1}^{n} A_{ilk}^T D_i^T \Sigma_i^{-1} D_i A_{it}, & r = p + 2l,\ l = 1, \ldots, m, \\ & s = p + 2t - 1,\ t = 1, \ldots, l \\ \sum_{i=1}^{n} A_{il}^T D_i^T \Sigma_i^{-1} D_i A_{itk}, & r = p + 2l - 1,\ l = 2, \ldots, m, \\ & s = p + 2t,\ t = 1, \ldots, l - 1 \end{cases}$$

$$\tilde{F}_{sr} = \tilde{F}_{rs}^T$$

In closed form one has

$$\begin{aligned} \mathrm{ps}(\beta) &= \frac{\partial \mathrm{pl}(\beta)}{\partial \beta} = \sum_{i=1}^{n} Z_i^T D_i^T(\beta)\, \Sigma_i^{-1}(\beta)\, (y_i - \mu_i) - \Lambda K \beta \\ &= Z^T D^T(\beta)\, \Sigma^{-1}(\beta)\, (y - \mu) - \Lambda K \beta \\ \tilde{F}(\beta) &= F(\beta) + \Lambda K\,, \end{aligned}$$

where

$$F(\beta) = \sum_{i=1}^{n} Z_i^T D_i^T(\beta)\, \Sigma_i^{-1}(\beta)\, D_i(\beta)\, Z_i = Z^T D^T(\beta)\, \Sigma^{-1}(\beta)\, D(\beta)\, Z$$

is the usual Fisher matrix with $D_i(\beta) = \partial h(\eta_i)/\partial \eta$, $\Sigma_i(\beta) = \mathrm{cov}(y_i)$, $D = \mathrm{Diag}(D_i(\beta))$, $\Sigma = \mathrm{Diag}(\Sigma_i(\beta))$ and

$$\begin{aligned} \Lambda &= \mathrm{Diag}(\mathbf{0}, \Lambda_1, \ldots, \Lambda_p, \Lambda_1^c, \ldots, \Lambda_m^c)\,, \\ K &= \mathrm{Diag}(\mathbf{0}, K_1, \ldots, K_p, K_1^c, \ldots, K_m^c) \end{aligned}$$

represent the penalty term.

One obtains a version of the Fisher scoring in the form

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \tilde{F}(\hat{\beta}^{(k)})^{-1}\, \mathrm{ps}(\hat{\beta}^{(k)})\,.$$

References

Abe, M. (1999). A Generalized Additive Model for Discrete-Choice Data. *Journal of Business and Economic Statistics* **17**, 271–284.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible Smoothing with B-Splines and Penalties. *Statistical Science* **11**, 89–121.

Fahrmeir, L. and Tutz, G. (2000). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman & Hall.

Hruschka, H. (1998). Market share models with semi-parametric additive brand attractions. *Regensburger Diskussionsbeiträge zur Wirtschaftswissenschaft* **16**.

Hruschka, H., Fettes, W., and Probst, M. (1999). Artificial neural net - multinomial logit model - a semiparametric approach to analyse brand choice. *Regensburger Diskussionsbeiträge zur Wirtschaftswissenschaft* **336**.

Marx, B. D. and Eilers, P. (1998). Direct Generalized Additive Modelling with Penalized Likelihood. *Comp. Stat. & Data Analysis* **28**, 193–209.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behaviour. In P. Zarembka (Ed.), *Frontiers in Econometrics*. New York: Academic Press.

McFadden, D. (1981). Econometric models of probabilistic choice. In C. F. Manski & D. McFadden (Eds.), *Structural Analysis of discrete data with econometric applications*, pp. 198–272. Cambridge, Mass.: MIT-Press.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society* **A 135**, 370–384.

Yee, T. W. and Wild, C. J. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society* **B**, 481–493.