

$p_t(\mathbf{x}_t)$. In most applications of interest, it is not possible to obtain exact samples from these evolving distributions, so the goal is to reuse an approximate sample, representative of $p_{t-1}(\mathbf{x}_{t-1})$, to obtain a good representation of $p_t(\mathbf{x}_t)$. Moreover, since inference is to be performed in real time as new observations arrive, it is necessary that the computational cost be fixed in t . We will see in the sequel that SMC methods are highly flexible, and widely applicable; we restrict our attention to a particular class of dynamic models called the *state-space model* (SSM).

Chapter 7

State-space models and the Kalman filter algorithm

7.1 Motivation

In many real-world applications, observations arrive sequentially in time, and interest lies in performing on-line inference about unknown quantities from the given observations. If prior knowledge about these quantities is available, then it is possible to formulate a Bayesian model that incorporates this knowledge in the form of a prior distribution on the unknown quantities, and relates these to the observations via a likelihood function. Inference on the unknown quantities is then based on the posterior distribution obtained from Bayes' theorem. Examples include tracking an aircraft using radar measurements, speech recognition using noisy measurements of voice signals, or estimating the volatility of financial instruments using stock market data. In these examples, the unknown quantities of interest might be the location and velocity of the aircraft, the words in the speech signal, and the variance-covariance structure, respectively. In all three examples, the data is modelled dynamically in the sense that the underlying distribution evolves in time; these models are known as *dynamic models*. Sequential Monte Carlo (SMC) methods are a non-iterative, alternative class of algorithms to MCMC, designed specifically for inference in dynamic models. A comprehensive introduction to these methods is the book by Doucet et al. (2001). We point out that SMC methods are applicable in settings beyond dynamic models, such as non-sequential Bayesian inference, rare events simulation, and global optimization, provided that it is possible to define an evolving sequence of artificial distributions from which the distribution of interest is obtained via marginalisation.

Let $p_t(\mathbf{x}_t)$ denote the distribution at time $t \geq 1$, where $\mathbf{x}_t = (x_1, \dots, x_t)$ typically increases in dimension with t , but it is possible that the dimension of \mathbf{x}_t be constant $\forall t \geq 1$, or that \mathbf{x}_t have one dimension less than \mathbf{x}_{t-1} . The particular feature of dynamic models is the evolving nature of the underlying distribution, where $p_t(\mathbf{x}_t)$ changes in time t as new observations are generated. Note that \mathbf{x}_t are the quantities of interest, not the observations; the observations up to time t determine the form of the distribution, and this is implied by the subscript t in $p_t(\cdot)$. This is in contrast to non-dynamic models where the distribution is constant as new observations are generated, denoted by $p(\mathbf{x})$. In the latter case, MCMC methods have proven highly effective in generating approximate samples from low-dimensional distributions $p(\mathbf{x})$, when exact simulation is not possible. In the dynamic case, at each time step t a different MCMC sampler with stationary distribution $p_t(\mathbf{x}_t)$ is required, so the overall computational cost would increase with t . Moreover, for large t , designing the sampler and assessing its convergence would be increasingly difficult.

SMC methods are a non-iterative alternative to MCMC algorithms, based on the key idea that if $p_{t-1}(\mathbf{x}_{t-1})$ does not differ much from $p_t(\mathbf{x}_t)$, then it is possible to reuse the samples from $p_{t-1}(\mathbf{x}_{t-1})$ to obtain samples from

7.2 State-space models

SSMs are a class of dynamic models that consist of an underlying Markov process, usually called the *state process*, X_t , that is hidden, i.e., unobserved, and an observed process, usually called the *observation process*, Y_t . Consider the following notation for a state-space model:

$$\begin{aligned} \text{observation: } & y_t = a(x_t, u_t) \sim g(\cdot|x_t, \phi) \\ \text{hidden state: } & x_t = b(x_{t-1}, v_t) \sim f(\cdot|x_{t-1}, \theta), \end{aligned}$$

where y_t and x_t are generated by functions $a(\cdot)$ and $b(\cdot)$ of the state and noise disturbances, denoted by u_t and v_t , respectively. Assume ϕ and θ to be known. Let $p(x_1)$ denote the distribution of the initial state x_1 . The state process is a Markov chain, i.e., $p(x_t|x_1, \dots, x_{t-1}) = p(x_t|x_{t-1}) = f(x_t|x_{t-1}, \theta)$, and the distribution of the observation y_t , conditional on x_t , is independent of previous values of the state and observation processes, i.e., $p(y_t|x_{1:t}, y_{1:t-1}) = p(y_t|x_t) = g(y_t|x_t, \phi)$. See Figure 7.1 for illustration.

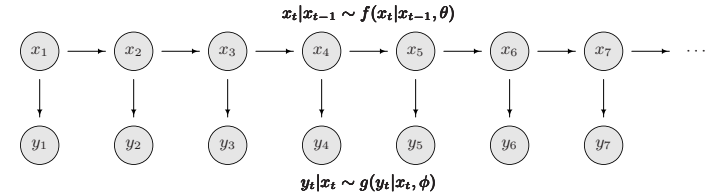


Figure 7.1. The conditional independence structure of the first few states and observations in a hidden Markov Model.

Note that we use the notation $x_{1:t}$ to denote x_1, \dots, x_t , and similarly for $y_{1:t}$. For simplicity, we drop the explicit dependence of the state transition and observation densities on θ and ϕ , and write $f(\cdot|x_{t-1})$, and $g(\cdot|x_t)$.

The literature sometimes distinguishes between state-space models where the state process is given by a discrete Markov chain, called *hidden Markov models* (HMM), as opposed to a continuous Markov chain. An extensive monograph on inference for state-space models is the book by Cappé et al. (2005), and a more recent overview is Cappé et al. (2007). In the present chapter and the following, we introduce several algorithms for inference in state-space models, and point out that the algorithms in Chapter 8 apply more generally to dynamic models.

7.2.1 Inference problems in SSMs

Under the notation introduced above, we have the joint density

$$p(x_{1:t}, y_{1:t}) = p(x_1)g(y_1|x_1) \prod_{i=2}^t p(x_i, y_i|x_{1:i-1}, y_{1:i-1}) = p(x_1)g(y_1|x_1) \prod_{i=2}^t f(x_i|x_{i-1})g(y_i|x_i),$$

and, by Bayes' theorem, the density of the distribution of interest

$$p(x_{1:t}|y_{1:t}) \propto p(x_{1:t}|y_{1:t-1})g(y_t|x_t) = p(x_{1:t-1}|y_{1:t-1})f(x_t|x_{t-1})g(y_t|x_t). \quad (7.1)$$

To connect this with the notation introduced for dynamic models, we can write $p_t(x_{1:t}) = p(x_{1:t}|y_{1:t})$, but we believe that stating the dependence on the observations explicitly leads to less confusion.

There exist several inference problems in state-space models that involve computing the posterior distribution of a collection of state variables conditional on a batch of observations:

- *filtering*: $p(x_t|y_{1:t})$
- *fixed lag smoothing*: $p(x_{t-l}|y_{1:t})$, for $0 \leq l \leq t-1$
- *fixed interval smoothing*: $p(x_{l:k}|y_{1:t})$, for $1 \leq l < k \leq t$
- *prediction*: $p(x_{l:k}|y_{1:t})$, for $k > t$ and $1 \leq l \leq k$.

The first three inference problems reduce to marginalisation of the full smoothing distribution $p(x_{1:t}|y_{1:t})$, i.e., integrating over the state variables that are not of interest, whereas the fourth reduces to marginalisation of

$$p(x_{1:k}|y_{1:t}) = p(x_{1:t}|y_{1:t}) \prod_{i=t+1}^k f(x_i|x_{i-1}).$$

So far we assumed that the state transition and observation densities are completely characterised, i.e., that the parameters θ and ϕ are known. If they are unknown, then Bayesian inference is concerned with the joint posterior distribution of the hidden states and the parameters:

$$p(x_{1:t}, \theta, \phi|y_{1:t}) \propto p(y_{1:t}|x_{1:t}, \theta, \phi) p(x_{1:t}|\theta, \phi) p(\theta, \phi) = p(\theta, \phi) p(x_1) g(y_1|x_1, \phi) \prod_{i=2}^t f(x_i|x_{i-1}, \theta) g(y_i|x_i, \phi).$$

If interest lies in the posterior distribution of the parameters, then the inference problem is called:

- *static parameter estimation*: $p(\theta, \phi|y_{1:t})$,

which reduces to integrating over the state variables in the joint posterior distribution $p(x_{1:t}, \theta, \phi|y_{1:t})$.

It is evident, then, that these inference problems depend on the tractability of the posterior distribution $p(x_{1:t}|y_{1:t})$, if the parameters are known, or $p(x_{1:t}, \theta, \phi|y_{1:t})$, otherwise. Notice that equation (7.1) gives the posterior distribution up to a normalising constant $\int p(x_{1:t}|y_{1:t-1}) g(y_t|x_t) dx_{1:t}$, and it is oftentimes the case that the posterior distribution is known only up to a constant. In fact, these posterior distributions can be computed in closed form only in a few specific cases, such as the hidden Markov model, i.e., when the state process is a discrete Markov chain, and the linear Gaussian model, i.e., when the functions $a()$ and $b()$ are linear, and the noise disturbances u_t and v_t are Gaussian.

For HMMs with discrete state transition and observation distributions, the tutorial of Rabiner (1989) presents recursive algorithms for the smoothing and static parameter estimation problems. The Viterbi algorithm returns the optimal sequence of hidden states, i.e., the sequence that maximises the smoothing distribution, and the Expectation-Maximization (EM) algorithm returns parameter values for which the likelihood function of the observations attains a local maximum. If the observation distribution is continuous, then it can be approximated by a finite mixture of Gaussian distributions to insure that the EM algorithm applies to the problem of parameter estimation. These recursive algorithms involve summations over all states in the model, so they are impractical when the state space is large.

For the linear Gaussian model, the normalising constant in (7.1) can be computed analytically, and thus the posterior distribution of interest is known in closed form; in fact, it is the Gaussian distribution. The *Kalman filter* algorithm (Kalman, 1960) gives recursive expressions for the mean and variance of the filtering distribution $p(x_t|y_{1:t})$, under the assumption that all parameters in the model are known. Kalman (1960) obtains recursive expressions for the optimal values of the mean and variance parameters via a least-squares approach. The algorithm alternates between two steps: a prediction step (i.e., predict the state at time t conditional on $y_{1:t-1}$), and an update step (i.e., observe y_t , and update the prediction in light of the new observation). Section 7.3 presents a Bayesian formulation of the Kalman filter algorithm following Meinhold and Singpurwalla (1983).

When an analytic solution is intractable, exact inference is replaced by inference based on an approximation to the posterior distribution of interest. Grid-based methods using discrete numerical approximations to these posterior distributions are severely limited by parameter dimension. Alternatively, sequential Monte Carlo methods are a simulation-based approach that offer greater flexibility and scale better with increasing dimensionality. The key idea of SMC methods is to represent the posterior distribution by a weighted set of samples, called *particles*, that are *filtered* in time as new observations arrive, through a combination of sampling and resampling steps. Hence SMC sampling algorithms are oftentimes called *particle filters* (Carpenter et al., 1999). Chapter 8 presents SMC methods for the problems of filtering and smoothing.

7.3 The Kalman filter algorithm

From (7.1), the posterior distribution of the state x_t conditional on the observations $y_{1:t}$ is proportional to $p(x_t|y_{1:t-1})g(y_t|x_t)$. The first term is the distribution of x_t conditional on the first $t-1$ observations; computing this distribution is known as the *prediction* step. The second term is the distribution of the new observation y_t conditional on the hidden state at time t . Updating $p(x_t|y_{1:t-1})$ in light of the new observation involves taking the product of these two terms, and normalising; this is known as the *update* step. The result is the distribution of interest:

$$p(x_t|y_{1:t}) = \int p(x_{1:t}|y_{1:t}) dx_1 \dots dx_{t-1} = \frac{p(x_t|y_{1:t-1})g(y_t|x_t)}{\int p(x_t|y_{1:t-1})g(y_t|x_t) dx_t}.$$

We now show how the prediction and update stages can be performed exactly for the linear Gaussian state-space model, which is represented as follows:

$$\text{observation: } y_t = A_t x_t + u_t \sim \mathcal{N}(A_t x_t, \Phi^2) \quad (7.2)$$

$$\text{hidden state: } x_t = B_t x_{t-1} + v_t \sim \mathcal{N}(B_t x_{t-1}, \Theta^2), \quad (7.3)$$

where $u_t \sim \mathcal{N}(0, \Phi^2)$ and $v_t \sim \mathcal{N}(0, \Theta^2)$ are independent noise sequences, and the parameters A_t , B_t , Φ^2 , and Θ^2 are known. It is also possible to let the noise variances Φ^2 and Θ^2 vary with time; the derivation of the mean and variance of the posterior distribution follows as detailed below. We assume that both states and observations are vectors, in which case the parameters are matrices of appropriate sizes.

The Kalman filter algorithm proceeds as follows. Start with an initial Gaussian distribution on x_1 : $x_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$. At time $t-1$, let μ_{t-1} and Σ_{t-1} be the mean and variance of the Gaussian distribution of x_{t-1} conditional on $y_{1:t-1}$. Looking forward to time t , we begin by predicting the distribution of x_t conditional on $y_{1:t-1}$.

Prediction step: From equation (7.3), $x_t = B_t x_{t-1} + v_t$, where $x_{t-1}|y_{1:t-1} \sim \mathcal{N}(\mu_{t-1}, \Sigma_{t-1})$, and $v_t \sim \mathcal{N}(0, \Theta^2)$ independently. By results in multivariate statistical analysis (Anderson, 2003), we have that

$$x_t|y_{1:t-1} \sim \mathcal{N}(B_t \mu_{t-1}, B_t \Sigma_{t-1} B_t^T + \Theta^2), \quad (7.4)$$

where the superscript T indicates matrix transpose. This can be thought of as the prior distribution on x_t .

Update step: Upon observing y_t , we are interested in

$$p(x_t|y_{1:t}) \propto p(y_t|x_t, y_{1:t-1})p(x_t|y_{1:t-1}).$$

Following equation (7.2) and the result in (7.4), consider predicting y_t by $\hat{y}_t = A_t B_t \mu_{t-1}$, where $B_t \mu_{t-1}$ is the prior mean on x_t . The prediction error is $e_t = y_t - \hat{y}_t = y_t - A_t B_t \mu_{t-1}$, which is equivalent to observing y_t . So it follows that $p(x_t|y_{1:t}) \propto p(e_t|x_t, y_{1:t-1})p(x_t|y_{1:t-1})$. Finally, from (7.2), $e_t = A_t(x_t - B_t \mu_{t-1}) + u_t$, where $u_t \sim \mathcal{N}(0, \Phi^2)$, so $e_t|x_t, y_{1:t-1} \sim \mathcal{N}(A_t(x_t - B_t \mu_{t-1}), \Phi^2)$.

We now use the following results from Anderson (2003). Let X_1 and X_2 have a bivariate normal distribution:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right). \quad (7.5)$$

If equation (7.5) holds, then the conditional distribution of X_1 given $X_2 = x_2$ is given by

$$X_1 | X_2 = x_2 \sim \mathcal{N}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}). \quad (7.6)$$

Conversely, if (7.6) holds, and $X_2 \sim \mathcal{N}(\mu_2, \Sigma_{22})$, then (7.5) is true.

Since $e_t | x_t, y_{1:t-1} \sim \mathcal{N}(A_t(x_t - B_t\mu_{t-1}), \Phi^2)$ and $x_t | y_{1:t-1} \sim \mathcal{N}(B_t\mu_{t-1}, B_t\Sigma_{t-1}B_t^T + \Theta^2)$, it follows that

$$\begin{pmatrix} x_t \\ e_t \end{pmatrix} \Big| y_{1:t-1} \sim \mathcal{N} \left(\begin{pmatrix} B_t\mu_{t-1} \\ 0 \end{pmatrix}, \begin{pmatrix} B_t\Sigma_{t-1}B_t^T + \Theta^2 & (B_t\Sigma_{t-1}B_t^T + \Theta^2)A_t^T \\ A_t(B_t\Sigma_{t-1}B_t^T + \Theta^2) & A_t(B_t\Sigma_{t-1}B_t^T + \Theta^2)A_t^T + \Phi^2 \end{pmatrix} \right)$$

Using the result above, the filtering distribution is $p(x_t | y_{1:t}) = p(x_t | e_t, y_{1:t-1}) = \mathcal{N}(\mu_t, \Sigma_t)$, since observing e_t is equivalent to observing y_t , where

$$\begin{aligned} \mu_t &= B_t\mu_{t-1} + (B_t\Sigma_{t-1}B_t^T + \Theta^2)A_t^T(A_t(B_t\Sigma_{t-1}B_t^T + \Theta^2)A_t^T + \Phi^2)^{-1}e_t \\ \Sigma_t &= B_t\Sigma_{t-1}B_t^T + \Theta^2 - (B_t\Sigma_{t-1}B_t^T + \Theta^2)A_t^T(A_t(B_t\Sigma_{t-1}B_t^T + \Theta^2)A_t^T + \Phi^2)^{-1}A_t(B_t\Sigma_{t-1}B_t^T + \Theta^2). \end{aligned}$$

Algorithm 1 The Kalman filter algorithm.

- 1: Input: μ_1 and Σ_1 .
- 2: Set $t = 2$.
- 3: Compute mean and variance of prediction: $\hat{\mu}_t = B_t\mu_{t-1}$, $\hat{\Sigma}_t = B_t\Sigma_{t-1}B_t^T + \Theta^2$.
- 4: Observe y_t and compute error in prediction: $e_t = y_t - A_t\hat{\mu}_t$.
- 5: Compute variance of prediction error: $R_t = A_t\hat{\Sigma}_tA_t^T + \Phi^2$.
- 6: Update the mean and variance of the posterior distribution:

$$\begin{aligned} \mu_t &= \hat{\mu}_t + \hat{\Sigma}_tA_t^TR_t^{-1}e_t \\ \Sigma_t &= \hat{\Sigma}_t - \hat{\Sigma}_tA_t^TR_t^{-1}A_t\hat{\Sigma}_t. \end{aligned}$$

- 7: Set $t = t + 1$. Go to step 3.
-

Example 7.1 (First-order, linear autoregressive (AR(1)) model observed with noise). Consider the following AR(1) model:

$$\begin{aligned} x_t &= \phi x_{t-1} + \sigma_U u_t \sim \mathcal{N}(\phi x_{t-1}, \sigma_u^2) \\ y_t &= x_t + \sigma_V v_t \sim \mathcal{N}(x_t, \sigma_v^2), \end{aligned}$$

where $u_t \sim \mathcal{N}(0, 1)$ and $v_t \sim \mathcal{N}(0, 1)$ are independent, Gaussian white noise processes. The Markov chain $\{X_t\}_{t \geq 1}$ is a Gaussian random walk with transition kernel $\mathbf{K}(x_{t-1}, x_t)$ corresponding to the $\mathcal{N}(\phi x_{t-1}, \sigma_u^2)$ distribution.

A normal distribution $\mathcal{N}(\mu, \sigma^2)$ is stationary for $\{X_t\}_{t \geq 1}$ if $X_{t-1} \sim \mathcal{N}(\mu, \sigma^2)$ and $X_t | X_{t-1} = x_{t-1} \sim \mathcal{N}(\phi x_{t-1}, \sigma_u^2)$ imply that $X_t \sim \mathcal{N}(\mu, \sigma^2)$. We require that $\mathbb{E}(X_t) = \phi\mu = \mu$ and $\text{Var}(X_t) = \phi^2\sigma^2 + \sigma_u^2 = \sigma^2$, which are satisfied by $\mu = 0$ and $\sigma^2 = \sigma_u^2/(1 - \phi^2)$, provided $|\phi| < 1$. In fact, the $\mathcal{N}(0, \sigma_u^2/(1 - \phi^2))$ distribution is the unique stationary distribution of the chain.

Start the Kalman filter algorithm with $\mu_1 = 0$ and $\Sigma_1 = \sigma_u^2/(1 - \phi^2)$. At time $t - 1$, $t \geq 2$, let μ_{t-1} and Σ_{t-1} denote the posterior mean and variance, respectively. Then the mean and variance of the prediction at time t are: $\hat{\mu}_t = \phi\mu_{t-1}$ and $\hat{\Sigma}_t = \phi^2\Sigma_{t-1} + \sigma_u^2$. The prediction error is $e_t = y_t - \hat{\mu}_t$ with variance $\hat{\Sigma}_t + \sigma_v^2$. Finally, update the mean and variance of the posterior distribution:

$$\begin{aligned} \mu_t &= \hat{\mu}_t + \hat{\Sigma}_t \frac{1}{\hat{\Sigma}_t + \sigma_v^2} (y_t - \hat{\mu}_t) = \left(1 - \frac{\hat{\Sigma}_t}{\hat{\Sigma}_t + \sigma_v^2} \right) \hat{\mu}_t + \frac{\hat{\Sigma}_t}{\hat{\Sigma}_t + \sigma_v^2} y_t \\ \Sigma_t &= \hat{\Sigma}_t - \hat{\Sigma}_t \left(\frac{\hat{\Sigma}_t}{\hat{\Sigma}_t + \sigma_v^2} \right) \hat{\Sigma}_t = \hat{\Sigma}_t \left(1 - \frac{\hat{\Sigma}_t}{\hat{\Sigma}_t + \sigma_v^2} \right). \end{aligned}$$

The Kalman filter algorithm (see Figure 1 for pseudo-code) is not robust to outlying observations y_t , i.e., when the prediction error e_t is large, because the mean μ_t is an unbounded function of e_t , and the variance Σ_t does not depend on the observed data y_t . Meinhold and Singpurwalla (1989) let the distributions of the error terms u_t and v_t be Student- t , and show that the posterior distribution of x_t given $y_{1:t}$ converges to the prior distribution of $p(x_t | y_{1:t-1})$ when e_t is large. In this case, the posterior distribution is no longer known exactly, but is approximated.

The underlying assumptions of the Kalman filter algorithm are that the state transition and observation equations are linear, and that the error terms are normally distributed. If the linearity assumption is violated, but the state transition and observation equations are differentiable functions, then the *extended Kalman filter* algorithm propagates the mean and covariance via the Kalman filter equations by linearizing the underlying non-linear model. If this model is highly non-linear, then this approach will result in very poor estimates of the mean and covariance. An alternative is the *unscented Kalman filter* which takes a deterministic sampling approach, representing the state transition distribution by a set of sample points that are propagated through the non-linear model. This approach improves the accuracy of the posterior mean and covariance; for details, see Wan and van der Merwe (2000).