Fahrmeir, Wagenpfeil:

Penalized likelihood estimation and iterative kalman smoothing for non-gaussian dynamic regression models

Sonderforschungsbereich 386, Paper 5 (1995)

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

Projektpartner

PENALIZED LIKELIHOOD ESTIMATION AND ITERATIVE KALMAN SMOOTHING FOR NON-GAUSSIAN DYNAMIC REGRESSION MODELS

BY LUDWIG FAHRMEIR AND STEFAN WAGENPFEIL

Ludwig Maximilians Universität, München

Address for correspondence:

Prof. Dr. Ludwig Fahrmeir
Institute of Statistics
Ludwigstr. 33 / II
80539 Munich, Germany

Tel.: +89 / 2180 - 2220
Fax: +89 / 2180 - 3804
Email: ua311aa@sunmail.lrz-muenchen.de

Part of this work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 386 „Statistische Analyse diskreter Strukturen”.

Abstract. Dynamic regression or state space models provide a flexible framework for analyzing non-Gaussian time series and longitudinal data, covering for example models for discrete longitudinal observations. As for non-Gaussian random coefficient models, a direct Bayesian approach leads to numerical integration problems, often intractable for more complicated data sets. Recent Markov chain Monte Carlo methods avoid this by repeated sampling from approximative posterior distributions, but there are still open questions about sampling schemes and convergence. In this article we consider simpler methods of inference based on posterior modes or, equivalently, maximum penalized likelihood estimation. From the latter point of view, the approach can also be interpreted as a nonparametric method for smoothing time-varying coefficients. Efficient smoothing algorithms are obtained by iteration of common linear Kalman filtering and smoothing, in the same way as estimation in generalized linear models with fixed effects can be performed by iteratively weighted least squares estimation. The algorithm can be combined with an EM-type method or cross-validation to estimate unknown hyper- or smoothing parameters. The approach is illustrated by applications to a binary time series and a multicategorical longitudinal data set.

Keywords. Discrete observations; hyperparameter estimation; non-Gaussian longitudinal data; smoothing; state space models; time-varying coefficients.

1. INTRODUCTION

Dynamic regression or state space models relate time series observations $\{y_t\}$ to a sequence of unknown states or parameters $\{\alpha_t\}$, typically including a trend component and time-varying coefficients of covariates. Given the observations $y_1,...,y_T$, estimation (filtering and smoothing) of the unknown sequence $\{\alpha_t\}$ is of primary interest. For Gaussian linear state space models, the relationship is

given by $y_t = Z_t\alpha_t + \varepsilon_t$, where Z_t is an observation or design matrix of appropriate dimension. It is supplemented by a linear transition equation $\alpha_t = F_t\alpha_{t-1} + \xi_t$ and the usual assumptions on the Gaussian noise processes. Due to linearity and normality, the posterior distribution of α_t given $y_1,...,y_T$ is also normal, $\alpha_t|y_1,...,y_T \sim N(a_{t|T}, V_{t|T})$, and the linear Kalman filter and smoother provides posterior means $a_{t|T}$, together with posterior covariances $V_{t|T}$, as optimal estimates for α_t given $y_1,...,y_T$ in a computationally efficient way.

For non-Gaussian time series or longitudinal data, the linear observation model has to be replaced by an appropriate non-Gaussian model. A broad class of generalized dynamic regression or exponential family state space models is obtained if the observation model for $y_t|\alpha_t$ is in the form of a generalized linear model with predictor $\eta_t = Z_t\alpha_t$. An important class of non-exponential family models are robust models with heavy-tailed error distributions that are resistant against additive outliers. Closed form updating formulas similar to linear Kalman filtering in the linear Gaussian model are only available for special models with appropriate conjugate prior-posterior distributions.

In this article, the Gaussian linear transition equation for $\{\alpha_t\}$ is retained, allowing simultaneous modelling and estimation of stochastic trends, seasonal components and time-varying covariate effects. This corresponds to the common assumption of Gaussian random effects in generalized linear mixed models as e.g. in Breslow and Clayton (1993).

Direct Bayesian approaches involve irreducibly high-dimensional integrations, which are generally intractable for more complicated problems. Recent Markov chain Monte Carlo methods avoid this by drawing repeated samples from approximative posterior distributions (e.g. Carlin, Polson and Stoffer, 1992; Carter and Kohn, 1994). However, there are still problems concerning choice of computationally efficient sampling schemes and convergence of the sampling to equilibrium.

Therefore, simpler approximative methods are still useful as an alternative, or supplement for exploratory data analysis, or to provide initial solutions for other methods as e.g. in Schnatter (1992), Frühwirth-Schnatter (1994). In this paper, as in Breslow and Clayton (1993) for generalized linear mixed models or in Fahrmeir and Kaufmann (1991), Fahrmeir (1992) for dynamic generalized linear models, estimation is based on posterior modes or, equivalently, maximum penalized likelihood estimation (Green, 1987). From the latter point of view, the approach can also be interpreted nonparametrically: Dropping the Bayesian smoothness prior imposed on $\{\alpha_t\}$ by the transition model and starting directly from the penalized likelihood criterion, the method yields an efficient procedure for discrete spline smoothing of time-varying coefficients (compare Hastie and Tibshirani, 1993). We show that maximum penalized likelihood smoothing estimates can be obtained by iterative application of linear Kalman filtering and smoothing to a working model, similarly as Fisher scoring in static generalized linear models can be performed by iteratively weighted least squares applied to working observations. This is a rather convenient result, since it allows to use any computationally efficient and available version of linear Kalman filters and smoothers in the iteration steps. For exponential family models, a related algorithm, though derived by different arguments, is contained in Durbin and Koopman (1992). Advantages of iterative Kalman filtering and smoothing, in comparison with common nonparametric procedures, are: It avoids an additional inner backfitting loop, directly provides error covariance matrices as elements of the blockdiagonal of the smoother matrix, and therefore can be combined with an EM-type algorithm or with cross-validation to estimate unknown hyper- or smoothing parameters.

The paper is organized as follows: Dynamic exponential family models are dealt with in Section 2, including specific models which are used in the simulations and illustrative applications of Section 4. Penalized likelihood

estimation by iterative Kalman smoothing is developed in Section 3. Extensions to general non-Gaussian dynamic regression models are given in Section 5.

2. EXPONENTIAL FAMILY STATE SPACE MODELS

We first consider the case of time series observations $\{y_t\}$. An extension to longitudinal data $\{y_{it}\}$ for a population of units i=1,...,n is given at the end of the section.

In the sequel responses y_t and states α_t have dimension q respectively p. Let us rewrite the Gaussian linear observation equation as

$$y_t|\alpha_t \sim N(\eta_t = Z_t\alpha_t, R_t) \ , \qquad (2.1)$$

where $R_t = var(y_t|\alpha_t)$ is the covariance matrix of y_t given α_t. The obvious modification to non-Gaussian exponential family observations is to specify the *observation model* for y_t given α_t by a q-dimensional distribution of the natural exponential family type:

$$y_t|\alpha_t \sim p(y_t|\alpha_t) = c_t(y_t)\exp\{\theta_t' y_t - b_t(\theta_t)\}, \qquad (2.2)$$

where θ_t , the natural parameter, is a function of $\eta_t = Z_t\alpha_t$, and $c_t(\cdot)$ and $b_t(\cdot)$ are known functions. For simplicity we assume that no unknown nuisance parameter is present. By the properties of exponential families the mean and variance functions are then

$$E(y_t|\alpha_t) = \mu_t(\alpha_t) = \partial b_t(\theta_t)/\partial\theta_t \ , \qquad (2.3)$$

$$var(y_t|\alpha_t) = \Sigma_t(\alpha_t) = \partial^2 b_t(\theta_t)/\partial\theta_t\partial\theta_t' . \qquad (2.4)$$

As in static generalized linear models the mean $\mu_t(\alpha_t)$ is related to the linear predictor $\eta_t = Z_t\alpha_t$ by

$$\mu_t(\alpha_t) = h(Z_t\alpha_t) \ , \qquad (2.5)$$

where h: $\mathbb{R}^q \to \mathbb{R}^q$ is a two-times continuously differentiable response function and Z_t is a qxp-matrix, which may depend on covariates x_t or also on past

responses y_s (s = 1,...,t-1). In the latter case densities, means etc. are to be understood conditionally upon past responses.

The exponential family assumption (2.2) together with the mean specification (2.5) replaces the observation equation (2.1) in linear Gaussian models. It is supplemented by a *state transition model*. We retain the assumption of a Gaussian linear transition equation

$$\alpha_t = F_t\alpha_{t-1} + \xi_t \quad (t = 1,\ldots,T) \tag{2.6}$$

with transition matrix F_t, Gaussian white noise $\{\xi_t\}$ with $\xi_t \sim N(0,Q_t)$, and initial state $\alpha_0 \sim N(a_0,Q_0)$.

To specify the models completely in terms of densities, the following conditional independence assumptions are added:

(A1) Conditional on α_t, current responses y_t are independent of past states $\alpha_{t-1},\ldots,\alpha_0$, i.e.

$$p(y_t|\alpha_t,\alpha_{t-1},\ldots,\alpha_o) = p(y_t|\alpha_t) \quad (t = 1,\ldots,T) .$$

Assumption (A1) is implied in Gaussian linear state space models by the assumption of mutual independence of the error sequences $\{\varepsilon_t\}$ and $\{\xi_t\}$. If the design matrix Z_t contains past responses or if covariates are stochastic, (A1) also has to be understood conditionally.

(A2) The sequence $\{\alpha_t\}$ is Markovian, i.e.

$$p(\alpha_t|\alpha_{t-1},\ldots,\alpha_0) = p(\alpha_t|\alpha_{t-1}) .$$

According to (2.6) we have $p(\alpha_t|\alpha_{t-1}) \sim N(F_t\alpha_{t-1},Q_t)$.

For scalar $(q = 1)$ responses, univariate dynamic generalized linear models are obtained. For counts loglinear Poisson models are a standard choice:

$$y_t|\alpha_t \sim Po(\lambda_t), \quad \lambda_t = \exp(\eta_t). \tag{2.7}$$

The linear predictor may be chosen as in simple structural time series models for Gaussian observations:

$$\eta_t = \tau_t + \gamma_t + x_t'\beta_t \quad (t = 1,\ldots,T),$$

where the states are unobserved stochastic trend and seasonal components τ_t, γ_t, and possibly time-varying effects β_t of covariates x_t. Simple nonstationary models for trend or time-varying effects are first or second order random walk models, e.g.

$$\tau_t = \tau_{t-1} + u_t \quad \text{resp.} \quad \tau_t = 2\tau_{t-1} - \tau_{t-2} + u_t, \quad u_t \sim N(0, \sigma_u^2) \ .$$

By appropriate definition of α_t, Z_t and F_t these models can be put into state space form. A seasonal component with period s can be modelled by

$$\sum_{j=0}^{s-1} \gamma_{t-j} = \omega_t, \quad \omega_t \sim N(0, \sigma_\omega^2),$$

see also Section 4.

Of course, a loglinear Poisson model will not always be appropriate, and other choices such as a negative binomial may also be considered. If the number of counts at time t is limited by n_t, say, binomial regression models, such as logit or probit models, are often appropriate:

$$y_t | \alpha_t \sim B(n_t, \pi_t), \quad \pi_t = h(\eta_t), \tag{2.8}$$

where $h(\cdot)$ is a response function, linking π_t to the predictor $\eta_t = Z_t \alpha_t$. For $h(\cdot) = \exp(\cdot)/\{1+\exp(\cdot)\}$ one obtains the logit model, for $h(\cdot) = \Phi(\cdot)$ the probit model. For $n_t = 1$, this is the most common way of modelling binary time series.

Extensions to time series of multicategorical or multinomial responses proceed along similar lines: If k is the number of categories, responses y_t can be described by a vector $y_t = (y_{t1}, ..., y_{tq})'$, with q=k-1 components. If only one multicategorical observation is made for each t, then $y_{tj} = 1$ if category j has been observed, and $y_{tj} = 0$ otherwise (j=1,...,q). Corresponding categorical response models are completely determined by response probabilities $\pi_t = (\pi_{t1}, ..., \pi_{tq})'$, with $\pi_{tj} = P(y_{tj} = 1)$ (j=1,...,q). If there are n_t independent repeated responses at t, then $y_t = (y_{t1}, ..., y_{tq})'$ is multinomial with parameters n_t, π_t, and y_{tj} is the absolute frequency of observations in category j.

For example, a dynamic multivariate logistic model with trend and covariates is specified by

$$n_{tj} = \frac{\exp(\eta_{tj})}{1 + \sum_{r=1}^{q} \exp(\eta_{tr})}, \quad \eta_{tj} = \tau_{tj} + x_t' \beta_{tj} \quad (j = 1, \ldots, q),$$

together with a transition model for trend and covariate components.

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

$$\pi_{tj} = G(\eta_{tj}) - G(\eta_{t,j-1}) \qquad (j = 1, \ldots, q), \tag{2.9}$$

with linear predictors

$$\eta_{tj} = \tau_{tj} - x_t' \beta_t,$$

ordered threshold parameters $-\infty = \tau_{t0} < \ldots < \tau_{tq} < \infty$, a global covariate effect β_t, and a known distribution function G, e.g. the logistic one.

Dynamic cumulative models can be written in state space form along the previous lines. In the simplest case threshold and covariate effects obey a first-order random walk or are partly constant in time. Then

$$\alpha_t = (\tau_{t1}, \ldots, \tau_{tq}, \beta_t')' = \alpha_{t-1} + \xi_t,$$

$$Z_t = \begin{bmatrix} 1 & & -x_t' \\ & \ddots & \vdots \\ & 1 & -x_t' \end{bmatrix},$$

and the response function can be appropriately defined. Dynamic versions of other models for ordered categories (see e.g. Fahrmeir and Tutz, 1994, ch.3) can be designed with analogous reasoning.

The models can be extended to longitudinal data, where time series $\{y_{it}\}$ are observed for each unit i (i = 1,...,n) of a population of size n, if we specify individual observation models $\eta_{it} = Z_{it}\alpha_t$, $\mu_{it} = h(\eta_{it})$ of the form (2.5). Design matrices Z_{it} are constructed as before and may be appropriate functions of covariates x_{it}. The states α_t have now to be interpreted as population parameters. As common for longitudinal data, we assume that individual responses y_{it} at time t are conditionally independent, given α_t, covariates and past responses. Collecting individual responses into the observation vector y_t =

(y_{it}, i = 1,...,n) at time t, the models above can be easily extended to the „time series“ $\{y_t\} = \{y_{it}, i = 1,...,n\}$.

3. POSTERIOR MODE SMOOTHING AND PENALIZED LIKELIHOOD ESTIMATION

3.1. *Fisher scoring by iterative Kalman smoothing*

In this subsection we derive smoothing algorithms, assuming that hyperparameters, e.g. Q_t , are known. Estimation of hyperparameters is dealt with in Subsection 3.2. For the ease of presentation, we first consider time series data $\{y_t\}$ and suppose that covariates are deterministic. Furthermore we denote histories of responses and states up to t by

$$y_t^* = (y_1', \ldots, y_t')', \quad \alpha_t^* = (\alpha_0', \ldots, \alpha_t')'$$

and set $\alpha = \alpha_T^*$. Then the posterior mode smoother $a \equiv (a_{0|T}', a_{1|T}', \ldots, a_{T|T}')' \in \mathbb{R}^m$ with m=(T+1)p is defined as

$$a \equiv (a_{0|T}', a_{1|T}', \ldots, a_{T|T}')' := \operatorname*{argmax}_{\alpha} \left\{ p\left(\alpha \middle| y_T^*\right) \right\} ,$$

i.e. as the mode of the posterior distribution of the entire sequence. The aim is to maximize $p\left(\alpha \middle| y_T^*\right)$. Repeated application of Bayes' theorem yields

$$p\left(\alpha \middle| y_T^*\right) = \frac{1}{p\left(y_T^*\right)} \left\{ \prod_{t=1}^{T} p\left(y_t \middle| \alpha_t^*, y_{t-1}^*\right) \prod_{t=1}^{T} p\left(\alpha_t \middle| \alpha_{t-1}^*, y_{t-1}^*\right) \cdot p(\alpha_0) \right\} .$$

With (A1), (A2), and as $p\left(y_T^*\right)$ does not depend on α,

$$p\left(\alpha \middle| y_T^*\right) \propto \prod_{t=1}^{T} p\left(y_t \middle| \alpha_t, y_{t-1}^*\right) \prod_{t=1}^{T} p\left(\alpha_t \middle| \alpha_{t-1}\right) \cdot p(\alpha_0) .$$

Taking logarithms and inserting the Gaussian densities of the transition model (2.6), we obtain the penalized log-likelihood function PL: $\mathbb{R}^m \to \mathbb{R}$

$$\begin{aligned} PL(\alpha) := & \sum_{t=1}^{T} \ln p\left(y_t \middle| \alpha_t, y_{t-1}^*\right) - \frac{1}{2} (\alpha_0 - a_0)' Q_0^{-1} (\alpha_0 - a_0) \\ & - \frac{1}{2} \sum_{t=1}^{T} (\alpha_t - F_t \alpha_{t-1})' Q_t^{-1} (\alpha_t - F_t \alpha_{t-1}) , \end{aligned} \qquad (3.1)$$

where the densities $p\left(y_t \middle| \alpha_t, y_{t-1}^*\right)$ are defined by the exponential family observation model.

Thus

$$a \equiv (a_{0|T}', a_{1|T}', \ldots, a_{T|T}')' = \arg\max_{\alpha} \{PL(\alpha)\}, \qquad (3.2)$$

i.e. maximizing $p\left(\alpha \middle| y_T^*\right)$ is equivalent to maximize the penalized log-likelihood (3.1) with respect to α. We may however also interprete (3.1), (3.2) without reference to the Bayesian smoothness prior defined by the transition model (2.6) for $\{\alpha_t\}$. From a nonparametric point of view, we may consider $\{\alpha_t\}$ as a fixed, but unknown sequence of states or parameters. Then the first term in (3.1) measures goodness of the fit obtained by the linear predictor $Z_t\alpha_t$. The second term penalizes roughness of the fit or, equivalently, smoothness of the sequence $\{\alpha_t\}$. This is in complete analogy to spline smoothing in generalized additive models (GAM), cf. Hastie and Tibshirani (1990), and most easily seen from a simple example, e.g. a binary dynamic logit model $\log\{\pi_t/(1-\pi_t)\} = \tau_t + \beta_t x_t$. If the trend τ_t and the time-varying effect β_t are assumed to obey first order random walks, then (3.1) becomes with $\alpha_t = (\tau_t, \beta_t)'$

$$PL(\alpha) = \sum_{t=1}^{T} \{y_t \log \pi_t + (1-y_t)\log(1-\pi_t)\} - \frac{1}{2\sigma_\tau^2}\sum_{t=1}^{T}(\tau_t - \tau_{t-1})^2 - \frac{1}{2\sigma_\beta^2}\sum_{t=1}^{T}(\beta_t - \beta_{t-1})^2,$$

neglecting the priors of τ_0 , β_0 for simplicity. While the first term measures goodness of fit in terms of the deviance, the other terms penalize roughness in trends $\{\tau_t\}$ and time-varying effects $\{\beta_t\}$. Compared to spline smoothing, we are therefore smoothing trends, seasonal components and covariate effects instead of covariate functions. The variances σ_τ^2, σ_β^2, or more general, the components of Q_t, play the role of smoothness parameters. This relationship is also pointed out in Hastie and Tibshirani (1993). For a linear Gaussian observation model $y_t = Z_t\alpha_t + \varepsilon_t$, the log-likelihood term in (3.1) specializes to

$$\sum_{t=1}^{T} \ln p\left(y_t \middle| \alpha_t, y_{t-1}^*\right) = -\frac{1}{2}\sum_{t=1}^{T}(y_t - Z_t\alpha_t)'R_t^{-1}(y_t - Z_t\alpha_t),$$

so that (3.1) becomes a penalized least squares criterion, and the nonlinear maximization problem (3.2) reduces to a quadratic programming problem. Since

posterior modes and means coincide for linear Gaussian state space models, the optimization problem is solved by common linear Kalman filters and smoothers. They exploit the special dynamic structure of the penalized least squares criterion very efficiently, resulting in recursive algorithms of complexity O(T).

To find a solution of (3.2) in the general case, i.e. the exponential family observation model, any nonlinear optimization code could be used in principle. For statistical purposes, Gauss-Newton or Fisher scoring is of advantage, just as for static GLM's. However, as in the case of linear Gaussian models, algorithms should take into account the special dynamic structure of the penalized log-likelihood criterion. In the following, we derive a single Fisher scoring step in analogy to static generalized linear models and show that it can be performed by applying linear Kalman filtering and smoothing to "working" observations, thus resulting in an algorithmic solution of complexity O(T). Let us first rewrite (3.1) in compact matrix notation as

$$\mathrm{PL}(\alpha) = l(\alpha) - \frac{1}{2}\alpha' \mathrm{K}\alpha, \qquad (3.3)$$

where

$$l(\alpha) = \sum_{t=0}^{T} l_t(\alpha_t), \quad l_t(\alpha_t) := \ln\, \mathrm{p}\left(\mathrm{y}_t \middle| \alpha_t, \mathrm{y}_{t-1}^*\right) \qquad (\mathrm{t} = 1,...,\mathrm{T}),$$

$l_0(\alpha_0) := -(\alpha_0 - \mathrm{a}_0)'\mathrm{Q}_0^{-1}(\alpha_0 - \mathrm{a}_0)/2$, and the penalty matrix K is symmetric and block-tridiagonal, with blocks easily obtained from (3.1):

$$\mathrm{K} = \begin{bmatrix} \mathrm{K}_{00} & \mathrm{K}_{01} & & & 0 \\ \mathrm{K}_{10} & \mathrm{K}_{11} & \mathrm{K}_{12} & & \\ & \mathrm{K}_{21} & \ddots & \ddots & \\ & & \ddots & \ddots & \mathrm{K}_{\mathrm{T-1,T}} \\ 0 & & & \mathrm{K}_{\mathrm{T,T-1}} & \mathrm{K}_{\mathrm{TT}} \end{bmatrix}$$

with

$$\begin{aligned} \mathrm{K}_{t-1,t} &= \mathrm{K}'_{t,t-1} \qquad (\mathrm{t}=1,\ldots,\mathrm{T}), \\ \mathrm{K}_{00} &= \mathrm{F}_1'\mathrm{Q}_1^{-1}\mathrm{F}_1, \\ \mathrm{K}_{tt} &= \mathrm{Q}_t^{-1} + \mathrm{F}'_{t+1}\mathrm{Q}_{t+1}^{-1}\mathrm{F}_{t+1} \quad (\mathrm{t}=1,\ldots,\mathrm{T}), \\ \mathrm{F}_{T+1} &= 0, \\ \mathrm{K}_{t-1,t} &= -\mathrm{F}_t'\mathrm{Q}_t^{-1} \quad (\mathrm{t}=1,\ldots,\mathrm{T}). \end{aligned}$$

To describe a Fisher scoring step in matrix notation, it is convenient to introduce the vector of observations

$$y' = (a_0', y_1', \ldots, y_T')$$

augmented by a_0 . Correspondingly we define the vector of expectations augmented by α_0 ,

$$\mu(\alpha)' = \{\alpha_0', \mu_1'(\alpha_1), \ldots, \mu_T'(\alpha_T)\},$$

where $\mu_t(\alpha_t) = h(Z_t\alpha_t)$, the block-diagonal covariance matrix

$$\Sigma(\alpha) = \text{diag}\{Q_0, \Sigma_1(\alpha_1), \ldots, \Sigma_T(\alpha_T)\},$$

the block-diagonal design matrix

$$Z = \text{diag}(I, Z_1, \ldots, Z_T),$$

with $I \in \mathbb{R}^{p,p}$ as the unit matrix 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, using properties (2.4), (2.5), the score function of $l(\alpha)$ in (3.3) is

$$s(\alpha) = \{s_0'(\alpha_0), s_1'(\alpha_1), \ldots, s_T'(\alpha_T)\}' := Z'D(\alpha)\Sigma^{-1}(\alpha)\{y - \mu(\alpha)\}, \qquad (3.4)$$

with components $s_0(\alpha_0) = Q_0^{-1}(a_0 - \alpha_0)$, $s_t(\alpha_t) = Z_t'D_t(\alpha_t)\Sigma_t^{-1}(\alpha_t)\{y_t - \mu_t(\alpha_t)\}$ (t = 1,...,T). The weight matrix

$$W(\alpha) = \text{diag}\{W_0, W_1(\alpha_1), \ldots, W_T(\alpha_T)\} := D(\alpha)\Sigma^{-1}(\alpha)D'(\alpha) \qquad (3.5)$$

with diagonal blocks $W_0 = Q_0^{-1}$, $W_t(\alpha_t) = D_t(\alpha_t)\Sigma_t^{-1}(\alpha_t)D_t'(\alpha_t)$ (t = 1,...,T), and the (expected) information matrix of $l(\alpha)$

$$S(\alpha) = \text{diag}\{S_0, S_1(\alpha_1), \ldots, S_T(\alpha_T)\} := Z'W(\alpha)Z \qquad (3.6)$$

with diagonal blocks $S_0 = Q_0^{-1}$, $S_t(\alpha_t) = Z_t'W_t(\alpha_t)Z_t$ (t = 1,...,T), are block-diagonal. The first derivative of $PL(\alpha)$ in (3.3) is

$$u(\alpha) = \partial PL(\alpha)/\partial\alpha = s(\alpha) - K\alpha,$$

and the block-tridiagonal expected information matrix is given by

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

A single Fisher scoring step from the current iterate $\alpha^0 \in \mathbb{R}^m$, say, to the next iterate $\alpha^1 \in \mathbb{R}^m$ is then

$$U(\alpha^0)\{\alpha^1 - \alpha^0\} = u(\alpha^0).$$

This can be rewritten as

$$\alpha^1 = \{Z'W(\alpha^0)Z + K\}^{-1} Z'W(\alpha^0)\tilde{y}(\alpha^0), \qquad (3.7)$$

with "working" observation

$$\tilde{y}(\alpha^0) = \{\tilde{y}_0', \tilde{y}_1'(\alpha_1^0), ..., \tilde{y}_T'(\alpha_T^0)\}' := \{D^{-1}(\alpha^0)\}'\{y - \mu(\alpha^0)\} + Z\alpha^0, \qquad (3.8)$$

where the components $\tilde{y}_0 = a_0$, $\tilde{y}_t(\alpha_t^0) = \{D_t^{-1}(\alpha_t^0)\}'\{y_t - \mu_t(\alpha_t^0)\} + Z_t\alpha_t^0$ $(t = 1,...,T)$. A similar formula, without the penalty matrix K which contains the information of the transition model, is obtained for the iteratively weighted least squares estimate applied to "working" observations in static GLM's.

Assume now the special case of a linear Gaussian state space model, defined by (2.1), (2.6). Then $\mu(\alpha) = Z\alpha$, $D(\alpha)$ is the identity matrix, and the score function becomes

$$s(\alpha) = Z'R^{-1}(y - Z\alpha),$$

with $R = \mathrm{diag}(Q_0, R_1, ..., R_T)$, $R_t = \mathrm{cov}(y_t|\alpha_t)$. The weight matrix $W(\alpha)$ reduces to R^{-1}, and the "working" observation to the actual observation y, since $D(\alpha^0) = I$, $\mu(\alpha^0) = Z\alpha^0$. Therefore (3.7) becomes

$$a = (Z'R^{-1}Z + K)^{-1} Z'R^{-1}y, \qquad (3.9)$$

where $a = (a_{0|T}', a_{1|T}', ..., a_{T|T}')'$ is the vector of smoothed estimates. As already remarked earlier, the classical linear Kalman filter and smoother solves (3.9) efficiently, without explicitly inverting the block-tridiagonal matrix $Z'R^{-1}Z + K$. Comparing now (3.9) and (3.7), we conclude the following: In order to solve (3.7), that is to carry out a single Fisher scoring step in the exponential family case, we can apply any convenient version of linear Kalman filtering and smoothing,

however replacing R_t by $W_t^{-1}(\alpha^0)$ from (3.5) and y by $\tilde{y}(\alpha^0)$ from (3.8). We will call this a "working" Kalman filter and smoother. In the following algorithm, $a_{t|t}$, $V_{t|t}$, $a_{t|t-1}$, $V_{t|t-1}$, $a_{t|T}$, $V_{t|T}$ are numerical approximations to filtered, predicted and smoothed values of α_t and corresponding approximate error covariance matrices.

Working Kalman filter and smoother (WKFS):

Initialization: $a_{0|0} = a_0$, $V_{0|0} = Q_0$.

For t = 1,...,T:

prediction step $a_{t|t-1} = F_t a_{t-1|t-1}$,

$$V_{t|t-1} = F_t V_{t-1|t-1} F_t' + Q_t.$$

correction step $a_{t|t} = a_{t|t-1} + K_t\{\tilde{y}_t(\alpha^0) - Z_t a_{t|t-1}\}$,

$$V_{t|t} = V_{t|t-1} - K_t Z_t V_{t|t-1},$$

with Kalman gain $K_t = V_{t|t-1} Z_t' \{Z_t V_{t|t-1} Z_t' + W_t^{-1}(\alpha^0)\}^{-1}$.

For smoothing we may use the classical fixed interval smoother

For t = T,...,1:

$$a_{t-1|T} = a_{t-1|t-1} + B_t(a_{t|T} - a_{t|t-1}),$$

$$V_{t-1|T} = V_{t-1|t-1} + B_t(V_{t|T} - V_{t|t-1})B_t' \text{ , where}$$

$$B_t = V_{t-1|t-1} F_t' V_{t|t-1}^{-1}$$

or any other computationally efficient version, yielding $\alpha^1 = (a_{0|T}', a_{1|T}', \ldots, a_{T|T}')'$.

Remarks:

(i) Note that for $\alpha^0 = (a_0', a_{1|0}', \ldots, a_{t|t-1}', \ldots, a_{T|T-1}')'$ (WKFS) specializes to (GKFS), the generalized extended Kalman filter and smoother in Fahrmeir (1992). Thus (GKFS) implicitly chooses a reasonable starting vector α^0, but it stops after only one iteration step.

(ii) Applying the matrix inversion lemma, e.g. Anderson and Moore (1979), and considering (3.4), (3.6), it can be shown that the correction step of (WKFS) can be written in scoring form as

correction step*: $V_{t|t} = \{V_{t|t-1}^{-1} + S_t(\alpha_t^0)\}^{-1}$,

$$a_{t|t} = a_{t|t-1} + V_{t|t}\tilde{s}_t(\alpha_t^0)$$

with the "working" score function $\tilde{s}_t(\alpha_t^0) := s_t(\alpha_t^0) - S_t(\alpha_t^0)\{a_{t|t-1} - \alpha_t^0\}$.

As we want to solve (3.2), we have to iterate (WKFS), where the solution $\alpha^{(k)}$ of the previous iteration is the starting vector for the next loop:

Iteratively weighted Kalman filter and smoother (IWKFS):

Initialization: Compute $\alpha^0 = (a_{0|T}^0, a_{1|T}^0, \ldots, a_{T|T}^0)'$ with (GKFS).

Set iteration index k = 0 .

Step 1: Starting with α^k, compute α^{k+1} by application of (WKFS).

Step 2: If a convergence criterion is fulfilled: STOP,

else set k = k+1 and go to Step 1 .

(IWKFS) is a complete Fisher scoring algorithm that makes efficient use of the block-tridiagonal form of $U(\alpha)$ as explicit inversion is avoided. At convergence, we obtain the posterior mode smoother $a = \alpha^{(\bar{k})}$. Moreover, the error covariances $V_{t|T}$ computed in (WKFS) at convergence are the curvatures of $PL(\alpha)$ at $\alpha = a$, i.e. the diagonal blocks of $U(a)^{-1}$, cf. Fahrmeir and Kaufmann (1991), and thus we do not need extra computational effort to get them. This is a very convenient result for hyperparameter estimation as will be seen in the next subsection. The iterative process is suitably initialized with (GKFS) since it does not require a starting vector α^0.

The estimation approach can be easily extended to longitudinal data. Due to the conditional independence of individual responses y_{it} within y_t, the log-likelihood $l(\alpha)$ in (3.3) is now the sum

$$l(\alpha) = \sum_{t=1}^{T}\sum_{i=1}^{n} \ln p\left(y_{it} \middle| \alpha_t y_{t-1}^*\right) = \sum_{t=1}^{T}\sum_{i=1}^{n} l_{it}(\alpha_t)$$

of individual log-likelihood contributions, and score functions and information matrices are also sums of individual contributions, i.e.

$$s_t(\alpha_t) = \sum_{i=1}^{n} s_{it}(\alpha_t),\ S_t(\alpha_t) = \sum_{i=1}^{n} S_{it}(\alpha_t),$$

with s_{it}, S_{it} as in (3.4), (3.6), with additional index i.

3.2. *Estimation of hyperparameters*

In the following we outline two methods for data-driven hyperparameter estimation. One way is to estimate by an EM-type algorithm, similarly as for linear Gaussian dynamic models and as already suggested in Fahrmeir (1992), Fahrmeir and Goss (1992). The procedure for joint estimation of α, Q_0, a_0 and $Q = Q_t$ (t = 1,...,T) can be summarized as follows:

EM-type algorithm:

1. Choose starting values $Q^{(0)}, Q_0^{(0)}, a_0^{(0)}$ and set iteration index p = 0 .
2. Smoothing: Compute $a_{t|T}^{(p)}, V_{t|T}^{(p)}$ (t = 1,...,T) by (GKFS) or (IWKFS), with un known parameters replaced by their current estimates $Q^{(p)}, Q_0^{(p)}$ and $a_0^{(p)}$.
3. EM step: Compute $Q^{(p+1)}, Q_0^{(p+1)}$ and $a_0^{(p+1)}$ by

$$a_0^{(p+1)} = a_{0|T}^{(p)},$$

$$Q_0^{(p+1)} = V_{0|T}^{(p)},$$

$$Q^{(p+1)} = \frac{1}{T}\sum_{t=1}^{T}\left\{\left(a_{t|T}^{(p)} - F_t a_{t-1|T}^{(p)}\right)\left(a_{t|T}^{(p)} - F_t a_{t-1|T}^{(p)}\right)' + V_{t|T}^{(p)} - F_t B_t^{(p)} V_{t|T}^{(p)} - V_{t|T}^{(p)\prime} B_t^{(p)\prime} F_t' + F_t V_{t-1|T}^{(p)} F_t'\right\}$$

with $B_t^{(p)}$ defined as in the fixed interval smoother.

4. If some termination criterion is reached: STOP, else set p = p+1 and go to 2.

A further way is to adopt the principle of cross-validation proposed by Kohn and Ansley (1989) for linear state space models and mentioned in Hastie and Tibshirani (1990), Fahrmeir and Tutz (1994) for static generalized additive models, to the present context. For simplicity we consider univariate responses (q = 1) and summarize the hyperparameters in the vector λ. Let $a = (a_{1|T}', \ldots, a_{T|T}')'$ be the (approximative) solution of (3.2) obtained with (GKFS) or (IWKFS) for a

given vector λ. Extending the generalized cross-validation criterion from static to dynamic generalized linear models, we define

$$GCV(\lambda) = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\left\{y_t - h\left(Z_t a_{t|T}\right)\right\}\Big/\Sigma_t^{1/2}\left(a_{t|T}\right)}{1 - tr\left(H_\lambda\right)/T}\right]^2$$

where H_λ is the "smoother" or "hat" matrix. It can be obtained by the same arguments as for static GLM's (see e.g. Fahrmeir and Tutz, 1994, ch.4): At convergence, the Fisher scoring step (3.7) has the form

$$a = \{Z'W(a)Z + K\}^{-1} Z'W(a)\tilde{y}(a),$$

and the estimated linear predictor is

$$Za = ZU(a)^{-1} Z'W(a)\tilde{y}(a) \ .$$

Suppressing the information connected with the initial prior $p(\alpha_0)$, the smoother matrix H_λ is therefore obtained by omitting the first row and column of $Z\{Z'W(a)Z + K\}^{-1}Z'W(a)$. As the diagonal blocks of $\{Z'W(a)Z + K\}^{-1} = U(a)^{-1}$ are the approximate error covariance matrices $V_{t|T}$ (t = 1,...,T), computed by (IWKFS) at convergence, the diagonal blocks of H_λ are $Z_t V_{t|T} Z_t' W_t(a_{t|T})$ (t = 1,...,T). Therefore

$$tr(H_\lambda) = \sum_{t=1}^{T} Z_t V_{t|T} Z_t' W_t\left(a_{t|T}\right)$$

can be obtained from (IWKFS) without additional computational effort. Unknown hyperparameters λ are estimated by minimizing $GCV(\lambda)$ numerically.

3.3. *Approximate posterior mean analysis*

In Subsection 3.1, the smoothing estimate **a** of the entire state vector α is defined and derived as the posterior mode of $p\left(\alpha \middle| y_T^*\right)$ and inverse information matrices are used as approximate error covariance matrices. Experience with simulated and real data sets indicate satisfactory approximation quality for practical purposes. Simulation results as in Fahrmeir (1992) also provide some

evidence that the posterior $p\left(\alpha \middle| y_T^*\right)$ is approximately Gaussian and, therefore, the posterior mode and associated error covariance matrices are reasonable and useful approximations to the posterior mean. In the following, we give an additional informal argument for approximate posterior normality. It is based on a Taylor expansion of the sampling log-likelihood $l(\alpha)$ about the mode **a** of the posterior, neglecting cubic and higher order terms, as used for Laplace's approximation (e.g. Tierney and Kadane, 1986; Breslow and Clayton, 1993). Carrying out such an expansion, we obtain

$$l(\alpha) = l(a) + (\alpha - a)' Z'D(a)\Sigma^{-1}(a)(y - \mu(a)) - \frac{1}{2}(\alpha - a)'(Z'W(a)Z + A)(\alpha - a)$$

$$+ \text{higher order terms.}$$

The remainder term A is 0 for natural link functions and has expectation 0 for general link functions. Omitting A and higher order terms and rearranging, we get

$$l(\alpha) \approx -\frac{1}{2}\left[Z\alpha - \left\{Za + D^{-1}(a)(y - \mu(a))\right\}\right]' W(a)\left[Z\alpha - \left\{Za + D^{-1}(a)(y - \mu(a))\right\}\right] + C$$

$$= -\frac{1}{2}(\tilde{y}(a) - Z\alpha)' W(a)(\tilde{y}(a) - Z\alpha) + C,$$

where C is independent of α, and $\tilde{y}(a) = Za + D^{-1}(a)(y - \mu(a))$. Thus, $l(\alpha)$ is approximated by a Gaussian sampling log-likelihood

$$\tilde{l}(\alpha) = -\frac{1}{2}(\tilde{y}(a) - Z\alpha)' W(a)(\tilde{y}(a) - Z\alpha) + C,$$

with mean $Z\alpha$, covariance matrix $W(a)$ and observations $\tilde{y}(a)$. Maximizing the approximate penalized likelihood $\tilde{l}(\alpha) - \alpha' K\alpha / 2$ yields the solution

$$\hat{a} = (Z'W(a)Z + K)^{-1} Z'W(a)\tilde{y}(a). \qquad (3.10)$$

Comparing with (3.7) and (3.9), we see that (3.10) corresponds to the solution of (3.7) at convergence. It can be obtained by the linear Kalman smoother. Thus, the posterior is (approximately) Gaussian, with mean $\hat{a}$ (approximately) equal to the mode.

The accuracy of the approximation depends on the data situation and the sample size. For longitudinal data the approximation can be justified asymptotically for $n \to \infty$ and T fixed, with arguments as for the Laplace method (Tierney and Kadane, 1986). The question of approximation quality becomes more difficult for small n, in particular n = 1 as in the pure time series situation. The simulation results in Subsection 4.1 indicate satisfactory behaviour even for this sparse data situation. A rigorous asymptotic theory for $T \to \infty$ and small n would be an interesting topic for further theoretical research.

4. ILLUSTRATIVE APPLICATIONS

The time series of rainfall data in the first application has already been analyzed in Kitagawa (1987) and with (GKFS) in Fahrmeir (1992), and is reanalyzed here for comparison with (IWKFS). Based on this example, a simulation study has been carried out to get some insight into estimation quality. In the second application, we analyze a larger longitudinal data set with multicategorical, ordinal responses from micro-economics.

4.1. *Binary rainfall data*

The data are given by the number of occurences of rainfall in the Tokyo area for each calendar day during the years 1983-1984. To obtain a smooth estimate of the probability π_t of occurence of rainfall on calendar day t (t = 1,...,366), Kitagawa (1987) chose the following dynamic binomial logit model:

$$y_t \sim \begin{cases} B(1,\pi_t),\ t = 60\ (\text{February } 29) \\ B(2,\pi_t),\ t \neq 60, \end{cases} \qquad (4.1)$$

$$\pi_t = h(\alpha_t) = \exp(\alpha_t)/\{1+\exp(\alpha_t)\},$$

$$\alpha_{t+1} = \alpha_t + \xi_t, \xi_t \sim N(0,\sigma^2),\ \ \xi_0 \sim N(a_0, q_0),$$

so that π_t = P(rain on day t) is parametrized by α_t. Up to a constant, the corresponding penalized log-likelihood is

$$PL(\alpha) = y_{60}\alpha_{60} - \ln\left(1+e^{\alpha_{60}}\right) + \sum_{\substack{t=1,\\ t\neq 60}}^{366}\left\{y_t\alpha_t - 2\ln\left(1+e^{\alpha_t}\right)\right\} -$$

$$- \frac{1}{2\sigma^2}\sum_{t=1}^{366}(\alpha_t - \alpha_{t-1})^2 - \frac{1}{2q_0}(\alpha_0 - a_0)^2 \ .$$

Figure 1 shows corresponding smoothed estimates $\hat{\pi}_t = h\left(a_{t|366}\right)$ based on (IWKFS) together with the data points. The random walk variance σ^2 was estimated by the EM-type algorithm and (GCV). Both methods provide the same estimate $\hat{\sigma}^2 = 0.032$. In this example (GKFS) and (IWKFS) lead to more or less the same pattern for the estimated probability of rainfall for calendar days.

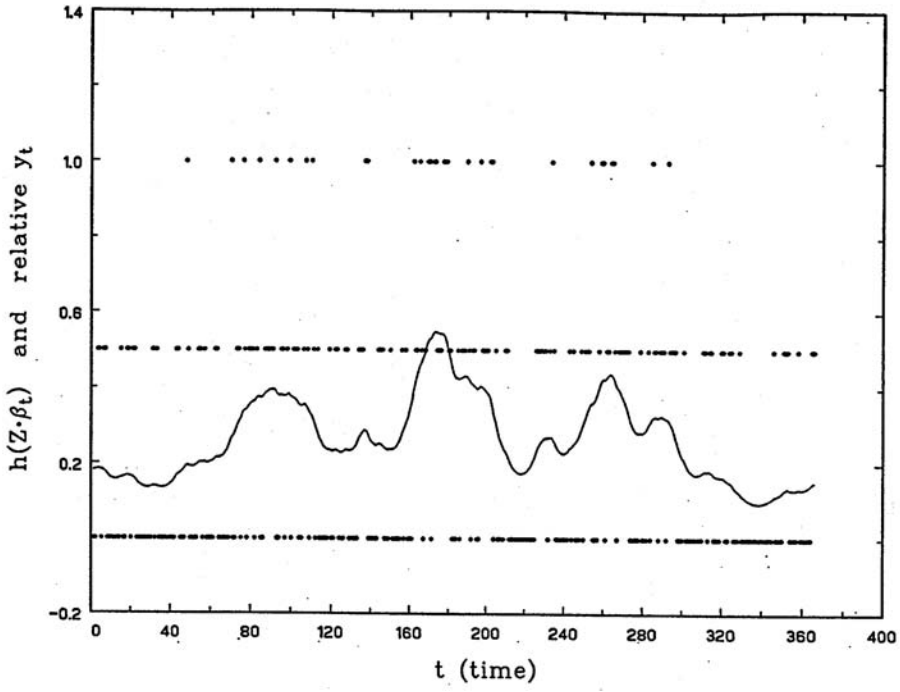

Figure 1 : Tokyo rainfall data, computed with (IWKFS) and RW1.

If we take the second order random walk as transition model, i.e.

$$\alpha_{t+1} = \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix}\alpha_t + \xi_t \ , \ \xi_t \sim N(0,Q) \ , \ Q = \begin{pmatrix} q & 0 \\ 0 & 0 \end{pmatrix},$$

then Figure 2 shows the (GCV)-function dependent on q, computed with (IWKFS). We can observe three local minima, approximately at $3\cdot 10^{-7}$, $3\cdot 10^{-5}$ and 0.008 .

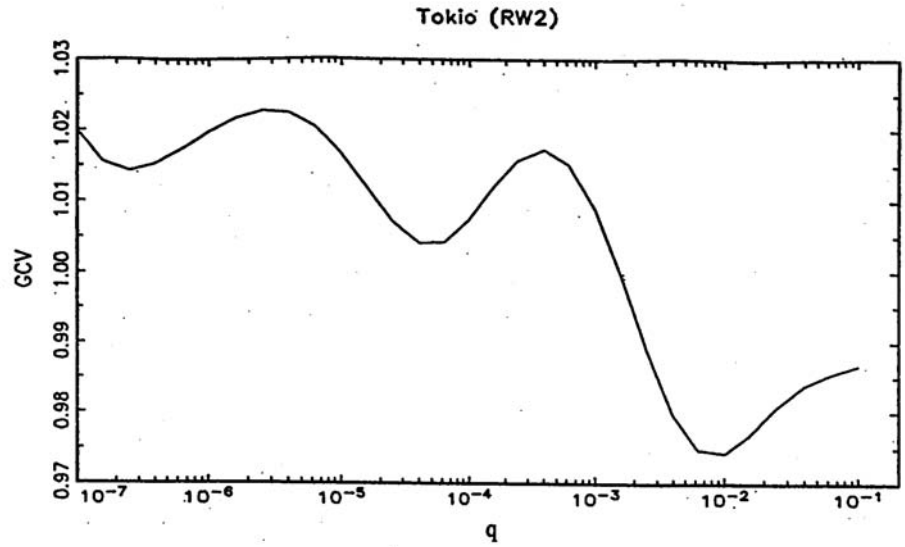

Figure 2 : Tokyo rainfall data. Values of (GCV) dependent on q, computed with (IWKFS) and RW2.

The corresponding estimates $\hat{\pi}_t$ are given in Figure 3. Dependent on the starting value of q, the EM-type algorithm yields the same estimates.

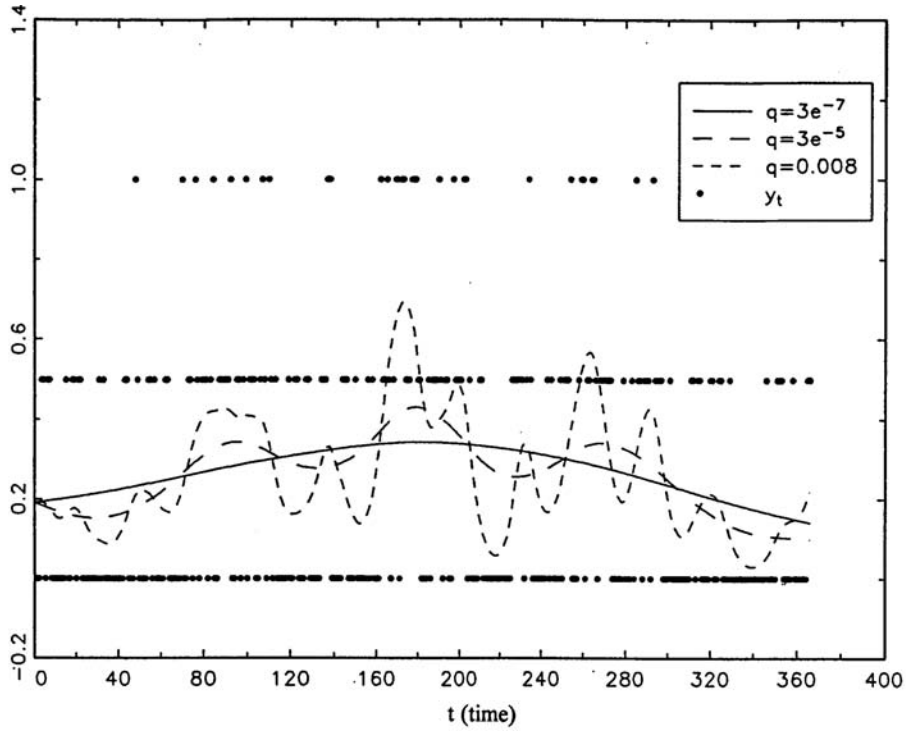

Figure 3 : Tokyo rainfall data, computed with (IWKFS), RW2 and different smoothing parameters q.

To provide some insight into estimation properties, we carried out the following Monte-Carlo experiment: Taking the estimated probabilities $\hat{\pi}_t$ of Figure 1 as

the „true“ probabilities π_t for rainfall on day t, 200 replications of binomial time series $\{y_t^b\}$, b = 1,...,B=200, were generated according to the model (4.1). For each replication $\{y_t^b\}$ smoothed estimates a_t^b and $\pi_t^b = h(a_t^b)$, together with approximate error variances V_t^b for a_t^b and transformed variances $\left(\sigma_t^b\right)^2$ for π_t^b, were computed by (IWKFS) combined with the EM-type algorithm.

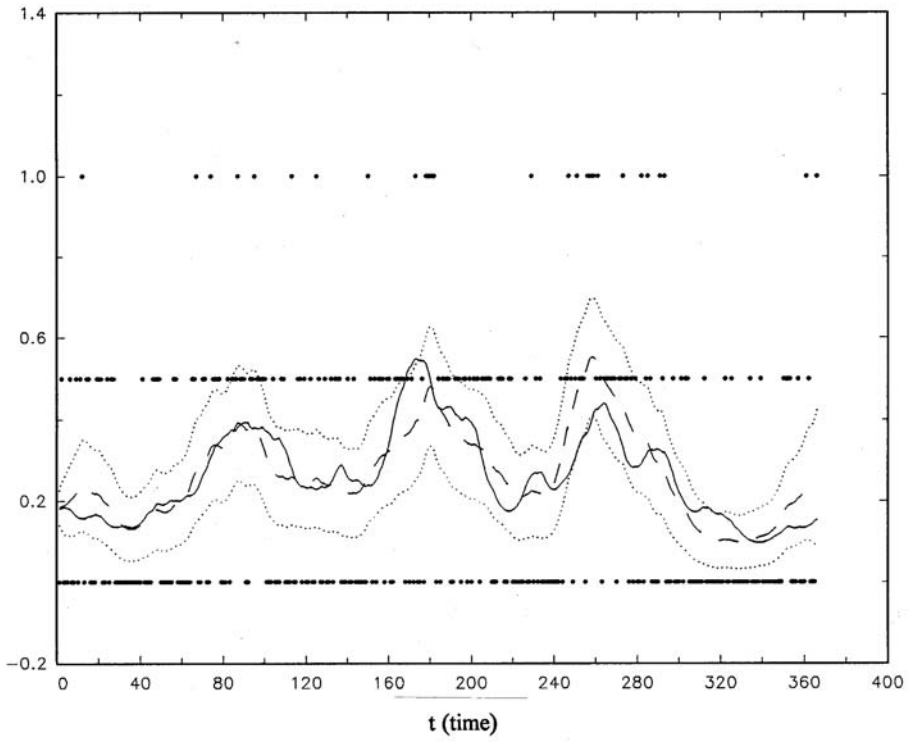

Figure 4 : Monte-Carlo experiment, ——— „true“ probabilities π_t, – – – – estimated probabilities π_t^1 (with ····· 90% confidence band) and the time series $\{y_t^1\}$.

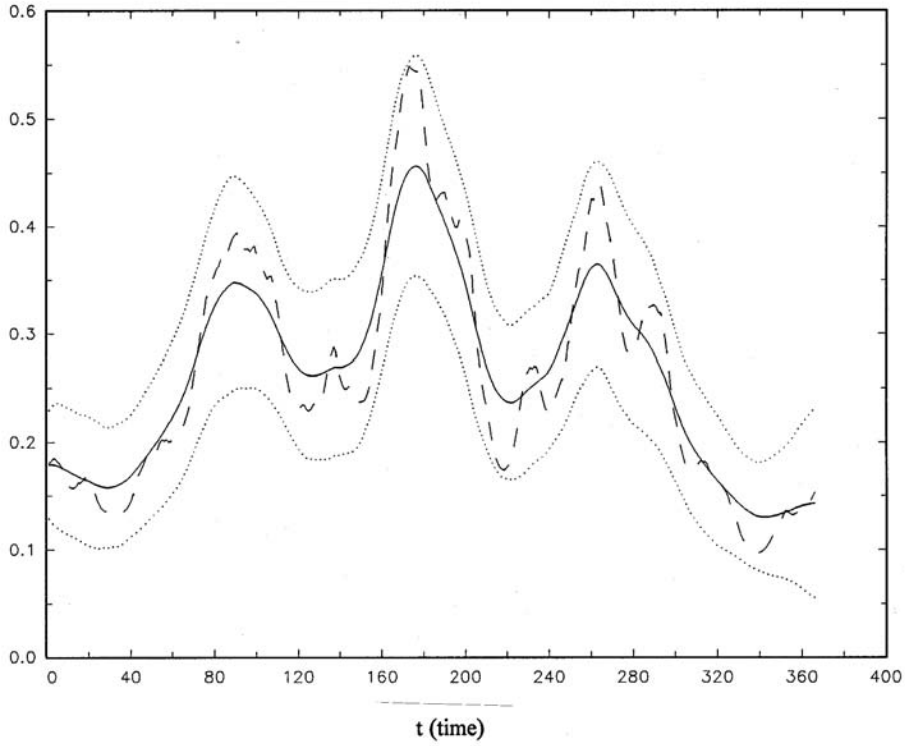

Figure 5 : Monte-Carlo experiment, ——— empirical mean $\tilde{\pi}_t$ (with ····· 90% confidence band) and – – – – „true“ probabilities

Figure 4 displays the „true“ probabilities π_t of Figure 1 (solid line), the time series $\{y_t^1\}$ from the first replicate, and the estimated probabilities π_t^1 together with pointwise 90% confidence bands $\pi_t^1 \pm 1.64\sigma_t^1$. In Figure 5, the „true“ curve is compared to the empirical mean $\tilde{\pi}_t = 1/200\left(\sum \pi_t^b\right)$ of the 200 smoothed estimates π_t^b, together with corresponding pointwise 90% confidence intervals. Both figures indicate that bias is associated with high curvature and that there is a tendency of oversmoothing. However, on the average, the „true“ curve is well covered by the pointwise confidence bands. This can also be seen from Figure 6, where the pointwise coverages out of the 200 replicates are plotted.

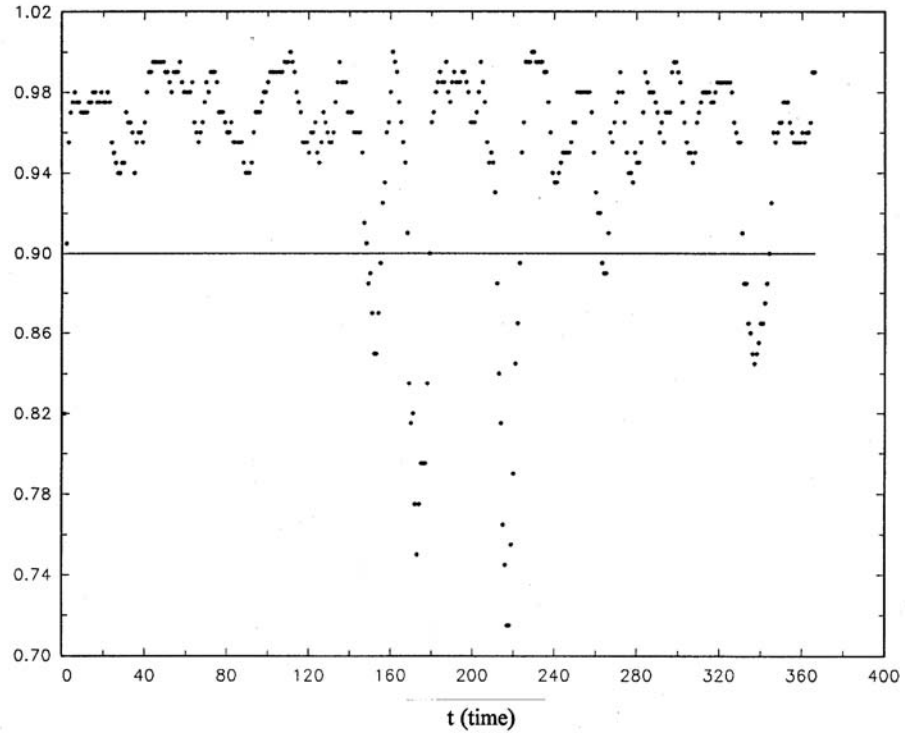

Figure 6 : Monte-Carlo experiment, ——— nominal value 90%. The points are pointwise coverages of 90% intervals.

There is again clear evidence that low coverage is associated with high curvature, which is in agreement with the simulations of Gu (1992) in the context of non-Gaussian spline smoothing. The average coverage probability, however, is about 95%. This indicates that approximate error variances $\left(\sigma_t^b\right)^2$ obtained from (IWKFS) tend to be larger than exact error variances, at least in this example. This is also confirmed by comparing the mean $\tilde{\sigma}_t^2 = 1/200\left\{\sum\left(\sigma_t^b\right)^2\right\}$ of the $\left(\sigma_t^b\right)^2$ to the empirical variances $\sum\left(\pi_t^b - \tilde{\pi}_t\right)^2 \Big/ 200$ obtained from the smoothed estimates out of the 200 replicates in Figure 7.

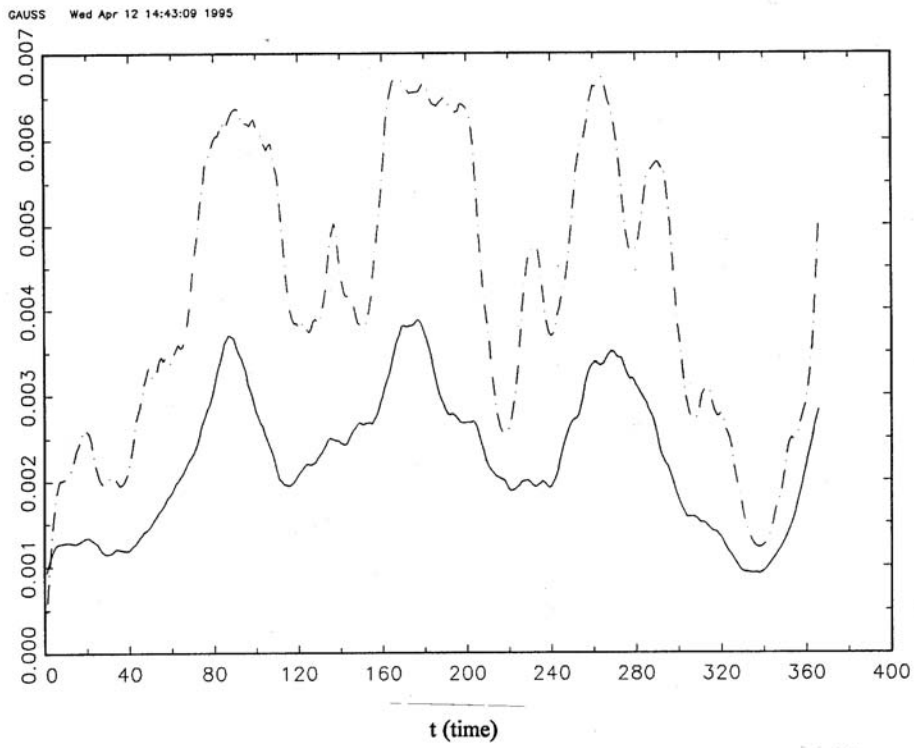

Figure 7 : Monte-Carlo experiment, ——— empirical variances $\sum\left(\pi_t^b - \tilde{\pi}_t\right)^2 \Big/ 200$ and

– – – – the mean of the transformed variances $\tilde{\sigma}_t^2$

4.2. *Business test*

The ordinal longitudinal data in this application are a subset of monthly business microdata collected by the IFO institute in Munich. The questionnaire contains questions on the tendency of successive change of realizations, plans and expectations of variables like production, orders in hand, demand, etc. Answers are categorical, most of them trichotomous and ordinal with categories like "increase" (+), "decrease" (-) and "no change" (=). Currently, several thousand firms from various industry branches participate in this survey on a voluntary basis. We analyze 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_{it} of each firm i (i = 1,...,55), for the t-th month. Its conditional distribution is supposed to depend on the covariates "expected business condition" D_{it}, "orders in hand compared to the

previous month" O_{it} and "production plans of the previous month" $P_{i,t-1}$. No interaction effects are included.

In the following 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 (-), respectively. The relevant dummies for (+) and (=) are abbreviated by $P_{it}^{+}, P_{it}^{=}$ etc. Then a cumulative logit model with stochastic trends τ_{1t}, τ_{2t}, yearly seasonal components γ_{1t}, γ_{2t} for both thresholds and global covariate effects $\beta_t = (\beta_{1t},...,\beta_{6t})'$ is specified by

$$pr\{P_{it} = (+)\} = h\left(\tau_{1t} + \gamma_{1t} + \beta_{1t}P_{i,t-1}^{+} + \beta_{2t}P_{i,t-1}^{=} + \beta_{3t}D_{it}^{+} + \beta_{4t}D_{it}^{=} + \beta_{5t}O_{it}^{+} + \beta_{6t}O_{it}^{=}\right)$$

$$pr\{P_{it} = (+)or(=)\} = h\left(\tau_{2t} + \gamma_{2t} + \beta_{1t}P_{i,t-1}^{+} + \beta_{2t}P_{i,t-1}^{=} + \beta_{3t}D_{it}^{+} + \beta_{4t}D_{it}^{=} + \beta_{5t}O_{it}^{+} + \beta_{6t}O_{it}^{=}\right)$$

,

where $pr\{P_{it} = (+)\}$ and $pr\{P_{it} = (+)or(=)\}$ stand for the probability of increasing and nondecreasing production plans, and h(·) is the logistic distribution function. Trends τ_{1t}, τ_{2t} and covariate effects $\beta_{1t},...,\beta_{6t}$ are modelled by independent random walks of first order, while seasonal components obey autoregressive transition models of order 12, i.e.

$$\gamma_{it} + \gamma_{i,t-1} + ... + \gamma_{i,t-11} = \varpi_{it},\ \varpi_{it} \sim N\left(0, \sigma_{\varpi}^{2}\right),\ i = 1,...,55\ .$$

Unknown hyperparameters were estimated by the EM-type algorithm.

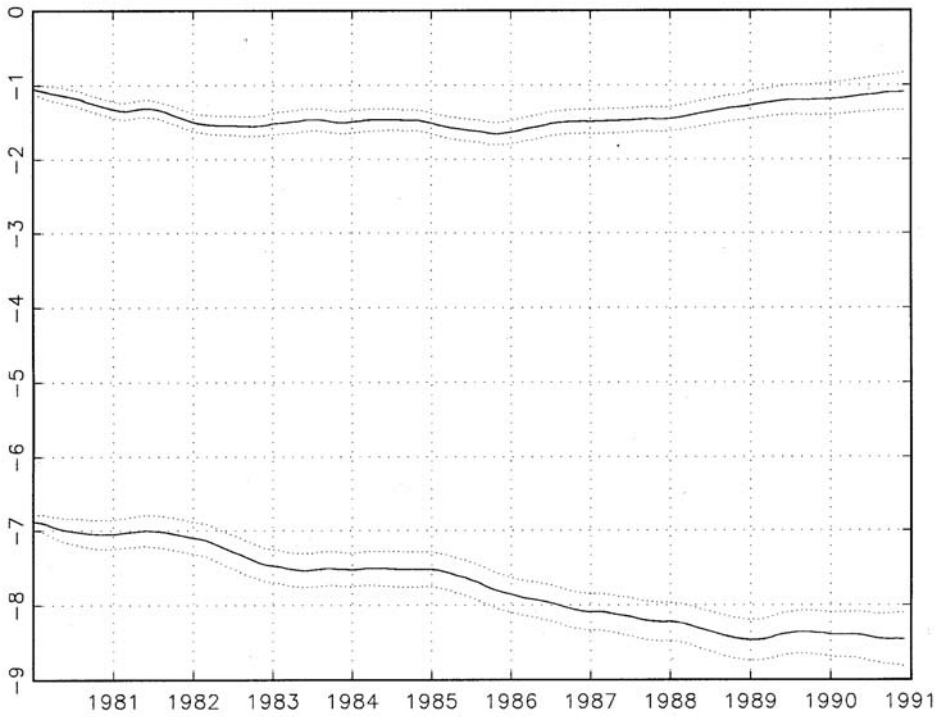

Figure 8 : Business test data. Estimated trends of both threshold parameters, computed with (IWKFS).

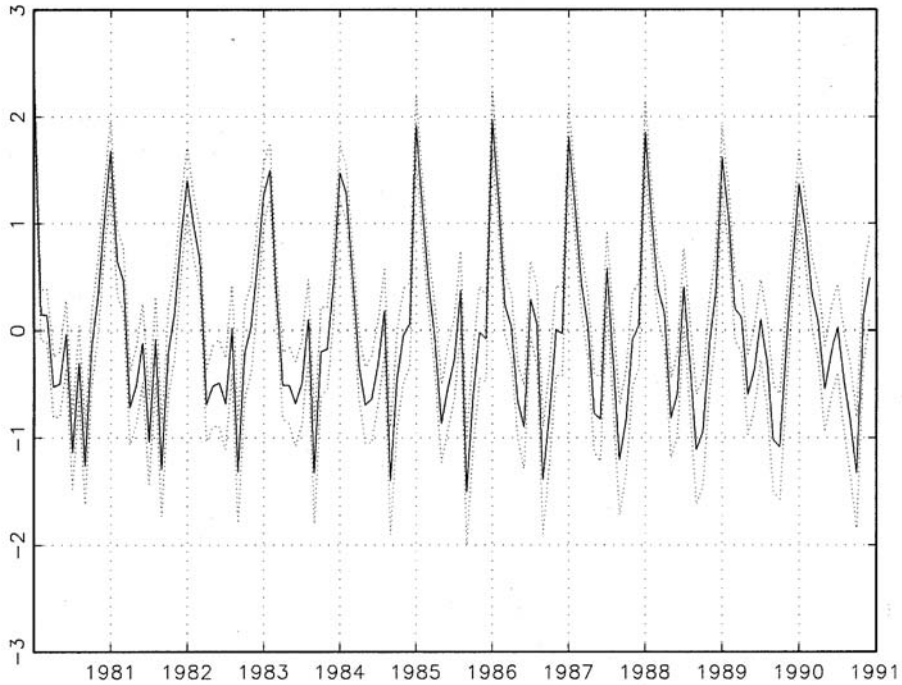

Figure 9a: Business test data. Estimated seasonal component of first threshold parameter, computed with (IWKFS).

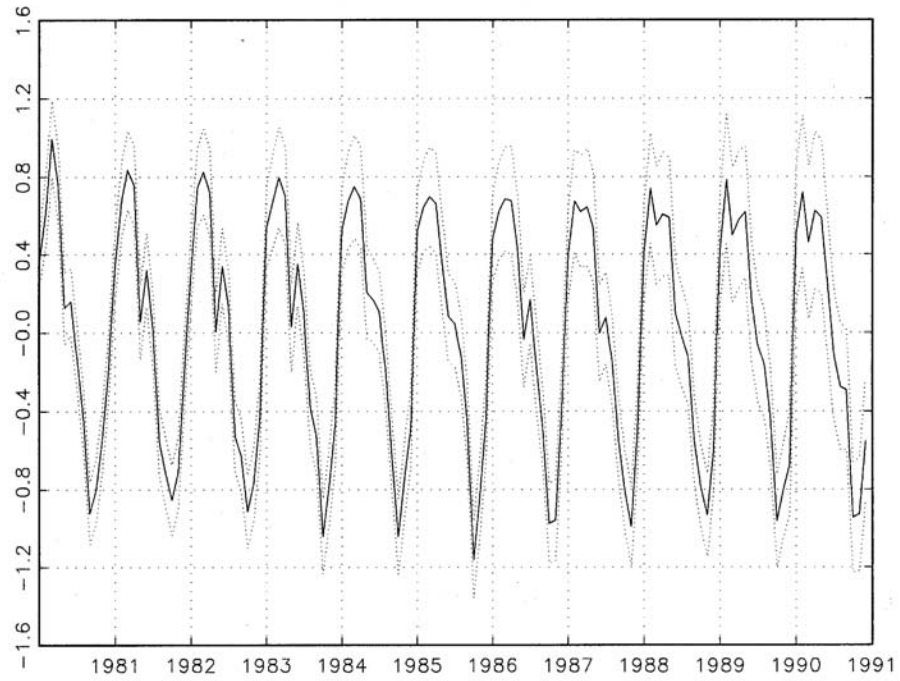

Figure 9b: Business test data. Estimated seasonal component of second threshold parameter, computed with (IWKFS).

Figure 8 gives the estimated trend parameters obtained from (IWKFS). The two trends are comparably smooth and stable, though slightly time-varying. Both seasonal components (Figures 9a and 9b) have a rather distinct pattern with clear at the beginning of the year and corresponding lows in autumn, coinciding with beginning of the new season in building trade industry after less busy months during winter. For the seasonal component γ_{1t}, an additional local peak appears about July to August, indicating plans for increased production after summer vacations. Figure 10 displays the smoothed estimates of the covariate parameters.

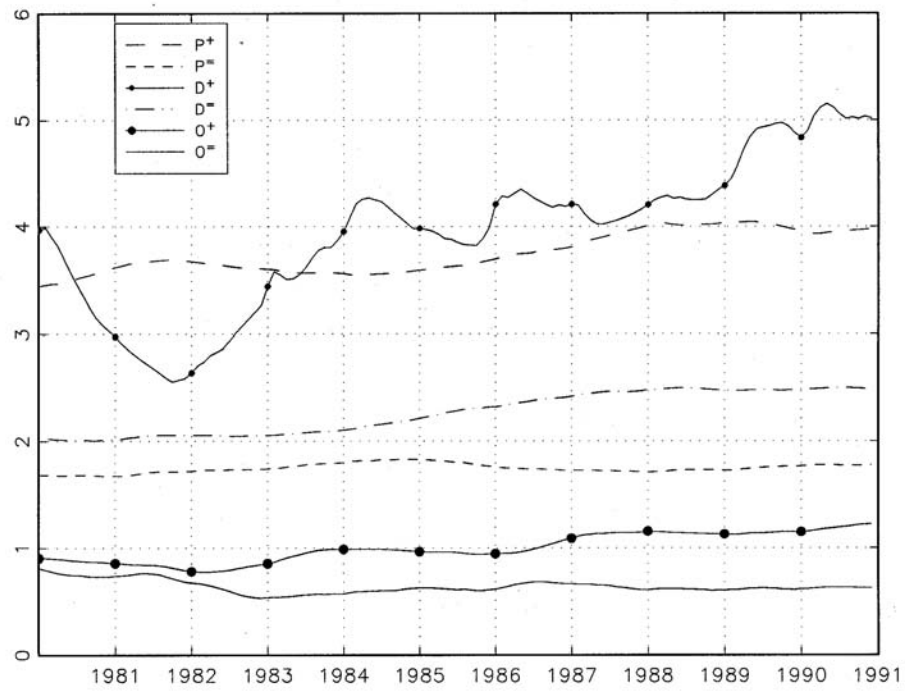

Figure 10: Business test data. Covariate effects, computed with (IWKFS).

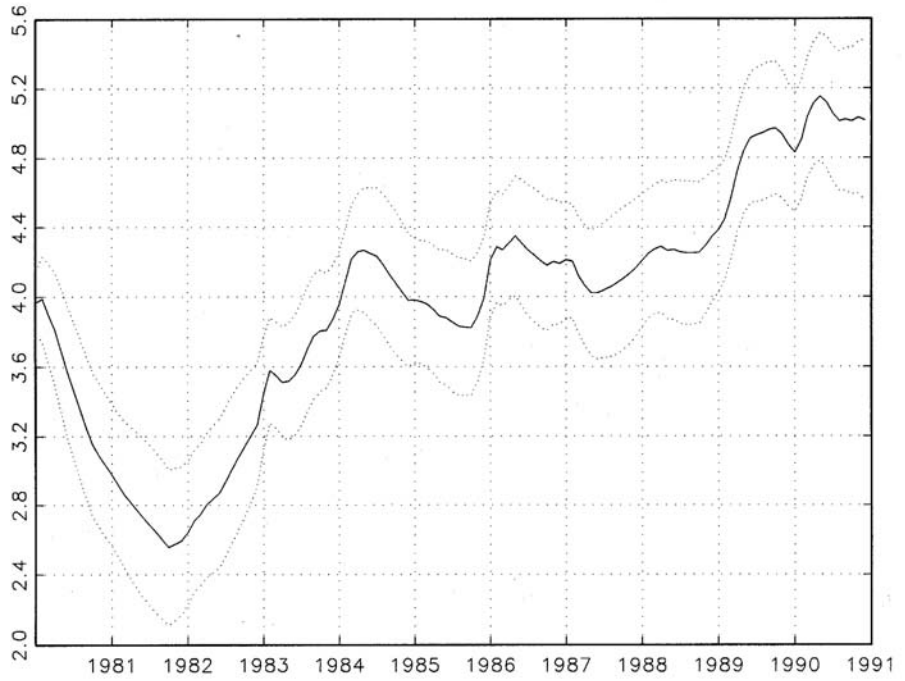

Figure 11: Business test data. Estimated D^+-effect with pointwise confidence bands, computed with (IWKFS).

Compared to the remaining effects, the parameter corresponding to the increase category D^+ of expected development of business has a remarkable temporal variation. Figure 11 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 up 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.

5. GENERAL NON-GAUSSIAN DYNAMIC REGRESSION

In Section 2, smoothing algorithms have been derived for state space models with observation densities from the exponential family. This leads to mathematically convenient expressions, but this restriction can be removed by admitting general non-exponential family densities with piecewise continuous first and second derivatives. A broad class of non-Gaussian models is obtained if we assume that the observation density for $y_t|\alpha_t$ has the general form $p\{y_t|\eta_t = \eta_t(\alpha_t)\}$, where η_t is any parameter of specific interest, for example the mean, and is parameterized as a possibly nonlinear function of the state vector α_t. An important subclass are robust models, where the errors ε_t in the observation equation $y_t = \eta_t(\alpha_t) + \varepsilon_t$ come from a heavy-tailed distribution p_ε, e.g. a Student distribution. Then $p\{y_t|\eta_t = \eta_t(\alpha_t)\}$ is given by $p_\varepsilon(y_t - \eta_t(\alpha_t))$. The incorporation of such a heavy-tail error distribution makes the model robust against additive outliers. Let $l_t\{\eta_t(\alpha_t); y_t\}$ denote the corresponding log-likelihood contribution. The score function contribution is then

$$s_t(\alpha_t) = M_t'(\alpha_t) \cdot g_t(\alpha_t) ,$$

where $M_t'(\alpha_t) = \partial\eta_t/\partial\alpha_t$, $g_t(\alpha_t) = \partial l_t/\partial\eta_t$ and

$$S_t(\alpha_t) = M_t'(\alpha_t) W_t(\alpha_t) M_t(\alpha_t),$$

with $W_t(\alpha_t) = E\{g_t(\alpha_t) g_t'(\alpha_t)\}$ is the expected information matrix contribution. Defining the augmented vector

$$g(\alpha) = \{a_0' - \alpha_0', g_1'(\alpha_1), \ldots, g_T'(\alpha_T)\}' ,$$

the block-diagonal matrix

$$M(\alpha) = \mathrm{diag}\{I, M_1(\alpha_1), \ldots, M_T(\alpha_T)\}$$

and the block-diagonal weight matrix

$$W(\alpha) = \mathrm{diag}\{Q_0^{-1}, W_1(\alpha_1), \ldots, W_T(\alpha_T)\},$$

we obtain, similarly as in Section 3, the (augmented) score vector

$$s(\alpha) = M'(\alpha)g(\alpha)$$

and the expected information matrix

$$S(\alpha) = M'(\alpha)W(\alpha)M(\alpha).$$

Proceeding as in Section 3, a Fisher scoring step from α^0 to α^1 can be written as

$$\alpha^1 = \left(M_0'W_0M_0 + K\right)^{-1} M_0'W_0\tilde{y}\left(\alpha^0\right), \qquad (5.1)$$

where M_0 and W_0 are $M(\alpha)$ and $W(\alpha)$ evaluated at α^0, and

$$\tilde{y}(\alpha^0) = W_0^{-1}g(\alpha_0) + M_0\alpha^0$$

is the working observation.

Comparing with (3.9), it is seen that a Fisher scoring step can be carried out by (WKFS), identifying Z_t with $M_{0,t}$ and $W_t(\alpha^0)$ with $W_{0,t}$.

ACKNOWLEDGEMENTS

We thank the German Science Foundation DFG for financial support.

REFERENCES

ANDERSON, B. D. O. and MOORE, J. B. (1979) *Optimal Filtering.* New Jersey: Prentice-Hall.

BRESLOW; N. E. and CLAYTON, D. G. (1993) Approximate Inference in Generalized Linear Mixed Models. *JASA* 88, 9-25.

CARLIN; B. P., POLSON, N. G. and STOFFER, D. S. (1992) A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modelling. *JASA* 87, 493-500.

CARTER, C. K. and KOHN, R. (1994) On Gibbs Sampling for State Space Models. *Biometrika* 81, 541-553.

DURBIN, J. and KOOPMAN, S. J. (1992) Kalman filtering and smoothing for non-Gaussian time series. Discussion paper: London School of Economics and Political Science.

FAHRMEIR, L. (1992) Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *JASA* 87, 501-509.

— and GOSS, M. (1992) On filtering and smoothing in dynamic models for categorial longitudinal data. In *Statistical Modelling* (eds. P. G. M. van der Heijden, W. Jansen , B. Francis and G. U. H. Seeber). Amsterdam: North-Holland, pp. 85-94.

— and KAUFMANN, H. (1991) On Kalman Filtering, Posterior Mode Estimation and Fisher Scoring in Dynamic Exponential Family Regression. *Metrika* 38, 37-60.

— and TUTZ, G. (1994) *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer.

FRÜHWIRTH-SCHNATTER, S. (1994) Applied State Space Modelling of Non-Gaussian Time Series Using Integration-Based Kalman-Filtering. *Statistics and Computing*, forthcoming.

GU, C. (1992) Penalized likelihood regression: A Bayesian analysis. *Statistica Sinica* 2, 255-264.

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

HASTIE, T. J. and TIBSHIRANI, R. J. (1993) Varying-coefficient Models. *Journal of the Royal Statistical Society* B 55, 757-796.

KITAGAWA, G. (1987) Non-gaussian state-space modeling of nonstationary time series. *JASA* 82, 1032-1063.

KOHN, R. and ANSLEY, C. F. (1989) A fast algorithm for signal extraction, influence and cross-validation in state space models. *Biometrika* 76, 65-79.

SCHNATTER, S. (1992) Integration-based Kalman-filtering for a dynamic generalized linear trend model. *Computational Statistics & Data Analysis* 13, 447-459.

TIERNEY, L. and KADANE, J. (1986) Accurate approximations for posterior moments and marginal densities. *JASA* 81, 82-86.