Tutz:

Generalized semiparametrically structured ordinal models

Sonderforschungsbereich 386, Paper 250 (2001)

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

Projektpartner

Generalized semiparametrically structured ordinal models

Gerhard Tutz
Ludwig-Maximilians-Universität München
Akademiestraße 1, 80799 München
tutz@stat.uni-muenchen.de

Summary

Semiparametrically structured models are defined as a class of models for which the predictors may contain parametric parts, additive parts of covariates with an unspecified functional form and interactions which are described as varying coefficients. In the case of an ordinal response the complexity of the predictor is determined by different sorts of effects. It is distinguished between global effects and category-specific effects where the latter allow that the effect varies across response categories. A general framework is developed in which global as well as category-specific effects may have unspecified functional form. The framework extends various existing methods of modeling ordinal responses. The Wilcoxon-Rogers notation is extended to incorporate smooth model parts and varying coefficient terms, the latter being important for the smooth specification of category-specific effects.

Keywords: Semiparametric model, ordinal regression, proportional odds model, varying coefficients, category-specific effects

1 Introduction

Since McCullagh's (1980) paper on regression models for ordinal data and the ensuing investigation of regression modeling with an ordered categorical response variable these models have become a standard tool in the statistical analysis of ordinal data. Basic models which have been studied are the cumulative type model

$$\mathrm{P}(Y \leq r|x) = F(\gamma_{0r} + x^T\gamma) \ , \ r = 1, \ldots, k-1, \tag{1}$$

the sequential type or continuation ratio model

$$\mathrm{P}(Y = r|Y \geq r|x) = F(\gamma_{0r} + x^T\gamma) \ , \ r = 1, \ldots, k-1, \tag{2}$$

and the adjacent-category type model

$$\mathrm{P}(Y = r|Y \in \{r,\ r+1\}, x) = F(\gamma_{0r} + x^T\gamma) \ , \ r = 1, \ldots, k-1, \tag{3}$$

where Y denotes the response variable taking values from ordered categories $1, \ldots, k$ and F is a fixed distribution function.

The most prominent members of these families are the proportional odds model (cumulative with F logistic) and the proportional hazards model (sequential with F logistic). Several extensions have been proposed where the effects of covariates or parts of the covariates are category-specific in the form $x'\gamma_r$ instead of $x'\gamma$, yielding partially proportional odds models(e.g. Cox,1988, Brant, 1990). The connections between the different types of model have been investigated (Läärä & Matthews,1985, Tutz,1991) and cumulative and sequential models have been compared extensively (e.g. Armstrong & Sloan,1989, Greenland,1994). Investigation of the adjacent-category models is found in Simon (1974), Goodman (1983). Overviews are found in Barnhart & Sampson (1994) and from a more general view in Agresti (1984), Agresti (1999).

All of these models may be considered within a closed framework as *multivariate generalized linear models* (MGLMs). The essential restriction is that the predictor is linear and therefore the functional form of the covariates is fixed in advance in a very simple parametric form which is convenient for statistical analysis. The rise of nonparametric regression techniques in recent decades makes much more flexible models accesible so that simple linear models seem almost like a crude (although often robust) first step in the modeling of regression data. Beside fully nonparametric models semiparametric models have been considered in the form of generalized additive models (Hastie & Tibshirani,1990), partially linear models (e.g. Green & Silverman,1994, Speckman,1988, Severini & Staniswalis,1994) and in the form of varying-coefficient models (Hastie & Tibshirani,1993). However, the focus of non- and semi-parametric methods has been on

onedimensional regression models, Gaussian or non-Gaussian, in the latter form as nonparametric extensions of univariate generalized linear models (GLMs).

Surprisingly little has been done in semiparametric ordinal modeling. When introducing generalized additive models Hastie & Tibshirani (1990) explicitly considered ordinal models by allowing an additive, but not necessarily linear functional form of predictors in the cumulative model. Yee & Wild (1996) derive the smoothing spline based estimators of Hastie & Tibshirani (1990) from the concept of vector splines. Partially linear models with a parametric term and an unspecific function of only one or two continuous variables have been investigated by Kauermann & Tutz (2000b) and varying-coefficients-models for ordinal data have been considered by Kauermann & Tutz (2000a).

In this paper a general framework for the extension of (multivariate) generalized linear models to incorporate nonparametric parts is given. Most of the non- and semiparametric approaches are based on an additive structure where the components may be functions of single covariates or interactions between two or more variables or may themselves have a multiplicative form, constructed from a variable and an unspecified function of another variable. For example an additive model is formed by a simple addition of unspecified functions of covariates. In a partially linear model the linear predictor is a sum of a linear term and an unspecified function whereas the basic structure of components in the the varying coefficient model is a multiplication of variables and smooth functions. It is straightforward to incorporate semiparametric parts into the predictor in the spirit of generalized linear additive smooth structures (Eilers & Marx, 1999). However, for ordinal models there is an additional structure which is important to model ordinal responses adequately, namely the modelling of effects that are specific for response categories. For a small number of response categories these effects may be modelled parametrically, for a large number of response categories, however, nonparametric modelling approaches are necessary. With the focus on ordinal models the challenge is how to specify smooth category-specific additive structures which are flexible enough to fit the data well but which also avoid overfitting and allow for robust estimation. The developed framework of

generalized semiparametrically structured models (GSSM) comprises all these approaches of specifying the predictor. In order to make the structure transparent the Wilkinson-Rogers notation is extended to incorporate semiparametric structures for multicategorical data.

In Section 2 multivariate GLMs as an unifying concept for ordinal models are considered and an alternative estimation concept based on penalized likelihood for basis-functions is developed for the ordinal additive model. In Section 3 the concept is extended to the modeling of category-specific effects. In Section 3.3 the extension of the Wilkinson-Rogers notation is outlined.

2 Ordinal additive models

A common framework for the ordinal models considered in the previous section is the multivariate generalized linear model. For observations $(Y_i,\ x_i),\ i = 1, \ldots, n$ with $Y_i \in \{1, \ldots, k\}$ and x_i a p-dimensional vector of covariates the models have the common form

$$\pi_i = h(Z_i\beta) \quad \text{or} \quad g(\pi_i) = Z_i\beta, \tag{4}$$

respectively, where $\pi_i^T = (\pi_{i1}, \ldots, \pi_{iq})$, $q = k - 1$, with components $\pi_{ir} = \mathrm{P}(Y_i = r|x_i)$ is the vector of response probabilities, $g : \mathbb{R}^q \mapsto \mathbb{R}^q$ is the link function and $h = g^{-1}$ is the inverse link or response function. The parameter vector is given by $\beta^T = (\gamma_{01}, \ldots, \gamma_{0q}, \gamma^T)$ and the design matrix by $Z_i = [I_{q\times q},\ 1_{q\times 1} \otimes x_i^T]$, where $I_{q\times q}$ is the unit matrix and $1_{q\times 1}^T = (1, \ldots, 1)$. While the linear predictor is the same the essential difference in models (1),(2),(3) is the use of different link functions which are easily derived (e.g. Fahrmeir & Tutz,2001). The link function depends on the type of the model (cumulative, sequential or adjacent categories) and the distribution function F. It is essentially a choice which determines the interpretation of parameters. In the cumulative model they may be interpreted in terms of an underlying metric regression model, for the sequential it is in terms of an Markov type process starting in response category 1 and stopping if one of the transitions to higher categories is not performed. For the adjacent categories model the parameters are interpreted locally as distinguishing between adjacent

categories.

2.1 Smooth components within the predictor

Apart from the choice of the link function the interesting part is the vector-valued linear predictor which determines how the covariates effect upon the response variables. The predictor $\eta_i^T = (\eta_{i1}, \ldots, \eta_{iq})$ in all of the models considered has components

$$\eta_{ir} = \gamma_{0r} + x_i^T \gamma$$

where the effect of x_i does not depend on the category.

For the extension to more structured models let the data be given by (Y_i, x_i, w_i) where $w_i^T = (w_{i1}, \ldots, w_{im})$ are additional covariates. Semiparametrically structured models are obtained by allowing the predictor to have the form

$$\eta_{ir} = \gamma_{0r} + x_i^T \gamma + \sum_{j=1}^{m} \alpha_{(j)}(w_{ij}). \tag{5}$$

where $\alpha_{(1)}, \ldots, \alpha_{(m)}$ are unspecified (global) functions. Thus one has partially linear models with $x_i^T \gamma$ representing the linear parametrization and a term which retains the additive form but does not specify the functional form of the components. For the cumulative type model with logit link and $\gamma = 0$ one obtains the generalized additive model considered by Hastie & Tibshirani (1990). But of course for all types of models this less restrictive modelling of the predictor is often more adequate in practical applications.

Before considering more general approaches a general method of estimation is considered which applies to all of the models (1),(2),(3) with semiparametrically structured predictor (5). The method is based on a penalized expansion in basis functions which can be seen as an extension of penalized regression splines. The approach is similar to smoothing splines but with fewer knots and a discrete roughness penalty. Having in mind that the function $\alpha_{(j)}$ may be a lth degree polynomial, let the function $\alpha_{(j)}(w)$ be given by regression splines in the truncated power series form

$$\alpha_{(j)}(w) = \alpha_{j1}^{(0)} w + \cdots + \alpha_{jl}^{(0)} w^l + \sum_{s=1}^{K_j} \alpha_{js} (w - w_{j(s)})_+^l \tag{6}$$

where $(w)_+^l = w^l I(w \geq 0)$ and $w_{j(1)} < \cdots < w_{j(K_j)}$ are fixed knots from the range of variable w_j and $(\alpha_{j1}^{(0)}, \ldots, a_{jl}^{(0)}, \alpha_{j1}, \ldots, \alpha_{jK_j})$ is a vector of regression coefficients (compare Ruppert & Carroll, 1999, Ruppert, 2000). The function given by (6) is a lth degree polynomial on each interval between two consecutive knots and $l-1$ derivatives are continuous on the whole range, but the lth derivative takes a jump of size $l!\alpha_{js}$ at knot $w_{j(s)}$. The number of knots K_j is chosen such that all functions of potential interest may be approximated. Ruppert (2000) demonstrates how a very limited number of knots yields approximations to quite complex functions which are visually not distinguishable from the underlying functions.

The truncated series (6) is natural if one suspects a polynomial form of the predictor but wants to allow for deviations from this parametric form. Alternatively one may use

$$\alpha_{(j)}(w) = \sum_{s=1}^{K_j} \alpha_{js} \Phi_{js}(w) \tag{7}$$

where Φ_{js} are basis functions connected to the knots $w_{j(s)}, s = 1, \ldots, K_j$, $\Phi_{js}, s = 1, \ldots, K_j$ may be choosen as B-splines (Eilers & Marx,1996) or radial basis functions of the form $\Phi_{js}(w) = \Phi(|w - w_{j(s)}|)$ where Φ is a fully specified function, e.g. the Gaussian kernel. Basis functions like B-splines seem to be more stable than the truncated power series as far as computation is concerned.

A common form for (6) and (7) is

$$\alpha_{(j)}(w) = \Phi_0(w) + \sum_{s=1}^{K_j} \alpha_{js} \Phi_{js}(w)$$

where $\Phi_0(w) = \sum_{s=1}^{l} \alpha_{js}^{(0)} w^s$ in (6) and $\Phi_0(w) = 0$ in (7). In (6) the basis functions are truncated lth degree polynomials.

Using (6) or (7) estimation is based on the maximization of the penalized log-likelihood

$$l = \sum_{i=1}^{n} l(Y_i;\, \eta_i) - \kappa(\alpha) \tag{8}$$

where $l(Y_i;\, \eta_i)$ is the log-likelihood contribution of the ith observation and $\kappa(\alpha)$ is a roughness penalty with $\alpha^T = (\alpha_1^T, \ldots, \alpha_m^T)$, $\alpha_j^T = (\alpha_{j1}, \ldots, \alpha_{jK_j})$. The rough-

ness penalty used here has the form

$$\kappa(\alpha) = \sum_{j=1}^{m} \lambda_j \sum_{s=1}^{K_j} (\Delta_s^d \alpha_{js})^2 \tag{9}$$

where Δ_s is the difference operator operating on adjacent coefficients of basis functions i.e. $\Delta_s \alpha_{js} = \alpha_{js} - \alpha_{j,s-1}, \Delta_s^2 \alpha_{js} = \Delta_s(\alpha_{js} - \alpha_{j,s-1}) = \alpha_{js} - 2\alpha_{j,s-1} + \alpha_{j,s-2}$, etc. In Δ_s the subscript s is used to indicate for which index the differences are built. The discrete penalty has strong connections to smoothing splines (see O'Sullivan, Yandell & Raynor, 1986). The form (9) has been used by Eilers & Marx (1996) for the B-spline basis. Ruppert & Carroll (1999) used the zero order difference $\Delta_s^0 \alpha_{js} = \alpha_{js}$, thus penalizing the jumps in the kth derivative of (6) by α_{js}^2.

Maximization of the penalized log-likelihood (8) may be performed by modification of the Fisher scoring as used in the framework of generalized multivariate models (see Appendix) In comparison to the backfitting algorithm used by Hastie & Tibshirani (1993) for the cumulative additive model it is less time consuming.

2.2 Application to retinopathy

In a 6-year follow up study on diabetes and retinopathy status reported by Bender & Grouven (1998) the interesting question is how the retinopathy status is associated with the risk factor smoking (SM = 1: smoker, SM = 0: non-smoker) adjusted for the risk factors diabetes duration (DIAB), glycosylated hemoglobin (GH) and diastolic blood pressure (BP). The fitted models have predictor

$$\eta_{ir} = \gamma_{0r} + SM \times \beta_{SM} + \alpha_D(DIAB) + \alpha_{GH}(GH) + \alpha_{BP}(BP)$$

with the retinopathy status measured on an ordinal scale having three categories. Table 1 shows the estimated parameters for various choices of the smoothing parameters where a B-spline basis of degree 2 with 22 equi-distant knots and a penalty of first difference was used. One sees that for the cumulative as well as for the sequential model estimates and standard errors are quite stable. Figure 1 shows the smooth effects of the three continuous variables for the sequential

		cumulative		sequential	
λ		est	std	est	std
3	β_{01}	-0.063	0.276	-0.041	0.262
	β_{02}	$+1.347$	0.279	$+0.342$	0.281
	SM	-0.260	0.196	-0.137	0.176
6	β_{01}	-0.120	0.244	-0.116	0.232
	β_{02}	$+1.265$	0.248	$+0.230$	0.253
	SM	-0.247	0.193	-0.126	0.173
12	β_{01}	-0.137	0.241	-0.148	0.205
	β_{02}	$+1.216$	0.219	$+0.153$	0.229
	SM	-0.235	0.190	-0.116	0.170
15	β_{01}	-0.135	0.205	-0.149	0.196
	β_{02}	$+1.207$	0.211	$+0.134$	0.221
	SM	-0.231	0.189	-0.113	0.169

Table 1: Parameter estimates for retinopathy status data

	cumulative		sequential	
coef	est	std	est	std
β_{01}	$+13.708$	1.372	$+12.160$	1.225
β_{02}	$+15.097$	1.401	$+12.504$	1.270
SM	-0.253	0.193	-0.128	0.172
BP	-0.067	0.014	-0.059	0.012
GH	-0.455	0.076	-0.416	0.068
DIAB	-0.371	0.061	-0.322	0.055
DIAB^2	$+0.006$	0.002	$+0.005$	0.001

Table 2: Estimates of parametric models for retinopathy status data

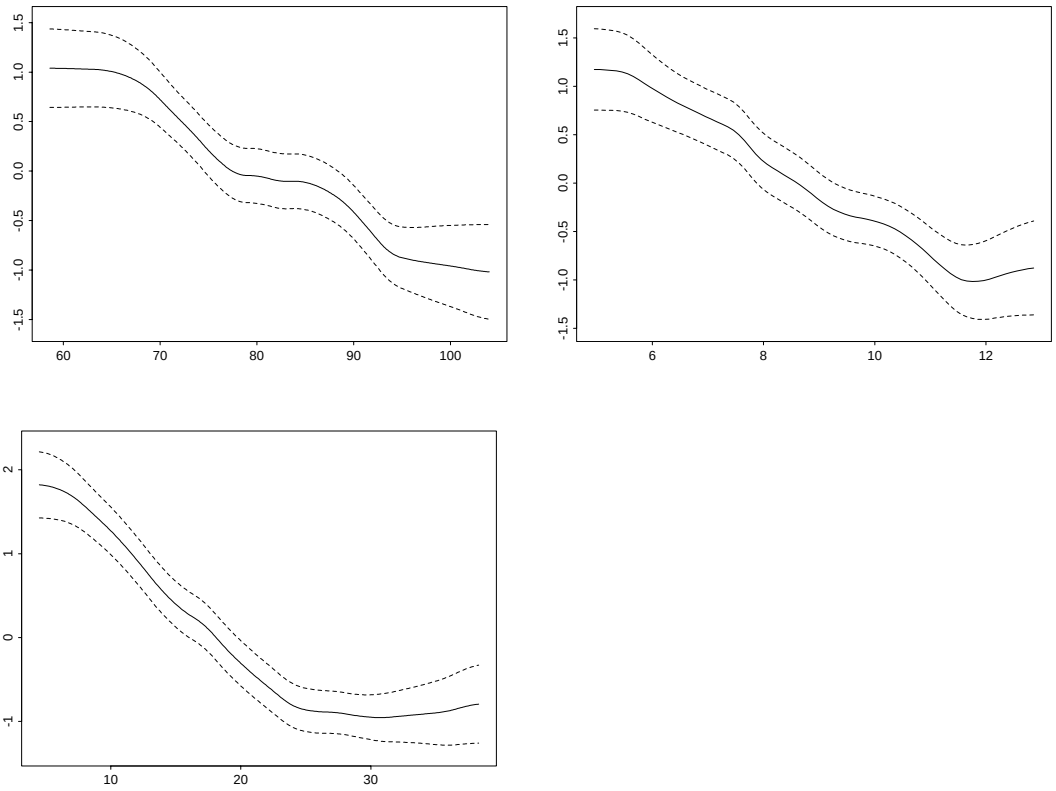

Figure 1: Smooth estimates for the sequential model for retinopathy status showing diastolic blood pressure (BP, top left), glycosylated hemoglobin (GH, top right) and duration of diabetes (DIAB, bottom left).

model. While blood pressure and glycosylated hemoglobin are approximately linear the estimated effect of duration of diabetes is nonlinear. With ongoing diabetes the probability for low scores on retinopathy status decreases. But after about 25 years the status remains stable. For comparison a parametric model is fitted where blood pressure and hemoglobin are linearly and duration is quadratically modelled. It is seen from Table 2 that the estimated effect of smoking status is comparable with the estimate based on the semiparametric model. Moreover, the quadratic effect of duration seems not be neglectable. The form of the smooth effects of the cumulative model (not given) are comparable with that of the sequential model, however on a different scale since effects have different interpretations.

3 Category-specific modeling of the predictor

Although the additive structure (5) is an important step into the direction of more flexible models, for ordinal models the challenge is the specification of category-specific additive structures which are flexible enough to fit the data well but which also avoid overfitting and allow for robust estimation. In a more general form than (5) the predictor has components

$$\eta_{ir} = \gamma_{0r} + x_i^T \gamma_r + \sum_{j=1}^{m} \alpha_{(jr)}(w_{ij})$$

where the parametric effect of x_i as well as the functions $\alpha_{(jr)}$ are category-specific meaning that the effect of the variables may have a different form for each of the response categories. When studying this model we start with the parametric term $x_i^T \gamma_r$ and consider the smooth term later. The reason for studying the parametric term in the context of semiparametric modeling is that in the case of many response categories it is useful and often necessary to specify it in a nonparametric way.

3.1 Category-specific parametric terms

In order to give some motivation for the more general form it is useful to look into the background of the cumulative model. The common derivation of the cumulative model (1) is based on an underlying latent variable $\tilde{Y}_i = -x_i^T \gamma + \varepsilon_i$ by assuming that Y is a coarser version of $\tilde{Y}$, or more precise $Y_i = r$ if $\gamma_{0,r-1} < \tilde{Y}_i \leq \gamma_{0r}$ where the parameters $-\infty = \gamma_{00} < \gamma_{01} < \cdots < \gamma_{0k} = \infty$ may be considered as thresholds on the latent scale. If one assumes that the noise variables ε_i follow distribution F one obtains the cumulative model

$$\mathrm{P}(Y_i \leq r | x_i) = F(\eta_{ir}),\ \eta_{ir} = \gamma_{0r} + x_i^T \gamma.$$

Due to the derivation from the latent regression model with slope γ the parameter γ is a so-called global effect that does not depend on the category. One obtains a property called strict stochastic ordering, meaning that if two populations characterized by covariates x_1, x_2 are compared then the difference

$$F^{-1}(Y \leq r | x_1) - F^{-1}(Y \leq r | x_2) = (x_1 - x_2)^T \gamma$$

does not depend on the category. If F is the logistic function the left hand side is

$$\log \frac{\mathrm{P}(Y \leq r|x_1)/\mathrm{P}(Y > r|x_1)}{\mathrm{P}(Y \leq r|x_2)/\mathrm{P}(Y > r|x_2)}$$

which is the logarithm of the proportion of cumulative odds. This property gives the model the name proportional odds model (McCullagh 1980). A similar property holds for the sequential and the adjacent type model. For the sequential type model it is given by $F^{-1}(\mathrm{P}(Y = r|Y \geq r,\ x_i) - F^{-1}(\mathrm{P}(Y = r| \geq r,\ x_2)) = (x_1 - x_2)^T\gamma$. However, strict stochastic ordering represents a strong restriction which is not always fulfilled in applications.

Extensions of the model have been given by considering predictors of the form

$$\eta_{ir} = \gamma_{0r} + x_i^T\gamma_r$$

where the parameter γ_r depends on the category. Since γ_r represents an effect which is category-specific, x_i is called a category-specific variable. The category-specific modeling of the cumulative type model is not derived from an underlying regression model. However, it still can be considered as a coarser version of a latent variable $\tilde{Y}_i = \varepsilon_i$ but now with the thresholds themselves given by $\gamma_{0r} + x_i^T\gamma_r$. Thus the covariates determine the thresholds instead of the slope of the regression. Since thresholds have to be ordered this has severe consequences for the cumulative model. The range of x-values where the model is valid, is strongly restricted since $\eta_{i1} \leq \eta_{i2} < \cdots < \eta_{iq}$ has to hold because probabilities are never negative. Thus one has to be careful if γ_r varies strongly across the response categories r. For the sequential and the adjacent categories model the generalization $\eta_{ir} = \beta_{0r} + x_i^T\gamma_r$ implies no restriction on the γ_r. However, for all types of models problems occur if the number of response categories is very high; for example if the sequential model is used in a discrete survival setting with the response measured in months the number of response categories may be 40 or 60. The effect is that the number of category-specific parameters γ_r is excessively high and only noise is fitted. Both problems demand to restrict the variability of γ_r across response categories r.

Often category-specific modeling cannot be avoided. If ordinal categories

are given by "healthy", "ill" and "dead" the transition between the first two categories will be governed by different mechanisms than the transition between the last two categories. The same holds for many categories in discrete duration models. For example if discrete time represents the response categories the effect of gender on the probability of finding a job usually decreases with the time of unemployment.

For the specification of the category-specific effects let the components of $\gamma_r^T = (\gamma_{1r}, \ldots, \gamma_{pr})$ as well as the intercept be split into

$$\gamma_{jr} = \gamma_{j.} + \bar{\gamma}_{jr}, \; j = 0, 1, \ldots, p \tag{10}$$

where for reasons of identifiability $\bar{\gamma}_{jr}$ is restricted by $\sum_{r=1}^{q} \bar{\gamma}_{jr} = 0$ or $\bar{\gamma}_{j1} = 0$. Then the predictor is given by

$$\eta_{ir} = \gamma_{0.} + \bar{\gamma}_{0r} + x_i^T \gamma_{.} + x_i^T \bar{\gamma}_r \tag{11}$$

where $\gamma_{.}^T = (\gamma_{1.}, \ldots, \gamma_{p.})$ is a global parameter and $\bar{\gamma}_r^T = (\bar{\gamma}_{1r}, \ldots, \bar{\gamma}_{pr})$ is the category-specific component. With the restriction $\gamma_{j1} = 0$ one has $\gamma_1 = (0, \ldots, 0)$.

Although one might estimate the category-specific effects γ_{jr} themselves in a nonparametric way, the separation into a global effect and a category-specific component has the advantage that smooth estimation of the latter form keeps the basic global effect constant. The components $\bar{\gamma}_{jr}$ represent just the (smoothly estimated) deviation from the parametric model with global effect $\gamma_{.}$. In order to obtain a smooth function over the response categories one specifies the category-specific components by

$$\bar{\gamma}_{jr} = \sum_{s=1}^{K_j} \tilde{\gamma}_{js} \Phi_{js}(r), \; j = 0, \ldots, p. \tag{12}$$

Depending on the number of knots used in(12) the category-specific components $\bar{\gamma}_{jr}$ are more or less smooth (across the discrete response categories r$\in \{1, \ldots, k\}$). However, in the spirit of penalized spline regression it is suggested that many knots are used and the amount of smoothing is determined by the penalization.

It should be noted that in (10) and(12) index j starts with 0, so that the intercept may also be estimated smoothly. Thereby implicitly for the intercept the

definition $\gamma_{0r} = \bar{\gamma}_{0r}$ is used. The smooth modeling of the intercept is important, since even if the covariates are considered global with parametric term $x_1^T \gamma_.$, for many categories the intercept which corresponds to the baseline hazard in duration models has to be smoothed. The corresponding penalized likelihood uses the roughness penalty

$$\kappa\left(\{\gamma_{jrs}\}\right) = \sum_{j=1}^{q} \lambda_j \left(\Delta_s^d \tilde{\gamma}_{js}\right)^2 .$$

If the number of response categories is less than 20 a simple choice of knots are the categories itself. Then for B-splines of first order (12) does not imply any restriction because $\Phi_s(s) = 1$, $\Phi_s(r) = 0$, $r \neq 0$ and therefore $\bar{\gamma}_{jr} = \tilde{\gamma}_{jr}$. However, the penalty, e.g. with $d = 1$, takes the form $\Sigma_j \lambda_j (\tilde{\gamma}_{js} - \tilde{\gamma}_{j,s-1})^2$ which, depending on the amount of smoothing imposed by λ_j, yields smooth variation across categories. For the extreme case $\lambda_j \to \infty$ one obtains $\tilde{\gamma}_{jr} = 0$ and therefore $\bar{\gamma}_{jr} = 0$.

3.2 Category-specific smooth functions

For continuous covariates the problem of specifying smooth functions which are also category-specific is more complex. In the general case the penalized basis functions approach uses the form

$$\alpha_{(jr)}(w) = \sum_{s=1}^{K_j} \alpha_{j0s} w^s + \sum_{s=1}^{K_j} \alpha_{jrs} \Phi_{js}(w) \tag{13}$$

where the polynomial term may be omitted by setting $\alpha_{j0s} = 0$, $j = 1, \ldots, m$, $s = 1, \ldots, K_j$. For simplicity the knots in (13) do not depend on the category. However, if the number of knots is chosen from a range between 10 and 30 for most functions of practical interest the flexibility is high enough without using different knot positions. The penalty term in the likelihood has to be modified since the weights α_{jrs} now depend on the response category r and the knot s. We suggest to use the penalized log-likelihood

$$l = \sum_{i=1}^{n} l\left(Y_i,\, \eta_i\right) + \sum_{j=1}^{m} \lambda_j \sum_{r=1}^{q} \sum_{s=d+1}^{K_j} \left(\Delta_s^d \alpha_{jrs}\right)^2 + \sum_{j=1}^{m} \delta_j \sum_{s=1}^{K_j} \sum_{r=d+1}^{q} \left(\Delta_r^d \alpha_{jrs}\right)^2 \tag{14}$$

where the subscript in Δ again refers to the index which is used in the construction of differences. Thus the differences are given by

$$\begin{aligned} \Delta_s \alpha_{jrs} &= \alpha_{jrs} - \alpha_{jr,\,s-1}, \\ \Delta_r \alpha_{jrs} &= \alpha_{jrs} - \alpha_{j,\,r-1,\,s}\ . \end{aligned}$$

The first penalty term penalizes the variation of adjacent knots and thus determines the smoothness of the function $\alpha_{jr}(w)$ for fixed response category r, i.e. the effect of covariates with respect to category r. The second penalty term reflects the variation of the curves $\alpha_{jr}(w)$, $r = 1, \ldots, q$ over response categories. For the extreme penalty $\delta_j \to \infty$ one obtains $\alpha_{jrs} = \alpha_{j.s}$ which does not depend on the category. In that case the simple additive model is fitted. For $\lambda_j \to \infty$ one has in the limit $\alpha_{jrs} = \alpha_{jr.}$ and therefore a category specific *parametric* model where the resulting form depends on the basis functions which are used.

When fitting model (13) the number of parameters is strongly increased. For the weights of basis functions one has now $(K_1 + \cdots + K_m)q$ parameters as compared of $K_1 + \cdots + K_m$ in (7). If the amount of smoothing is strong enough then there are no problems in estimating the parameters although the dimension of the estimation procedure is increased. A more severe drawback of model (13) is interpretation. Although a model based on (13) is extremely flexible the price is that the interpretation of the effects suffers. In particular for a high number of response categories one has as many curves as response categories to evaluate the effect of a variable. The straightforward interpretation of effects which is one of the strongest advantages of the semiparametric model (5) is lost. While model (5) makes use of the ordinal response categories to obtain the same influence structure as in an univariate response model like the binary logit model, this advantage is lost by the general model (13) which in this general form does not use that the response is ordinal. It is not a parsimonious model in the spirit of Occam's razor.

In order to obtain a model which allows a simple interpretation of effects let α_{jrs} be decomposed into

$$\alpha_{jrs} = \bar{\alpha}_{jr}\alpha_{js}$$

where only the first term depends on the response category r (apart from the dependence on the variable j). Omitting the polynomial term one obtains from (13)

$$\begin{aligned} \alpha_{(jr)}(w) &= \bar{\alpha}_{jr} \sum_{s=1}^{K_j} \alpha_{js} \Phi_{js}(w) \\ &= \bar{\alpha}_{jr} \alpha_{(j)}(w) \end{aligned} \tag{15}$$

where $\alpha_{(j)}(w) = \sum_s \alpha_{js} \Phi_{js}(w)$ has the same form as in simple additive models. The simple structure (15) retains a global additive term that represents the basic effect of variable w_j on the response. By centering $\alpha_{(j)}$ around zero the constants $\bar{\alpha}_{jr}$ represent factors which intensify or flatten the basic curve reflecting whether the basic effect $\alpha_{(j)}(w)$ increases or decreases over response categories. Identifiability demands some restrictions e.g. $\bar{\alpha}_{j1} = 1$. The penalty now includes the terms

$$\sum_{j=1}^{m} \lambda_j \sum_{s=d+1}^{K_j} \left(\Delta_s^d \alpha_{js}\right) \quad \text{and} \quad \sum_{j=1}^{m} \delta_j \sum_{r=d+1}^{q} \left(\Delta_r^d \bar{\alpha}_{jr}\right)^2 .$$

The fixed term is the same as in additive modeling given in (9), the second term corresponds to the category-specific parameters $\bar{\alpha}_{jr}$ which are penalized with smoothing parameter δ_j. The number of parameters used in (15) is strongly reduced in comparison to the general category-specific model (13). But, regretably the predictor (15) does not yield a GLM since parameters are connected by multiplication. This makes a different fitting procedures necessary. However, this cannot be circumvented by additive decomposition since an additive decomposition of the form $\alpha_{jrs} = \bar{\alpha}_{jr} + \alpha_{js}$ yields the simple model where $\alpha_{(jr)} = \alpha_{(j)}$ because shifting of the influence term is always confounded with the intercepts. Thus the multiplicative form is necessary what may already be seen from the special case where $\alpha_{(j)}$ is a linear function since then one obtains $\alpha_{(jr)}(w) = \bar{\alpha}_{jr} w$ which corresponds to model (10).

3.3 An extended Wilkinson-Rogers notation

In the following Wilkinson-Rogers coding of influential terms is extended to account for ordinal models, which contain category-specific effects as well as smooth

structures. We will start with parametric terms and then proceed to nonparametric structures. The usual Wilkinson-Rogers notation uses "$*$" for hierarchical interaction terms and "." for single interaction terms. Thus the term $x_1 * x_2$ means that the interaction $x_1.x_2$ between the variables x_1 and x_2 as well as the main effects x_1 and x_2 are found in the predictor. If x_1 and x_2 are categorical variables with x_1, x_2 having k_1, k_2 values, respectively, then $x_1 * x_2$ means that the predictor includes $\sum_{j=1}^{k_1-1} x_1^{(j)}\gamma_1^{(j)} + \sum_{j=1}^{k_2-1} x_2^{(j)}\gamma_2^{(j)} + \sum_{j_1=1}^{k_1-1}\sum_{j_2=1}^{k_2-1} x_1^{(j1)}x_2^{(j2)}\gamma_{12}^{(j_1 j_2)}$ where $x_1^{(j)}$, $x_2^{(j)}$ are dummy variables, e.g. in 0-1 coding with $x_1^{(j)} = 1$ if $x_1 = j$, $x_1^{(j)} = 0$ if $x_1 \neq j$. The first term represents x_1, the second x_2 and the third $x_1.x_2$.

In the ordinal setting the multinomial response $Y_i \in \{1, \ldots, k\}$ actually is a multivariate response represented by $y_i^T = (y_{i1}, \ldots, y_{iq})$, $q = k - 1$ with $y_{ir} = 1$ if $Y_i = r$, $y_{ir} = 0$ if $Y_i \neq r$. Since the predictor $\eta_{ir} = \gamma_{0r} + x_i^T\gamma$ may also be written as $\eta_{ir} = y_i^T\gamma_0 + x_i^T\gamma$ with $\gamma_0^T = (\gamma_{01}, \ldots, \gamma_{0q})$ it is quite natural to use the Wilkinson-Rogers abbreviation $y + x_1 + \cdots + x_p$ for this predictor. Although at first sight it seems strange that the response variable occurs in the predictor, since y is a categorical, multinomially distributed response it corresponds exactly to the parameterization.

The more general model $\eta_{ir} = \gamma_{0r} + x_i^T\gamma_r = y_i^T\gamma_0 + x_i^T\gamma_r$ seems to be nonhierarchical. But consider the reparametrization (10) given by $\gamma_{jr} = \gamma_{j.} + \bar{\gamma}_{jr}$ for the components of $\gamma_r^T = (\gamma_{1r}, \ldots, \gamma_{pr})$ with restriction $\sum_r \bar{\gamma}_{jr} = 0$ or $\gamma_{j1} = 0$. Then with $\bar{\gamma}_r^T = (\gamma_{1r}, \ldots, \gamma_{pr})$, $\gamma_.^T = (\gamma_{1.}, \ldots, \gamma_{p.})$ one obtains $\eta_{ir} = y_i^T\gamma_0 + x_i^T\bar{\gamma}_r + x_i^T\gamma_. = y_i^T\beta_0 + x_i^T\gamma + y_{ir}x_i^T\bar{\gamma}_r$. This shows that the model is hierarchical and the corresponding abbreviation is $y * x_1 + \cdots + y * x_p$ or in expanded form $y + x_1 + \cdots + x_p + y.x_1, \ldots, y.x_p$.

In Wilkinson-Rogers notation the terms are listed which yield the parametric predictor. In order to distinguish between a parametric term and an unspecified functional form of a variable we suggest to use brackets for the latter. Thus the simple additive model with predictor $\gamma_{0r} + \sum_j \alpha_{(j)}(w_j)$ is abbreviated by $y + (w_1) + \cdots + (w_m)$. A more difficult case arises if the category-specific part is considered as smooth. In order to obtain flexible models of low dimension in Section 3.1, equ (12), the parameters in the category-specific term $x_i^T\bar{\gamma}_r$ have been

	Predictor η_{ir}	Abbreviation
Parametric	$\gamma_{0r} + x^T\gamma$	$y + x_1 + \cdots + x_p$
	$\gamma_{0r} + x^T\gamma_r$	$y + x_1 + \cdots + x_p + y.x_1 \cdots + y.x_p$ or $y * x_1 + \cdots + y * x_p$
Additive	$\gamma_{0r} + x^T\gamma_{\cdot} + x^T\bar{\gamma}_r$ with $\bar{\gamma}_r$ smooth	$y + x_1 + \cdots + x_p + x_1.(y) + \cdots + x_p.(y)$
	With γ_{0r} and $\bar{\gamma}_r$ smooth	$(y) + x_1 + \cdots + x_p + x_1.(y) + \cdots + x_p.(y)$
	$\gamma_{0r} + \sum_{j=1}^{m} \alpha_{(j)}(w_j)$	$y + (w_1) + \cdots + (w_m)$
	$\gamma_{0r} + \sum_{j=1}^{m} \alpha_{(jr)}(w_j)$	$y + (y.w_1) + \cdots + (y.w_m)$
	$\gamma_{0r} + \sum_{j=1}^{m} \bar{\alpha}_{jr}\alpha_{(j)}(w_j)$	$y + y.(w_1) + \cdots + y.(w_m)$
	Same with $\bar{\alpha}_{jr}$ smooth	$y + (y).(w_1) + \cdots + (y).(w_m)$
Varying coefficients	$\gamma_{0r} + \sum_{j=1}^{t} \alpha_{(j)}(u_j) + \sum_{j=1}^{t} v_{ij}\nu_{(j)}(u_j)$	$y + u_1 + \ldots u_t + v_1.(u_1) + \cdots + v_t.(u_t)$

Table 3: Extended Wilkinson-Rogers notation for semiparametrically structured ordinal models

smoothed across the categories. Because this corresponds to a smooth modelling of y within the interactions the corresponding model is abbreviated by $y+x_1+\cdots+x_p+x_1.(y)+\cdots+x_1.(y)$. This should not be confounded with the model where the total interaction is considered a smooth unknown function. The corresponding model $\beta_{0r} + \sum_j \alpha_{(jr)}(w_{ij})$ has abbreviation $y + (y.w_1) + \cdots + (y.w_m)$. The case of interactions where only one component is considered smooth is also found in the modelling of varying-coefficients (see next section). The nonparametric modeling of the intercept which may be incorporated in any model is abbreviated by (y) which is in total accordance with the notation. An overview of the extended notation, including varying-coefficients is given in Table 3.

3.4 Application to injuries of the knee

In a clinical study focusing on the healing of sports related injuries of the knee $n = 123$ patients have been treated. By random design one of two therapies were chosen. In the treatment group an anti-inflammatory spray was used while in the placebo group a spray without active ingredients was used. After ten days of treatment with the spray, the mobility of the knee was investigated in a standardized experiment during which the knee was actively moved by the patient. The pain Y occurring during the movement was assesed on a four point scale ranging from 1 for no pain to 4 giving severe pain (for further information on the data see Tutz, 2000). In addition to treatment the covariate age was measured. Since the form of the influence of age is unknown it is modelled as a smooth function by using B-splines on a grid of 20 equidistant knots. We consider the cumulative and sequential logit model with predictor

$$\eta_{ir} = \gamma_{0r} + \text{treatment} \times \gamma_T + \alpha_r(\text{age})$$

where treatment is a binary variable with 1 denoting treatment groups and 0 denoting placebo groups. The models have the structure $y+$ treatment $+(y.\text{treatment})$. Since the effect of age may depend on the category one obtains a curve for the effect of age for each response category. For the cumulative model the restriction $\eta_{ir} \leq \eta_{i,r+1}$ has to be fulfilled which causes serious troubles when fitting the model for smaller values of the smoothing parameters. For the sequential model predictors have not to be ordered and thus the model may be fitted also when smoothing produces a fit closer to the data. Figure 2 shows the fit for the sequential model together with the fit of the simple model

$$\eta_{ir} = \gamma_{0r} + \text{treatment} \times \gamma_T + \alpha(\text{age})$$

which assumes that the effect does not depend on the response category. Since the curves are rather similar, the simpler model seems adequate for this data set. This holds in particular when looking at the confidence intervals given for the estimated curve of the simpler model.

The effect of therapy is not affected by the category specific modelling of age. For the cumulative model obtains $\gamma_T = 0.971$ with standard error 0.336 (global,

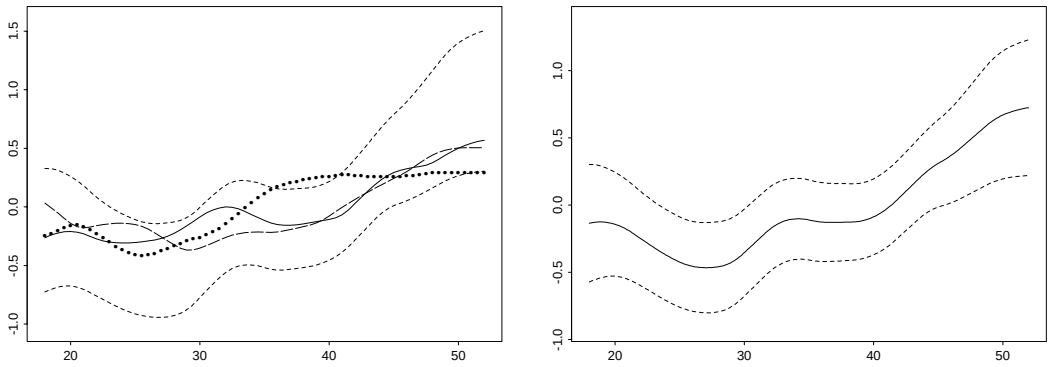

Figure 2: Effect of age for knee injury data in the sequential model (left: category-specific effects of age with confidence intervals of the global effect, right: global effect of age)

$\lambda = 10$) in comparison to $\gamma_T = 0.949$ with standard error 0.335 (category specific, $\lambda = 10$); for the sequential one obtains $\gamma_T = 0.918$ with standard error 0.276 (global, $\lambda = 10$) and $\gamma_T = 0.918$ with standard error 0.274 (category-specific, $\lambda = 10$). The estimates for the global model are given in Table 4.

model	coefficient	estimate	standard deviation
cumulative	γ_{01}	-1.386	0.296
	γ_{02}	-0.124	0.264
	γ_{03}	$+0.891$	0.272
	γ_T	$+0.959$	0.339
sequential	γ_{01}	-1.353	0.272
	γ_{02}	-0.807	0.273
	γ_{03}	-0.293	0.302
	γ_T	$+0.918$	0.276

Table 4: Parameter estimates for knee injury data

4 Smooth modelling of interactions

4.1 Varying-coefficients modelling

As a last step we will introduce a simple form of nonparametric interaction in the form of varying coefficient models. Omitting the parametric and additive terms and considering additional covariates (v_{ij}, u_{ij}), $i = 1, \ldots, n$, $j = 1, \ldots, t$ a global

varying coefficient term is given by

$$\eta_{ir} = \sum_{j=1}^{t} v_{ij}\nu_{(j)}(u_{ij}). \tag{16}$$

where the effect of variable v_{ij} is modified in a non-specified functional form by the effect modifier u_{ij} (see Hastie & Tibshirani, 1993). The advantage of specifying the interaction between a categorical or continuous variable v and a continuous variable u is that the modification may be investigated by plotting $\nu(u)$. The more general alternative, namely to incorporate a smooth interaction $\alpha(v, u)$ yields a two-dimensional form which is much harder to interpret. Without going into any details it is obvious that the unspecified functions $\nu_{(1)}, \ldots \nu_{(t)}$ may be modeled by basis functions. The abbreviation for (16) given in Table 3 is $v_{1.}(u_1), \ldots, v_{t.}(u_p)$ illustrating that it is a semiparametric model by retaining the form of a multiplication between variables and parameters but with the parameter being a function of non-specified form.The similarity in abbreviated notation to the category-specific modelling reveals the connection between the modelling approaches. Category-specific smooth effects may be seen as varying coefficients models where the interaction between the response and a continuous variable is modeled.

It is noteworthy that if (16) is part of the linear predictor, then the smooth functions $(u_1), \ldots (u_p)$ should also be incorporated in order to avoid identification problems (see discussion in Hastie & Tibshirani, 1993).

4.2 Applications to recovery scores

Davis (1991) investigated data collected in a randomized study in which 60 children undergoing surgery were treated with one of four dosages of an anaesthetic. Upon admission to the recovery room and at 5, 15 and 30 following admissions, recovery scores were assigned on a categorical scale ranging from 1 (least favourable) to 7 (most favourable). From the four repetitions of a variable having 7 categories only the first one is used here. In addition to the response covariates were considered, namely dose (4 categories: 15 mg/kg, 20 mg/kg, 25 mg/kg, 30 mg/kg), age ($9 - 70$ months), and duration ($35 - 190$ minutes). For the data see Davis

	cumulative		sequential	
coef	est	std	est	std
β_{01}	-0.988	0.527	-1.161	0.465
β_{02}	$+0.993$	0.511	$+0.490$	0.443
β_{03}	$+1.595$	0.533	-0.196	0.553
β_{04}	$+2.517$	0.591	$+0.726$	0.616
β_{05}	$+3.001$	0.640	$+0.109$	0.828
β_{06}	$+3.459$	0.707	$+0.284$	0.983
do[1]	-0.842	0.674	-0.574	0.509
do[2]	-0.836	0.678	-0.577	0.512
do[3]	-0.271	0.672	-0.122	0.514

Table 5: Estimates of parameters for recovery scores data (cumulative and sequential model)

(1991). Since dose is given in four levels it is considered as a categorical covariate with four categories. When analyzing the data as repeated measurements Davis (1991) found little evidence of significant effects of dosage. However, this effect may be due to missing interaction effects. Within a parametric model including interactions dose.age and dose.dur one finds that dose.dur may be omitted (likelihood ratio test yields 2.24 on a 3 degrees of freedom) but a further reduction by omitting dose.age is not adequate (likelihood ratio test yields 12.04 on 3 degrees of freedom). Thus the parametric modelling suggests to include interaction effects. Allowing age and duration to have an unknown functional form the model with predictor

$$\eta_{ir} = \gamma_{0r} + \text{dose}_1 * \gamma_{d1} + \text{dose}_2 * \gamma_{d2} + \text{dose}_3 * \gamma_{d3} + \alpha_{(a)}(\text{age}) + \alpha_{(d)}(\text{duration}) + \text{dose}_1 * u_{(1)}(\text{age}) * \text{dose}_2 * u_{(2)}(\text{age}) + \text{dose}_3 * u_{(3)}(\text{age})$$

is fitted where dose_1, dose_2, dose_3 are (0-1)-dummy variables. The underlying structure is that of the varying coefficient model

$$y + \text{dose} + (\text{age}) + (\text{duration}) + \text{dose}.(\text{age}).$$

Table 5 gives the estimated parameters for the cumulative and sequential models and Figure 3 and 4 show the smooth effects. The smooth effects are only given for the cumulative model since they are quite similar for both types of models. The

duration of surgery increases the probability for low categories on the recovery scale with a linear trend in the middle of the data and flat ends at the boundaries. From parametric modelling it has been inferred that the interaction between dose and age should not be omitted. It is interesting to see from Figure 4 that these effects are not linear. For two of the categories of dose 40 months seem to be some sort of a change point with an increase of the effect above 40 months or below 40 months. Only for dose = 3 the effect is rather flat around zero.

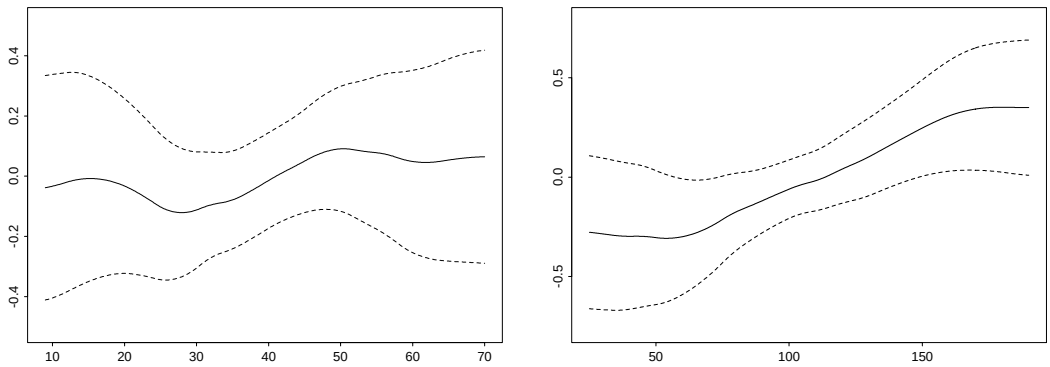

Figure 3: Smooth effects of cumulative model for surgery data (left: main effect of age, right: main effect of duration)

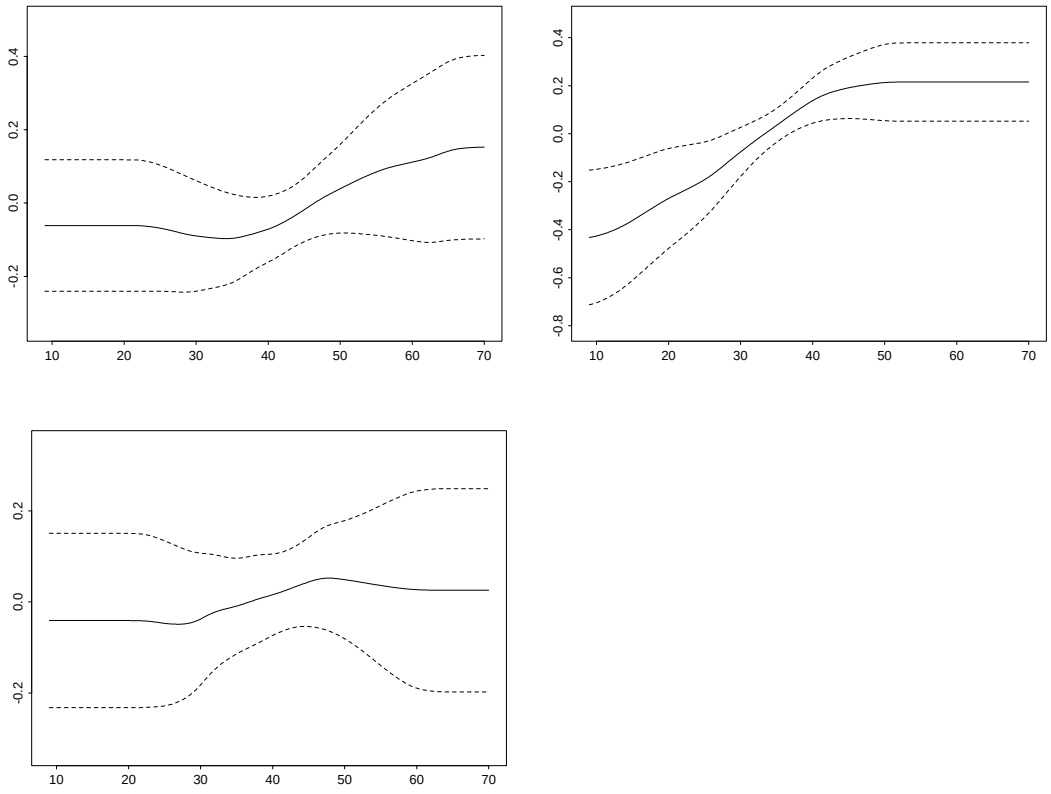

Figure 4: Smooth effects (interaction of duration and dose) of cumulative model for surgery data 1 (top left), 2 (top right), 3 (bottom left).

Appendix

In the following additional indices are used to distinguish between basic functions and number of knots for parametric, smooth additive and varying coefficient terms.

A.1 Models with additive effects

Let us first consider the model which includes potentially smooth category specific effects of categorical variables $x_1, \ldots, x_p$, additive effects of $w_1, \ldots, w_p$ and varying coefficients. The predictor of the corresponding model $x_1.(y), \ldots, x_p.(y), (w_1), \ldots, (w_m), v_1.(u_1), \ldots v_t.(u_t)$ has the form

$$\begin{aligned}\eta_{ir} &= \gamma_0. + x_i^T \gamma. + x_i^T \bar{\gamma}_r + \alpha_{(j)}(w_{ij}) + \sum_{j=1}^{t} v_{ij} \nu_{(j)}(u_{ij}) \\ &= \gamma_0. + x_i^T \gamma. + \sum_{j=1}^{p} x_{ij} \sum_{s=1}^{K_j(\gamma)} \tilde{\gamma}_{js} \Phi_{js}^{(\gamma)}(r) + \sum_{s=1}^{K_j(\alpha)} \alpha_{js} \Phi_{js}^{(\alpha)}(w_{ij}) + \sum_{j=1}^{t} v_{ij} \sum_{s=1}^{K_j(\nu)} \nu_{js} \Phi_{js}^{(\nu)}(v_{ij})\end{aligned}$$

with $\bar{\gamma}_1 = 0$. One obtains

$$\eta_{ir} = z_{ir}^T \beta$$

where

$$z_{ir} = \left(1, x_i, x_{i1}\Phi_1^{(\gamma)}(r), \ldots, x_{ir}\Phi_p^{(\gamma)}(r), \Phi_1^{(\alpha)}(w_{i1}), \ldots, \Phi_m^{(\alpha)}(w_{ip}), v_{i1}\Phi_1^{(\nu)}(u_{i1}), \ldots, v_{it}\Phi_t^{(\nu)}(u_{it})\right)$$

with

$$\begin{aligned}\Phi_j^{(\gamma)}(r) &= \left(\Phi_{j1}^{(\gamma)}(r), \ldots, \Phi_{jK_j(\gamma)}^{(\gamma)}(r)\right)^T, j = 1, \ldots, p, \\ \Phi_{js}(1) &= 0, j = 1, \ldots, p, \; s = 1, \ldots, K_j(\gamma), \\ \Phi_j^{(\alpha)}(w_{ij}) &= \left(\Phi_{j1}^{(\alpha)}(w_{ij}), \ldots, \Phi_{jK_j(\alpha)}^{(\alpha)}(w_{ij})\right)^T, \; j = 1, \ldots, p, \\ \Phi_j^{(\nu)}(u_{ij}) &= \left(\Phi_{j1}^{(\nu)}(u_{ij}), \ldots, \Phi_{jK_j(\nu)}^{(\nu)}(u_{ij})\right)^T, \; j = 1, \ldots, t,\end{aligned}$$

and

$$\beta^T = \left(\gamma_{0.}, \gamma_{.}, \tilde{\gamma}_1^T, \ldots, \tilde{\gamma}_p^T, \alpha_1^T, \ldots, \alpha_m^T, \nu_1^T, \ldots, \nu_t^T\right)$$

with

$$\begin{aligned}
\tilde{\gamma}_j^T &= (\tilde{\gamma}_{j1}, \ldots, \tilde{\gamma}_{jK_j(\gamma)}),\ j = 1, \ldots, p, \\
\alpha_j^T &= (\alpha_{j1}, \ldots, \alpha_{jK_j}),\ j = 1, \ldots, m, \\
\nu_j^T &= (\nu_{j1}, \ldots, \nu_{jK_j}),\ j = 1, \ldots, t.
\end{aligned}$$

Thus one obtains for the predictor $\eta_i^T = (\eta_{i1}, \ldots, \eta_{iq})$

$$\eta_i = Z_i \beta$$

where

$$Z_i = \left(1, 1 \otimes x_i^T, \Phi_{i1}^{(\gamma)}, \ldots, \Phi_{ip}^{(\gamma)}, \Phi_{i1}^{(\alpha)}, \ldots, \Phi_{im}^{(\alpha)}, \Phi_{i1}^{(\nu)}, \ldots, \Phi_{it}^{(\nu)}\right)$$

with

$$\begin{aligned}
\Phi_{ij}^{(\gamma)} &= \begin{pmatrix} x_{ij}\left(\Phi_j^{(\gamma)}(1)\right)^T \\ \vdots \\ x_{ij}\left(\Phi_j^{(\gamma)}(q)\right)^T \end{pmatrix}, j = 1, \ldots, p, \\
\Phi_{ij}^{(\alpha)} &= 1 \otimes \left(\Phi_j^{(\alpha)}(w_{ij})\right)^T,\ j = 1, \ldots, m, \\
\Phi_{ij}^{(\nu)} &= 1 \otimes \left(v_{ij}\Phi_j^{(\nu)}(u_{ij})\right)^T,\ j = 1, \ldots, t.
\end{aligned}$$

The special case of unrestricted category-specific variables $x_1, \ldots, x_p$ results if for the number of knots $K_j(\gamma) = q$ is chosen and the parameters $\tilde{\gamma}_{js}$ are not penalized. The case of global variables $x_1, \ldots, x_p$ results if $\Phi_{ij}^{(\gamma)}$ is omitted. The penalized log likelihood has the form

$$l = \sum_{i=1}^{n} l(y_i, \eta_i) + r(\tilde{\beta})$$

where $\tilde{\beta}^T = (\tilde{\gamma}_1^T, \ldots, \tilde{\gamma}_p^T, \alpha_1^T, \ldots, \alpha_m^T, \nu_1^T, \ldots, \nu_t^T)$ represents the penalized parameters with

$$r(\tilde{\beta}) = \sum_{j=1}^{p} \lambda_j^{(\gamma)} \sum_{s=d+1}^{K_j(\gamma)} \left(\Delta_s^d \gamma_{js}\right)^2 + \sum_{j=1}^{m} \lambda_j^{(\alpha)} \sum_{s=d+1}^{K_j(\alpha)} \left(\Delta_s^d \alpha_{js}\right)^2 + \sum_{j=1}^{t} \lambda_j^{(\gamma)} \sum_{s=d+1}^{K_j(\nu)} \left(\Delta_s^d \nu_{js}\right)^2$$
$$= \sum_{j=1}^{p} \lambda_j^{(\gamma)} M_j^{(\gamma)} \gamma_j + \sum_{j=1}^{m} \lambda_j^{(\alpha)} M_j^{(\alpha)} \alpha_j + \sum_{j=1}^{l} \lambda_j^{(\gamma)} M_j^{(\nu)} \nu_j$$

where $M_j^{(\gamma)} = (D_\Delta^d)^T (D_\Delta^d)$, and D_Δ represents the contrast matrix corresponding to the difference $\Delta_s \gamma_{js} = \sum_{s=1}^{K_j} (\gamma_{js} - \gamma_{j,s-1})$.

The corresponding penalized score function is given by

$$s(\beta) = \partial l(\beta)/\partial \beta = \sum_{i=1}^{n} Z_i D_i^T \Sigma_i^{-1} (y_i - \mu_i) - M\beta$$

where $D_i = \partial h(\eta_i)/\partial \eta, \Sigma_i = \mathrm{cov}_{\eta_i}(y_i)$ and M is a block diagonal matrix given by

$$M = \mathrm{Diag}\left(0_{p+1,p+1}, \lambda_1^{(\gamma)} M_1^{(\gamma)}, \ldots, \lambda_t^{(\gamma)} M_t^{(\nu)}\right).$$

where the first block of dimension $(p+1) \times (p+1)$ contains only zeros.

The (penalized) Fisher matrix

$$F = \sum_{i=1}^{n} Z_i D_i^T \Sigma_i^{-1} D_i Z_i^T - M$$

yields the Fisher scoring

$$\hat{\beta}^{(\nu+1)} = \hat{\beta}^{(\nu)} + F^{-1}\left(\hat{\beta}^{(\nu)}\right) s\left(\hat{\beta}^{(\nu)}\right), \ \nu = 1, 2, \ldots$$

A.2 Models with multiplicative effects

In the following the model with multiplicative effects is considered where for simplicity the other terms are omitted. The predictor of $(y), y.(w_1), \ldots, y.(w_m)$ has the form

$$\eta_{ir} = \gamma_{0r} + \sum_{j=1}^{m} \bar{\alpha}_{jr} \sum_{s=1}^{K_j} \alpha_{js} \Phi_{js}(w_{ij}).$$

The penalized likelihood has the form

$$l = \sum l(y_i; \eta_{ir}) - \tau\left(\{\gamma_{0r}\}, \{\alpha_{js}\}, \{\bar{\alpha}_{jr}\}\right)$$

where

$$\tau\left(\{\gamma_{0r}\},\{\alpha_{js}\}\right)=\lambda^{(\gamma)}\sum_{r=d+1}^{q}(\Delta_r^d\gamma_{0r})^2+\sum_{j=1}^{m}\lambda_j\sum_{s=d+1}^{K_j}(\Delta_s^d\alpha_{js})^2+\sum_{j=1}^{m}\delta_j\sum_{r=d+1}^{q}(\Delta_r^d\bar{\alpha}_{jr})^2.$$

We suggest a two step estimation procedure:

Step 1

For fixed parameters $\bar{\alpha}_j, j=1,\ldots,m,\ r=1,\ldots,q$ the total predictor $\eta_i^T=(\eta_{i1},\ldots,\eta_{iq})$ is given by

$$\eta_i=\gamma_0+\tilde{D}_i(\bar{\alpha}_1)\Phi_{i1}\alpha_1+\cdots+\tilde{D}_i(\bar{\alpha}_m)\Phi_{im}\alpha_m$$

with

$$\begin{aligned}\gamma_0^T&=(\gamma_{01},\ldots,\gamma_{0q}),\\ \alpha_j^T&=(\alpha_{j1},\ldots,\alpha_{jK_j}),\ j=1,\ldots,m,\\ \bar{\alpha}_j^T&=(\bar{\alpha}_{j1},\ldots,\bar{\alpha}_{jq}),\ j=1,\ldots,m,\\ \tilde{D}(\bar{\alpha}_j)&=\mathrm{Diag}(\bar{\alpha}_j),\ j=1,\ldots,m,\end{aligned}$$

one obtains a MGLM with predictor $\eta_i=Z_i\beta$ where

$$Z_i=\left(I,\tilde{D}(\bar{\alpha}_1)\Phi_{i1},\ldots,\tilde{D}(\bar{\alpha}_m)\Phi_{im}\right)$$

and $\beta^T=(\gamma_0^T,\alpha_1^T,\ldots,\alpha_m^T)$ is penalized. The corresponding penalized likelihood and score function are given by

$$\begin{aligned}l(\beta)&=\sum_{i=1}^{n}l(y_i;\eta_i)-\lambda^{(\gamma)}\left(\Delta_r^d\gamma_{0r}\right)^2-\sum_{j=1}^{m}\lambda_j\sum_{s=d+1}^{K_j}\left(\Delta_s^d\alpha_{js}\right)^2,\\ s(\beta)&=\partial l/\partial\beta=\sum_{i=1}^{n}Z_iD_i^T\Sigma_i^{-1}(y_i-\eta_i)-\lambda^{(\gamma)}M^{(\gamma)}-\sum_{j=1}^{m}\lambda_jM_j\bar{\alpha}_j,\end{aligned}$$

where $D_i=\partial h(\eta_i)/\partial\eta,\Sigma_i=\mathrm{cov}_{\eta_i}(y_i)$, and $M^{(\gamma)}$ is the matrix corresponding to the penalization of γ_{0r} and M_j corresponds to the jth component of the second penalization term.

The penalized Fisher matrix is obtained by

$$F(\beta)=\sum_{i=1}^{n}Z_iD_i^T\Sigma_i^{-1}D_iZ_i^T+M$$

with $M = \left(\lambda^{(\gamma)} M^{(\gamma)}, \lambda_1 M_1, \ldots, \lambda_m M_m\right)$ yielding Fisher scoring

$$\hat{\beta}^{(\nu+1)} = \hat{\beta}^{(\nu)} + F\left(\hat{\beta}^{(\nu)}\right)^{-1} s\left(\hat{\beta}^{(\nu)}\right).$$

Step 2:

Given $\gamma_0, \alpha_1, \ldots, \alpha_m$ one obtains a MGLM with predictor $\eta_i = \gamma_0 + Z_i\alpha$ where γ_0 is a known offset and

$$Z_i = \left(\sum_{s=1}^{K_j} \alpha_{1s}\Phi_{1s}(w_{i1})I_{q\times q}, \ldots, \sum_{s=1}^{K_m} \alpha_{ms}\Phi_{ms}(w_{im})I_{q\times q}\right),$$

where $I_{q\times q}$ is the $(q \times q)$-unit matrix and $\alpha^T = \left(\bar{\alpha}_1^T, \ldots, \bar{\alpha}_m^T\right)$.

The corresponding penalized likelihood and score function have the form

$$l(\bar{\alpha}) = \sum_{i=1}^{n} l(y_i; \eta_i) - \sum_{j=1}^{m} \delta_j \sum_{r=d+1}^{q} \left(\Delta_r^d \bar{\alpha}_{jr}\right),$$

$$s(\alpha) = \sum_{i=1}^{n} Z_i D_i^T \Sigma_i^{-1}(y_i - \eta_i) - \sum_{j=1}^{m} \delta_j M_j \bar{\alpha}_j,$$

where M_j is the matrix corresponding to the jth component of the penalization term.

One obtains with

$$F(\alpha) = \sum_{i=1}^{n} Z_i D_i^T \Sigma_i^{-1} D_i Z_i^T + M$$

with $M = (\delta_1 M_1, \ldots, \delta_m M_m)$ the Fisher scoring

$$\hat{\alpha}^{(\nu+1)} = \hat{\alpha}^{(\nu)} + F^{-1}\left(\hat{\alpha}^{(\nu)}\right) s\left(\hat{\alpha}^{(\nu)}\right).$$

Step 1 and Step 2 are iterated until convergence.

Acknowledgment:

Support from the SFB 386 funded by Deutsche Forschungsgemeinschaft is gratefully acknowledged. The first version of this paper was written while the author was visiting the University of Florida, Gainesville – thanks to A. Agresti. I am grateful for computational work done by Torsten Scholz. Thanks to R. Bender who let us use the retinopathy data.

References

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. New York: Wiley.

Agresti, A. (1999). Modelling ordered categorical data: Recent advances and future challenges. *Statistics in Medicine* **18**, 2191–2207.

Armstrong, B. and Sloan, M. (1989). Ordinal regression models for epidemiologic data. *American Journal of Epidemiology* **129**, 191–204.

Barnhart, H. and Sampson, A. (1994). Overviews of multinomial models for ordinal data. *Comm. Stat-Theory & Methods* **23(12)**, 3395–3416.

Bender, R. and Grouven, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *J. Clin. Epidemiol.* **51**, 809–816.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics* **46**, 1171–1178.

Cox, C. (1988). Multinomial regression models based on continuation ratios. *Statistics in Medicine* **7**, 433–441.

Davis, C. S. (1991). Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials. *Statistics in Medicine* **10**, 1959–1980.

Eilers, P. H. and Marx, B. D. (1999). Generalized linear additive smooth structures. Preprint.

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. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models* (2nd ed.). New York: Springer.

Goodman, L. A. (1983). The analysis of dependence in cross-classification having ordered categories, using log-linear models for frequencies and log-linear models for odds. *Biometrika* **39**, 149–160.

Green, D. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman

& Hall.

Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statistics in Medicine* **13**, 1665–1677.

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

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society* **B 55**, 757–796.

Kauermann, G. and Tutz, G. (2000a). Local likelihood estimation in varying coefficient models including additive bias correction. *Journal of Nonparametric Statistics* **12**, 343–371.

Kauermann, G. and Tutz, G. (2000b). Testing generalized and semiparametric models against smooth alternatives. *J. Roy. Stat. Soc. B* (to appear).

Läärä, E. and Matthews, J. N. (1985). The equivalence of two models for ordinal data. *Biometrika* **72**, 206–207.

McCullagh, P. (1980). Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society* **B 42**, 109–127.

O'Sullivan, F., Yandell, B. S., and Raynor, W. J. (1986). Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association* **81**, 96–103.

Ruppert, D. (2000). Selecting the number of knots for penalized splines. Preprint.

Ruppert, D. and Carroll, R. J. (1999). Spatially-adaptive penalties for spline fitting. *Australian Journal of Statistics* **42**, 205–223.

Severini, T. A. and Staniswalis, J. G. (1994). Quasi–likelihood estimation in semiparametric models. *Journal of the American Statistical Association* **89**, 501–511.

Simon, G. (1974). Alternative analyses for singly-ordered contingency table. *Journal of the American Statistical Association* **69**, 971–976.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society* **B 50**, 413–436.

Tutz, G. (1991). Sequential models in ordinal regression. *Computational Statistics & Data Analysis* **11**, 275–295.

Tutz, G. (2000). *Die Analyse kategorialer Daten – eine anwendungsorientierte Einführung in Logit-Modellierung und kategoriale Regression.* München: Oldenbourg Verlag.

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