# Sequential Bayesian Updating

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lectures 14 and 15, Hilary Term 2009

May 28, 2009

We consider data arriving sequentially $X_1, \dots, X_n, \dots$ and wish to update inference on an unknown parameter $\theta$ online.

In a Bayesian setting, we have a prior distribution $\pi(\theta)$ and at time $n$ we have a density for data conditional on $\theta$ as

$$f(x_1, \dots, x_n \mid \theta) = f(x_1 \mid \theta) f(x_2 \mid x_1, \theta) \cdots f(x_n \mid \mathbf{x_{n-1}}, \theta)$$

where we have let $\mathbf{x_i} = (x_1, \dots, x_i)$. Note that we are not assuming $X_1, \dots, X_n, \dots$ to be independent conditionally on $\theta$.

At time $n$, we may have updated our distribution of $\theta$ to its posterior

$$\pi_n(\theta) = f(\theta \mid \mathbf{x_n}) \propto \pi(\theta) f(\mathbf{x_n} \mid \theta).$$

If we obtain a new observation $X_{n+1} = x_{n+1}$ we may either start afresh and write

$$\pi_{n+1}(\theta) = f(\theta \,|\, \mathbf{x_{n+1}}) \propto \pi(\theta) f(\mathbf{x_{n+1}} \,|\, \theta)$$

or we could claim that just before time $n+1$, our knowledge of $\theta$ is summarized in the distribution $\pi_n(\theta)$ so we just use this as a prior distribution for the new piece of information and update as

$$\tilde{\pi}_{n+1}(\theta) \propto \pi_n(\theta) f(x_{n+1} \,|\, \mathbf{x_n}, \theta).$$

Indeed, *these updates are identical* since

$$\begin{aligned} \tilde{\pi}_{n+1}(\theta) &\propto \pi_n(\theta) f(x_{n+1} \,|\, \mathbf{x_n}, \theta) \\ &\propto \pi(\theta) f(\mathbf{x_n} \,|\, \theta) f(x_{n+1} \,|\, \mathbf{x_n}, \theta) \\ &= \pi(\theta) f(\mathbf{x_{n+1}} \,|\, \theta) \propto \pi_{n+1}(\theta). \end{aligned}$$

We may summarize these facts by replacing the usual expression for a Bayesian updating scheme

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

with

$$\textit{revised} \propto \textit{current} \times \textit{new likelihood}$$

represented by the formula

$$\pi_{n+1}(\theta) \propto \pi_n(\theta) \times L_{n+1}(\theta) = \pi_n(\theta) f(x_{n+1} \mid \mathbf{x_n}, \theta).$$

In this dynamic perspective we notice that at time $n$ we only need to keep a representation of $\pi_n$ and otherwise can ignore the past.

*The current $\pi_n$ contains all information needed to revise knowledge when confronted with new information $L_{n+1}(\theta)$.*

We sometimes refer to this way of updating as *recursive*.

The previous considerations take on a particular dynamic form when also the parameter or *state* $\theta$ is changing with time. More precisely, we consider a Markovian model for the *state dynamics* of the form

$$f(\theta_0) = \pi(\theta_0), \quad f(\theta_{i+1} \,|\, \theta_{\mathbf{i}}) = f(\theta_{i+1} \,|\, \theta_i)$$

where the evolving states $\theta_0, \theta_1, \ldots$ are not directly observed, but information about them are available through sequential *observations* $X_i = x_i$, where

$$f(x_i \,|\, \theta_{\mathbf{i}}, \mathbf{x_{i-1}}) = f(x_i \,|\, \theta_i)$$

so the joint density of states and observations is

$$f(\mathbf{x_n}, \theta_{\mathbf{n}}) = \pi(\theta_0) \prod_{i=1}^{n} f(\theta_i \,|\, \theta_{i-1}) f(x_i \,|\, \theta_i).$$

This type of model is common in robotics, speech recognition, target tracking, and steering/control, for example of large ships, airplanes, and space ships.

The natural tasks associated with inference about the evolving state $\theta_i$ are known as

- *Filtering:* Find $f(\theta_n \mid \mathbf{x_n})$. What is the current state?

This type of model is common in robotics, speech recognition, target tracking, and steering/control, for example of large ships, airplanes, and space ships.

The natural tasks associated with inference about the evolving state $\theta_i$ are known as

- *Filtering:* Find $f(\theta_n \mid \mathbf{x_n})$. What is the current state?
- *Prediction:* Find $f(\theta_{n+1} \mid \mathbf{x_n})$. What is the next state?

This type of model is common in robotics, speech recognition, target tracking, and steering/control, for example of large ships, airplanes, and space ships.

The natural tasks associated with inference about the evolving state $\theta_i$ are known as

- *Filtering:* Find $f(\theta_n \mid \mathbf{x_n})$. What is the current state?
- *Prediction:* Find $f(\theta_{n+1} \mid \mathbf{x_n})$. What is the next state?
- *Smoothing:* Find $f(\theta_j \mid \mathbf{x_n}), j < n$. What was the past state at time $j$?

If the filter distribution $f(\theta_n \,|\, \mathbf{x_n})$ is available we may calculate the *predictive distribution* as

$$f(\theta_{n+1} \,|\, \mathbf{x_n}) = \int_{\theta_n} f(\theta_{n+1} \,|\, \theta_n) f(\theta_n \,|\, \mathbf{x_n}) \, d\theta_n \tag{1}$$

which uses the current filter distribution and the dynamic model. When a new observation $X_{n+1} = x_{n+1}$ is obtained, we can use

$$\text{revised} \propto \text{current} \times \text{new likelihood}$$

to update the filter distribution as

$$f(\theta_{n+1} \,|\, \mathbf{x_{n+1}}) \propto f(\theta_{n+1} \,|\, \mathbf{x_n}) f(x_{n+1} \,|\, \theta_{n+1}), \tag{2}$$

i.e. the *updated filter distribution is found by combining the current predictive with the incoming likelihood.* The predictive distributions can now be updated to yield a general recursive scheme of *predict-observe-filter-predict-observe-filter...*

When we have more time, we may similarly look retrospectively and try to reconstruct the movements of $\theta$. This calculation is slightly more subtle than filtering. We first get

$$\begin{aligned} f(\theta_{j-1} \mid \mathbf{x_n}) &= \int_{\theta_j} f(\theta_{j-1} \mid \theta_j, \mathbf{x_n}) f(\theta_j \mid \mathbf{x_n})\, d\theta_j \\ &= \int_{\theta_j} f(\theta_{j-1} \mid \theta_j, \mathbf{x_{j-1}}) f(\theta_j \mid \mathbf{x_n})\, d\theta_j, \end{aligned}$$

where we have used that

$$\begin{aligned} f(\theta_{j-1}, \theta_j, \mathbf{x_n}) &= f(\theta_{j-1}, \theta_j, \mathbf{x_{j-1}}) f(x_j, \ldots, x_n \mid \theta_j, \theta_{j-1}) \\ &= f(\theta_{j-1}, \theta_j, \mathbf{x_{j-1}}) f(x_j, \ldots, x_n \mid \theta_j) \end{aligned}$$

and

$$f(\theta_j, \mathbf{x_n}) = f(\theta_j, \mathbf{x_{j-1}}) f(x_j, \ldots, x_n \mid \theta_j)$$

so

$$f(\theta_{j-1} \mid \theta_j, \mathbf{x_n}) = f(\theta_{j-1} \mid \theta_j, \mathbf{x_{j-1}}).$$

Since we can think of $f(\theta_j \mid \theta_{j-1})$ as a likelihood in $\theta_j$ we further get

$$f(\theta_{j-1} \mid \theta_j, \mathbf{x_{j-1}}) \propto f(\theta_j \mid \theta_{j-1}) f(\theta_{j-1} \mid \mathbf{x_{j-1}})$$

we thus get

$$\begin{aligned} f(\theta_{j-1} \mid \mathbf{x_n}) &\propto \int_{\theta_j} f(\theta_j \mid \theta_{j-1}) f(\theta_{j-1} \mid \mathbf{x_{j-1}}) f(\theta_j \mid \mathbf{x_n})\, d\theta_j \\ &\propto f(\theta_{j-1} \mid \mathbf{x_{j-1}}) \int_{\theta_j} f(\theta_j \mid \theta_{j-1}) f(\theta_j \mid \mathbf{x_n})\, d\theta_j, \end{aligned}$$

Which is the basic *smoothing recursion*:

$$f(\theta_{j-1} \mid \mathbf{x_n}) \propto f(\theta_{j-1} \mid \mathbf{x_{j-1}}) \int_{\theta_j} f(\theta_j \mid \theta_{j-1}) f(\theta_j \mid \mathbf{x_n})\, d\theta_j. \qquad (3)$$

It demands that we have stored a representation of the filter distributions $f(\theta_{j-1} \mid \mathbf{x_{j-1}})$ as well as the dynamic state model.

A special case of the previous is traditionally attributed to Kalman from a result in 1960, and known as the *Kalman filter and smoother* but was in fact developed in full detail by the Danish statistician T.N. Thiele in 1880.

It is based on the Markovian state model

$$\theta_{i+1} \mid \theta_i \sim \mathcal{N}(\theta_i, \sigma^2_{i+1}), \quad \theta_0 = 0$$

and the simple observational model

$$X_i \mid \theta_i \sim \mathcal{N}(\theta_i, \tau_i^2), i = 1, \ldots$$

where typically $\sigma_i^2 = (t_i - t_{i-1})\sigma^2$ and $\tau_i^2 = \tau^2$ with $t_i$ denoting the time of the $i$th observation. For simplicity we shall assume $t_i = i$ and $w_i = 1$ in the following.

The filtering relations become particularly simple, since the conditional distributions all are normal, and we are only concerned with expectations and variances.

We repeat Thiele's argument as an instance of the general theory developed.

Suppose at time $n$ we have the filter distribution of $\theta_n$ as $\mathcal{N}(\mu_n, \omega_n^2)$. Then the predictive distribution of $\theta_{n+1}$ is

$$\theta_{n+1} \,|\, \mathbf{x_n} \sim \mathcal{N}(\mu_n, \omega_n^2 + \sigma^2).$$

We can think of $\mu_n$ as our *current best measurement* of $\theta_{n+1}$, with this variance.

The contribution from the observation is a measurement of $\theta_{n+1}$ with a value of $x_{n+1}$ and a variance $\tau^2$. The best way of combining these estimates is to take a weighted average with the inverse variances as weights.

It follows that our new filter distribution has expectation

$$\mu_{n+1} = \frac{\mu_n/(\omega_n^2+\sigma^2) + x_{n+1}/\tau^2}{(\omega_n^2+\sigma^2)^{-1}+\tau^{-2}} = \frac{\tau^2\mu_n + (\sigma_2+\omega_n^2)x_{n+1}}{\tau^2+\sigma^2+\omega_n^2}$$

and variance

$$\omega_{n+1}^2 = \frac{1}{(\omega_n^2+\sigma^2)^{-1}+\tau^{-2}} = \frac{\tau^2(\sigma^2+\omega_n^2)}{\tau^2+\sigma^2+\omega_n^2}.$$

Clearly this result could also have been obtained from expanding the sum of squares in the expression for the filter distribution (2)

$$f(\theta_{n+1}\,|\,\mathbf{x_n}) \propto \exp\left\{-\frac{(\theta_{n+1}-\mu_n)^2}{2(\sigma^2+\omega_n^2)} + \frac{(\theta_{n+1}-x_{n+1})^2}{2\tau^2}\right\}.$$

We may elaborate the expression for $\mu_{n+1}$ and write it as a correction of $\mu_n$ or of $x_{n+1}$ as

$$\mu_{n+1} = \mu_n + \frac{\sigma^2 + \omega_n^2}{\tau^2 + \sigma^2 + \omega_n^2}(x_{n+1} - \mu_n)$$

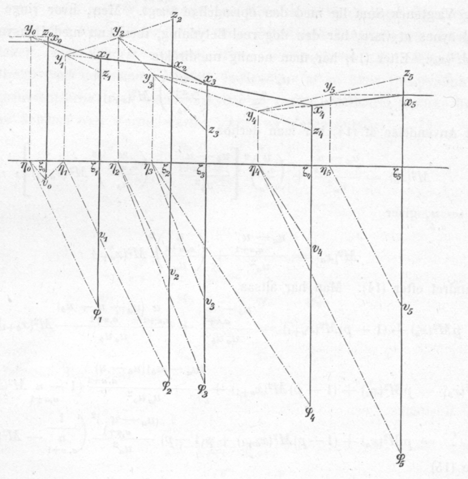
or

$$\mu_{n+1} = x_{n+1} - \frac{\tau^2}{\tau^2 + \sigma^2 + \omega_n^2}(x_{n+1} - \mu_n)$$

showing how at each stage $n$ the filtered value is obtained by modifying the observed and predicted values when the prediction is not on target.

The Kalman filter readily generalizes to the multivariate case and more complex models for the state evolution and observation equation. We abstain from further details.

This geometric construction of the Kalman filter and smoother is taken from Thiele (1880).



One of the most recent developments in modern statistics is using Monte Carlo methods for representing the predictive and filtered distributions.

We assume that we at time $n$ have represented the filter distribution (2) by a sample

$$f(\theta_n \,|\, \mathbf{x_n}) \sim \{\theta_n^1, \ldots, \theta^M\}$$

so that we would approximate any integral w.r.t. this density as

$$\int h(\theta_n) f(\theta_n \,|\, \mathbf{x_n})\, d\theta_n \approx \sum_{i=1}^{M} h(\theta_n^i).$$

The values $\{\theta_n^1, \ldots, \theta_n^M\}$ are generally referred to as *particles*.

More generally, we may have the particles associated with weights

$$f(\theta_n \,|\, \mathbf{x_n}) \sim \{(\theta_n^1, w_n^1), \ldots, (\theta^M, w_n^M)\}$$

with $\sum_{i=1}^M w_n^i = 1$, so that the integral is approximated by

$$\int h(\theta_n) f(\theta_n \,|\, \mathbf{x_n})\, d\theta_n \approx \sum_{i=1}^M h(\theta_n^i) w_n^i. \tag{4}$$

Typically, $w_i$ will reflect that we have been sampling from a *proposal distribution* $g(\theta_n)$ rather than the *target distribution* $f(\theta_n \,|\, \mathbf{x_n})$ so the weights are calculated as

$$w_n^i = f(\theta_n^i \,|\, \mathbf{x_n}) / g(\theta_n^i).$$

When filtering to obtain particles representing the next stage of the filtering distribution we *move* each particle a random amount by drawing $\theta^i_{n+1}$ at random from a proposal distribution $g_{n+1}(\theta \,|\, \theta^i_n, \mathbf{x_{n+1}})$ and subsequently *reweight* the particle as

$$w^i_{n+1} \propto w^i_n \frac{f(\theta^i_{n+1} \,|\, \theta^i_n) f(x_{n+1} \,|\, \theta^i_{n+1})}{g_{n+1}(\theta^i_{n+1} \,|\, \theta^i_n, \mathbf{x_{n+1}})}$$

the numerator being proportional to $f(\theta^i_{n+1} \,|\, \theta^i_n, x_{n+1})$.

There are many possible proposal distributions but a common choice is a normal distribution with an approximately correct mean and slightly enlarged variance.

The approximate inverse variance of the integral (4) is for the constant function $h \equiv 1$ equal to

$$\tilde{M}_n = \frac{1}{\sum_i (w_n^i)^2}$$

which is known as *effective number of particles*. It is maximized for $w^i \equiv 1/M$ which represents weights obtained when sampling from the correct distribution.

As the filtering evolves, it may happen that some weights become very small, reflecting bad particles, which are placed in areas of small probability. This leads to the effective number of particles becoming small.

To get rid of these, $M$ new particles are *resampled* with replacement, the probability for choosing particle $i$ at each sampling being equal to $w^i$ so that bad particles have high probability of not being included. This creates now a new set of particles which now all have weight $1/M$.
However, some particles will now be repeated in the sample and when this has been done many times, there may be only few particles left.

Various schemes then exist for *replenishing* and sampling new particles.

This can also be done routinely at each filtering, for example by first sampling two new particles for every existing one and subsequently resampling as above to retain exactly $M$ particles.