Wagenpfeil:

Hyperparameter Estimation in Exponential Family State Space Models

Sonderforschungsbereich 386, Paper 6 (1995)

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

Hyperparameter Estimation in Exponential Family State Space Models

Stefan Wagenpfeil

Institut für Statistik, Ludwig-Maximilians-Universität München
Ludwigstr. 33/II, 80539 München, Germany

Summary

Data-driven hyperparameter estimation or automatic choice of the smoothing parameter is of great importance, especially in the applications. This article presents and compares three methods for hyperparameter estimation in the framework of exponential family state space models: First, we motivate and derive a formula for an approximative likelihood, and an alternative, yet mathematical equivalent, expression proves to be a generalized version of a proposal in Durbin and Koopman (1992). Second, the EM-type algorithm suggested in Fahrmeir (1992) is restated here for reasons of comparison and third, the idea of cross-validation proposed by Kohn and Ansley (1989) for linear state space models is extended to the present context, in particular for multicategorical and multidimensional responses. Finally, we compare the three methods for hyperparameter estimation by applying each on three real data sets.

Keywords: Approximative likelihood, choice of the smoothing parameter, cross-validation, EM-type algorithm, penalized likelihood, posterior mode smoothing.

1 Introduction

An important and general tool for modelling time series observations y_t at discrete time $t = 1, 2, \ldots, T$ with fixed or stochastic covariates x_t is the state space approach. To estimate the unobservable structural parameters α_t in the framework of exponential state space models by posterior mode smoothing, a penalized log likelihood criterion can be maximized equivalently. Therefor Fahrmeir (1992) proposed the generalized extended Kalman filter and smoother (GKFS) combined with an EM-type algorithm for hyperparameter estimation, and, as an alternative, Fahrmeir and Wagenpfeil (1994) present an iteratively weighted Kalman filter and smoother (IWKFS). Variances within the penalized log likelihood criterion play, from a nonparametric point of view, the role of smoothing parameters. Data-driven estimation of these hyperparameters is an essential problem, especially in real data applications. For illustration, let us consider the following example:

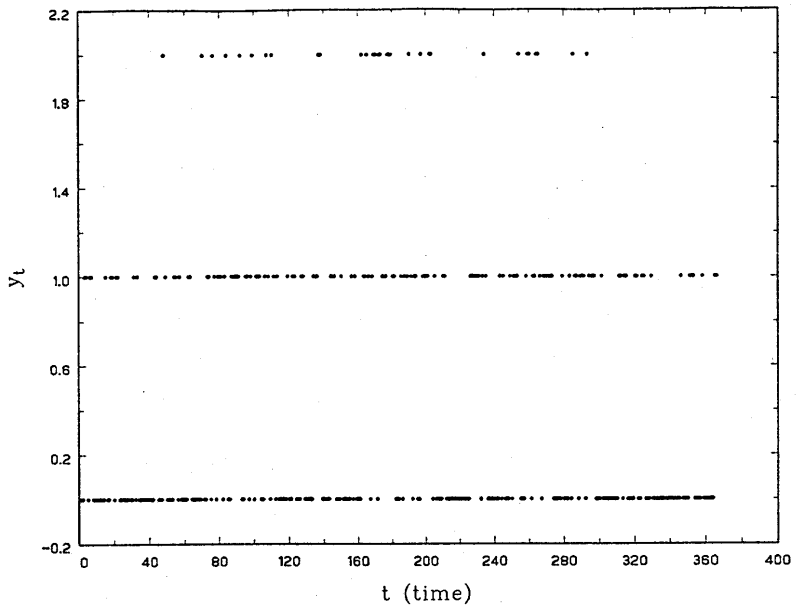

Figure 1: Tokyo rainfall data. Data points.

Figure 1 displays the number of occurences of rainfall in the Tokyo area for each calendar day during the years 1983-1984. With π_t as the probability of occurence of rainfall on calendar day t, $t = 1, \ldots, 366$, Kitagawa (1987)

chose the following dynamic binomial logit model:

$$\begin{aligned} y_t &\sim \begin{cases} \mathrm{B}(1,\pi_t), & t=60 \quad \text{(February 29)} \\ \mathrm{B}(2,\pi_t), & t\neq 60 \end{cases}, \\ \pi_t &= \mathrm{h}(\alpha_t) = \exp(\alpha_t)/1+\exp(\alpha_t), \\ \alpha_{t+1} &= \alpha_t + \xi_t, \quad \xi_t \sim \mathrm{N}(0,q), \quad \xi_0 \sim \mathrm{N}(a_0,q_0), \end{aligned}$$

so that π_t = probability (rain on day t) is parametrized by α_t. Here, a_0, q_0 and q are unknown hyperparameters. Setting $a_0 = -1.51, q_0 = 0.0019$ and $q = 0.5$ fixed, Figure 2 shows corresponding estimates $\hat{\pi}_t = h(a_{t|366})$ based on (GKFS) together with the data points. The estimation is rough and adjusted to the data. Retaining a_0 and q_0 as above and using $q = 0.001$, the estimates $(a_{0|366}, a_{1|366}, \ldots, a_{t|366}, \ldots, a_{366|366})'$ obtained with (GKFS) yield an extremely smooth data-fit, displayed in Figure 3. Comparison of Figure 2 with Figure 3 shows that q acts as a smoothing parameter.

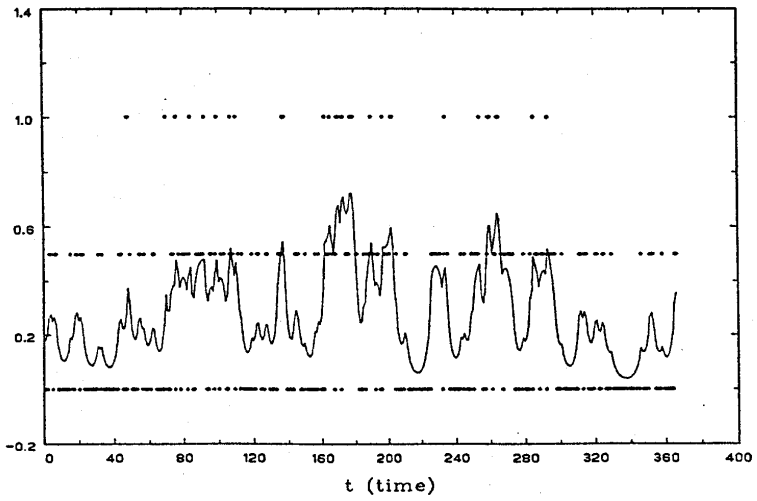

Figure 2: Tokyo rainfall data. Rough fit.

This example illustrates the necessity of procedures for data-driven hyperparameter estimation. In particular, automatically choosen hyperparameters can be a useful starting point for further subjective selections.

In larger simulation studies, Kohn and Ansley (1991) compare the performance of the marginal likelihood estimate with generalized cross-validation GCV and cross-validation CV for Gaussian stae space models. The result is that the marginal likelihood estimate yields often better results than GCV, and GCV itself is better or equal than CV.

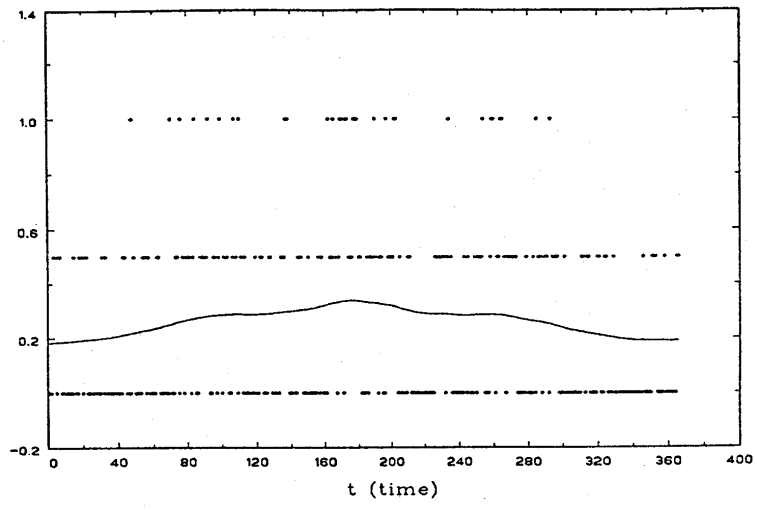

Figure 3: Tokyo rainfall data. Smooth fit.

In this paper, after restating the concept of penalized likelihood estimation for notational purposes in Section 2, we describe three methods for hyperparameter estimation in the framework of exponential family state space models. The approximative likelihood approach as direct Bayesian variant is motivated and derived in Section 3.1. We give a rigorous proof to show that our version is a generalization of a proposal from Durbin and Koopman (1992) allowing for the use of non-natural link functions. The EM-type algorithm as indirect Bayesian method is given in Section 3.2. In Section 3.3, the idea of cross-validation as nonparametric approach is extended to the present exponential family state space context. At this stage, a very useful and from the numerical point of view very desirable property of the estimation procedures (GKFS) and (IWKFS) become apparent: Both algorithms give direct access to the diagonal blocks of the inverse Fisher information matrix, yielding to an efficient computation of the trace of the smoother matrix.

To compare and illustrate the properties of these data-driven methods for hyperparameter estimation empirically, real data applications from the literature are given in Section 4.

2 Penalized likelihood estimation

Our basis for modelling discrete-valued time series observations $y_t \in \mathbb{R}^r, t = 1, 2, \ldots, T$, is the exponential family state space model. Thus we specify the observation model for y_t given the states $\alpha_t \in \mathbb{R}^p$ by the density of a

r-dimensional distribution of the natural exponential family type:

$$y_t|\alpha_t \sim \mathrm{p}(y_t|\alpha_t) = c_t(y_t)\exp\{\theta_t' y_t - b_t(\theta_t)\}, \tag{2.1}$$

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. Z_t is a $q \times r$ design-matrix, maybe dependent on covariates x_t or also on past responses $y_1, \ldots, y_{t-1}$. In the latter case densities, means etc. are to be understood conditionally. By the properties of exponential families the mean and variance functions are then

$$\begin{aligned} \mathrm{E}\,(y_t|\alpha_t) &= \mu_t(\alpha_t) = \partial b_t(\theta_t)/\partial\theta_t, \\ \mathrm{var}\,(y_t|\alpha_t) &= \Sigma_t(\alpha_t) = \partial^2 b_t(\theta_t)/\partial\theta_t\partial\theta_t'. \end{aligned}$$

As in static generalized linear models GLM's the mean μ_t is linked to the linear predictor $\eta_t = Z_t\alpha_t$ by

$$\mu_t = h(Z_t\alpha_t), \tag{2.2}$$

where $h : \mathbb{R}^r \to \mathbb{R}^r$ is an appropriate response function. The exponential family assumption (2.1) together with the mean specification (2.2) is our observation model. Note that for the classical linear state space model, (2.1) and (2.2) specialize to

$$y_t|\alpha_t \sim \mathrm{N}\,(\eta_t = Z_t\alpha_t, R_t) \tag{2.3}$$

where $R_t = \mathrm{var}\,(y_t|\alpha_t)$ is the covariance matrix of y_t given α_t and $h(\cdot)$ the identity function. The observation model is supplemented by a Gaussian transition model with Markov property for α_t:

$$\alpha_t|\alpha_{t-1} \sim \mathrm{N}\,(F_t\alpha_{t-1}, Q_t), \quad t = 1, \ldots, T \tag{2.4}$$

with transition matrix $F_t \in \mathbb{R}^{p,p}$, initial state $\alpha_0 \sim \mathrm{N}\,(a_0, Q_0)$. We summarize the hyperparameters a_0, Q_0, Q_t in the vector λ. Let λ be fixed and known for the moment.

The exponential family state space model (2.1), (2.2), (2.4) covers many well-known time series models, cf. Fahrmeir and Tutz (1994) chapter 8. In this framework we want to estimate the unobservable states α_t via penalized likelihood estimation which could be motivated by posterior mode smoothing outlined in Fahrmeir and Wagenpfeil (1994). With $\alpha = (\alpha_0', \alpha_1', \ldots, \alpha_T')'$, the penalized likelihood estimate $a \in \mathbb{R}^m, m = (T+1)p$, is defined as

$$a := \underset{\alpha}{\arg\max}\ \{\mathrm{PL}(\alpha)\}, \tag{2.5}$$

where

$$\mathrm{PL} : \mathbb{R}^m \to \mathbb{R}, \quad \mathrm{PL}(\alpha) := \sum_{t=1}^{T} l_t(\alpha_t) + \sum_{t=1}^{T} \ln \mathrm{p}(\alpha_t|\alpha_{t-1}) + \ln p(\alpha_0) \tag{2.6}$$

is the penalized log likelihood function with $l_t(\alpha_t) := \ln \mathrm{p}(y_t|\alpha_t)$ for $1 \leq t \leq T$ and the densities from (2.2) and (2.4). Note that $\mathrm{PL}(\alpha)$ in (2.6) reduces to a quadratic function for the linear Gaussian state space model (2.3), (2.4).

To see the connection between hyperparameters in the framework of state space models and smoothing parameters in nonparametric regression, let us regard the simple case where, in addition, $p = 1, Q_t = q \in \mathbb{R}$ for $1 \leq t \leq T$ and $Q_0 = q_0 \in \mathbb{R}$. Then $\mathrm{PL}(\alpha)$ specializes to

$$\begin{aligned} \mathrm{PL}(\alpha) &= -\frac{1}{2}\sum_{t=1}^{T}(y_t - Z_t\alpha_t)' R_t^{-1}(y_t - Z_t\alpha_t) \\ &\quad -\frac{1}{2q}\sum_{t=1}^{T}(\alpha_t - F_t\alpha_{t-1})^2 - \frac{1}{2q_0}(\alpha_0 - a_0)^2. \end{aligned}$$

From a Bayesian point of view, the first term is the log likelihood and the second part acts as a smoothness prior defined by the transition model (2.4) for $\{\alpha_t\}$ with variances q and q_0. If we hold a nonparametric viewpoint, we may consider $\{\alpha_t\}$ not as random variables but as a sequence of unknown states or parameters. Then the first part in $\mathrm{PL}(\alpha)$ measures the goodness of fit obtained by $Z_t\alpha_t$ via weighted euclidean distances and the second one penalizes roughness of the fit. The hyperparameters q and q_0 play the role of smoothing parameters. The problem of hyperparameter estimation is considered in chapter 3.

To compute the penalized likelihood estimate $a \in \mathbb{R}^m$ in the general case, i.e. in the framework of our exponential family state space model, we have to solve (2.5). A numerical solution of the nonlinear programming problem involved in (2.5) could be obtained by various algorithms from optimization theory. To denote one explicit Fisher scoring step in compact matrix notation, the following abbreviations are introduced: the observation vector, augmented by a_0,

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

the vector of expectations augmented by α_0,

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

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

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

the block-diagonal design matrix

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

with $\mathrm{I} \in \mathbb{R}^{p,p}$ as the unit matrix and the block-diagonal matrix

$$H(\alpha) = \mathrm{diag}\,\{\mathrm{I}, H_1(\alpha_1), \ldots, H_T(\alpha_T)\},$$

where $H_t(\alpha_t) = \partial h(\eta_t)/\partial\eta$ is the first derivative of the response function $h(\eta)$ evaluated at $\eta_t = Z_t\alpha_t$. Then the score function of $\sum_{t=1}^{T} l_t(\alpha_t)$ is

$$s(\alpha) = Z'H(\alpha)\Sigma^{-1}(\alpha)\left\{y - \mu(\alpha)\right\},$$

and the block-diagonal (expected) information matrix

$$S(\alpha) = Z'W(\alpha)Z$$

with the weight matrix

$$W(\alpha) = \text{diag}\left\{Q_0^{-1}, W_1(\alpha_1), \ldots, W_T(\alpha_T)\right\} := H(\alpha)\Sigma^{-1}(\alpha)H'(\alpha). \quad (2.7)$$

Defining the symmetric and block-tridiagonal penalty matrix M easily obtained from (2.4), (2.6), as

$$M := \begin{bmatrix} M_{00} & M_{01} & & & 0 \\ M_{10} & M_{11} & M_{12} & & \\ & M_{21} & \ddots & \ddots & \\ & & \ddots & \ddots & M_{T-1,T} \\ 0 & & & M_{T,T-1} & M_{T,T} \end{bmatrix}$$

where

$$\begin{aligned} M_{t-1,t} &:= M'_{t,t-1}, \quad 1 \le t \le T, \\ M_{00} &:= F_1'Q_1^{-1}F_1', \\ M_{tt} &:= Q_t^{-1} + F_{t+1}'Q_{t+1}^{-1} + F_{t+1}, \quad 1 \le t \le T, \\ M_{T+1} &:= 0, \\ M_{t-1,t} &:= -F_t'Q_t^{-1}, \quad 1 \le t \le T, \end{aligned}$$

the first derivative of $\text{PL}(\alpha)$ in (2.6) is

$$u(\alpha) = \partial\text{PL}(\alpha)/\partial\alpha = s(\alpha) - M\alpha$$

and the block-tridiagonal expected information matrix is

$$U(\alpha) = -\text{E}\left\{\partial^2\text{PL}(\alpha)/\partial\alpha\partial\alpha'\right\} = S(\alpha) + M. \quad (2.8)$$

A single Fisher-scoring step from the current iterate $\alpha^0 = (\alpha_0^0, \alpha_1^0, \ldots, \alpha_T^0)' \in \mathbb{R}^m$, say, to the next iterate $\alpha^1 = (\alpha_0^1, \alpha_1^1, \ldots, \alpha_T^1)' \in \mathbb{R}^m$ is then

$$U(\alpha^0)[\alpha^1 - \alpha^0] = u(\alpha^0).$$

This can be rewritten as

$$\alpha^1 = \left\{U(\alpha)\right\}^{-1} Z'W(\alpha^0)\tilde{y}\left(\alpha^0\right) \quad (2.9)$$

with "working" observation

$$\tilde{y}\left(\alpha^0\right) := \left\{a_0', \tilde{y}_1(\alpha_1^0), \ldots, \tilde{y}_T(\alpha_T^0)\right\}' := \left\{H^{-1}(\alpha^0)\right\}' \left\{y - \mu(\alpha^0)\right\} + Z\alpha^0.$$

To solve (2.9) in a numerical efficient way, that is without explicitly inverting the block-tridiagonal expected information matrix $U(\alpha)$, Fahrmeir and Wagenpfeil (1994) propose the "working Kalman filter and smoother". In the following algorithm, $a_{t|t}, V_{t|t}, a_{t|t-1}, V_{t|t-1}, a_{t|T}, V_{t|T}$ denote 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 $t = 1, \ldots, T$:

$$\begin{aligned}
\text{prediction step:} \quad a_{t|t-1} &= F_t a_{t-1|t-1}, \\
V_{t|t-1} &= F_t V_{t-1|t-1} F_t' + Q_t. \qquad (2.10) \\
\text{correction step a):} \quad a_{t|t} &= a_{t|t-1} + K_t \left\{\tilde{y}_t(\alpha_t^0) - Z_t a_{t|t-1}\right\}, \\
V_{t|t} &= V_{t|t-1} - K_t Z_t V_{t|t-1}, \\
\text{with Kalman gain} \quad K_t &= V_{t|t-1} Z_t' \left\{Z_t V_{t|t-1} Z_t' + W_t^{-1}(\alpha_t^0)\right\}^{-1}.
\end{aligned}$$

For smoothing one may use the classical fixed interval smoother.

For $t = T, ..., 1$:

$$\begin{aligned}
a_{t-1|T} &= a_{t-1|t-1} + B_t(a_{t|T} - a_{t|t-1}), \\
V_{t-1|T} &= V_{t-1|t-1} + B_t(V_{t|T} - V_{t|t-1})B_t', \text{ where} \\
B_t &= V_{t-1|t-1} F_t' V_{t|t-1}^{-1} \qquad (2.11)
\end{aligned}$$

or any other computationally more efficient version. The result is $\alpha^1 = (a_{0|T}', a_{1|T}', \ldots, a_{T|T}')' \in \mathbb{R}^m$. Note that, underlying the linear Gaussian state space model (2.3), (2.4), $\alpha^1 = a$, and $a \in \mathbb{R}^m$ is also the posterior mean estimate since posterior modes and means coincide in the normal distribution case. Furthermore (WKFS) reduces to the classical linear Kalman filter and smoother in Kalman gain form. Setting $\bar{y}_t(\alpha_t^0) = H_t'(\alpha_t^0)\tilde{y}_t(\alpha_t^0)$ and supposing that $H_t(\alpha_t^0)$ is regular, we may rewrite the correction step of (WKFS) as

$$\begin{aligned}
\text{correction step b):} \quad a_{t|t} &= a_{t|t-1} + K_t \left\{\bar{y}_t(\alpha_t^0) - H_t'(\alpha_t^0) Z_t a_{t|t-1}\right\}, \\
V_{t|t} &= V_{t|t-1} - K_t H_t'(\alpha_t^0) Z_t V_{t|t-1} \qquad (2.12) \\
\text{with Kalman gain} \quad K_t &= V_{t|t-1} Z_t' H_t(\alpha_t^0) \left\{H_t'(\alpha_t^0) Z_t V_{t|t-1} Z_t' H_t(\alpha_t^0)\right. \\
&\quad \left. + \Sigma_t(\alpha_t^0)\right\}^{-1} \qquad (2.13)
\end{aligned}$$

To solve the nonlinear programming problem (2.5) we have to iterate (WKFS) yielding (IWKFS) as proposed in Fahrmeir and Wagenpfeil (1994):

Iteratively weighted Kalman filter and smoother (IWKFS):

Initialization: Compute $\alpha^0 = (a^0_{0|T}, a^0_{1|T}, \ldots, a^0_{T|T})'$ with (GKFS) from Fahrmeir (1992).

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, e.g. α^{k+1} is very close to α^k: STOP, else set $k = k + 1$ and go to Step 1.

3 Three methods for hyperparameter estimation

So far we assumed the vector of hyperparameters λ to be fixed and known. In the following we describe three methods for data-driven hyperparameter estimation in the framework of our exponential family state space model.

3.1 Approximative likelihood

In the following we motivate and derive an approximative formula for the likelihood $p(y^*)$ with $y^* = (y_1', \ldots, y_T')'$, which proves to be a generalization of a proposal from Durbin and Koopman (1992) allowing for the use of non-narutal link functions. For natural link functions our formula can be regarded as an alternative, however mathematical equivalent, expression to Durbin and Koopman's proposal.

The idea for derivation is as follows: The joint density of α and y^* is

$$p(\alpha, y^*) = p(y^*)p(\alpha|y^*). \tag{3.1}$$

Let the numerical solution $a(\lambda) := \{\bar{a}_{0,T}(\lambda), \bar{a}_{1,T}(\lambda), \ldots, \bar{a}_{T,T}(\lambda)\}' \in \mathbb{R}^m$ of (2.5) be obtained with (GKFS) or (IWKFS). Approximating $p(\alpha|y^*)$ by the normal distribution with expectation $a(\lambda)$ and variance $V(\lambda)$ where

$$V^{-1}(\lambda) := -\mathrm{E}\left[\frac{\partial^2 \ln p\{a(\lambda), y^*\}}{\partial\alpha\partial\alpha'}\right] \overset{(3.1)}{=} -\mathrm{E}\left[\frac{\partial^2 \ln p\{a(\lambda)|y^*\}}{\partial\alpha\partial\alpha'}\right], \tag{3.2}$$

we get

$$p(\alpha|y^*) \approx \frac{1}{(2\pi)^{m/2}\sqrt{\det\{V(\lambda)\}}} \exp\left[-\frac{1}{2}\{\alpha - a(\lambda)\}'V^{-1}(\lambda)\{\alpha - a(\lambda)\}\right].$$

Considering $p(\alpha|y^*)$ as a function of α, we have

$$p\left\{a(\lambda)|y^*\right\} \approx \frac{1}{(2\pi)^{m/2}\sqrt{\det\{V(\lambda)\}}}. \tag{3.3}$$

(3.1) and (3.3) yield

$$p(y^*) \approx f(\lambda) := (2\pi)^{m/2} \left[\det\{V(\lambda)\}\right]^{1/2} p\{a(\lambda), y^*\}, \tag{3.4}$$

where $f(\lambda)$ is the approximative likelihood function. Note that for the linear Gaussian state space model $f(\lambda) = p(y^*)$.

The aim is to maximize the approximative likelihood $f(\lambda)$ in (3.4) with respect to λ. Therefore we give explicit formulae for $[\det\{V(\lambda)\}]^{1/2}$ and $p\{a(\lambda), y^*\}$. Repeated application of Bayes' theorem, using (2.1), (2.2), (2.4) and further independence assumptions, cf. Fahrmeir and Tutz (1994) chapter 8, yields

$$\begin{aligned} \ln p\,\{a(\lambda), y^*\} &= \ln\left\{(2\pi)^{-m/2}\right\} + \ln(\det Q_0)^{-1/2} + \\ &\quad + \sum_{t=1}^{T} \ln(\det Q_t)^{-1/2} + \mathrm{PL}\{a(\lambda)\} \end{aligned} \tag{3.5}$$

with the penalized log likelihood $\mathrm{PL}(\cdot)$ from (2.6) and the densities from (2.1), (2.4),

$$\begin{aligned} \mathrm{PL}\{a(\lambda)\} &= \sum_{t=1}^{T} l_t\{\bar{a}_t(\lambda)\} - \frac{1}{2}\{\bar{a}_0(\lambda) - a_0\}' Q_0^{-1}\{\bar{a}_0(\lambda) - a_0\} \\ &\quad -\frac{1}{2}\sum_{t=1}^{T}\{\bar{a}_t(\lambda) - F_t\bar{a}_{t-1}(\lambda)\}' Q_t^{-1}\{\bar{a}_t(\lambda) - F_t\bar{a}_{t-1}(\lambda)\} \end{aligned} \tag{3.6}$$

Lemma 1 in Appendix A shows that

$$\{\det V(\lambda)\}^{1/2} = \left\{\det Q_0 \prod_{t=1}^{T} G_t(\lambda)\right\}^{1/2}, \tag{3.7}$$

with $G_t(\lambda) := \det V_{t|t} \det(\mathrm{I} - F_t' V_{t|t-1}^{'-1} F_t V_{t-1|t-1}')$, $\quad 1 \le t \le T$, where $V_{t|t}$ and $V_{t|t-1}$ are numerical approximations to filtered and predicted approximate error covariance matrices obtained from (GKFS) or (IWKFS). Considering (3.4), (3.5) and (3.7), the approximative likelihood is thus

$$\begin{aligned} f(\lambda) &= \prod_{t=1}^{T} \left\{\det(Q_t)^{-1/2} \det(V_{t|t})^{1/2} \det(\mathrm{I} - F_t' V_{t|t-1}^{'-1} F_t V_{t-1|t-1}')^{1/2}\right\} \\ &\quad \exp[\mathrm{PL}\{a(\lambda)\}]. \end{aligned}$$

Durbin and Koopman (1992) give a different yet mathematical equivalent expression for $\{\det V(\lambda)\}^{1/2}$. The following formula (3.8) is more general than the original version of Durbin and Koopman (1992) as we do not presume the natural link function well-known from static GLM's:

$$\{\det V(\lambda)\}^{1/2} = \left\{\det Q_0 \prod_{t=1}^{T} A_t(\lambda)\right\}^{1/2} \tag{3.8}$$

with $A_t(\lambda) := \det(Q_t)\det\{\Sigma_t(\alpha_t^0)\}\det\{H_t'(\alpha_t^0)Z_tV_{t|t-1}Z_t'H_t(\alpha_t^0)+\Sigma_t(\alpha_t^0)\}^{-1}$, $1 \le t \le T$. Note that $\Sigma_t(\alpha_t^0) = H_t(\alpha_t^0), t = 1,\ldots,T$, if $h : \mathbb{R}^r \to \mathbb{R}^r$ is the natural response function, that is the inverse of the natural link function, and then we have Durbin and Koopman's formula. Supposing that F_t and $Z_t, 1 \le t \le T$, are regular, Appendix B gives the proof that (3.7) and (3.8) coincide. Considering (3.4), (3.5) and (3.8), the approximative likelihood is then

$$f(\lambda) = \prod_{t=1}^{T}\left[\det\left\{\Sigma_t(\alpha_t^0)\right\}^{1/2}\det\left\{H_t'(\alpha_t^0)ZtV_{t|t-1}Z_t'H_t(\alpha_t^0)+\Sigma_t(\alpha_t^0)\right\}^{-1/2}\right] \exp[\mathrm{PL}\{a(\lambda)\}].$$

The maximization of $f(\lambda)$ with respect to λ can be achieved by various algorithms. In our test examples we used the BFGS algorithm decribed e.g. in Gill, Murray and Wright (1981).

3.2 EM-type algorithm

An indirect Bayesian method for estimation of unknown hyperparameters summarized in λ is the EM algorithm proposed by Dempster, Laird and Rubin (1977). Considering y^* as the observable but incomplete data and $(y'^*, \alpha')'$ as the non-observable, however complete data, the idea is to compute the conditional expectation of the log likelihood given the observations and the current iterate $\lambda^{(k)}$

$$\mathrm{E}\left(\lambda|\lambda^{(k)}\right) := \mathrm{E}\left(\ln p(\alpha, y^*;\lambda)|y^*, \lambda^{(k)}\right).$$

The next iterate $\lambda^{(k+1)}$ is obtained as the maximizer of $\mathrm{E}\left(\lambda|\lambda^{(k)}\right)$ with respect to λ. In the linear Gaussian state space context this optimization problem can be solved analytically. More details are given in Goss (1990). For the exponential family state space model (2.1), (2.2), (2.4), Fahrmeir (1992) suggests to replace posterior expectations by posterior modes $a_{t|T}$ obtained from (GKFS) or (IWKFS). Moreover, as $V_{t|T}$ are the diagonal blocks of $U^{-1}(\alpha)$ in (8), cf. Fahrmeir and Kaufmann (1991), covariance matrices may be replaced by $V_{t|T}$ from (GKFS) or (IWKFS) yielding the following formulae for the

EM-type algorithm:

1. Choose starting values $Q^{(0)}, Q_0^{(0)}, a_0^{(0)}$ and set iteration index $k = 0$.
2. Smoothing: Compute $a_{t|T}^{(k)}, V_{t|T}^{(k)}, \quad t = 1,...,T$ by (GKFS) or (IWKFS), with unknown parameters replaced by their current estimates $Q^{(k)}, Q_0^{(k)}, a_0^{(k)}$.

3. EM step: Compute $Q^{(k+1)}, Q_0^{(k+1)}, a_0^{(k+1)}$ by

$$\begin{aligned} a_0^{(k+1)} &= a_{0|T}^{(k)} \\ Q_0^{(k+1)} &= V_{0|T}^{(k)} \\ Q^{(k+1)} &= \frac{1}{T}\sum_{t=1}^{T}\Big[\left(a_{t|T}^{(k)} - F_t a_{t-1|T}^{(k)}\right)\left(a_{t|T}^{(k)} - F_t a_{t-1|T}^{(k)}\right)' + V_{t|T}^{(k)} \\ &\quad -F_t B_t^{(k)} V_{t|T}^{(k)} - V_{t|T}^{'(k)} B_t^{'(k)} F_t' + F_t V_{t-1|T}^{(k)} F_t'\Big] \end{aligned}$$

   with $B_t^{(k)}$ defined as in (2.11).

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

Note that the EM-type algorithm jointly estimates the structural and hyperstructural parameters α and λ.

3.3 Cross-validation

A further, nonparametric way for hyperparameter estimation is to adjust 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, Chapter 4) for static generalized additive models to the present situation. Let now $a(\lambda) := \{\bar{a}_{1,T}(\lambda), \ldots, \bar{a}_{T,T}(\lambda)\}' \in \mathbb{R}^{m-p}$ be the (approximative) solution of (2.5) obtained with (GKFS) or (IWKFS) for fixed λ. Adopting the idea of cross-validation from static generalized linear models to (dynamic) exponential family state space models and weighting the Pearson residuals as in the generalized cross-validation criterion, we arrive at the generalized cross-validation function

$$GCV(\lambda) = \frac{1}{T}\sum_{t=1}^{T}\frac{\left[y_t - h\left\{Z_t\bar{a}_{t|T}(\lambda)\right\}\right]' \Sigma_t^{-1}\left\{\bar{a}_{t|T}(\lambda)\right\}\left[y_t - h\left\{Z_t\bar{a}_{t|T}(\lambda)\right\}\right]}{\left\{1 - \mathrm{tr}\,(S_\lambda)/T\right\}^2}, \tag{3.9}$$

where S_λ is the smoother or hat matrix. The trace of the smoother matrix $\mathrm{tr}\,(S_\lambda)$ can be computed as follows: Considering (2.9), the estimated weighted linear predictor is

$$\begin{aligned} [W'\{a(\lambda)\}]^{1/2} Z a(\lambda) &= [W'\{a(\lambda)\}]^{1/2} Z [U\{a(\lambda)\}]^{-1} Z' [W\{a(\lambda)\}]^{1/2} \cdot \\ &\quad \cdot [W'\{a(\lambda)\}]^{1/2} \tilde{y}\{a(\lambda)\}. \end{aligned} \tag{3.10}$$

As the approximate error covariance matrices $V_{t|T}, t = 1, \ldots, T$, conveniently and without extra computational effort obtained with (GKFS) or (IWKFS), are the diagonal blocks of the inverse Fisher information matrix $[U\{a(\lambda)\}]^{-1}$

(cf. Fahrmeir and Kaufmann, 1991) and suppressing the information connected with $p(\alpha_0)$, we get from (3.10)

$$S_\lambda = \begin{bmatrix} [W_1'\{\bar{a}_{1|T}(\lambda)\}]^{1/2} Z_1 V_{1|T} Z_1' [W_1\{\bar{a}_{1|T}(\lambda)\}]^{1/2} & & * \\ & \ddots & \\ * & & [W_T'\{\bar{a}_{T|T}(\lambda)\}]^{1/2} Z_T V_{T|T} Z_T' [W_T\{\bar{a}_{T|T}(\lambda)\}]^{1/2} \end{bmatrix}.$$

Thus $\mathrm{tr}\,(S_\lambda) = \sum_{t=1}^{T} \mathrm{tr}\,\left[[W_t'\{\bar{a}_{t|T}(\lambda)\}]^{1/2} Z_t V_{t|T} Z_t' [W_t\{\bar{a}_{t|T}(\lambda)\}]^{1/2}\right]$. To maximize GCV($\lambda$) in (3.9) with respect to λ we used in our test examples the BFGS algorithm with numerical differentiation. However, any nonlinear programming method from optimization theory can be used in principle.

4 Comparison of the three methods

In the following we give an empirical comparison of the three methods for hyperparameter estimation described above.

4.1 Tokyo rainfall data

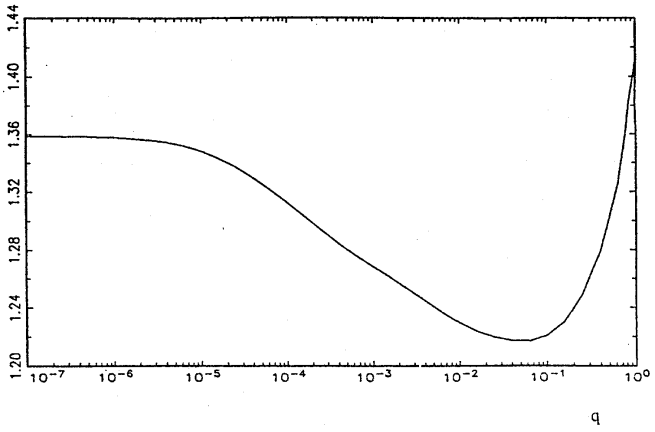

Figure 4: Tokyo rainfall data. GCV-function.

We come back to the example of daily rainfall data from the introduction. The dynamic binomial logit model supplemented with a random walk of order 1 for the parameter process is retained:

$$\begin{aligned} y_t &\sim \begin{cases} \mathrm{B}(1,\pi_t), & t = 60 \quad \text{(February 29)} \\ \mathrm{B}(2,\pi_t), & t \neq 60 \end{cases}, \\ \pi_t &= h(\alpha_t) = \exp(\alpha_t)/1+\exp(\alpha_t), \\ \alpha_{t+1} &= \alpha_t + \xi_t, \quad \xi_t \sim \mathrm{N}(0,q), \quad \xi_0 \sim \mathrm{N}(a_0, q_0), \end{aligned}$$

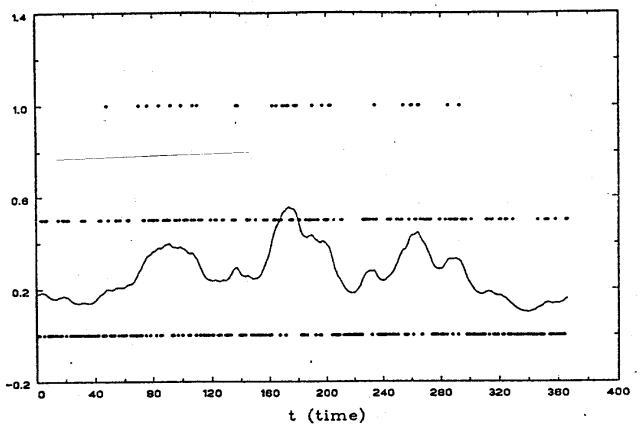

Figure 5: Tokyo rainfall data. Computed with (GKFS) and $\hat{q} = 0.032$

With $a_0 = -1.51, q_0 = 0.0019$ as in the introduction, Figure 4 displays the GCV-function dependent on q. The computed estimate is $\hat{q} = 0.032$. The EM-type algorithm and maximizing the approximative likelihood f(λ) yield the same result. What strikes is the slow convergence rate of the EM-type algorithm in comparison to the other methods. Figure 5 shows the estimates $\hat{\pi}_t = h(a_{t|366})$ computed with (GKFS) and $\hat{q} = 0.032$. The fit is smoother than in Figure 1 and rougher than in Figure 2.

4.2 Advertising data

West, Harrison and Migon (1985) analyzed weekly counts y_t of the number of people, out of a sample of $n = 66$, who give a positive response to the advertisement of a chocolate bar. As a measure of advertisement influence, an "adstock coefficient" serves as a covariate x_t. Our framework for estimation is the following dynamic binomial logit model, with

$$y_t \sim \mathrm{B}(66, \pi_t), \quad \pi_t = h(\tau_t + x_t\beta_t), \quad \alpha_{t+1} = \alpha_t + \xi_t,$$

with $\alpha_t = (\tau_t, \beta_t)'$ and $\mathrm{cov}\,\xi_t = \mathrm{diag}\,(q_1, q_2)$. The EM-type algorithm yields $\hat{q}_1 = 0.0016$ and $\hat{q}_2 = 0.00031$ whereas the result from the GCV-criterion is different: $\hat{q}_1 = 0.0006362$ and $\hat{q}_2 = 0.000239$.

Figure 6 displays the smoothed estimates $\hat{\pi}_t$ obtained with (GKFS) and EM-type algorithm. The fit for GCV, however, shows no remarkable differences. The estimation of $\hat{q}_1$ and $\hat{q}_2$ with the approximative likelihood failed due to numerical problems during the optimization procedure.

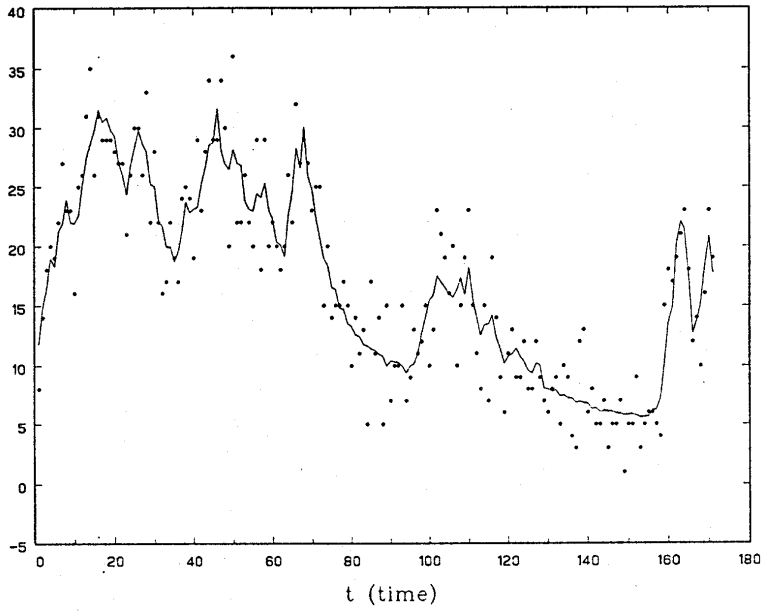

Figure 6: Advertising data. Computed with (GKFS) and EM-type algorith

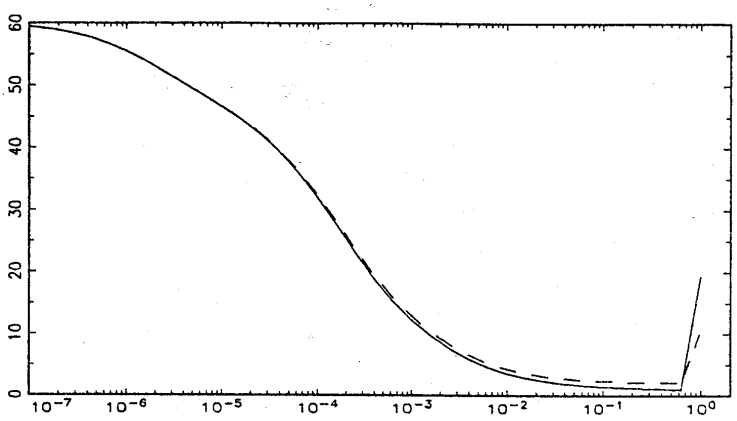

Figure 7: Phone calls. GCV-function

4.3 Phone calls

The data, analyzed in West, Harrison and Migon (1985), consist in counts of phone calls, registrated within successive periods of 30 minutes, at the University of Warwick, from Monday, September 6, 1982, 0.00 to Sunday, September 12, 1982, 24.00 . We analyze the data with a dynamic loglinear Poisson model:

$$y_t \sim \mathrm{Po}\,(\exp(\alpha_t)), \quad \mu_t = \exp(\alpha_t)$$

$$\alpha_t = \alpha_{t-1} + \xi_t, \quad \xi_t \sim N(a_0\,, q), \quad \alpha_0 \sim N(0, q_0).$$

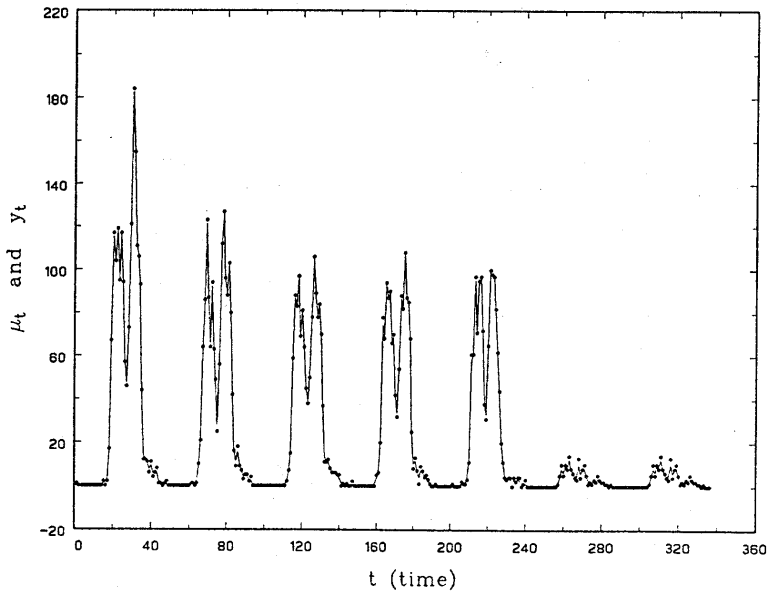

Figure 8: Phone calls. Computed with (IWKFS) and $\hat{q} = 0.44$

The EM-type algorithm with (IWKFS) yields the following hyperparameter estimates: $\hat{q}_0 = 0.015, \hat{a}_0 = -0.864$ and $\hat{q} = 0.44$. With GCV the same estimated $\hat{q}$ is obtained as can be seen from Figure 7 displaying the GCV-function dependent on q. Figure 8 shows the corresponding fit computed with (IWKFS) in combination with the data points. The result is adjusted to the data and provides only moderate smoothing. Maximizing the approximative likelihood yields different estimates: $\hat{q} = 0.0077$ or $\hat{q} = 0.0976$, dependent on the starting value of q. Figure 9, computed with (IWKFS) and $\hat{q} = 0.0077$, shows a quite smooth estimation without neglecting the cyclical structure of the data.

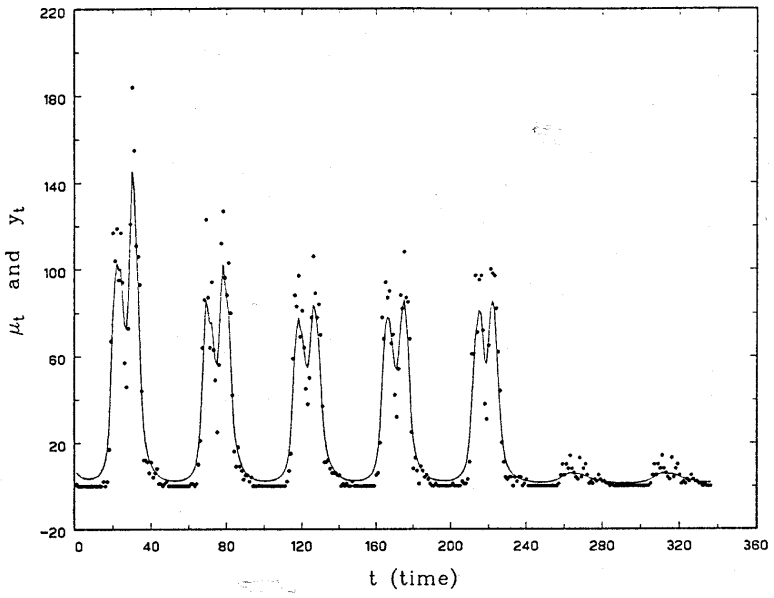

Figure 9: Phone calls. Computed with (IWKFS) and $\hat{q} = 0.0077$

Conclusion

The EM-type algorithm is a very robust method for hyperparameter estimation. However, convergence is slow and sometimes the result seems to depend on the starting point and on the value of the stopping accuracy. Thus estimation algorithms with a higher rate of convergence should be a point of further research. The GCV-criterion as well as maximizing the approximative likelihood could be an alternative since these methods use the convergence rate of nonlinear programming algorithms. In most situations of our study GCV worked well, whereas hyperparameter estimation with the approximative likelihood often ran into numerical problems for multidimensional λ. Summarizing we could say: EM is robust but slow, GCV and approximative likelihood are faster if they work.

Acknowledgements:

I want to thank the German Science Foundation DFG for financial support. Furthermore I am very grateful to Christian Kastner for carefully typing the TEX-version of this manuskript. Special thanks go to Prof. Dr. L. Fahrmeir for his support and the encouragement of my work.

Appendix A

Lemma 1:
Let $V_{t|t}$ and $V_{t|t-1}$ denote numerical approximations to filtered and predicted approximate error covariance matrices obtained with (GKFS) or (IWKFS). Then

$$\det V(\lambda) = \det Q_0 \cdot \prod_{t=1}^{T} \det V_{t|t} \cdot \prod_{t=0}^{T-1} \det \left(\mathrm{I} - F'_{t+1} V'^{-1}_{t+1|t} F_{t+1} V'_{t|t} \right).$$

Proof:
Since the normalizing terms in (3.5) are independent of α, (2.8) and (3.2) show that $V^{-1}(\lambda) = U\{a(\lambda)\}$. Furthermore $U\{a(\lambda)\}$ can be uniquely factorized according to Fahrmeir and Kaufmann (1991) into

$$U\{a(\lambda)\} = LDL'$$

with lower triangular matrix $L = \begin{bmatrix} \mathrm{I} & & & 0 \\ -B'_1 & \mathrm{I} & & \\ & \ddots & \ddots & \\ 0 & & -B'_T & \mathrm{I} \end{bmatrix}$, B_t from (2.11), $1 \le t \le T, \mathrm{I} \in \mathbb{R}^{p,p}$ as the unit matrix, $D = \mathrm{diag}\,(D_0, D_1, \ldots, D_T)$,

$$D_t^{-1} = V_{t|t} - B_{t+1} V_{t+1|t} B'_{t+1}, 0 \le t \le T-1, \quad D_T^{-1} = V_{T|T}, \qquad \text{(A.1)}$$

$V_{t|t}, 0 \le t \le T$, from (2.12) and $V_{t|t-1}, 1 \le t \le T$, from (2.10). Although $V_{t|t}$ and $V_{t|t-1}$ can be regarded as functions of λ, we suppress this dependence for notational convenienc. Thus

$$\begin{aligned}
\det V(\lambda) &= \det U^{-1}\{a(\lambda)\} = \frac{1}{\det U\{a(\lambda)\}} \\
&\overset{\det(L)=1}{=} \frac{1}{\prod_{t=0}^{T} \det D_t} = \prod_{t=0}^{T} \det D_t^{-1} \\
&\overset{(A.1)}{=} \det V_{T|T} \prod_{t=0}^{T-1} \det \left(V_{t|t} - B_{t+1} V_{t+1|t} B'_{t+1} \right) \\
&\overset{B_t \text{ from } (2.11)}{=} \det V_{T|T} \prod_{t=0}^{T-1} \left\{ \det V_{t|t} \det \left(\mathrm{I} - F'_{t+1} V'^{-1}_{t+1|t} F_{t+1} V'_{t|t} \right) \right\} \\
&\overset{V_{0|0}=Q_0}{=} \det Q_0 \prod_{t=1}^{T} \det V_{t|t} \prod_{t=0}^{T-1} \det \left(\mathrm{I} - F'_{t+1} V'^{-1}_{t+1|t} F_{t+1} V'_{t|t} \right).
\end{aligned}$$

Appendix B

To show that (3.7) and (3.8) in Chapter 3.1 coincide, we have to proof

$$G_t(\lambda) = A_t(\lambda) \text{ for } 1 \le t \le T, \qquad \text{(B.1)}$$

assuming that F_t and Z_t, $1 \leq t \leq T$, are regular. Therefor we use

$$F_t V'_{t-1|t-1} = (V_{t|t-1} - Q_t)' F_t^{'-1} \tag{B.2}$$

obtained from (2.10). Moreover we need

$$\begin{aligned}
\det\left(\mathrm{I} - F_t' V_{t|t-1}^{'-1} F_t V'_{t-1|t-1}\right) &\overset{(B.2)}{=} \det\left\{\mathrm{I} - F_t' V_{t|t-1}^{'-1}(V'_{t|t-1} - Q'_t) F_t^{'-1}\right\} \\
&= \det\left\{\mathrm{I} - \left(\mathrm{I} - F_t' V_{t|t-1}^{'-1} Q'_t F_t^{'-1}\right)\right\} \\
&= \det(F'_t)\det(F_t^{'-1})\det(V_{t|t-1}^{'-1})\det(Q'_t) \\
&= \det(V_{t|t-1})^{-1}\det(Q_t). \qquad \text{(B.3)}
\end{aligned}$$

To proof (B.1) we use the definition of $G_t(\lambda)$ and get

$$\begin{aligned}
G_t(\lambda) &\overset{(2.12)}{=} \det\left\{V_{t|t-1} - K_t H_t'(\alpha_t^0) Z_t V_{t|t-1}\right\} \cdot \\
&\quad \det\left(\mathrm{I} - F_t' V_{t|t-1}^{'-1} F_t V'_{t-1|t-1}\right) \\
&\overset{(B.3)}{=} \det(V_{t|t-1})\det\left[\mathrm{I} - Z_t' H_t(\alpha_t^0)\left\{H_t'(\alpha_t^0) Z_t V_{t|t-1} Z_t' H_t(\alpha_t^0) + \right.\right. \\
&\quad \left.\left. \Sigma_t(\alpha_t^0)\right\}^{-1} H_t'(\alpha_t^0) Z_t V_{t|t-1}\right] \det(V_{t|t-1})^{-1}\det(Q_t) = \\
&\overset{Z_t \text{ regular}}{=} \det(Q_t)\det\left[Z_t' H_t(\alpha_t^0)\left\{H_t^{'-1}(\alpha_t^0) Z_t^{'-1} V_{t|t-1}^{-1} Z_t^{-1} H_t^{'-1}(\alpha_t^0) - \right.\right. \\
&\quad \left.\left. \left(H_t'(\alpha_t^0) Z_t V_{t|t-1} Z_t' H_t(\alpha_t^0) + \Sigma_t(\alpha_t^0)\right)^{-1}\right\} H_t'(\alpha_t^0) Z_t V_{t|t-1}\right] \\
&= \det(Q_t)\det\left\{H_t'(\alpha_t^0) Z_t V_{t|t-1} Z_t' H_t(\alpha_t^0) + \Sigma_t(\alpha_t^0)\right\}^{-1} \cdot \\
&\quad \det\left\{H_t'(\alpha_t^0) Z_t V_{t|t-1} Z_t' H_t(\alpha_t^0) + \Sigma_t(\alpha_t^0)\right\} \cdot \\
&\quad \det\left\{H_t'(\alpha_t^0) Z_t V_{t|t-1} Z_t' H_t(\alpha_t^0)\right\} \cdot \\
&\quad \det\left[H_t^{-1}(\alpha_t^0) Z_t^{'-1} V_{t|t-1}^{-1} Z_t^{-1} H_t^{'-1}(\alpha_t^0) - \right. \\
&\quad \left. \left\{H_t'(\alpha_t^0) Z_t V_{t|t-1} Z_t' H_t(\alpha_t^0) + \Sigma_t(\alpha_t^0)\right\}^{-1}\right] \cdot \\
&= \det(Q_t)\det\left\{H_t'(\alpha_t^0) Z_t V_{t|t-1} Z_t' H_t(\alpha_t^0) + \Sigma_t(\alpha_t^0)\right\}^{-1} \cdot \\
&\quad \det\left\{H_t'(\alpha_t^0) Z_t V_{t|t-1} Z_t' H_t(\alpha_t^0) + \Sigma_t(\alpha_t^0)\right\} \cdot \\
&\quad \det\left\{\Sigma_t(\alpha_t^0) H_t^{-1}(\alpha_t^0) Z_t^{'-1} V_{t|t-1}^{-1} Z_t^{-1} H_t^{'-1}(\alpha_t^0)\right\} \\
&= \det(Q_t)\det\left\{H_t'(\alpha_t^0) Z_t V_{t|t-1} Z_t' H_t(\alpha_t^0) + \Sigma_t(\alpha_t^0)\right\}^{-1} \cdot \\
&\quad \det\left\{\Sigma_t(\alpha_t^0)\right\}\det(\mathrm{I}) \\
&\overset{(3.8)}{=} A_t(\lambda).
\end{aligned}$$

References

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977): Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society* B **39**, 1-38.

Durbin, J. and Koopman, S.J. (1992): Kalman filtering and smooting 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.

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

Fahrmeir, L. and Tutz, G. (1994): *Multivariate Statistical Modelling Based On Generalized Linear Models*, Springer-Verlag, New York.

Fahrmeir, L. and Wagenpfeil S. (1994): Iteratively weighted Kalman filtering and smoothing for exponential family state space models, *Journal of Time Series Analysis*, submitted.

Gill, P.E., Murray, W. and Wright, M.H. (1981): *Practical Optimization*, Academic Press, London, San Diego, New York.

Goss, M. (1990): *Schätzung und Identifikation von Struktur- und Hyperstrukturparametern in Dynamischen Generalisierten Linearen Modellen. Theoretische Grundlagen und empirische Auswertungen*, Dissertation, Wirtschaftswissenschaftliche Fakultät der Universität Regensburg, Regensburg.

Hastie, T.J. and Tibshirani, R.J. (1990): *Generalized Additive Models*, Chapman and Hall, London, New York, Tokyo, Melbourne, Madras.

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.

Kohn, R. and Ansley, C.F. (1991): A Signal Extraction Approach to the Estimation of Treatment and Control Curves, *JASA* **86**, 1034-1041.

Kohn, R., Ansley, C.F. and Tharm, D. (1991): The Performance of Cross-Validation and Maximum Likelihood Estimation of Spline Smoothing Parameters, *JASA* **86**, 1042-1050.

West, M., Harrison, P.J. and Migon, H.S. (1985): Dynamic generalized linear models and Bayesian forecasting, *JASA* **80**, 73-83.