AALBORG UNIVERSITY

Formulating state space models in R with focus on longitudinal regression models

by

Claus Dethlefsen and Søren Lundbye-Christensen

R-2005-21 May 2005

Department of Mathematical Sciences
Aalborg University
Fredrik Bajers Vej 7 G ▪ DK - 9220 Aalborg Øst ▪ Denmark
Phone: +45 96 35 80 80 ▪ Telefax: +45 98 15 81 29
URL: http://www.math.aau.dk

ISSN 1399–2503 ▪ On-line version ISSN 1601–7811

Formulating State Space Models in R with Focus on Longitudinal Regression Models

Claus Dethlefsen
Aalborg Hospital, Aarhus University Hospital

Søren Lundbye-Christensen
Aalborg University

Resumé

We provide a language for formulating a range of state space models. The described methodology is implemented in the R-package **sspir** available from `cran.r-project.org`. A state space model is specified similarly to a generalized linear model in R, by marking the time-varying terms in the formula. However, the model definition and the model fit are separated in different calls. The model definition creates an object with a number of associated functions. The model object may be edited to incorporate extra features before it is fitted to data. The formulation of models does not depend on the implemented method of inference. The package is demonstrated on three datasets.

Keywords: dynamic models, exponential family, generalized linear models, iterated extended Kalman smoothing, Kalman filtering, seasonality, time series, trend.

1. Introduction

Generalized linear models, see McCullagh and Nelder (1989), are used when analyzing data where response-densities are assumed to belong to the exponential family. Time series of counts may adequately be described by such models. However, if serial correlation is present or if the observations are overdispersed, these models may not be adequate, and several approaches can be taken. The book by Diggle, Heagerty, Liang, and Zeger (2002) gives an excellent review of many approaches incorporating serial correlation and overdispersion in generalized linear models. Dynamic generalized linear models (DGLM), often called state space models, also address those problems and are treated in a paper by West, Harrison, and Migon (1985) in a conjugate Bayesian setting. They have been subject to further research by *e.g.* Zeger (1988) using generalized estimating equations (GEE), Gamerman (1998) using Markov chain Monte Carlo (MCMC) methods and Durbin and Koopman (1997) using iterated extended Kalman filtering and importance sampling.

Standard statistical software does not not include procedures for DGLMs and only sparse support for Gaussian state space models. There is a need for a simple, yet flexible way of specifying complicated non-Gaussian state space models. Often, one need to tailor make software for each specific application in mind. A function, `StructTS`, has been developed for analysis of a subclass of Gaussian state space models, see Ripley (2002). The freely available `ssfpack` for Ox provide a tool set for analysis of Gaussian state space models with some support for non-Gaussian models, see Koopman, Shephard, and Doornik (1999). The interface is very flexible, but not as easy to use as a `glm` call in R.

Section 2 describes Gaussian state space models and shows how generalized linear models can naturally be extended to allow the parameters to evolve over time. We define components (*e.g.* trend and seasonal components) that separate the time series into parts that may be inspected individually after analysis. In Section 3 the syntax for defining objects describing the proposed state space models are described as a simple, yet powerful, extension to the `glm`-call. The techniques are illustrated on three examples in Section 4.

2. State space models

The Gaussian state space model for univariate observations involves two processes, namely the state process (or latent process), $\{\boldsymbol{\theta}_k\}$, and the observation process, $\{y_k\}$. The random variation in the state space model is specified through descriptions of the sampling distribution, the evolution of the state vector and, the initialization of the state vector.

Let $\{y_k\}$ be measured at timepoints t_k for $k = 1, \dots, n$. The state space model is defined by

$$y_k = \mathbf{F}_k^\top \boldsymbol{\theta}_k + \nu_k, \qquad \nu_k \sim \mathcal{N}(0, V_k) \tag{1}$$

$$\boldsymbol{\theta}_k = \mathbf{G}_k \boldsymbol{\theta}_{k-1} + \boldsymbol{\omega}_k, \qquad \boldsymbol{\omega}_k \sim \mathcal{N}_p(\mathbf{0}, \mathbf{W}_k) \tag{2}$$

$$\boldsymbol{\theta}_0 \sim \mathcal{N}_p(\mathbf{m}_0, \mathbf{C}_0). \tag{3}$$

We assume that the disturbances $\{\nu_k\}$ and $\{\boldsymbol{\omega}_k\}$ are both serially independent and also independent of each other. The possible time-dependent quantities $\mathbf{F}_k$, $\mathbf{G}_k$, V_k and $\mathbf{W}_k$ may depend on a parameter vector, but this is suppressed in the notation.

We now consider the case where the state process is Gaussian and the sampling distribution belongs to the exponential family,

$$p(y_k|\eta_k) = \exp\left\{y_k\eta_k - b_k(\eta_k) + c_k(y_k)\right\}. \tag{4}$$

The density (4) contains the Gaussian, Poisson, gamma and the binomial distributions as special cases. The natural parameter η_k is related to the linear predictor λ_k by the equation $\eta_k = v(\lambda_k)$ or equivalently $\lambda_k = u(\eta_k)$. The linear predictor in a generalized linear model is of the form $\lambda_k = \mathbf{Z}_k\boldsymbol{\beta}$, where $\mathbf{Z}_k$ is a row vector of explanatory variables and $\boldsymbol{\beta}$ is the vector of regression parameters. The link function, g, relates the mean, $\mathsf{E}(y_k) = \mu_k$, and the linear predictor, λ_k, as $g(\mu_k) = \lambda_k$. The inverse link function, h, is defined as $\mu_k = \tau(\eta_k) = h(\lambda_k)$, where τ is the mean value mapping. The following relations hold $\eta_k = v(\lambda_k) = \tau^{-1}(h(\lambda_k))$ and $\lambda_k = u(\eta_k) = g(\tau(\eta_k))$, where u is the inverse of v. The link function is said to be canonical if $\eta_k = \lambda_k$, *i.e.* if $g = \tau^{-1}$.

2.1. Dynamic extension

The static generalized linear model is extended by adding a dynamic term, $\mathbf{X}_k\boldsymbol{\beta}_k$, to the linear predictor, where $\boldsymbol{\beta}_k$ is varying randomly over time according to a first order Markov process. Hence,

$$\lambda_k = \mathbf{Z}_k\boldsymbol{\beta} + \mathbf{X}_k\boldsymbol{\beta}_k, \tag{5}$$

where $\boldsymbol{\beta}$ is the coefficient of the static component and $\{\boldsymbol{\beta}_k\}$ are the time-varying coefficients of the dynamic component.

For notational convenience, we will use the notation

$$\lambda_k = \mathbf{F}_k^\top \boldsymbol{\theta}_k, \qquad \boldsymbol{\theta}_k = \left(\begin{array}{c} \boldsymbol{\beta} \\ \boldsymbol{\beta}_k \end{array} \right). \tag{6}$$

The evolution through time of the state vector, $\boldsymbol{\theta}_k$, is modelled by the relation

$$\boldsymbol{\theta}_k = \mathbf{G}_k\boldsymbol{\theta}_{k-1} + \boldsymbol{\omega}_k, \tag{7}$$

for an evolution matrix $\mathbf{G}_k$, determined by the model. The error terms, $\{\boldsymbol{\omega}_k\}$, are assumed to be independent Gaussian variables with zero mean and variance $\mathsf{VAR}(\boldsymbol{\omega}_k)$, with non-zero entries corresponding to the entries of the time-varying coefficients, $\boldsymbol{\beta}_k$, and zero elsewhere.

The model is fully specified by the initializing parameters $\mathbf{m}_0$ and $\mathbf{C}_0$, the matrices $\mathbf{F}_k$, $\mathbf{G}_k$, and the variance parameters V_k and $\mathsf{VAR}(\boldsymbol{\omega}_k)$. The variances may be parametrized as *e.g.* $\mathsf{VAR}(\boldsymbol{\omega}_k) = \psi \cdot \mathrm{diag}(1, 0, 0, 1, 1)$ or $\mathsf{VAR}(\boldsymbol{\omega}_k) = \mathrm{diag}(\psi_1, \psi_2, \psi_2)$.

2.2. Inferential procedures

For a Gaussian state space model, we write $\boldsymbol{\theta}_k|D_k \sim \mathcal{N}_p(\mathbf{m}_k, \mathbf{C}_k)$, where D_k is all information available at time t_k. The Kalman filter recursively yields $\mathbf{m}_k$ and $\mathbf{C}_k$ with the recursion starting in $\boldsymbol{\theta}_0 \sim \mathcal{N}_p(\mathbf{m}_0, \mathbf{C}_o)$.

Assessment of the state vector, $\boldsymbol{\theta}_k$, using all available information, D_n, is called Kalman smoothing and we write $\boldsymbol{\theta}_n|D_n \sim \mathcal{N}_p(\widetilde{\mathbf{m}}_k, \widetilde{\mathbf{C}}_k)$. Starting with $\widetilde{\mathbf{m}}_n = \mathbf{m}_n$ and $\widetilde{\mathbf{C}}_n = \mathbf{C}_n$, the Kalman smoother is a backwards recursion in time, $k = n-1, \ldots, 1$.

For exponential family sampling distributions, *the iterated extended Kalman filter* yields an approximation to the conditional distribution of the state vector given D_n, see *e.g.* Durbin and Koopman (2000). By Taylor expansion, the sample distribution (4) is approximated with a Gaussian density, giving an approximating Gaussian state space model. The conditional distribution of the state vector given D_n in the exact model and in the Gaussian approximation have the same mode. The iterated extended Kalman filter is used as filter and smoother method in **sspir**.

2.3. Decomposition

The variation in the linear predictor, random or not, may be decomposed into four components: a time trend (T_k), harmonic seasonal patterns (H_k), unstructured seasonal patterns (S_k), and a regression with possibly time-varying covariates (R_k).

Each component may contain static and/or dynamic components, which is specified by zero and non-zero diagonal elements in $\mathsf{VAR}(\boldsymbol{\omega}_k)$, respectively, as described in the following.

The block-diagonal evolution matrix takes the form

$$\mathbf{G}_k = \begin{bmatrix} \mathbf{G}_k^{(1)} & & & \\ & \mathbf{I} & & \\ & & \mathbf{G}_k^{(3)} & \\ & & & \mathbf{I} \end{bmatrix},$$

where $\mathbf{G}_k^{(1)}$ is defined in (9), and $\mathbf{G}_k^{(3)}$ in (12). The components are only present if the model includes the corresponding terms.

The linear predictor,

$$\begin{aligned} \lambda_k &= \mathbf{T}_k\boldsymbol{\theta}_k^{(1)} + \mathbf{H}_k\boldsymbol{\theta}_k^{(2)} + \mathbf{S}_k\boldsymbol{\theta}_k^{(3)} + \mathbf{R}_k\boldsymbol{\theta}_k^{(4)} \\ &= T_k + H_k + S_k + R_k. \end{aligned}$$

will be detailed in the following.

Time trend

The long term trend is usually modelled by a sufficiently smooth function. In static regression models, this can be done by *e.g.* a high degree polynomial, a spline, or a generalized additive model. In the dynamic setting, however, a low degree polynomial with time-varying coefficient may suffice.

By stacking a polynomial, $q(t) = b_0 + b_1 t + \cdots + b_p t^p$, and the first p derivatives, the transition from t_{k-1} to t_k obeys the relation

$$\begin{bmatrix} q(t_k) \\ q'(t_k) \\ \vdots \\ q^{(p)}(t_k) \end{bmatrix} = \mathbf{G}_k^{(1)} \begin{bmatrix} q(t_{k-1}) \\ q'(t_{k-1}) \\ \vdots \\ q^{(p)}(t_{k-1}) \end{bmatrix}, \tag{8}$$

where $\Delta t_k = t_k - t_{k-1}$, and the upper triangular transition matrix is given by

$$\mathbf{G}_k^{(1)} = \begin{bmatrix} 1 & \Delta t_k & \cdots & \Delta t_k^p/p! \\ & 1 & \cdots & \Delta t_k^{p-1}/(p-1)! \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix}. \tag{9}$$

Using $\boldsymbol{\theta}_k^{(1)}$ for the left hand side of (8), a polynomial growth model with time-varying coefficients can be written as $\boldsymbol{\theta}_k^{(1)} = \mathbf{G}_k^{(1)}\boldsymbol{\theta}_{k-1}^{(1)} + \boldsymbol{\omega}_k^{(1)}$. The error term has variance $\mathsf{VAR}(\boldsymbol{\omega}_k^{(1)}) = \Delta t_k \mathbf{W}^{(1)}$, where $\mathbf{W}^{(1)}$ is diagonal in the case with independent random perturbations in each of the derivatives.

The trend component is the first element in $\boldsymbol{\theta}_k^{(1)}$, *i.e.*

$$T_k = \mathbf{T}_k\boldsymbol{\theta}_k^{(1)} = [\ 1 \quad 0 \quad \cdots \quad 0\ ]\ \boldsymbol{\theta}_k^{(1)}.$$

Alternatively, the time trend may be modelled as a random function, $q(t)$, for which the increments over time are described by a random walk, resulting in a cubic spline, see Kitagawa and Gersch (1984). The transition is the same as in (8) with $p = 2$, but only one variance parameter is necessary as,

$$\mathsf{VAR}(\boldsymbol{\omega}_k^{(1)}) = \sigma_w^2 \begin{bmatrix} \Delta t_k^3/3 & \Delta t_k^2/2 \\ \Delta t_k^2/2 & \Delta t_k \end{bmatrix}. \tag{10}$$

Harmonic seasonal pattern

Seasonal patterns with a given period, m, can be described by the following dth degree trigonometric polynomial

$$\begin{aligned} H_k &= \mathbf{H}_k \boldsymbol{\theta}_k^{(2)} \\ &= \sum_{i=1}^{d} \left\{ \theta_{c,i} \cos\left(i \cdot \frac{2\pi}{m} t_k \right) + \theta_{s,i} \sin\left(i \cdot \frac{2\pi}{m} t_k \right) \right\} \\ &= \begin{bmatrix} c_{1k} & \cdots & c_{dk} & s_{1k} & \cdots & s_{dk} \end{bmatrix} \boldsymbol{\theta}_k^{(2)}, \end{aligned} \tag{11}$$

where $c_{ik} = \cos(i \cdot 2\pi t_k/m)$ and $s_{ik} = \sin(i \cdot 2\pi t_k/m)$. This component can be used to describe seasonal effects showing cyclic patterns. Further seasonal components may be added for each period of interest.

The random fluctuations in $\boldsymbol{\theta}_k^{(2)}$ is modelled by a random walk, $\boldsymbol{\theta}_k^{(2)} = \boldsymbol{\theta}_{k-1}^{(2)} + \boldsymbol{\omega}_k^{(2)}$ with $\mathsf{VAR}(\boldsymbol{\omega}_k^{(2)}) = \Delta t_k \mathbf{W}^{(2)}$.

Unstructured seasonal component

For equidistant observations, a commonly used parameterization for the seasonal component is to let the effects, γ_k, for each period sum to zero in the static case, or to a white noise error sequence in the time-varying case, see Kitagawa and Gersch (1984). For an integer period, m, the sum-to-zero constraint can be expressed as $\sum_{i=0}^{m-1} \gamma_{k-i} = 0$ in the static case, and in the dynamic case, $\sum_{i=0}^{m-1} \gamma_{k-i} = \omega_k^{(3)}$, with $\omega_k^{(3)} \sim \mathcal{N}(0, \sigma_w^2)$. This is expressed in matrix form by letting $\boldsymbol{\theta}_k^{(3)} = [\gamma_k, \gamma_{k-1}, \ldots, \gamma_{k-m+2}]^\top$, and defining the $(m-1) \times (m-1)$ matrix

$$\mathbf{G}_k^{(3)} = \begin{bmatrix} -1 & -1 & \cdots & -1 \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}. \tag{12}$$

Then, $\boldsymbol{\theta}_k^{(3)} = \mathbf{G}_k^{(3)} \boldsymbol{\theta}_{k-1}^{(3)} + \boldsymbol{\omega}_k^{(3)}$, with $\mathsf{VAR}(\boldsymbol{\omega}_k^{(3)}) = \mathbf{W}^{(3)} = \operatorname{diag}(\sigma_w^2, 0, \ldots, 0)$ defines the evolution of the seasonal component. The corresponding term in the linear predictor is extracted by

$$S_k = \mathbf{S}_k \boldsymbol{\theta}_k^{(3)} = [\ 1 \quad 0 \quad \cdots \quad 0\ ] \boldsymbol{\theta}_k^{(3)}.$$

Regression component

Observed time-varying covariates, $\mathbf{R}_k$, enter the model through the usual regression term

$$R_k = \mathbf{R}_k \boldsymbol{\theta}_k^{(4)},$$

with $\boldsymbol{\theta}_k^{(4)} = \boldsymbol{\theta}_{k-1}^{(4)} + \boldsymbol{\omega}_k^{(4)}$ and $\mathsf{VAR}(\boldsymbol{\omega}_k^{(4)}) = \Delta t_k \mathbf{W}^{(4)}$. The structure of $\mathbf{W}^{(4)}$ is specified by the modeller and depends on the context.

3. Specification of state space objects

The package **sspir** can be downloaded and installed from `cran.r-project.org` and is then activated in R by `library(sspir)`. Assuming that the data are available either in a dataframe or in the current environment, then a state space model is setup using `glm`-style formula and family arguments. Terms are considered static unless embraced by the special function `tvar()`, described further in Section 3.2.

3.1. State space model objects

In **sspir**, a state space model is defined as an object from the class `ssm`. The object defines the model and contains the slots that are needed for the subsequent statistical analysis.

The definition of a state space model object has the following generic syntax

```
ssm(formula, family=gaussian, data, subset, time)
```

The call is designed to be closely connected to the `glm` call. A very important distinction is that it only *defines* the model and does not do any inferential calculations. There are many reasons for this. One reason is that models may be combined into more complex models. Another reason is that the models may be fitted using different inferential engines, and the model formulation should be independent of the choice of algorithm. The elements in the call are

formula a specification of the linear predictor (5) of the model. The syntax is defined in Section 3.2.

family a specification of the observation error distribution and link function to be used in the model, as in a `glm`-call. This can be a character string naming a family function, a family function or the result of a call to a family function. Currently, only Poisson with log-link, binomial with logit-link, and Gaussian with identity-link have been implemented. But it is straightforward to expand with further combinations within the exponential family.

data an optional data frame containing the variables in the model. By default the variables are taken from 'environment(formula)', typically the environment from which 'ssm' is called.

subset an optional vector specifying a subset of observations to be used in the fitting process.

time a numeric vector giving the time points, t_k, for the cases selected by `subset`.

3.2. Model formulas

A model formula is built up as in an `lm` or `glm` call in R. The response appears on the left hand side of a tilde (`~`) and on the right hand side the explanatory variables, factors and continuous

variables, appear. However, to specify time-varying regression coefficients, we have defined a special notation, `tvar()`, in which these are enclosed.
For example, the formula

```
y ~ z + tvar(x)
```

will correspond to covariates, `z` and `x`, of which `z` has a static parameter and `x` has a dynamic parameter. An implicit intercept is also included in the model, unless the term `-1` appears in the formula.When `tvar` enters a formula and `-1` is *not* included, the intercept will always be time-varying, *i.e.* a random walk is added to the linear predictor. Thus, this model corresponds to the state space model with the linear predictor specified as $\lambda_k = \mathbf{Z}_k\boldsymbol{\beta} + \mathbf{X}_k\boldsymbol{\beta}_k$, $\mathbf{Z}_k$ being the kth row in the $n \times 1$ matrix $\mathbf{Z} = [\mathrm{z}]$ and $\mathbf{X}_k$ the kth row of the $n \times 2$ matrix $\mathbf{X} = [1 \quad \mathrm{x}]$. The R command `model.matrix` applied to the formula `y ~ z + x` yields the $n \times 3$ matrix $[1 \quad z \quad x]$, in which the rows are $\mathbf{F}_k^\top$.
The polynomial time trend, (9), is specified using the function,

```
polytime(time,degree=p)
```

Note that `polytime` is different than the built-in R-function `poly` since the latter produces a design matrix with orthonormal columns.
The harmonic seasonal pattern, (11), is specified using the function,

```
polytrig(time,period=m,degree=d)
```

whereas the unstructured seasonal pattern, (12), is specified using the function,

```
sumseason(time,period=m)
```

Regression components are specified using the usual Wilkinson-Rogers formula notation in R.

The model matrix does not contain information about which variables are time-varying. This distinction is implemented by specifying the variance matrix, $\mathsf{VAR}(\boldsymbol{\omega}_k)$, with zeros in entries corresponding to static parameters and non-zero entries otherwise.

3.3. Inference

When a model has been defined, the function `kfs(ssm)` applies the *iterated extended Kalman smoother*, see Durbin and Koopman (2000), to yield an object containing the estimated mean and variance of the state vector, $\boldsymbol{\theta}_k$, as well as the approximate log-likelihood based on the Gaussian approximation to the state space model.

4. Examples

In this section, three examples of specification and application of state space models will be presented. The examples include Gaussian and Poisson observation densities. The time series are decomposed into components of trend and seasonality and also inclusion of external covariates is illustrated. The main focus will be on formulation of the state space object, how a relevant data analysis can be performed, and how to present the output from the analysis, based on this object.

Example 4.1 (Gas consumption)
A dataset provided with R *is the quarterly UK gas consumption from 1960 to 1986, in millions of therms (Durbin and Koopman 2001, p.233). As response, we use the (base 10) logarithm*

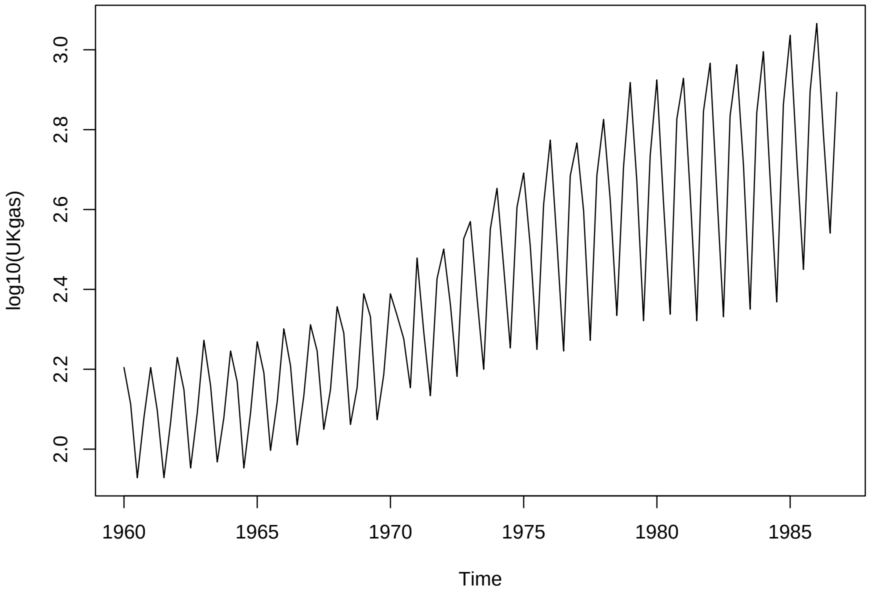

Figure 1: Log-transformed UK gas consumption, recorded quarterly from 1960 to 1986.

of the UK gas consumption (displayed in Figure 1), which we assume is normal distributed. We fit a model with a first order polynomial trend with time-varying coefficients and an unstructured seasonal component, also varying over time. The model is specified by

```
gasmodel <- ssm(log10(UKgas) ~ -1 + tvar(polytime(time,1)) +
                tvar(sumseason(time,4)), time=1:length(UKgas))
```

Here, the estimated variances are taken from an external maximum likelihood algorithm provided by the function `StructTS`*, Ripley (2002), which is standard in* R*. The decomposition in trend, slope and season components is displayed in Figure 2. In 1971, the slope increases from approximately 0.005 to approximately 0.014 and returns to this level in 1979. At the end of the observation period, the slope increases again. Similarly, it is seen, that the amplitude of the seasonal component is fairly constant from 1960-1971, after which it increases in the period 1971-1979 and then it stabilizes. The analysis can be reproduced in* **sspir** *by* `demo(gas)`*.*

Example 4.2 (Vandrivers)
Let y_t be the monthly numbers of light goods van drivers killed in road accidents, from January 1969 to December 1984 (192 observations). On January 31st, 1983, a seat belt law was introduced. The interest is to quantify the effect of the seat belt legislation law. For further information about the data set consult Harvey and Durbin (1986).

Here we use a state space model for Poisson data with a 13-dimensional latent process, consisting of an intervention parameter, `seatbelt`*, changing value from zero to one in February*

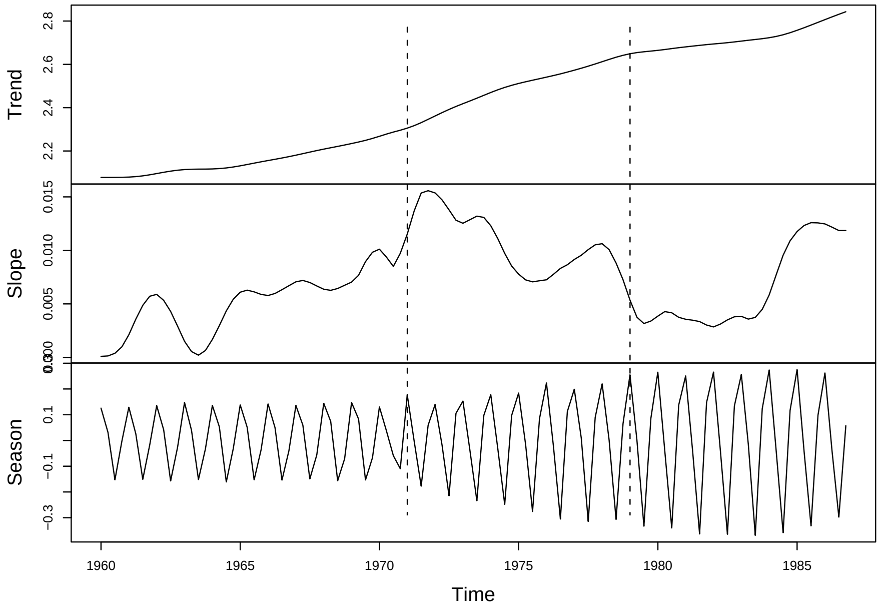

Figure 2: Time-varying trend, slope, and seasonal components in UK gas consumption.

1983, a constant monthly seasonal, and a trend modeled as a random walk.
The model can be defined in **sspir** *by*

```
vd <- ssm( y ~ tvar(1) + seatbelt + sumseason(time,12), time=time,
          family=poisson(link="log"), data=vandrivers)
```

The plot displayed in Figure 3 shows the observations together with the estimated trend and intervention. The superimposed confidence limits are ± 2 standard deviations, based on the conditional variance of the latent process, $\widetilde{\boldsymbol{C}}_t$, see Section 2.2. The trend over time is generally decreasing and the intervention effect corresponds approximately to a 25% reduction of casualties. The analysis can be reproduced in **sspir** *by* `demo(vandrivers)`.

Example 4.3 (Mumps)
Monthly registered cases of mumps in New York City, January 1928 through June 1972 has been studied by Hipel and McLeod (1994). The incidence of mumps are known to show seasonal behavior. In the study period the incidence also show variation in trend. The monthly sample variance grow with the monthly average, although substantial overdispersion is clearly present.

Fitting a Poisson generalized linear model with a quadratic trend and an monthly seasonal pattern, yields an overdispersion of 89.7, a significant trend and a significant seasonal variation. Changing the seasonal pattern to a harmonic pattern is in accordance with the data but does not substantially change the overdispersion.

We model the mumps incidence with a first order polynomial trend with time-varying coefficients and a time-varying harmonic seasonal component. This is done by the call

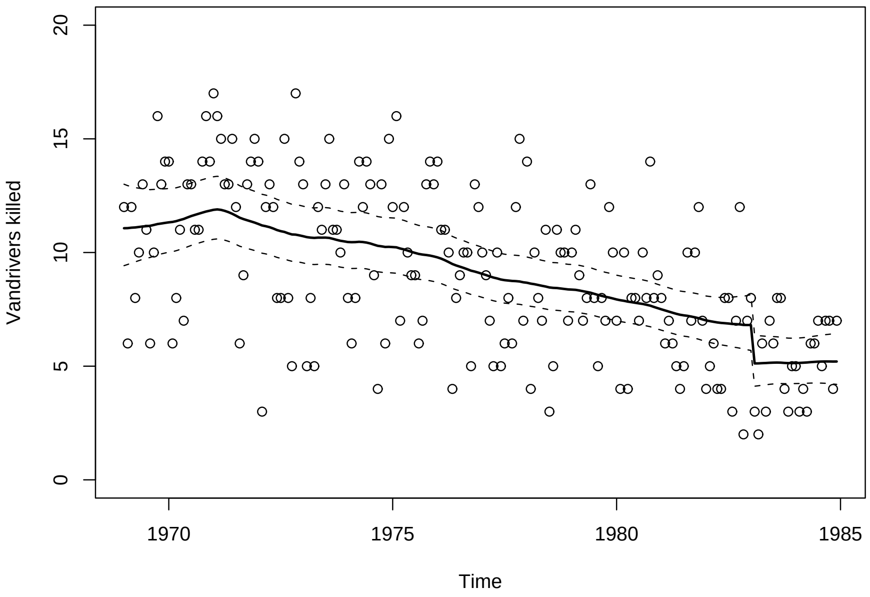

Figure 3: Estimated trend and intervention (solid line) for the vandrivers data. The dashed lines are ± 2 standard deviations.

```
ssm( mumps ~ -1 + tvar(polytime(time,1)) +
           tvar(polytrig(time,12,1)), family=poisson(link=log) )
```

The choice of a first order sinusoid gives the possibility to express the seasonal variation via the peak-to-trough ratio (yearly max/min) and the location of the peak. The output in Figure 4 shows a graduately changing seasonal pattern with a decreasing peak-to-trough ratio and a peak location slowly changing. The location of the peak is changing from late April in 1928–1936, where after the location of the peak stabilizes around May 1st until 1964, when the peak wanders off to late May, see Figure 4. It is also seen that the peak-to-trough ratio is varying between 6 to 9 until around 1948, when the ratio graduately decreases to about 4 in 1971. Epidemic episodes are seen irregularly each 3 to 5 years. The analysis can be reproduced in **sspir** *by* `demo(mumps)`.

5. Discussion

The main contribution of the **sspir** package is to give a formula language for specifying dynamic generalized linear models. That is, an extension of `glm` formulae by marking terms with `tvar` to specify that the corresponding coefficients are time-varying. The package also provides (extended) Kalman filter and Kalman smoother for models within the Gaussian, Poisson and binomial families. The output from the Kalman smoother leaves many possibilities for designing a suitable presentation of features of the latent process.

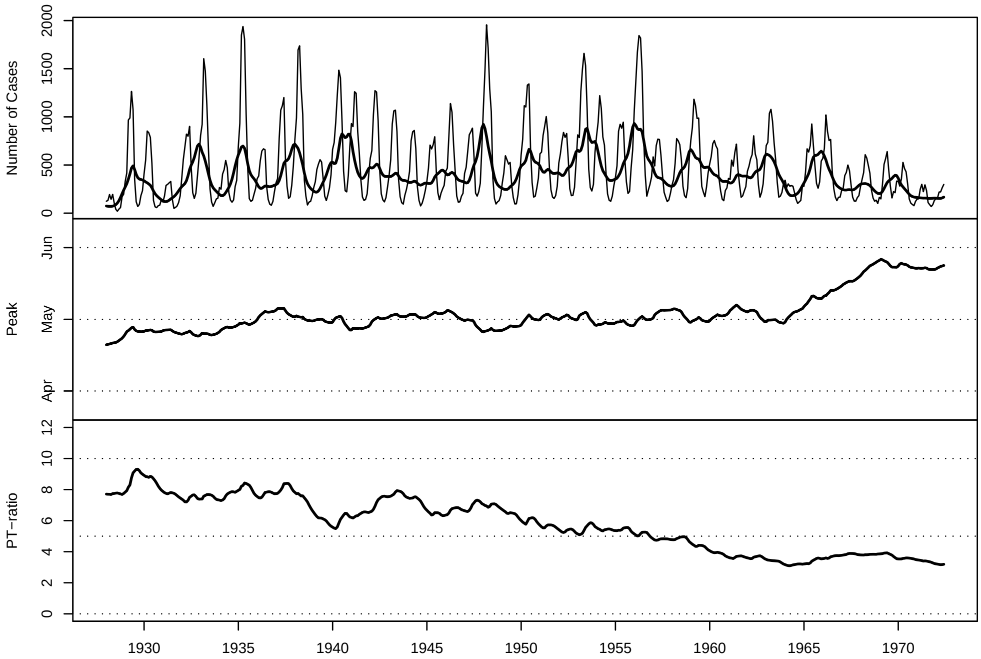

Figure 4: The variation in the incidence in mumps, NYC, 1927 – 1972. The upper frame shows the observed number of cases with the de-seasonalized trend superimposed. The middle frame shows the location of the peak of the seasonal pattern. The lower frame depicts the variation in the peak-to-trough ratio over the period.

The Kalman filter is initialized by the values of $\mathbf{m}_0$ and $\mathbf{C}_0$, see (3). The modeller can set entries in $\mathbf{C}_0$ to accommodate prior knowledge. In cases where the prior information about $\boldsymbol{\theta}_0$ is sparse, a diffuse initialization may be adequate, see Durbin and Koopman (2001). This feature has not yet been implemented.

The present framework does not allow the modeller to estimate the unknown variance parameters automatically. The modeller can, though, combine numerical maximization algorithms with the output of the iterated extended Kalman smoother. Hence, the formulation in **sspir** does not rely on any specific implementation of an inferential procedure.

References

Diggle PJ, Heagerty PJ, Liang KY, Zeger SL (2002). *Analysis of Longitudinal Data*. Oxford University Press, 2nd edition.

Durbin J, Koopman SJ (1997). "Monte Carlo maximum likelihood estimation for non-Gaussian state space models." *Biometrika*, **84**(3), 669–684.

Durbin J, Koopman SJ (2000). "Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with discussion)." *Journal of the Royal Statistical Society, Series B*, **62**(1), 3–56.

Durbin J, Koopman SJ (2001). *Time series analysis by state space methods*. Oxford University Press.

Gamerman D (1998). "Markov Chain Monte Carlo for Dynamic Generalised Linear Models." *Biometrika*, **85**, 215–227.

Harvey AC, Durbin J (1986). "The Effects of Seat Belt Legislation on British Road Casualties: A Case Study in Structural Time Series Modelling." *Journal of the Royal Statistical Society, Series A*, **149**(3), 187–227.

Hipel KW, McLeod IA (1994). *Time Series Modeling of Water Resources and Environmental Systems*. Elsevier Science Publishers B.V. (North-Holland).

Kitagawa G, Gersch W (1984). "A smoothness priors-state space modeling of time series with trend and seasonality." *Journal of the American Statistical Association*, **79**(386), 378–389.

Koopman SJ, Shephard N, Doornik JA (1999). "Statistical algorithms for models in state space using SsfPack 2.2." *Econometrics Journal*, **2**, 113–166.

McCullagh P, Nelder JA (1989). *Generalised Linear Models*. Chapman and Hall, London, 2nd edition.

Ripley B (2002). "Time Series in R 1.5.0." *R News*, **2**(2), 2–7.

West M, Harrison PJ, Migon HS (1985). "Dynamic generalized linear models and Bayesian forecasting (with discussion)." *Journal of the American Statistical Association*, **80**, 73–97.

Zeger SL (1988). "A Regression Model for Time Series of Counts." *Biometrika*, **75**(4), 621–629.