LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR STATISTIK
SONDERFORSCHUNGSBEREICH 386

Fahrmeir, Künstler:

Penalized likelihood smoothing in robust state space models

Sonderforschungsbereich 386, Paper 111 (1998)

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

Projektpartner

Penalized likelihood smoothing in robust state space models

Ludwig Fahrmeir and Rita Künstler
Seminar für Statistik
Universität München
Ludwigstr.33
80539 München
fahrmeir@stat.uni-muenchen.de
kuenstler@stat.uni-muenchen.de

March 13, 1998

Abstract In likelihood-based approaches to robustify state space models, Gaussian error distributions are replaced by non-normal alternatives with heavier tails. Robustified observation models are appropriate for time series with additive outliers, while state or transition equations with heavy-tailed error distributions lead to filters and smoothers that can cope with structural changes in trend or slope caused by innovations outliers. As a consequence, however, conditional filtering and smoothing densities become analytically intractable. Various attempts have been made to deal with this problem, reaching from approximate conditional mean type estimation to fully Bayesian analysis using MCMC simulation. In this article we consider penalized likelihood smoothers, this means estimators which maximize penalized likelihoods or, equivalently, posterior densities. Filtering and smoothing for additive and innovations outlier models can be carried out by computationally efficient Fisher scoring steps or iterative Kalman-type filters. Special emphasis is on the Student family, for which EM-type algorithms to estimate unknown hyperparameters are developed. Operational behaviour is illustrated by simulation experiments and by real data applications.

KEYWORDS: Additive outliers, EM algorithm, innovations outliers, iterative Kalman Filtering, non-Gaussian state space models.

1 INTRODUCTION

Robustification of state space models and of filtering and smoothing algorithms has been considered by various authors. In this paper we follow the approach of Martin (1979), West (1981, 1984), Meinhold and Singpurwalla (1989) among others, where errors are assumed to be non-Gaussian with longer than normal tails. As is well-known, exact closed-form solutions to the filtering and smoothing problem are generally no longer available. Approximate filtering and smoothing algorithms have therefore been given already in early work on robustified state space modelling, for example approximate conditional mean (ACM) type smoothers (see Martin, 1979, or Martin and Raftery, 1987). Kitagawa (1987) uses numerical integration for computing

posterior means, but the method becomes infeasible for higher state dimension. More recently, fully Bayesian MCMC simulation methods for models with finite Gaussian mixtures have been developed to tackle this problem, see for instance Carter and Kohn (1996a, 1996b). Shephard and Pitt (1997), and Durbin and Koopman (1997) discuss models with Student errors for additive outliers.

In this paper we consider posterior mode filters and smoothers as an alternative or supplementary tool that avoids numerical or Monte Carlo integration. Computational solutions can be based on well understood, efficient algorithms for nonlinear maximization problems. This approach leads to Gauss-Newton or Fisher scoring smoothing algorithms which maximize posterior densities or, equivalently, a certain penalized likelihood criterion, by modifying and extending arguments in Fahrmeir and Kaufmann (1991). Alternatively, these algorithms can we written as iteratively weighted Kalman filters and smoothers applied to working observations in a similar way as for dynamic generalized linear models (compare Fahrmeir and Tutz, 1994, ch.8; Fahrmeir and Wagenpfeil, 1997). For models with heavy-tailed observation error distribution we obtain filters and smoothers that are robust against additive outliers. Innovations outliers, leading for instance to distinct changes in level or slope of a time series, can be modelled by heavy-tailed error distributions in the transition equation. Resulting smoothers are 'edge preserving', that is they react quite flexibly to change points or edges, but still provide smooth fits in other regions.

Our approach is useful for a large class of heavy-tailed error distributions but special emphasis is on the Student family. This concerns, in particular, estimation of unknown hyperparameters such as scale factors or degrees of freedom. We suggest an EM-type algorithm that is tailored to the Student family and can be combined with smoothing algorithms for joint estimation of state and hyperparameters. We illustrate performance by some simulation experiments and by application to real data in Section 5.

2 Robust state space models

For simplicity we will consider only the standard linear state space model for univariate observations. However, extensions to more complex models for instance nonlinear models and multivariate observations are obvious. The model consists of a linear *observation equation*

$$y_t = z_t'\beta_t + \epsilon_t \qquad (t = 1, 2, \ldots) \tag{2.1}$$

for the observations $y_1, y_2, \ldots$ given the states $\beta_1, \beta_2, \ldots$, which is supplemented by a linear *transition equation*

$$\begin{aligned} \beta_t &= F_t\beta_{t-1} + v_t \qquad (t = 1, 2, \ldots) \\ \beta_0 &= a_0 + v_0. \end{aligned} \tag{2.2}$$

The design vectors $z_1, z_2, \ldots$ and the transition matrices $F_1, F_2, \ldots$ as well as the vector a_0 are nonrandom.

The errors ϵ_t, v_t, $t \geq 1$, and v_0 are assumed to have zero mean densities f, g and g_0, which are twice piecewise differentiable. Furthermore errors are mutually independent. If these densities are normal, we have the common linear Gaussian state space model. We say that (2.1) and (2.2) form a *robust state space model* if at least one of the densities f or g is heavy-tailed. Models for *additive outliers* (AO), where the observation densities f are heavy-tailed while g and g_0 are Gaussian, form an important subclass. However, we can also deal with *innovations outliers* (IO) by choice of heavy-tailed densities g for the errors v_t in the transition equation (2.2). Such IO robust state models are quite useful for fitting time series with change points, for instance sudden shifts of level or slope. Resulting filters or smoothers are 'edge preserving': they provide smooth fits for regions with only small local variation but do not blur edges or change points.

Well-known univariate examples with heavy-tailed densities are the Cauchy distribution, the logistic distribution, discrete mixtures of normals, the Student family, or the Huber family. Multivariate distributions can be handled as either generated by independent univariate variables or e.g. as a

multivariate t-distribution discussed by Meinhold and Singpurwalla (1989) and Lange, Little and Taylor (1989). However, as pointed out by Meinhold and Singpurwalla (1989, Appendix 1) there may be serious problems concerning estimation of the dispersion parameter. Our focus will be on the Student family, in particular concerning estimation of hyperparameters. Large parts of the development are valid more generally, however.

For derivations and formulations of filters and smoothers it is convenient to introduce (negative) log-densities, first derivatives (influence or score function) and second derivatives (random information):

$$\begin{aligned} \rho(z) &= -\log f(z)\ ,\ \psi(z) = \partial\rho(z)/\partial z, \\ \Psi(z) &= \partial^2\rho(z)/\partial z\partial z' = \partial\psi(z)/\partial z', \end{aligned}$$

$$\begin{aligned} r(z) &= -\log g(z)\ ,\ c(z) = \partial r(z)/\partial z, \\ C(z) &= \partial^2 r(z)/\partial z\partial z' = \partial c(z)/\partial z', \end{aligned}$$

and r_0, c_0, C_0 defined analogously. To ensure positive definiteness, it may be necessary to consider expected information $\mathrm{E}(\Psi(z))$, $\mathrm{E}(C(z))$ instead of $\Psi(z), C(z)$.We will use Ψ and C as generic symbols for observed and expected second derivatives. For the t-distribution with scale factor σ and ν degrees of freedom, the density is up to a normalizing constant

$$f(z) = \left(1 + z^2/\nu\sigma^2\right)^{-(\nu+1)/2} \quad , \quad \nu, \sigma > 0 \quad . \tag{2.3}$$

Score function and random information are given by

$$\begin{aligned} \psi(z) &= \frac{\nu+1}{\nu + z^2/\sigma^2}\sigma^{-2}z \\ \Psi(z) &= \frac{\nu+1}{\left(\nu + z^2/\sigma^2\right)^2}\sigma^{-2}(\nu - z^2/\sigma^2) \end{aligned}$$

and the expected information is (see Lange, Little and Taylor, 1989):

$$\mathrm{E}\Psi(z) = \frac{\nu+1}{\nu+3}\sigma^{-2}.$$

Throughout the paper we assume that design vectors z_t and transition matrices are known. However, unknown hyperparameters of the densities f

and g, for instance the scale factor σ and the degrees of freedom ν of the t-distribution, have to be estimated in most practical applications along with the sequence of unknown states. A number of data driven methods for choosing hyperparameters are conceivable, for instance simple heuristic methods as in ACM-type smoothing (Martin and Yohai, 1985) or cross-validation. We develop an EM-type algorithm that combines suggestions of Lange, Little and Taylor (1989) for static robust regression and of Fahrmeir (1992) for dynamic generalized linear models.

3 Penalized likelihood estimation

For the following let $y = (y_1, \ldots, y_T)'$, $\beta = (\beta_0', \beta_1', \ldots, \beta_T')'$ denote the whole vector of observations or parameters up to time T. Smoothing is based on the posterior density $p(\beta|y)$. Fully Bayesian methods based on MCMC simulation have been developed recently to tackle this problem, see for instance Shephard and Pitt (1997) and Carter and Kohn (1996a, 1996b). As pointed out in the introduction, posterior mode smoothers are still a useful alternative. They are obtained by maximizing $p(\beta|y)$ or, equivalently, $p(y|\beta)p(\beta)$. Taking logarithms and using the model assumptions of Section 2, we obtain the penalized log-likelihood criterion

$$\begin{aligned} pl(\beta) &= \log p(y|\beta) + \log p(\beta) \qquad (3.1) \\ &= \sum_{t=1}^{T} \log f(y_t - z_t'\beta_t) + \log g_0(\beta_0 - a_0) + \sum_{t=1}^{T} \log g(\beta_t - F_t\beta_{t-1}) \end{aligned}$$

With $\rho = -\log g$, $r = -\log f$ and $r_0 = -\log f_0$ defined in Section 2 maximization of (3.1) is equivalent to minimizing

$$\sum_{t=1}^{T} \rho(y_t - z_t'\beta_t) + r_0(\beta_0 - a_0) + \sum_{t=1}^{T} r(\beta_t - F_t\beta_{t-1}), \qquad (3.2)$$

The first term in (3.2) is a robust measure for the distance between data and fit and is familiar from M-estimation in static robust regression. The second term acts as a robust smoothness prior penalizing roughness of the sequence of states. For $\rho(x) = r(x) = x^2$ we get a penalized least squares criterion,

leading to non-robust classical linear Kalman filtering and smoothing, see for instance Fahrmeir and Tutz (1994, Section 8.1).

The following should be noted: We have arrived at the penalized log-likelihood criterion in a Bayesian framework by maximizing the posterior density $p(\beta|y)$. However, we might forget about this Bayesian approach and start directly from (3.1) regarding $\{\beta_t\}$ as a fixed but unknown sequence which has to be estimated subject to smoothness restrictions. Furthermore, we may allow that ρ is not a proper (negative) log-density but any of the ρ-functions as they are popular in robust statistics, leading to posterior M-estimation.

In maximizing (3.1) or minimizing (3.2) the score function

$$u(\beta) = \partial pl(\beta)/\partial\beta$$

and the observed or the expected information matrix

$$U(\beta) = -\partial^2 pl(\beta)/\partial\beta\partial\beta' \quad \text{or} \quad \tilde{U}(\beta) = \mathrm{E}\, U(\beta)$$

are of interest. The score function can be partitioned as $u = (u_0', \ldots, u_t', \ldots, u_T')'$ with $u_t = \partial pl(\beta)/\partial\beta_t$, $t = 0, \ldots, T$. To avoid special formulas for $t = 0$ and $t = T$, we define $z_0 = 0$ and $F_{T+1} = 0$. Straightforward differentiation shows that

$$u_t = z_t'\Psi_t - c_t + F_{t+1}'c_{t+1} \quad (t = 0, \ldots, T) \tag{3.3}$$

where Ψ_t and c_t are the first derivatives of ρ and r evaluated at $y_t - z_t'\beta_t$ and $\beta_t - F_t\beta_{t-1}$. The information matrix is block-tridiagonal,

$$U = \begin{pmatrix} U_{00} & U_{01} & 0 & \cdots & 0 \\ U_{01}' & U_{11} & \ddots & & \vdots \\ 0 & \ddots & \ddots & & 0 \\ \vdots & & & \ddots & U_{T-1,T} \\ 0 & \cdots & 0 & U_{T-1,T}' & U_{TT} \end{pmatrix} \tag{3.4}$$

with

$$\begin{aligned} U_{tt} &= z_t\Psi_t z_t' + C_t + F_{t+1}'C_{t+1}F_{t+1} \\ U_{t-1,t} &= F_t'C_t \end{aligned} \quad (t = 1, \ldots, T) \tag{3.5}$$

where Ψ_t and C_t are (expected) second derivatives of ρ and r evaluated at $y_t - z_t'\beta_t$ and $\beta_t - F_t\beta_{t-1}$.

Setting $r_t := z_t'\Psi_t$, $R_t := z_t\Psi_t z_t'$ and $Q_t^{-1} := C_t$, the expressions (3.3) and (3.5) for first and second derivatives are formally identical to formulas (4.7) and (4.9) for exponential family state space models in Fahrmeir and Kaufmann (1991). Therefore factorization and inversion of the information matrix U and the covariance matrix recursion developed in that paper remain formally identical.

4 Filtering, Smoothing and Estimation of Hyperparameters

In the following we first summarize the resulting Fisher scoring or Gauss-Newton filters and smoothers for given or known hyperparameters.

Gauss-Newton smoother

Initialize: Choose a starting sequence $\beta^0 = ({\beta^0_{0|T}}', {\beta^0_{1|T}}', \ldots, {\beta^0_{t|T}}', \ldots, {\beta^0_{T|T}}')'$, for example by an ACM-type smoother.

Iterate the Gauss-Newton steps $\beta^0 \to \beta^1$:

1. $\gamma_0 = u_0 \quad , \quad \Sigma_{0|0} = C_0^{-1} \quad$, with u_0 and C_0 evaluated at β^0.

2. Compute for $t = 1, \ldots, T$

$$\begin{aligned} \Sigma_{t|t-1} &= F_t \Sigma_{t-1|t-1} F_t' + C_t^{-1} \\ \Sigma_{t|t} &= [\Sigma_{t|t-1}^{-1} + z_t \Psi_t z_t']^{-1} \\ B_t &= \Sigma_{t-1|t-1} F_t' \Sigma_{t|t-1}^{-1} \end{aligned} \tag{4.1}$$

   and u_t by (3.3), all expressions evaluated at β^0. Set $\gamma_t = u_t + B_t'\gamma_{t-1}$.

3. Filter correction: $\beta^1_{T|T} = \beta^0_{T|T} + \Sigma_{T|T}\gamma_T$

4. Smoother corrections: For $t = T, \ldots, 1$

$$\Sigma_{t-1|T} = \Sigma_{t-1|t-1} + B_t(\Sigma_{t|T} - \Sigma_{t|t-1})B_t'$$

$$\beta^1_{t-1|T} \quad = \quad \beta^0_{t-1|T} + B_t(\beta^1_{t|T} - \beta^0_{t|T}) + (\Sigma_{t-1|T} + B_t\Sigma_{t|T}B_t')\gamma_{t-1}.$$

Iterate steps 1.- 4. till convergence to obtain conditional mode smoothers $(\beta'_{0|T}, \ldots, \beta'_{t|T}, \ldots, \beta'_{T|T})'$ together with curvatures $(\Sigma_{0|T}, \ldots, \Sigma_{t|T}, \ldots, \Sigma_{T|T})$ as approximate error covariance matrices.

An equivalent but computationally alternative form for filtering and smoothing are iterative Kalman filters and smoothers applied to working observations. They can be derived along the line of argument in Fahrmeir and Wagenpfeil (1997), but are not presented here.

Up to now we assumed hyperparameters of the error distributions, such as scale factors or degrees of freedom, as known. Estimation of hyperparameters can be based on general concepts such as cross-validation or maximum likelihood. We developed an EM-type algorithm for (approximate) ML estimation. It is tailored to the Student family, using the fact that a t-distributed random variable t can be generated as a mixture $t = x/\sqrt{z/\nu}$ with x as zero-mean normal and the mixture variable z as χ^2-distributed with ν degrees of freedom. Therefore we can treat the states in an approximative EM algorithm together with the mixture variables as missing. E(xpectation)-steps are then analogous to robust regression models (see Lange, Little and Taylor, 1989), but posterior expectations are substituted by posterior modes. Compared to the EM-type algorithm for dynamic generalized linear models (see e.g. Fahrmeir and Tutz, 1994), further Taylor series expansions are necessary. Details are given in the appendix.

Then, the complete algorithm can be summarized as follows:

1. Set hyperparameters $\theta = \theta^{(0)}$.

2. Compute penalized likelihood smoother, with $\theta = \theta^{(0)}$.

3. Compute $\theta^{(1)}$ by EM steps, using (6.4), (6.5) for updating of variances and maximization of (6.6) for degrees of freedom.

4. Set $\theta^{(0)} = \theta^{(1)}$.

Iterate steps 1.- 4. till convergence.

5 Simulations and Applications

To gain experience with practical performance, the smoothing algorithm was applied to a number of simulated and real data. Gauss–Newton smoothing was combined with Fisher scoring by using expected information whenever the observed information matrix was not positive definite. To combine states and parameter estimation a complete Gauss-Newton algorithm and a single EM-type step were alternated until convergence. Subsection 5.1 and 5.2 report on typical simulation results. Real data examples follow in Subsection 5.3.

5.1 Simulation 1: additive outliers

One-dimensional states were computed according to $\beta_t = \sin(2t\pi/60 + 0.3)$, $t = 0, \ldots, 60$ and held fixed throughout 100 simulation runs. Scalar observations were obtained from $y_t = \beta_t + \epsilon_t$, $t = 1, \ldots, 60$ with errors ϵ_t drawn from a t-distribution with 2 d.f. and scale 0.1. Gauss-Newton smoothing estimates $\{\beta_{t|60}\}$ were computed based on a second-order random walk model for AO, i.e.

$$\begin{bmatrix} \beta_t \\ \beta_{t-1} \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_{t-1} \\ \beta_{t-2} \end{bmatrix} + \begin{bmatrix} v_t \\ 0 \end{bmatrix}, \quad y_t = \beta_t + \epsilon_t,$$

with $v_t \sim N(0, q)$ and ϵ_t as t-distributed with unknown d.f. ν and scale σ_ϵ. Since positive definiteness of C_t is required, we set $C_t = \mathrm{diag}(q, 1e^{-15})$. Approximative confidence bands $\{\beta_{t|60} \pm 2 \cdot \sigma_{t|60}\}$ were computed using corresponding diagonal elements $\sigma^2_{t|60}$ of curvatures $\Sigma_{t|60}$.

To illustrate advantages of robust smoothing over linear smoothing under normality assumptions, we pick out run 46, which was the 14th best according to the mean squared error criterion. Results are shown in Figure 1. Gauss-Newton estimates are not affected by the additive outlier at $t = 13$ and

	EM-type for robust smoother		EM for linear smoother	
	Bias	MSE	Bias	MSE
σ_ϵ^2	-0.00700	0.00005	0.06343	0.02834
ν	-1.24588	1.55816	–	–

Table 1: Hyperparameter estimation for simulation 1.

confidence bands are considerably smaller. The EM-type algorithm yielded $q = 0.0004$, $\sigma_\epsilon^2 = 0.0053$ and $\nu = 0.78$. The EM algorithm combined with the linear smoother computed $q = 0.0011$ and $\sigma_\epsilon^2 = 0.05$. Overestimation of σ_ϵ^2 is typical for linear smoothers in the case of AO, compare Table 1.

The boxplots in Figures 2 and 3 show the empirical distributions of Gauss-Newton resp. linear smoothing estimates $\beta_{t|60}^{(i)}$ from simulation runs $i = 1, \ldots, 100$. Points indicate outlying estimates beyond the whiskers which are drawn to the nearest value not beyond one and a half times the inter quartile range. Comparing both figures with respect to bias and, in particular, variability provides clear evidence for MSE superiority of robust smoothing, in agreement with Table 1.

5.2 Simulation 2: innovations outliers

For analyzing IO, we chose

$$\beta_t = \begin{cases} -0.5\ , & t = 0, \ldots, 20 \\ 0.5\ , & t = 21, \ldots, 40 \\ -1.25, & t = 41, \ldots, 50 \\ 0\ , & t = 51, \ldots, 55 \\ -1.25, & t = 56, \ldots, 60 \end{cases}$$

fixed throughout 100 simulations runs and generated scalar observations $y_t \sim N(\beta_t, 0.05)$. Gauss-Newton smoothing estimates were computed assuming a steady state model for IO:

$$\beta_t = \beta_{t-1} + v_t, \quad y_t = \beta_t + \epsilon_t$$

with v_t as t-distributed with unknown κ and q, and $\epsilon_t \sim N(0, \sigma_\epsilon^2)$.

Figure 4 shows run 7 which was no.50 according to the mean squared error criterion. In comparison to the linear smoother under normality assumption the Gauss Newton algorithm is able to track the level shifts quite well and yields smooth estimates in between with smaller confidence bands. The EM-type algorithm yielded $q = 0.0039, \sigma_\epsilon^2 = 0.0677$ and $\kappa = 1.28$. The EM algorithm combined with the linear smoother computed $q = 0.1042$ and $\sigma_\epsilon^2 = 0.0934 - q$ is typically greater than the robust estimate in case of IO. The boxplots in Figures 5 and 6 were constructed in analogy to Simulation 1 and enlighten the behaviour for all 100 simulation runs. They show that dynamic models with robust smoothness priors clearly outperform Gaussian dynamic models in the presence of discontinuities and are promising candidates for edge preserving smoothing.

5.3 Real Data Examples

Penalized likelihood smoothing was applied to the suspended deposit data of Tukey (1977), see also Martin and Raftery (1987), which show an IO in the year 1934 after the foundation of the Federal Deposit Insurance Corporation in the USA. The data and the results are illustrated in Figure 7. Based on a steady state model for IO, the EM type algorithm computed the estimates $q = 218$, $\sigma_\epsilon^2 = 1179$ and $\kappa = 3.67$. Gauss Newton smoothing exhibits the level shift immediately and yields a smooth track before and after the year 1934.

The monthly CP6 sales data (West and Harrison, 1989) shown in Figure 8 contain an AO in December 1955, indicating also a change point, as well as IO in January 1957 and 1958. Assuming again a steady state model for IO penalized likelihood smoothing clearly indicates the level shifts and gives smooth estimates in between, especially almost ignoring the AO. Hyperparameter estimates were $q = 94$, $\sigma_\epsilon^2 = 586$, $\kappa = 2.55$.

6 Conclusion and Outlook

Linear state space models with heavy-tailed error distributions provide a flexible tool for curve estimation in the presence of additive outliers. The proposed penalized likelihood or posterior mode smoothers avoid numerical integration or Monte Carlo techniques and provide a useful alternative or supplement to MCMC simulation. Special emphasis was laid on the Student distribution. For this case, an EM-type algorithm for data driven estimation of unknown scale factors and degrees of freedom has been developed. State space models for innovations outliers lead to robust smoothness priors and to edge preserving smoothing algorithms that can cope with discontinuities or change points in the underlying curve. Extensions to spatial models, in particular for image analysis seem to be promising and will be considered in future research.

Appendix

We assume independent univariate t-distributions for the observation errors ϵ_t and the components $v_{tj},\ j = 1,\ldots,p$ of the errors $v_t = (v_{t1},\ldots,v_{tp})$ of the transition equation. Then $\epsilon_t|u_t \sim N(0,\sigma_\epsilon^2/u_t)$, $v_{tj} \sim N(0,q_j/w_{tj})$ with mixture variables $u_t \sim \chi^2_\nu/\nu$ and $w_{tj} \sim \chi^2_{\kappa_j}/\kappa_j,\ j = 1,\ldots,p$. If we assume, for simplification, starting values a_0, Q_0 to be known, then $\theta = (\sigma_\epsilon^2,\nu,q_1,\kappa_1,\ldots,q_p,\kappa_p)$ is the vector of unknown hyperparameters. Given the current iterate $\theta^{(0)}$, the EM algorithm computes the next iterate $\theta^{(1)}$ by maximizing the posterior expectation of the complete data log–likelihood $\mathrm{E}\left\{\log p(y,u,\beta,w)|y;\theta^{(0)}\right\}$, where y,β,u and w are the vectors of all observations, state vectors and mixture variables respectively. Due to the model assumptions this is equivalent to

$$\mathrm{E}\left\{\log p(y|u,\beta) + \log p(u) + \log p(\beta|w) + \log p(w)\,\middle|y;\theta^{(0)}\right\} \longrightarrow \max_\theta . \quad (6.1)$$

This implies separate maximization problems for the components of θ. Suppressing the index j, we outline the derivation of our EM-type algorithm for the unknown scale factor q and degrees of freedom κ. Omitting constants,

we have to consider the maximization problems

$$S(q) = -\frac{T}{2}\log q - \frac{1}{2q}\sum_{t=1}^{T} \mathrm{E}\left\{\alpha_t^2 w_t | y; \theta^{(0)}\right\} \longrightarrow \max_q \tag{6.2}$$

with α_t as the j–the component of $\beta_t - F_t\beta_{t-1}$, and

$$S(\kappa) = \sum_{t=1}^{T} \mathrm{E}\left\{\log p(w_t) | y; \theta^{(0)}\right\} \longrightarrow \max_\kappa . \tag{6.3}$$

Using iterated conditional expectations, the t-the summand in $S(q)$ can be written as

$$\begin{aligned} \mathrm{E}\left\{\alpha_t^2 w_t \,\middle|\, y; \theta^{(0)}\right\} &= \mathrm{E}\left\{\mathrm{E}(\alpha_t^2 w_t | \beta, y; \eta^{(0)}) \,\middle|\, y; \theta^{(0)}\right\} \\ &= \mathrm{E}\left\{\alpha_t^2 \Delta^{(0)}(\alpha_t) \,\middle|\, y; \theta^{(0)}\right\} \end{aligned}$$

with $\Delta^{(0)}(\alpha_t) = \{(\kappa^{(0)}+1)/(\kappa^{(0)}+\alpha_t^2/q^{(0)})\}$, compare Lange, Little and Taylor (1989, property 3). By a Taylor series expansion around $\hat{\alpha}_t = \mathrm{E}(\alpha_t | y; \theta^{(0)})$ we get

$$\mathrm{E}(\alpha_t^2 w_t | \beta, y; \theta^{(0)}) \approx \hat{\alpha}_t^2 \Delta^{(0)}(\hat{\alpha}_t) - \frac{1}{2}\Lambda^{(0)}(\hat{\alpha}_t)\,\mathrm{var}(\alpha_t | y; \theta^{(0)})$$

with $\Lambda^{(0)}(\hat{\alpha}_t) = \{\kappa^{(0)}(\kappa^{(0)}+1)(\kappa^{(0)} - 3\hat{\alpha}_t^2/q^{(0)})/(\kappa^{(0)} + \hat{\alpha}_t^2/q^{(0)})^3\}$. Setting the first derivative of $S(q)$ to zero, we get

$$q^{(1)} = \frac{1}{T}\sum_{t=1}^{T} \hat{\alpha}_t^2 \Delta^{(0)}\,\mathrm{var}(\alpha_t | y; \theta^{(0)}). \tag{6.4}$$

Approximating posterior expectations and variances by posterior modes and curvatures, available from our smoothing algorithm for given $\theta^{(0)}$, this is an EM-type step for estimating q. Similarly we get the iteration step

$$\sigma_\epsilon^{2(1)} = \frac{1}{T}\sum_{t=1}^{T} \hat{\gamma}_t^2 \Delta^{(0)}(\hat{\gamma}_t) + \Lambda^{(0)}(\hat{\gamma}_t)\, z_t \Sigma_{t|T} z_t' \tag{6.5}$$

for σ_ϵ^2 with $\hat{\gamma}_t = y_t - z_t'\beta_t$ and $\Delta^{(0)}(\hat{\gamma}_t)$ and $\Lambda^{(0)}(\hat{\gamma}_t)$ defined analogous to the above.

Similar approximations are made in the E-step for $\kappa^{(1)}$. Since the w_t are independently $\kappa\chi_\kappa^2$ distributed, we have

$$\begin{aligned} S(\kappa) &= \frac{T\kappa}{2}\log(\frac{\kappa}{2}) - T\ \log(\Gamma(\frac{\kappa}{2})) + (\frac{\kappa}{2} - 1)\ \sum_{t=1}^{T} \mathrm{E}\left\{\log w_t | y; \theta^{(0)}\right\} \\ &\quad -\frac{\kappa}{2}\ \sum_{t=1}^{T} \mathrm{E}\left\{w_t | y; \theta^{(0)}\right\}. \end{aligned} \tag{6.6}$$

Suppressing t, iterated conditional expectations now yield

$$\begin{aligned} \mathrm{E}\left\{w \left| y;\theta^{(0)}\right.\right\} &= \mathrm{E}\left\{\mathrm{E}(w|\beta,y;\theta^{(0)}) \left| y;\theta^{(0)}\right.\right\} \\ &= \mathrm{E}\left\{(\kappa^{(0)}+1)/(\kappa^{(0)}+\alpha^2/q^{(0)}) \left| y;\theta^{(0)}\right.\right\}. \end{aligned}$$

By Taylor series expansion of $h(\alpha) = (\kappa^{(0)}+1)/(\kappa^{(0)}+\alpha^2/q^{(0)})$ around α we obtain

$$\mathrm{E}\left\{w|y;\theta^{(0)}\right\} \approx h(\hat{\alpha}) - (\kappa^{(0)}+1)\frac{(\kappa^{(0)}-3\hat{\alpha}^2/q^{(0)})}{q^{(0)}(\kappa^{(0)}+\hat{\alpha}^2/q^{(0)})^3}\mathrm{var}(\alpha|y;\theta^{(0)}). \quad (6.7)$$

Once more iterating conditional expectations gives

$$\begin{aligned} \mathrm{E}\left\{\log w|y;\theta^{(0)}\right\} &= \mathrm{E}\left\{\mathrm{E}(\log w|\beta,y;\theta^{(0)}) \left| y;\theta^{(0)}\right.\right\} \\ &= \mathrm{E}\left\{\mathrm{E}(\log w|\beta;\theta^{(0)}) \left| y;\theta^{(0)}\right.\right\} \\ &= \mathrm{DG}(\frac{\kappa^{(0)}+1}{2}) - \mathrm{E}\left\{\log(\frac{1}{2}(\kappa^{(0)}+\alpha^2/q^{(0)})) \left| y;\theta^{(0)}\right.\right\}, \end{aligned}$$

where $\mathrm{DG}(\cdot)$ is the Digamma function. The last equation is given by Lange, Little and Taylor (1989). By Taylor series expansion of $g(\alpha) = \log((\kappa^{(0)}+\alpha^2/q^{(0)})/2)$ we obtain the final approximation

$$\mathrm{E}\left\{\log w|y;\theta^{(0)}\right\} \approx \mathrm{DG}(\frac{\kappa^{(0)}+1}{2}) - g(\hat{\alpha}) - \frac{\kappa^{(0)}-\hat{\alpha}^2/q^{(0)}}{q^{(0)}(\kappa^{(0)}+\hat{\alpha}^2/q^{(0)})}\mathrm{var}(\alpha|y;\theta^{(0)}) \quad (6.8)$$

Conditional variances $\mathrm{var}(\alpha|y;\eta^{(0)})$ in (6.7) and (6.8) are again approximated by curvatures and after differentiation of $S(\kappa)$ the next estimate $\kappa^{(1)}$ can be found by a one dimensional search algorithm.

To obtain an estimate for the degrees of freedom ν of the observation error's distribution we can proceed analogously, especially using that u conditional on β *and* y is $\chi^2_{\nu+1}/(\nu+\gamma^2/\sigma^2_\epsilon)$–distributed and therefore $\mathrm{E}\{u|y,\beta\} = (\nu+1)/(\nu+\gamma^2/\sigma^2_\epsilon)$.

References

Carter, C.K., Kohn, R. (1996a). Markov Chain Monte Carlo in conditionally Gaussian State Space Models. *Biometrika*, **83**, 589–601.

Carter, C.K., Kohn, R. (1996b). Robust Bayesian Nonparametric Regression. In: *Statistical Theory and Computational Aspects of Smoothing.* (W.Härdle and M.G.Schimek, eds.) Heidelberg: Physika-Verlag, 128–148.

Durbin, J. and Koopman, S.J. (1997). Monte Carlo Maximum Likelihood Estimation for Non-Gaussian State Space Models. *Biometrika*, **84**, 669–684.

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

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

Fahrmeir, L. and Wagenpfeil, S. (1997). Penalized Likelihood Estimation and Iterative Kalman Smoothing for Non-Gaussian Dynamic Regression Models. *Computational Statistics & Data Analysis.* **24**, 295–320.

Kitagawa, G. (1987). Non-Gaussian State Space Modeling of Non-Stationary Time Series (with Comments). *Journal of the American Statistical Association*, **82**, 1032–1063.

Lange, K.L., Little, R.J.A. and Taylor, J.M.G. (1989). Robust Statistical Modeling Using the t Distribution. *Journal of the American Statistical Association*, **84**, 881–896.

Martin, R.D. (1979). Approximate Conditional-Mean Type Smoothers and Interpolators. In: *Smoothing Techniques for Curve Estimation*, edited by T. Gasser and M. Rosenblatt. Berlin: Springer, 117–143.

Martin, R.D. and Yohai, V.J. (1985). Robustness in Time Series and Estimating ARMA Models. In: *Handbook of Statistics*, edited by E.J. Hannan, P.R. Krishnaiah and M.M. Rao. Amsterdam: Elsevier, **5**, 119–155.

MARTIN, R.D. and RAFTERY, A.E. (1987). Robustness, Computation, and Non-Euclidian Models (Comment). *Journal of the American Statistical Association*, **82**, 1044–1050.

MEINHOLD, R.J. and SINGPURWALLA, N.D. (1989). Robustification of Kalman Filter Models. *Journal of the American Statistical Association*, **84**, 479–486.

SHEPHARD, N. and PITT, M.K.(1997). Likelihood Analysis of non-Gaussian Measurement Time Series. *Biometrika*, **84**, 653–667.

TUKEY, J.W. (1977). *Exploratory Data Analysis.* London: Addison-Wesley.

WEST, M. (1981). Robust Sequential Approximate Bayesian Estimation. *Journal of the Royal Statistical Society*, **B 43**, 157–166.

WEST, M. (1984). Outlier Models and Prior Distributions in Bayesian Linear Regression. *Journal of the Royal Statistical Society*, **46**, 431–439.

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

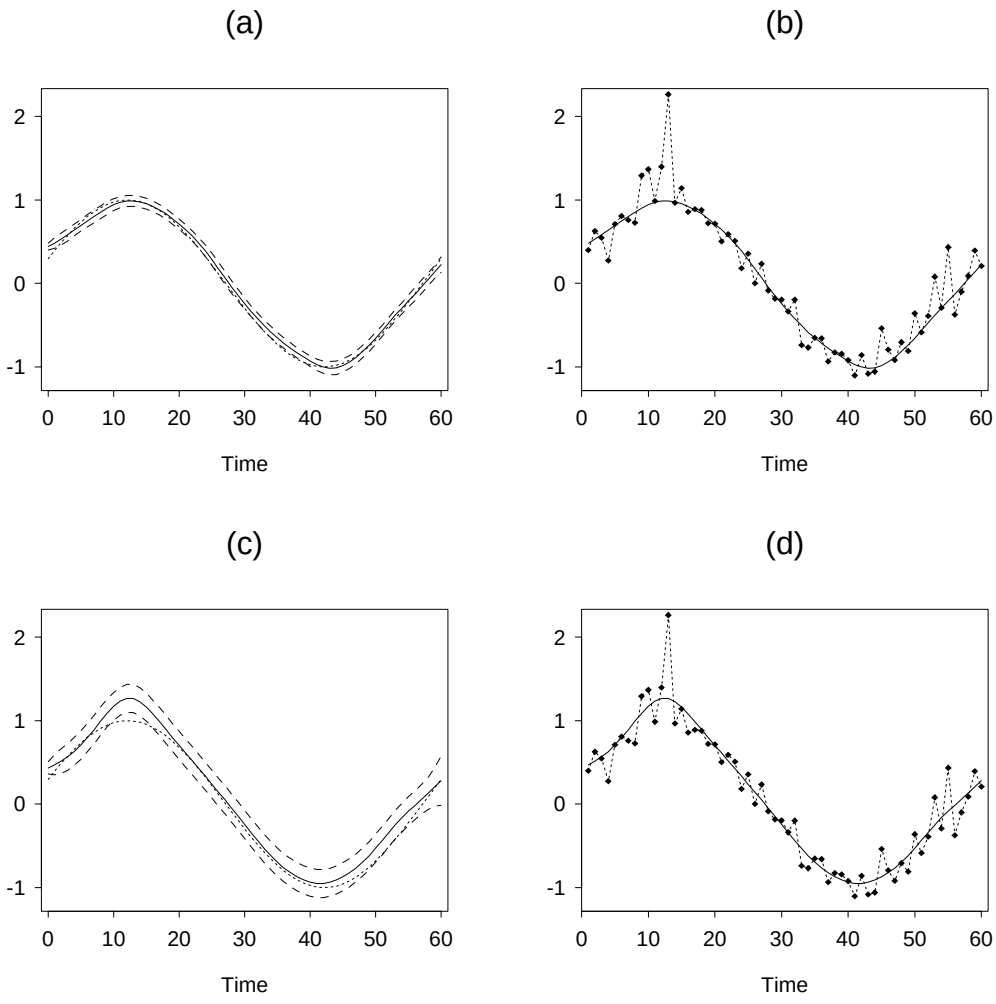

Figure 1: True parameters $\{\beta_t\}$ indicated by (- - -) and smoothing estimates (——) together with naive 2-σ-confidence bands (– –) obtained by the robust smoother (a) and by the linear smoother (c). Observations $\{y_t\}$ indicated by diamonds and fitted values $\{\hat{y}_t\}$ (——) resulting from the robust smoother (b) and from the linear smoother (d).

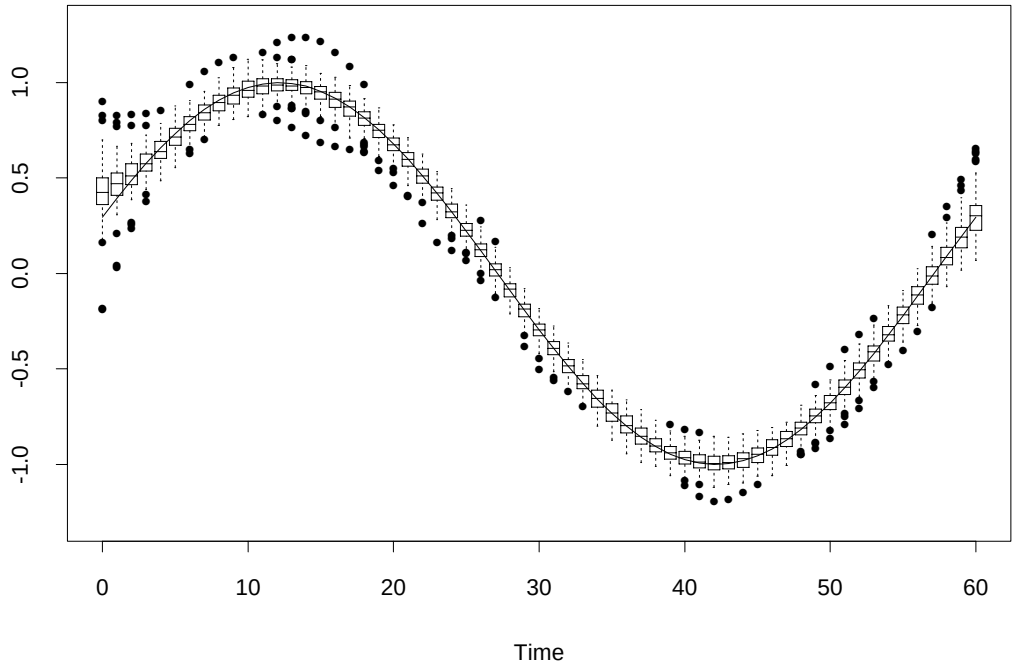

Figure 2: Boxplots visualizing the empirical distribution of Gauss-Newton smoothing estimates for simulation 1. True values $\{\beta_t\}$ indicated by (——).

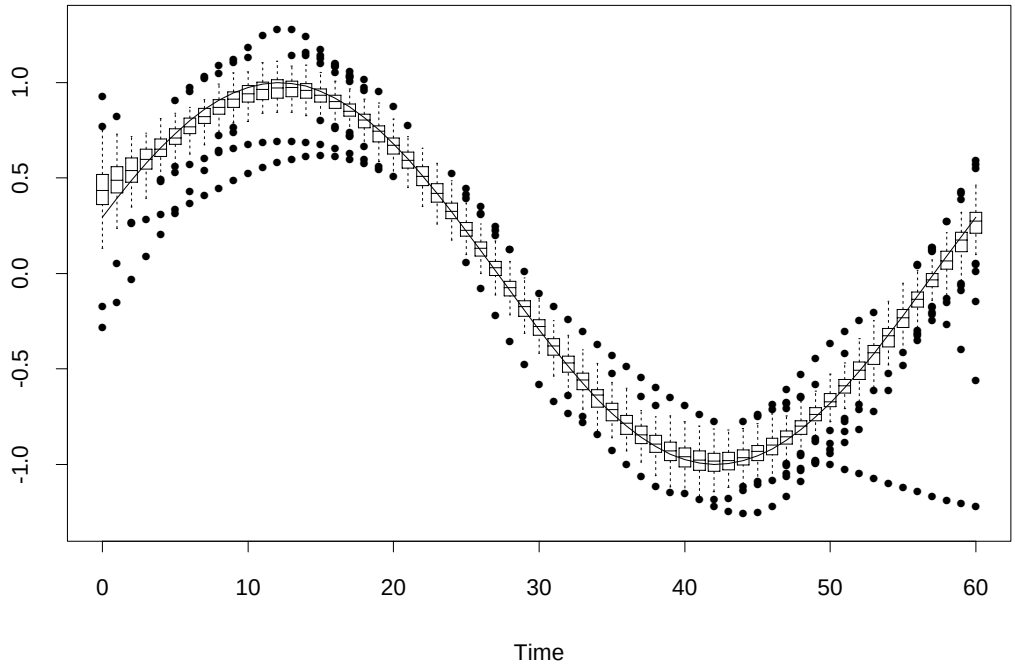

Figure 3: Boxplots visualizing the empirical distribution of linear smoothing estimates for simulation 1. True values $\{\beta_t\}$ indicated by (——).

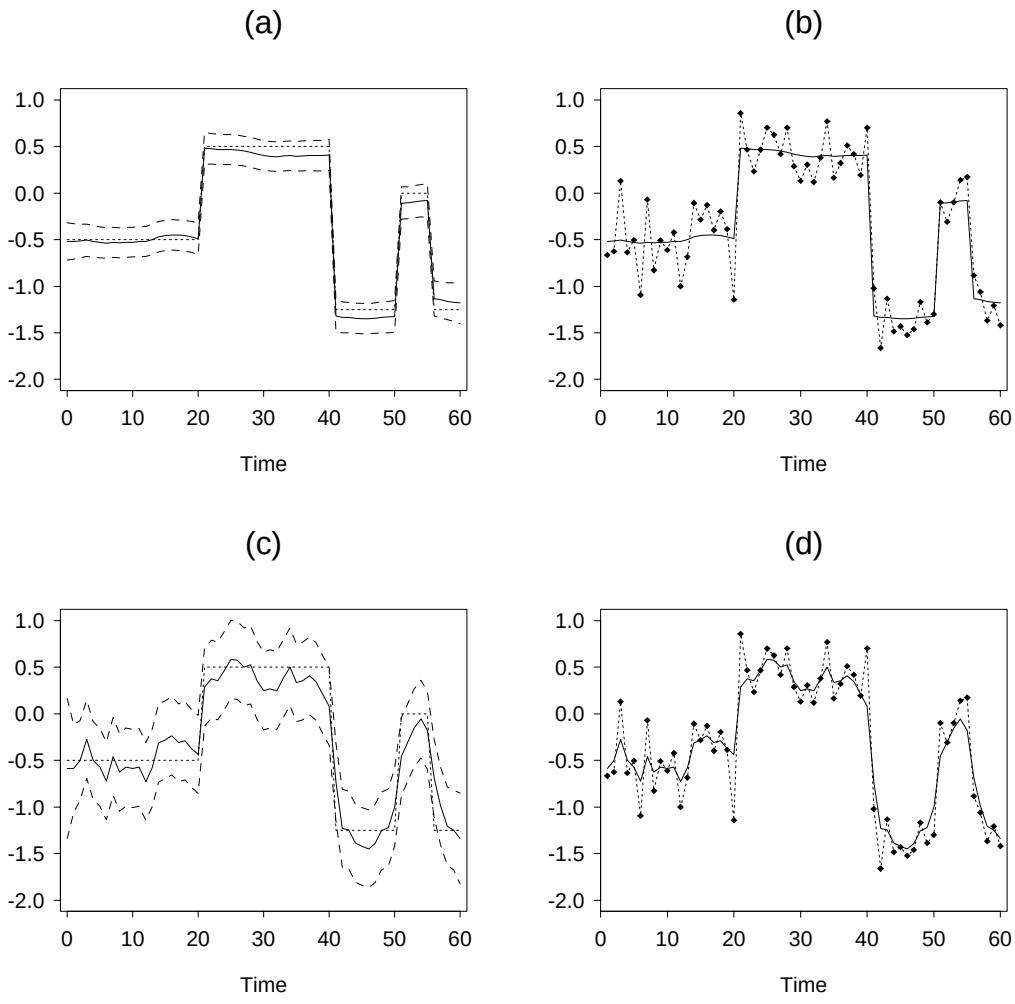

Figure 4: True parameters $\{\beta_t\}$ indicated by (- - -) and smoothing estimates (——) together with naive 2-σ-confidence bands (– –) obtained by the robust smoother (a) and by the linear smoother (c). Observations $\{y_t\}$ indicated by diamonds and fitted values $\{\hat{y}_t\}$ (——) resulting from the robust smoother (b) and from the linear smoother (d).

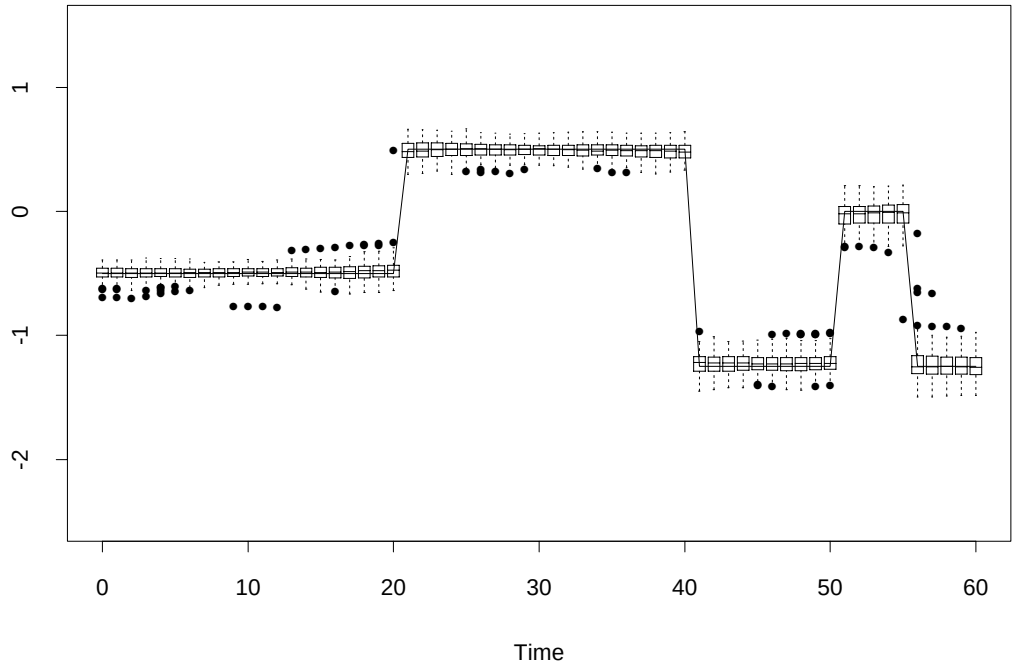

Figure 5: Boxplots visualizing the empirical distribution of Gauss-Newton smoothing estimates for simulation 2. True values $\{\beta_t\}$ indicated by (——).

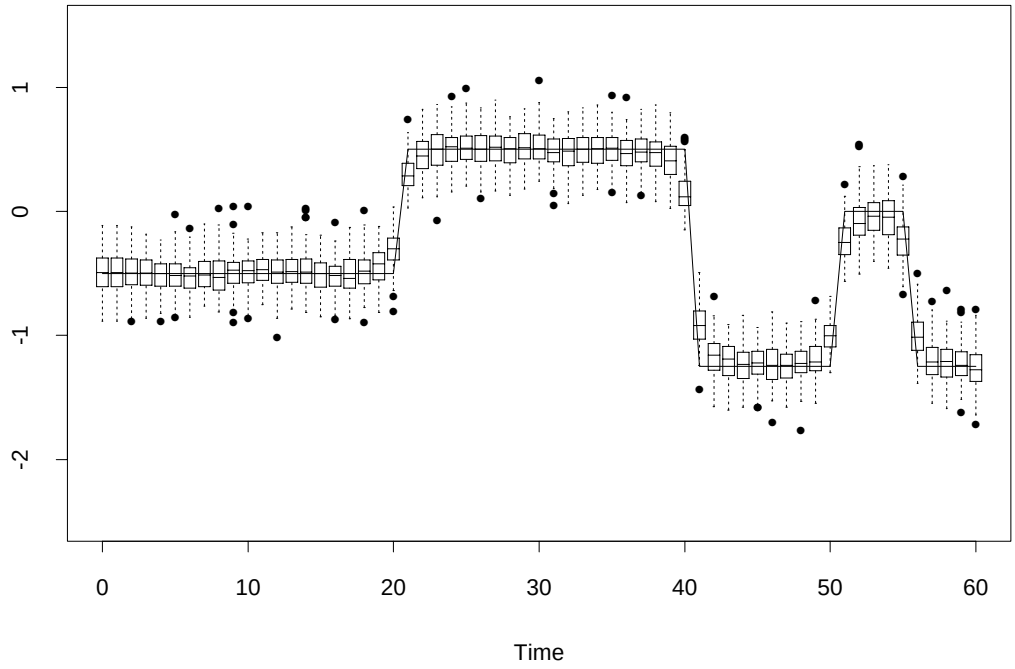

Figure 6: Boxplots visualizing the empirical distribution of linear smoothing estimates for simulation 2. True values $\{\beta_t\}$ indicated by (——).

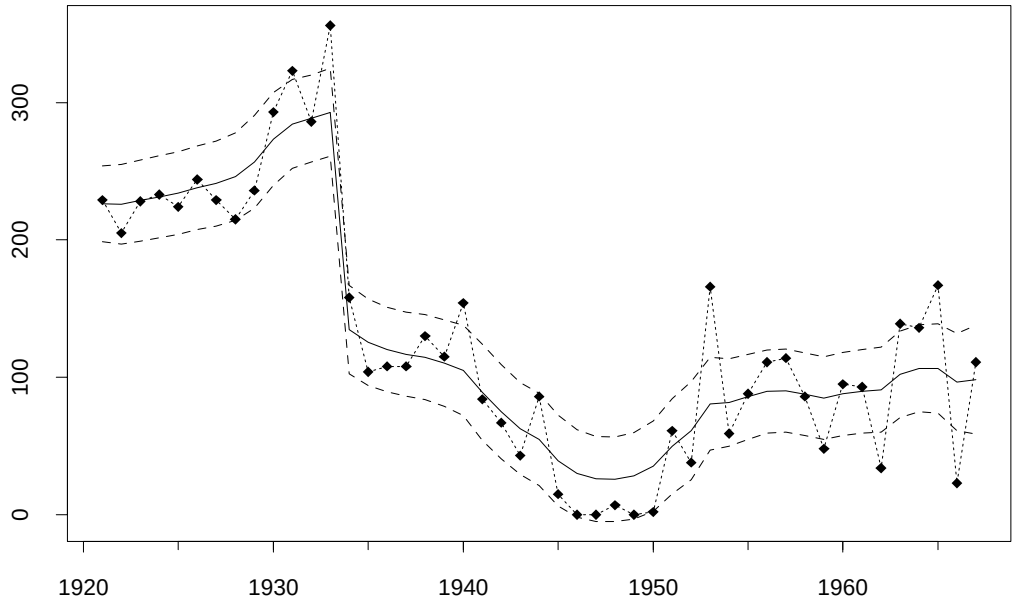

Figure 7: Transformed suspended deposit data – diamonds indicating $100 \log(c + \text{suspended deposits})$ for $c = 1$Mio. Dollar – and robust smoothing estimates (——) with 2-σ-confidence bands (– –).

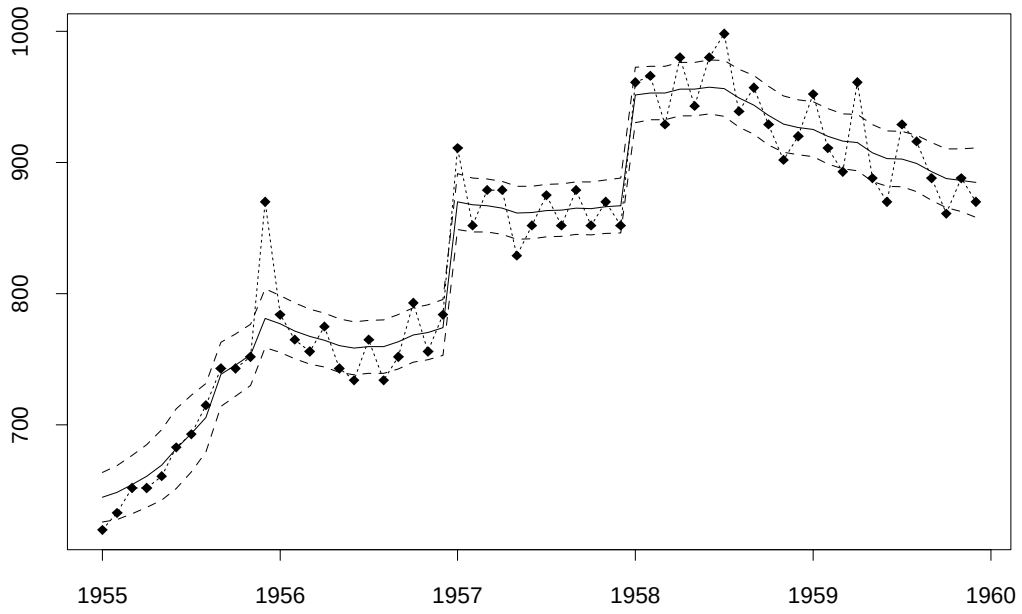

Figure 8: CP6 sales indicated by diamonds and robust smoothing estimates (——) with 2-σ-confidence bands (– –).