Environmental and Ecological Statistics **3**, 291–309 (1996)

Recursive residuals and model diagnostics for normal and non-normal state space models

SYLVIA FRÜHWIRTH-SCHNATTER

Department of Statistics, University of Economics and Business Administration, Vienna, Austria

Received May 1995. Revised February 1996

Model diagnostics for normal and non-normal state space models are based on recursive residuals which are defined from the one-step ahead predictive distribution. Routine calculation of these residuals is discussed in detail. Various diagnostic tools are suggested to check, for example, for wrong observation distributions and for autocorrelation. The paper also discusses such topics as model diagnostics for discrete time series and model discrimination via Bayes factors. The case studies cover environmental applications such as analysing a time series of the number of daily rainfall occurrences and a time series of daily sulfur dioxide emissions.

Keywords: autocorrelation, Bayes factors, Kalman filtering, model discrimination, rainfall occurrences, sulfur dioxide emissions, time series

1. Introduction

Model diagnostics is understood as a more or less formal check of properties that certain residuals should have under the assumption that the data were generated by the model under investigation. Although the statistical literature on diagnostics for more specific models such as regression models or generalized linear models is vast, this issue is somewhat neglected for state space models.

For normal state space models some useful material may be found in the books of Schneider (1986) and Harvey (1989). They both derive appropriate statistics from the recursive residuals

$$r_t = \frac{y_t - \mathrm{E}(y_t|y^{t-1})}{\sqrt{\mathrm{var}(y_t|y^{t-1})}} \tag{1}$$

where $\mathrm{E}(y_t|y^{t-1})$ and $\mathrm{var}(y_t|y^{t-1})$ are the expectation and the variance of the one-step-ahead predictive distribution of a future value y_t of a time series given observations $y^{t-1} = \{y_1, \ldots, y_{t-1}\}$ up to $t-1$. As the residuals $r_1, \ldots, r_t, \ldots$ are i.i.d. standard normal, if the model is correct, it is easy to develop significance tests to decide whether to reject the model.

Harvey (1989) also deals with diagnostics for non-normal state space models, where he sticks to diagnostics based on the recursive residuals (1). These 'Pearson residuals' unfortunately lose their normality within non-normal models, and significance tests based on them are somewhat doubtful. Smith (1985), following Dawid (1984), defines recursive residuals for time series of continuous observations by

$$u_t = \Pr(Y_t \leq y_t|y^{t-1}) \tag{2}$$

and

$$v_t = \Phi^{-1}(u_t) \tag{3}$$

where $\Pr(Y_t \leq y_t|y^{t-1})$ is the one-step-ahead predictive distribution of a future value y_t of a time series given observations up to $t-1$ and Φ is the standard normal distribution. u_t will be called the P-score and v_t will be called the transformed P-score in this paper. What makes the P-scores and their transformed version so useful is the following property: if Y_t is a continous random variable and if the model is correct, the P-scores are i.i.d. uniform on $[0, 1]$ and therefore the transformed P-scores are i.i.d. standard normal (Rosenblatt, 1952). Furthermore for normal state space models the transformed P-scores coincide with the recursive residuals (1). Thus the P-scores and their transformed version seem to be the appropriate extension of recursive residuals to non-normal state space models.

A general method of routine calculation of P-scores for non-normal state space models is still missing – only some special cases have been treated in Smith (1985). In the present paper we discuss computation of P-scores for a rather general class of non-normal state space models, namely the dynamic generalized linear model (DGLM). The DGLM combines the linear Gaussian transition equation for the state vector $\boldsymbol{x}_t$,

$$\boldsymbol{x}_t = \boldsymbol{F}_t \cdot \boldsymbol{x}_{t-1} + \boldsymbol{w}_t, \qquad \boldsymbol{w}_t \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{Q}_t) \tag{4}$$

with the usually non-normal observation density $p(y_t|\boldsymbol{x}_t) = p(y_t|\lambda_t)$ which depends on the state vector $\boldsymbol{x}_t$ only through the linear predictor $\lambda_t = \boldsymbol{H}_t\boldsymbol{x}_t$ (by 'density' we mean the density of the distribution $P(y_t|\lambda_t)$ either with respect to the Lebesgue measure for y_t continuous or with respect to the counting measure for y_t discrete). The observation density need not belong to an exponential family. For more details the reader is referred to Fahrmeir (1992). For the sake of simplicity we confine ourselves to univariate time series.

Although we adopt a Bayesian approach to derive the predictive distributions through which the P-scores are defined, we do not hesitate to use sampling theory methods to derive diagnostic tools from the P-scores and to interpret them. From a dogmatic Bayesian viewpoint sampling theory methods in diagnostic checking are thought inappropriate (O'Hagan, 1980). Box (1980), however, in a rather convincing paper, has suggested accepting a Bayes/non-Bayes interplay in the dual processes of model estimation and model criticism. It is the intention of the present paper to illustrate how very useful such a Bayes/non-Bayes marriage turns out to be when dealing with state space models.

The outline of the paper is as follows: Section 2 shows how to supplement routine methods of filtering for DGLM (West *et al.*, 1985; Fahrmeir, 1992; Frühwirth-Schnatter, 1994) by a systematic calculation of P-scores. In Section 3 we derive various diagnostic tools from the P-scores. Section 4 deals with model diagnostics for time series of binary and count data. Section 5 discusses the difference between model choice via Bayes factors and model diagnostics. Finally, in Section 6 we illustrate iterative model building based on residual diagnostics for a time series of daily sulfur dioxide emissions.

2. Approximate computation of recursive residuals for DGLM

The recursive residuals (2) and (3) are defined through the one-step-ahead predictive distribution $\Pr(Y_t \leq y_t|y^{t-1})$. This distribution is known analytically only if the observation density $p(y_t|\lambda_t)$ happens to be $\mathrm{N}(\lambda_t, R_t)$ with R_t independent of λ_t and if all hyperparameters are known. In this

section we discuss how to approximate the one-step-ahead predictive distribution for DGLM with possibly unknown hyperparameters. y_t may be continuous or discrete.

Let us assume for the moment that the hyperparameters are known. We represent the one-step-ahead predictive distribution as an infinite mixture:

$$\Pr(Y_t \leq y_t|y^{t-1}) = \int_{-\infty}^{\infty} P(y_t|\lambda_t)p(\lambda_t|y^{t-1})\,\mathrm{d}\,\lambda_t \tag{5}$$

where $P(y_t|\lambda_t)$ is the distribution function of $p(y_t|\lambda_t)$ and $p(\lambda_t|y^{t-1})$ is the one-step-ahead predictive density of the linear predictor $\lambda_t = \boldsymbol{H}_t \cdot \boldsymbol{x}_t$. $p(\lambda_t|y^{t-1})$ as well as the posterior $p(\boldsymbol{x}_{t-1}|y^{t-1})$ are not known analytically. Routine filtering methods such as the WHM algorithm (West *et al.*, 1985), posterior mode filtering (Fahrmeir, 1992) or integration-based Kalman filtering (Frühwirth-Schnatter, 1994) lead to estimates $\hat{\boldsymbol{x}}_{t-1|t-1}$ and $\boldsymbol{P}_{t-1|t-1}$ of the first two moments of the posterior $p(\boldsymbol{x}_{t-1}|y^{t-1})$. It is obvious from the linear transition Equation (4) and the linear relationship between state vector and predictor that the first two moments $\hat{\lambda}_{t|t-1}$ and $\hat{\Lambda}_{t|t-1}$ of $p(\lambda_t|y^{t-1})$ are given by:

$$\hat{\lambda}_{t|t-1} = \boldsymbol{H}_t \cdot \boldsymbol{F}_t \cdot \hat{\boldsymbol{x}}_{t-1|t-1} \qquad \hat{\Lambda}_{t|t-1} = \boldsymbol{H}_t(\boldsymbol{F}_t \cdot \boldsymbol{P}_{t-1|t-1}\boldsymbol{F}_t^{\mathrm{T}} + \boldsymbol{Q}_t)\boldsymbol{H}_t^{\mathrm{T}}$$

To approximate the infinite mixture distribution (5) by a finite mixture distribution we substitute the exact, but unknown mixing density $p(\lambda_t|y^{t-1})$ by a normal density with the same first two moments and use Gauss–Hermite integration after applying the transformation $z_t = (2 \cdot \hat{\Lambda}_{t|t-1})^{-1/2}(\lambda_t - \hat{\lambda}_{t|t-1})$:

$$\Pr(Y_t \leq y_t|y^{t-1}) \approx \frac{1}{\sqrt{\pi}} \sum_{i=1}^{M} P(y_t|\lambda_t^{(i)})\omega_M^{(i)} \tag{6}$$

with

$$\lambda_t^{(i)} = \hat{\lambda}_{t|t-1} + \sqrt{2 \cdot \hat{\Lambda}_{t|t-1}} \cdot \tau_M^{(i)} \tag{7}$$

where $\omega_M^{(i)}$ and $\tau_M^{(i)}$ are the grid points and the weights of univariate Gauss–Hermite integration of order M as tabulated for example in Abramowitz and Stegun (1970). Although the approximation of the predictive distribution by means of (6) may be combined with any filtering method which computes the first two posterior moments, we use integration-based Kalman filtering for the rest of the paper (with the exception of Case Study 1).

The finite mixture distribution (6) is approximate in two senses. First, it approximates an analytical integral by a numerical one. The corresponding error can be kept small by choosing M sufficiently large, e.g. $M = 10$. Second, it substitutes the exact mixing density by a normal density. This error will be small if the posterior $p(\boldsymbol{x}_{t-1}|y^{t-1})$ is close to a normal density. Experimental results reported in Schnatter (1992) and Fahrmeir (1992) indicate that with t increasing the posterior in fact tends to be normal even in cases where the observation density is extremely non-normal.

If hyperparameters $\boldsymbol{\theta}$ are unknown, one may proceed in two ways (Dawid, 1984). The first method is to assume that $\boldsymbol{\theta}$ is unknown but fixed ($\boldsymbol{\theta} = \boldsymbol{\theta}_0$) and to use the plug-in approach of estimating $\boldsymbol{\theta}_0$, e.g. by the ML estimate $\hat{\boldsymbol{\theta}}$. An approximation of the likelihood function results automatically as a by-product of integration-based Kalman filtering – see Frühwirth-Schnatter (1994).

The second method is to assume that $\boldsymbol{\theta}$ is a random variable and to use a hierarchical model with prior $p(\boldsymbol{\theta}|y^0)$. The exact P-scores u_t which are given by the infinite mixture

$$u_t = \int \Pr(Y_t \leq y_t|y^{t-1}, \boldsymbol{\theta})p(\boldsymbol{\theta}|y^{t-1})\,\mathrm{d}\,\boldsymbol{\theta} \tag{8}$$

are approximated via multiprocess-filtering by a finite mixture:

$$u_t \approx \sum_{j=1}^{G} \Pr(Y_t \leq y_t | y^{t-1}, \theta^{(j)}) p(\theta^{(j)} | y^{t-1}) \tag{9}$$

The posterior weights of the grid points $\theta^{(1)}, \ldots, \theta^{(G)}$ are determined from Bayes' theorem:

$$p(\theta^{(j)} | y^t) \propto p(y_t | y^{t-1}, \theta^{(j)}) p(\theta^{(j)} | y^{t-1}) \tag{10}$$

An approximation of the 'likelihood' $p(y_t | y^{t-1}, \theta^{(j)})$ is directly available from the approximation (6) of the predictive distribution $\Pr(Y_t \leq y_t | y^{t-1}, \theta^{(j)})$:

$$p(y_t | y^{t-1}, \theta^{(j)}) \approx \frac{1}{(\sqrt{\pi})} \sum_{i=1}^{M} p(y_t | \lambda_t^{(i)}) \omega_M^{(i)} \tag{11}$$

where $\lambda_t^{(i)}$ is the same as in (7). It should be kept in mind that the notation in (11) does not reflect the fact that $\hat{\lambda}_{t|t-1}$ and $\hat{\Lambda}_{t|t-1}$ in (7) depend on the grid point $\theta^{(j)}$. For further details on estimating hyperparameters via multiprocess filtering the reader is referred to Harrison and Stevens (1976) for normal state models and to Frühwirth-Schnatter (1994) for non-normal state space models.

The statistical properties of the P-scores under the assumption of a correct model differ for both methods. Given a consistent estimate $\hat{\theta}$ of θ_0 the 'plug-in' P-scores approximated from $\Pr(Y_t \leq y_t | y^{t-1}, \hat{\theta})$ rather than from $\Pr(Y_t \leq y_t | y^{t-1}, \theta_0)$ are i.i.d. uniform only asymptotically, even if the model is correct (Dawid, 1984). Advantageously, the exact 'hierarchical' P-scores (8) are i.i.d. uniform, if the model is correct. The approximate 'hierarchical' P-scores (9) will be close to an i.i.d. uniform sequence, if G is large. In contrast to the plug-in-approach, however, not only the model structure given by $\boldsymbol{F}_t, \boldsymbol{Q}_t, \boldsymbol{H}_t$ and the conditional observation density $p(y_t | \lambda_t)$ but also the prior $p(\theta | y^0)$ define the hierarchical model. A model with correct model structure and correct conditional observation density may be rejected simply because this prior has been poorly chosen.

A last remark concerns the influence of the prior $p(\boldsymbol{x}_0 | y^0)$ of the state vector $\boldsymbol{x}_0$. The first d P-scores $u_1, \ldots, u_d$, with $d = \dim(\boldsymbol{x}_0)$, are well defined only if $p(\boldsymbol{x}_0 | y^0)$ is a proper density. If the prior of the state vector is improper, the first d observations are needed to build up a proper prior for all components of the state vector. It is at $n_0 = d + 1$ that computation of the P-scores starts.

3. Basic tools of diagnostic checking for state space models

From our practical experience with analysing numerous time series by state space models we have learnt that the following diagnostic tools are 'standard tools' which should be plotted and computed in any case. Some of them are well-known diagnostic tools for regression models, generalized linear models, and normal state space models.

The most simple graphic device is a *plot of the transformed P-scores* v_t as a function of t. Some types of departure from the assumed model such as the presence of outliers, autocorrelation or heterogeneity might be obvious from a visual inspection of this graph. A rough check of the distribution of the P-scores is provided by the *empirical distribution function* of u_t. It is often more instructive to produce a *normal plot of the transformed P-scores*: a plot of the ordered $v_{(t)}$'s against normal order statistics should be close to a straight line. Such a plot, however based on the deviance residuals (Pregibon, 1981), has been discussed by Davison and Gigli (1989) to check outliers and distributional assumptions in generalized linear models.

A difficulty in the inspection of normal plots is sometimes to decide whether the variation in the plot is too far from a straight line. Simulation envelopes (Atkinson, 1981, 1982) could be generated

to provide a statistical test. In the present paper we use the following indices derived from the first four moments of these residuals about their mean:

$$m_r = \sum_{t=n_0}^{n} (v_t - m_1)^r / N, \qquad r \geq 2$$

$$m_1 = \sum_{t=n_0}^{n} v_t / N, \qquad N = n - n_0 + 1$$

where n is equal to the number of observations and n_0 is equal to 1, if the prior of the state vector x_0 is a proper density. For improper priors see the remark at the end of Section 2.

These moments are used to construct four indices which are asymptotically standard normal. The *bias index* B_N is defined by

$$B_N = \sqrt{N} m_1$$

the *dispersion index* D_N by

$$D_N = \frac{N m_2 - N + 1}{\sqrt{2(N-1)}}$$

the *skewness index* S_N by

$$S_N = \sqrt{\frac{(N+1)(N+3)}{6(N-2)}} \frac{m_3}{(m_2)^{3/2}}$$

and the *tail index* T_N by

$$T_N = \frac{(N+1)\sqrt{(N+3)(N+5)}}{\sqrt{24(N-2)(N-3)N}} \left(\frac{m_4}{m_2^2} - \frac{3(N-1)}{(N+1)} \right)$$

B_N obviously is standard normal, if the model is correct. D_N, S_N and T_N are standard normal asymptotically, if the model is correct (see Appendix).

We refer to these four indices as the first four moment indices. We use these indices in a more or less explorative way by looking for 'surprising values'. As 'surprisingly low' and 'surprisingly high' we qualify indices which, for N not too small, are smaller than say -2 and bigger than say 2, respectively. For N small we compare these indices with the lower and upper quantiles of their exact distribution under a correct model to get an idea of what 'surprisingly low' and 'surprisingly high' means. These quantiles may be also used to reject the model at a given significance level.

The quantiles $D_{N,\alpha}$ of the exact distribution of the dispersion index D_N under the assumption of a correct model may be derived for each N from $D_{N,\alpha} = (\chi^2_{N-1,\alpha} - (N-1))/\sqrt{2(N-1)}$ where the quantile $\chi^2_{N-1,\alpha}$ of the χ^2_{N-1}-distribution is computed by the approximation of Wilson and Hilferty. The quantiles $S_{N,\alpha}$ and $T_{N,\alpha}$ of the exact distribution of S_N and T_N may be estimated by empirical quantiles obtained from Monte Carlo simulation. Such simulation experiments can be carried out by the algorithm suggested in Hatzinger and Panny (1993). In Table 1 we report the results from such a simulation experiment. As the quantiles are insensitive to minor changes of the sample size N, we report the quantiles for typical sample sizes only.

If the bias index B_N is surprisingly high or low, the observations tend to be bigger or smaller than predicted. If the dispersion index D_N is surprisingly high or low, the observations are overdispersed and underdispersed, respectively. The skewness and the tail index are valuable tools when checking

Table 1. Quantiles of the exact distribution of the various indices under the assumption of a correct model ($S_{N,\alpha}$, $T_{N,\alpha}$, $J_{N,\alpha}$, and $A_{N,\alpha}$ are empirical quantiles from a Monte Carlo sample of size 3×10^6)

		α							
Quantiles	N	0.01	0.02	0.025	0.05	0.95	0.975	0.98	0.99
B_α		−2.33	−2.05	−1.96	−1.65	1.65	1.96	2.05	2.33
$D_{N,\alpha}$	60	−2.06	−1.85	−1.78	−1.53	1.74	2.13	2.25	2.59
	80	−2.09	−1.88	−1.81	−1.55	1.73	2.11	2.22	2.56
	100	−2.12	−1.90	−1.82	−1.56	1.72	2.09	2.20	2.53
	120	−2.14	−1.91	−1.84	−1.57	1.72	2.08	2.19	2.52
	150	−2.16	−1.93	−1.85	−1.58	1.71	2.07	2.18	2.50
	200	−2.18	−1.95	−1.86	−1.59	1.70	2.05	2.16	2.47
	400	−2.22	−1.98	−1.89	−1.60	1.68	2.03	2.13	2.43
$S_{N,\alpha}$	60	−2.40	−2.08	−1.98	−1.64	1.64	1.98	2.08	2.40
	80	−2.39	−2.08	−1.97	−1.64	1.64	1.98	2.08	2.39
	100	−2.38	−2.08	−1.97	−1.64	1.64	1.97	2.08	2.38
	120	−2.37	−2.07	−1.97	−1.64	1.64	1.97	2.07	2.37
	150	−2.37	−2.07	−1.97	−1.64	1.64	1.97	2.07	2.37
	200	−2.36	−2.07	−1.97	−1.64	1.64	1.97	2.07	2.36
	400	−2.34	−2.06	−1.97	−1.64	1.64	1.96	2.06	2.35
$T_{N,\alpha}$	60	−1.56	−1.44	−1.40	−1.25	1.84	2.44	2.64	3.29
	80	−1.61	−1.48	−1.44	−1.28	1.83	2.41	2.60	3.22
	100	−1.66	−1.52	−1.47	−1.31	1.83	2.39	2.57	3.16
	120	−1.69	−1.55	−1.50	−1.33	1.82	2.37	2.55	3.12
	150	−1.73	−1.59	−1.53	−1.35	1.81	2.34	2.52	3.06
	200	−1.79	−1.63	−1.57	−1.38	1.80	2.31	2.47	2.99
	400	−1.91	−1.73	−1.66	−1.45	1.78	2.24	2.39	2.83
$J_{N,\alpha}$	60	–	–	–	–	6.49	9.92	11.24	16.12
	80	–	–	–	–	6.39	9.60	10.83	15.32
	100	–	–	–	–	6.32	9.36	10.51	14.75
	120	–	–	–	–	6.26	9.14	10.25	14.26
	150	–	–	–	–	6.20	8.94	9.97	13.68
	200	–	–	–	–	6.14	8.66	9.61	13.00
	400	–	–	–	–	6.06	8.13	8.90	11.60
$A_{N,\alpha}$	60	−2.34	−2.04	−1.94	−1.60	1.60	1.93	2.03	2.33
	80	−2.34	−2.04	−1.94	−1.61	1.61	1.94	2.04	2.33
	100	−2.34	−2.04	−1.94	−1.62	1.62	1.94	2.04	2.33
	120	−2.33	−2.04	−1.95	−1.62	1.62	1.95	2.04	2.33
	150	−2.33	−2.04	−1.95	−1.63	1.63	1.95	2.04	2.33
	200	−2.33	−2.05	−1.95	−1.63	1.63	1.95	2.05	2.33
	400	−2.33	−2.05	−1.95	−1.64	1.64	1.96	2.05	2.33

the observation distribution. The observations tend to be skewer to the right or skewer to the left than assumed by the model, if the skewness index S_N is either surprisingly high or surprisingly low. The observations have longer tails than assumed by the model, if the tail index T_N is surprisingly high, and the observations have shorter tails than assumed by the model, if T_N is surprisingly low. The higher moment indices, however, should be interpreted only together with the normal plot since they are sensitive to outliers which are easy to recognize in the normal plot.

Instead of judging the skewness and the tail index individually one could construct a *joint index* J_N

by summing their squares:

$$J_N = S_N^2 + T_N^2 \tag{12}$$

and look for 'surprisingly high' values. J_N is asympotically equivalent to well-known tests for normality (Bowman and Shenton, 1975; Pearson *et al.*, 1977; Harvey, 1989, p. 260) which similarly make joint use of the sample moment ratio statistics and are known to have an asymptotic χ^2-distribution with two degrees of freedom. For N small J_N is compared with empirical upper quantiles (see Table 1).

For a check for serial correlation we use the *empirical autocorrelogram of the transformed P-scores* which is compared with the asymptotic confidence band $[-1/(N-1) - 2/\sqrt{N},\ -1/(N-1) + 2/\sqrt{N}]$ (see for example Chatfield, 1989). Diagnostic statistics may be defined for example by the *AC(1)-index* A_N,

$$A_N = \sqrt{N}\left(\rho_1 + \frac{1}{N-1}\right)$$

where ρ_1 is the first-order empirical autocorrelation coefficient or by some portmanteau statistic comparable to the Box–Pierce statistic. A_N is known to be standard normal asymptotically. A_N is used in the same explorative way as the first four moment indices.

4. Diagnostics for time series of binary and count data

Time series of binary or count data are such that $p(y_t|\lambda_t)$ is a discrete distribution on the integers (sometimes including 0). It is quite common to use the Pearson residuals for diagnostics even of binary data (see for example Harvey and Fernandez, 1989; Aitkin *et al.*, 1989, p. 171). As the distribution of the Pearson residuals is highly non-normal for binary and small count data they can not be used to check the correctness of the observation distribution. This drawback is avoided by the use of P-scores.

For discrete distributions the residuals u_t defined by (2) directly are not i.i.d. uniform but follow a discrete distribution on [0,1]. Residuals which are distributed uniformly on [0,1] are obtained from the predictive distribution via randomization (see also Smith, 1985). Let α_t be a sequence of i.i.d. uniform random variables and let y_t be the actual observation. For discrete observations the P-score u_t is then defined by the random interpolation

$$u_t := (1-\alpha_t)\Pr(Y_t \le (y_t - 1)|y^{t-1}) + \alpha_t \Pr(Y_t \le y_t|y^{t-1}) \tag{13}$$

These P-scores actually are i.i.d. uniform on [0,1].

If the hyperparameters θ are unknown then again the predictive distribution $\Pr(Y_t \le y_t|y^{t-1})$ may be substituted either by the conditional distribution $\Pr(Y_t \le y_t|y^{t-1}, \hat{\theta})$ or by a finite mixture as in (9):

$$u_t \approx (1-\alpha_t)\sum_{j=1}^{G} \Pr(Y_t \le (y_t - 1)|y^{t-1}, \theta^{(j)}) p(\theta^{(j)}|y^{t-1})$$

$$+ \alpha_t \sum_{j=1}^{G} \Pr(Y_t \le y_t|y^{t-1}, \theta^{(j)}) p(\theta^{(j)}|y^{t-1})$$

Case Study 1: Rainfall occurrence in Tokyo. We consider model diagnostics for a time series y_t, $t = 1, \ldots, 366$, where y_t takes the values 0, 1 and 2 depending on the number of rainfall occurrences

in Tokyo on the day t in the years 1983 and 1984 (Kitagawa, 1987). This time series has been modelled by Kitagawa (1987) and Fahrmeir (1992) by a dynamic binomial logit model: $y_t \sim \mathrm{B}(n_t, \pi_t)$, $n_t = 2$ for $t \neq 60$ and $n_t = 1$ for $t = 60$, $\pi_t = \mathrm{logit}(a_t)$, $a_t = a_{t-1} + w_t$, $w_t \sim \mathrm{N}(0, \sigma_\eta^2)$. For the prior parameters and the process variance σ_η^2 we take the values estimated by Fahrmeir *et al.* (1992, p.85): $\hat{a}_{0|0} = -1.58$, $P_{0|0} = 0.1$, $\sigma_\eta^2 = 0.33$ – thus $N = 366$. The upper and the middle part of Fig. 1 show diagnostics based on the P-scores u_t defined by (13) and $v_t = \Phi^{-1}(u_t)$. The predictive distribution has been approximated from (6) with the posterior moments estimated by posterior mode filtering (Fahrmeir, 1992) in the upper and integration-based Kalman filtering (Frühwirth-Schnatter, 1994) in the middle part. For both filtering methods diagnostics are satisfactory (compare the indices with the quantiles given in Table 1). There exists, however, a slight tendency toward bias for posterior mode filtering which vanishes for integration-based Kalman filtering.

For illustration we include diagnostics based on Pearson residuals in the lower part of the figure. The indices B_N and D_N indicate that the first two moments do not differ significantly from 0 and 1. The normal plot as well as the indices S_N and T_N, however, show the extreme non-normality of the Pearson residuals. Therefore Pearson residuals cannot be used to check the observation distribution.

5. Model diagnostics and Bayes factors

The diagnostic methods discussed in Section 3 may be compared with other Bayesian methods of examining the adequacy of a model. O'Hagan (1980) in discussing Box (1980) argues that the correct Bayesian solution to model diagnostics is the use of Bayes factors which are the ratios of the overall predictive densities (model likelihood $L(y^n|\mathcal{M}_j)$) given two different models $\mathcal{M}_1$ and $\mathcal{M}_2$:

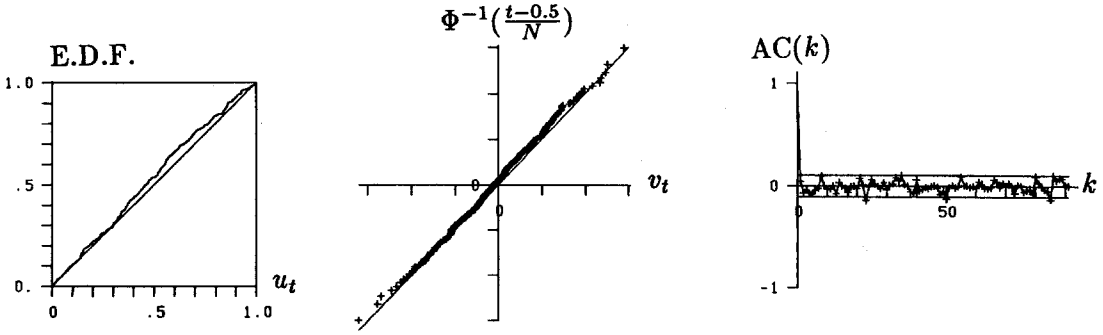
$$B = \frac{L(y^n|\mathcal{M}_1)}{L(y^n|\mathcal{M}_2)}, \qquad L(y^n|\mathcal{M}_j) = p(y_{m_0}, \ldots, y_n|\mathcal{M}_j) = \prod_{t=m_0}^{n} p(y_t|y^{t-1}, \mathcal{M}_j)$$

m_0 is the smallest integer such that all predictive densities $p(y_t|y^{t-1}, \mathcal{M}_j)$ are proper densities for $t \geq m_0$ for all models $\mathcal{M}_j$. The Bayesian approach of selecting one of two models from the Bayes factor has been extensively discussed in the literature (see among many others: Smith and Spiegelhalter, 1980; Spiegelhalter and Smith, 1982; Berger and Delambady, 1987). To select one of more than two models one usually computes posterior probabilities from the model likelihood (e.g. Geisser and Eddy, 1979): $P(\mathcal{M}_j|y^n) \propto L(y^n|\mathcal{M}_j)P(\mathcal{M}_j)$.

The Bayes factor is a measure of the relative performance of one model compared to another. It has been already pointed out by Box (1980) in his reply to the discussion that a large Bayes factor alone does not guarantee that the preferred model is appropriate. Model diagnostics, for instance, may be extremely poor. On the other hand, in practical case studies it often turns out that more than one model passes the global diagnostic examination and model diagnostics alone is not sensitive enough for model discrimination. In such cases Bayes factors – or more general Bayesian model discrimination rules – turn out to be an appropriate method for model choice. In the two following case studies we demonstrate how such a combination of model diagnostics, Bayes factors and model discrimination works in practice.

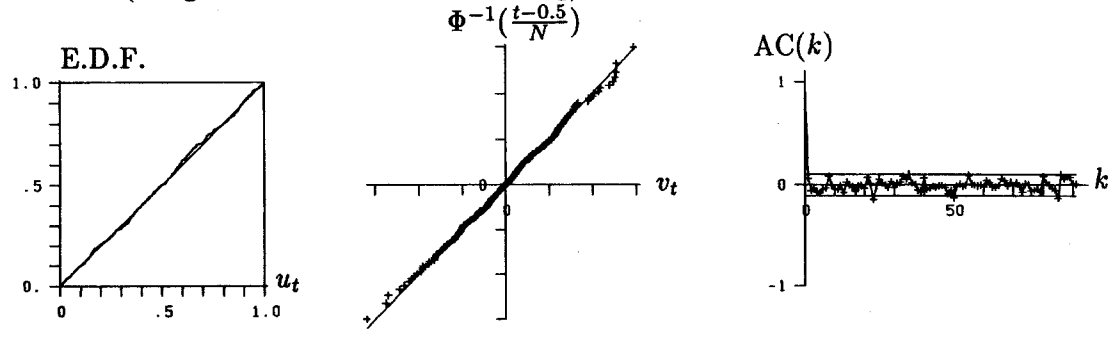
Computing Bayes factors is standard for linear models (see e.g. Smith and Spiegelhalter, 1980). If the model is non-linear in the parameter – as is the case for state space models with unknown hyperparameters or non-normal state space models – only approximate Bayes factors are available. Bayes factors for normal state space models with unknown hyperparameter may be computed by Markov chain Monte Carlo methods (Frühwirth-Schnatter, 1995). For non-normal state space such a Markov chain Monte Carlo approximation has not yet been derived.

P-scores (approximate posterior mode filtering)



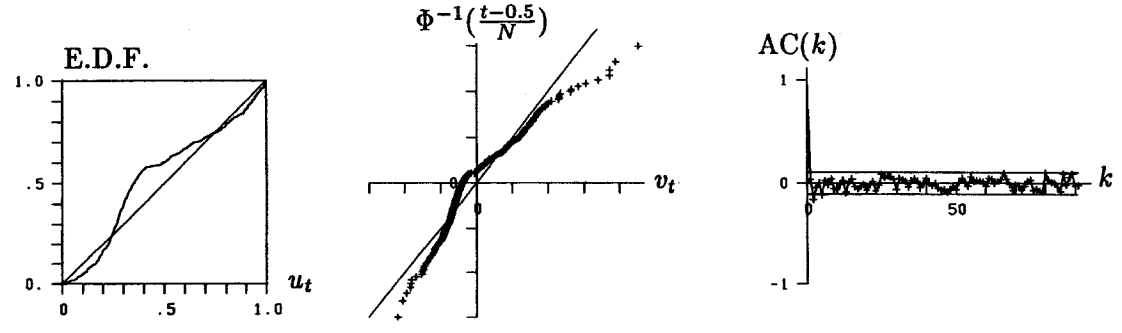
$B_N = -1.91$ $D_N = -0.85$ $S_N = -0.01$ $T_N = 0.81$ $A_N = 0.85$

P-scores (integration-based Kalman-filtering)



$B_N = -0.31$ $D_N = 0.12$ $S_N = -0.10$ $T_N = 0.55$ $A_N = 1.23$

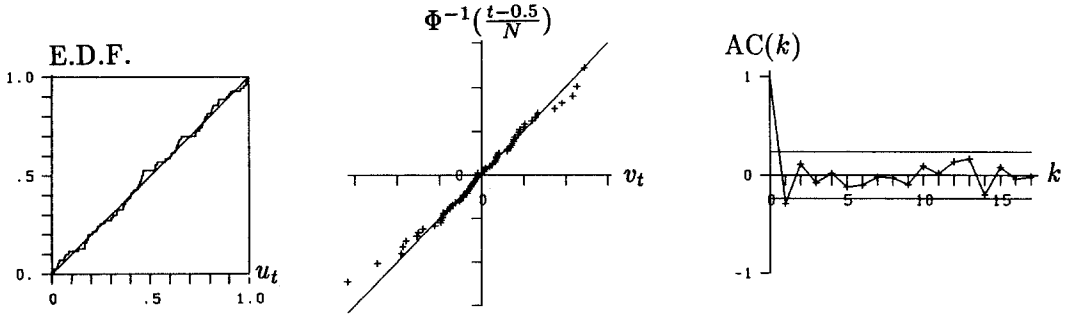
Pearson residuals



$B_N = 0.03$ $D_N = 1.76$ $S_N = 8.61$ $T_N = 5.74$ $A_N = 0.72$

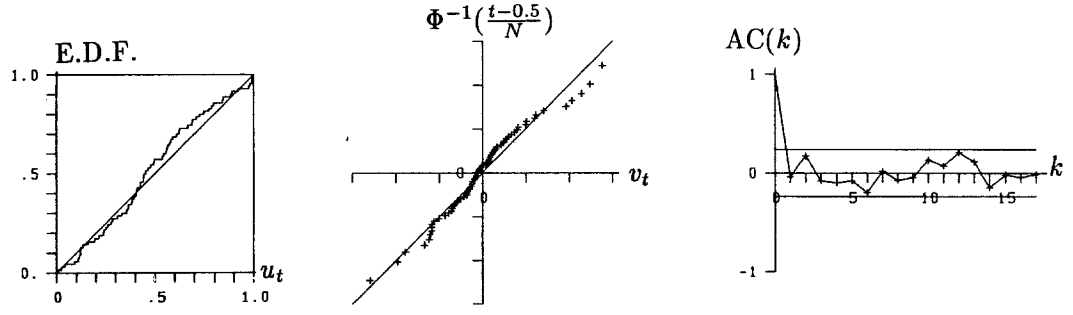
Fig. 1. Diagnostics for rainfall occurence in Tokyo.

$\mathcal{M}_1$: dynamic model (Poisson distribution)



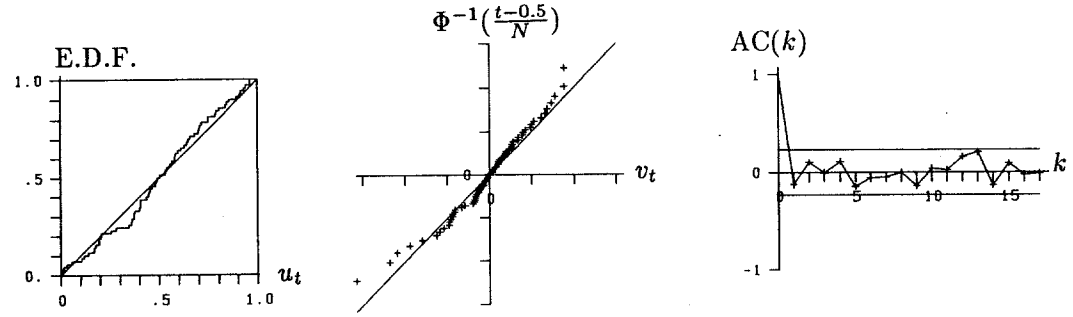
$B_N = -0.26$ $D_N = 0.89$ $S_N = -0.59$ $T_N = 1.05$ $J_N = 1.46$ $A_N = -2.28$

$\mathcal{M}_2$: dynamic model (normal distribution)



$B_N = -0.13$ $D_N = -0.08$ $S_N = 1.63$ $T_N = 1.80$ $J_N = 5.90$ $A_N = -0.19$

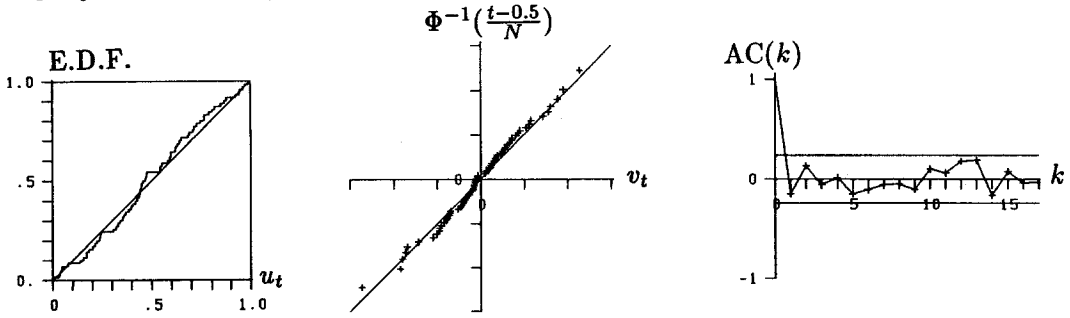
$\mathcal{M}_3$: dynamic model (log-normal distribution)



$B_N = -0.35$ $D_N = -0.60$ $S_N = -2.30$ $T_N = 2.03$ $J_N = 9.40$ $A_N = -0.94$

Fig. 2. Diagnostics for time series of purse snatching.

$\mathcal{M}_4$: dynamic model (negative binomial distribution)



$B_N = -0.21$ $D_N = -0.73$ $S_N = -0.37$ $T_N = 0.96$ $J_N = 1.05$ $A_N = -1.13$

Fig. 2. Continued

An approximation method for the model likelihood $L(y^n|\mathcal{M}_j)$ follows immediately from the results of the previous section. If all hyperparameters θ are known, $p(y_t|y^{t-1}, \mathcal{M}_j)$ is approximated simply by (11) with $\theta^{(i)} = \theta$. For unknown hyperparameters we use:

$$p(y_t|y^{t-1}, \mathcal{M}_j) = \int p(y_t|y^{t-1}, \theta, \mathcal{M}_j) p(\theta|y^{t-1}, \mathcal{M}_j) \mathrm{d}\theta$$

$$\approx \sum_{i=1}^{G} p(y_t|y^{t-1}, \theta^{(i)}, \mathcal{M}_j) p(\theta^{(i)}|y^{t-1}, \mathcal{M}_j), \qquad m_0 \le t \le n$$

where we compute $p(y_t|y^{t-1}, \theta^{(i)}, \mathcal{M}_j)$ from (11) and $p(\theta^{(i)}|y^{t-1}, \mathcal{M}_j)$ from (10).

A final remark concerns diagnostics and discrimination of models where one or more of them are built up for the original time series y_t and others for a transformed version $y_t^* = f(y_t)$ of y_t. The predictive distribution which we need for computing the P-scores is invariant to invertible transformations of the time series. Thus in contrast to other residuals such as the Pearson residuals which depend on the chosen scale, a direct comparison of models for original and transformed time series is possible with P-scores. When discriminating such models by a Bayes factor it should be kept in mind that the functional value of the predictive density which we need for the Bayes factor is not invariant to data transformations. Before carrying out model discrimination we have to compute the model likelihood $L(y^n|\mathcal{M}_j)$ of the original time series from the model likelihood $L^*((y^*)^n|\mathcal{M}_j)$ of the transformed time series by the following formula:

$$L(y^n|\mathcal{M}_j) = L^*((y^*)^n|\mathcal{M}_j) \prod_{t=n_0}^{n} \frac{\mathrm{d}f}{\mathrm{d}y}(y_t)$$

This result follows directly from the law of transformation of densities.

Case Study 2: Purse snatchings in Chicago. Here we reanalyse a time series of reported purse snatchings over a period of 71 weeks in the Hyde Park neighbourhood of Chicago, taken from Harvey (1989, pp. 217, 516). The local level model $a_t = a_{t-1} + w_t$, $w_t \sim \mathrm{N}(0, \sigma_\eta^2)$, $\lambda_t = a_t$, is combined with various observation distributions such as the Poisson, the normal, the log-normal and the negative binomial distribution:

$$\mathcal{M}_1 : y_t|\lambda_t \sim P_{\mu_t}, \qquad \ln \mu_t = \lambda_t$$

$$\mathcal{M}_2 : y_t|\lambda_t \sim \mathrm{N}(\mu_t, R), \qquad \mu_t = \lambda_t$$

Table 2. Model discrimination for time series of purse snatching

$\mathcal{M}_j$	$\log L(y^n\|\mathcal{M}_j)$	$P(\mathcal{M}_j)$	$P(\mathcal{M}_j\|y^n)$
$\mathcal{M}_1$	−220.52	0.25	0.250
$\mathcal{M}_2$	−227.72	0.25	4.8×10^{-4}
$\mathcal{M}_3$	−221.42	0.25	0.102
$\mathcal{M}_4$	−219.57	0.25	0.647

$$\mathcal{M}_3 : \ln y_t|\lambda_t \sim \mathrm{N}(\mu_t, R), \qquad \mu_t = \lambda_t$$
$$\mathcal{M}_4 : y_t|\lambda_t \sim \mathrm{NB}(\nu, \pi_t), \qquad \pi_t = \mathrm{logit}(\lambda_t)$$

Estimation is carried out via integration-based Kalman filtering with improper priors (thus $N = 70$). The unknown hyperparameters σ_η^2, R and ν are estimated via multi-process filtering.

Figure 2 shows model diagnostics for each of these distributions (compare the indices with the quantiles given in Table 1). The Poisson distribution is not satisfactory because of a surprisingly high autocorrelation index A_N at lag 1. The log-normal distribution has a surprisingly low skewness index S_N and a surprisingly high joint index J_N. Only the normal and the negative binomial distribution pass the examination. From model diagnostics alone it is not possible to decide which of these two distributions explains the time series better.

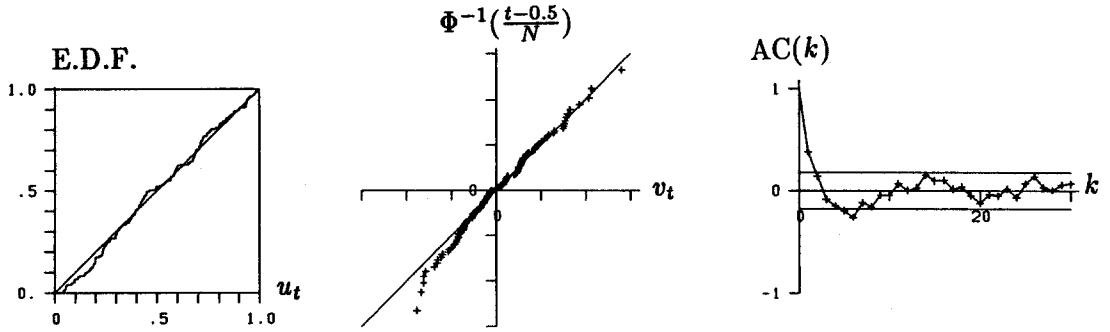
In Table 2 the various observation distributions are compared via model likelihoods and posterior probabilities based on equal prior weights for each model. The negative binomial distribution ($\mathcal{M}_4$) is the observation distribution with the biggest posterior probability, the posterior probability of the normal distribution ($\mathcal{M}_2$) is extremely low. The Bayes factor of model $\mathcal{M}_2$ versus $\mathcal{M}_4$ is equal to 0.00029 and highly favours the negative binomial distribution. Surprisingly the Poisson distribution ($\mathcal{M}_1$) which has been rejected because of autocorrelation in the residuals has considerable posterior probability. It would have been highly favoured if compared with the normal distribution ($\mathcal{M}_1$) via the Bayes factor which is equal to 1339.4.

6. Iterative model building based on model diagnostics

Applied model building has often been viewed as an iterative process of acquisition of knowledge (e.g. Box, 1980; Sharefkin, 1983). A major problem with model building lies in the complex way different misspecifications may interact. As the methods suggested in this paper allow for diagnostics without the need to specify an alternative model, we start with some simple models. The question, whether a model is appropriate or whether and how it should be modified, in practice is investigated by inspecting residuals. It will be illustrated by the following case study that the results of diagnostics might contain valuable hints how to design alternative and hopefully improved models.

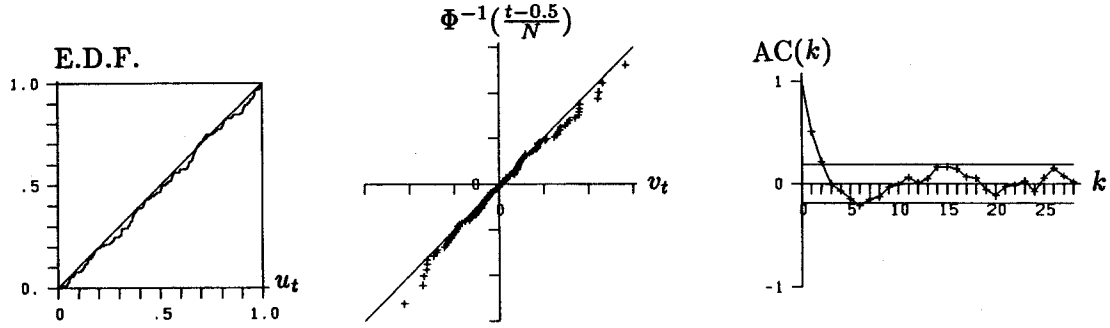
Case Study 3: Sulfur dioxide emissions. Here we reanalyse a time series of daily sulfur dioxide emission in Brotjachtriegel (FRG) over a period of 4 months ($n = 122$) which was published in Frühwirth-Schnatter (1991). Various questions have to be answered in order to build an appropriate state space model. What is the conditional observation distribution $p(y_t|\lambda_t)$ of the time series y_t given the predictor λ_t? Which effects are present (choice of the components $\boldsymbol{x}_t$, $\mathbf{F}_t$ and $\boldsymbol{H}_t$ of the structural part of the model)? Are these effects static (fixed) or dynamic (random) (choice of $\boldsymbol{Q}_t$)?

$\mathcal{M}_1$: gamma distribution with mixed link; $\lambda_t = a_t$; $\sigma_\eta^2 > 0$



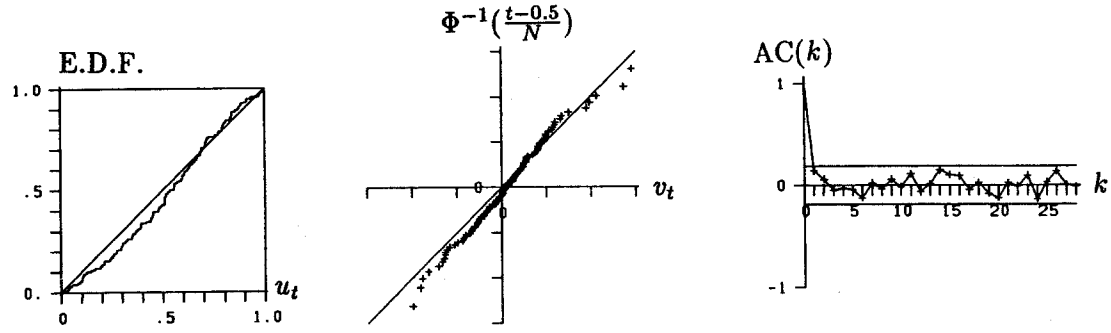
$B_N = 0.48$ $D_N = -1.46$ $S_N = 1.69$ $T_N = -0.14$ $J_N = 2.86$ $A_N = 4.28$

$\mathcal{M}_2$: gamma distribution with mixed link; $\lambda_t = a_t + s_t$; $\sigma_\eta^2 > 0, \sigma_\omega^2 > 0$



$B_N = 0.99$ $D_N = -0.08$ $S_N = 1.18$ $T_N = -0.54$ $J_N = 1.68$ $A_N = 5.47$

$\mathcal{M}_3$: gamma distribution with mixed link; $\lambda_t = a_t + s_t + l_t$; $\sigma_\eta^2 > 0, \sigma_\omega^2 > 0$



$B_N = 1.13$ $D_N = -1.51$ $S_N = 1.24$ $T_N = 1.51$ $J_N = 3.80$ $A_N = 1.59$

Fig. 3. Diagnostics for time series of sulfur dioxide emissions.

Table 3. Diagnostics for time series of sulfur dioxide emissions ('+' indicates $\sigma^2 > 0$, '0' indicates $\sigma^2 = 0$)

Effects	$p(y_t\|\lambda_t)$	σ_η^2	σ_ω^2	B_N	D_N	S_N	T_N	J_N	A_N
$\lambda_t = a_t$	N	+		0.01	0.81	4.90	5.31	52.60	1.96
	log-N	+		0.12	−0.97	−1.59	0.48	2.75	4.52
	γ-mixed	+		0.48	−1.46	1.69	−0.14	2.86	4.28
	γ-log	+		0.64	−0.57	1.57	0.56	2.79	4.17
	t_4	+		1.80	1.57	4.19	0.83	18.27	4.93
	log-t_4	+		−0.37	−1.65	−0.65	−1.46	2.55	4.76
$\lambda_t = a_t + s_t$	log-N	+	+	−0.14	−0.85	−1.47	0.89	2.94	4.72
	γ-mixed	+	+	0.99	−0.08	1.18	−0.54	1.68	5.47
	γ-log	+	+	0.46	−0.47	1.20	0.17	1.48	4.93
	log-t_4	+	+	−0.38	−1.55	−0.80	−0.34	0.75	5.04
$\lambda_t = a_t + s_t + l_t$	log-N	+	+	−0.03	0.40	−1.20	2.19	6.22	0.74
	γ-mixed	+	+	1.13	−1.51	1.24	1.51	3.80	1.59
	γ-mixed	0	+	0.89	−1.12	0.80	0.97	1.57	1.71
	γ-mixed	+	0	0.88	−1.13	1.70	2.17	7.62	1.33
	γ-mixed	0	0	0.81	−0.75	1.20	1.43	3.48	1.65
	γ-log	+	+	0.03	−1.26	1.99	1.89	7.53	2.32
	log t_4	+	+	−0.08	−2.28	−0.76	0.49	0.82	0.87
	log t_4	0	+	0.20	−1.96	−1.39	0.02	1.94	1.17
	log t_4	0	0	0.21	−1.84	−1.49	0.06	2.23	1.04
$\lambda_t = a_t + l_t$	log-N	+		0.40	1.29	−0.68	1.88	3.99	0.48
	γ-mixed	+		0.34	−1.88	2.41	2.48	11.97	0.73
	γ-log	+		0.53	−0.85	2.77	3.00	16.69	1.28
	log-t_4	+		0.16	−1.87	−0.73	−0.15	0.55	0.46
	log-t_4	0		0.62	−1.65	−1.40	−0.38	2.09	0.69

For the present time series we start with various distributions

normal distribution: $y_t|\lambda_t \sim \mathrm{N}(\lambda_t, R)$

log-normal distribution: $\log y_t|\lambda_t \sim N(\lambda_t, R)$

gamma distribution with mixed link: $y_t|\lambda_t \sim \gamma\left(\alpha, \dfrac{\alpha}{\mu_t(\lambda_t)}\right)$

$$\mu_t(\lambda_t) = g^{-1}(\lambda_t) = \begin{cases} \lambda_t, & \lambda_t \geq 1 \\ \exp(\lambda_t - 1), & \lambda_t \leq 1 \end{cases}$$

gamma distribution with log-link: $y_t|\lambda_t \sim \gamma\left(\alpha, \dfrac{\alpha}{\mu_t(\lambda_t)}\right)$

$$\mu_t(\lambda_t) = g^{-1}(\lambda_t) = \exp(\lambda_t)$$

Student t_4-distribution: $y_t|\lambda_t \sim t_4(\lambda_t, R)$

log-Student t_4-distribution: $\log y_t|\lambda_t \sim t_4(\lambda_t, R)$

and the most simple structural model, namely a local level model which is assumed to be dynamic: $\lambda_t = a_t$, $a_t = a_{t-1} + w_t$, $w_t \sim \mathrm{N}(0, \sigma_\eta^2)$ and $\sigma_\eta^2 > 0$.

Diagnostics for these models (n = 121) are summarized in Table 3 (compare these indices with the quantiles in Table 1). The higher moment indices indicate objections against the normal and the t_4-distribution and no objections against the gamma distribution (for both link functions), the log-normal and the log-t_4 distribution. The later models are, however, rejected because of autocorrelation of the residuals at lag 1. Furthermore for each of these models the complete autocorrelogram exhibits a slight periodic behaviour – see for example the autocorrelogram of the residuals from the gamma distribution with mixed link in the upper part of Fig. 3. This might be due to autocorrelation only, but it might also be an indication of a weekly seasonal effect.

First we refine the structural part by including a form-free weekly seasonal effect:

$$\begin{aligned} \lambda_t &= a_t + s_t \\ a_t &= a_{t-1} + w_{t,1}, \qquad w_{t,1} \sim \mathrm{N}(0, \sigma_\eta^2) \\ s_t &= -\sum_{j=1}^{6} s_{t-j} + w_{t,2}, \qquad w_{t,2} \sim \mathrm{N}(0, \sigma_\omega^2) \end{aligned} \tag{14}$$

This effect is dynamic if $\sigma_\omega^2 > 0$. Model structure (14) is combined with both gamma distributions, the log-normal and the log-t_4 distribution. All models are rejected once more because short-term autocorrelation still is present (compare the indices in Table 3 – n = 115 – with the quantiles in Table 1; see also Fig. 3). Thus we decided to extend model structure (14) by introducing lagged values of the time series into the predictor λ_t:

$$\begin{aligned} \lambda_t &= a_t + s_t + l_t \\ l_t &= \beta(g(y_{t-1}) - a_{t-1} - s_{t-1}) \end{aligned} \tag{15}$$

a_t and s_t are modelled in the same way as in (14), $g(\cdot)$ is equal to the link function for the gamma distributions and equal to the identity for the log-normal and the log-t_4 distribution. β is treated as an unknown hyperparameter and is estimated from the data in the same way as $\sigma_\eta^2, \sigma_\omega^2$ and α by multiprocess filtering. Such Markov models where the linear predictor depends on past outcomes have been applied to generalized linear models by various authors (e.g. Cox, 1970; Fahrmeir and Kaufmann, 1987; Zeger and Qaqish, 1988) and are extended to dynamic generalized linear models by (15).

Model (15) is combined with the various observation distributions. Only for the gamma distribution with the log-link is autocorrelation still present; for the gamma distribution with the mixed link, the log-normal and the log-t_4 distribution short-term autocorrelation of the residuals vanishes. However, model diagnostics from the other indices is satisfactory for the gamma distribution with the mixed link, only (compare the indices in Table 3 – N = 114 – with the quantiles in Table 1; see also Figure 3).

To get an idea if it is really necessary to include the weekly seasons, we have dropped s_t from (15), but have kept the lagged observations in the predictor:

$$\lambda_t = a_t + l_t \tag{16}$$

This model is rejected for both gamma distributions and is not rejected for the log-normal and the log-t_4 distribution (compare the indices in Table 3 – n = 120 – with the quantiles in Table 1).

For the moment we end up with three candidates for a suitable model, one of them with and two of them without a dynamic seasonal effect. It is not possible to decide whether a dynamic weekly effect is present or not by the standard tools of model diagnostics. If we compute Bayes factors from the model likelihoods given in Table 4 we would much prefer the gamma distribution with weekly seasonal effects.

Table 4. Model discrimination for time series of sulfur dioxide emissions (only reporting models with $P(\mathcal{M}_j|y^n) > 10^{-6}$)

Effects	$p(y_t\|\lambda_t)$	σ_η^2	σ_ω^2	$logL(y^n\|\mathcal{M}_j)$	$P(\mathcal{M}_j)$	$P(\mathcal{M}_j\|y^n)$
$\lambda_t = a_t + s_t + l_t$	γ-mixed	+	+	−362.88	1/48	0.2069
	γ-mixed	0	+	−361.84	1/48	0.5877
	γ-mixed	+	0	−364.68	1/48	0.0342
	γ-mixed	0	0	−363.28	1/48	0.1394
	log t_4	+	0	−374.86	1/48	1.3×10^{-6}
	log t_4	0	+	−374.27	1/48	2.3×10^{-6}
	log t_4	0	0	−373.35	1/48	5.8×10^{-6}
$\lambda_t = a_t + l_t$	log-N	+		−371.84	1/48	2.6×10^{-5}
	log-N	0		−372.41	1/48	1.5×10^{-5}
	γ-mixed	+		−366.18	1/48	0.0076
	γ-mixed	0		−365.10	1/48	0.0225
	γ-log	+		−369.33	1/48	3.2×10^{-4}
	γ-log	0		−368.28	1/48	9.4×10^{-4}
	log-t_4	+		−370.79	1/48	7.6×10^{-5}
	log-t_4	0		−369.68	1/48	2.3×10^{-4}

It still remains to decide whether some of the effects are static or dynamic. For the most general model, namely model (15), four cases are possible: $\sigma_\omega^2 > 0, \sigma_\eta^2 > 0$ or $\sigma_\omega^2 = 0, \sigma_\eta^2 > 0$ or $\sigma_\omega^2 > 0, \sigma_\eta^2 = 0$ or $\sigma_\omega^2 = 0, \sigma_\eta^2 = 0$. For the gamma distribution with mixed link only the combination $\sigma_\omega^2 = 0, \sigma_\eta^2 > 0$ can be rejected by the standard tools of model diagnostics (compare the indices in Table 3 with the quantiles in Table 1). If we compute Bayes factors from the model likelihoods reported in Table 4, we would slightly prefer the model with the fixed level and dynamic seasonal effects.

One might wonder if any other of the 48 possible combinations of observations distributions, effects and assumptions on the variances σ_η^2 and σ_ω^2 would lead to a model which is not rejected or which is even preferable to the chosen model. For completeness we would like to mention that by allowing for fixed effects we found three further models – all of them based on the log-t_4 distribution – which could not be rejected, namely two seasonal models (15) with $\sigma_\eta^2 > 0$, $\sigma_\omega^2 = 0$ and $\sigma_\eta^2 = 0$, $\sigma_\omega^2 = 0$, respectively, and the seasonal free model (16) with $\sigma_\eta^2 = 0$ (compare the indices in Table 3 with the quantiles in Table 1). All other 42 models are rejected by model diagnostics.

To select a possible candidate from this group of various possible models we carried out Bayesian model discrimination among all 48 models with $P(\mathcal{M}_j) = 1/48$ and $m_0 = 10$. The results are summarized in Table 4. The model with the highest posterior probability is the gamma distribution with mixed link, fixed level and dynamic seasonal effects. It is highly preferred compared with each of those models based on the log-t_4 distributions which passed model diagnostics. This result is not surprising if we take into account that the estimated shape parameter α of the gamma-distribution is very close to 1: $\mathrm{E}(\alpha|y^n) = 1.065$, indicating an extremely skew observation distribution. It seems not to be possible to eliminate this skewness by taking the logarithm of the data.

7. Concluding remarks

Although we have found the diagnostic tools defined in Section 3 of this paper sufficient for a routine diagnostic check of non-normal state space models, we would like to mention that more

specific tests are also easily extended to non-normal state space models, if they are based on the transformed P-scores. This is especially true of the CUSUM technique which has been discussed for normal state space models by Brown *et al.* (1975), as well as of the test for heteroscedasticity discussed for regression models in Hedayat and Robson (1970) and for normal state space model in Harvey (1989, p. 259).

In the present paper we have confined ourselves to univariate time series. There are two ways of extending the methods of this paper to multivariate time series y_t. The first method is to define univariate P-scores u_t and transformed P-scores v_t from the joint predictive distribution: $u_t = \Pr(\boldsymbol{Y}_t \leq \boldsymbol{y}_t | \boldsymbol{y}^{t-1})$ and $v_t = \Phi^{-1}(u_t)$. Computational methods for approximating the predictive distribution for multivariate time series are covered by Frühwirth-Schnatter (1994). $\{u_t\}$ and $\{v_t\}$ are i.i.d. uniform on [0,1] and standard normal, respectively, if the model is correct. Sometimes, for instance for longitudinal data, it is more interesting to check the model for each component $\{y_{t,i}\}$, $t = 1, \ldots, N$ of the multivariate time series individually. P-scores $\{u_{t,i}\}$ and $\{v_{t,i}\}$ defined from the marginal predictive distribution by $u_{t,i} = \Pr(Y_{t,i} \leq y_{t,i} | \boldsymbol{y}^{t-1})$ and $v_{t,i} = \Phi^{-1}(u_{t,i})$ should be i.i.d. uniform on [0,1] and standard normal, respectively, if the model is correct. For normal linear state space models $\boldsymbol{v}_t = (v_{t,1} \cdots v_{t,r})$ is equal to a multivariate version of the Pearson residual.

Acknowledgements

I want to thank Professor W. Panny (Vienna University of Economics and Business Administration) and Dr R. Frühwirth (Austrian Academy of Science) for their support when I tried to determine the empirical quantiles of the various diagnostic indices by Monte Carlo Simulation.

References

Abramowitz, M. and Stegun, I. (1970) *Handbook of Mathematical Functions*. Dover, New York.

Aitkin, M., Anderson D., Francis, B. and Hinde, J. (1989) *Statistical Modelling in GLIM*. Clarendon Press, Oxford.

Atkinson, A.C. (1981) Two graphical displays for outlying and influential observations in regression. *Biometrika*, **68**, 13–20.

Atkinson, A.C. (1982) Regression diagnostics, transformations and constructed variables. *Journal of the Royal Statistical Society* B, **44**, 1–36.

Berger, J. and Delambady, M. (1987) Testing precise hypothesis (with discussion). *Statistical Science*, **2**, 317–35.

Bowman, K.O. and Shenton, L.R. (1975) Omnibus contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika*, **62**, 243–50.

Box, G.E.P. (1980) Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society* A, **143**, 383–430.

Brown, R.L., Durbin, J. and Evans, J.M. (1975) Techniques of testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society* B, **37**, 141–92.

Chatfield, C. (1989) *The Analysis of Time Series*. Chapman & Hall, London.

Cox, D.R. (1970) *The Analysis of Binary Data*. Chapman & Hall, London.

Davison, A.C. and Gigli, A. (1989) Deviance residuals and normal score plots. *Biometrika*, **76**, 211–21.

Dawid, A.P. (1984) Statistical theory – the prequential approach. *Journal of the Royal Statistical Society* A, **147**, 278–92.

Fahrmeir, L. (1992) Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association*, **87**, 501–9.

Fahrmeir, L. and Kaufmann, H. (1987) Regression models for non-stationary categorical time series. *Journal of Time Series Analysis*, **8**, 147–60.

Fahrmeir, L., Hennevogel, W. and Klemme, K. (1992) Smoothing in dynamic generalized linear models by Gibbs sampling. In *Advances in GLIM and Statistical Modelling*, Fahrmeir, L., Francis, B., Gilchrist, R. and Tutz, G. (eds), *Lecture Notes in Statistics*, **78**, Springer-Verlag, Berlin pp. 85–90.

Frühwirth-Schnatter, S. (1991) Monitoring von ökologischen und biometrischen Prozessen mit statistischen Filtern. In *Multivariate Modelle. Neue Ansätze für biometrische Anwendungen*, Seeber, G.U.H. and Minder, Ch.E. (eds), Springer-Verlag, Berlin, pp. 89–122.

Frühwirth-Schnatter, S. (1994) Applied state space modelling of non-Gaussian time series using integration-based Kalman-filtering. *Statistics and Computing*, **4**, 259–69.

Frühwirth-Schnatter, S. (1995) Bayesian model discrimination and Bayes factors for linear Gaussian state space models. *Journal of the Royal Statistical Society*, B, **57**, 237–46.

Geisser, S. and Eddy, W.F. (1979) A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153–60.

Harrison, P.J. and Stevens, C.F. (1976) Bayesian forecasting (with discussion). *Journal of the Royal Statistical Society* B, **38**, 205–47.

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

Harvey, A. and Fernandes, C. (1989) Time series models for count or qualitative observations (with discussion). *Journal of Business and Economic Statistics*, **7**, 407–22.

Hatzinger, R. and Panny, W. (1993) Single and twin-heaps as natural data structures for percentile point simulation algorithm. *Statistics and Computing*, **3**, 163–70.

Hedayat, A. and Robson, D.S. (1970) Independent stepwise residuals for testing homoscedasticity. *Journal of the American Statistical Association*, **65**, 1573–81.

Kitagawa, G. (1987) Non-Gaussian state space modelling of nonstationary time series (with comments). *Journal of the American Statistical Association*, **82**, 1032–63.

O'Hagan, A. (1980) Discussion of Professor Box's paper. *Journal of the Royal Statistical Society* A, **143**, 408.

Pearson, E.S., D'Agostino, R.B. and Bowman, K.O. (1977) Tests for departure from normality: comparison of powers. *Biometrika*, **64**, 231–46.

Pregibon, D. (1981) Logistic regression diagnostics. *Annals of Statistics*, **9**, 705–24.

Rosenblatt, M. (1952) Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, **23**, 470–2.

Schnatter, S. (1992) Integration-based Kalman-filtering for a dynamic generalized linear trend model. *Computational Statistics and Data Analysis*, **13**, 447–59.

Schneider, W. (1986) *Der Kalmanfilter als Instrument zur Diagnose und Schätzung variabler Parameter in ökonometrischen Modellen*. Physica, Heidelberg.

Sharefkin, M. (1983) Reflections of an Ignorant Bayesian. In *Uncertainty and Forecasting of Water Quality*, Beck, M.B. and van Straten, G. (eds), Springer-Verlag, Berlin, pp. 373–9.

Smith, J.Q. (1985) Diagnostic check of non-standard time series models. *Journal of Forecasting*, **4**, 283–91.

Smith, A.F.M. and Spiegelhalter, D.J. (1980) Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society B*, **42**, 213–20.

Spiegelhalter, D.J. and Smith, A.F.M. (1982) Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society B*, **44**, 377–87.

Stuart, A. and Ord, J.K. (1987) *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*. Griffin, London.

West, M., Harrison, P.J. and Migon, H. (1985) Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association*, **80**, 741–50.

Zeger, S.L. and Qaqish, B. (1988) Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, **44**, 1019–31.

Appendix

$s = Nm_2$ obviously is χ^2_{N-1} with $\mathrm{E}(s) = N - 1$ and $\mathrm{var}(s) = 2(N-1)$, if the residuals $v_{n_0}, \ldots,$

v_n are i.i.d. standard normal (as is the case for a correct model). Thus $D_N = (s - \mathrm{E}(s))/\sqrt{\mathrm{var}(s)}$ is standard normal asymptotically.

The observed moment ratios $\sqrt{b_1}$ and b_2 are related to the observed ratios of cumulants k_2, k_3 and k_4 in the following way (Stuart and Ord, 1987, pp. 410, 422):

$$\sqrt{b_1} = \frac{m_3}{(m_2)^{3/2}} = \frac{N-2}{\sqrt{N(N-1)}} \frac{k_3}{(k_2)^{3/2}}$$

$$b_2 = \frac{m_4}{m_2^2} = \frac{(N-2)(N-3)}{N^2-1} \frac{k_4}{k_2^2} + \frac{3(N-1)}{N+1}$$

The ratio of cumulants from the standard normal sample $v_{n_0}, \ldots, v_n$ have zero mean and variances given by Stuart and Ord (1987, p. 422):

$$\mathrm{var}\left(\frac{k_3}{(k_2)^{3/2}}\right) = \frac{6N(N-1)}{(N-2)(N+1)(N+3)}$$

$$\mathrm{var}\left(\frac{k_4}{k_2^2}\right) = \frac{24N(N-1)^2}{(N-3)(N-2)(N+3)(N+5)}$$

It is easy to verify that S_N and T_N are defined via:

$$S_N = (\sqrt{b_1} - E(\sqrt{b_1}))/\sqrt{\mathrm{var}(\sqrt{b_1})}$$

$$T_N = (b_2 - E(b_2))/\sqrt{\mathrm{var}(b_2)}$$

Therefore S_N and T_N are standard normal asympotically, if the residuals $v_{n_0}, \ldots, v_n$ are i.i.d. standard normal.

Biographical sketch

Dr Sylvia Frühwirth-Schnatter took her diploma degree in 1982 and Ph.D. in mathematics in 1989 at the Vienna University of Technology. In 1982–88 she was a research assistant at the Department of Hydraulics, Hydrology and Water Resources Research at the Vienna University of Technology. In 1988–90 she was a research assistant at the Department of Statistics and Probability Theory at the Vienna University of Technology. Since 1990 she has been assistant professor at the Department of Statistics at the Vienna University of Economics and Business Adiministration.