An Introduction to Probabilistic Graphical Models

Michael I. Jordan
University of California, Berkeley

June 30, 2003

Chapter 15

Kalman filtering and smoothing

Thus far we have presented two major categories of latent variable models: *mixture models*, which are based on a discrete latent variable, and *factor analysis models*, which are based on a continuous latent variable. The graphs underlying these models are identical—two-node graphs in which a single latent variable is connected to a single observable variable.

Chapter 12 presented a dynamical generalization of mixture models—the hidden Markov model (HMM). Graphically, the HMM was obtained by copying the two-node mixture model as a spatial array, connecting successive state nodes in the array. It is natural to wonder if a similar generalization of factor analysis might be worth considering. In fact the dynamical generalization of factor analysis is well worth considering—it yields an interesting and important methodology for time series analysis known as the *Kalman filter*. In fact, in an attempt to develop a consistent terminology, we reserve the term "Kalman filter" for the recursive inference algorithm that is the analog of the "alpha" algorithm in the HMM setting. The underlying model, which we refer to as the "state space model (SSM)," is structurally identical to the HMM; only the type of the nodes (real-valued vectors) and the probability model (linear-Gaussian) changes. The model has exactly the same Markov properties as the HMM, and its states are hidden in exactly the same way as in the HMM.

Historically, the HMM and the Kalman filtering methodology were developed in separate research communities and their close relationship has not always been widely appreciated. This is partly due to the fact that the general framework of graphical models came later than the HMM and the Kalman filter. Without the graphical framework, the algorithms underlying the inference calculation in the two cases look rather different (as we will see). This is, however, simply a reflection of the differences between the multinomial distribution and the Gaussian distribution, and it is imperative that we not let these details–important as they may be in practice–obscure the fundamental similarity between the two models.

We will develop the inference procedures for the SSM in some detail in this chapter. This is not only to acknowledge the historical importance of the Kalman filter, but also to provide an additional concrete example of the solution of the inference problem for a reasonably complex graphical model. Once we develop a general perspective on graphical models in Chapter 15, we will return to the SSM and the HMM, not only to provide concrete examples to ground our general theory, but also

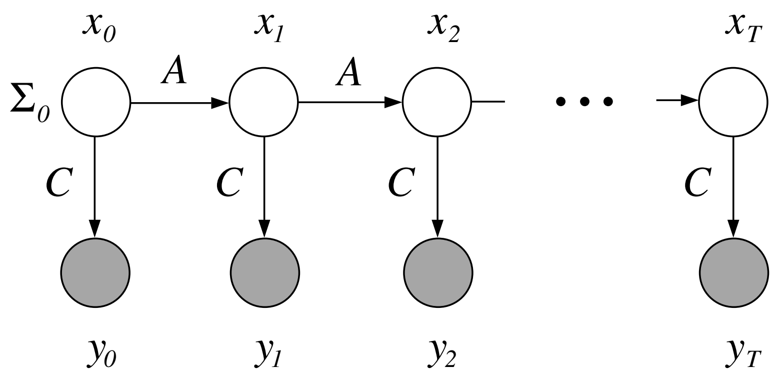

Figure 15.1: The SSM as a graphical model. Each vertical slice represents a time step. The top node in each slice represents the state variable x_t and the bottom node in each slice represents the observable output variable y_t.

to indicate that both are best viewed as jumping-off points for a much larger class of models.

15.1 The state space model

As we have already discussed, the model underlying Kalman filtering is a graphical model in the form of a chain (see Figure 15.1). We copy the two-node factor analysis model as an array and we link successive state nodes.

The independence relationships that characterize the SSM are identical to those that characterize the HMM. In particular, given the state at one moment in time, the states in the future are conditionally independent of those in the past. Moreover, the observation of the output nodes fails to separate any of the state nodes, and in general we expect for there to be a probabilistic relationship in the posterior distribution between all of the state nodes. As in the HMM, we hope that we can calculate these relationships recursively.

The state nodes in the factor analysis model are continuous, vector-valued nodes endowed with a Gaussian probability distribution. To develop a dynamical generalization of the factor analysis model we must represent the transition between the nodes at successive moments in time. Perhaps the simplest choice that we can make is to allow the mean of the state at time $t+1$ to be a linear function of the state at time t. Thus we write:

$$x_{t+1} = Ax_t + Gw_t, \tag{15.1}$$

where w_t is a "noise" term—a Gaussian random variable that is independent of w_s for $s < t$, and thus independent of x_t. We assume that w_t has zero mean and covariance matrix Q. Given that the sum of Gaussian variables is Gaussian, we have that x_{t+1} is indeed Gaussian. Conditional on x_t, its mean is Ax_t and its covariance is GQG^T.

In the factor analysis model, the output is endowed with a Gaussian distribution having a mean that is a linear function of the state. We continue to use this model for the output of the SSM:

$$y_t = Cx_t + v_t, \tag{15.2}$$

where v_t is a Gaussian random variable with zero mean and covariance matrix R. Conditional on x_t, y_t is a Gaussian with mean Cx_t and covariance R.

Finally we endow the initial state, x_0, with a Gaussian distribution having mean 0 and covariance Σ_0. The assumption of zero mean is without loss of generality (a non-zero mean gives rise to a deterministic component that can added to the probabilistic solution; see Exercise XXX for the details).

15.2 The unconditional distribution

Before beginning our investigation of the inference problem for the SSM, it is of interest to study the unconditional distribution of the states x_t.

The unconditional mean of x_t is clearly zero. This follows from the assumption that x_0 has zero mean, and via the dynamical equation (Eq. 15.1) each successive state has zero mean.

Turning to the unconditional covariance, which we denote Σ_t, we have:

$$\begin{aligned}
\Sigma_{t+1} &\triangleq E[x_{t+1}x_{t+1}^T] && (15.3)\\
&= E[(Ax_t + Gw_t)(Ax_t + Gw_t)^T] && (15.4)\\
&= AE[x_t x_t^T]A^T + GE[w_t w_t^T]G^T && (15.5)\\
&= A\Sigma_t A^T + GQG^T, && (15.6)
\end{aligned}$$

where we have made use of our independence assumptions. This equation, a dynamical equation for the evolution of the unconditional covariance, is referred to as the *Lyapunov equation.*

It can also be verified that the unconditional covariance between neighboring states x_t and x_{t+1} is given by $\Sigma_t A^T$.

15.3 Inference

The inference problem for the SSM is the same as it was for the HMM—that of calculating the posterior probability of the states given an output sequence. Based on our experience with the HMM, we hope to be able to calculate such posterior probabilities recursively.

In the case of the HMM, we were able to decompose the inference problem into a "forward" problem and a "backward" problem. In the forward problem the evidence consisted of a partial sequence of outputs—all those outputs up to time t. The backward problem also utilized a partial sequence—all those outputs after time t. We will find that this same decomposition will yield recursive algorithms for the SSM.

As in the case of the HMM we distinguish between "filtering" and "smoothing"—two classes of problem that arise in this graphical model when we introduce evidence. We develop algorithms for solving both problems.

15.4 Filtering

The problem is to calculate an estimate of the state x_t based on a partial output sequence $y_0, \ldots, y_t$. That is, we wish to calculate $P(x_t|y_0, \ldots, y_t)$.[1]

Sums of Gaussian variables are Gaussian, and thus, considering all of the variables in the SSM jointly, we have a (large) multivariate Gaussian distribution. Conditionals of Gaussians are Gaussian (see Chapter 13) and thus the probability distribution $P(x_t|y_0, \ldots, y_t)$ must be Gaussian. This implies that we need only calculate a mean vector and a covariance matrix (or the corresponding canonical parameters). As we will see, inference in the SSM involves finding a recursion linking these conditional means and conditional covariances at neighboring moments in time.

We use a simplified notation for the conditional means and conditional covariances that emphasizes the particular output sequence being conditioned on. We write $\hat{x}_{t|t}$ to denote the mean of x_t conditioned on the partial sequence $y_0, \ldots, y_t$. The covariance matrix of x_t conditioned on $y_0, \ldots, y_t$ is denoted $P_{t|t}$; thus:

$$\hat{x}_{t|t} \triangleq E[x_t|y_0, \ldots, y_t] \tag{15.7}$$

$$P_{t|t} \triangleq E[(x_t - \hat{x}_{t|t})(x_t - \hat{x}_{t|t})^T|y_0, \ldots, y_t]. \tag{15.8}$$

In our derivation of the algorithm, we will find that it is useful as an intermediate step to compute the probability distribution of x_t conditioned on $y_0, \ldots, y_{t-1}$. In our new notation, this distribution has mean $\hat{x}_{t|t-1}$ and covariance matrix $P_{t|t-1}$.

To uncover the recursion behind the Kalman filter, let us refer to the graphical model fragments in Figure 15.2. In the fragment on the left, where we condition on the outputs $y_0, \ldots, y_t$, we assume that we have already calculated $P(x_t|y_0, \ldots, y_t)$; that is, we have calculated $\hat{x}_{t|t}$ and $P_{t|t}$. We wish to carry this distribution forward into the fragment on the right, where we condition on $y_0, \ldots, y_{t-1}$. We decompose the transformation into two steps:

time update:	$P(x_t|y_0, \ldots, y_t) \rightarrow P(x_{t+1}|y_0, \ldots, y_t)$
measurement update:	$P(x_{t+1}|y_0, \ldots, y_t) \rightarrow P(x_{t+1}|y_0, \ldots, y_{t+1})$

Thus, in the *time update* step, we simply propagate the distribution forward one step in time, calculating the new mean and covariance based on the old mean and covariance, but based on no new measurements (i.e., no new outputs). In the *measurement update* step, we incorporate the new measurement y_{t+1} and update the probability distribution for x_{t+1}. The overall result is a transformation from $\hat{x}_{t|t}$ and $P_{t|t}$ to $\hat{x}_{t+1|t+1}$ and $P_{t+1|t+1}$.

Let us first consider the time update step. Recall the dynamic equation (Eq. 15.1):

$$x_{t+1} = Ax_t + Gw_t. \tag{15.9}$$

[1]Note that this quantity is analogous to the normalized alpha variable from the HMM—the alpha variables themselves are *joint* probabilities: $P(x_t, y_0, \ldots, y_t)$. The alphas and normalized alphas differ from each other, however, only by the normalization constant. In the Gaussian case we represent probability distributions by storing only the mean and covariance matrix (or the corresponding canonical parameters); the normalization factor is implicit. Thus there is no difference between "alphas" and "normalized alphas" in the SSM setting; all probabilities are implicitly normalized.

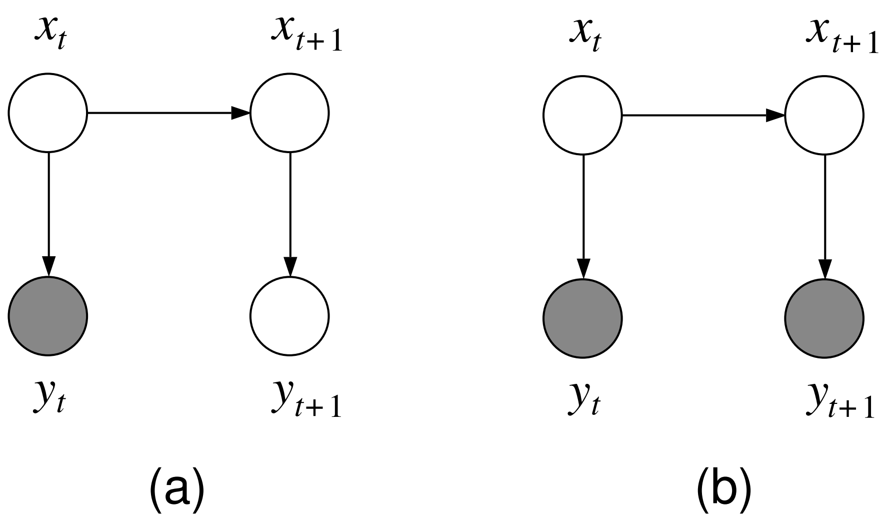

Figure 15.2: (a) A fragment of an SSM before a measurement update and (b) after a measurement update.

We take the conditional expectation on both sides of this equation. Given that w_t is independent of the conditioning variables $y_0, \ldots, y_t$, the second term vanishes, and we have:

$$\hat{x}_{t+1|t} = A\hat{x}_{t|t}. \tag{15.10}$$

Similarly, taking the conditional covariance of both sides of the dynamic equation, we have:

$$\begin{aligned} P_{t+1|t} &= E\left[(x_{t+1} - \hat{x}_{t+1|t})(x_{t+1} - \hat{x}_{t+1|t})^T | y_0, \ldots, y_t\right] && (15.11)\\ &= E\left[(Ax_t + Gw_t - A\hat{x}_{t|t})(Ax_t + Gw_t - A\hat{x}_{t|t})^T | y_0, \ldots, y_t\right] && (15.12)\\ &= AP_{t|t}A^T + GQG^T, && (15.13) \end{aligned}$$

where we have used the facts that $\hat{x}_{t+1|t}$ is a constant in the conditional distribution, w_t has zero mean, and w_t and x_t are independent.

Now that we know the conditional distribution of x_{t+1} we proceed further in the graphical model fragment and calculate the conditional mean and covariance of y_{t+1}, as well as the conditional covariance of x_{t+1} and y_{t+1}. These calculations allow us to write down the joint conditional distribution of x_{t+1} and y_{t+1}, at which point the measurement update becomes a simple matter of "reversing the arrow"—calculating the conditional distribution of x_{t+1} given y_{t+1}.

The calculations are straightforward:

$$\begin{aligned} E\left[y_{t+1} | y_0, \ldots, y_t\right] &= E\left[Cx_{t+1} + v_{t+1} | y_0, \ldots, y_t\right] && (15.14)\\ &= C\hat{x}_{t+1|t} && (15.15) \end{aligned}$$

$$E\left[(y_{t+1} - \hat{y}_{t+1|t})(y_{t+1} - \hat{y}_{t+1|t})^T | y_0, \ldots, y_t\right]$$

$$
\begin{aligned}
&= E\left[(Cx_{t+1} + v_{t+1} - C\hat{x}_{t+1|t})(Cx_{t+1} + v_{t+1} - C\hat{x}_{t+1|t})^T | y_0, \ldots, y_t\right] && (15.16)\\
&= CP_{t+1|t}C^T + R && (15.17)
\end{aligned}
$$

and

$$
\begin{aligned}
&E\left[(y_{t+1} - \hat{y}_{t+1|t})(x_{t+1} - \hat{x}_{t+1|t})^T | y_0, \ldots, y_t\right] \\
&\quad = E\left[(Cx_{t+1} + v_{t+1} - \hat{y}_{t+1|t})(x_{t+1} - \hat{x}_{t+1|t})^T | y_0, \ldots, y_t\right] && (15.18)\\
&\quad = CP_{t+1|t}, && (15.19)
\end{aligned}
$$

where we have made use of the various independence assumptions.

We summarize these results as follows. Conditioned on the past outputs $y_0, \ldots, y_t$, the variables x_{t+1} and y_{t+1} have a joint Gaussian distribution, with mean and covariance matrix:

$$
\begin{bmatrix} \hat{x}_{t+1|t} \\ C\hat{x}_{t+1|t} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} P_{t+1|t} & P_{t+1|t}C^T \\ CP_{t+1|t} & CP_{t+1|t}C^T + R \end{bmatrix} \tag{15.20}
$$

This leaves us in a situation which is familiar to us from factor analysis. Making reference to Figure 15.2(b), we have a Gaussian graphical model fragment in which we wish to reverse the arrow; that is, we wish to compute the conditional distribution of x_{t+1} given y_{t+1}, where x_{t+1} and y_{t+1} have a joint Gaussian distribution. The only difference in the current situation is that the joint distribution is itself a conditional distribution, conditioned on the past outputs $y_0, \ldots, y_t$.

Utilizing Eq. 13.26 and 13.27 from Chapter 13, we obtain:

$$
\begin{aligned}
\hat{x}_{t+1|t+1} &= \hat{x}_{t+1|t} + P_{t+1|t}C^T(CP_{t+1|t}C^T + R)^{-1}(y_{t+1} - C\hat{x}_{t+1|t}) && (15.21)\\
P_{t+1|t+1} &= P_{t+1|t} - P_{t+1|t}C^T(CP_{t+1|t}C^T + R)^{-1}CP_{t+1|t}. && (15.22)
\end{aligned}
$$

We summarize the filtering equations that we have obtained. At time t we assume that we have available the mean estimate $\hat{x}_{t|t}$ and the covariance estimate $P_{t|t}$. Based on these estimates we calculate $\hat{x}_{t+1|t+1}$ and $P_{t+1|t+1}$ recursively as follows:

$$
\begin{aligned}
\hat{x}_{t+1|t} &= A\hat{x}_{t|t} && (15.23)\\
P_{t+1|t} &= AP_{t|t}A^T + GQG^T && (15.24)\\
\hat{x}_{t+1|t+1} &= \hat{x}_{t+1|t} + P_{t+1|t}C^T(CP_{t+1|t}C^T + R)^{-1}(y_{t+1} - C\hat{x}_{t+1|t}) && (15.25)\\
P_{t+1|t+1} &= P_{t+1|t} - P_{t+1|t}C^T(CP_{t+1|t}C^T + R)^{-1}CP_{t+1|t}. && (15.26)
\end{aligned}
$$

These recursions constitute the Kalman filter. They are initialized with $\hat{x}_{0|-1} = 0$ and $P_{0|-1} = P_0$.

The update in Eq. 15.25 is often summarized in more a compact form by defining the *Kalman gain matrix*:

$$
K_{t+1} \triangleq P_{t+1|t}C^T(CP_{t+1|t}C^T + R)^{-1}. \tag{15.27}
$$

Using this notation we have:

$$
\hat{x}_{t+1|t+1} = \hat{x}_{t+1|t} + K_{t+1}(y_{t+1} - C\hat{x}_{t+1|t}). \tag{15.28}
$$

Moreover, we can use the matrix inversion formulas to write the gain matrix in an alternative form. In particular, using Eq. 13.17 and Eq. 13.18, we obtain:

$$\begin{aligned} K_{t+1} &= P_{t+1|t}C^T(CP_{t+1|t}C^T + R)^{-1} && (15.29)\\ &= (P_{t+1|t}^{-1} + C^TRC)^{-1}C^TR^{-1} && (15.30)\\ &= (P_{t+1|t} + P_{t+1|t}C^T(CP_{t+1|t}C^T + R)^{-1}CP_{t+1|t})C^TR^{-1} && (15.31)\\ &= P_{t+1|t+1}C^TR^{-1}, && (15.32) \end{aligned}$$

which expresses the gain matrix in terms of the updated matrix $P_{t+1|t+1}$.

15.5 Interpretation and relationship to LMS

The Kalman filtering equations have an appealing interpretation as an error-correcting algorithm. Let us write a single equation for the update of the mean by combining Eq. 15.23 and Eq. 15.25:

$$\hat{x}_{t+1|t+1} = A\hat{x}_{t|t} + K_{t+1}(y_{t+1} - CA\hat{x}_{t|t}). \tag{15.33}$$

Eq. 15.33 describes an error-correcting algorithm for estimating the state x_{t+1}. In particular, at time t, our best estimate of the state x_t is $\hat{x}_{t|t}$. Imitating the dynamical equation we produce an estimate $A\hat{x}_{t|t}$ of the state at time $t+1$. This estimate is then corrected based on the observation y_{t+1}; in particular, we adjust our estimate by a term $(y_{t+1} - CA\hat{x}_{t|t})$ that is proportional to the error between the observed output and our prediction of the output.

This error-correction procedure is reminiscent of the LMS algorithm. To clarify the relationship, consider a simplified situation in which the matrix A is the identity matrix and the noise term w_t is zero. In this case, the "dynamical equation" $x_{t+1} = x_t + Gw_t$ reduces to the statement that the "state" is a constant. Let θ denote this constant. Furthermore, let the matrix C in Eq. 15.2 be replaced by the (time-varying) vector x_t^T (as in Section XXX). In this case, Eq. 15.2 reduces to:

$$y_t = x_t^T\theta + v_t. \tag{15.34}$$

We are back in the world of linear regression, in which the outputs y_t are a sequence of iid observations that provide information about the parameter vector θ. In this case the Kalman filtering equation becomes:

$$\hat{\theta}_{t+1} = \hat{\theta}_t + P_{t+1}R^{-1}(y_{t+1} - x_t^T\hat{\theta})x_t, \tag{15.35}$$

where we have used the fact that R^{-1} is a scalar and have dropped the unnecessary second time subscript on the P matrix.

We have derived an equation which, when combined with the update for P_{t+1}, is referred to as the *recursive least squares (RLS) algorithm*. RLS is a special case of the Kalman filter and, as such, provides the optimal least-squares estimate of θ based on data y_t up to and including time t.

If we proceed further and approximate the matrix $P_{t+1}R^{-1}$ with a scalar multiplier μ, Eq. 15.35 reduces to the LMS algorithm (Eq. 6.6). Thus LMS can be viewed as an approximation to the Kalman filter. We have gained in simplicity—no longer needing to carry forward a covariance matrix—but we have lost in accuracy. The LMS algorithm requires multiple passes through a data set to converge to the least-squares estimate of the parameter; the Kalman filter converges in a single pass.

Although this connection between the Kalman filter and the LMS algorithm is a interesting and useful relationship to be aware of, the approximation of $P_{t+1}R^{-1}$ by a scalar multiplier receives no particular justification within the theory of Kalman filtering. Rather it requires a different theoretical framework (that of stochastic approximation) for its justification.

15.6 Information filter

Recall that the multivariate Gaussian distribution can be described using either the moment parameterization or the canonical parameterization. Our derivation of the Kalman filter used the moment parameterization of the Gaussian, but it is also of interest to define a filtering algorithm in terms of the canonical parameterization. The result is an algorithm known as an *information filter*.

We can derive the information filter either from first principles or by transforming the equations that we have already obtained. We pursue the former approach in Chapter 18, where we reconsider the SSM from the perspective of the junction tree framework. In the current section we pursue the latter approach. This is essentially an exercise in the use of the matrix inversion lemmas (Eq. 13.17 and Eq. 13.18).

Recall from Chapter 13 that the canonical parameters of a Gaussian distribution can be obtained from the moment parameters by the following transformation (cf. Eq. 13.5): $\Lambda = \Sigma^{-1}$ and $\xi = \Sigma^{-1}\mu$. Define $\hat{\xi}_{t|t-1}$ and $S_{t|t-1}$ to be the canonical parameters of the distribution of x_t conditioned on $y_{1,\ldots,t-1}$ and let $\hat{\xi}_{t|t}$ and $S_{t|t}$ to be the canonical parameters of the distribution of x_t conditioned on $y_{1,\ldots,t}$. We obtain a set of recursions for these quantities by substituting from Eqs. 15.23 to 15.26.

Let us begin with the inverse covariance matrices. Defining $H \triangleq GQG^T$ to simplify the notation, we have:

$$\begin{aligned}
S_{t+1|t} &= P_{t+1|t}^{-1} && (15.36)\\
&= (AP_{t|t}A^T + H)^{-1} && (15.37)\\
&= H^{-1} - H^{-1}A(P_{t|t}^{-1} + A^TH^{-1}A)^{-1}A^TH^{-1} && (15.38)\\
&= H^{-1} - H^{-1}A(S_{t|t} + A^TH^{-1}A)^{-1}A^TH^{-1}. && (15.39)
\end{aligned}$$

A further application of the matrix inversion lemma yields:

$$\begin{aligned}
S_{t+1|t+1} &= P_{t+1|t+1}^{-1} && (15.40)\\
&= (P_{t+1|t} - P_{t+1|t}C^T(CP_{t+1|t}C^T + R)^{-1}CP_{t+1|t})^{-1} && (15.41)\\
&= P_{t+1|t}^{-1} + C^TR^{-1}C && (15.42)\\
&= S_{t+1|t} + C^TR^{-1}C. && (15.43)
\end{aligned}$$

Turning now to the ξ parameters, we have:

$$\begin{aligned}
\hat{\xi}_{t+1|t} &= P_{t+1|t}^{-1}\hat{x}_{t+1|t} &(15.44)\\
&= P_{t+1|t}^{-1}A\hat{x}_{t|t} &(15.45)\\
&= P_{t+1|t}^{-1}AP_{t|t}\hat{\xi}_{t|t} &(15.46)\\
&= (AP_{t|t}A^T + H)^{-1}AP_{t|t}\hat{\xi}_{t|t} &(15.47)\\
&= H^{-1}A(P_{t|t}^{-1} + A^TH^{-1}A)^{-1}\hat{\xi}_{t|t} &(15.48)\\
&= H^{-1}A(S_{t|t} + A^TH^{-1}A)^{-1}\hat{\xi}_{t|t}, &(15.49)
\end{aligned}$$

and

$$\begin{aligned}
\hat{\xi}_{t+1|t+1} &= P_{t+1|t+1}^{-1}\hat{x}_{t+1|t+1} &(15.50)\\
&= P_{t+1|t+1}^{-1}(\hat{x}_{t+1|t} + P_{t+1|t+1}C^TR^{-1}(y_{t+1} - C\hat{x}_{t+1|t})) &(15.51)\\
&= (P_{t+1|t+1}^{-1} - C^TR^{-1}C)P_{t+1|t}^{-1}\hat{\xi}_{t+1|t} + C^TR^{-1}y_{t+1} &(15.52)\\
&= (P_{t+1|t}^{-1} + C^TR^{-1}C - C^TR^{-1}C)P_{t+1|t}^{-1}\hat{\xi}_{t+1|t} + C^TR^{-1}y_{t+1} &(15.53)\\
&= \hat{\xi}_{t+1|t} + C^TR^{-1}y_{t+1}. &(15.54)
\end{aligned}$$

We summarize the information filter equations. At time t we assume that we have available $\hat{\xi}_{t|t}$ and $S_{t|t}$. Based on these estimates we calculate $\hat{\xi}_{t+1|t+1}$ and $S_{t+1|t+1}$ recursively as follows:

$$\begin{aligned}
\hat{\xi}_{t+1|t} &= H^{-1}A(S_{t|t} + A^THA)^{-1}\hat{\xi}_{t|t} &(15.55)\\
\hat{\xi}_{t+1|t+1} &= \hat{\xi}_{t+1|t} + C^TR^{-1}y_{t+1} &(15.56)\\
S_{t+1|t} &= H^{-1} - H^{-1}A(S_{t|t} + A^TH^{-1}A)^{-1}A^TH^{-1} &(15.57)\\
S_{t+1|t+1} &= S_{t+1|t} + C^TR^{-1}C. &(15.58)
\end{aligned}$$

These recursions are initialized with $\hat{\xi}_{0|-1} = \bar{\xi}_0$ and $S_{0|-1} = S_0$.

The Kalman filter and the information filter are mathematically equivalent; the major practical difference between them is essentially numerical. Recall that the condition number of a matrix is the reciprocal of the condition number of its inverse; this implies that poor conditioning for one set of recursions generally implies good conditioning for the other set. A related issue concerns the initial conditions. If we are quite certain about the initial state, then we would set P_0 to zero, in which case S_0 is undefined and we would be forced to use the Kalman filter. On the other hand, if we are quite uncertain about the initial state, we would set S_0, in which case P_0 is undefined and we would be forced to use the information filter.

15.7 Smoothing

We now turn to the issue of obtaining estimates of the state at time t based on data up to and including a later time T. As in the case of the HMM, the calculation of this state estimate requires

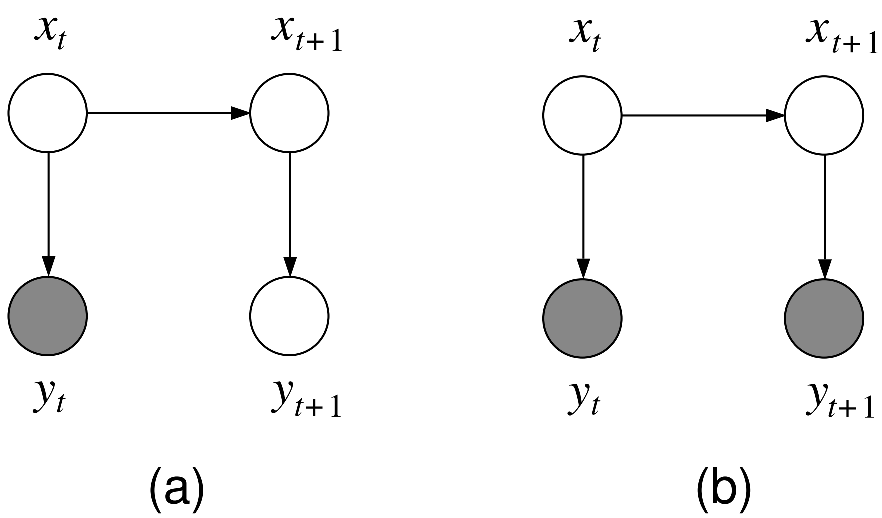

Figure 15.3: (a) A fragment of an SSM in which the observations up to and including y_t are available, and (b) the same fragment in which observations y_{t+1} to y_T are available.

us to combine a forward recursion with a backward recursion. Furthermore, we once again have the choice between an algorithm that computes backward-filtered estimates and combines them with the forward-filtered estimates (an "alpha-beta algorithm"), or an algorithm that recurses directly on the filtered-and-smoothed estimates (an "alpha-gamma algorithm"). Both kinds of algorithm are available in the literature on state-space models, but the latter approach appears to dominate (as opposed to the HMM literature, where the former approach dominates). In this section we begin with the "alpha-gamma" approach, deriving the the "Rauch-Tung-Striebel (RTS) smoothing algorithm," and then turn to an alternative "alpha-beta" approach.

15.7.1 The Rauch-Tung-Striebel (RTS) smoother

Our approach to deriving the RTS smoothing algorithm will once again be based on the graphical model fragment shown in Figure 15.2, which we reproduce in Figure 15.3. We begin by writing down the joint distribution of x_t and x_{t+1}, conditional on $y_0, \ldots, y_t$. Recall that $\hat{x}_{t+1|t} = A\hat{x}_{t|t}$, which implies:

$$E[(x_t - \hat{x}_{t|t})(x_{t+1} - \hat{x}_{t+1|t})^T | y_0, \ldots, y_t] = P_{t|t}A^T. \tag{15.59}$$

Thus our distribution has the following mean and covariance matrix:

$$\begin{bmatrix} \hat{x}_{t|t} \\ \hat{x}_{t+1|t} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} P_{t|t} & P_{t|t}A^T \\ AP_{t|t} & P_{t+1|t} \end{bmatrix}, \tag{15.60}$$

where all of the quantities are available to us after a forward Kalman filtering pass.

We now introduce a "backwards" computation. In particular, we condition on x_{t+1} and calculate the probability of x_t, still conditioning on $y_0, \ldots, y_t$. Using the by-now familiar Gaussian conditioning rule (Eq. 13.26), we obtain:

$$\begin{aligned} E[x_t|x_{t+1}, y_0, \ldots, y_t] &= \hat{x}_{t|t} + P_{t|t}A^T P_{t+1|t}^{-1}(x_{t+1} - \hat{x}_{t+1|t}) && (15.61) \\ &= \hat{x}_{t|t} + L_t(x_{t+1} - \hat{x}_{t+1|t}), && (15.62) \end{aligned}$$

where we have introduced the notation $L_t \triangleq P_{t|t}A^T P_{t+1|t}^{-1}$, and

$$\begin{aligned} \mathrm{Var}[x_t|x_{t+1}, y_0, \ldots, y_t] &= P_{t|t} - P_{t|t}A^T P_{t+1|t}^{-1} A P_{t|t} && (15.63) \\ &= P_{t|t} - L_t P_{t+1|t} L_t^T. && (15.64) \end{aligned}$$

The purpose of conditioning on x_{t+1} is to render x_t independent of the future observations $y_{t+1}, \ldots, y_T$. That is, we can use conditional independence to write:

$$\begin{aligned} E[x_t|x_{t+1}, y_0, \ldots, y_T] &= E[x_t|x_{t+1}, y_0, \ldots, y_t] && (15.65) \\ &= \hat{x}_{t|t} + L_t(x_{t+1} - \hat{x}_{t+1|t}) && (15.66) \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}[x_t|x_{t+1}, y_0, \ldots, y_T] &= \mathrm{Var}[x_t|x_{t+1}, y_0, \ldots, y_t] && (15.67) \\ &= P_{t|t} - L_t P_{t+1|t} L_t^T. && (15.68) \end{aligned}$$

The quantities on the left-hand side of these equations are almost what we want; indeed, if we could drop x_{t+1} we would have the desired filtered-and-smoothed quantities.

The remainder of the derivation is an exercise in conditional expectation. Recall from Appendix XXX the following fundamental facts about conditional expectations:

$$E[X|Z] = E[E[X|Y, Z]|Z] \tag{15.69}$$

and

$$\mathrm{Var}[X|Z] = \mathrm{Var}[E[X|Y, Z]|Z] + E[\mathrm{Var}[X|Y, Z]|Z] \tag{15.70}$$

which show us how to compute unconditional expectations using conditional expectations. We will substitute x_t for X, x_{t+1} for Y, and $y_0, \ldots, y_T$ for Z in these equations.

Beginning with Eq. 15.66, we take the conditional expectation on both sides, conditioning with respect to $y_0, \ldots, y_T$:

$$\begin{aligned} \hat{x}_{t|T} &\triangleq E[x_t|y_0, \ldots, y_T] && (15.71) \\ &= E[E[x_t|x_{t+1}, y_0, \ldots, y_T]|y_0, \ldots, y_T] && (15.72) \\ &= E[\hat{x}_{t|t} + L_t(x_{t+1} - \hat{x}_{t+1|t})|y_0, \ldots, y_T] && (15.73) \\ &= \hat{x}_{t|t} + L_t(x_{t+1|T} - \hat{x}_{t+1|t}), && (15.74) \end{aligned}$$

where we have used the fact that all of the quantities in Eq. 15.74 other than x_{t+1} are constants when we condition on $y_0, \ldots, y_T$.

Eq. 15.74 is the basic update equation in the RTS smoothing algorithm. We see that a estimate of x_t based on all of the data can be obtained by correcting the filtered estimate $\hat{x}_{t|t}$ by an error term composed of a smoothed estimate of x_{T+1} and the filtered estimate $\hat{x}_{t+1|t}$. The gain matrix L_t is a quantity that depends only on matrices computed during the forward pass.

We now work on the conditional variance equation (Eq. 15.68). Using Eq. 15.70, we have:

$$
\begin{aligned}
P_{t|T} &\triangleq \mathrm{Var}[x_t|y_0,\ldots,y_T] && (15.75)\\
&= \mathrm{Var}[E[x_t|x_{t+1},y_0,\ldots,y_T]|y_0,\ldots,y_T] + E[\mathrm{Var}[x_t|x_{t+1},y_0,\ldots,y_T]|y_0,\ldots,y_T] && (15.76)\\
&= \mathrm{Var}[\hat{x}_{t|t} + L_t(x_{t+1} - \hat{x}_{t+1|t})|y_0,\ldots,y_T] + E[P_{t|t} - L_t P_{t+1|t} L_t^T|y_0,\ldots,y_T] && (15.77)\\
&= L_t \mathrm{Var}[(x_{t+1} - \hat{x}_{t+1|t})|y_0,\ldots,y_T] L_t^T + P_{t|t} - L_t P_{t+1|t} L_t^T && (15.78)\\
&= L_t \mathrm{Var}[x_{t+1}|y_0,\ldots,y_T] L_t^T + P_{t|t} - L_t P_{t+1|t} L_t^T && (15.79)\\
&= L_t P_{t+1|T} L_t^T + P_{t|t} - L_t P_{t+1|t} L_t^T && (15.80)\\
&= P_{t|t} + L_t(P_{t+1|T} - P_{t+1|t}) L_t^T, && (15.81)
\end{aligned}
$$

where at several junctures we have used the fact that expectations taken with respect to $y_0,\ldots,y_t$ are constant when conditioning with respect to the larger conditioning set $y_0,\ldots,y_T$.

We summarize the RTS smoothing algorithm. Based on the quantities $\hat{x}_{t+1|t}$, $P_{t|t}$ and $P_{t+1|t}^{-1}$ from the filtering algorithm, we compute:

$$
\begin{aligned}
\hat{x}_{t|T} &= \hat{x}_{t|t} + L_t(x_{t+1|T} - \hat{x}_{t+1|t}), && (15.82)\\
P_{t|T} &= P_{t|t} + L_t(P_{t+1|T} - P_{t+1|t}) J_t^T, && (15.83)
\end{aligned}
$$

where $L_t \triangleq P_{t|t} A^T P_{t+1|t}^{-1}$. The algorithm is initialized by using $\hat{x}_{T|T}$ and $P_{T|T}$ from the filtering pass.

15.7.2 The two-filter smoother

In this section we describe an alternative approach to smoothing in the SSM which is the analog of the alpha-beta algorithm for the HMM. In this approach, known as the "two-filter algorithm," the idea is to combine the "forward" conditional probability $P(x_t|y_0,\ldots,y_t)$ with the "backward" conditional probability $P(x_t|y_{t+1},\ldots,y_T)$. Note that the latter quantity, like the former quantity, is a "filtered estimate"; that is, a conditional probability of the state given a (partial) output sequence. This differs from the traditional beta variable in the HMM, which is the conditional probability of the output sequence given the state. Clearly we can move from one to the other, however, by multiplying or dividing by the unconditional probability of the state, $P(x_t)$, which is available via the Lyapunov equation. Thus, the difference is minor and it is appropriate to think of the two-filter algorithm as the analog of the alpha-beta algorithm.

Given that we want filtered estimates in the backward direction, a simple approach to deriving the backward algorithm is to "invert the dynamics" and apply a forward filtering algorithm to the inverted dynamics. In graphical model terms, we invert the arrows in the graph. This is in itself a useful exercise.

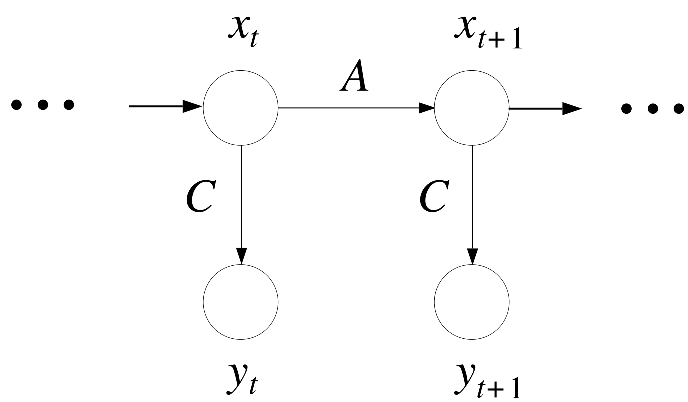

Figure 15.4: (a) A fragment of an SSM with no observations.

In this section we assume that the matrix A is invertible and make use of A^{-1} in our derivation of the algorithm. In fact this assumption is not necessary, and a lookahead at the algorithm that we derive shows that A^{-1} does not appear. (In Chapter 18 we present an alternative derivation of the algorithm from the point of view of the junction tree algorithm, and in that derivation we do not make use of A^{-1}).

The naive approach to inverting the dynamical equation is to simply write:

$$x_t = A^{-1}x_{t+1} - A^{-1}Gw_t, \tag{15.84}$$

and let t run backwards in time. This approach is not viable, however, because w_t is not independent of the "past" values of the state; i.e., $x_{t+1}, \ldots, x_T$. Indeed, these states are all a function of w_t. Thus one of the assumptions that we used in deriving the Kalman filter is not valid and we cannot simply apply the Kalman filter to Eq. 15.84.

To obtain a more useful inverse of the dynamics, consider the graphical model fragment shown in Figure 15.4. The forward dynamics yields a joint probability distribution on (x_t, x_{t+1}) characterized by the Lyapunov equation $\Sigma_{t+1} = A\Sigma_t A^T + GQG^T$. Indeed the covariance matrix of (x_t, x_{t+1}) is given by:

$$\begin{bmatrix} \Sigma_t & \Sigma_t A^T \\ A\Sigma_t & A\Sigma_t A^T + GQG^T \end{bmatrix} \tag{15.85}$$

We can invert the relationship between x_t and x_{t+1} by solving for Σ_t in terms of Σ_{t+1} and rewriting the covariance matrix in terms of Σ_{t+1}. Thus:

$$\Sigma_t = A^{-1}\Sigma_{t+1}A^{-T} - A^{-1}GQG^T A^{-T}, \tag{15.86}$$

where we assume that A is invertible.[2] This equation also implies:

$$A\Sigma_t = \Sigma_{t+1}A^{-T} - GQG^T A^{-T}, \tag{15.87}$$

[2] In fact this assumption is not necessary, see Exercise XXX.

and we can rewrite the covariance matrix as follows:

$$\left[\begin{array}{cc} A^{-1}\Sigma_{t+1}A^{-T} - A^{-1}GQG^TA^{-T} & A^{-1}\Sigma_{t+1} - A^{-1}GQG^T \\ \Sigma_{t+1}A^{-T} - GQG^TA^{-T} & \Sigma_{t+1} \end{array}\right] \quad (15.88)$$

Noting that the upper-right-hand corner of this matrix can be written as $A^{-1}(I - A^{-1}GQG^T\Sigma_{t+1}^{-1})\Sigma_{t+1}$, we see that if we define:

$$\tilde{A} = A^{-1}(I - A^{-1}GQG^T\Sigma_{t+1}^{-1}) \quad (15.89)$$

then we obtain $\tilde{A}\Sigma_{t+1}$ and $\Sigma_{t+1}\tilde{A}^T$ in the corners of the matrix, and the matrix begins to take the form of a forward covariance matrix. This suggests that we define the inverse dynamics via:

$$x_t = \tilde{A}x_{t+1} + \tilde{G}\tilde{w}_{t+1}, \quad (15.90)$$

with $\tilde{G}$ and $\tilde{w}_t$ chosen appropriately so as to match the forward dynamics (Eq. 15.1). Indeed, choosing

$$\begin{aligned} \tilde{G} &= -A^{-1}G && (15.91) \\ \tilde{w}_{t+1} &= w_t - QG^T\Sigma_{t+1}^{-1}x_{t+1} && (15.92) \end{aligned}$$

Eq. 15.90 matches Eq. 15.1. Moreover, we have:

$$\tilde{Q} \triangleq E[\tilde{w}_{t+1}\tilde{w}_{t+1}^T] = Q - QG^T\Sigma_{t+1}^{-1}GQ, \quad (15.93)$$

and substituting Eqs. 15.89, 15.92 and 15.93 in the backward Lyapunov equation:

$$\Sigma_t = \tilde{A}\Sigma_{t+1}\tilde{A}^T + \tilde{G}\tilde{Q}\tilde{G}^T \quad (15.94)$$

we recover the forward Lyapunov equation.

Finally, it can also be verified (see Exercise XXX) that $\tilde{w}_{t+1}$ is independent of the "past" values of the state $x_{t+1}, \ldots, x_T$.

We have therefore succeeded in obtaining a version of the inverse dynamics to which standard filtering algorithms can be applied. If we use the canonical parameterization (i.e., the information filter in Eqs. 15.39, 15.43, 15.49, and 15.54), utilizing the inverse dynamical equation and noting that the output equation $y_t = Cx_t + v_t$ has not changed, we obtain:

$$\begin{aligned} S_{t|t+1} &= A^THA + \Sigma_t^{-1} - A^TH^{-1}(S_{t+1|t+1} + H^{-1} - \Sigma_{t+1}^{-1})^{-1}H^{-1}A && (15.95) \\ S_{t|t} &= S_{t|t+1} + C^TR^{-1}C && (15.96) \\ \hat{\xi}_{t|t+1} &= A^TH^{-1}(S_{t+1|t+1} + H^{-1} - \Sigma_{t+1}^{-1})^{-1}\hat{\xi}_{t+1|t+1} && (15.97) \\ \hat{\xi}_{t|t} &= \hat{\xi}_{t|t+1} + C^TR^{-1}y_t, && (15.98) \end{aligned}$$

where t and $t+1$ have been interchanged to reflect the fact that we are filtering backward in time. This filter calculates the canonical representation of $P(x_t|y_{t+1}, \ldots, y_T)$. Thus, converting to the moment representation, we have $\hat{x}_{t|t+1} = S_{t|t+1}^{-1}\hat{\xi}_{t|t+1}$ and $P_{t|t+1} = S_{t|t+1}^{-1}$.

The final issue that we must address involves the fusing of the probability distributions $P(x_t|y_0,\ldots,y_t)$ and $P(x_t|y_{t+1},\ldots,y_T)$ to obtain the posterior probability $P(x_t|y_0,\ldots,y_T)$. This problem is not unique to the filtering and smoothing domain, but arises in many other settings as well. It is therefore worth posing and solving the problem in full generality; this we do in the following section. Anticipating the result, we have the following fusion rule for $\hat{x}_{t|T}$, the estimate of x_t based on all of the data:

$$\hat{x}_{t|T} = P_{t|T}\ (P_{t|t}^{-1}\hat{x}_{t|t} + P_{t|t+1}^{-1}\hat{x}_{t|t+1}), \tag{15.99}$$

where the covariance matrix $P_{t|T}$ is computed as follows:

$$P_{t|T} = \left(P_{t|t}^{-1} + P_{t|t+1}^{-1} - \Sigma_t^{-1}\right)^{-1}. \tag{15.100}$$

The appearance of Σ_t^{-1} in the latter equation should not be a surprise. The filtering process and the smoothing process both make use of the prior statistics on x_t; in the latter case this is because we have inverted the dynamics. When the covariance matrices of these two processes are combined we have included the prior covariance twice. To avoid double-counting Σ_t^{-1} must be subtracted in the combination rule.

15.7.3 Fusion of Gaussian posterior probabilities

Let us consider three sets of random variables: x, z_1 and z_2. Suppose that these variables are characterized by a multivariate Gaussian distribution and suppose moreover that z_1 and z_2 are conditionally independent given x. We wish to fuse the posteriors $P(x|z_1)$ and $P(x|z_2)$ into an overall posterior $P(x|z_1, z_2)$.

Let us assume, without loss of generality, that x, z_1, and z_2 have zero means. Non-zero means can be subtracted away and added back at the end of the analysis.

Under the conditional independence assumption, there are three ways to represent the distribution of x, z_1, and z_2 as a directed graphical model. The representation given in Figure 15.5(a) is particularly useful for our purposes. To parameterize the graph, we require the marginal $P(x)$, and conditionals $P(z_1|x)$ and $P(z_2|x)$. For the marginal, we endow x with a zero mean and covariance Σ. For the conditionals, recall that Gaussian conditionals are linear functions of the conditioning variable (cf. Eq. refeq:Gaussian-conditional-mean). Thus we can write:

$$\begin{aligned} z_1 &= M_1 x + v_1 \qquad (15.101)\\ z_2 &= M_2 x + v_2, \qquad (15.102)\end{aligned}$$

for appropriately chosen matrices M_1 and M_2 and zero-mean Gaussian variables v_1 and v_2 having covariance matrices R_1 and R_2. Note moreover that v_1 and v_2 are independent of x and are conditionally independent of each other given x.

Let us now consider a generic linear equation $z = Mx + v$, where v is independent of x and has covariance R. To calculate the conditional expectation of x given z we first obtain the covariance matrix of the pair (x, z):

$$\begin{bmatrix} \Sigma & \Sigma M^T \\ M\Sigma & M\Sigma M^T + R \end{bmatrix}. \tag{15.103}$$

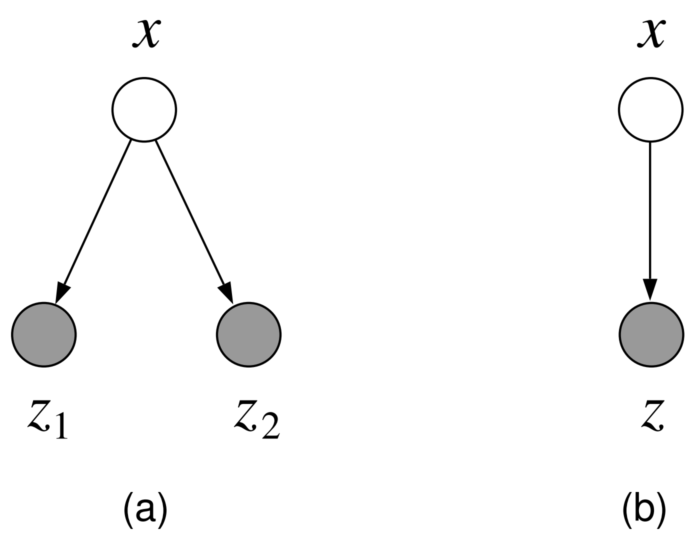

Figure 15.5: A graphical model representation of the fusion problem. (a) The observables z_1 and z_2 are assumed conditionally independent given x. The conditional probabilities of z_i are parameterized as linear functions of x with additive, independent noise terms. (b) Conjoining z_1 and z_2 into a single observable vector z.

We then apply the usual Gaussian conditioning formulas (Eqs. 13.26 and 13.27) to obtain the conditional distribution of x given z. Denoting the mean of this conditional distribution as $\hat{x}$ and the covariance as P, we have:

$$\begin{aligned} \hat{x} &= \Sigma M^T \left(M \Sigma M^T + R \right)^{-1} z && (15.104) \\ &= \left(M^T R^{-1} M + \Sigma^{-1} \right)^{-1} M^T R^{-1} z, && (15.105) \end{aligned}$$

where we have used a matrix inversion identity (Eq. 13.18) in the second step, and:

$$\begin{aligned} P &= \Sigma - \Sigma M^T \left(M \Sigma M^T + R \right)^{-1} M \Sigma && (15.106) \\ &= \left(\Sigma^{-1} + M^T R^{-1} M \right)^{-1}, && (15.107) \end{aligned}$$

where again we use a matrix inversion identity (Eq. 13.17) to simplify the result.

The individual conditionals of x given z_1 and z_2 are special cases of the foregoing equations. Defining $\hat{x}_i \triangleq E(x|z_i)$ and letting P_i denote the corresponding conditional covariance, we have:

$$\begin{aligned} \hat{x}_1 &= \left(M_1^T R_1^{-1} M_1 + \Sigma^{-1} \right)^{-1} M_1^T R_1^{-1} z_1 && (15.108) \\ \hat{x}_2 &= \left(M_2^T R_2^{-1} M_2 + \Sigma^{-1} \right)^{-1} M_2^T R_2^{-1} z_2, && (15.109) \end{aligned}$$

and

$$\begin{aligned} P_1 &= \left(M_1^T R_1^{-1} M_1 + \Sigma^{-1} \right)^{-1} && (15.110) \\ P_2 &= \left(M_2^T R_2^{-1} M_2 + \Sigma^{-1} \right)^{-1}. && (15.111) \end{aligned}$$

Now let us consider the overall posterior of x given both z_1 and z_2. Grouping z_1 and z_2 into a single variable z (cf. Figure 15.5(b)), we can apply Eqs. 15.105 and 15.107 where:

$$M \triangleq \begin{bmatrix} M_1 \\ M_2 \end{bmatrix} \quad \text{and} \quad R \triangleq \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix}. \tag{15.112}$$

From these definitions we obtain:

$$\begin{aligned}
\hat{x} &= \left(\begin{bmatrix} M_1^T & M_2^T \end{bmatrix} \begin{bmatrix} R_1^{-1} & 0 \\ 0 & R_2^{-1} \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \end{bmatrix} + \Sigma^{-1} \right)^{-1} \begin{bmatrix} M_1^T & M_2^T \end{bmatrix} \begin{bmatrix} R_1^{-1} & 0 \\ 0 & R_2^{-1} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \\
&= \left(M_1^T R_1^{-1} M_1 + M_2^T R_2^{-1} M_2 + \Sigma^{-1} \right)^{-1} \left(M_1^T R_1^{-1} z_1 + M_2^T R_2^{-1} z_2 \right) \qquad (15.113) \\
&= \left(P_1^{-1} + P_2^{-1} - \Sigma^{-1} \right)^{-1} \left(P_1^{-1} \hat{x}_1 + P_2^{-1} \hat{x}_2 \right) \qquad (15.114)
\end{aligned}$$

We can similarly expand Eq. 15.107 to obtain the overall conditional covariance P:

$$P = \left(P_1^{-1} + P_2^{-1} - \Sigma^{-1} \right)^{-1}, \tag{15.115}$$

thus allowing us to rewrite Eq. 15.114 as:

$$\hat{x} = P(P_1^{-1} \hat{x}_1 + P_2^{-1} \hat{x}_2). \tag{15.116}$$

Eqs. 15.116 and 15.115 are our general solution to the Gaussian fusion problem.

Let us relate these results back to the two-filter smoothing problem. We collect the observations up to and including time t into a single "past" vector $z_1 \triangleq (y_0, \ldots, y_t)$, and collect the "future" observations into a single vector $z_2 \triangleq (y_{t+1}, \ldots, y_T)$. Let $x \triangleq x_t$. These definitions fit the problem specification of the current section; in particular (x, z_1, z_2) are characterized by a multivariate Gaussian distribution (a marginal of the larger Gaussian distribution that includes the other state variables), and moreover z_1 and z_2 are independent given x. The estimate $\hat{x}_{t|t}$ is the conditional expectation of x given z_1, and must therefore have the form in Eq. 15.109, for matrices M_1 and R_1 that we do not bother to calculate. Similarly $\hat{x}_{t|t+1}$ must be of the form in Eq. 15.109, and the conditional covariances $P_{t|t}$ and $P_{t|t+1}$ must have the form of Eqs. 15.111 and 15.111. Substituting into Eqs. 15.116 and 15.115 we obtain the fusion rules at the end of the previous section (Eqs. 15.99 and 15.100).

15.8 Parameter estimation

We follow the by now familiar recipe for developing an EM algorithm for parameter estimation for the SSM. We write out the expected complete log likelihood, identify the expected sufficient statistics, solve for maximum likelihood estimates in terms of these expected sufficient statistics. This latter problem is simply linear regression.

[Section not yet finished].

15.9 Historical remarks and bibliography