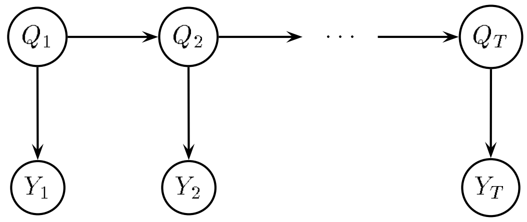
Filtering, Smoothing and the Junction Tree Algorithm

Kevin P. Murphy

19 November 1998
Revised 30 September 1999*

1 Introduction

It is well known that the forwards-backwards algorithm for Hidden Markov Models (HMMs) is equivalent to the junction tree (jtree) algorithm applied to the graphical model shown below [SHJ97]. It is also known that the Kalman filter-smoother algorithm for discrete time Linear Dynamical Systems (LDSs) must be equivalent to the jtree algorithm, although we believe this paper is the first to show the correspondence explicitely. We use a representation of a Gaussian distribution called canonical characteristics [Lau92], instead of the more familiar moment characteristics, to derive a simple two-filter ($\alpha/\ \beta$) approach to smoothing in LDSs, as is commonly done in HMMs, without having to assume that the dynamics are invertible or that time is continuous c.f.,[LK76]. To complete the symmetry, we also present a method for smoothing in HMMs which is analogous to the standard ($\alpha/\ \gamma$) method of smoothing in LDSs.



2 The forwards-backwards algorithm for HMMs

We start by reviewing the forwards-backwards algorithm for HMMs [Rab89], which will serve as a basis for future generalizations. The parameterization of the model is as follows, where Q_t is the discrete hidden state variable at time t. (For simplicity we assume that the parameters are time-invariant.)

$$\begin{aligned} \Pr(Q_1 = i) &= \pi(i) \\ \Pr(y_t|Q_t = i) &= B_t(i) \\ \Pr(Q_t = j|Q_{t-1} = i) &= A(i,j) \end{aligned}$$

Given a fixed observation sequence $y_{1:T} = y_1, \ldots, y_T$, the goal is to compute the posterior probability of the hidden variables $\gamma_t(i) = \Pr(Q_t = i, y_{1:T})$. This can then be normalized to yield the *smoothed* estimate,

*Many of these revisions were inspired by enlightening discussions with Michael Jordan.

$\alpha_t(i) = \Pr(Q_t = i, y_{1:t})$	$\hat{\alpha}_t(i) = \Pr(Q_t = i \vert y_{1:t})$
$\beta_t(i) = \Pr(y_{t+1:T} \vert Q_t = i)$	
$\beta'_{t\vert t+1}(i) = \Pr(Q_t = i, y_{t+1:T})$	$\hat{\beta}'_{t\vert t+1}(i) = \Pr(Q_t = i \vert y_{t+1:T})$
$\beta'_{t\vert t}(i) = \Pr(Q_t = i, y_{t:T})$	$\hat{\beta}'_{t\vert t}(i) = \Pr(Q_t = i \vert y_{t:T})$
$\gamma_t(i) = \Pr(Q_t = i, y_{1:T})$	$\hat{\gamma}_t(i) = \Pr(Q_t = i \vert y_{1:T})$

Table 1: Definition of the variables used for the one-slice distributions in the HMM. On the left are the joint forms, and on the right, the normalized forms.

$\hat{\gamma}_t(i) = \Pr(Q_t = i|y_{1:T})$, where the normalization constant, $\Pr(y_{1:T})$, is the likelihood of the data. A summary of the variables we are using is shown in Table 1.

The recursions for computing the α and β are well-known, and will just be stated here without proof.

$$\begin{aligned}
\alpha_1(i) &= \pi(i)B_1(i), \qquad \alpha_t(j) = \left[\sum_i \alpha_{t-1}(i)A(i,j)\right] B_t(j) \\
\beta_T(i) &= 1, \qquad \beta_t(i) = \sum_j A(i,j)B_{t+1}(j)\beta_{t+1}(j)
\end{aligned}$$

Note that the backwards algorithm (which computes β) can be run independently of the forwards algorithm (which computes α), but requires the evidence stream in the form of B.

After computing α and β, we can combine them to compute γ:

$$\gamma_t(i) \stackrel{\text{def}}{=} \Pr(Q_t = i, y_{1:T}) = \Pr(y_{t+1:T}|Q_t = i)\Pr(Q_t = i, y_{1:t}) = \alpha_t(i)\beta_t(i)$$

This is the standard approach to smoothing in HMMs; we call it the α / β algorithm.

Note that we can compute *filtered* estimates $\hat{\alpha}_t(i) = \Pr(Q_t = i|y_{1:t})$ on the forwards pass just by normalizing:

$$\hat{\alpha}_t(j) = (1/c_t)\left[\sum_i \hat{\alpha}_{t-1}(i)A(i,j)\right] B_t(j)$$

where the normalization factor is $c_t = \Pr(y_t|y_{1:t-1})$ (and hence, by the chain rule, $\Pr(y_{1:T}) = \prod_t c_t$). In practice, one must always use the normalized version to avoid underflow. We would like to develop an analogous backwards filter recursion for $\hat{\beta}'_{t|\tau}(i) = \Pr(Q_t = i|y_{\tau:T})$, where $\tau = t$ includes y_t in the estimate and $\tau = t+1$ excludes it. (When we combine the forwards and backwards filtered estimates, we must be sure not to "count" y_t twice, which is why it is useful to consider the $\tau = t+1$ variant.) We can do this using the standard recurrence for β providing we have access to the unconditional marginals $\Pr(Q_t = i) = \pi_t(i)$, since

$$\beta'_{t|t+1}(i) = \Pr(Q_t, y_{t+1:T}) = \Pr(y_{t+1:T}|Q_t = i)\Pr(Q_t = i) = \beta_t(i)\pi_t(i)$$

Hence we can compute γ_t from the forwards and backwards filters as follows:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta'_{t|t+1}(i)}{\pi_t(i)}$$

We call this version the α / β', or two-filter, approach to smoothing.

Finally, we consider a variation in which we first compute the α's and then compute the γ_t's directly, without first having to compute β_t or β'_t. This is analogous to the standard RTS smoothing equations for LDSs. We call this approach the α / γ algorithm. In this modified backwards algorithm, we do not need access to the evidence. We derive the recurrence using the fact that $Q_t \perp y_{t+1:T} | Q_{t+1}$.

$$\begin{aligned}
\gamma_t(i) &= \sum_j \Pr(Q_t = i | Q_{t+1} = j, y_{1:t}, y_{t+1:T}) \Pr(Q_{t+1} = j, y_{1:T}) \\
&= \sum_j \frac{\Pr(Q_t = i, Q_{t+1} = j, y_{1:t})}{\Pr(Q_{t+1} = j, y_{1:t})} \Pr(Q_{t+1} = j, y_{1:T}) \\
&= \sum_j \frac{\alpha_t(i) A(i,j)}{\sum_i \alpha_t(i) A(i,j)} \gamma_{t+1}(j)
\end{aligned}$$

2.1 Two-slice version

For parameter estimation, we need the joint probability distributions on two consecutive time-slices,

$$\xi_{t,t+1}(i,j) \stackrel{\text{def}}{=} \Pr(Q_t = i, Q_{t+1} = j, y_{1:T})$$

The standard method is to compute it as follows

$$\xi_{t,t+1}(i,j) = \alpha_t(i) B_{t+1}(j) \beta_t(j) A(i,j)$$

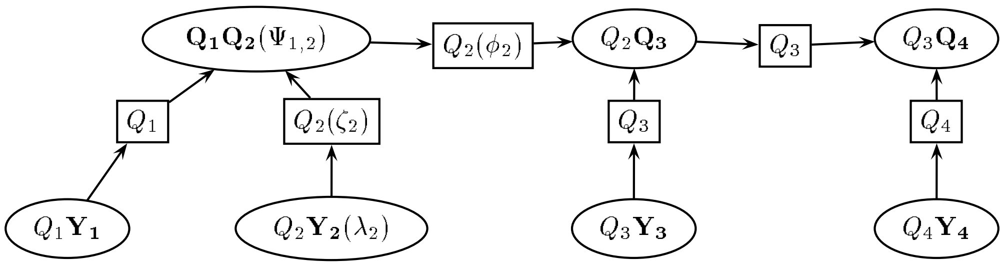
However, by analogy with the α / γ, α / β and α / β' recursions, we can derive a recursion directly in terms of ρ / ξ, ρ/ σ and ρ/ σ' respectively, where

$$\begin{aligned}
\rho_{t-1,t}(i,j) &= \Pr(Q_{t-1} = i, Q_t = j, y_{1:t}) \\
\sigma_{t-1,t}(i,j) &= \Pr(Q_t = j, y_{t+1:T} | Q_{t-1} = i) \\
\sigma'_{t-1,t}(i,j) &= \Pr(Q_{t-1} = i, Q_t = j, y_{t:T})
\end{aligned}$$

It will turn out that the jtree always works with the two-slice version, and computes the one-slice version by marginalization.

3 The junction tree algorithm for HMMs

Consider the junction tree shown below, where the arrows indicate the direction in which "messages" will flow when we perform the collect evidence operation with $Q_{T-1}Q_T$ as the root, greek letters denote the potentials, and bold letters indicate which nodes are assigned to each clique. (We say a node Q is a assigned to a clique C if Q's conditional probability distribution (CPD) gets multiplied onto to C's potential. Note that, although a node can appear in several cliques, it only contributes its CPD to one of them, to avoid double-counting.)



3.1 Unconditional inference

First we initialize all clique and separator potentials to 1s. Then for each node in the original graph, we multiply its conditional probability distribution (CPD) onto the clique to which it is assigned. If we assume initially that the Y nodes are hidden, the λ potentials at the "leaves" will be all 1s (since $\lambda_t(i) = \Pr(\emptyset|Q_t = i) = 1$, where $\emptyset$ denotes the empty set of evidence), and the potentials along the "backbone" will be as follows:

$$\begin{aligned}
\Psi_{1,2}(i,j) &= \Pr(Q_2 = j|Q_1 = i)\Pr(Q_1 = i) = A(i,j)\pi(i) = \Pr(Q_1 = i, Q_2 = j) \\
\Psi_{t,t+1}(i,j) &= \Pr(Q_{t+1} = j|Q_t = i) = A(i,j)
\end{aligned}$$

Now we propagate messages from children to their parents in the rooted jtree, where propagating a message from clique V to W via separator S means performing the following steps:

- $P_S^* = \sum_{V \setminus S} P_V$.
- $P_W^* = P_W \frac{P_S^*}{P_S}$.

(The star superscript denotes that the potential has been updated once.) If we do this in the "evidence-free" jtree, it is simple to show by induction that, after the forwards pass (i.e., after collecting to root), $\Psi_{t,t+1}^*(i,j) = \Pr(Q_t = i, Q_{t+1} = j)$, $\phi_t^*(i) = \Pr(Q_t = i)$, and all the other potentials are unchanged (i.e., all 1s). (We will see how to derive such equations in more detail in the next section.) After the backwards pass (i.e., after distributing from the root), we have

$$\Psi_{t-1,t}^{**}(i,j) = \Pr(Q_{t-1} = i, Q_t = j)$$

and

$$\zeta_t^{**}(i) = \lambda_t^{**}(i) = \phi_t^{**}(i) = \Pr(Q_t = i)$$

Thus all potentials have been converted to marginals. (Note that the terms along the backbone did not change during the backwards pass.) Of course, a more direct way to compute the unconditional marginals is by "powering up" the transition matrix. Specifically, define the vector $\pi_t(i) = \Pr(Q_t = i)$. Then $\pi_t = A^{t-1}\pi$.

3.2 Forwards pass

Now suppose we initialise the leaf potentials with evidence

$$\lambda_t^*(i) \stackrel{\text{def}}{=} \Pr(y_t|Q_t = i) = B_t(i)$$

and run the forwards algorithm. It is clear that $\zeta_t^*(i) = \lambda_t^*(i)$, since we are just marginalizing out the constant term y_t. Along the backbone, we get the following potentials:

$$\begin{aligned}
\Psi_{t-1,t}^*(i,j) &= \Pr(Q_{t-1}=i, Q_t=j, y_{1:t}) = \rho_{t-1,t}(i,j) \\
\phi_t^*(i) &= \Pr(Q_t=i, y_{1:t}) = \alpha_t(i)
\end{aligned}$$

We prove this by induction. First the base case.

$$\begin{aligned}
\Psi_{1,2}^*(i,j) &= \Psi_{1,2}(i,j)\frac{\zeta_1^*(i)}{1}\frac{\zeta_2^*(j)}{1} \\
&= \Pr(Q_2=j|Q_1=i)\Pr(Q_1=i)\Pr(y_1|Q_1=i)\Pr(y_2|Q_2=j) \\
&= \Pr(Q_1=i, Q_2=j, y_{1:2})
\end{aligned}$$

Marginalizing yields

$$\Phi_2^*(j) = \sum_i \Psi_{1,2}^*(i,j) = \Pr(Q_2=j, y_{1:2})$$

Now for the induction step.

$$\begin{aligned}
\Psi_{t-1,t}^*(i,j) &= \Psi_{t-1,t}(i,j)\frac{\phi_{t-1}^*(i)}{1}\frac{\zeta_t^*(j)}{1} \\
&= \Pr(Q_t=j|Q_{t-1}=i)\Pr(Q_{t-1}=i, y_{1:t-1})\Pr(y_t|Q_t=j) \\
&= A(i,j)\alpha_{t-1}(i)B_t(j) \\
&= \Pr(Q_{t-1}=i, Q_t=j, y_{1:t})
\end{aligned}$$

This is the recursion for ρ. Marginalizing yields the recursion for α:

$$\phi_t^*(j) = \sum_i \Psi_{t-1,t}^*(i,j) = \sum_i A(i,j)\alpha_{t-1}(i)B_t(j) = \Pr(Q_t=j, y_{1:t})$$

3.3 Backwards pass

There are several variants on the backwards pass.

- Distribute evidence from the $Q_{T-1}Q_T$ root in the jtree that was just updated by the forwards pass. This yields the γ recursion.
- Collect evidence to a different root Q_1Q_2 This yields the β recursion.
- Collect to the Q_1Q_2 root as in the β algorithm, but first initialize all the potentials to marginals (as in Section 3.1). This yields the β' recursion.

3.4 Backwards pass: γ version

In the backwards pass, the direction that the messages flow is exactly the reverse of what is shown in the previous diagram. We find that the potentials are

$$\begin{aligned}
\zeta_t^{**}(i) = \phi_t^{**}(i) = \lambda_t^{**}(i) &= \Pr(Q_t = i, y_{1:T}) = \gamma_t(i) \\
\Psi_{t-1,t}^{**}(i,j) &= \Pr(Q_{t-1} = i, Q_t = j, y_{1:T}) = \xi_{t-1,t}(i,j)
\end{aligned}$$

We can prove these formulas by induction as follows. The base case is

$$\Psi_{T-1,T}^{**}(i,j) = \Psi_{T-1,T}^{*}(i,j) = \Pr(Q_{T-1} = i, Q_T = j, y_{1:T})$$

The inductive step uses the fact that $Q_{t-1} \perp y_{t+1:T} | Q_t$ as follows

$$\begin{aligned}
\Psi_{t-1,t}^{**}(i,j) &= \frac{\Psi_{t-1,t}^{*}(i,j)}{\phi_t^{*}(j)} \phi_t^{**}(j) \\
&= \frac{\Pr(Q_{t-1} = i, Q_t = j, y_{1:t})}{\Pr(Q_t = j, y_{1:t})} \Pr(Q_t = j, y_{1:T}) \\
&= \Pr(Q_{t-1} = i | Q_t = j, y_{1:t}) \Pr(Q_t = j, y_{1:t}, y_{t+1:T}) \\
&= \Pr(Q_{t-1} = i, Q_t = j, y_{1:T})
\end{aligned}$$

This corresponds to the recursion for ξ_t. Marginalizing yields the recursion for γ_t:

$$\phi_t^{**}(j) = \sum_i \Psi_{t-1,t}^{**}(i,j) = \Pr(Q_t = j, y_{1:T})$$

3.5 Backwards pass: β version

Now we consider collecting to the root $Q_1 Q_2$. The arrows along the backbone will be reversed, but the arrows up from the leaves remain unchanged.

We initialize the cliques just as in the forwards pass, and we find that the potentials are

$$\begin{aligned}
\lambda_t^{*}(i) = \zeta_t^{*}(i) &= \Pr(y_t | Q_t = i) \\
\phi_t^{*}(i) &= \Pr(y_{t+1:T} | Q_t = i) = \beta_t(i) \\
\Psi_{t-1,t}^{*}(i,j) &= \Pr(y_{t:T}, Q_t = j | Q_{t-1} = i) = \sigma_{t-1,t}(i,j)
\end{aligned}$$

We can prove these formulas by induction as follows. We use the following conditional independence properties: $y_{t+1} \perp Q_t | Q_{t+1}$ and $y_{t+2:T} \perp y_{t+1}, Q_t | Q_{t+1}$.

The base case is

$$\begin{aligned}
\Psi_{T-1,T}^{*}(i,j) &= \Psi_{T-1,T}(i,j) \frac{\zeta_T^{*}(j)}{1} \\
&= \Pr(Q_T = j | Q_{T-1} = i) \Pr(y_T | Q_T = j) \\
&= \Pr(Q_t = j, y_T | Q_{T-1} = i)
\end{aligned}$$

Marginalizing yields

$$\phi^*_{T-1}(i) = \sum_j \Psi^*_{T-1,T}(i,j) = \Pr(y_T|Q_{T-1}=i)$$

The induction step is

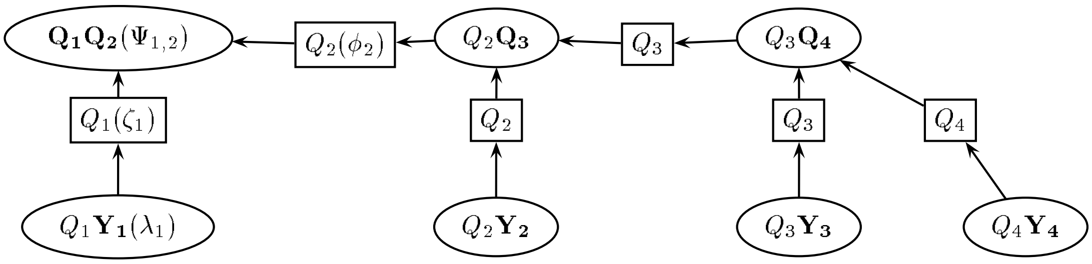
$$\begin{aligned}
\Psi^*_{t,t+1}(i,j) &= \Psi_{t,t+1}(i,j)\frac{\zeta^*_{t+1}(i)}{1}\frac{\phi^*_{t+1}(j)}{1} \\
&= \Pr(Q_{t+1}=j|Q_t=i)\Pr(y_{t+1}|Q_{t+1}=j)\Pr(y_{t+2:T}|Q_{t+1}=j) \\
&= A(i,j)B_{t+1}(j)\beta_{t+1}(j) \\
&= \Pr(Q_{t+1}=j,y_{t+1}|Q_t=i)\Pr(y_{t+2:T}|Q_{t+1}=j,y_{t+1},Q_t=i) \\
&= \Pr(Q_{t+1}=j,y_{t+1:T}|Q_t=i)
\end{aligned}$$

This is the same as the recursion for σ. Marginalizing yields the recursion for β:

$$\phi^*_t(i) = \sum_j A(i,j)B_{t+1}(j)\beta_{t+1}(j) = \Pr(y_t|Q_t=i)$$

3.6 Backwards pass: β' version

To derive the β' version, we consider the following jtree. We start by doing a forwards and backwards pass to convert all the potentials to marginals, as in Section 3.1. Then we collect to the root Q_1Q_2.



The potentials will turn out to be

$$\begin{aligned}
\Psi^*_{t,t+1}(i,j) &= \Pr(Q_t=i,Q_{t+1}=j,y_{t:T}) = \sigma'_{t,t+1}(i,j) \\
\phi^*_t(i) &= \Pr(Q_t=i,y_{t:T}) = \beta'_{t|t}(i)
\end{aligned}$$

Note that both of these terms include y_t, but we would like to exclude it, to compute $\beta'_{t|t+1}(i) = \Pr(Q_t = i, y_{t+1:T})$. Hence we consider what happens when $\Psi^*_{t,t+1}$ absorbs from ϕ^*_{t+1} but before it absorbs from ζ^*_t.

$$\begin{aligned}
\Psi_{t,t+1}(i,j)\frac{\phi^*_{t+1}(j)}{\phi_{t+1}(j)} &= \Pr(Q_t=i,Q_{t+1}=j)\Pr(Q_{t+1}=j,y_{t+1:T})/\Pr(Q_{t+1}=j) \\
&= \Pr(Q_t=i,Q_{t+1}=j)\Pr(y_{t+1:T}|Q_{t+1}=j) \\
&= \Pr(Q_t=i,Q_{t+1}=j,y_{t+1:T})
\end{aligned}$$

Marginalizing out Q_{t+1} yields

$$\beta'_{t|t+1}(i) = \sum_j \Pr(Q_t = i, Q_{t+1} = j, y_{t+1:T}) = \Pr(Q_t = i, y_{t+1:T})$$

Now we absorb from ζ_t^*:

$$\begin{aligned}
\Psi^*_{t,t+1}(i,j) &= \Psi_{t,t+1}(i,j)\frac{\phi^*_{t+1}(j)}{\phi_{t+1}(j)}\frac{\zeta^*_t(i)}{\zeta_t(i)} \\
&= \Pr(Q_t = i, Q_{t+1} = j, y_{t+1:T})\Pr(y_t|Q_t = i) \\
&= \Pr(Q_t = i, Q_{t+1} = j, y_{t:T})
\end{aligned}$$

(Note that, if we used the previous jtree, we would get

$$\Psi^*_{t,t+1}(i,j) = \Psi_{t,t+1}(i,j)\frac{\phi^*_{t+1}(j)}{\phi_{t+1}(j)}\frac{\zeta^*_{t+1}(j)}{\zeta_{t+1}(j)}$$

and hence we would be dividing by $\Pr(Q_{t+1} = j)$ twice.) Marginalizing yields

$$\phi^*_t(i) = \sum_j \Psi^*_{t,t+1}(i,j) = \Pr(Q_t = i, y_{t:T})$$

Although this seems rather complicated, it will turn out to give us a much easier way to compute the backwards filter for LDSs than by using standard methods.

4 The junction tree algorithm for LDSs

A Linear Dynamical System corresponds to the same graphical model as an HMM, but the parameterization of the nodes is different. In particular, we define

$$\begin{aligned}
P(X_1 = x) &= N(x; x_{1|0}, V_{1|0}) \\
P(Y_t = y|X_t = x) &= N(y; Cx, R) \\
P(X_t = x_t|X_{t-1} = x_{t-1}) &= N(x_t; Ax_{t-1}, Q)
\end{aligned}$$

where $x_{1|0}$ and $V_{1|0}$ are the initial mean and variance, and X_t is the continuous hidden state variable at time t. We will also need a different representation for the clique potentials. Following [LW89, Lau92], we use the following representation in terms of *canonical characteristics*:

$$\phi(x) = \exp[g + x'h - \tfrac{1}{2}x'Kx]$$

If K is full rank, we can convert this to the more familiar form

$$\phi(x) = p \times \exp[-\tfrac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)]$$

where the *moment characteristics* are as follows:

$$\Sigma = K^{-1}, \qquad \mu = \Sigma h, \qquad p = \exp[g + \tfrac{1}{2}\mu'\Sigma^{-1}\mu] \tag{1}$$

If the potential contains no evidence, p will be equal to the normalizing constant

$$p = (2\pi)^{-|x|/2}|\Sigma|^{-\frac{1}{2}}$$

so that the potential integrates to 1.

4.1 Operations on canonical potentials

We now show how to create canonical potentials from linear Gaussian CPDs, how to add evidence to them, and how to multiply, divide and marginalize them, which is necessary for implementing the fundamental operation of message passing. (Note that division can be thought of as conditioning, $\phi(V)/\phi(W) = P(V|W)$, and multiplication as a way of computing joints, $\phi(V|W)\phi(W) = P(V,W)$.) Once we have defined these operators, the inference algorithm is identical to before.

Note that there is no single best representation for Gaussian potentials: multiplication/division is easy in canonical form and hard in moment form, and marginalization is hard in canonical form and easy in moment form. Here we work exclusively in canonical form, and then convert to the more intuitive moment form at the end.

4.1.1 Creating a canonical potential from a CPD

Consider a node Y with parent X whose CPD is $N(y; \mu + BX, \Sigma)$.

$$\begin{aligned}
P(y|x) &= p\exp\left[-\tfrac{1}{2}\left((y-\mu-Bx)'\Sigma^{-1}(y-\mu-Bx)\right)\right] \\
&= \exp\left[-\tfrac{1}{2}\begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} B'\Sigma^{-1}B & -B'\Sigma^{-1} \\ -\Sigma^{-1}B & \Sigma^{-1} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} -B'\Sigma^{-1}\mu \\ \Sigma^{-1}\mu \end{pmatrix} - \tfrac{1}{2}\mu'\Sigma^{-1}\mu + \log p\right]
\end{aligned}$$

where $p = (2\pi)^{-n/2}|\Sigma|^{-\frac{1}{2}}$. Hence we set the canonical characteristics to

$$g = -\tfrac{1}{2}\mu'\Sigma^{-1}\mu - \frac{n}{2}\log(2\pi) - \tfrac{1}{2}\log|\Sigma|, \quad h = \begin{pmatrix} -B'\Sigma^{-1}\mu \\ \Sigma^{-1}\mu \end{pmatrix}, \quad K = \begin{pmatrix} B'\Sigma^{-1}B & -B'\Sigma^{-1} \\ -\Sigma^{-1}B & \Sigma^{-1} \end{pmatrix} \tag{2}$$

This generalizes the result in [Lau92] to the vector case. In the scalar case, $\Sigma^{-1} = 1/\sigma^2$, $\vec{\mu} = \mu$, $B = b'$ and $n = 1$, so the above becomes

$$g = \frac{-\mu^2}{2\sigma^2} - \tfrac{1}{2}\log(2\pi\sigma), \qquad h = \frac{\mu}{\sigma^2}\begin{pmatrix} -b \\ 1 \end{pmatrix}, \qquad K = \frac{1}{\sigma^2}\begin{pmatrix} bb' & -b \\ -b' & 1 \end{pmatrix}$$

4.1.2 Entering evidence

If we observe that a continuous variable Y takes on a specific value y, we must modify the potentials of all the cliques/separators that contain Y, since their dimensionality will be reduced. Consider a clique that contains X and Y. The new potential is

$$\begin{aligned}\phi^*(x) &= \exp[g + \begin{pmatrix} x^T & y^T \end{pmatrix} \begin{pmatrix} h_X \\ h_Y \end{pmatrix} - \tfrac{1}{2} \begin{pmatrix} x^T & y^T \end{pmatrix} \begin{pmatrix} K_{XX} & K_{XY} \\ K_{YX} & K_{YY} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\ &= \exp[(g + h_Y^T y - \tfrac{1}{2} y^T K_{YY} y) + x^T (h_X - K_{XY} y) - \tfrac{1}{2} x^T K_{XX} x]\end{aligned}$$

i.e.,

$$g^* = g + h_Y^T y - \tfrac{1}{2} y^T K_{YY} y, \qquad h^* = h_X - K_{XY} y, \qquad K^* = K_{XX} \tag{3}$$

4.1.3 Multiplication and division

To multiply $\phi_1(x_1, \ldots, x_k; g_1, h_1, K_1)$ by $\phi_2(x_{k+1}, \ldots, x_n; g_2, h_2, K_2)$, we extend them both to the same domain $x_1, \ldots, x_n$ by adding zeros to the appropriate dimensions, and compute

$$(g_1, h_1, K_1) * (g_2, h_2, K_2) = (g_1 + g_2, h_1 + h_2, K_1 + K_2) \tag{4}$$

Division is similar: we just replace the $+$ operator with $-$.

4.1.4 Marginalization

Let ϕ_W be a potential over a set W of variables. We can compute the potential over a subset $V \subset W$ of variables by marginalizing, denoted $\phi_V = \sum_{W \setminus V} \phi_W$. Let

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \qquad h = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \qquad K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}$$

with y_1 having dimension p and y_2 having dimension q. It can be shown (by completing the square and using nice properties of multidimensional Gaussians) that

$$\int \phi(y_1 y_2) dy_1 = \phi[y_2; \hat{g}, \hat{h}, \hat{K}]$$

where

$$\hat{g} = g + \tfrac{1}{2}\left(p \log(2\pi) - \log|K_{11}| + h_1^T K_{11}^{-1} h_1\right), \quad \hat{h} = h_2 - K_{21} K_{11}^{-1} h_1, \quad \hat{K} = K_{22} - K_{21} K_{11}^{-1} K_{12} \tag{5}$$

4.2 Unconditional inference

We start by considering the case when all the Y_t nodes are unobserved. We will just keep track of the quadratic form defined by (h, K), and ignore the g terms, which are just constant factors. (We order the entries in each vector/matrix such that parents preceed children (i.e., X_t before Y_t and X_{t-1} before X_t.) We will state the results for this section without proof. In the next section, we derive the forwards equations in the case where we have evidence.

Without loss of generality, we assume the initial node has zero mean (since we can always add a constant to all means later), and define the initial variance to be Σ_1. Since $X_t = AX_{t-1} + w_t$, where $w_t \sim N(0, Q)$, we have $EX_t = 0$ for all t, and $\Sigma_{t+1} = A\Sigma_t A' + Q$. Hence $\phi_t^* = \left[0, \Sigma_t^{-1}\right]$ in canonical form. The moment representation of the two-slice potential $\Psi_{t,t+1}^* = P(X_t, X_{t+1})$ is

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} \Sigma_t & -\Sigma_t A \\ -A'\Sigma_t & A\Sigma_t A' + Q \end{pmatrix}$$

In canonical form, this is

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} A'Q^{-1}A + \Sigma_t^{-1} & -A'Q^{-1} \\ -Q^{-1}A & Q^{-1} \end{pmatrix}$$

4.3 Forwards pass

First we initialize all clique and separator potentials to 0s (which is the identity element for this domain), and then for each node in the original graph, we multiply its CPD onto the clique to which it is assigned. For the $Y_t X_t$ cliques, we have (using Equation 2 with $\mu = 0$)

$$\lambda_t = P(Y_t|X_t) = \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} C'R^{-1}C & -C'R^{-1} \\ -R^{-1}C & R^{-1} \end{pmatrix}\right]$$

When we add evidence, we get (using Equation 3)

$$\lambda_t^* = P(y_t|X_t) = [(\, C'R^{-1}y_t\,), (\, C'R^{-1}C\,)]$$

For the transition nodes, we get (using Equation 2)

$$\Psi_{t,t+1} = P(X_{t+1}|X_t) = \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} A'Q^{-1}A & -A'Q^{-1} \\ -Q^{-1}A & Q^{-1} \end{pmatrix}\right]$$

We define $\phi_t^* = P(X_t, y_{1:t}) = [h_t, K_t]$. The update equation becomes

$$\begin{aligned} \Psi_{t,t+1}^* &= \Psi_{t,t+1} \frac{\phi_t^*}{1} \frac{\zeta_{t+1}^*}{1} \\ &= \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} h_t \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ C'R^{-1}y_{t+1} \end{pmatrix}, \begin{pmatrix} A'Q^{-1}A & -A'Q^{-1} \\ -Q^{-1}A & Q^{-1} \end{pmatrix} + \begin{pmatrix} K_t & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & C'R^{-1}C \end{pmatrix}\right] \end{aligned}$$

Marginalizing this yields

$$\Phi_{t+1}^* = [h_{t+1}, K_{t+1}] = \left[C'R^{-1}y_{t+1} + Q^{-1}AM_t^{-1}h_t, \qquad Q^{-1} + C'R^{-1}C - Q^{-1}AM_t^{-1}A'Q^{-1}\right]$$

where

$$M_t = K_t + A'Q^{-1}A$$

This is just the Kalman filter in information form.

4.4 Backwards pass: β' version

Just as in HMMs, we can derive a γ, a β, or a β' version of the backwards pass, which correspond to the RTS smoothing equations, a slight variant of the backwards filter, and the standard backwards filter equations respectively. We just consider the last case; the others are analogous.

The potentials are initialized to marginals, as in Section 4.2. We define $\phi_t^* = P(X_t, y_{t:T}) = [h_t, K_t]$. As in Section 3.6, we consider the case where $\Psi_{t,t+1}^*$ absorbs from ϕ_{t+1}^* but not ζ_t^*:

$$\begin{aligned}
\Psi_{t,t+1}\frac{\phi_{t+1}^*}{\phi_{t+1}} &= \Bigg[\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ h_{t+1} \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \\
&\quad \begin{pmatrix} A'Q^{-1}A + \Sigma_t^{-1} & -A'Q^{-1} \\ -Q^{-1}A & Q^{-1} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & K_{t+1} \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{t+1}^{-1} \end{pmatrix}\Bigg]
\end{aligned}$$

Marginalizing this yields

$$\beta'_{t|t+1} = P(X_t, y_{t+1:T}) = [h_{t|t+1}, K_{t|t+1}] = [A'Q^{-1}Z_{t+1}^{-1}h_{t+1}, A'Q^{-1}Z_{t+1}^{-1}Q^{-1}A]$$

where

$$Z_{t+1} = K_{t+1} + Q^{-1} - \Sigma_{t+1}^{-1}$$

This is the information form of the backwards filter. Now we include y_t:

$$\begin{aligned}
\Psi_{t,t+1}^* &= \left(\Psi_{t,t+1}\frac{\phi_{t+1}^*}{\phi_{t+1}}\right)\frac{\zeta_t^*}{\zeta_t} \\
&= \Bigg[\begin{pmatrix} 0 \\ h_{t+1} \end{pmatrix} + \begin{pmatrix} C'R^{-1}y_t \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \\
&\quad \begin{pmatrix} A'Q^{-1}A + \Sigma_t^{-1} & -A'Q^{-1} \\ -Q^{-1}A & Q^{-1} + K_{t+1} - \Sigma_{t+1}^{-1} \end{pmatrix} + \begin{pmatrix} C'R^{-1}C & 0 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} \Sigma_t^{-1} & 0 \\ 0 & 0 \end{pmatrix}\Bigg]
\end{aligned}$$

Marginalizing yields

$$\Phi_t^* = [h_{t|t}, K_{t|t}] = \left[h_{t|t+1} + C'R^{-1}y_t, K_{t|t+1} + C'R^{-1}C\right]$$

References

[Lau92] S. L. Lauritzen. Propagation of probabilities, means and variances in mixed graphical association models. *J. of the Am. Stat. Assoc.*, 87(420):1098–1108, December 1992.

[LK76] L. Ljung and T. Kailath. A unified approach to smoothing formulas. *Automatica*, 12:147–157, 1976.

[LW89] S. L. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, 17:31–57, 1989.

[Rab89] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286, 1989.

[SHJ97] P. Smyth, D. Heckerman, and M. I. Jordan. Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9(2):227–269, 1997.