Overview over Dynamical Statistical Models – with a view towards application in biological sciences

Søren Højsgaard

March 5, 2009

Contents

1 Introduction

This paper is intended to give an overview over some aspects of dynamical statistical models (hereafter abbreviated DSMs) with a view towards applications in biological sciences. The specific view we have is to applications where data are measured online and where interest is utilizing data as they arrive (as opposed to waiting until a whole batch of data is available for an analysis).

To set the scene, we let y_t denote a measurement at time t. The observations available up to ancluding time t is denoted either as D_t or as $y_{1:t}$. Observations avaliable up to but not including time t are denoted $D_{<t}$ or $y_{<t}$. In the models we consider there are unknown parameters, generally denoted by θ. Following this, we are interested in models in which a new observation y_t measured at time t can readily be incorporated together with previous observations $y_{<t}$ to provide an updated estimate of a quantity of interest to be used in e.g. a prediction problems

The approaches to DSMs in the engineering, statistical and biological communities differ in many respects, but there is a gain to be made by combining the views.

2 Example: New Hampshire temperatures

The data are recordings of the mean annual temperature in degrees Fahrenheit in New Haven, Connecticut, from 1912 to 1971. Data are available in R. A plot of data is shown in Figure 1.

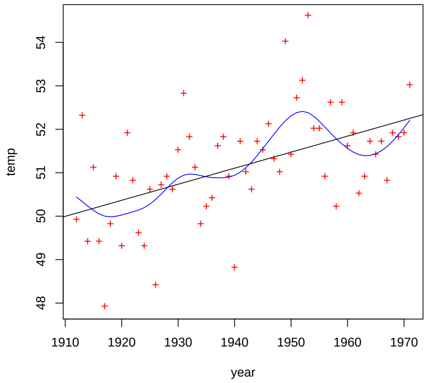

Figure 1: New Hampshire temperatures data with linear trend and smoothed spline added.

For simplicity we recode the year as $t = 1, 2, \ldots, 60$.

3 Regression and recursive least squares

Consider a regression setting model,

$$y_t \approx \alpha + \beta t \tag{1}$$

Suppose the task is to make predictions, and we focus on making 1-step forecasts. That is given information D_t we wish to predict the temperature in year $t+1$.

3.1 Recursive least squares

A naive approach is to refit the model to all available observations D_t every time a new measurement y_t is made. Let $\theta = (\alpha, \beta)$ and let $\hat{\theta}_t$ denote the estimate of θ based on data D_t. Thus $\hat{\theta}$ is obtained by minimizing the sum–of–squares,

$$SS = \sum_{t'=1}^{t} (y_{t'} - (\alpha + \beta t'))^2. \tag{2}$$

In a naive implementation of this approach we will have to store all data, which might be a problem in some applications. However it is possible (and in linear models it is quite simple) to avoid this: Instead we can find $\hat{\theta}_t$ from $\hat{\theta}_{t-1}$ and (y_t, t). This is called *recursive least squares* method.

3.2 Recursive least squares with forgetting factor

Note that with this method is that observations are "weighted equally" in (2); that is observations made in the far past are given the same weight as recent observations. This may not be appropriate. A very practical alternative is instead to minimize a weighted sum–of–squares,

$$SS = \sum_{t'=1}^{t} \kappa^{t-t'} (y_{t'} - (\alpha + \beta t'))^2. \tag{3}$$

for some constant κ where $0 < \kappa \leq 1$. This constant sometimes called a *forgetting factor* as it describes "how fast the past is forgotten" (down weighted) in the regression. Small values of κ means that the past is forgotten quickly. It is worth noticing that κ can not be estimated directly as part of a least squares procedure (letting κ go to zero means that SS goes to zero no matter the choice of θ). However, on can estimate κ along these lines: Suppose a batch of data is available. Let $\hat{y}_{t+1} = \hat{\alpha}_t + \hat{\beta}_t(t+1)$ be the 1–step prediction of y_{t+1}. Note that $\hat{y}_{t+1}$ depends on θ, κ so we might write $\hat{y}_{t+1} = \hat{y}_{t+1}(\theta, \kappa)$. Then the squared prediction error $(y_{t+1} - \hat{y}_{t+1}(\theta, \kappa))^2$ is a function of (θ, κ). We can then estimate (θ, κ) by minimizing (the non–linear) function:

$$SS = \sum_{t'=2}^{t} (y_{t+1} - \hat{y}_{t+1}(\theta, \kappa))^2. \tag{4}$$

Based on this we can obtain an estimate $\hat{\kappa}$ which can be used in the dynamic update in (3). To some people this approach may appear too much *ad hoc* while other find it perfectly acceptable.

See [Wikipedia] for a details and references.

4 Exponential smoothing and the Holt–Winters filter

In this section we discuss exponential smoothing and the Holt–Winters filter.

4.1 Exponential smoothing

Suppose there is no obvious trend in data (which does not apply to the New Hampshire temperature data), such that data can be described as "level + noise". A *one sided exponential smoothing* is obtained as

$$\alpha_t = \kappa y_t + (1-\kappa)\alpha_{t-1} \text{ with } \alpha_1 = y_1 \tag{5}$$

for some constant κ where $0 < \kappa \leq 1$. The 1–step ahead prediction is then simply $\hat{\alpha}t + 1 = \alpha_t$. It is illustrative to write (5) as

$$\alpha_t = \alpha_{t-1} + \kappa(y_t - \hat{\alpha}_t) \text{ with } \alpha_1 = y_1 \tag{6}$$

where $(y_t - \hat{\alpha}_t)$ can be regarded as the 1–step prediction error. Note that (5) has the property that data are incorporated as they arrive. That is, once α_{t-1} is calculated there is no need to store y_t (or any of the preceeding values for that matter). (The term exponential comes from the fact that observations from the past are downweight at an exponential rate.) Clearly, κ can not be estimated by minimizing the sum–of–squares, $\sum_t(y_t - \alpha_t)$ because that sum is zero if κ is zero. The parameter κ can instead be estimated by minimizing the 1–step prediction error, i.e. by minimizing $\sum_t(y_t - \hat{\alpha}_t)^2$ where $\hat{\alpha}_t$ is the 1–step prediction based on data up to time $t-1$.

See Brockwell and Davis (2002) and [Wikipedia] for a details and references.

4.2 Holt–Winters filter

If instead there is a (linear) trend in data (as seems to be the case for the New Hampshire data), an alternative is the Holt–Winters filter. (This filter applies to other situations as well, but we shall here focus on the linear case).

Consider a situation where we have data D_t and wish to make prediction at time $t+h$. The Holt–Winters filter does so as:

$$\hat{y}_{t+h} = \alpha_t + \beta_t h \tag{7}$$

The parameters α and β are updated as

$$\begin{aligned}\alpha_t &= \kappa y_t + (1-\kappa)(\alpha_{t-1} + \beta_{t-1}) && (8)\\ \beta_t &= \delta(\alpha_t - \alpha_{t-1}) + (1-\delta)(\beta_{t-1}) && (9)\end{aligned}$$

Hence there are now two smoothing parameters/forgetting factors, namely κ and δ. Note that setting $\delta = 0$ implies that (8) reduces to (5).

The smoothing parameters are usually estimated from a batch of data based on minimizing the 1–step prediction error.

See Winters (1960) and Brockwell and Davis (2002) for details and references.

5 State space models

Linear state space models (SSMs) are traditionally specified as consisting of

$$\begin{aligned}y_t &= F_t^\top \theta_t + v_t, \quad v_t \sim N(0, V_t) && (10)\\ \theta_t &= G_t \theta_{t-1} + w_t, \quad w_t \sim N(0, W_t) && (11)\\ \theta_0 &\sim N(m_0, C_0) && (12)\end{aligned}$$

known as respectively *observation equation*, the *system equation* and the *initial distribution* (of θ_0). The term $\lambda_t = F_t^\top \theta_t = \mathbb{E}(y_t|\theta_t)$ is the *signal*. Thus, technically the 6–tuple $(F_t, V_t, G_t, W_t, m_0, C_0)$ defines a linear state space model.

The state space model is a *probabilistic model* for all θ_ts and all y_ts. An alternative (but equivalent) formulation is

$$p(y, \theta) = p(\theta_0) \prod_t p(\theta_t|\theta_{t-1}) p(y_t|\theta_t)$$

where

$$\begin{aligned}p(\theta_t|\theta_{t-1}) &: N(G_t\theta_{t-1}, W_t), && (13)\\ p(y_t|\theta_t) &: N(F_t^\top \theta_t, V_t), && (14)\end{aligned}$$

5.1 Examples of SSMs

Many classical statistical models can be formulated as an SSM. Consider an autoregression of order 1, i.e.

$$y_t = \kappa y_{t-1} + v_t \qquad (15)$$

This model can be put in state space form as

$$\begin{aligned}y_t &= \alpha_t && (16)\\ \alpha_t &= \kappa \alpha_{t-1} + v_t && (17)\end{aligned}$$

6 The Kalman filter

Based on observed data we can then calculate (the distribution of) the parameters θ_t. The Kalman filter (named after its inventor, Rudolf Kalman) is, roughly speaking, an efficient algorithm for recursively estimating θ_t when new observations become available. The filter was developed in papers by Kalman (1960) and Kalman and S. (1961).

A wide variety of Kalman filters have now been developed, from Kalman's original formulation, now called the simple Kalman filter, to the extended Kalman filter, the information filter, a variety of square-root filters, unscented filter, particle filters.

See [Wikipedia] for an overview and references.

The Kalman filter works by "updating θ_t as new observations arrive", i.e. for $t = 1, \dots, T$: Given is that $\theta_{t-1}|D_{t-1} \sim N(m_{t-1}, C_{t-1})$. Then

$$\begin{aligned} \theta_t|D_{t-1} &\sim N(\overbrace{G_t, m_{t-p}}^{a_t}, \overbrace{G_t C_{t-1} G_t^\top + W_t}^{R_t}) && (18) \\ y_t|D_{t-1} &\sim N(\overbrace{F_t^\top a_t}^{f_t}, \overbrace{F_t^\top R_t F_t + V_t}^{Q_t}) && (19) \\ \theta_t|D_t &\sim N(\overbrace{a_t + \underbrace{R_t F_t Q_t^{-1}}_{A_t} \underbrace{(y_t - f_t)}_{e_t}}^{m_t}, \overbrace{R_t - A_t Q_t A_t^\top}^{C_t}) && (20) \end{aligned}$$

Hence, the only thing needed to be stored from time $t-1$ to time t is (m_{t-1}, C_{t-1}). Here f_t is the 1–step forecast, e_t is the *1–step forecast error* while A_t is the *Kalman gain.*

For the normal distribution an alternative formulation is as follows: Since $y_t = F_t^\top \theta_t + v_t$ we have $\mathbb{C}\text{ov}(y_t, \theta_t) = F_t \mathbb{V}\text{ar}(\theta_t)$ and so $\mathbb{C}\text{ov}(y_t, \theta_t|D_{t-1}) = F_t^\top R_t$. Hence

$$\left(\begin{matrix} y_t \\ \theta_t \end{matrix} |D_{t-1} \right) \sim N\left(\begin{bmatrix} f_t \\ a_t \end{bmatrix} \begin{bmatrix} Q_t & F_t^\top R_t \\ R_t F_t & R_t \end{bmatrix} \right) \tag{21}$$

6.1 The likelihood

The state space model depends on unknown parameters through V_t and W_t. There are several ways of estimating these; a brute force approach is the following: We have

$$p(y) = \prod_t p(y_t|y_{t-1}, \dots, y_1) = \prod_t p(y_t|D_{t-1})$$

We have $y_t|D_{t-1} \sim N(f_t, Q_t)$ so the log likelihood $\log L_t$ for a single observation is

$$\log L_t = -\frac{d}{2}\log(2\pi) - \frac{1}{2}(det(Q_t) + e_t^\top Q_t^{-1} e_t)$$

because $L_t = (2\pi)^{-d/2} det(Q_t)^{-1/2} \exp(-\frac{1}{2} e_t^\top Q_t^{-1} e_t)$. Hence the entire log likelihood is

$$\log L = -n\frac{d}{2}\log(2\pi) - \frac{1}{2}\sum_t (det(Q_t) + e_t^\top Q_t^{-1} e_t)$$

This likelihood can be maximized numerically, but it is often not entirely trivial to get this to work in specific applications.

6.2 Online outlier detection

Suppose y_t is p dimensional. Based on D_{t-1} the distribution of y_t is $y_t | D_{t-1} \sim N(f_t, Q_t)$ so the prediction error is

$$e_t = y_t - f_t \sim N(0, Q_t)$$

That means that when y_t is observed but before we update the equations, we can see how "likely" y_t is by noticing

$$ssd_t = e_t'(Q_t)^{-1} e_t \sim \chi_p^2$$

This provides a tool for on–line detection of outliers: A large value of ssd_t suggests that y_t is an "unlikely observation", that is, it might be erroneous.

7 Connecting exponential smoothing and state space models

There is a close connection between exponential smoothing and SSMs. Consider exponential smoothing as set out in Section 4 with

$$\alpha_t = \alpha_{t-1} + \kappa(y_t - \hat{\alpha}_t) \text{ with } \alpha_1 = y_1 \tag{22}$$

We can think of $e_t = (y_t - \hat{\alpha}_t)$ as a prediction error (which initially is zero). It is then trivially true that $y_t = \alpha_{t-1} + (y_t - \alpha_{t-1}) = \alpha_{t-1} + e_t$. Combining this with (22) gives

$$y_t = \alpha_{t-1} + e_t \tag{23}$$

$$\alpha_t = \alpha_{t-1} + \kappa e_t \tag{24}$$

which has a resemblence with the general for of a linear state space model in (12). There are however two important differences: The error term e_t appears both in the observation and system equation and there has been made no assumption about the distribution of the error term. If we assume the error terms to be normal and independent, then we are almost at (12). The fact that the error term appears in both the observation and system equation only changes the filter equations slightly; the Kalman filter can still be used.

This can be put in a more general setting: Let

$$y_t = F_t^\top \theta_{t-1} + e_t, \quad e_t \sim N(0, V_t) \tag{25}$$
$$\theta_t = G_t \theta_{t-1} + \kappa e_t \tag{26}$$

where κ now is a vector in which each element is in $[0;1]$. From this we find that

$$y_t = F_t^\top \theta_{t-1} + e_t, \quad e_t \sim N(0, V_t) \tag{27}$$
$$\theta_t = G_t \theta_{t-1} + \kappa(y_t - F_t^\top \theta_{t-1}) = D\theta_{t-1} + \kappa y_t, \tag{28}$$

say. So θ_t is hence a weighted sum of θ_{t-1} and t_t.

8 Discussion

Purists would often feel uncomfortable with the methods set out in Sections 3 and 4. They are not per se based on an assumption of an underlying statistical model. We can obtain parameter estimates, but the way the are obtained appears quite ad hoc and hence we do not know "how good they are". We can make predictions, but we do not know how good they are either. On the other hand, the methods are quite easy to work with in practice and are also easily implemented.

The state space models on the other hand appear attractive to some people because they are based on a proper statistical model, the parameter estimates have well understood interpretations (for example, (20) gives the conditional distribution of θ_t given D_t), the distribution of the prediction errors is known etc. In practice however, it is usually tricky to estimate the unknown parameters and therefore one often ends up by stipulating reasonable values.

With a view towards biology there are some additional reservations to be made. In physical/engineering applications there are often (but certainly not always) differential equations etc. which describe a natural form of the system equation. For example that $d\theta/dt = h(\theta, t)$ which can be translated into the approximation that $\theta_t \approx \theta_{t-1} + h(\theta_{t-1}, t-1)$ which yields the system equation. In biological sciences such quantitative descriptions are more rare. Therefore, it is often necessary to stipulate a system equation which essentially provides a smoothing of the state vector. It is in some cases therefore not clear whether it is worthwhile to invoke the entire "Kalman machinery" for that purpose (other simpler methods exist for that).

9 Acknowledgements

This work has been carried out in connection with the ILSORM project which is partly funded by The Danish National Advanced-Technology Foundation.

References

Peter J. Brockwell and Richard A. Davis. *Introduction to time series and forecasting.* Springer New York, 2 edition, 2002.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering*, 82:35–45, 1960.

R. E. Kalman and Bucy R. S. New results in linear filtering and prediction theory. *Tranactions of the ASME - Journal of Basic Engineering*, 83:95–107, 1961.

P.R. Winters. Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3):324–342, 1960.