Fahrmeir, Gieger, Klinger:

Additive, Dynamic and Multiplicative Regression

Additive, Dynamic and Multiplicative Regression

Ludwig Fahrmeir, Christian Gieger and Artur Klinger
Institut für Statistik, Universität München

We survey and compare model-based approaches to regression for cross-sectional and longitudinal data which extend the classical parametric linear model for Gaussian responses in several aspects and for a variety of settings. Additive models replace the sum of linear functions of regressors by a sum of smooth functions. In dynamic or state space models, still linear in the regressors, coefficients are allowed to vary smoothly with time according to a Bayesian smoothness prior. We show that this is equivalent to imposing a roughness penalty on time-varying coefficients. Admitting the coefficients to vary with the values of other covariates, one obtains a class of varying-coefficient models (Hastie and Tibshirani, 1993), or in another interpretation, multiplicative models. The roughness penalty approach to non- and semiparametric modelling, together with Bayesian justifications, is used as a unifying and general framework for estimation. The methodological discussion is illustrated by some real data applications.

1. Introduction

Consider first the case of a Gaussian response y which is observed together with regressors $\{x_1, \ldots, x_p\}$. The classical Gaussian linear model assumes

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \tag{1.1}$$

with the usual assumptions on the error variable ε.

In additive models (Hastie and Tibshirani, 1990), some or all of the linear functions $\beta_j x_j$ of the covariates are replaced by smooth functions $f_j(x_j)$, modelled and estimated in some nonparametric way, e.g. by kernel and nearest neighborhood methods or splines. We focus on penalized least-squares methods, which lead to cubic smoothing splines and related estimators.

Dynamic models are useful for analyzing time series and longitudinal data, where the variables $\{y, x_1, \ldots, x_p\}$ are observed over time. In linear dynamic models, some or all of the coefficients $\{\beta_0, \beta_1, \ldots, \beta_p\}$ are allowed to vary over time and (1.1) is modified to

$$y(t) = \beta_0(t) + \beta_1(t)x_1(t) + \ldots + \beta_p(t)x_p(t) + \varepsilon(t)\,. \tag{1.2}$$

The time-varying intercept $\beta_0(t)$ is often additively splitted up into a trend component $m(t)$ and a seasonal component $s(t)$, and sometimes no covariates are present in the model. In the state space approach to dynamic models (Harvey, 1989; West and Harrison, 1989) the parameters or 'states' $\{\beta_0(t), \ldots, \beta_p(t)\}$ obey a linear Markovian transition model or, in other words, a Bayesian smoothness prior. Following Bayesian arguments, the sequence of 'states' is estimated by the well-known linear Kalman filtering and smoothing algorithms. We show in Section 3 that this is equivalent to minimizing a penalized least-squares criterion, so that dynamic modelling methods can also be interpreted as a model-based semiparametric roughness penalty approach.

If the parameters are allowed to vary with the values of other covariates than time, say $v_0, \ldots, v_p$, one arrives at varying coefficient models

$$y = \beta_0(v_0) + \beta_1(v_1)x_1 + \ldots + \beta_p(v_p)x_p + \varepsilon\,, \tag{1.3}$$

as introduced by Hastie and Tibshirani (1993) from the nonparametric point of view. Since the terms $\beta_j(v_j)x_j$ in (1.3) may also be interpreted as special forms of multiplicative interaction between v_j and x_j, we also say that (1.3) is a multiplicative regression model.

For non-Gaussian responses y, for example discrete or categorical responses, generalized linear models extend the linear model (1.1) to a much broader class. However, they still are parametric and retain an essential feature of linear models by relating the mean Ey to a linear predictor $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ via a link function. Obviously, generalized additive, dynamic and multiplicative models can be defined in the same way as before by appropriate modification of η.

Section 2 describes the models in more detail, accompanied by real data examples. Estimation by the roughness penalty approach is dealt with in Section 3, and Section 4 contains applications of the methods to the real data examples. Section 5 concludes with some remarks on some topics where further research would be useful and interesting.

2. Generalized regression models

2.1 Additive models

Consider the common situation of cross-sectional regression analysis with a response variable y and a vector $x = (x_1, \ldots, x_p)$ of covariates. The observations (y_i, x_i), $i = 1, \ldots, n$ on (y, x) are assumed to be independent. In the simplest case of linear Gaussian regression one assumes model (1.1), where y is normally distributed and $E(\varepsilon) = 0$, $\text{var}(\varepsilon) = \sigma^2$. In other words, the (conditional) mean $\mu = E(y|x)$ of y is specified as a linear predictor $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$.

Generalized linear models provide a comprehensive parametric framework for regression analysis with non-Gaussian responses, including categorical and counted responses. In their original version (e.g. Mc Cullagh and Nelder, 1989), generalized linear models assume that the distribution of y given x comes from an exponential family and that the mean $\mu = E(y|x)$ is related to the linear predictor η by a response or link function h in the form

$$\mu = h(\eta) = h(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p).$$

Due to the distributional assumptions the variance function $\text{var}(y|x)$ is then determined by choice of the specific exponential family. Common models are logistic models, with $\mu = \exp(\eta)/\{1 + \exp(\eta)\}$ and y a binary variable, and log-linear models with $\mu = \exp(\eta)$ and y a Poisson variable. Dropping the exponential family assumption, η may be any reasonable parameter of interest of the likelihood or some quasi-likelihood of the observations, as for example in the Cox model where η parametrizes a part of the hazard function. Also μ, η and h may be multidimensional if the response variable is a vector $y = (y_1, \ldots, y_q)$, as for example in multinomial models for multicategorical responses, where y_j is a dummy variable representing category j. Then, generally, a vector of predictors $\beta_{0j} + \beta_{1j} x_1 + \ldots + \beta_{pj} x_p$ will be necessary, see e.g. Fahrmeir and Tutz (1994a, ch.3).

In generalized additive models (Hastie and Tibshirani, 1990) all or a part of the linear functions $\beta_j x_j$ of the regressors are replaced by smooth functions $f_j(x_j)$, so that

$$\eta = f_1(x_1) + \ldots + f_p(x_p)\,, \tag{2.1}$$

or, for example,

$$\eta = f_1(x_1) + \beta_2 x_2 + \ldots + \beta_p x_p\,, \tag{2.2}$$

if only x_1 is metrical and $x_2, \ldots, x_p$ are binary. The smooth functions can be modelled by flexible parametric forms, e.g. by piecewise polynomials or orthogonal series, or nonparametrically, e.g. by using kernel, nearest neighborhood or penalized likelihood methods. In this paper we will focus on penalized least squares and likelihood methods as a unifying modelling and estimation approach. From this point of view, the smooth functions $f_j()$ are unknown, but fixed. It should be noted, however, that appropriate Bayesian formulations of smoothness lead to the same estimate, see e.g. Wahba (1978) and Green and Silverman (1994, Section 3.8).

Example 1: Credit-Scoring Revisited

In credit business banks are interested in estimating the risk that consumers will pay back their credits as agreed upon by contract or not. The aim of credit-scoring systems is to model or predict the probability that a client with certain covariates ("risk factors") is to be considered as a potential risk. We will analyze the effect of covariates on the binary response "creditability" by a logit model. Other tools currently used in credit scoring are (linear) discriminance analysis, classification and regression trees, and neural networks.

The data set consists of 1000 consumer's credits from a South German bank. The response variable of interest is "creditability",which is given in dichotomous form ($y = 0$ for creditworthy, $y = 1$ for not creditworthy). In addition, 20 covariates that are assumed to influence creditability were collected. The raw data are recorded in Fahrmeir and Hamerle (1984, see p. 334 ff. and p. 751 ff.) and are available on electronic file. In Fahrmeir and Kredler (1984, p. 285-86) and Fahrmeir and Tutz (1994a, Ch. 2) a logit model was used to analyze a subset of these data containing only the following covariates, which are partly metrical and partly categorical:

X1 running account, trichotomous with categories "no running account" (=1), "good running account" (=2), "medium running account" ("less than 200 DM" = 3 = reference category)

X3 duration of credit in months, metrical

X4 amount of credit in DM, metrical

X5 payment of previous credits, dichotomous with categories "good", "bad" (=reference category)

X6 intended use, dichotomous with categories "private" or "professional" (=reference category)

X8 marital status, with reference category "living alone".

Assuming a logit model with linear predictor

$$\eta = \beta_0 + \beta_1 X1[1] + \beta_2 X1[2] + \beta_3 X3 + \beta_4 X4 + \beta_5 X5 + \beta_6 X6 + \beta_8 X8$$

for the probability $pr(y = 1|x)$ of being "not creditworthy", one obtains the following maximum likelihood estimates of the covariate effects:

	Intercept	X1[1]	X1[2]	X3	X4	X5	X6	X8
value	0.01	0.62	-1.32	0.03	0.00	-0.98	-0.46	-0.54
t-value	0.03	3.55	-6.53	4.45	0.86	-3.89	-2.90	-3.38

This leads to the somewhat surprising conclusion that the covariate X4 "amount of credit" has no significant influence on the risk. In Section 4, the data are reanalyzed by an additive logit model, with the linear functions $\beta_3 X3$ and $\beta_4 X4$ replaced by smooth functions $f_3(X3)$ and $f_4(X4)$. The results obtained there lead to a different conclusion.

2.2 Dynamic models

Suppose now that the data consist of repeated observations of the response y and, possibly, a vector $x = (x_1, \ldots, x_p)$ of covariates at T time points $t_1 < t_2 < \ldots < t_T$. To simplify notation, we write $t \in \{1, \ldots, T\}$, but equidistant time points are not a necessary prerequisite.

Linear Gaussian dynamic models relate the observations $\{y(t), x_1(t), \ldots, x_p(t)\}$ in additive form, including a trend component $m(t)$ and perhaps a seasonal component $s(t)$:

$$y(t) = m(t) + s(t) + \beta_1(t)x_1(t) + \ldots + \beta_p(t)x_p(t) + \varepsilon(t)\,, \qquad (2.3)$$

where $y(t)$ is normally distributed, and $E\,\varepsilon(t) = 0$, $\mathrm{var}\,\varepsilon(t) = \sigma_t^2$. The effects $\beta_1(t), \ldots, \beta_p(t)$ may be time-varying or not. If no covariates are present, (2.3) reduces to a simple additive structural time series model, where $m(t)$, $s(t)$ are unknown sequences or functions of time. Traditional descriptive methods for analyzing trends and seasonal components are based on moving averages or the method of graduation (Whittaker, 1923), which imposes a certain roughness penalty on the trend function. We follow here the state space approach to structural time series analysis (e.g. Harvey, 1989), where the observation model (2.3) is supplemented by a linear Gaussian transition model for $m(t)$, $s(t)$ and $\beta(t)$. Gathering $m(t)$, $s(t)$ and $\beta(t)$ in a 'state' vector $\alpha(t)$, the general form is

$$\alpha(t) = F(t)\alpha(t-1) + \xi(t) , \quad t = 1, \ldots , T \tag{2.4}$$

with a non-random transition matrix $F(t)$, Gaussian white noise $\{\xi_t\}$ with $\xi_t \sim N(0, Q_t)$ and initial state $\alpha(0) \sim N(a_0, Q_0)$.

Admitting multivariate observations $y(t)$, the observation model (2.3) may be rewritten in the form

$$y(t) = Z(t)\alpha(t) + \varepsilon(t) , \quad t = 1, \ldots , T , \tag{2.5}$$

where $Z(t)$ is an observation or design matrix of appropriate dimension, reducing to a design vector $z'(t)$ if $y(t)$ is scalar. The Gaussian white noise sequence $\{\varepsilon(t) \sim N(0, \Sigma_t)\}$ is assumed to be uncorrelated with $\{\xi(t)\}$ and $\alpha(0)$. Simple nonstationary models for trend or time-varying effects are first or second order random walks

$$m(t) = m(t-1) + u(t) , \quad m(t) = 2m(t-1) - m(t-2) + u(t) \tag{2.6}$$

with $u(t) \sim N(0, q_t^2)$. By appropriate definition of $Z(t)$ and $F(t)$ they can be put in state space form as well as more complicated seasonal components, see e.g. Harvey (1989, pp. 40-43) and Fahrmeir and Tutz (1994a, Section 8.1). From a Bayesian perspective, the transition models (2.4), (2.6) can be interpreted as 'smoothness priors' for $\{\alpha(t)\}$ or $\{m(t), s(t), \beta(t)\}$. In fact it turns out, see Section 3, that these 'smoothness priors' are the Bayesian justification for the roughness penalty approach.

The obvious modification for observations $y(t)$ with exponential family densities are dynamic generalized linear models (e.g. West, Harrison and Migon, 1985; Fahrmeir,1992; Fahrmeir and Tutz, 1994a, ch. 8). The observation models (2.3) or (2.5) are now specified by an exponential family density for $y(t)$, given $\alpha(t)$ and $x(t)$, with conditional mean

$$E(y(t)|\alpha(t), x(t)) = \mu(t) = h(\eta(t)) , \tag{2.7}$$

the predictor

$$\eta(t) = m(t) + s(t) + x'(t)\beta(t) \quad \text{resp.} \quad \eta(t) = Z(t)\alpha(t) \tag{2.8}$$

and one of the common response functions h. The observation model (2.7) is again supplemented by a linear Gaussian transition model (2.4) or (2.6).

For time series of counts, loglinear Poisson models $y(t)|\alpha(t),\, x(t) \sim Po(\lambda(t))$, $\lambda(t) = \exp(\eta(t))$ are a standard choice. If the number of counts at t is limited by $n(t)$, say, binomial regression models, in particular logit or probit models, are often appropriate: $y(t)|\alpha(t),\, x(t) \sim B(n(t), \pi(t))$; $\pi(t) = h(\eta(t) = z'(t)\alpha(t))$, with h the logistic or standard normal distribution function. For $n(t) = 1$, this is a common way for modelling binary time series.

Extensions to time series of multicategorical or multinomial responses proceed along similar lines. Let k be the number of categories and $y(t) = (y_1(t), \ldots, y_q(t))$ be a vector of $q = k - 1$ dummy variables, with $y_j(t) = 1$ if category j has been observed, $y_j(t) = 0$ otherwise. Dynamic categorical response models are specified by relating response probabilities $\pi_j(t) = pr(y_j(t) = 1)$, $j = 1, \ldots, q$, to a q-dimensional predictor

$$\eta(t) = (\eta_1(t), \ldots, \eta_q(t))' = Z(t)\alpha(t)\,. \tag{2.9}$$

The most common models for ordered categories are dynamic cumulative models. They can be derived from a threshold mechanism for an underlying linear dynamic model. The resulting response probabilities are

$$\pi_j(t) = F(\eta_j(t)) - F(\eta_{j-1}(t))\,, \quad j = 1, \ldots, q \tag{2.10}$$

with linear predictors

$$\eta_j(t) = m_j(t) + x'(t)\beta(t)\,,$$

ordered threshold parameters $-\infty = m_0(t) < \ldots < m_q(t) < \infty$, a vector $\beta(t)$ of global covariate effects, and a known distribution function F, e.g. the logistic one. The thresholds may also contain additive seasonal components $s_j(t)$. Dynamic versions of other models for ordered categories discussed e.g. in Fahrmeir and Tutz (1994a, Section 3.4) can be designed with analogous reasoning.

In many applications, more than one individual or object is observed sequentially over time. Let us consider longitudinal or panel data which consist of observations $(y_i(t),\, x_i(t))$, $i = 1, \ldots, n$, $t = 1, \ldots, T$, for a population of n units observed across time. The state space modelling approach to longitudinal data allows, in principle, to deal with random effects ('states') across units and across time, like stochastic trend and seasonal components. We will confine attention to the case where states are constant across units. In this case it is assumed that the predictor for observation $(y_i(t),\, x_i(t))$ is

$$\eta(i, t) = m(t) + s(t) + x_i'(t)\beta(t)\,. \tag{2.11}$$

This means that $m(t)$, $s(t)$ and $\beta(t)$ are population-averaged effects over time. Random effects across units could be modelled in additive form, e.g. by

$$\eta(i,t) = \gamma(i) + m(t) + x_i'(t)\beta(t)\,, \tag{2.12}$$

together with a Gaussian prior $\gamma(i) \sim N(0, G)$.

Example 2: IFO business test

The IFO institute for economic research in Munich collects categorical monthly data of firms in various industrial branches. The questionnaire contains questions on expectations and realizations of variables like production, orders in hand, demand etc. Most answers are in categories like increase (+), decrease (−), or no change (=). Considering all firms within a certain branch we have categorical longitudinal data.

We apply a dynamic cumulative model to data collected in the industrial branch "Steine und Erden", for the period of January 1980 to December 1990. Firms in this branch manufacture initial products for the building trade industry.

The response variable is formed by the production plans $P(t)$. Its conditional distribution is assumed to depend on the covariates "orders in hand" $O(t)$ and "expected business condition" $D(t)$, and on the production plans $P(t-1)$ of the previous month. No interaction effects are included. Each trichotomous variable is described by two ($q = 2$) dummy variables, with "−" as the reference category. Thus (1,0), (0,1) and (0,0) stand for the responses "+","=" and "−". The relevant dummies for "+" and "=" are shortened by $P(t)^+$, $P(t)^=$, etc. Then a cumulative logistic model with time-varying thresholds $m_1(t)$, $m_2(t)$ and global covariate effects $\beta_1(t)$ to $\beta_6(t)$ is specified by

$$\begin{aligned} pr(P(t) = \text{`+'}) &= h(m_1(t) + \beta_1(t)P(t-1)^+ + \beta_2(t)P(t-1)^= + \beta_3(t)D(t)^+ \\ &\quad +\beta_4(t)D(t)^= + \beta_5(t)O(t)^+ + \beta_6(t)O(t)^=)\,, \\ pr(P(t) = \text{`+'or`='}) &= h(m_2(t) + \beta_1(t)P(t-1)^+ + \beta_2(t)P(t-1)^= + \beta_3(t)D(t)^+ \\ &\quad +\beta_4(t)D(t)^= + \beta_5(t)O(t)^+ + \beta_6(t)O(t)^=)\,, \end{aligned}$$

where $pr(P(t) = \text{`+'})$ and $pr(P(t) = \text{`+' or `='})$ stand for the probability of increasing and nondecreasing production plans, and h is the logistic distribution function. The time-varying parameters $m_1(t)$, $m_2(t)$, $\beta_1(t), \ldots, \beta_6(t)$ are modelled by an eight-dimensional first order random walk. More details on this and a second example can be found in Fahrmeir and Nase (1994).

Example 3: Dynamic Pair Comparisons for the German Fußball-Bundesliga.

In paired comparisons, treatments, players or teams $\{a_1, \ldots, a_n\}$ are compared with each other in pairs. Let y_{ij} denote the observed response when the pair (a_i, a_j) meets. For soccer teams, y_{ij} is trichotomous where the categories 1, 2, 3 stand for "a_i wins", "draw", "a_j wins". Based on latent random utilities and thresholds, Tutz (1986) derives the ordinal logistic paired comparison model

$$\begin{aligned} pr(y_{ij} = 1) &= F(\theta_1 + \alpha_i - \alpha_j)\,, \\ pr(y_{ij} = 2) &= F(\theta_2 + \alpha_i - \alpha_j) - F(\theta_1 + \alpha_i - \alpha_j)\,, \\ pr(y_{ij} = 3) &= 1 - pr(y_{ij} = 1) - pr(y_{ij} = 2)\,, \end{aligned}$$

where F is the logistic distribution function. The parameters α_i represent the unobserved "ability" of team a_i. The role of thresholds refers to the home court advantage. In the German Fußball-Bundesliga teams meet twice within each season giving each team the home court advantage once. For competing teams the pair (a_i, a_j) implies that the game is played on the home court of a_i. The home court advantage is most obvious in the case where the abilities of teams are equal i.e. $\alpha_i = \alpha_j$. Then the probabilities $pr(y_{ij} = r) = F(\theta_r) - F(\theta_{r-1})$ depend only on the thresholds. Since the teams have equal abilities the probability of response categories reflects the home court advantage which of course is specific for the game. In our soccer example it turns out that home court advantage is rather stable over the years, yielding the thresholds $\hat{\theta}_1 = -0.358$ and $\hat{\theta}_2 = 1.039$. For $\alpha_i = \alpha_j$ that means $pr(y_{ij} = 1) = 0.411$, $pr(y_{ij} = 2) = 0.328$, $pr(y_{ij} = 3) = 0.261$. Therefore a soccer team will beat another team of equal ability on their home court with probability 0.411 and will be beaten only with probability 0.261.

Since we analyze results of pair comparisons of soccer teams for the seasons 1966 to 1987, it is not to be expected that abilities remain constant over time. Fahrmeir and Tutz (1994b) introduce dynamic models for time-dependent ordered pair comparisons for responses $y_{ij}(t)$ observed at time t and possibly time-varying latent thresholds $\theta_1(t)$, $\theta_2(t)$ and abilities $\alpha_i(t)$, $\alpha_j(t)$. The observation model is then

$$\begin{aligned} pr(y_{ij}(t) = 1) &= F(\theta_1(t) + \alpha_i(t) - \alpha_j(t)) \\ pr(y_{ij}(t) = 2) &= F(\theta_2(t) + \alpha_i(t) - \alpha_j(t)) - F(\theta_1(t) + \alpha_i(t) - \alpha_j(t))\,, \end{aligned}$$

and is supplemented by a transition model, e.g. random walk models, for $\theta_1(t)$, $\theta_2(t)$, $\alpha_i(t)$ and $\alpha_j(t)$.

2.3 Multiplicative models

Dynamic models with predictors (2.3), (2.9) or (2.11) are commonly interpreted as extensions of (generalized) linear models with time-varying intercepts and covariate effects. Another way to look at them is to consider time as another, though special covariate. Then a term $x_j(t)\beta_j(t)$ has the form of a multiplicative interaction term between the possibly time-varying covariate x_j and a smooth function of the 'covariate' time. Admitting other covariates, say $v_0, v_1, \ldots, v_p$, than time, we arrive at multiplicative models of the form

$$\eta = \beta_0(v_0) + \beta_1(v_1)x_1 + \ldots + \beta_p(v_p)x_p \,, \tag{2.13}$$

where terms $\beta_j(v_j)x_j$ can be seen as a special kind of interaction between v_j and x_j. Another way is to look at (2.13) as a model linear in the regressors $x_1, \ldots, x_p$, but with parameters changing smoothly with the values of $v_0, v_1, \ldots, v_p$, and to call it a 'varying-coefficient model', as introduced by Hastie and Tibshirani (1993). Although looking apparently special, multiplicative or varying-coefficient models are quite general: For $\beta_j(v_j) = \beta_j$, i.e. constant functions $\beta_j()$, one gets back generalized linear models, for $x_1 = \ldots = x_p = 1$ additive models and for $v_0 = v_1 = \ldots = v_p = t =$ time dynamic models. Many other particular models can be written in the form (2.13), see Hastie and Tibshirani (1993) and the discussion following the paper. In the following Example 4, we will consider a specific application. In all cases, the unspecified functions $\beta_j()$ may be modelled in various ways, e.g. using kernel methods, penalized least squares and likelihoods, or other nonparametric approaches as in additive models, or imposing Bayesian smoothness priors as in dynamic models. In Section 3, we will deal with the estimation problem under the general framework of roughness penalties.

Example 4: Rental tables ("Mietspiegel")

Surveys on rents for lodging, paid according tenancy agreements between letters and tenants of rented flats or appartments, are conducted regularly in larger communities or cities. Based on a sample of tenancies, traditional rental tables contain average rents in form of contingency tables with cells determined by categories of floor space, year of construction and perhaps site of the flat. According to the German "Mieterhöhungsgesetz", rental tables may be used to determine adequate raising of rents.

As an alternative to contingency tables, regression may be a useful tool for analyzing how rents depend on floor space, year of construction and factors characterizing site, type and equipment of the flat. For our example we use a sample of 1969 tenancies for flats in Munich, with floor space from 30 to 120 square meters and year of construction between 1890 and 1989. The response variable is the net rent, which does not contain operating costs. Covariates are

F floor space in square meters,
A age (= year of construction),
S^+ site above average, binary, with $S^+ = 0 =$ average as reference category,
S^- site below average, binary, with $S^- = 0 =$ average as reference category,
H no central heating, binary,
B no bathroom, binary,
L bathroom, with equipment above average.

A linear additive regression model $y = \eta + \varepsilon$ with

$$\eta = \beta_0 + \beta_1 F + \beta_2 A + \beta_3 S^+ + \beta_4 S^- + \beta_5 H + \beta_6 B + \beta_7 L$$

will not be adequate since increase or decrease of the average rent η due to one of the factors age, site or equipment would be independent of floor space of the flat, leading to implausible results. Instead, multiplicative models with interaction terms like $F{\cdot}H$ are more realistic. Also it is unclear wether the metrical covariates F and A can modelled appropriately by linear functions. Therefore, a multiplicative model with predictor

$$\eta = \beta_1(F) + F\beta_2(A) + \beta_3(F)S^+ + \beta_4(F)S^- + \beta_5(F)H + \beta_6(F)B + \beta_7(F)L$$

can be useful for exploratory data analysis.

3. Estimation

In this section, the focus is on the roughness penalty approach. Methods for selecting smoothing parameters are only mentioned and Bayesian posterior mean estimation will be addressed to only briefly.

3.1 Penalized least squares

Smoothed estimators of regression curves may be considered as compromises between faith with the data and reduced roughness caused by the noise in the data. This view is made explicit in the construction of smoothing splines. For bivariate observations $(y_i,\, x_i)$, $i = 1, \ldots, n$ of the continuous variables (y, x), the starting point is the following minimization problem: Find the twice continuously differentiable function $f()$ that minimizes the penalized sum of squares

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int\limits_a^b f''(u)^2 \, du \,, \tag{3.1}$$

where $[a, b]$ contains the covariate values $x_1 < \ldots < x_n$. The first term in (3.1) is the residual sum of squares, which is used as a distance function between data and estimator. The second term penalizes roughness of the function by taking the integrated squared second derivative $\int f''(u)^2 \, du$ as a global measure for curvature or roughness. The parameter $\lambda \geq 0$ is a smoothing parameter that controls the trade-off between smoothness of the curve and faith with the data: Large values of the smoothing parameter λ give large weight to the penalty term, therefore enforcing smooth functions with small variance but possibly high bias. For rather small λ, the function $f()$ will nearly interpolate the data. The function $\hat{f}()$ minimizing (3.1) is a natural cubic smoothing spline with knots at $x_1 < \ldots < x_n$ (Reinsch, 1967). Since cubic splines are actually defined by a finite number of parameters, the minimization problem with respect to a set of functions reduces to a finite-dimensional optimization problem. It can be shown that minimization of (3.1) is equivalent to minimizing the penalized least-squares criterion

$$PS(f) = (y - f)'(y - f) + \lambda f' K f \,, \tag{3.2}$$

where $y = (y_1, \ldots, y_n)$ are the data and $f = (f(x_1), \ldots, f(x_n))$ denotes now the vector of evaluations of the function $f()$. The penalty matrix K has a special structure and can be written as the product of tridiagonal band matrices, see e.g. Green and Silverman (1994, ch. 2). The minimizer $\hat{f}$ of $PS(f)$ is obtained by equating the vector of first derivatives to zero. This yields the linear smoother

$$\hat{f} = (I + \lambda K)^{-1} y \tag{3.3}$$

with smoothing matrix $S = (I + \lambda K)^{-1}$.

For computational reasons, $\hat{f}$ and the smoothing matrix S are generally not computed directly by inversion of $I + \lambda K$ (note that S is an full $(n \times n)$-matrix). Instead, $\hat{f}$ is computed indirectly, e.g. by the Reinsch algorithm.

In (3.2) the distance between data and estimator is measured by a simple quadratic function. More generally a weighted quadratic distance may be used. For given diagonal weight matrix W a weighted penalized least squares criterion is given by

$$(y - f)'W(y - f) + \lambda f'Kf\,. \tag{3.4}$$

The solution is again a cubic smoothing spline, with the vector $\hat{f}$ of fitted values now given by

$$\hat{f} = (W + \lambda K)^{-1} W y\,. \tag{3.5}$$

In (3.2) and (3.4), the smoothing parameter was assumed to be known or given. In practice it is either obtained by a subjective choice or by an automatic data-driven method, e.g. by minimizing some cross-validation score, see Härdle (1990) for details.

The integrated squared curvature $\int f''(u)^2\,du$ and the resulting penalty matrix K are not the only way to penalize roughness of the estimator. Simple roughness penalties are the sums of squared first or second differences

$$D_1(f) = \sum_{i=2}^{n} \{f(x_i) - f(x_{i-1})\}^2\,,\; D_2(f) = \sum_{i=3}^{n} \{f(x_i) - 2f(x_{i-1}) + f(x_{i-2})\}^2. \tag{3.6}$$

If the differences $x_i - x_{i-1}$ are small and almost equidistant, second differences are good approximations to $f''(x)$, and the resulting smooth estimate $\hat{f}$ is very similar to a cubic spline. However, the penalty matrices K satisfying $f'Kf = D_1(f)$ and $D_2(f)$ are now tridiagonal and pentadiagonal. Using band matrix manipulations, this makes computation of $\hat{f}$ in $O(n)$ operations quite easy.

For additive models

$$y = f_1(x_1) + \ldots + f_p(x_p) + \varepsilon$$

the penalized sum of squares is generalized to

$$\sum_{i=1}^{n} w_i(y_i - f_1(x_{i1}) - \ldots - f_p(x_{ip}))^2 + \lambda_1 \int f_1''(u)^2\,du + \ldots + \lambda_p \int f_p''(u)^2\,du\,. \tag{3.7}$$

The minimizing functions are again cubic splines. Parameterizing by the vectors $f_j = (f_j(x_{1j}), \ldots, f_j(x_{nj}))$, $j = 1, \ldots, p$, (3.7) can be written as the penalized least squares criterion

$$
\begin{aligned}
PS(f_1,\ldots,f_p) &= (y-f_1-\ldots-f_p)'W(y-f_1-\ldots-f_p)+ \\
&\quad +\lambda_1 f_1'K_1 f_1+\ldots+\lambda_p f_p'K_p f_p\,,
\end{aligned}
\tag{3.8}
$$

where $W=\operatorname{diag}(w_1,\ldots,w_n)$ and the penalty matrices K_j are defined analogously to K. The minimizing functions now satisfy the system of equations

$$
\begin{bmatrix} W+\lambda_1K_1 & \cdots & W \\ \vdots & \ddots & \vdots \\ W & \cdots & W+\lambda_pK_p \end{bmatrix}
\begin{bmatrix} f_1 \\ \vdots \\ f_p \end{bmatrix}
=
\begin{bmatrix} Wy \\ \vdots \\ Wy \end{bmatrix}
$$

or equivalently

$$
\begin{aligned}
f_1 &= (W+\lambda_1K_1)^{-1}W(y-f_2-\ldots-f_p) \\
\vdots & \qquad\qquad \vdots \\
f_p &= (W+\lambda_pK_p)^{-1}W(y-f_1-\ldots-f_{p-1})\,.
\end{aligned}
$$

The solutions $\hat{f}_1,\ldots,\hat{f}_p$ are obtained iteratively by 'backfitting', a Gauss-Seidel type algorithm, see Buja, Hastie and Tibshirani (1989) and Hastie and Tibshirani (1990) for details. Automatic choice of the smoothing parameters $\lambda_1,\ldots,\lambda_p$, based on cross-validation, is now far more demanding, since it would require that the diagonal or the trace of the global smoother matrix were available with reasonable amount of effort. It seems that additional research is necessary here.

A non-iterative and simpler solution avoiding backfitting can be obtained for semiparametric models (2.2) with

$$
\eta_i = f_1(x_{i1})+\beta_2x_{i2}+\ldots+\beta_px_{ip} = f_1(x_{i1})+z_i'\beta\,,\ i=1,\ldots,n\,,
$$

$z_i'=(x_{i2},\ldots,x_{ip})\,,\beta'=(\beta_2,\ldots,\beta_p)$. Defining the design matrix $Z=(z_1,\ldots,z_n)'$, one obtains

$$
\begin{aligned}
\hat{\beta} &= \{Z'W(I-S)Z\}^{-1}Z'W(I-S)y \\
\hat{f}_1 &= S(y-Z\hat{\beta})\,,\quad S\ =\ (W+\lambda K)^{-1}W\,.
\end{aligned}
$$

Consider now Gaussian dynamic linear models (2.4), (2.5). Given the observations $y=(y(1),\ldots,y(T))$, estimation of $\alpha(t)$ is traditionally called filtering for $t=T$ and smoothing for $t<T$. Due to the linearity and normality assumptions in (2.4), (2.5), the posterior distribution of $\alpha(t)$ is also Gaussian

$$
\alpha(t)|y \sim N(a_{t|T},V_{t|T})
$$

with posterior mean $a_{t|T} = E(\alpha(t)|y)$ and posterior covariance matrix $V_{t|T} = E((\alpha(t) - a_{t|T})(\alpha(t) - a_{t|T})')$. Linear Kalman filters and smoothers provide $a_{t|T}$, $V_{t|T}$ in a computationally efficient, recursive way. Very short proofs are based on Bayesian arguments using conjugate prior-posterior properties of Gaussian distributions. In the following, we will sketch the lines of argument for a derivation which corresponds to the historically first derivation (Thiele, 1880) and shows that Kalman filtering and smoothing is actually equivalent to penalized least squares estimation.

Consider the joint posterior $p(\alpha|y)$, with $\alpha = (\alpha(0), \alpha(1), \ldots, \alpha(T))$. Since this posterior is Gaussian, posterior means and posterior modes are equal and can therefore be obtained by maximizing the posterior density. Repeated application of Bayes' theorem, thereby making use of the model assumptions and taking logarithms shows that this maximization is equivalent to minimization of the penalized least-squares criterion

$$\begin{aligned} PS(\alpha) = \sum_{t=1}^{T}(y(t) - Z(t)\alpha(t))'\Sigma_t^{-1}(y_t - Z(t)\alpha(t)) + (\alpha(0) - a_0)'Q_0^{-1} \\ (\alpha(0) - a_0) + \sum_{t=1}^{T}(\alpha(t) - F(t)\alpha(t-1))'Q_t^{-1}(\alpha(t) - F(t)\alpha(t-1)) \end{aligned} \tag{3.9}$$

with respect to α. For simplicity, we have assumed that Σ_t, Q_t are nonsingular. One may, however, drop this assumption.

As an example, consider the model $y(t) = m(t) + x(t)\beta(t) + \varepsilon(t)$ with independent second-order random walks for $m(t)$ and $\beta(t)$. Setting $\lambda_1 = \sigma_\varepsilon^2/q_m^2$, $\lambda_2 = \sigma_\varepsilon^2/q_\beta^2$, where q_m^2, q_β^2 are the variances of the random walk error variables, and omitting priors for $m(0)$, $m(-1)$, $\beta(0)$, $\beta(-1)$, criterion (3.9) reduces to

$$\begin{aligned} PS(\alpha) &= \sum_{t=1}^{T}(y(t) - m(t) - x(t)\beta(t))^2 \\ &+ \lambda_1 \sum_{t=1}^{T}(m(t) - 2m(t-1) + m(t-2))^2 \\ &+ \lambda_2 \sum_{t=1}^{T}(\beta(t) - 2\beta(t-1) + \beta(t-2))^2 . \end{aligned} \tag{3.10}$$

Introducing $m = (m(1), \ldots, m(T)), \beta = (\beta(1), \ldots, \beta(T)), X = \text{diag}(x(1), \ldots, x(T))$ and defining the pentadiagonal penalty matrix K appropriately, (3.10) can be rewritten as

$$PS(\alpha) = (y - m - X\beta)'(y - m - X\beta) + \lambda_1 m'Km + \lambda_2 \beta' K \beta \,,$$

which is in complete correspondence to the penalized sum of squares for additive models. For dynamic models, however, it is more useful to gather m and β in the 'state' vector α and to rewrite (3.9) in matrix notation as follows: To incorporate initial conditions, we define $y(0) := a_0$, $Z(0) := I$ and redefine $y = (y(0), \ldots, y(T))$. Introducing the (block-)diagonal design matrix

$$Z = \begin{bmatrix} Z(0) & & \cdots & 0 \\ & Z(1) & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & Z(T) \end{bmatrix}$$

and the (block-)diagonal weight matrix

$$W = \begin{bmatrix} Q_0^{-1} & & \cdots & 0 \\ & \Sigma_1^{-1} & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \Sigma_T^{-1} \end{bmatrix}$$

criterion (3.9) can be rewritten as

$$PS(\alpha) = (y - Z\alpha)'W(y - Z\alpha) + \alpha' \tilde{K} \alpha \,, \tag{3.11}$$

with a block-tridiagonal and symmetric penalty matrix $\tilde{K}$. The minimizer $\hat{\alpha}$ of $PS(\alpha)$ is given by

$$\hat{\alpha} = (Z'WZ + \tilde{K})^{-1} Z'Wy \,. \tag{3.12}$$

Since it is the mode of the Gaussian posterior $p(\alpha|y)$, it coincides with the posterior mean $(a_{0|T}, \ldots, a_{t|T}, \ldots, a_{T|T})$, which is computed by the linear Kalman filter and smoother. It computes $\hat{\alpha}$ without explicitly inverting $Z'WZ + \tilde{K}$, by making efficitient use of its block-banded structure and *avoiding any backfitting iterations.* Moreover, as a side product, the block-diagonals $V_{t|T}$ of the smoother matrix are provided. This is useful, for example, to compute cross-validation scores for automatic data-driven choice of smoothing parameters, or in Bayesian terminology, hyperparameters, such as error variances in dynamic models. The Bayesian view is also useful for defining likelihood-based procedures to estimate hyperparameters, see e.g. Harvey (1989).

For multiplicative or varying-coefficient models (2.13), Hastie and Tibshirani (1993) propose to estimate the unknown smooth functions $\beta_0(v_0), \ldots, \beta_p(v_p)$ by minimization of the penalized least squares criterion

$$\begin{aligned} &\sum_{i=1}^{n} w_i(y_i - \beta_0(v_{i0}) - \beta_1(v_{i1})x_{i1} - \ldots - \beta_p(v_{ip})x_{ip})^2 \\ &\qquad + \lambda_0 \int \beta_0''(u)^2 du + \ldots + \lambda_p \int \beta_p''(u)^2 du \,. \end{aligned} \tag{3.13}$$

Criterion (3.13) reduces to the criterion (3.7) for additive models by identifying $v_{i0}, \ldots, v_{ip}$ in (2.13) with the covariates $x_{i1}, \ldots, x_{ip}$ in (3.7) and setting $x_{i1} = \ldots = x_{ip} = 1$ in (3.13). The criterion is also closely related to the penalized least squares criteria (3.9) and (3.11) for dynamic models: In (3.9) and (3.11), covariates $v_0, \ldots, v_p$ are equal the 'covariate' time t, and the penalty terms are discrete time versions of the penalty terms in (3.13), for example second differences in (3.10) compared to second derivatives in (3.13).

To derive the estimation algorithm for multiplicative models let us first consider a simple Gaussian multiplicative model

$$y_i = \beta(v_i)x_i + \varepsilon_i. \tag{3.14}$$

This model is useful when observations $y = (y_1, \ldots, y_n)'$, $x = (x_1, \ldots, x_n)'$ and $v_1, \ldots, v_n$ are metrical, and the ratio y_i/x_i is assumed to vary smoothly over v. Let $v_1 < \ldots < v_u < \ldots < v_U$ be the uniquely ordered sequence of the v_i's, so a $n \times U$ design matrix Z can be defined by its components

$$Z_{iu} = \begin{cases} x_i & \text{if } (y_i, x_i) \text{ is observed at } v_u \\ 0 & \text{else.} \end{cases} \tag{3.15}$$

Using parametrization (3.15) with the coefficients $\beta = (\beta(v_1), \ldots, \beta(v_U))'$, model (3.14) is written as $y = Z\beta + \varepsilon$. Note, that the resulting weighted penalized least squares criterion

$$(y - Z\beta)'W(y - Z\beta) + \lambda\beta' K\beta, \tag{3.16}$$

with W and K defined as above has the same form as for dynamic models in (3.11), but the design matrix is generally different. Equating the first derivatives of (3.16) to zero yields the equation

$$Z'WZ\beta + \lambda K\beta = Z'Wy \tag{3.17}$$

to obtain the estimations $\hat{\beta}$. The corresponding 'ratio–type' smoothing matrix projecting y onto Zf is $\widetilde{S} = Z(Z'WZ + \lambda K)^{-1}Z'W$, where $Z'WZ$ is a diagonal matrix. If the 'discrete' roughness penalties described in (3.6) are used, equation (3.17) can again be solved directly by efficient band–matrix manipulation algorithms. For smoothing splines, the Reinsch algorithm has to be extended by some modifications to get an $O(U)$ algorithm for solving (3.17). Details are given in Klinger (1993) and Hastie and Tibshirani (1993). Using an intercept vector $x_0 = (1, \ldots, 1)'$ to build the matrix Z as defined in (3.15), one obtains a matrix Z_0 which allows simple handling of tied predictor values for related scatterplot smoothers.

With the formulations stated above, criterion (3.13) can be written similarly as for additive models. The weighted penalized least squares criterion

$$\begin{aligned} &PS(\beta_0, \ldots, \beta_p) = \\ &\quad (y - Z_0\beta_0 - Z_1\beta_1 - \ldots - Z_p\beta_p)'W(y - Z_0\beta_0 - Z_1\beta_1 - \ldots - Z_p\beta_p) \\ &\quad +\lambda_0\beta_0'K_0\beta_0 + \ldots + \lambda_p\beta_p'K_p\beta_p \end{aligned} \tag{3.18}$$

yields an analogous system of equations for the minimizing functions $\beta_1, \ldots, \beta_p$ given by

$$\begin{bmatrix} Z_0'WZ_0 + \lambda_0 K_0 & \cdots & Z_0'WZ_p \\ \vdots & \ddots & \vdots \\ Z_p'WZ_0 & \cdots & Z_p'WZ_p + \lambda_p K_p \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} Z_0'Wy \\ \vdots \\ Z_p'Wy \end{bmatrix}. \tag{3.19}$$

Due to the special structure of system (3.19), the backfitting algorithm is again feasible to compute the solutions $\hat{\beta}_0, \ldots, \hat{\beta}_p$. In each backfitting step a 'ratio–type' smoothing matrix

$$Z_j\beta_j^{(1)} = \widetilde{S}_j \left(y - \sum_{h=0}^{j-1} Z_h\beta_h^{(1)} - \sum_{h=j+1}^{p} Z_h\beta_h^{(0)} \right)$$

is applied to actual partial residuals, $\beta^{(0)}$ denotes the results of the previous loop and $\beta^{(1)}$ corresponds to the actual loop. These steps are repeated for $j = 1, \ldots, p, 1, \ldots, p, \ldots$ until convergence in $\beta_0, \ldots, \beta_p$.

3.2 Penalized likelihood estimation

Up to constants, the sums of squares in the penalized least squares criteria (3.7), (3.9) and (3.13) are identical to the sums of (negative) Gaussian log-likelihood contributions of the observations. For generalized additive, dynamic or multiplicative models, these sums of squares are replaced by the sums of non-Gaussian log-likelihoods $l_i(y_i, \eta_i)$ for generalized additive and multiplicative models or $l_t(y(t), \eta(t))$ for generalized dynamic models, with predictors η_i or $\eta(t)$ as in Section 2. For generalized additive or multiplicative models, the minimizing functions are again natural cubic splines and are now obtained by a Fisher scoring or Gauss-Newton algorithm. This can be written as an iteratively weighted least squares algorithm, with an inner backfitting loop in each iteration step, applied to 'working' observations, see Hastie and Tibshirani (1990) for generalized additive models and Klinger (1993) for generalized multiplicative models.

Similarly, filtering and smoothing in generalized dynamic models can be carried out by iteratively weighted Kalman filtering and smoothing algorithms, applied to working observations (Fahrmeir and Tutz, 1994a, ch. 8; Fahrmeir and Wagenpfeil, 1994). The penalized least squares criterion $PS\,(\alpha)$ in (3.9) or (3.11) is replaced by the penalized log-likelihood criterion

$$PL\,(\alpha) = l(\alpha) - \frac{1}{2}\alpha' K\alpha\,,$$

with α and K as in Section 3.1, and

$$l\,(\alpha) = -\frac{1}{2}(\alpha(0) - a_0)Q_0^{-1}(\alpha(0) - a_0) - \sum_{t=1}^{T} l_t(y(t), \eta(t))\,,$$

with individual log-likelihoods l_t and linear predictors $\eta(t) = Z(t)\alpha(t)$.

We define $y = (y(0), \ldots, y(T))$ and $Z = \text{diag}(Z(0), \ldots, Z(T))$ as in Section 3.1. Furthermore we introduce the vector of expectations

$$\mu(\alpha) = (\alpha(0), \mu_1(\alpha(1)), \ldots, \mu_T(\alpha(T))\,,$$

with $\mu_t(\alpha(t)) = h(Z(t)\alpha(t))$, the block diagonal covariance matrix

$$\Sigma(\alpha) = \text{diag}\,(Q_0, \Sigma_1(\alpha(1)), \ldots, \Sigma_T(\alpha(T)))\,,$$

and the block-diagonal matrix

$$D(\alpha) = \text{diag}\,(I, D_1(\alpha(1)), \ldots, D_T(\alpha(T)))\,,$$

where $D_t(\alpha(t)) = \partial h(\eta(t))/\partial\eta$ is the first derivative of the response function $h(\eta)$ evaluated at $\eta(t) = Z(t)\alpha(t)$. Then the first derivative of $PL(\alpha)$ is given by

$$u(\alpha) = \partial PL(\alpha)/\partial\alpha = Z'D(\alpha)\Sigma^{-1}(\alpha)(y - \mu(\alpha)) - K\alpha\,.$$

The expected information matrix is

$$U(\alpha) = -E(\partial^2 PL(\alpha)/\partial\alpha\partial\alpha') = Z'W(\alpha)Z + K$$

with the weight matrix $W(\alpha) = D(\alpha)\Sigma^{-1}(\alpha)D(\alpha)$. A Fisher-scoring step from the current iterate α^0, say, to the next iterate α^1 is then

$$(Z'W(\alpha^0)Z + K)(\alpha^1 - \alpha^0) = Z'D(\alpha^0)\Sigma^{-1}(\alpha^0)(y - \mu(\alpha^0)) - K\alpha^0\,.$$

This can be rewritten as

$$\alpha^1 = (Z'W(\alpha^0)Z + K)^{-1}Z'W(\alpha^0)\tilde{y}^0\,, \tag{3.20}$$

with "working" observation

$$\tilde{y}^0 = D^{-1}(\alpha^0)(y - \mu(\alpha^0)) + Z\alpha^0\,.$$

Comparing (3.20) with (3.12), we see that α^1 can be obtained from the current iterate by applying common linear Kalman filtering and smoothing to the "working" observation $\tilde{y}^0$. In contrast to the iteratively weighted least squares algorithms for additive or multiplictive models, no inner backfitting loop is necessary. Also, the block-diagonal of the smoother matrix, which is required for obtaining confidence bands or cross-validated choice of hyperparameters, is obtained directly from the algorithm.

4. Applications

4.1 Credit-Scoring Revisited

In section 2 we applied a logistic regression model with a linear predictor to analyze consumer's creditworthiness. The maximum likelihood estimates led to the surprising conclusion that the variable 'amount of credit' has no significant influence on the risk of borrowers not paying back their credits. Alternatively we treat it as a generalized additive regression problem, regarding $X3$ ('duration of credit') and $X4$ ('amount of credit') as splined variables. This leads to the additive predictor

$$\eta = \beta_0 + \beta_1 X1[1] + \beta_2 X1[2] + f_3(X3) + f_4(X4) + \beta_5 X5 + \beta_6 X6 + \beta_8 X8.$$

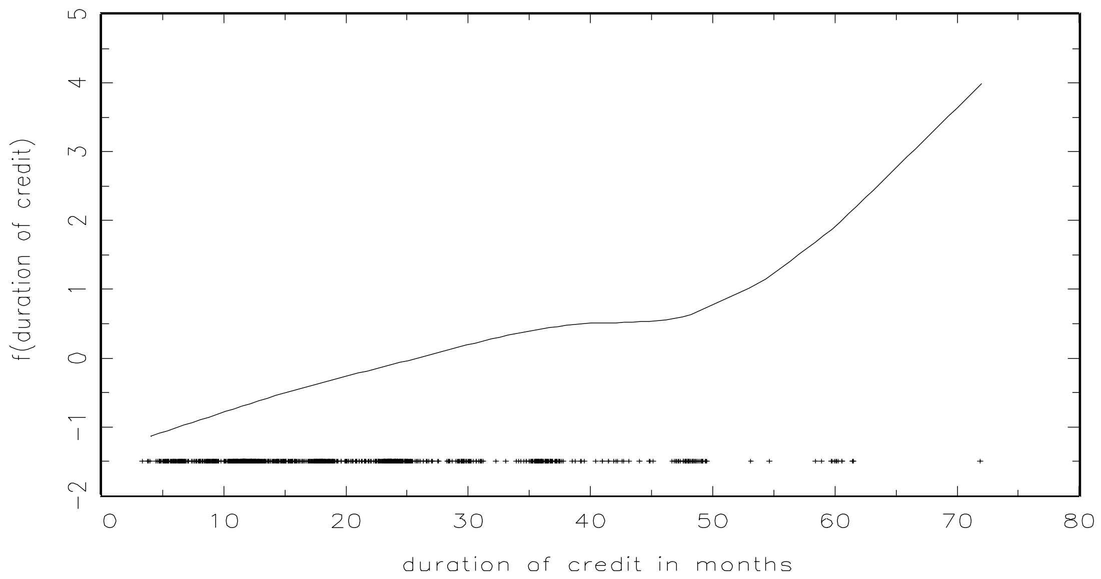

Figure 4.1: Estimated dependence on 'duration of credit'

This nonparametric approach avoids the issue of selecting a particular parametric dependence, e.g. 'linearity', of the response 'creditability' on 'duration of credit' and 'amount of credit'. The point of view we take is: Let the data show us the appropriate form by a smooth curve. The analysis gives the following maximum penalized likelihood estimates of the categorical variables:

	Intercept	X1[1]	X1[2]	X5	X6	X8
value	0.77	0.65	-1.19	-0.91	-0.49	-0.59

In comparison with the linear logistic model the estimated coefficients change only slightly. The estimated curves are shown in Figure 4.1 and Figure 4.2 (solid line). While the variable 'duration' is not far away from linearity, the estimate of 'amount of credit' is clearly not linear. The curve shows that not only high credits but also low credits (below 4000 DM) increase the risk. The smoothing parameters have been chosen by vision. A data-driven choice of the smoothing parameters, e.g. by generalized cross-validation, is possible in principle. However, efficient computation would be required.

Since the curve of 'duration' in the logistic additive model is almost linear, we reanalyze the data with a logistic semiparametric model of the form (2.2), with predictor

$$\eta = \beta_0 + \beta_1 X1[1] + \beta_2 X1[2] + \beta_3 X3 + f_4(X4) + \beta_5 X5 + \beta_6 X6 + \beta_8 X8.$$

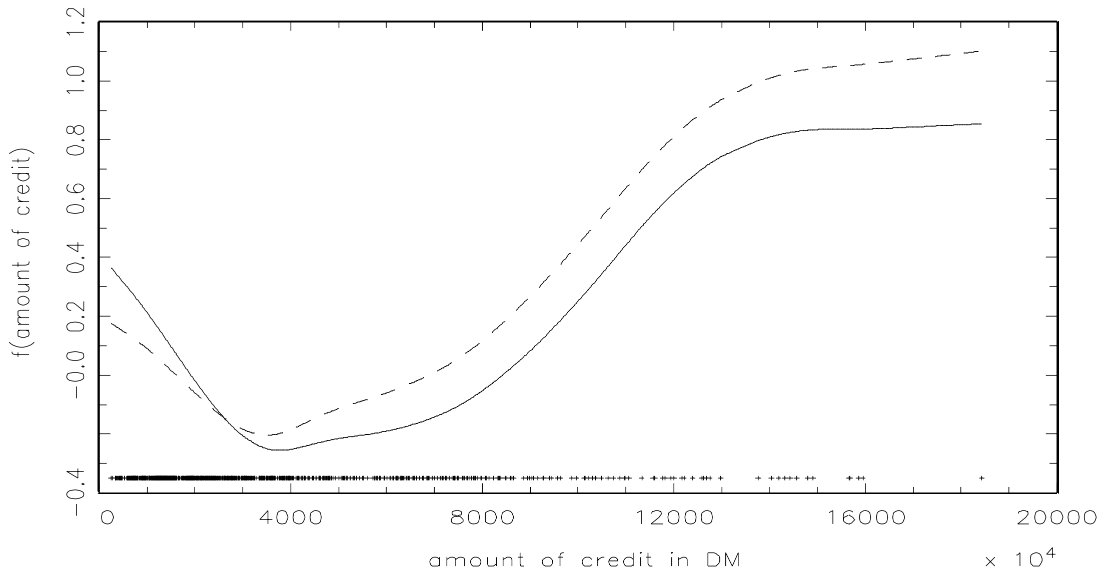

Figure 4.2: Estimated dependence on 'amount of credit'

The advantage is that we can avoid the backfitting loop and are able to compute the generalized cross-validation score in a simple way (see Green and Silverman, 1994, ch.4). Unfortunately the minimization of the cross-validation criterion yields only a global minimum $\lambda = 0$, which corresponds to a very rough estimate of the variable 'amount of credit'. So we have chosen the same smoothing parameter as above. We get the estimates of the fixed coefficients:

	Intercept	X1[1]	X1[2]	X3	X5	X6	X8
value	0.02	0.66	-1.24	0.03	-0.82	-0.50	-0.52

They are again not far away from the estimates of the logistic linear model. The estimated dependence on 'amount of credit' is shown in Figure 4.2 (dashed line). The form is very similar to the logistic additive model.

It seems that the logistic semiparametric model itself is a good model for the credit scoring data. If someone is interested in getting a parametric model, the semiparametric model can be used as a starting point for further analysis.

4.2 IFO business test

In Example 2, time-varying thresholds $m_1(t)$, $m_2(t)$ and covariate effects $\beta_1(t), \ldots, \beta_6(t)$ were modelled by an eight-dimensional random walk of first order. Smoothing estimates of the covariate parameters are displayed in Figure

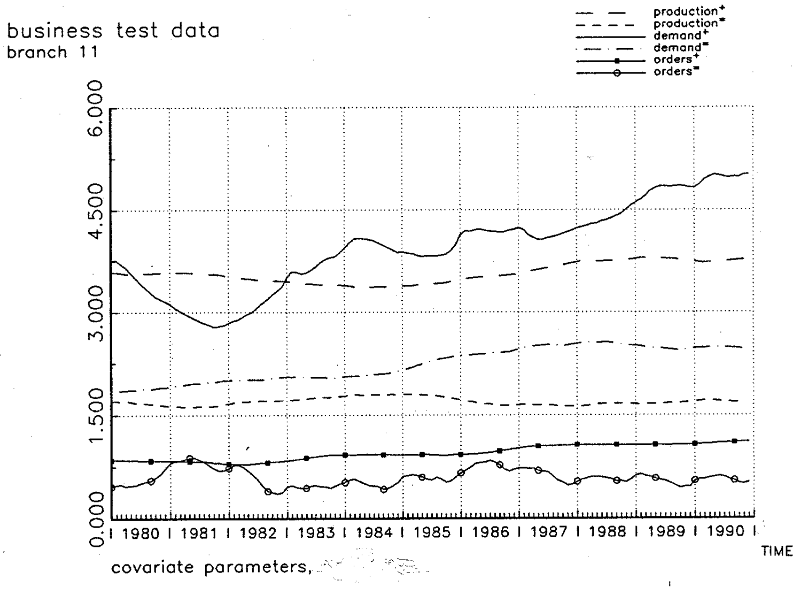

Figure 4.3: Covariate effects

4.3. Apart from the $D^+(t)$–parameter all effects are nearly constant in time. An increase of production plans in the previous month $(P^+(t-1))$ has a high positive influence on current production plans, while the effect of $P^=(t-1)$ is still positive but distinctly smaller. Both effects are in agreement with continuity in planning production. Compared to the effects of $D^+(t), D^=(t)$, which are both clearly positive on the average, the effects of increasing or constant orders in hand $(O^+(t), O^=(t))$ are still positive but surprisingly small. This result, which is in agreement with previous findings, can be explained as follows: The variable D serves as a substitute for expected demand. For the purpose of short–range production planning, expected demand is more relevant than current orders at hand, which are more relevant for current production.

Compared to the remaining effects, the parameter $\beta_3(t)$ corresponding to the increase category D^+ of expected development of business has a remarkable temporal variation. It exhibits a clear decline to a minimum at the beginning, and a distinct increase period coincides with the first months of the new German government in autumn 1982, ending with the elections to the German parliament in 1983. The growing positive effect of a positive state of business to the "increase" category of production plans indicates positive reactions of firms to the change of government.

In Figure 4.4 both thresholds (solid line) exhibit seasonal variation correspon-

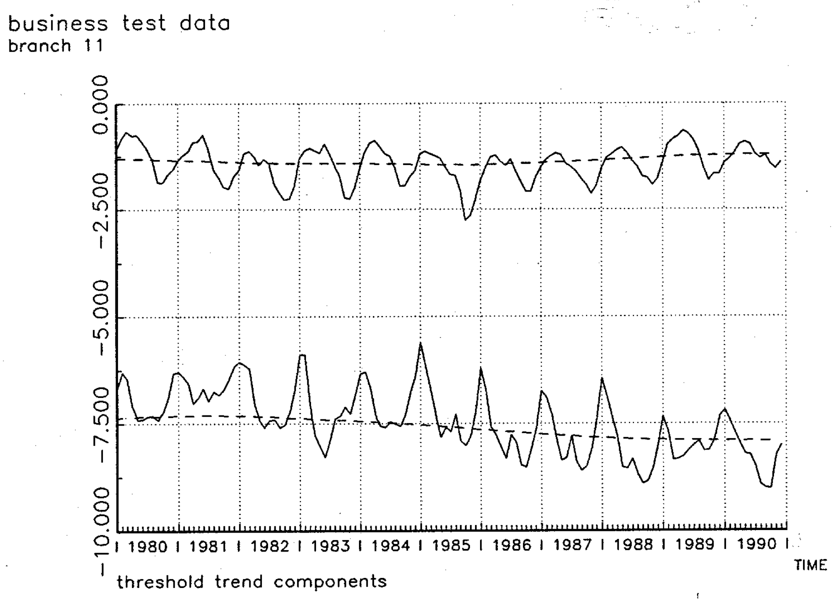

Figure 4.4: Trend parameters

ding to successive years. Threshold parameter $m_1(t)$ has peaks, mostly rather distinct, in December or January, and low values in the summer months. An explanation for this seasonal behaviour, which is not captured by covariate effects, may be the following: Firms in this specific branch manufacture initial products for the building industry. To be able to satisfy the increasing demand for their products in late winter/early spring, production plans are increased 2 to 3 months earlier. This is in agreement with the model, since higher values of $m_1(t)$ result in higher probabilities for increasing production plans, keeping covariate effects fixed. Similarly, decreasing values in spring and low values in summer reflect the tendency not to increase an already comparably high level of production any further. The ups and downs of the second threshold parameter appear some months later. Interpretation is analogous and corresponds to seasonal ups and downs in the tendency of firms not to change their current production plans. To specify this seasonal effect more explicitly, a seasonal component in trigonometric form was included additionally. Since seasonal variation is now modelled by these components, the trend parameters are now more or less constant in time (dashed line in Figure 4.4).

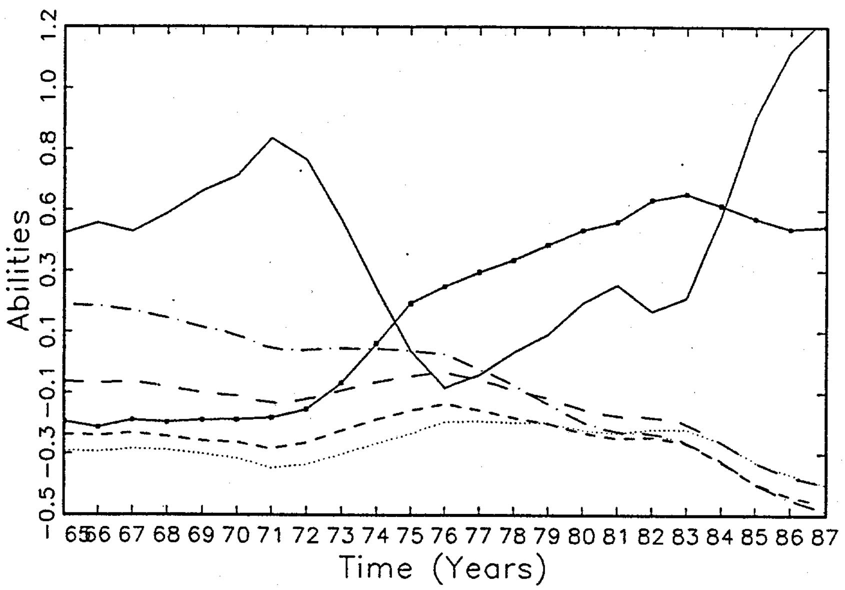

Figure 4.5: Kalman filter and smoother for soccer data based on a random walk of first order. Teams are Bayern München (——), 1.FC Köln (– · –), VfB Stuttgart (· · ·), 1.FC Kaiserslautern (- - -), Hamburger SV (—•—) and Eintracht Frankfurt (– –)

4.3 Dynamic Pair Comparisons for the German Fußball-Bundesliga

We apply the ordinal logistic paired comparison model of Example 3 to data for the teams Bayern München, 1.FC Köln, VfB Stuttgart, 1.FC Kaiserslautern, Hamburger SV and Eintracht Frankfurt for the years 1966 to 1987. Thresholds and abilities are modelled by first order random walks. For the thresholds the estimated variances are 0.001 and 0.008. That means the thesholds in fact remain rather stable over years. For the abilities the estimated variances in Q are 0.124, 0.006, 0.005, 0.002 and 0.027. Figure 4.5 shows the smoothed abilities for the six teams based on these estimated hyperparameters. The large variance of the first team (0.124) and the fifth team (0.027) may also be seen from the picture which shows strong fluctuation for Bayern München (team 1) and comparatively high fluctuation for Hamburger SV (team 5) whereas the other teams are quite constant. The highs and lows of Bayern München are in good agreement with the development, coming and going of important players and coaches. For example the peak about 1970–1972 coincides with the most successful years of the team with Franz Beckenbauer as captain and other important members of the national team at that time. While still successful in European cup finals till 1974, success

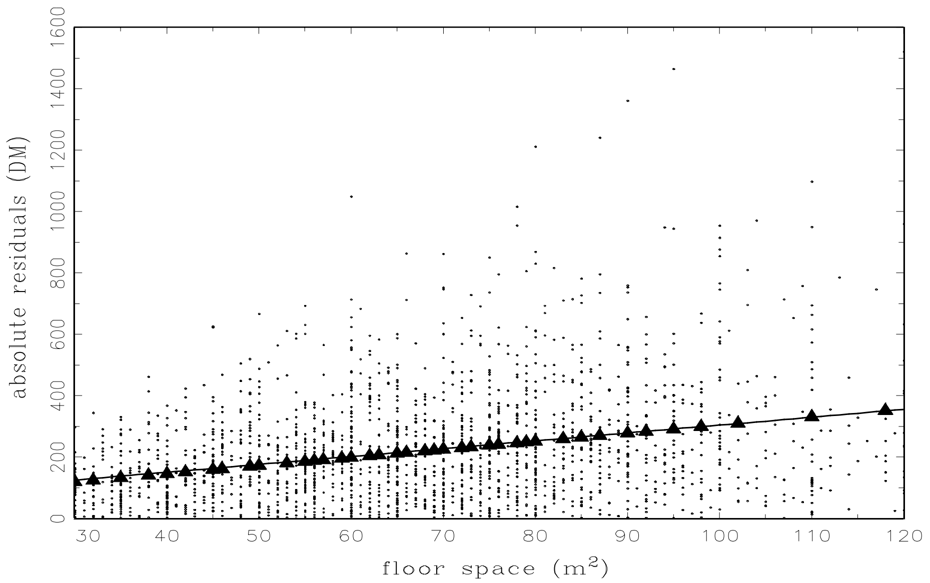

Figure 4.6: Absolute residuals of the unweighted estimation for the tenancy data. The solid line is the linear regression used to determine the weights. The triangles correspond to a linear regression for the absolute residuals computed from the weighted estimation.

was steadily declining in the German national league, eventually leading to a distinct low when Franz Beckenbauer went to Cosmos New York and others left the team. It took some time to form a new team which became better and eventually very successful again in the late 80's. In this later period Hamburger SV, which had become more and more powerful, and Bayern München were the dominating teams in the national soccer league. An alternative analysis with local linear trend models gives rather similar results, see Fahrmeir and Tutz (1994b).

4.4 Rental tables

As introduced in Section 2, a seven component multiplicative model

$$y = \beta_1(F)+F\beta_2(A)+\beta_3(F)S^-+\beta_4(F)S^++\beta_5(F)H+\beta_6(F)B+\beta_7(F)L+\epsilon \quad (0.1)$$

is suggested to analyse the tenancy survey. Since, in contrast to floor space (F), the variable age (A) has no meaningful origin, we use the interaction term $F\beta_2(A)$ instead of $\beta_2(F)A$. For penalizing the roughness of each effect β_j the integrated squared curvature is applied again.

The smoothing parameters $\lambda_1, \dots, \lambda_7$ for the cubic smoothing splines were selected automatically by an adaptive backfitting algorithm similar to BRUTO as

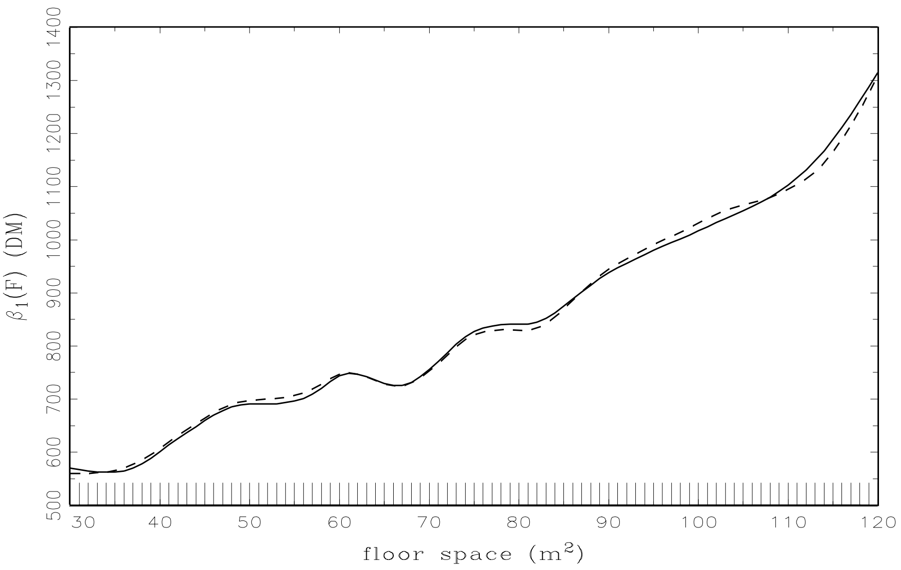

Figure 4.7: Basic rent depending on floor space. (—— weighted regression, – – – unweighted regression)

proposed by Hastie (1989). Each backfitting step is divided into two steps (i) and (ii). In step (i) the (univariate) trade–off parameter λ_j is chosen to minimise a generalized cross–validation score (GCV) depending on partial residuals within a given range $I_j = [\lambda_j(l), \lambda_j(h)]$. In a following step (ii) the actual smoothing parameters $\lambda_1, \ldots, \lambda_j, \ldots, \lambda_7$ are considered to be fixed and a backfitting algorithm is applied to update all coefficients of the model simultaneously. By initialising the 'inner' backfitting in step (ii) with the estimation result of step (i), convergence is usually reached after the first or during the second loop. Step (i) and step (ii) are alternated for $j = 1, \ldots, 7, 1 \ldots, 7, \ldots$ until any convergence criterion in $\lambda_1, \ldots, \lambda_7$ and $\beta_1, \ldots, \beta_7$ is reached. When arrived after a full 'outer' loop at the j-th covariate again, the interval I_j is shifted, depending on the location of the GCV-minimal λ_j in I_j found in the previous loop. Hence the algorithm is capable to find smoothing parameters within a total range $(0, \infty)$. Details of this method and extensions to the non–Gaussian case are described in Klinger (1993).

The absolute residuals computed from unweighted penalized least squares estimation shown in Fig. 4.6 are indicating a heterogeneous error variance depending on floor space. Therefore we estimate the coefficients in two steps, similarly as in linear models (see e.g. Carroll and Ruppert, 1988). To obtain weights for a

component	$\mathrm{tr}(\widetilde{S}_j)$	
estimation:	*unweighted*	*weighted*
$\beta_1(F)$	13.0486	13.0885
$F\beta_2(A)$	6.05547	5.66461
$\beta_3(F)S^-$	2.31410	2.35575
$\beta_4(F)S^+$	6.07736	2.00000
$\beta_5(F)H$	4.61022	4.18226
$\beta_6(F)B$	2.35605	2.35024
$\beta_7(F)L$	2.00000	2.00000
WRSS	3122.76	3131.46

Table 4.1: Traces of the components smoother matrices for the rental–table model.

weighted penalized least squares estimation, a linear regression of the form

$$|r_i| = \gamma_0 + \gamma_1 F + \epsilon$$

is applied to the absolute residuals $|r_i|$ resulting from the unweighted estimation. The weights used in the final estimation are then given by

$$w_i = (\hat{\gamma}_0 + \hat{\gamma}_1 F)^{-2}.$$

As shown in the two linear regressions in Fig. 4.6 a further estimation step would use almost the same weights, and therefore no great differences in estimation results could be expected.

A comparison of the fit to the data by the weighted residual sum of squares (WRSS) in Tab. 4.1, shows that the unweighted estimation with automatically chosen smoothing parameters has even a slightly better fit. Viewing the trace of the 'ratio–type'–smoothing matrices given in Tab. 4.1 as an approximation to 'individual' degrees of freedom, the weighted regression seems to compensate the loss of fit by stronger smoothness restrictions.

The estimated functions for basic rent, depending on floor space and year of construction, obtained by weighted and unweighted regression are quite similar (Fig. 4.7 and Fig. 4.8). An interesting result is the rent reduction for flats constructed in the post-war era during the 1950's and the steep ascent for recently built apartments shown in Fig. 4.8. For a careful investigation of this fact additional covariates describing type and equipment of flats, like renovation or balcony, would have to be included.

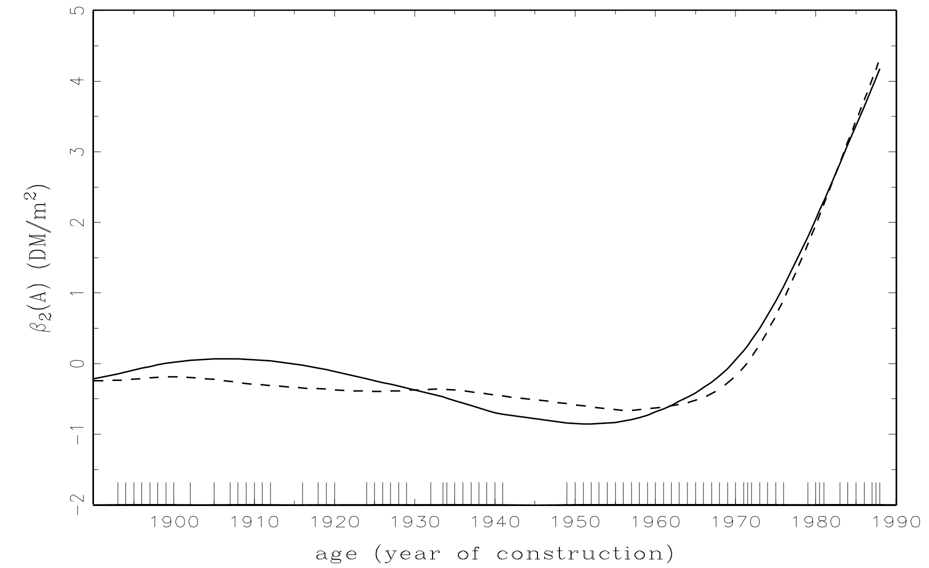

Figure 4.8: Correction of basic rent in DM/m^2 depending on the year of construction. (—— weighted regression, – – – unweighted regression)

An example for the improvement obtained from weighted regression is given in Fig. 4.10 and Fig. 4.12. For the two terms $\beta_4(F)S^+$ and $\beta_6(F)B$ unweighted regression leads to no plausible results since the effects are expected to increase or decrease monotonously in floor space. Interestingly both traces of the component $\beta_6(F)B$ are nearly the same (see Tab. 4.1). Therefore the different results shown in Fig. 4.12 are due to a reduction of weights for the few bigger flats without bathroom.

Advantages of this nonparametric approach can be studied by the influence of no central heating (H) in Fig. 4.11. Here it seems that a less efficient heating system is more disadvantageous in bigger apartments than in smaller ones. In addition the discount on site below average in Fig. 4.9 and on missing bathroom in Fig. 4.12 is decreasing less than linearly in floor space as has to be supposed by linear regression analysis. For the influence due to equipment of bathroom (L) automatic selection of smoothing parameters indicates linear effects in both estimates (Fig. 4.13).

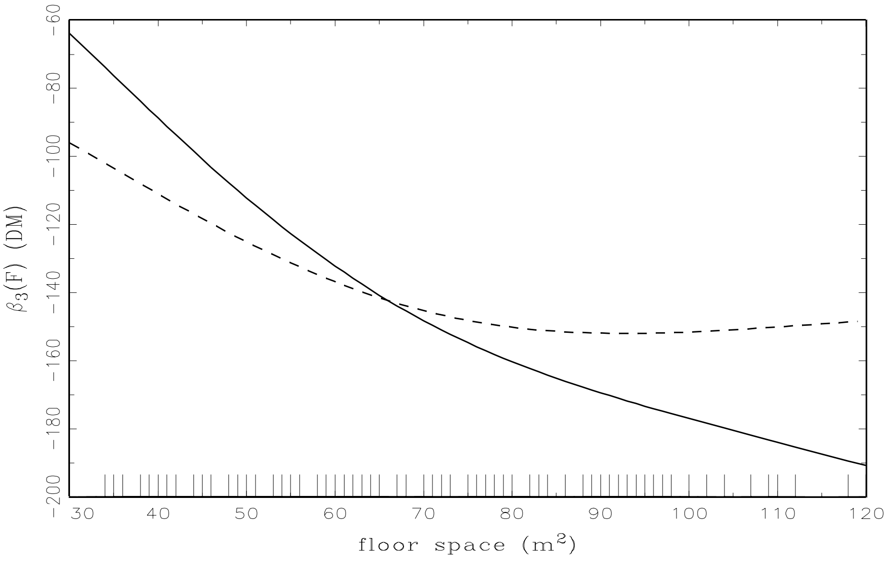

Figure 4.9: Reduction on rent for apartments located in sites below average depending on floor space. (—— weighted regression, – – – unweighted regression)

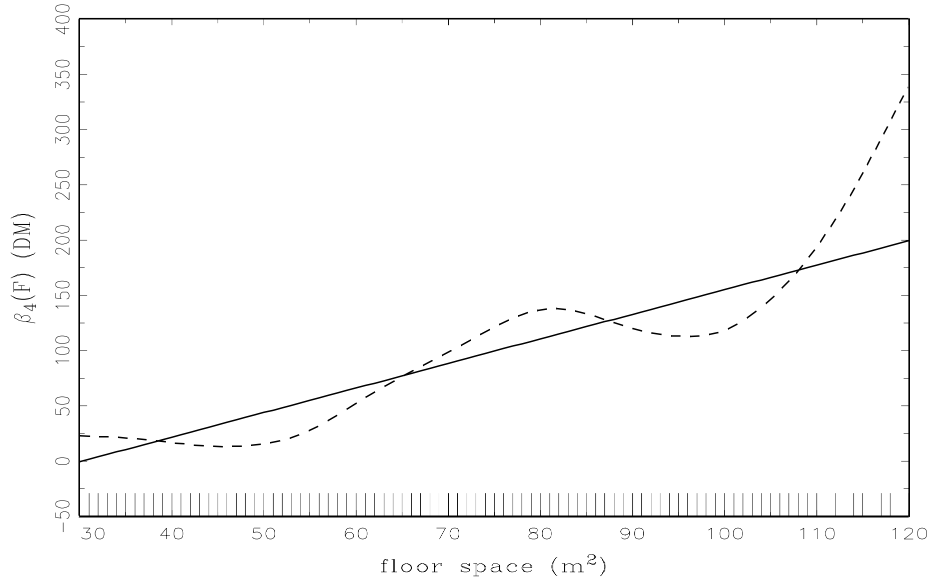

Figure 4.10: Surcharge on rent for apartments located in sites above average depending on floor space. (—— weighted regression, – – – unweighted regression)

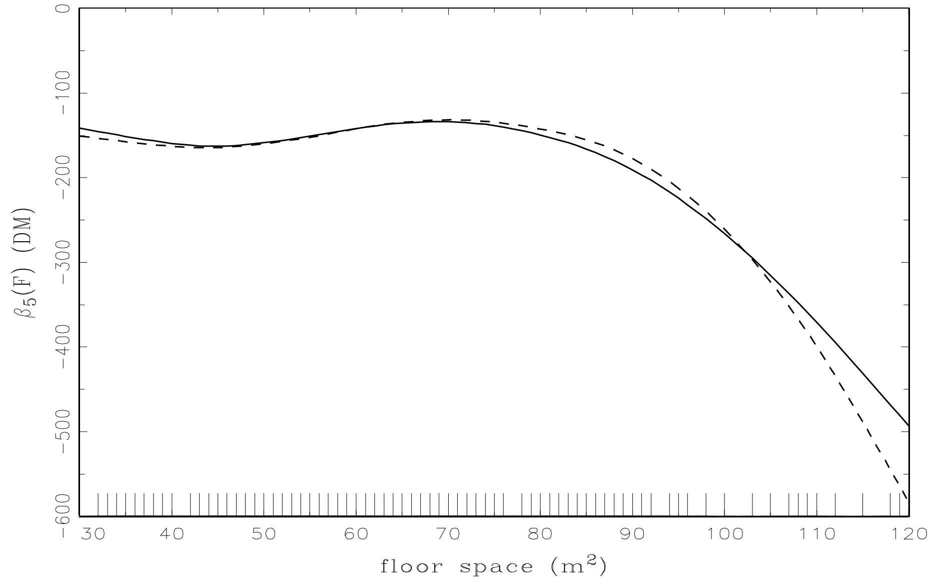

Figure 4.11: Reduction on rent for apartments without central heating depending on floor space. (—— weighted regression, – – – unweighted regression)

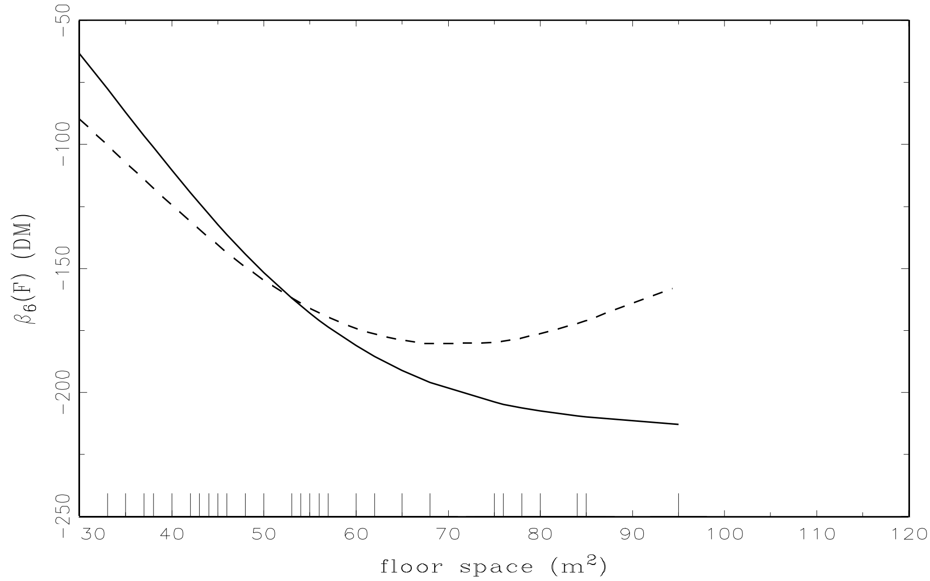

Figure 4.12: Reduction on rent for apartments without bathroom depending on floor space. (—— weighted regression, – – – unweighted regression)

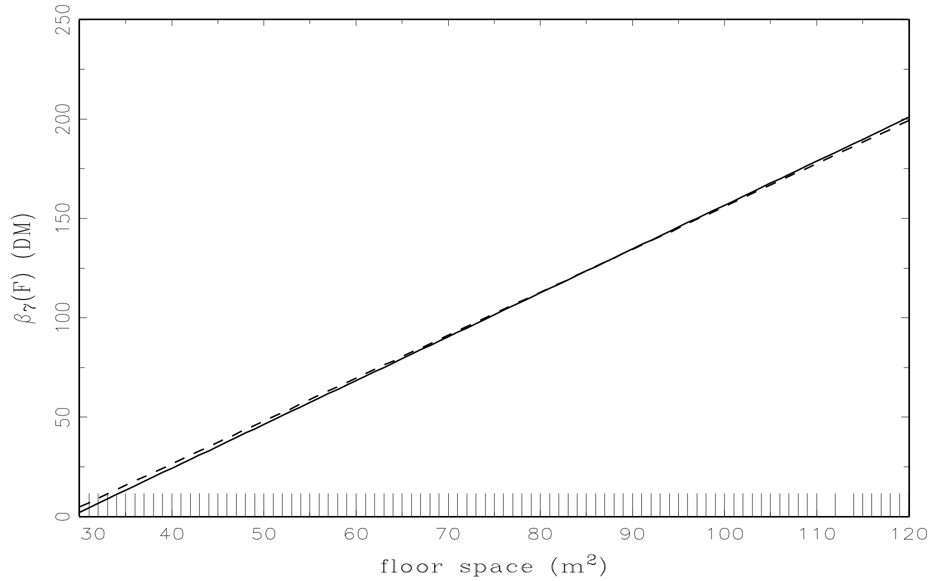

Figure 4.13: Surcharge on rent for apartments with luxury fitted bathrooms depending on floor space. (—— weighted regression, – – – unweighted regression)

5. Concluding remarks

Due to its flexibility and versatility, the roughness penalty approach provides a unifying framework for non- and semiparametrically modelling and estimation in various settings of regression analysis. Dynamic or state space models can be viewed as Bayesian versions of varying-coefficient or multiplicative models if estimation is based on maximization of posterior densities. We conclude by pointing out some topics for further research.

- Extensions to multicategorical or multivariate correlated responses, e.g. semiparametric marginal models for clustered data or repeated measurements, are possible by introducing appropriate (quasi-)likelihoods.

- Monotonicity or concavity of functions $f(x)$ can be accounted for by appropriate modification of penalty functions.

- Identification and choice of models needs to be further developed.

- Still more efficient algorithms, for example avoiding the backfitting loops, would be useful, in particular in combination with data-driven selection of smoothing parameters or hyperparameters.

- For mixed continuous and discrete covariates, more flexible approaches than a semiparametric additive model like (2.2) should be available. This might be accomplished by combining the features of classification and regression trees (CART) and smoothing techniques.

- If one is willing to adopt Bayesian formulations in form of state space models, full posterior analysis or at least posterior mean estimation will be the ultimate goal. It seems that Gibbs sampling or related data augmentation techniques are most promising and general tools for Bayesian estimation.

References

BUJA, A., HASTIE, T., TIBSHIRANI, R. (1989). Linear Smoothers and Additive Models (with discussion). *Annals of Statistics* 17, 453-555.

CARROLL R.J., RUPPERT, D. (1988) *Transformations and Weighting in Regression.* New York: J. Wiley & Sons.

FAHRMEIR, L. (1992). Posterior Mode Estimation by Extended Kalman Filtering for Multivariate Dynamic Generalized Linear Models. *Journal of the American Statistical Association*, 87, 501–509.

FAHRMEIR, L., HAMERLE, A. (1984, Eds.). *Multivariate statistische Verfahren.* Berlin: De Gruyter.

FAHRMEIR, L., KREDLER, CH. (1984). Verallgemeinerte Lineare Modelle. In: Fahrmeir, L., Hamerle, A. (Eds.). *Multivariate statistische Verfahren.* Berlin: De Gruyter.

FAHRMEIR, L., NASE, H. (1994). Dynamische Modellierung und Analyse von Mikrodaten des Konjunkturtests. *IFO-Studien* 40, S.1-22.

FAHRMEIR, L., TUTZ, G. (1994a). *Multivariate Statistical Modelling Based on Generalized Linear Models.* New York: Springer-Verlag.

FAHRMEIR, L., TUTZ, G. (1994b). Dynamic Stochastic Models for Time-Dependent Ordered Paired Comparison Systems. *Journal of the American Statistical Association*, to appear.

FAHRMEIR, L., WAGENPFEIL, S. (1994). Iteratively Weighted Kalman-Filtering and Smoothing for Exponential State Space Models. Technical Report, Universität München.

GREEN, P.J., SILVERMAN, B.W. (1994). *Nonparametric Regression and Generalized Linear Models.* London: Chapman & Hall.

HÄRDLE, W. (1990). *Smoothing Techniques.* With Implementation in S. New York: Springer Verlag.

HARVEY, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press.

HASTIE, T.(1989) Discussion of "Flexible parsimonious smoothing and additive modelling" by J. Friedman and B. Silverman. *Technometrics* **31**, 3–39.

HASTIE, T., TIBSHIRANI, R. (1990). *Generalized Additive Models.* London: Chapman and Hall.

HASTIE, T., TIBSHIRANI, R. (1993) Varying–coefficient Models. *J. R. Statist. Soc. B* **55** 757–496.

KLINGER, A. (1993) Spline–Glättung in zeitdiskreten Verweildauermodellen. Diplomarbeit, Institut für Statistik, Universität München.

MC CULLAGH, P., NELDER, J.A. (1983, 1989 2d ed.). *Generalized Linear Models.* New York: Chapman and Hall.

REINSCH, C. (1967). Smoothing by Spline Functions. *Numerische Mathematik*, 10, 177–183.

THIELE, T. (1880). Sur la Compensation de Quelques Erreurs Quasi–Systematiques par la Methode des Moindres Carrees. Copenhagen: Reitzel.

TUTZ, G. (1986). Bradley–Terry–Luce Models with an Ordered Response. *J. of Mathematical Psychology*, 30, 306–316.

WAHBA, G. (1978). Improper Prior, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression. *Journal of the Royal Statistical Society*, B 44, 364–372.

WEST, M., HARRISON, P.J. (1989). *Bayesian Forecasting and Dynamic Models.* New York: Springer.

WEST, M., HARRISON, P.J., MIGON, M. (1985). Dynamic Generalized Linear Models and Bayesian Forecasting. *Journal of the American Statistical Association*, 80, 73–97.

WHITTAKER, E.T. (1923). On a New Method of Graduation. *Proc. Edinborough Math. Assoc.*, 78, 81–89.