Fahrmeir, Knorr-Held:

Dynamic and semiparametric models

Sonderforschungsbereich 386, Paper 76 (1997)

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

Dynamic and semiparametric models

Ludwig Fahrmeir and Leonhard Knorr–Held
Universität München
Institut für Statistik
Ludwigstr. 33, 80539 München

1 Introduction

This chapter surveys dynamic or state space models and their relationship to non– and semiparametric models that are based on the roughness penalty approach. We focus on recent advances in dynamic modelling of non–Gaussian, in particular discrete–valued, time series and longitudinal data, make the close correspondence to semiparametric smoothing methods evident, and show how ideas from dynamic models can be adopted for Bayesian semiparametric inference in generalized additive and varying coefficient models. Basic tools for corresponding inference techniques are penalized likelihood estimation, Kalman filtering and smoothing and Markov chain Monte Carlo (MCMC) simulation. Similarities, relative merits, advantages and disadvantages of these methods are illustrated through several applications.

Section 2 gives a short introductory review of results for the classical situation of Gaussian time series observations. We start with Whittaker's (1923) "method of graduation" for estimating trends and show that it is equivalent to the posterior mean estimate from a linear Kalman filter model with known smoothing or variance parameters. We sketch extensions to general Gaussian linear dynamic or state space models and to continuous time analogues like the Bayesian version of cubic spline smoothing (Wahba, 1978). For more detailed expositions of the equivalence between Bayesian smoothness priors and penalized least squares we refer the reader to Kohn and Ansley (1988) and previous work cited there and to van der Linde (1995, 1996) for a thorough discussion of splines from a Bayesian point of view. This equivalence also suggests alternative ways of estimating unknown smoothing or variance pa-

rameters: Within a semiparametric approach, estimation by optimizing some cross–validated criterion is a common choice. Empirical Bayes models, also treating hyperparameters as fixed or unknown, lead to marginal likelihood estimation. Maximization can be done by EM–type algorithms. Fully Bayesian models put a weakly informative prior on the hyperparameters and make a complete posterior analysis with MCMC techniques feasible.

We then turn briefly to so–called conditionally Gaussian dynamic models that are still linear but with errors distributed as scale mixtures of normals. Already with this seemingly moderate generalization, penalized least squares and posterior mean estimates are no longer equivalent. Beyond various approximate Kalman filters and smoothers, fully Bayesian approaches based on MCMC are available that make efficient use of the conditionally Gaussian structure.

Fundamentally non–Gaussian time series and longitudinal data, in particular for categorical and count data, are considered in Section 3. Dynamic binomial and Poisson models are important members of the family of dynamic generalized linear models. Semiparametric counterparts based on penalized likelihood estimation can be derived as posterior mode estimators, with extended or iterative Kalman–type smoothing algorithms as efficient computational tools (Fahrmeir 1992; Fahrmeir and Tutz, 1994, ch. 8, Fahrmeir and Wagenpfeil, 1996). However, the equivalence between posterior mean and penalized likelihood estimation is lost. Fully Bayesian inference is possible with recently developed MCMC techniques for non–Gaussian dynamic models, an area of intensive current research. In Section 3.2, we outline the ideas for Metropolis–Hastings algorithms suggested by Knorr–Held (1996). These algorithms are used for the applications and are generalized in Section 3.3 to non–normal longitudinal data with additional unobserved population heterogeneity across units.

Ideas from non–Gaussian dynamic modelling, in particular for non–equally spaced or continuous–time parameter models, can be transferred to semiparametric regression models for cross–sectional data (Section 4). This leads to Bayesian spline–type smoothing for generalized additive and varying coefficient models using MCMC techniques as a supplement and alternative to penalized likelihood or, equivalently, posterior mode estimation (Hastie and Tibshirani, 1990, 1993).

Finally Section 5 summarizes conclusions and indicates extensions to other data situations and statistical models.

2 Linear dynamic models and optimal smoothing for time series data

This section gives a brief survey on the correspondence between linear dynamic or state space models and semiparametric optimal smoothing methods based on the roughness penalty approach. We illustrate this correspondence by some simple and commonly–used examples and review more general and recent work.

2.1 Gaussian models

In the classical smoothing problem treated by Whittaker (1923), time series observations $y = (y_1, \ldots, y_T)$ are assumed to be the sum

$$y_t = \tau_t + \epsilon_t, \quad t = 1, \ldots, T \tag{1}$$

of a smooth trend function τ and an irregular noise component ϵ. Whittaker suggested to estimate τ by minimizing the penalized least squares criterion

$$\mathrm{PLS}(\tau) = \sum_{t=1}^{T}(y_t - \tau_t)^2 + \lambda \sum_{t=3}^{T}(\tau_t - 2\tau_{t-1} + \tau_{t-2})^2 \tag{2}$$

with respect to $\tau = (\tau_1, \ldots, \tau_T)$. Minimization of $\mathrm{PLS}(\tau)$ tries to hold the balance between fit of the data, expressed by the sum of squares on the left side and smoothness of the trend, expressed by the roughness penalty term in form of the sum of squared differences. The smoothness parameter λ, assumed to be given or fixed, weights the two competing goals data fit and smoothness.

The trend model (1) and the PLS criterion (2) can be generalized in a flexible way. Inclusion of a seasonal component $\gamma = (\gamma_1, \ldots, \gamma_T)$ with period m leads to the additive trend–seasonal

model

$$y_t = \tau_t + \gamma_t + \epsilon_t, \quad t = 1, \ldots, T \tag{3}$$

and the PLS criterion

$$\text{PLS}(\tau, \gamma) = \sum_{t=1}^{T}(y_t - \tau_t - \gamma_t)^2 + \lambda_1 \sum_{t=3}^{T}(\tau_t - 2\tau_{t-1} + \tau_{t-2})^2 + \lambda_2 \sum_{t=m}^{T}(\gamma_t + \gamma_{t-1} + \ldots + \gamma_{t-m+1})^2 \to \min_{\tau,\gamma} \tag{4}$$

for estimating the trend function τ and the seasonal component γ. More generally, the influence of covariates can be taken into account by extending the additive predictor $\tau_t + \gamma_t$ to

$$\eta_t = \tau_t + \gamma_t + x_t'\beta_t + w_t'\delta, \tag{5}$$

with time–varying effects β_t for x_t and constant effects δ for w_t.

This penalized least squares approach is reasonable if time series observations are – at least approximately – Gaussian. This is made explicit by assuming that the errors ϵ_t in (1) and (3) are i.i.d. $N(0, \sigma^2)$ random variables. Then the fit term in (4) corresponds to the log–likelihood of the additive Gaussian observation model (3), and the PLS approach appears as a semiparametric method for estimating the fixed, unknown sequences $\tau = (\tau_1, \ldots, \tau_T)$, $\gamma = (\gamma_1, \ldots, \gamma_T)$.

A dynamic model corresponding to (3) and (4) considers τ and γ as sequences of random variables. It is hierarchical and consists of two stages: The first stage is the Gaussian observation model (3) for y given τ and γ. In the second stage, a transition model corresponding to the roughness penalty term in (4) is given by the difference equations

$$\tau_t - 2\tau_{t-1} + \tau_{t-2} = u_t, \quad t = 3, \ldots, T \tag{6}$$

$$\gamma_t + \gamma_{t-1} + \ldots + \gamma_{t-m+1} = w_t, \quad t = m, \ldots, T \tag{7}$$

The errors u_t and w_t are i.i.d. $N(0, \sigma_u^2)$– and $N(0, \sigma_w^2)$– distributed. Initial values are specified for example by

$$(\tau_1, \tau_2)' \sim N(0, k_\tau I), \quad (\gamma_1, \ldots, \gamma_{m-1})' \sim N(0, k_\gamma I). \tag{8}$$

All errors and initial values are assumed as mutually independent. The difference equation (6), also called a random walk of second order, penalizes deviations from the linear trend $\tau_t =$

$2\tau_{t-1} - \tau_{t-2}$. The seasonal model (7) prefers "small" values of the sum $\gamma_t + \gamma_{t-1} + \ldots + \gamma_{t-m+1}$, so that the seasonal pattern does not change too much over periods. From a Bayesian point of view (6) and (7), together with (8), define a multivariate normal prior $\mathrm{p}(\tau, \gamma) = \mathrm{p}(\tau)\mathrm{p}(\gamma)$ for $(\tau, \gamma) = (\tau_1, \ldots, \tau_T, \gamma_1, \ldots, \gamma_T)$, the so–called smoothness prior. Also, the observation model (3) defines a multivariate normal distribution $\mathrm{p}(y|\tau, \gamma)$ for the data y given τ and γ. Here the hyperparameters, i.e. the variances σ^2, σ_u^2, σ_w^2, are regarded as known or given constants. Thus, the posterior

$$\mathrm{p}(\tau, \gamma|y) = \frac{\mathrm{p}(y|\tau, \gamma)\mathrm{p}(\tau)\mathrm{p}(\gamma)}{\mathrm{p}(y)} \propto \mathrm{p}(y|\tau, \gamma)\mathrm{p}(\tau)\mathrm{p}(\gamma) \tag{9}$$

is also normal and characterized by the posterior expectation $E(\tau, \gamma|y)$ and covariance $Var(\tau, \gamma|y)$. Due to normality, the posterior expectation and the posterior mode, i.e. the maximizer of (9), coincide. Taking logarithms, using the (conditional) independence assumptions and ignoring constant factors leads to the criterion

$$\begin{aligned} & \frac{1}{\sigma^2}\sum_{t=1}^{T}(y_t - \tau_t - \gamma_t)^2 + \frac{1}{k_\tau}(\tau_1^2 + \tau_2^2) + \frac{1}{\sigma_u^2}\sum_{t=3}^{T}(\tau_t - 2\tau_{t-1} + \tau_{t-2})^2 \\ + \quad & \frac{1}{k_\gamma}(\gamma_1^2 + \ldots + \gamma_{m-1}^2) + \frac{1}{\sigma_w^2}\sum_{t=m}^{T}(\gamma_t + \ldots + \gamma_{t-m+1})^2 \to \min_{\tau,\gamma} \end{aligned} \tag{10}$$

for estimating τ, γ. Setting $\lambda_1 = \sigma^2/\sigma_u^2$; $\lambda_2 = \sigma^2/\sigma_w^2$ and choosing diffuse priors (8) with $k_\tau \to \infty$, $k_\gamma \to \infty$, the criteria (4) and (10) are identical so that the semiparametric PLS estimate $\hat{\tau}$, $\hat{\gamma}$ is identical to the posterior mode estimate and, due to posterior normality, the posterior mean:

$$(\hat{\tau}, \hat{\gamma}) = E(\tau, \gamma|y). \tag{11}$$

This equivalence remains valid for more general linear Gaussian observation and transition models, see Kohn and Ansley (1988) for a thorough treatment. Collecting trend, season and other parameters as in (5) in a so–called state vector α_t, e.g.

$$\alpha_t = (\tau_t, \tau_{t-1}, \gamma_t, \ldots, \gamma_{t-m+1}, \beta_t, \delta),$$

most linear dynamic models can be put in the form of Gaussian linear state space models

$$y_t = z_t'\alpha_t + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma^2) \tag{12}$$

$$\alpha_{t+1} = F_t\alpha_t + v_t, \qquad v_t \sim N(0, Q_t) \tag{13}$$

by appropriate definition of design vectors z_t and transition matrices F_t, see for example Harvey (1989), West and Harrison (1989) or Fahrmeir and Tutz (1994, ch. 8.1). The well–known classical Kalman filter and smoother or recent variants like the diffuse filter (de Jong, 1991) can be used for efficiently computing posterior expectations $\hat{\alpha}_t = E(\alpha_t|y)$ and variances $Var(\alpha_t|y)$. Because of the equivalence with a corresponding PLS criterion, the Kalman filter and smoother can also be regarded as an algorithmic tool for computing semiparametric PLS estimates, without any need for a Bayesian interpretation, see Fahrmeir and Tutz (1994, ch. 8.1). Using Kalman smoothers for semiparametric additive models like (3) avoids backfitting and provides diagonal bands of smoother matrices as a by–product. However, forcing dynamic models into state space form can result in high–dimensional state vectors with singular multivariate priors, causing unnecessary algorithmic complications.

Up to now it was tacitly assumed that the time series is equally spaced in time. Extensions to non–equally spaced data are possible either by modified difference priors or by continuous time models. For example, a first order random walk $\tau_t - \tau_{t-1} = u_t$, $u_t \sim N(0, \sigma_u^2)$ is generalized to

$$\tau_t - \tau_{t-1} = u_t, \quad u_t \sim N(0, \delta_t \sigma_u^2) \tag{14}$$

where δ_t is the time between observation y_{t-1} and y_t. A second order random walk for non–equally spaced data can be defined by

$$\tau_t - \left(1 + \frac{\delta_t}{\delta_{t-1}}\right) \tau_{t-1} + \frac{\delta_t}{\delta_{t-1}} \tau_{t-2} = u_t, \quad u_t \sim N(0, k_t \sigma_u^2), \tag{15}$$

where k_t is a weight function depending on δ_t and δ_{t-1}.

A simple and straightforward choice is $k_t = \delta_t$ as for first order random walk priors. There are other reasonable, but more complex forms of k_t that are consistent with the equally–spaced case, see Knorr–Held (1996). Corresponding PLS criteria are easily derived from these priors.

For continuous–time models, trend, season and other time–varying parameters are considered as smooth functions of time. With a slight change in notation, the simple trend model (1) becomes

$$y_s = \tau(t_s) + \epsilon_s, \quad s = 1, \ldots, T,$$

with observation times $t_1 < \ldots < t_s < \ldots < t_\tau$, a smooth trend function $\tau(t)$ and i.i.d. errors $\epsilon_s \sim N(0, \sigma^2)$. A continuous time version of the PLS criterion (2) is: Find τ as a twice–differentiable function that minimizes

$$\sum_{s=1}^{T}(y_s - \tau(t_s))^2 + \lambda \int (\tau''(t))^2 dt. \tag{16}$$

The minimizing function $\widehat{\tau}$ is a cubic smoothing spline, see Green and Silverman (1994) for a recent treatment and Eubank (1996, this volume).

Wahba (1978) showed that (16) has a Bayesian justification by placing the solution of the stochastic differential equation

$$\frac{d^2\tau(t_s)}{ds^2} = \lambda^{-1/2}\sigma\frac{dW(s)}{ds}, \quad s \geq t_1 \tag{17}$$

as a smoothness prior over τ. Such a differential equation of order two is the continuous time version of a second order random walk (6). Here $W(s)$ is a standard Wiener process with $W(t_1) = 0$, independent of the errors ϵ_s. For diffuse initial conditions

$$(\tau(t_1), \tau'(t_1)) \sim N(0, kI), \quad k \to \infty$$

the cubic smoothing spline $\widehat{\tau}(s)$ at s coincides with the posterior expectation of $\tau(s)$ given the data, i.e.

$$\widehat{\tau}(s) = E(\tau(s)|y).$$

This equivalence can also be established for more general types of splines where second derivatives are replaced by linear differential operators, see e.g. Kohn and Ansley (1987, 1988). They also derive a discrete–time stochastic difference equation from (17) and use state space techniques for computation of the smoothing spline. Again, pointwise Bayesian confidence bands can be computed as a by–product. For a recent discussion of splines from a Bayesian point of view we refer to van der Linde (1995).

In practice, smoothing parameters λ or hyperparameters like $\sigma^2, \sigma_u^2, \sigma_w^2$ are usually unknown. Within the semiparametric roughness penalty approach, data–driven choice of smoothing parameters is often done by cross-validated optimization of some selection criterion. Already for a small number of smoothing parameters problems may occur because the selection

criterion can be a rather flat function of $\lambda = (\lambda_1, \lambda_2, \ldots)$. Whithin an empirical Bayes approach, hyperparameters in dynamic models are treated as unknown constants. Then the method of maximum likelihood is a natural choice. Maximization can be carried out directly by numerical optimization routines or indirectly via the EM algorithm, see Harvey (1989, ch. 4). If the likelihood is rather flat, then ML estimation also performs poorly. Fully Bayesian approaches can avoid these problems by providing additional information about hyperparameters in form of "hyperpriors". A traditional approach are discrete priors leading to multiprocess Kalman filters (Harrison and Stevens, 1976). More recently Markov chain Monte Carlo (MCMC) techniques have been developed to estimate hyperparameters by simulation from their posteriors (Carlin, Polson and Stoffer, 1992, Carter and Kohn, 1994, Frühwirth–Schnatter, 1994). An advantage of these simulation methods is that their basic concepts are also useful in conditionally non–Gaussian situations as below and in the following sections.

2.2 Conditionally Gaussian models

Gaussian models are not robust against outliers in the observation errors and change points in the trend function or other unobserved components. One way to robustify linear dynamic models is to assume that error distributions are scale mixtures of normals. For given values of the mixture variables the linear dynamic model is then conditionally Gaussian. Mixture variables may be discrete or continuous. A popular choice are χ^2–mixture variables, leading to t–distributions for the errors. A conditionally Gaussian version of the simple trend model (1) with a second order random walk model for the trend is

$$\begin{aligned} y_t &= \tau_t + \epsilon_t, \qquad & \epsilon_t|\omega_{1t} &\sim N(0, \sigma^2/\omega_{1t}) \\ \tau_t - 2\tau_{t-1} + \tau_{t-2} &= u_t, \qquad & u_t|\omega_{2t} &\sim N(0, \sigma_u^2/\omega_{2t}). \end{aligned}$$

Assuming $\omega_{1t}\nu_1$ and $\omega_{2t}\nu_2$ to be independently χ^2–distributed with ν_1 and ν_2 degrees of freedom, then ϵ_t and u_t are independently $t(\nu_1)$ and $t(\nu_2)$ distributed. Although Kalman filters and smoothers are still best linear estimators, they perform poorly for small degrees of freedom ν_1 and ν_2. Various approximate filtering and smoothing algorithms have therefore been

given already in early work on robustified state space modelling (Masreliez, 1975, Masreliez and Martin, 1977, Martin and Raftery, 1987). More recently, fully Bayesian MCMC methods have been developed to tackle this problem. Carlin, Polson and Stoffer (1992) suggest a Gibbs sampling algorithm adding the mixture variables ω_{1t} and ω_{2t} to the set of unknown parameters. Their approach applies to rather general nonnormal dynamic models, but can be inefficient with respect to mixing and convergence properties. Carter and Kohn (1994, 1996) and Shephard (1994) propose a modified Gibbs sampling algorithm, that updates the whole "state vector" $\tau = (\tau_1, \ldots, \tau_T)$ all at once. This modification makes the algorithm much more efficient. The parameters $\tau_1, \ldots, \tau_T$ are often highly correlated, so updating τ_t, $t = 1, \ldots, T$ one at a time, which is done in Carlin, Polson and Stoffer, often results in poor mixing, i.e. the corresponding Markov chain is not moving rapidly throughout the support of the posterior distribution. Consequently, Monte–Carlo standard errors of sample averages will be large.

As an alternative to these fully Bayesian methods one may also consider posterior mode estimation. Let $\rho_1(\epsilon_t)$ and $\rho_2(u_t)$ denote the negative log–densities of the i.i.d. errors ϵ_t and u_t. Taking logarithms and using (conditional) independence assumptions, a robustified version of the PLS criterion (2)

$$\sum_{t=1}^{T} \rho_1(y_t - \tau_t)^2 + \sum_{t=3}^{T} \rho_2(\tau_t - 2\tau_{t-1} + \tau_{t-2})^2 \to \min_{\tau}, \tag{18}$$

can be derived. Computation of the minimizer $\widehat{\tau}$ can be carried out by iterative Kalman–type algorithms, see Künstler (1996). An advantage of posterior mode estimation is that it can also be extended to other ρ–functions, for example Huber functions or $\rho(u) = |u|$. Also, one may start directly from criterion (18), without Bayesian interpretation, to obtain robust semiparametric estimators, and transfer this approach to robust continuous–time spline–type estimation. It should be noted, however, that already for conditionally Gaussian dynamic linear models posterior mean estimates, obtained from a fully Bayesian approach, and posterior mode or spline–type estimators are no longer equivalent. This property holds only for linear Gaussian models with known hyperparameters as in Section (2.1).

3 Non–Gaussian observation models

This section deals with fundamentally non–Gaussian time series and longitudinal data. We progress from simple examples for discrete–valued time series to general non–Gaussian situations.

3.1 Non–Gaussian time series

Figure 1 displays the number y_t of occurrences of rainfall over 1 mm in the Tokyo area for each calendar day during the years 1983–1984. The data, presented in Kitagawa (1987) and reanalyzed later on by several authors, is an example of a discrete–valued time series. Responses y_t, $t = 1, \ldots, 366$, are assumed as binomial:

$$y_t \sim B(n_t, \pi_t) \text{ with } \begin{cases} n_t = 2 \text{ for } t \neq 60 \\ n_t = 1 \text{ for } t = 60 \quad (\text{ February 29}), \end{cases}$$

and π_t the probability of rainfall on calendar day t. To compare it to similar data from other areas or other years, and to see some seasonal pattern, the probabilities $\pi = (\pi_1, \ldots, \pi_T)$, $T = 366$, will be estimated as a smooth curve. For the following we reparametrize π_t by a logit link to τ_t:

$$\tau_t = \log \frac{\pi_t}{1 - \pi_t}, \quad \pi_t(\tau_t) = \frac{\exp(\tau_t)}{1 + \exp(\tau_t)}.$$

A semiparametric discrete–time roughness penalty approach will start from a penalized log–likelihood criterion like

$$PL(\tau) = \sum_{t=1}^{T} \{y_t \log \pi_t(\tau_t) + (n_t - y_t) \log(1 - \pi_t(\tau_t))\} - \lambda \sum_{t=3}^{T} (\tau_t - 2\tau_{t-1} + \tau_{t-2})^2 \to \max_{\tau} \quad (19)$$

to obtain smooth estimates $\widehat{\tau}$ and $\widehat{\pi}$ of the fixed, unknown sequences τ and π. Comparison with the penalized least squares criterion (2) shows that essentially the Gaussian log–likelihood of the observation model (1) is replaced by the sum of binomial log–likelihood contributions. Instead of second order differences, one might also use a sum $\sum(\tau_t - \tau_{t-1})^2$ of squared first order differences.

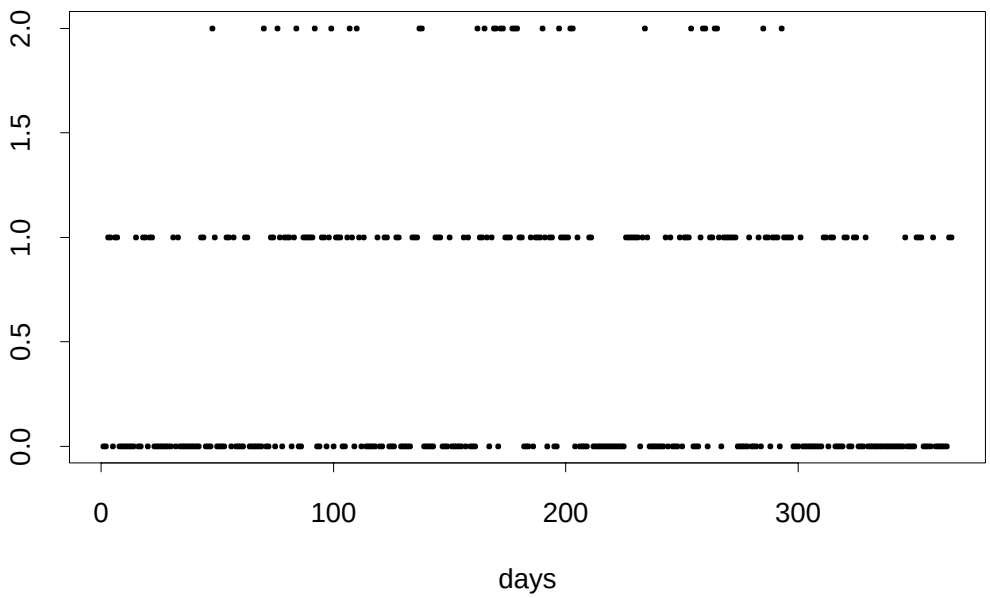

Figure 1: Tokyo rainfall data.

Using the same notation as in Section 2 the continuous–time version of (19) is: Find $\tau = \{\tau(t)\}$ as a twice–differentiable function that solves

$$PL(\tau) = \sum_{s=1}^{T} \{y_s \log \pi(\tau(t_s)) + (n_s - y_s) \log(1 - \pi_t(\tau(t_s)))\} - \lambda \int (\tau''(t))^2 dt \to \max_{\tau}. \quad (20)$$

For a given smoothing parameter λ, the solution is again a cubic smoothing spline, see Hastie and Tibshirani (1990) and Green and Silverman (1994). For equally–spaced data as in the example, the discrete– and continuous–time spline solutions to (19) and (20) are usually in quite close agreement. Algorithmic efficient solutions of the high–dimensional nonlinear optimization problems (19) and (20) are usually carried out by iteratively applying smoothers for penalized least squares estimation to working observations.

For a Bayesian version of the semiparametric approach (19) in form of a non–Gaussian dynamic model, we take

$$y_t|\tau_t \sim B(n_t, \pi_t(\tau_t)), \quad \pi_t(\tau_t) = \frac{\exp(\tau_t)}{1 + \exp(\tau_t)} \quad (21)$$

as the observation model. We supplement it as in (6) by a random walk model of first or

second order

$$\tau_t = \tau_{t-1} + u_t \text{ or } \tau_t = 2\tau_{t-1} + \tau_{t-2} + u_t \tag{22}$$

as a smoothness prior for τ. The errors u_t are i.i.d. $N(0, \sigma_u^2)$ distributed, and initial values τ_1 resp. τ_1, τ_2 are specified as in (8). Variances are assumed to be known. In addition, conditional independence is assumed among all $y_t|\tau$.

In contrast to Gaussian models, the posterior

$$\mathrm{p}(\tau|y) = \frac{\mathrm{p}(y|\tau)\mathrm{p}(\tau)}{\mathrm{p}(y)} \propto \mathrm{p}(y|\tau)\mathrm{p}(\tau) \tag{23}$$

is now non–normal. Thus, posterior expectations and posterior modes are no longer equivalent. With a diffuse prior for initial values, the posterior mode $\widehat{\tau}$ is the maximizer of (19) with smoothing parameter λ equal to $1/2\sigma_\tau^2$. Algorithmic solutions can be efficiently obtained by extended or iterative Kalman filtering and smoothing, see Fahrmeir (1992), Fahrmeir and Tutz (1994) and Fahrmeir and Wagenpfeil (1996). As in the Gaussian case, these techniques may also be viewed as convenient computational tools for computing penalized likelihood estimators, without Bayesian interpretation. For a fully Bayesian analysis, including computation of posterior moments and quantiles, simulation based estimation, in particular MCMC methods, are generally most appropriate. Details are given in Section 3.2.

A continuous–time dynamic model corresponding to (20) is obtained by placing the stochastic differential equation (17) as a smoothness prior over τ. Again, posterior modes are still equivalent to cubic smoothing splines, but different from posterior expectations. Fully Bayesian spline–type smoothing will also be based on MCMC for dynamic models. For this purpose, it is useful to rewrite the continuous–time prior (17) as a stochastic difference equation for the state vector

$$\alpha_t = (\tau(t), d\tau(t)/dt)$$

of the trend τ and its derivative. Following Kohn and Ansley (1987), the sequence $\alpha_s := \alpha(t_s)$ of evaluations at t_s, $s = 1, \ldots, T$ obeys the stochastic difference equation

$$\alpha_{s+1} = F_s\alpha_s + u_s, \quad s = 1, \ldots, T, \tag{24}$$

with transition matrices

$$F_s = \begin{pmatrix} 1 & \delta_{s+1} \\ 0 & 1 \end{pmatrix}, \quad \delta_{s+1} = t_{s+1} - t_s$$

and independent errors

$$u_s \sim N(0, \sigma^2 U_s), \quad U_s = \begin{pmatrix} \delta_{s+1}^3/3 & \delta_{s+1}^2/2 \\ \delta_{s+1}^2/2 & \delta_{s+1} \end{pmatrix}.$$

Together with the observation model

$$y_s|\tau(t_s) \sim B(n_s, \pi(t_s))$$

we obtain a binomial dynamic, or state space, model. Higher order splines can also be written in state space form, see Kohn & Ansley (1987).

As a second example, we consider a time series of counts y_t of the weekly incidence of acute hemorragic conjunctivitis (AHC) in Chiba–prefecture in Japan during 1987. Kashiwagi and Yanagimoto (1992) analyze this data, assuming a loglinear Poisson model

$$y_t|\lambda_t \sim Po(\lambda_t), \quad \lambda_t = \exp(\tau_t)$$

and a first order random walk prior for τ. They obtain a posterior mean estimate based on numerical integrations similar as in Kitagawa (1987). Of course, other smoothness priors as second order random walks or the continuous–time analogue (17) might be used as well. A penalized likelihood approach would start directly from

$$PL(\tau) = \sum_{t=1}^{T} (y_t \log \lambda_t - \lambda_t) - \lambda \sum_{t=2}^{T} (\tau_t - \tau_{t-1})^2 \to \max_{\tau},$$

or with other forms of the roughness penalty term. Again, penalized likelihood estimators are equivalent to posterior mode estimators, but different from corresponding posterior means.

Both examples belong to the class of dynamic generalized linear models. The general *observation model* is as following:

The conditional density of y_t, given the unknown state vector α_t is of the linear exponential family type with conditional expectation

$$E(y_t|\alpha_t) = \mu_t = h(\eta_t)$$

related to the linear predictor $\eta_t = z_t'\alpha_t$ by a suitable link h. As in the Gaussian case the components of α_t may consist of trend τ_t, season γ_t and possibly time–varying effects β_t of covariates x_t and z_t is a suitable design vector. For example an additive predictor

$$\eta_t = \tau_t + \gamma_t + x_t'\beta_t$$

can be written in this form. Although time–constant effects δ can be incorporated formally by setting $\delta_t = \delta_{t-1}$, it is often advantageous to split up the predictor in

$$\eta_t = \tau_t + \gamma_t + x_t'\beta_t + w_t'\delta.$$

For the second stage, smoothness priors $p(\alpha)$ are put on the sequence $\alpha = (\alpha_1, \ldots, \alpha_T)$ in form of a *transition model*. Linear Gaussian transition models like difference equation (6), (7) or the state space form $\alpha_{t+1} = F\alpha_t + u_t$ are often retained as a common choice, but we will also use priors for non–equally spaced observations or continuous times priors.

As for the examples, we can always write down a corresponding semiparametric model and an associated penalized likelihood criterion

$$PL(\alpha) = \sum_{t=1}^{T} l_t(y_t|\alpha_t) - \sum_{j=1}^{p} \lambda_j I(\alpha_j) \to \max_{\alpha}. \tag{25}$$

Here $\alpha_j = (\alpha_{j1}, \ldots, \alpha_{jT})$ is the j–th component of α, $I(\alpha_j)$ a penalty function and λ_j a smoothing parameter. For given smoothing parameters λ_j, estimates α_j are obtained by iterative smoothing, with backfitting in an inner loop, see Hastie and Tibshirani (1990, 1993), Green and Silverman (1994), Fahrmeir and Tutz (1994, ch. 5). As in the examples, (25) can always be derived from the corresponding dynamic model relying on the principle of posterior mode estimation.

Estimation of unknown smoothing parameters λ_j or corresponding hyperparameters σ_j^2 can be based on the same principles as for Gaussian models. Relying on the roughness penalty approach, smoothing parameters are selected as minimizer of a generalized cross–validation criterion, see O'Sullivan, Yandell and Raynor (1986), Wahba, Wang, Gu, Klein and Klein (1995). Empirical Bayes approaches consider hyperparameters σ_j^2 as unknown but fixed and use (approximate) maximum likelihood estimation, for example an EM–type algorithm as

suggested in Fahrmeir (1992). Wagenpfeil (1996) compares some of these approaches. In a fully Bayesian setting, hyperparameters σ_j^2 are considered as random and independent inverse gamma priors

$$\sigma_j^2 \sim IG(a_j, b_j), \quad j = 1, \ldots, p \tag{26}$$

are a common choice for hyperpriors. By appropriate choice of a_j, b_j, these priors can be made more or less informative.

3.2 MCMC in non–Gaussian dynamic models

The design of efficient MCMC algorithms in dynamic models with non–Gaussian observation model is currently an intense research area. For easier presentation, we first discuss several MCMC algorithms for simple non–Gaussian dynamic trend models, like the dynamic binomial or Poisson models in the examples above. Extensions to the general case are outlined at the end of this subsection. Supplementing model (21), (22) with a hyperprior $\mathrm{p}(\sigma_u^2)$ for σ_u^2, for example an inverse gamma prior as in (26), the posterior distribution of the parameters τ and σ_u^2 is given by:

$$\mathrm{p}(\tau, \sigma_u^2|y) \propto \mathrm{p}(y|\tau)\mathrm{p}(\tau|\sigma_u^2)\mathrm{p}(\sigma_u^2). \tag{27}$$

MCMC methods construct Markov chains that converge to a given distribution, here the posterior. Once the chain has reached equilibrium, it provides (dependent) samples from that posterior distribution. Quantities of interest, such as the posterior mean or median, can now be estimated by the appropriate empirical versions.

The well–known Gibbs sampling algorithm (e.g. Gelfand and Smith, 1990) is based on samples from the full conditional distributions of all parameters. In general, a full conditional distribution is proportional to the posterior (27) but often considerable simplifications can be done. To implement the Gibbs sampler in dynamic trend models, we have to sample from

$$\mathrm{p}(\tau_t|\sigma_\tau^2, y) \quad \propto \quad \mathrm{p}(y_t|\tau_t)\mathrm{p}(\tau_t|\tau_{s\neq t}, \sigma_u^2) \tag{28}$$

$$\text{and } \mathrm{p}(\sigma_\tau^2|\tau, y) \quad \propto \quad \mathrm{p}(\tau|\sigma_t^2)\mathrm{p}(\sigma_u^2). \tag{29}$$

If inverse gamma priors are assigned to σ_u^2, (29) is still inverse gamma and samples can be generated easily using standard algorithms.

Suppose we could also easily generate samples from (28), $t = 1, \ldots, T$. The Gibbs sampling algorithm iteratively updates $\tau_1, \ldots, \tau_T$ and σ_u^2 by samples from their full conditionals. Markov chain theory shows, that the so generated sequence of random numbers converges to the posterior (27) for any starting value of the Markov chain. Such an algorithm is proposed in Fahrmeir, Hennevogl and Klemme (1992), following suggestions of Carlin, Polson and Stoffer (1992). However, there are some drawbacks of pure Gibbs sampling in non–Gaussian dynamic models. Firstly, samples from (28), which is non–standard for non–Gaussian observation models, can only be obtained by carefully designed rejection algorithms which may require already a considerable amount of computation time in itself. Fortunately, instead of sampling from the full conditional distribution, a member of the more general class of Hastings algorithms (Hastings, 1970) can be used to update τ_t, $t = 1, \ldots, T$. Here so–called proposals are generated from an arbitrary distribution and a specific accept–reject step is added. Such a Hastings step is typically easier to implement and more efficient in terms of CPU time. A thorough discussion of the Hastings algorithm is given in Tierney (1994) and Besag, Green, Higdon and Mengersen (1995).

For example, to update (28), it is sufficient to generate a proposal τ_t^* from the conditional prior distribution $\mathrm{p}(\tau_t|\tau_{s\neq t}, \sigma_u^2)$ and to accept the proposal as the new state of the Markov chain with probability

$$\delta = \min\left\{1, \frac{\mathrm{p}(y_t|\tau_t^*)}{\mathrm{p}(y_t|\tau_t)}\right\},$$

here τ_t denotes the current state of the chain. The resulting algorithm requires less computation time than pure Gibbs sampling since the conditional prior distribution is Gaussian with known moments so proposals are easy to generate.

However, the generated Markov chain might show signs of slow convergence and does not mix rapidly. That is, the Markov chain is not moving rapidly throughout the support of the posterior distribution so that subsequent samples are highly dependent and Monte Carlo estimates become imprecise. This is a consequence of the underlying single move strategy, i.e. parameters $\tau_t, t = 1, \ldots, T$ are updated one by one. Various attempts have been made to design algorithms that converge fast and mix rapidly. A fruitful idea is the use of blocking; here blocks of parameters, say $\tau_{(a,b)} = (\tau_a, \tau_{a+1}, \ldots, \tau_{b-1}, \tau_b)$, are updated simultaneously

rather than step by step. Such a blocking strategy is a compromise between updating τ all at once, which is infeasible for fundamentally non–Gaussian time series, and updating τ one at a time. The algorithms of Shephard and Pitt (1995) and Knorr–Held (1996) are based on blocking; Knorr–Held generalizes of the conditional prior proposal above to block move algorithms: Generate a proposal $\tau^*_{(a,b)}$ form the conditional prior distribution $\mathrm{p}(\tau_{(a,b)}|\tau_{(1,a-1)}, \tau_{(b+1,T)}, \sigma_u^2)$ and accept the proposal as the new state of the Markov chain with probability

$$\delta = \min \left\{ 1, \frac{\prod_{t=a}^{b} \mathrm{p}(y_t|\tau_t^*)}{\prod_{t=a}^{b} \mathrm{p}(y_t|\tau_t)} \right\}.$$

One of the advantages of MCMC is the possibility to calculate exact posterior distributions of functionals of parameters. For the Tokyo rainfall data, the posterior estimates of the probabilities

$$\pi_t = \frac{\exp(\tau_t)}{1 + \exp(\tau_t)} \tag{30}$$

are of main interest. Instead of plugging an estimate for $\{\tau_t\}$ in (30), we calculate posterior samples from $\{\pi_t\}$, using the original samples from $\mathrm{p}(\tau|y)$. The posterior distributions $\mathrm{p}(\pi|y)$ can now be explored in detail without any approximation. In contrast, posterior mode or splines estimation do not have this feature. Here plug–in estimates, especially confidence bands, are typically biased due to the non–linearity in (30). Similar considerations apply to the AHC example, where $\lambda_t = \exp(\tau_t)$ is to be estimated.

Figure 2 shows the posterior estimates of the probabilities $\{\pi_t\}$ for the Tokyo rainfall data, calculated by a conditional prior block–MCMC algorithm. A highly dispersed but proper inverse gamma hyperprior (26) with $a = 1$, $b = 0.00005$ was assigned to σ_u^2. This prior has a mode at 0.000025. The estimated posterior median was 0.0001. The pattern in Figure 2, with peaks for wet seasons, nicely reflects the climate in Tokyo. It would be difficult to see this by looking only at the raw data (Figure 1). In Fahrmeir and Tutz (1994, ch. 5.3), the probabilities $\{\pi_t\}$ are fitted by a cubic smoothing spline, with the smoothing parameter estimated by generalized cross–validation criterion. This criterion had two local minima at $\lambda = 32$ and $\lambda = 4064$. The smoothing spline for $\lambda = 4064$ is quite close to the posterior median fit in 2, while the smoothing spline for $\lambda = 32$ is much rougher. Such rougher

posterior median estimates are also obtained if the parameter b for the inverse gamma prior is set to higher values. For example, with $a = 1$, $b = 0.005$, the prior mode equals 0.0025. This prior is in favor of larger values for σ_u^2, so that posterior median estimates for $\{\pi_t\}$ become rougher. As a third approach, posterior mode estimation, with an EM–type algorithm for estimating σ_u^2 by maximization of the marginal likelihood, also gives estimates that are in good agreement. These results correspond to empirical evidence experienced in other applications: If smoothing and variance parameters are properly adjusted, posterior mean and medians are often rather close to posterior modes or penalized likelihood estimates. Also, estimation of hyperparameters by cross-validation or marginal likelihood can be helpful for the choice of parameters of the hyperprior in a fully Bayesian model. Similar evidence is provided by the next example.

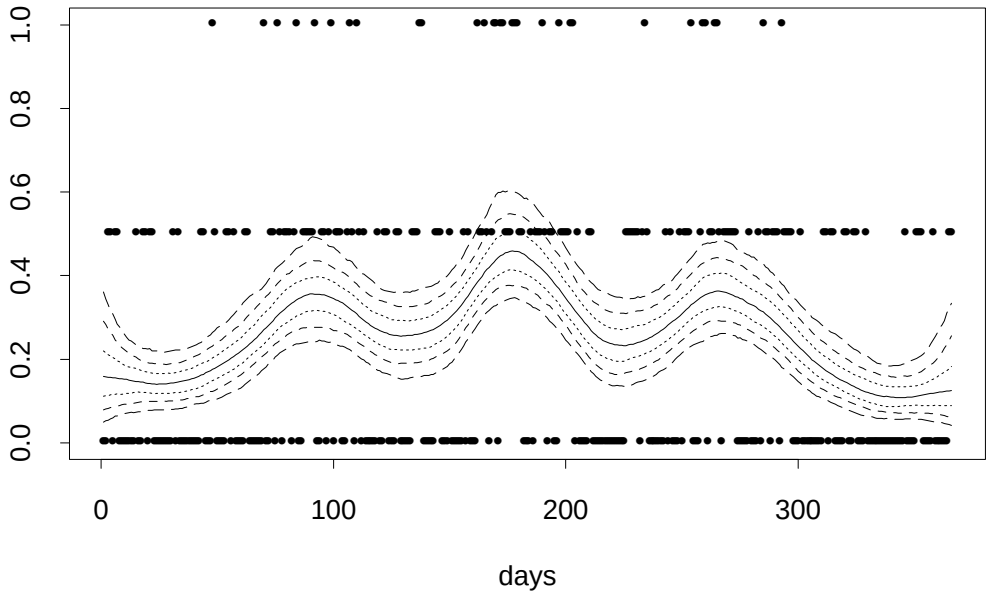

Figure 2: Tokyo rainfall data. Data and fitted probabilities (posterior median within 50, 80 and 95 % credible regions). The data is reproduced as relative frequencies with values 0, 0.5 and 1.

Estimates for the AHC data are shown in Figure 3a) and b) for both first and second order random walk priors. The posterior distribution of the intensities $\{\lambda_t\}$ shows a peak around weak 33 similar to the results of Kashiwagi and Yanagimoto (1992). Compared to the model with second order random walk priors, estimates in Figure 3a) are somewhat rougher and the

peak around week 33 is lower and more flat. This reflects the fact that first order random walk priors are in favor of horizontal, locally straight lines. Figure 3c) shows Bayesian cubic spline–type estimates with thec continuous–time prior (17). As was to be expected with equally spaced observations, these estimates are in very close agreement with those in Figure 3b). Figure 3d) shows displays the cubic smoothing spline, which is the posterior mode estimator from the Bayesian point of view. As with the rainfall data example, it is again quite close to the posterior median in 3c).

In more general dynamic models, response y_t is related to some unknown parameter vector α_t, see for example the state space representation (24) of the spline–type prior (17). MCMC simulation in dynamic models can be performed similarly as for the simple dynamic trend model, where $\alpha_t = \tau_t$ is a scalar, by single– or block–move algorithms. Shephard and Pitt (1995) make specific Fisher scoring type steps to construct a proposal that approximates the full conditional distribution taking the observation into account. In contrast, conditional prior proposals are built independently of the observation and are therefore easier to construct. Their performance is good for situations, where the posterior is not very different from the conditional prior. This is typically the case for discrete valued observations such as bi– or multinomial logistic models as in our examples. Sometimes components α_j of $\alpha = (\alpha_1, \ldots, \alpha_T)$ (compare the notation in (25)) are a priori independent and a componentwise updating strategy with conditional prior proposals can have advantages. Componentwise updating becomes inevitable in problems with multiple time scales or, more general, generalized additive models, see Section 4.

Finally we note, that it is possible to combine robust transition models with non–Gaussian observation models similarly as in Section 2.2. For example, on may use random walk priors with t–distributed errors for trend components, allowing for abrupt large jumps. MCMC simulation in such models is often straightforward, since error terms are still Gaussian, given unknown mixture values. An example is given in Knorr–Held (1996) with $t(\nu)$–distributed errors and an additional hyperprior on the degrees of freedom ν.

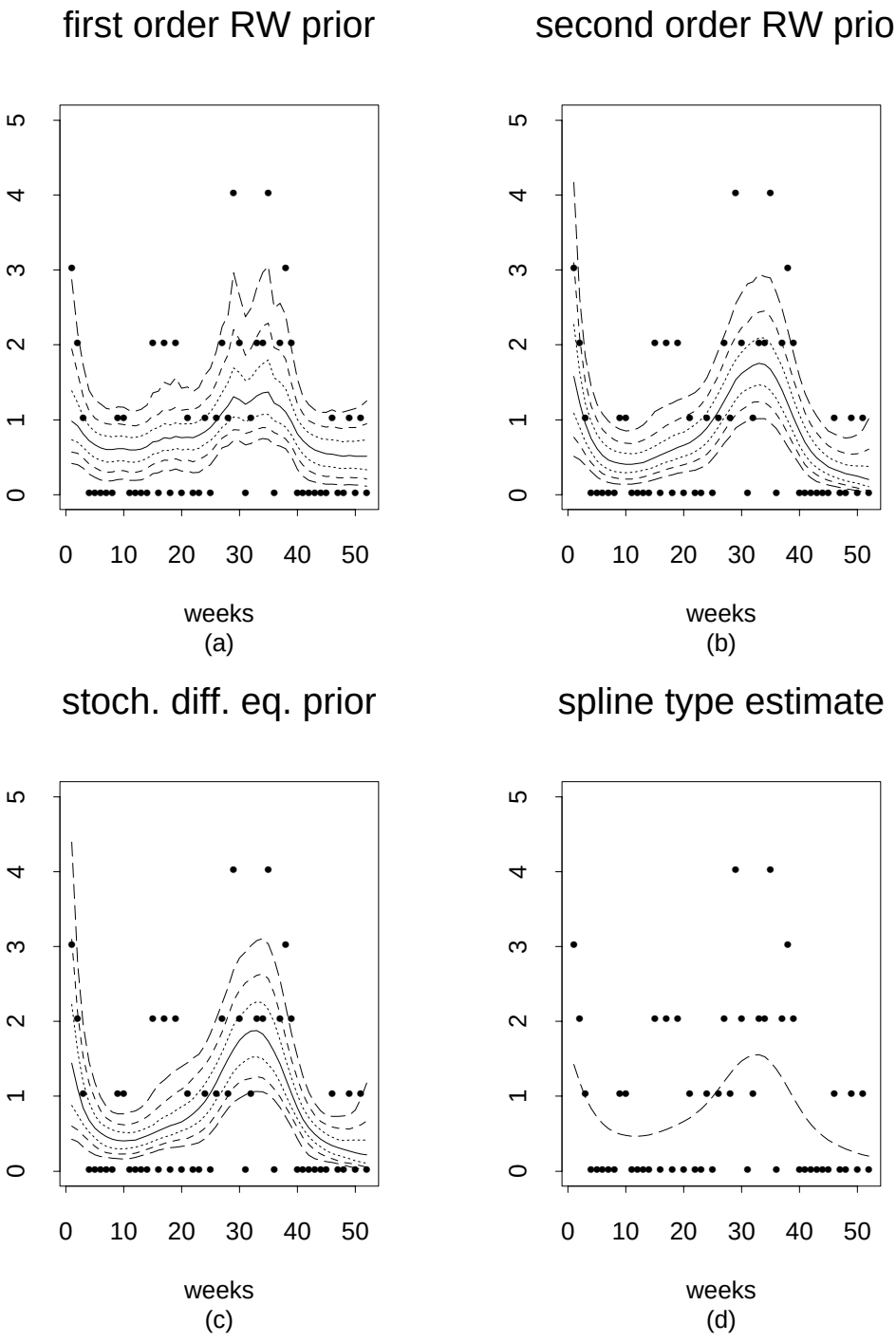

Figure 3: AHC data. Data and fitted probabilities (posterior median within 50, 80 and 95% credible regions).

3.3 Non–Gaussian longitudinal data

In this section, we consider longitudinal data where observations

$$(y_{ti}, x_{ti}), \quad t = 1, \ldots, T, \quad i = 1, \ldots, n,$$

on a response variable y and a vector x of covariates are made for a cross–section of n units at the same time points $t = 1, \ldots, T$. Models for Gaussian outcomes y_{ti} have been treated already extensively, but much less has been done in the non–Gaussian case. As an example, we will consider monthly business test data collected by the IFO institute in Munich for a large cross–section of firms. Answers given in a monthly questionnaire are categorical, most of them trichotomous with categories like "increase" (+), "no change" (=) or "decrease" (−), compare Fahrmeir and Tutz (1994, Examples 6.3, 8.5). Selecting a specific response variable y, say answers on production plans, we obtain categorical longitudinal data.

Observation models for longitudinal data can be defined by appropriate extensions of models for time series data. A straightforward generalization within the exponential family framework is as follows: For given covariates x_{ti} and a possibly time–varying parameter vector α_t, the q–dimensional response y_{ti} comes from a linear exponential density with conditional mean

$$E(y_{ti}|x_{ti}, \alpha_t) = h(\eta_{ti}) \tag{31}$$

and linear predictor

$$\eta_{ti} = Z_{ti}\alpha_t. \tag{32}$$

Here h: $I\!R^q \to I\!R^q$ is a q–dimensional link and the matrix Z_{ti} is a function of the covariates x_{ti} and possibly past responses. Individual responses are assumed to be conditionally independent. A dynamic model for longitudinal data is obtained by supplementing the observation model (31) and (32) with transition models as smoothness priors for α as in Section 3.1. Just as for time series, some subvector $\tilde{\alpha}_t$ of α_t may indeed be time–constant. Such partially dynamic models are formally covered by (32) with the additional restriction $\tilde{\alpha}_t = \tilde{\alpha}_{t-1}$ or by making this explicit and rewriting the predictor in additive form $\eta_{ti} = Z_{ti}\alpha_t + V_{ti}\beta$.

The posterior mode or penalized likelihood approach leads to

$$PL(\alpha) = \sum_{t=1}^{T} \sum_{i=1}^{n} l_{ti}(\alpha_t) - \sum_{j=1}^{p} \lambda_j I(\alpha_j) \rightarrow \max_{\alpha}. \tag{33}$$

Here, $l_{ti}(\alpha_t) = \log f(y_{ti}|x_{ti}, \alpha_t)$ is the conditional likelihood contribution of observation y_{ti}. Computationally efficient solutions can be obtained, for example, by extended or iterative Kalman–type smoothers, see Fahrmeir and Tutz (1994, ch. 8.4) and Wagenpfeil (1996).

Observation models of the form (31), (32) may be appropriate if heterogeneity among units is sufficiently described by observed covariates. This will not always be the case, in particular for larger cross–sections. A natural way to deal with this problem is an additive extension of the linear predictor to

$$\eta_{ti} = Z_{ti}\alpha_t + W_{ti}b_i,$$

where b_i are unit–specific parameters and W_{ti} an appropriate design matrix. A dynamic mixed model is obtained with usual transition models for α and a "random effects" model for the unit–specific parameter. A common assumption is to assume the $b_i's$ are i.i.d. Gaussian,

$$b_i \sim N(0, D), \tag{34}$$

with covariance matrix D. For posterior mode or penalized likelihood estimation of $\alpha = (\alpha_1, \ldots, \alpha_T)$ and $b = (b_1, \ldots, b_n)$, a further penalty term

$$I(b) = \sum_{i=1}^{n} b_i' D b_i,$$

corresponding to the Gaussian prior (34) is added to (33). An algorithmic solution for the resulting joint posterior mode or penalized likelihood estimates $(\widehat{\alpha}, \widehat{b})$ is worked out in Biller (1993), also in combination with an EM–type algorithm for estimation of smoothing parameters. However, computation times become large for multicategorical responses. Moreover, serious bias may occur, see Breslow and Clayton (1993), Breslow and Lin (1995).

MCMC techniques are more attractive for dynamic mixed models through their model flexibility. The additional parameters $b_1, \ldots, b_n$ are added to the set of unknown parameters and are updated with some well designed proposals, for example with Metropolis random walk

proposals, in every MCMC cycle. Besides, a hyperprior for D has to be introduced. The usual choice is the inverted Wishart distribution

$$\mathrm{p}(D) \propto |D|^{-\zeta-(m+1)/2} \exp(-\mathrm{tr}(BD^{-1}))$$

with parameters $\zeta > (m-1)/2$ and $|B| > 0$; here m is the dimension of b_i. A Gibbs step can then be used to update D.

Turning to the IFO business test example, we investigate the dependency of current production plans on demand and orders in hand in the specific branch "Vorprodukte Steine und Erden". We have complete longitudinal observations of 51 firms for the period from 1980 to 1994. Our model allows for time–changing effects of covariates and for trend and seasonal variation of threshold parameters, which represent corresponding probabilities of the response categories. Additional unit–specific parameters b_i are introduced to allow for firm–specific differences of these probabilities.

The response variable "production plans" is given in three ordered categories: "increase" $(+)$, "no change" $(=)$ and "decrease" $(-)$. Its conditional distribution is assumed to depend on the covariates "orders in hand", "expected business conditions" as well as on the production plans of the previous month. All these covariates are trichotomous. We used a dummy coding approach for comparison with previous analyses with the category $(-)$ as reference category. The corresponding dummies are denoted by A^+, $A^=$ (orders in hand), G^+, $G^=$ (expected business conditions) and P^+, $P^=$ (production plans of the previous month) and define the covariate vector x_{ti}. The inclusion of P as a covariate reduces the panel length by 1 to $T = 179$ (February 1980 to December 1994).

A cumulative logistic model (e.g. Fahrmeir & Tutz, 1994a, ch. 3) was used due to the ordinal nature of the response variable: Let $\tilde{y}_{ti} = 1$ and $\tilde{y}_{ti} = 2$ denote the response categories "increase" and "no change" respectively. Then

$$P(\tilde{y}_{ti} \leq j) = F(\theta_{tij} + x'_{ti}\beta_t), \quad , j = 1, 2,$$

is assumed with $x_{ti} = (G^+, G^=, P^+, P^=, A^+, A^=)'$ and $F(x) = 1/(1+\exp(-x))$.

We decompose both threshold parameters θ_{ti1} and θ_{ti2} into trend parameters τ_t, seasonal

parameters γ_t and unit specific parameters b_i, one for each threshold:

$$\theta_{tij} = \tau_{tj} + \gamma_{tj} + b_{ij}, \quad j = 1, 2.$$

Note that the threshold parameters have to follow the restriction $\theta_{ti1} < \theta_{ti2}$ for all combinations of t and i. A seasonal model (7) with period $m = 12$ was chosen for the seasonal parameters of both thresholds. First order random walk priors are assigned to all covariate effect parameters β_t and to both trend parameters θ_{t1}, θ_{t2}. All time–changing parameters are assumed to be mutually independent with proper but highly dispersed inverse gamma hyperpriors (a=1, b=0.005). The firm–specific parameters $b_i = (b_{i1}, b_{i2})'$ are assumed to follow a Gaussian distribution with mean zero and dispersion D. We used the parameter values $\zeta = 1$ and $B = \text{diag}(0.005, 0.005)$ for the inverted Wishart hyperprior specification for D.

This model can be written as a dynamic mixed model with

$$\pi_{ti} = h(\eta_{ti}) = h(Z_{ti}\alpha_t + W_{ti}b_i),$$

where $\alpha_t' = (\tau_{t1}, \gamma_{t1}, \tau_{t2}, \gamma_{t2}, \beta_t')$,

$$W_{ti} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and

$$Z_{ti} = \begin{pmatrix} 1 & 1 & 0 & 0 & x_{ti}' \\ 0 & 0 & 1 & 1 & x_{ti}' \end{pmatrix}.$$

The responses variable y_{ti} is multinomially distributed

$$y_{ti} \sim M_2(1, \pi_{ti})$$

where $y_{ti} = (1,0)'$, $(0,1)'$ or $(0,0)'$, if the first $(+)$, second $(=)$ or third $(-)$ category is observed. The link function h is given by

$$h(\eta_{ti}) = \begin{pmatrix} F(\eta_{ti1}) \\ F(\eta_{ti2}) - F(\eta_{ti1}) \end{pmatrix}.$$

Figure 4 displays the temporal pattern of the trend parameters τ_{tj}, $j = 1, 2$, and of both threshold parameters $\theta_{tj} = \tau_{tj} + \gamma_{tj}$, $j = 1, 2$. The first trend parameter is slightly decreasing

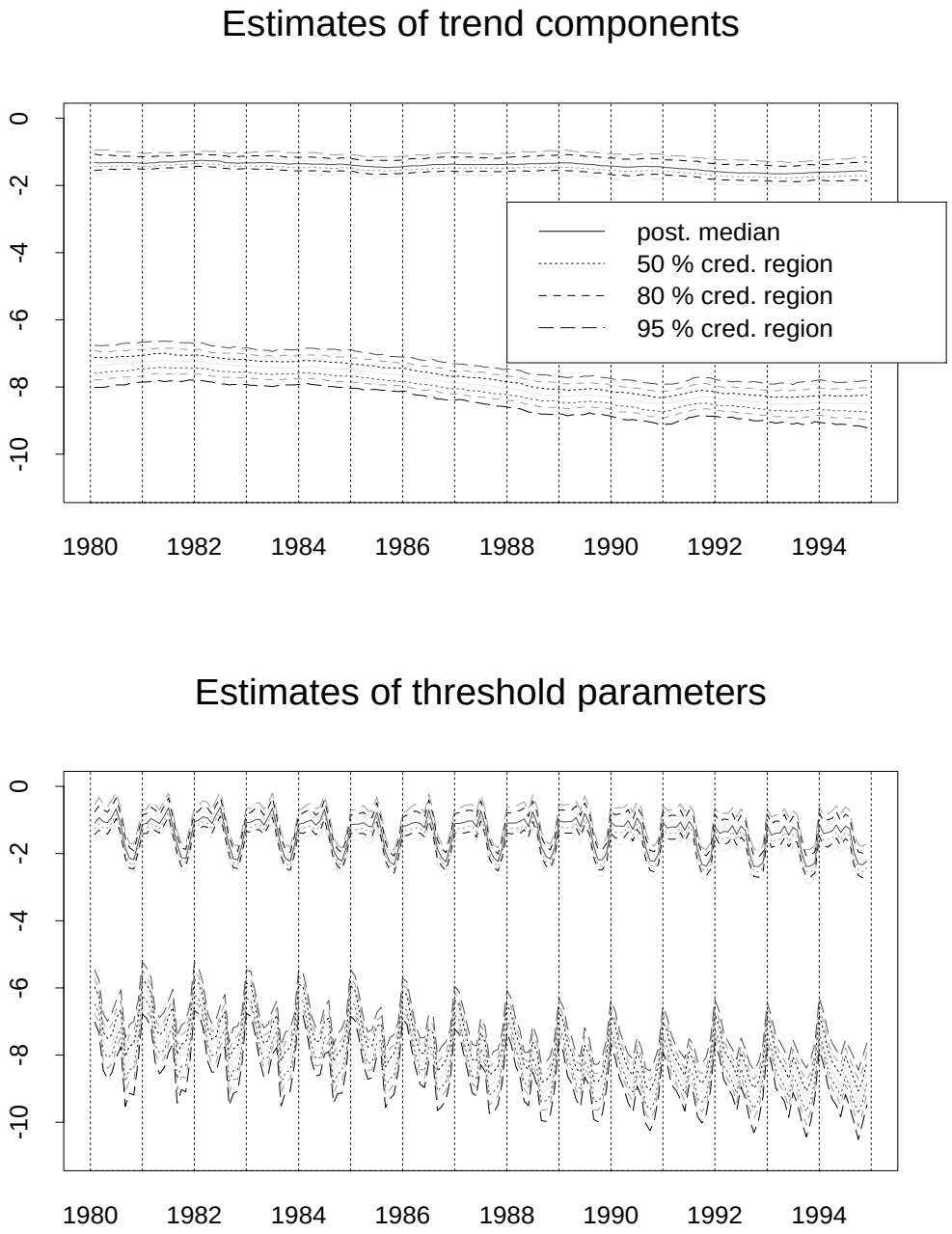

Figure 4: Estimates of trends and thresholds. Dashed vertical lines represent the month January of each year.

while the second remains constant over the whole period. A distinct seasonal pattern can be seen with higher probabilities of positive response in spring and negative response in fall. However, firm–specific deviations from this pattern are substantial as can be seen from Figure 5. Here, posterior median estimates of the first and second firm–specific parameter b_{i1} and b_{i2} are plotted against each other for all 51 firms. Interestingly, these two parameters

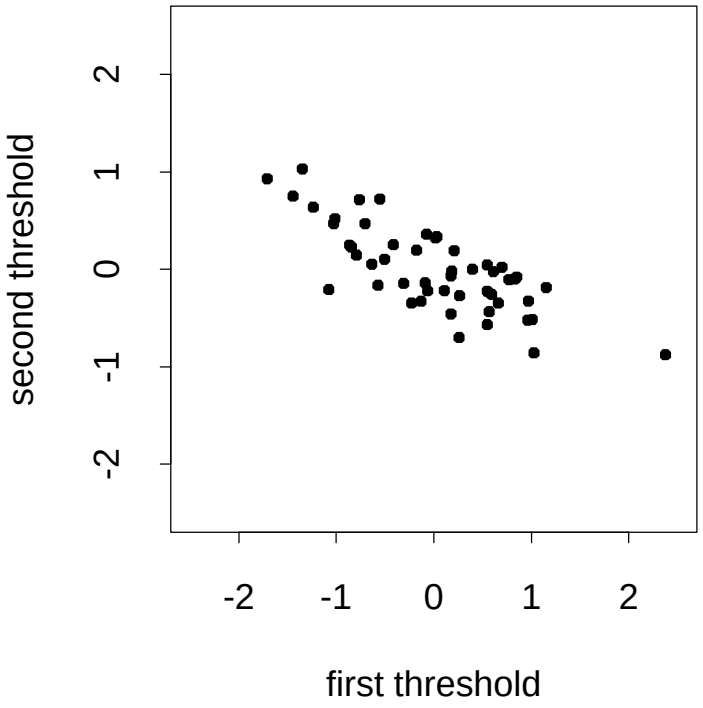

Figure 5: Plot of the estimates of b_{i1} against b_{i2} for each unit.

are often highly negatively correlated. The estimated dispersion matrix of the random effect distribution is

$$\widehat{D} = \begin{pmatrix} 0.78 & -0.28 \\ -0.28 & 0.23 \end{pmatrix},$$

and the estimated correlation, based on posterior samples of the corresponding functional of

D, is -0.67. Both estimates are posterior median estimates. It seems that some firms are more conservative in their answers and often choose "no change" for the response variable, relative to the overall frequencies. Such firms have negative values for b_{i1} and positive values for b_{i2}. Other firms avoid the category "no change" and answer often more extremely with "decrease" or "increase". For these firms b_{i1} is positive and b_{i2} negative.

The estimated patterns of time–dependent covariate effects (Figure 6) show an interesting temporal pattern, in particular the effect of the dummy $G+$ (Figure 7), which stands for expected improved business conditions, relative to $G-$: A distinct low can be seen end at the of 1991, when the German economy was shaken by a recession. In 1982 a new government under the leadership of chancellor Helmut Kohl was established. From that time onwards the effect increases until 1989/1990 with some additional variation and can be interpreted as a growing trust in the government.

The peak in 1989/1990 coincidences with the German reunification, which was expected to have a catalytic effect on the economy due to the sudden opening of the market in former east Germany. In the years 1986, 90 and 94, parliament elections were held in fall. In these years the effect is always decreasing towards the end of the year, which may be due to the uncertainty regarding the election results.

4 Generalized additive and varying coefficient models

Let us now turn to a cross–sectional regression situation where observations $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$ on a response y and a vector $(x_1, \ldots, x_p)$ of covariates are given. Generalized additive models (Hastie and Tibshirani, 1990) assume that, given $x_i = (x_{i1}, \ldots, x_{ip})$, the distribution of y_i belongs to an exponential family with mean $\mu_i = E(y_i|x_i)$ linked to an additive predictor η_i by

$$\mu_i = h(\eta_i), \quad \eta_i = f_1(x_{i1}) + \ldots + f_p(x_{ip}). \tag{35}$$

Here $f_1, \ldots, f_p$ are unknown, smooth functions of continuous covariates $x_1, \ldots, x_p$. If some covariates are assumed to have a linear effect on the predictor, then semiparametric or

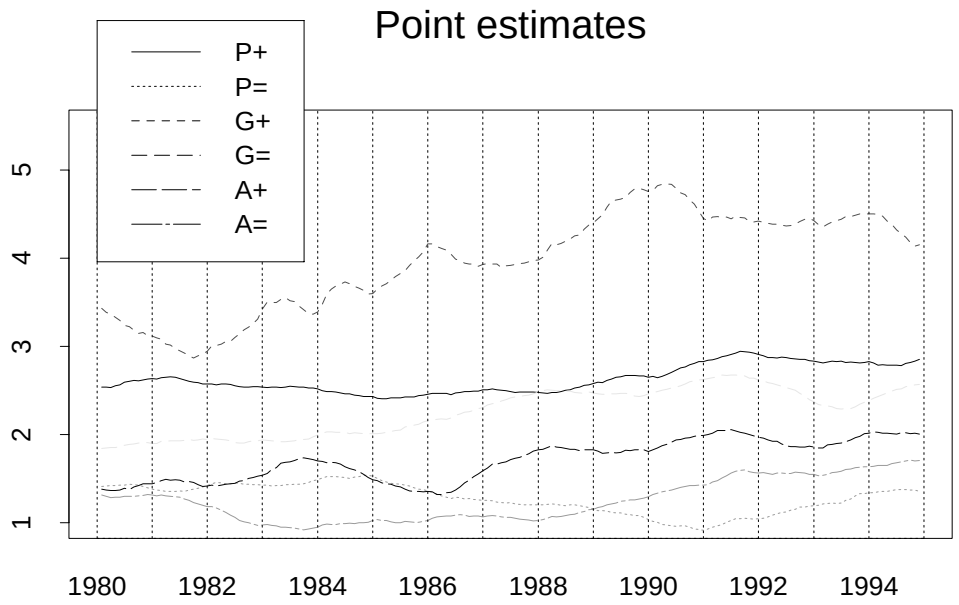

Figure 6: Estimated time–changing covariate effects. Dashed vertical lines represent the month January of each year.

generalized partial linear models like

$$\eta_i = f_1(x_{i1}) + \ldots + \beta_p x_{ip} \tag{36}$$

are appropriate modifications of (35). In (36), x_{ip} could also be a binary or categorical covariate. In the following, we will focus on model (35).

Nonparametric estimation of the functions $f_1, \ldots, f_p$ can be based on the penalized likelihood criterion

$$PL(f_1, \ldots, f_p) = \sum_{i=1}^{n} l_i(y_i|\eta_i) - \sum_{j=1}^{p} \lambda_j \int (f_j''(u))^2 du \rightarrow \max_{f_1,\ldots,f_p} \tag{37}$$

with individual likelihood contributions l_i from $y_i|x_i$. The maximizing functions are cubic smoothing splines $\widehat{f}_1, \ldots, \widehat{f}_p$. Other types of penalty terms may also be used, replacing for example second derivatives by m–th order derivatives $f_j^{(m)}(u)$ or using discretized penalty terms. Computation is usually carried out by backfitting algorithms, see Hastie and Tibshirani (1990) or Fahrmeir and Tutz (1994, ch. 5). As a drawback, construction of confidence bands relies on conjectures of approximate normality of penalized likelihood estimators,

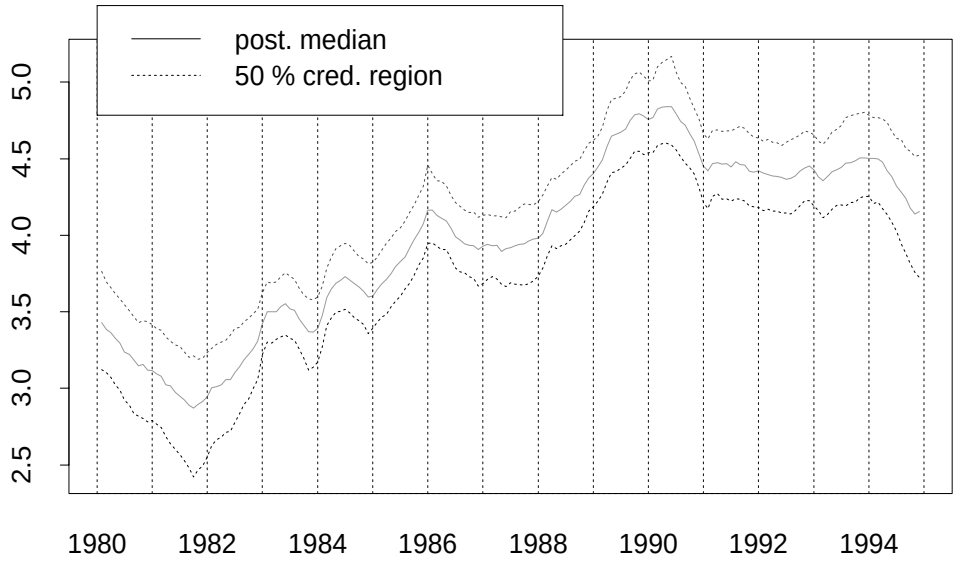

Figure 7: Estimated time–changing covariate effect of G^+. Dashed vertical lines represent the month January of each year.

with a rigorous proof still missing. Also, data–driven choice of smoothing parameters can be problematic.

Bayesian inference in generalized additive models, as outlined in the sequel, uses ideas from dynamic models for time series data. Basically, time is replaced by metrical covariates, with different covariates x_j, $j = 1, \ldots, p$, corresponding to different time scales t_j.

For each covariate, let

$$t_{j1} < \ldots < t_{js} < \ldots < t_{jT_j}, \quad T_j \leq n$$

denote the strictly ordered, different values of observations $x_{ij}, i = 1, \ldots, n$. Bayesian smoothness priors for the unknown values

$$f(t_{j1}) < \ldots < f(t_{js}) < \ldots < f(t_{jT_j})$$

can now be defined by adapting random walk models (5), (12) for non–equally spaced time–points or continuous–time priors like (17) to the present situation. Setting $\alpha_{js} = f_j(t_{js})$, and

$\delta_{js} = t_{js} - t_{j,s-1}$, first and second order random walk priors are given by

$$\alpha_{js} - \alpha_{j,s-1} = u_{js}, \quad u_{js} \sim N(0, \delta_{js}\sigma_j^2)$$

and

$$\alpha_{js} - \left(1 + \frac{\delta_{js}}{\delta_{j,s-1}}\right)\alpha_{j,s-1} + \frac{\delta_{js}}{\delta_{j,s-1}}\alpha_{j,s-2} = u_{js}, \quad u_{js} \sim N(0, k_{js}\sigma_j^2), \tag{38}$$

with mutually independent errors u_{js}. Priors for Bayesian cubic spline–type smoothing, corresponding to the penalized log–likelihood (37) are given by the stochastic differential equations

$$\frac{d^2 f_j(s)}{ds^2} = \sigma_j \frac{dW_j(s)}{ds}, \quad s \geq t_{j1} \tag{39}$$

with mutually independent standard Wiener processes, $W_j(t_{j1}) = 0$, and diffuse initial conditions for

$$\alpha_{js} = (f_j(s), f_j'(s)).$$

In complete analogy to Section 3.1, the priors (39) can be written in state space form (24) and some hyperpriors are assigned to σ_j.

The likelihood $\mathrm{p}(y|\alpha, \sigma)$, the priors $\mathrm{p}(\alpha)$, $\mathrm{p}(\sigma)$ and as a consequence, the posterior $\mathrm{p}(\alpha|y)$ have the same structure as in Section (3.2). Therefore single– or block–move schemes as outlined there can be used to simulate from the posterior. Details and some generalizations are given in Lang (1996) for random walk priors and Biller and Fahrmeir (1997) for stochastic differential equation priors.

As an application, we consider the credit–scoring problem described in Fahrmeir & Tutz (1996, ch. 2.1). 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 is to model or predict the probability that a client with certain covariates ("risk factors") is to be considered as a potential risk. The data set consists of 1000 consumers'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. As in Fahrmeir and Tutz, we will use a subset of these data, containing only the following covariates, which are

partly metrical and partly categorical:

x_1	running account, trichotomous with categories "no running account" (= 1), "good running account" (= 2), "medium running account" ("less than 200 DM" = 3 = reference category)
x_3	duration of credit in months, metrical
x_4	amount of credit in DM, metrical
x_5	payment of previous credits, dichotomous with categories "good", "bad" (=reference category)
x_6	intended use, dichotomous with categories "private" or "professional" (=reference category)
x_8	marital status, with reference category "living alone".

A parametric logit model for the probability $\mathrm{P}(y = 1|x)$ of being not creditworthy leads to the somewhat surprising conclusion that the covariate "amount of credit" has no significant influence on the risk. Here, we reanalyze the data with a partial linear logit model

$$\log \frac{\mathrm{P}(y = 1|x)}{1 - \mathrm{P}(y = 1|x)} = \beta_0 + \beta_1 x_1^1 + \beta_2 x_1^2 + f_3(x_3) + f_4(x_4) + \beta_5 x_5 + \beta_6 x_6 + \beta_8 x_8.$$

Here x_1^1 and x_1^2 are dummies for the categories "good" and "medium" running account, respectively. The predictor has semiparametric or partial linear form: The smooth functions $f_3(x_3)$, $f_4(x_4)$ of the metrical covariates "duration of credit" and "amount of credit", are estimated by usual cubic splines and by Bayesian spline–type smoothing using second order random walk models (38) for non–equally spaced observations. The constant β_0 and the effects $\beta_1, \ldots, \beta_8$ of the remaining categorical covariates are considered as fixed for penalized likelihood estimation and estimated jointly with the curves f_3 and f_4. For Bayesian estimation, diffuse priors are chosen for $\beta_0, \beta_1, \beta_2, \beta_5, \beta_6, \beta_8$, and additional M–H–steps with random walk proposals are included for MCMC simulation.

Figure 8 shows the estimates for the curves f_3 and f_4. Again, the posterior mean of the spline–type smoother and the posterior mode or penalized likelihood estimator (full line) are not far away from each other. While the effect of the variable "duration of credit" is not

too far away from linearity, the effect of "amount of credit" is clearly nonlinear. The curve has bathtub shape and indicates that not only high credits but also low credits increase the risk, compared to "medium" credits between 3000–6000 DM. Apparently, if the influence is misspecified by assuming a linear function $\beta_4 x_4$ instead of $f_4(x_4)$, the estimated effect $\widehat{\beta}_4$ will be near zero, corresponding to an almost horizontal line $\widehat{\beta}_4 x_4$ near zero, and falsely considered as nonsignificant. Table 1 gives the posterior means together with 80% credible intervals. and maximum likelihood estimates of the remaining effects. Both estimates are in good agreement.

	posterior mean	80% CI		ML estimator
x_1^1	0.662	–0.004	1.335	0.633
x_1^2	–1.468	–2.243	–0.733	–1.324
x_5	–1.085	–2.051	–0.135	–0.998
x_6	–0.442	–1.035	0.209	–0.440
x_8	–0.578	–1.180	0.016	–0.516

Table 1: Estimates of constant parameters in the credit–scoring data.

Finally we note, that the whole approach can be extended to varying coefficient models (Hastie and Tibshirani, 1993), where the predictor has the form

$$\eta_i = f_1(x_{i1})z_{i1} + \ldots + f_p(x_{ip})z_{ip}, \tag{40}$$

with $z_{i1}, \ldots, z_{ip}$ as further "factors". For the special case $z_{i1} = \ldots = z_{ip} = 1$, (40) reduces to a generalized additive model (35). If $z_1, \ldots, z_p$ are further covariates, possibly including some components of x, then a term $f_j(x_{ij})z_{ij}$ can be interpreted as an interaction term between x and z, or $f_j(x_{ij})$ can be considered as an effect of z, varying over the "effect–modifier" $f_j(x_j)$. For $x_j \equiv t$, i.e. if covariate x_j is time t, $f_j(t)$ is a time–varying effect, and for $x_1 = \ldots = x_p = t$ the linear predictor has the same form as in dynamic models for time series or longitudinal data.

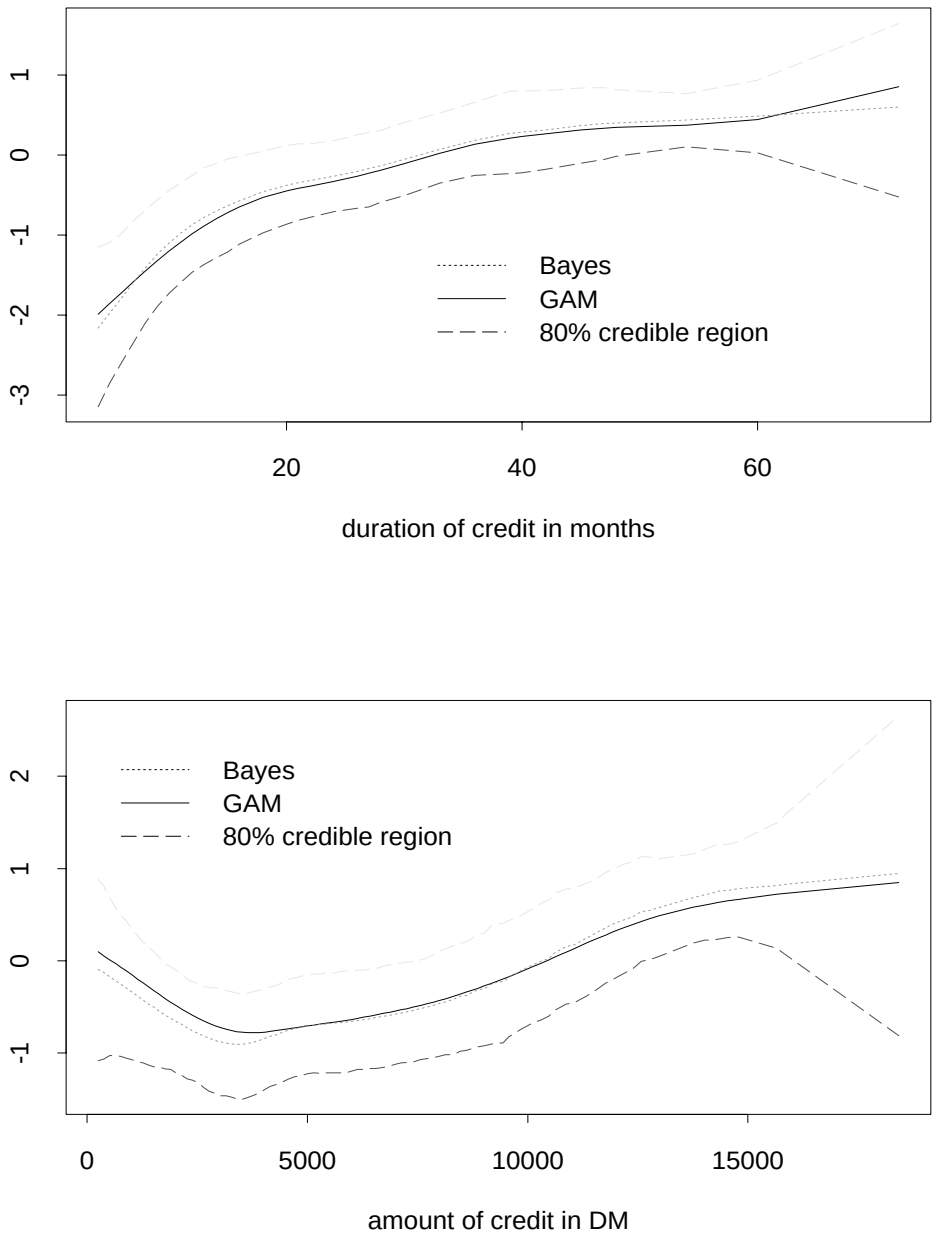

Figure 8: Estimated functions of the covariate "duration of credit" and "amount of credit".

5 Conclusions

In this chapter we showed that dynamic models with Bayesian smoothness priors and semiparametric models based on the roughness penalty approach provide supplementary ways for nonparametric function estimation. Semiparametric Bayesian smoothing has some attractive features: It provides a natural framework for Bayesian analysis beyond posterior mode or MAP estimation and recent advances in MCMC techniques allow to estimate posterior means, medians, quantiles and other functionals of regression functions or other parameters. No approximation, based on conjectures of asymptotic normality, have to be made. Bayesian data–driven choice of smoothing parameters is automatically incorporated in the model. Due to the hierarchical model formulation and modular estimation techniques, the Bayesian approach offers much flexibility in modifying or extending methods to other situations, for example to dynamic mixed models for longitudinal data (Section 3.3), to generalized additive and varying coefficient models (Section 4) or to data with missing values, an issue not treated here.

To some extent, of course, one has to pay for these advantages: MCMC techniques produce a rich output, but computation times can also be quite high. Metropolis–Hastings algorithms provide a wide variety of possibilities for updating steps, but convergence and mixing of the so constructed Markov chain has also to be checked empirically. Careful convergence diagnostics deserve much attention in particular applications. Above all, the choice of reasonable priors on the unknown functions remains subjective and may not be easily accepted.

Semiparametric models based on the roughness penalty approach are useful supplementary tools for data analysis: Roughness penalties corresponding to smoothness priors can be interpreted without any underlying Bayesian framework. Thus, if the roughness penalty looks reasonable it supports the choice of the smoothness prior. As we have shown, the penalized likelihood estimator can always be interpreted as a corresponding posterior mode estimator from a Bayesian point of view. Computation is done by numerically efficient solutions of a nonlinear maximization problem, relying on commonly accepted and well–understood optimization routines. As we demonstrated by examples, the posterior mode is

often quite near to posterior means or medians and, therefore, can be quite useful to check convergence of MCMC simulations.

We focused on non–Gaussian models for times series, longitudinal and regression data within the set up of generalized linear models with a prespecified link functions of known parametric form, as for example the logistic or the exponential functions. Tis restriction could be relaxed by defining a generalized parametric family of link functions (as for example in Stukel, 1988, Czado, 1992) and estimating unknown parameters in the link function jointly with unknowns in the predictor. A non–parametric Bayesian approach, avoiding any parametric specification of a link function, has been proposed by Arjas and Gasbarra (1994) and Arjas and Liu (1996) in the related context of hazard regression. Generally, we believe that in situations with many covariates flexible non– or semiparametric modelling and exploration of the predictor is more important compared to nonparametric choice of the link function while retaining linear parametric predictors. For Gaussian models, Bayesian analysis of regression splines with adaptive knot selection has been recently proposed by Smith and Kohn (1994), Smith, Wong and Kohn (1996) and Denison, Mallick and Smith (1996). It would be interesting to adapt these methods for non–Gaussian regression models.

Extensions to other data structures are possible by choosing other observation models and smoothness assumptions. In particular, event history analysis and spatial statistics are a wide and promising field of research, e.g. Fahrmeir & Knorr–Held (1997), Arjas and Liu (1996), Arjas and Heikkinen (1996) and Besag, York and Mollie (1991). Also, problems of model diagnostics and model choice have to dealt with convincingly. Here again, Bayesian and non–Bayesian data analyses could complement one another in a productive way.

References

[1] Arjas, E. & Gasbarra, D. (1994). Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statistica Sinica*, **4**, 505–524.

[2] Arjas, E. & Heikkinen, J. (1996). An algorithm for nonparametric Bayesian estimation

of a Poisson intensity. *Computional Statistics*, to appear.

[3] Arjas, E. & Liu, L. (1996). A Non–parametric Bayesian Approach to Hazard Regression: A Case Study with a Large Number of Covariate Values. *Statistics in Medicine*, **15**, 1757–1770.

[4] Besag, J. E., Green, P. J., Higdon, D. & Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science*, **10**, 3–66.

[5] Besag, J. E., York, J. & Mollie, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.

[6] Biller, C. (1993). Modelle mit zufälligen Effekten für kategoriale Längschnittdaten. Diplomarbeit, Universität München.

[7] Biller, C. & Fahrmeir, L. (1997). Bayesian Spline–type Smoothing in Generalized Regression Models, to appear in Computational Statistics.

[8] Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

[9] Breslow, N. E. & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, **82**, 81–91.

[10] Carlin, B. P., Polson, N. G. & Stoffer, D. S. (1992). A Monte Carlo approach to nonnormal and nonlinear state-space-modeling. *Journal of the American Statistical Association*, **87**, 493–500.

[11] Carter, C. K. & Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, **81**, 541–553.

[12] Carter, C. K. & Kohn, R. (1996). Robust Bayesian nonparametric regression. In: *Statistical Theory and Computational Aspects of Smoothing* (W. Härdle & M. G. Schimek, eds.). Heidelberg: Physica–Verlag, 128–148.

[13] Czado, C. (1992). On Link Selection in Generalized Linear Models. In: *Advances in GLIM and Statistical Modelling* (L. Fahrmeir, B. Francis, R. Gilchrist & G. Tutz, eds.). Heidelberg: Springer, 60–65.

[14] de Jong, P. (1991). The diffuse Kalman filter. *The Annals of Statistics*, **19**, 1073–1083.

[15] Denison, D. G. T., Mallick, K. B. & Smith, A. F. M. (1996). Automatic Bayesian curve fitting. Unpublished Manuscript.

[16] 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.

[17] Fahrmeir, L., Hennevogl, W. & Klemme, K. (1992). Smoothing in Dynamic Generalized Linear Models by Gibbs Sampling. In: *Advances in GLIM and Statistical Modelling* (L. Fahrmeir, B. Francis, R. Gilchrist & G. Tutz, eds.). Heidelberg: Springer, 85–90.

[18] Fahrmeir, L. & Knorr–Held, L. (1997). Dynamic discrete time duration models, *Sociological Methodology 1997*, to appear.

[19] Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer–Verlag.

[20] Fahrmeir, L. & Wagenpfeil, S. (1996). Penalized Likelihood Estimation and Iterative Kalman Smoothing for Non–Gaussian Dynamic Regression Models. *Computational Statistics and Data Analysis*, to appear.

[21] Frühwirth–Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, **15**, 183–202.

[22] Gelfand, A. E. & Smith, A. F. M. (1990). Sampling–Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85**, 398-409.

[23] Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* London: Chapman & Hall.

[24] Harrison, P. J. & Stevens, C. F. (1976). Bayesian Forecasting (with Discussion). *Journal of the Royal Statistical Society* B, **38**, 205–247.

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

[26] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

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

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

[29] Kashiwagi, & Yanagimoto (1992). Smoothing Serial Count Data Through a State–Space Model. *Biometrics*, **48**, 1187–1194.

[30] Kitagawa, G. (1987). Non–Gaussian State–Space Modelling of Non–stationary Time Series (with discussion). *Journal of the American Statistical Association*, **82**, 1032–1063.

[31] Knorr–Held, L. (1996). Hierarchical Modelling of Discrete Longitudinal Data: Applications of Markov Chain Monte Carlo. Ph.D. dissertation, Universität München.

[32] Künstler, R. (1996). Robuste Zustandsraummodelle. Ph.D. dissertation, Universität München.

[33] Kohn, R. & Ansley, C. (1987). A New Algorithm for Spline Smoothing Based on Smoothing a Stochastic Process. *SIAM Journal Scientific Statistical Computing*, **8**, 33–48.

[34] Kohn, R. & Ansley, C. (1988). Equivalence between Bayesian smoothing priors and optimal smoothing for function estimation. In *Bayesian Analysis of Time Series and Dynamic Models* (J. C. Spall, eds.). New York: Dekker, 393–430.

[35] Lang, S. (1996). Bayesianische Inferenz in Modellen mit variierenden Koeffizienten. Diplomarbeit, Universität München.

[36] Martin, R. D. & Raftery (1987). Robustness, Computation, and Non–Euclidean Models (Comment). *Journal of the American Statistical Association*, **82**, 1044–1050.

[37] Masreliez, C. J. (1975). Approximate Non–Gaussian Filtering with Linear State and Observation Relations. *IEEE Transactions on Automatic Control*, **AC–20**, 107–110.

[38] Masreliez, C. J. & Martin, R. D. (1977). Robust Bayesian Estimation for the Linear Model and Robustifying the Kalman Filter. *IEEE Transactions on Automatic Control*, **AC–22**, 361–371.

[39] Shephard, N. (1994). Partial non–Gaussian state space. *Biometrika*, **81**, 115–131.

[40] Shephard, N. & Pitt, M. K. (1995). Parameter–driven exponential family models. Unpublished manuscript, Nuffield College, Oxford, UK.

[41] Smith, M. & Kohn, R. (1994). Nonparametric Regression using Bayesian Variable Selection. *Journal of Econometrics*, to appear.

[42] Smith, M., Wong, C.–M. & Kohn, R. (1996). Additive Nonparametric Regression for Time Series. Preprint, Australian Graduate School of Management.

[43] Stukel, T. A. (1988). Generalized Logistic Models. *Journal of the American Statistical Association*, **83**, 426–431.

[44] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics*, **22**, 1701–1762.

[45] O'Sullivan, F., Yandell, B. & Raynor, W. (1986). Automatic Smoothing of Regression Functions in Generalized Linear Models. *Journal of the American Statistical Association*, **81**, 96–103.

[46] van der Linde, A. (1995). Splines from a Bayesian point of view. *Test*, **4**, 63–81.

[47] van der Linde, A. (1996).

[48] Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regressions. *Journal of the Royal Statistical Society* B, **40**, 364–372.

[49] Wahba, G. Wang, Y., Gu, C., Klein, R. & Klein, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. *The Annals of Statistics*, **23**, 1865–1896.

[50] Wagenpfeil, S. (1996). Dynamische Modelle zur Ereignisanalyse. Ph.D. dissertation, Universität München.

[51] West, M. & Harrison, P. J. (1989). *Bayesian Forecasting and Dynamic Models*. New York: Springer–Verlag.

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