Switching Kalman Filters

Kevin P. Murphy

21 August 1998

Abstract

We show how many different variants of Switching Kalman Filter models can be represented in a unified way, leading to a single, general-purpose inference algorithm. We then show how to find approximate Maximum Likelihood Estimates of the parameters using the EM algorithm, extending previous results on learning using EM in the non-switching case [DRO93, GH96a] and in the switching, but fully observed, case [Ham90].

1 Introduction

Dynamical systems are often assumed to be linear and subject to Gaussian noise. This model, called the Linear Dynamical System (LDS) model, can be defined as

$$\begin{aligned}\mathbf{x}_t &= A_t\mathbf{x}_{t-1} + \mathbf{v}_t \\ \mathbf{y}_t &= C_t\mathbf{x}_t + \mathbf{w}_t\end{aligned}$$

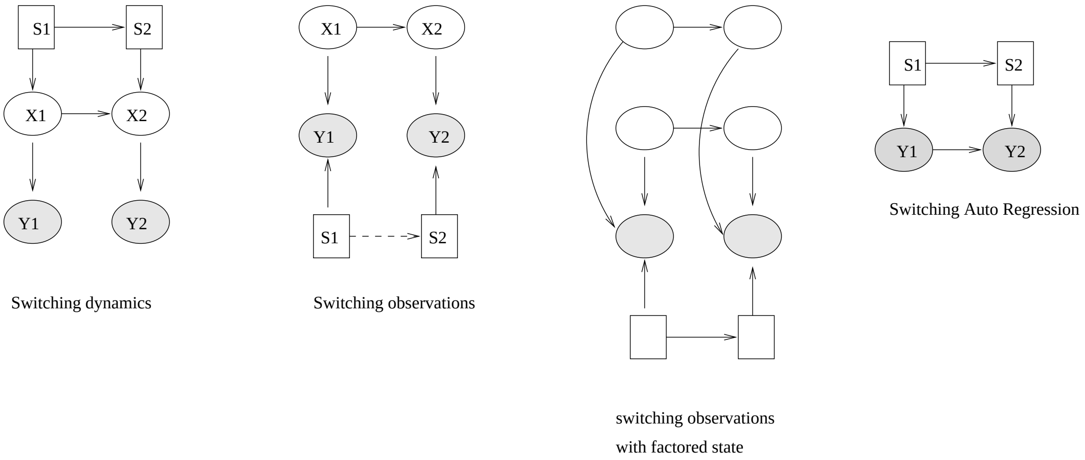
where $\mathbf{x}_t$ is the hidden state variable at time t, $\mathbf{y}_t$ is the observation at time t, and $\mathbf{v}_t \sim N(0, Q_t)$ and $\mathbf{w}_t \sim N(0, R_t)$ are independent Gaussian noise sources. Typically the parameters of the model $\Theta = \{(A_t, C_t, Q_t, R_t)\}$ are assumed to be time-invariant, so that they can be estimated from data using e.g., EM [GH96a]. One of the main advantages of this model is that there is an efficient algorithm for performing inference (i.e., computing the belief state $P(\mathbf{X}_t|\mathbf{y}_{1:t})$), the well-known Kalman filter, and its generalization to the offline case, the Rauch-Tung-Strieber smoother (for computing $P(\mathbf{X}_t|\mathbf{y}_{1:T})$, where $\mathbf{y}_{1:T}$ is all the observed data).

Unfortunately, most systems are not linear and are subject to non-Gaussian noise. One approach to this problem is to discretize the (hidden) state variables, resulting in Dynamic Bayesian Networks [DW91, Gha97], of which the Hidden Markov Model (HMM) [Rab89] is the simplest example. However, the resulting system will in general have a belief state that is exponential in the number of hidden state variables, resulting in intractable inference. In addition, it may also have a large number of parameters, resulting in inefficient learning (i.e., a lot of data is needed).

Another approach, and the one we take in this paper, is to have a bank of M different linear models, and to switch between them or take some linear combination of them.

Let us first consider the case where the dynamics are piecewise linear. We have a discrete switch variable S_t which specifies which A/Q matrix to use at time t. We assume S_t has Markovian dynamics (with transition matrix Z and initial distribution π [1]). This model is shown in Figure 1(a).

[1] Thus $1/Z(i,i)$ is the expected time we spend in state/mode i before switching. We can change the distribution on the length of each linear segment by explicitely modeling how long we spend in each state [Rab89, KRHE96].



Switching dynamics

Switching observations

switching observations
with factored state

Switching Auto Regression

Figure 1: Some switching Kalman filter models represented as Bayesian networks [Pea88]. Square nodes are discrete, oval ones are Gaussian. Shaded nodes are observed, clear nodes are hidden.

If S_t were observed, we would know when to apply each submodel (i.e., the segmentation would be known), but since S_t is hidden, we use a weighted combination of each sub-model, where the weights are given by $\Pr(S_t = i|\mathbf{y}_{1:t})$. This is called "soft switching". Hence the resulting system can be thought of as a mixture of Kalman filters. For example, we might be interested in tracking a maneuvering airplane. If the two basic models cover horizontal and vertical motion, then we can represent turns using a convex combination. SKF's have been shown [BSL93] to give superior performance to online adapative methods (such as Input Estimation) for problems such as these.

Let us now consider the case where S_t specifies which observation matrices C/R to use at time t. This can be used to model non-Gaussian observation noise, by approximating it as a mixture of Gaussians. For example, we might take Q_1 to be the covariance of the observation process, and Q_2 to be a very broad covariance (e.g., approximately uniform). The prior probability of S_t reflects how often we expect outliers to occur. This is a widely used technique for making linear regression more robust, see e.g., [PG88], and for modelling sensor failure [Wil76].

Recently, Ghahrhamani et al. [GH96b] have proposed the model shown in Figure 1(c). This also has switching observations, but the interpretation is different. The switch variable in this case can be thought of as "selecting" one of the sub-processes to pass through to the output variable or as choosing a permutation matrix C_t to apply, to model the fact that we are uncertain about which process causes which observation [SS91]; this is called data association ambiguity [BSF88].[2]

Of course, we can make both the dynamics and the observation model dependent on S_t (or on two separate Markov chains). This is the most general case that we will assume for the rest of this paper. We will also be concerned with the special case in which $C = I$, so we get to observe X directly (see Figure 1(d)). This is called a Switching Auto Regression model. SAR models are computationally much simpler than SKFs (no approximations are necessary to do inference, as we will see), since the only hidden node is discrete.

[2] Of course, in practice, we need to deal with the fact that the number of objects we are tracking, and the number of measurements we receive at each time step, is not constant.

2 Inference

The fundamental problem with SKFs is that the belief state grows exponentially with time. To see this, suppose that the initial distribution $p(\mathbf{X}_1)$ is a mixture of M Gaussians, one for each value of S_1. Then each of these must be propogated through M different equations (one for each value of S_2), so that $p(\mathbf{X}_2)$ will be a mixture of M^2 Gaussians. In general, at time t, the belief state $p(\mathbf{X}_t|\mathbf{y}_{1:t})$ will be a mixture of M^t Gaussians, one for each possible model history $S_1, \ldots, S_t$. There are several general approaches to dealing with this exponential growth [SM80]:

- Collapsing: approximate the mixture of M^t Gaussians with a mixture of r Gaussians. This is called the Generalized Pseudo Bayesian algorithm of order r (GPB(r)) (see e.g., [BSL93, Kim94]). When $r = 1$, we approximate a mixture of Gaussians with a single Gaussian using moment matching; this can be shown (e.g., [Lau96]) to be the best (in the KL sense) single Gaussian approximation. When $r = 2$, we "collapse" Gaussians which differ in their history two steps ago; in general, these will be more similar than Gaussians that differ in their more recent history. The Interacting Multiple Models (IMM) algorithm [BSL93] is a good approximation to GPB2 at the cost of only M (instead of M^2) filters (see Figure 2), although it cannot be used for smoothing. Not surprisingly, the more history we keep, the better the approximation [SM80]. See Figure 2 for a comparison of GPB1, GPB2, and the IMM filtering algorithms.
- Selection: only keep the high-probability paths in the tree of model histories. This technique is widely used for filtering when there is data association ambiguity, when it is called Multiple Hypothesis Tracking [BSF88].
- Iterative: we can sample the missing values using MCMC methods and collect averaged statistics [CK96, BMR98]. More simply, we can alternate between picking good segmentations (i.e., the most likely sequence of S_t's, c.f. the Viterbi algorithm for HMMs) and doing inference using a fixed segmentation. Alternatively, we could use weighted combinations of the matrices instead of the "best" matrix at each step [SM80].
- Variational: essentially we break all the vertical links in the model, but introduce new variational parameters to couple them together in as tight a way as possible. Using EM with such a model will maximize a lower bound on the likelihood: see [GH96b] for details.

In this paper, we focus on the collapsing approximation. One worry is that the errors introduced at each time step by approximating the posterior might accumulate over time, leading to very poor performance. However, as shown in [BK98b, BK98a], the stochasticity of the process ensures that the true distribution "spreads out" and (with high probability) "overlaps" the approximate distribution; hence they are able to prove that the error remains bounded.

Before delving into the SKF case, we "warm up" by considering the simpler case of the SAR model, for which we can perform exact inference.

2.1 Switching AR model

We define inference as computing the posterior probabilities $\Pr(S_t = j|\mathbf{x}_{1:T})$, where $\mathbf{x}_{1:T} = \mathbf{y}_{1:T}$ is the sequence of observations. We do this in two passes. In the forwards pass we proceed as follows.

$$\Pr(S_t = j|\mathbf{x}_t, \mathbf{x}_{1:t-1}) \quad = \quad \frac{1}{c}\Pr(\mathbf{x}_t|S_t = j, \mathbf{x}_{1:t-1})\Pr(S_t = j|\mathbf{x}_{1:t-1})$$

$$
\begin{aligned}
&= \frac{1}{c}\Pr(\mathbf{x}_t|S_t=j,\mathbf{x}_{t-1})\sum_i \Pr(S_t=j|S_{t-1}=i,\mathbf{x}_{1:t-1})\Pr(S_{t-1}=i|\mathbf{x}_{1:t-1}) \\
&= \frac{1}{c}L_t(j)\sum_i Z(i,j)\Pr(S_{t-1}=i|\mathbf{x}_{1:t-1})
\end{aligned}
$$

where c is the normalization constant

$$
c=\Pr(\mathbf{x}_t|\mathbf{x}_{1:t-1})=\sum_j L_t(j)\sum_i Z(i,j)M_{t-1|t-1}(i)
$$

and

$$
L_t(j)=N(\mathbf{x}_t;A_j\mathbf{x}_{t-1},Q_j)
$$

is the likelihood of the innovation (prediction error) at time t given model j. On the backwards pass, we have

$$
\begin{aligned}
\Pr(S_t=j|\mathbf{x}_{1:T}) &= \sum_k \Pr(S_t=j|S_{t+1}=k,\mathbf{x}_{1:T})\Pr(S_{t+1}=k|\mathbf{x}_{1:T}) \\
&= \sum_k \Pr(S_t=j|S_{t+1}=k,\mathbf{x}_{1:t})\Pr(S_{t+1}=k|\mathbf{x}_{1:T}) \qquad * \\
&= \sum_k \frac{\Pr(S_t=j|\mathbf{x}_{1:t})\Pr(S_{t+1}=k|S_t=j)}{\Pr(S_{t+1}=k|\mathbf{x}_{1:t})}\Pr(S_{t+1}=k|\mathbf{x}_{1:T})
\end{aligned}
$$

where the line marked * follows since the effect of future evidence on S_t is blocked by observing all its children (S_{t+1} and $\mathbf{X}_t$). On a practical note, we remark that, since we are computing the conditional probability $\Pr(S_t|\mathbf{x}_{1:t})$, as opposed to the joint probability $\Pr(S_t,\mathbf{x}_{1:t})$ as in an HMM, we do not need to worry about underflow and hence do not need scaling [Rab89].

2.2 Switching Kalman filter model

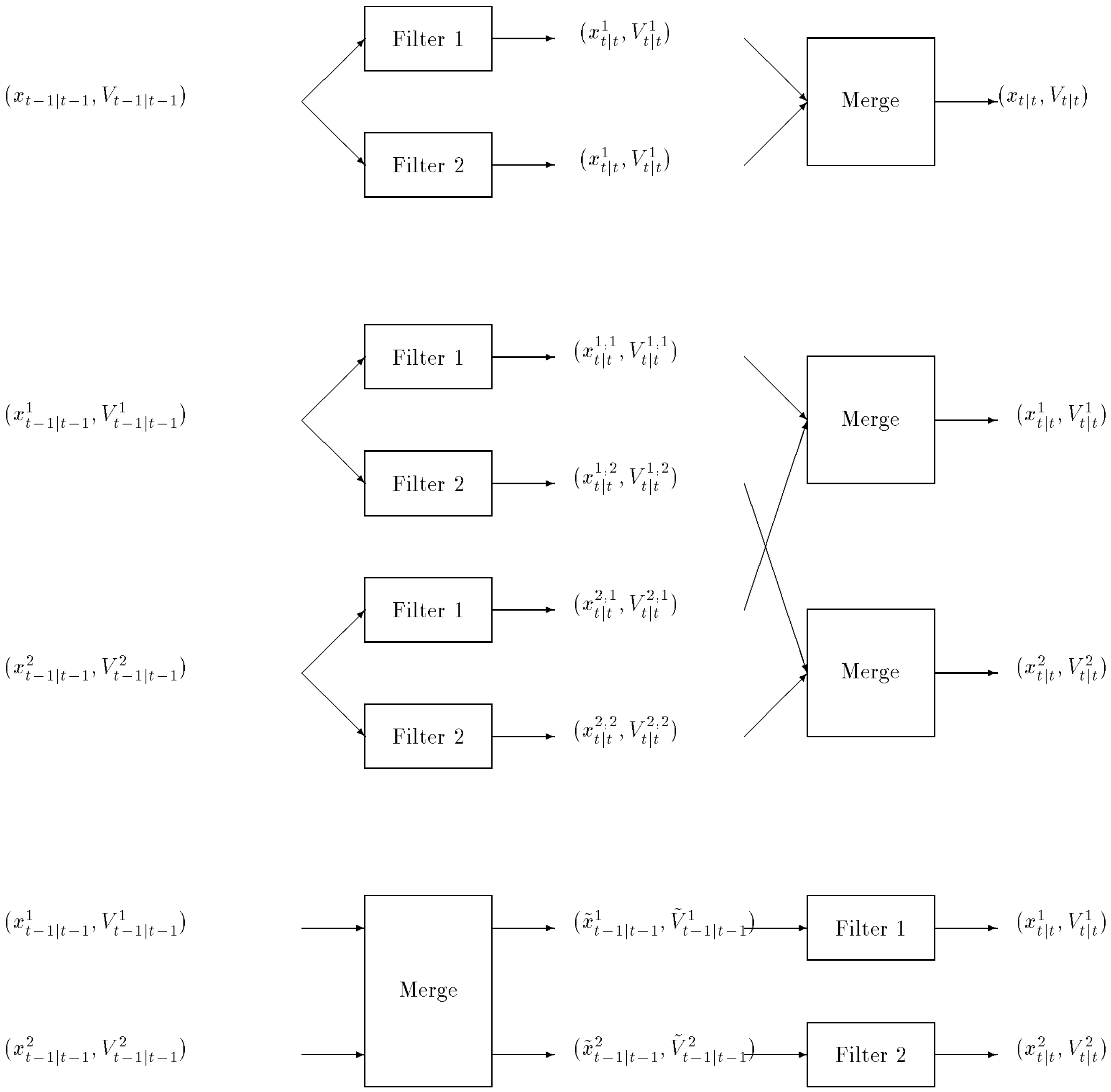
In what follows, we review the GPB2 algorithm; GPB1 and IMM cannot be used since we need to reason about two consecutive time steps to calculate the cross-variance term needed for EM.

Let us start by defining some notation.

$$
\begin{aligned}
\mathbf{x}^{i(j)}_{t|\tau} &= \mathrm{E}\left[\mathbf{X}_t|\mathbf{y}_{1:\tau},S_{t-1}=i,S_t=j\right] \\
\mathbf{x}^{(j)k}_{t|\tau} &= \mathrm{E}\left[\mathbf{X}_t|\mathbf{y}_{1:\tau},S_t=j,S_{t+1}=k\right] \\
\mathbf{x}^{j}_{t|\tau} &= \mathrm{E}\left[\mathbf{X}_t|\mathbf{y}_{1:\tau},S_t=j\right]
\end{aligned}
$$

If $\tau=t$, these are called filtered statistics; if $\tau>t$, they are called smoothed statistics; and if $\tau<t$, they are called predicted statistics. Notice how the superscript inside the brackets is the value of the switch node at time t (the subscript value); the superscript to the left is the value of S_{t-1}, and to the right, S_{t+1}. We need these subtle distinctions to handle the cross-variance terms correctly. We also define the following.

$$
\begin{aligned}
V^j_{t|\tau} &= \mathrm{Cov}\left[\mathbf{X}_t|\mathbf{y}_{1:\tau},S_t=j\right] \\
V^j_{t,t-1|\tau} &= \mathrm{Cov}\left[\mathbf{X}_t,\mathbf{X}_{t-1}|\mathbf{y}_{1:\tau},S_t=j\right] \\
V^{i(j)}_{t,t-1|\tau} &= \mathrm{Cov}\left[\mathbf{X}_t,\mathbf{X}_{t-1}|\mathbf{y}_{1:\tau},S_{t-1}=i,S_t=j\right] \\
M_{t-1,t|\tau}(i,j) &= \Pr(S_{t-1}=i,S_t=j|\mathbf{y}_{1:\tau}) \\
M_{t|\tau}(j) &= \Pr(S_t=j|\mathbf{y}_{1:\tau}) \\
L^j_t &= \Pr(\mathbf{y}_t|\mathbf{y}_{1:t-1},S_t=j)
\end{aligned}
$$



$(x_{t-1|t-1}, V_{t-1|t-1})$
Filter 1
Filter 2
$(x^1_{t|t}, V^1_{t|t})$
$(x^1_{t|t}, V^1_{t|t})$
Merge
$(x_{t|t}, V_{t|t})$
$(x^1_{t-1|t-1}, V^1_{t-1|t-1})$
$(x^2_{t-1|t-1}, V^2_{t-1|t-1})$
Filter 1
Filter 2
Filter 1
Filter 2
$(x^{1,1}_{t|t}, V^{1,1}_{t|t})$
$(x^{1,2}_{t|t}, V^{1,2}_{t|t})$
$(x^{2,1}_{t|t}, V^{2,1}_{t|t})$
$(x^{2,2}_{t|t}, V^{2,2}_{t|t})$
Merge
Merge
$(x^1_{t|t}, V^1_{t|t})$
$(x^2_{t|t}, V^2_{t|t})$
$(x^1_{t-1|t-1}, V^1_{t-1|t-1})$
$(x^2_{t-1|t-1}, V^2_{t-1|t-1})$
Merge
$(\tilde{x}^1_{t-1|t-1}, \tilde{V}^1_{t-1|t-1})$
$(\tilde{x}^2_{t-1|t-1}, \tilde{V}^2_{t-1|t-1})$
Filter 1
Filter 2
$(x^1_{t|t}, V^1_{t|t})$
$(x^2_{t|t}, V^2_{t|t})$

Figure 2: The GPB1, GPB2 and IMM algorithms. Each of the filters takes an extra input, $\mathbf{y}_t$, and returns an extra output, L_t, which is not shown for clarity.

(L_t^j is the likelihood of the innovation at time t, given that the current model is j.)

2.3 Filtering

We perform the following steps in sequence.

$$\begin{aligned}
(\mathbf{x}_{t|t}^{i(j)}, V_{t|t}^{i(j)}, V_{t,t-1|t}^{i(j)}, L_t^{i(j)}) &= \text{Filter}(\mathbf{x}_{t-1|t-1}^{i}, V_{t-1|t-1}^{i}, \mathbf{y}_t; F_j, H_j, Q_j, R_j) \\
M_{t-1,t|t}(i,j) &= \Pr(S_{t-1} = i, S_t = j | \mathbf{y}_{1:t}) = \frac{L_t(i,j) Z(i,j) M_{t-1|t-1}(i)}{\sum_i \sum_j L_t(i,j) Z(i,j) M_{t-1|t-1}(i)} \\
M_{t|t}(j) &= \sum_i M_{t-1,t|t}(i,j) \\
W^{i|j} &= \Pr(S_{t-1} = i | S_t = j, \mathbf{y}_{1:t}) = M_{t-1,t|t}(i,j) / M_{t|t}(j) \\
(\mathbf{x}_{t|t}^{j}, V_{t|t}^{j}) &= \text{Collapse}(\mathbf{x}_{t|t}^{i(j)}, V_{t|t}^{i(j)}, W_t^{i|j})
\end{aligned}$$

The definitions of the filter, smoother and collapse operators are given in Appendix A; for a derivation, see e.g., [BSL93]; for the derivation of the cross-variance term, see [DRO93]; the deriviation of the mode update equation is as follows.

$$\begin{aligned}
\Pr(S_{t-1} = i, S_t = j | \mathbf{y}_t, \mathbf{y}_{1:t-1}) &= \frac{1}{c} \Pr(S_{t-1} = i, S_t = j, \mathbf{y}_t | \mathbf{y}_{1:t-1}) \\
&= \frac{1}{c} \Pr(\mathbf{y}_t | S_{t-1} = i, S_t = j, \mathbf{y}_{1:t-1}) \Pr(S_{t-1} = i, S_t = j | \mathbf{y}_{1:t-1}) \\
&= \frac{1}{c} \Pr(\mathbf{y}_t | S_{t-1} = i, S_t = j, \mathbf{y}_{1:t-1}) \Pr(S_t = j | S_{t-1} = i, \mathbf{y}_{1:t-1}) \Pr(S_{t-1} = i, \mathbf{y}_{1:t-1}) \\
&= \frac{1}{c} L_t(i,j) Z(i,j) M_{t-1|t-1}(i)
\end{aligned}$$

where c is the normalization constant

$$c = \sum_i \sum_j L_t(i,j) Z(i,j) M_{t-1|t-1}(i)$$

The initial conditions are as follows. We set the predicted mean and variance based on no evidence to be $\mathbf{x}_{1|0}^j = \mathrm{E}\left[\mathbf{X}_1 | S_1 = j\right] = \mu^j$ and $V_{1|0}^j = \mathrm{Cov}\left[\mathbf{X}_1 | S_1 = j\right] = \Sigma^j$, and we set $M_{0|0} = \pi$.

2.4 Smoothing

We perform the following steps in sequence.

$$\begin{aligned}
(\mathbf{x}_{t|T}^{(j)k}, V_{t|T}^{(j)k}, V_{t+1,t|T}^{j(k)}) &= \text{Smooth}(\mathbf{x}_{t+1|T}^{k}, V_{t+1|T}^{k}, \mathbf{x}_{t|t}^{j}, V_{t|t}^{j}, V_{t+1|t+1}^{k}, V_{t+1,t|t+1}^{j(k)}; F_k, Q_k) \\
U_t^{j|k} &= \Pr(S_t = j | S_{t+1} = k, \mathbf{y}_{1:T}) \approx \frac{M_{t|t}(j) Z(j,k)}{\sum_{j'} M_{t|t}(j') Z(j',k)} \qquad * \\
M_{t,t+1|T}(j,k) &= U_t^{j|k} M_{t+1|T}(k) \\
M_{t|T}(j) &= \sum_k M_{t,t+1|T}(j,k)
\end{aligned}$$

$$\begin{aligned}
W_t^{k|j} &= \Pr(S_{t+1} = k|S_t = j, \mathbf{y}_{1:T}) = M_{t,t+1|T}(j,k)/M_{t|T}(j) \\
(\mathbf{x}_{t|T}^j, V_{t|T}^j) &= \text{Collapse}(\mathbf{x}_{t|T}^{(j)k}, V_{t|T}^{(j)k}, W_t^{k|j}) \\
(\mathbf{x}_{t|T}, V_{t|T}) &= \text{Collapse}(\mathbf{x}_{t|T}^j, V_{t|T}^j, M_{t|T}(j)) \\
\mathbf{x}_{t+1|T}^{j(k)} &= \text{E}\left[\mathbf{x}_{t+1}|\mathbf{y}_{1:T}, S_{t+1} = k, S_t = j\right] \approx \mathbf{x}_{t+1|T}^k \\
V_{t+1,t|T}^k &= \text{CollapseCross}(\mathbf{x}_{t+1|T}^{j(k)}, \mathbf{x}_{t|T}^{(j)k}, V_{t+1,t|T}^{j(k)}, U_t^{j|k}) \\
\mathbf{x}_{t|T}^{()k} &= \text{E}\left[\mathbf{X}_t|\mathbf{y}_{1:T}, S_{t+1} = k\right] = \sum_j \mathbf{x}_{t|T}^{(j)k} U_t^{j|k} \\
V_{t+1,t|T} &= \text{CollapseCross}(\mathbf{x}_{t+1|T}^k, \mathbf{x}_{t|T}^{()k}, V_{t+1,t|T}^k, M_{t+1|T}(k))
\end{aligned}$$

The line marked * is a standard approximation [Kim94], derived as follows.

$$\begin{aligned}
\Pr(S_t = j|S_{t+1} = k, \mathbf{y}_{1:T}) &\approx \Pr(S_t = j|S_{t+1} = k, \mathbf{y}_{1:t}) \\
&= \frac{\Pr(S_t = j|\mathbf{y}_{1:t}) \Pr(S_{t+1} = k|S_t = j)}{\Pr(S_{t+1} = k|\mathbf{y}_{1:t})}
\end{aligned}$$

where the approximation arised because S_t is *not* conditionally independent of the future evidence $\mathbf{y}_{t+1}, \ldots, \mathbf{y}_T$ given S_{t+1}.[3] This approximation will not be too bad provided future evidence does not contain much information about S_t beyond than that contained in S_{t+1}.

3 Learning

Once again, we start by considering the simpler SAR case before extending it to the SKF case.

3.1 Switching AR model

We are interested in finding the Maximum Likelihood estimate of the parameters. If we knew the segmentation (i.e., which model to apply at each time step), we could solve this using linear regression. Since the S_t values are unobserved, we shall use EM (c.f., [Ham90]).

The complete data log likelihood is

$$L = \log P(\mathbf{x}_{1:T}, S_{1:T}) = \sum_{t=2}^{T} \log P(\mathbf{x}_t|\mathbf{x}_{t-1}, S_t) + \sum_{t=2}^{T} \log \Pr(S_t|S_{t-1}) + \log P(\mathbf{x}_1|S_1) + \log \Pr(S_1)$$

where

$$P(S_t = j|S_{t-1} = i) = Z(i,j)$$

$$P(\mathbf{x}_t|\mathbf{x}_{t-1}, S_t = j) = \exp\left(-\tfrac{1}{2}[\mathbf{x}_t - A_j\mathbf{x}_{t-1}]' Q_j^{-1} [\mathbf{x}_t - A_j\mathbf{x}_{t-1}]\right) (2\pi)^{-n/2} |Q_j|^{-\frac{1}{2}}$$

$$P(S_1 = j) = \pi_j$$

[3] The best way to see this is in terms of the directed graphical model. Recall that a node is conditionally independent of its non-descendants given its parents [Pea88]. Hence S_t is independent of past evidence given only its parent S_{t-1}. However, all of S_t's children need to be observed to block the effect of future evidence (observing S_{t+1} is not enough because the arrow points the "wrong" way). In the AR model, all of S_t's children (namely $\mathbf{Y}_t$ and S_{t+1}) *are* observed, but this is not the case in the general model (since $\mathbf{X}_t$ is not observed).

$$P(\mathbf{x}_1 | S_1 = j) = N(\mu_j, \Sigma_j)$$

In EM, we iteratively maximize (w.r.t. the parameters θ) the expected value (w.r.t. the parameters θ^{old}) of the complete data log likelihood:

$$\begin{aligned}
\hat{L} &= \mathrm{E}_{P(S_{1:T}, \mathbf{x}_{1:T})}[L] \\
&= \sum_{S_1} \cdots \sum_{S_T} P(\mathbf{x}_{1:T}, S_{1:T}; \theta^{\text{old}}) \log P(\mathbf{x}_{1:T}, S_{1:T}; \theta) \\
&= P(\mathbf{x}_{1:T}) \sum_{t=2}^{T} \sum_{S_t} \left(\sum_{\{S_\tau, \tau \neq t\}} P(S_{1:T} | \mathbf{x}_{1:T}) \right) \log P(\mathbf{x}_t | \mathbf{x}_{t-1}, S_t) + \cdots \\
&= P(\mathbf{x}_{1:T}) \sum_{t=2}^{T} \sum_{S_t = j} W_t^j \log P(\mathbf{x}_t | \mathbf{x}_{t-1}, S_t = j) + \cdots
\end{aligned}$$

where the weights $W_t^j = \Pr(S_t = j | \mathbf{x}_{1:T})$ were computed during the inference step, and we have used the fact that $\mathbf{X}_t \perp S_\tau | \mathbf{X}_{t-1}, S_t$ for all $\tau \neq t$ (since a node is independent of its non-descendants given its parents).

To maximize this equation, we take derivatives and set to 0; intuitively, the derivative will kill all terms but one in the summation over S_t, resulting in a weighted version of the standard formulas: see Appendix B for details. Assuming we have N iid sequences, indexed by ℓ, we find (where $\hat{\mathbf{x}}_t = \mathbf{x}_t$, $P_t = \mathbf{x}_t \mathbf{x}_t'$ and $P_{t,t-1} = \mathbf{x}_t \mathbf{x}_{1-1}'$)

$$\begin{aligned}
A_i &= \left(\sum_\ell \sum_{t=2}^{T} W_t^i P_{t,t-1} \right) \left(\sum_\ell \sum_{t=2}^{T} W_t^i P_{t-1} \right)^{-1} \\
Q_i &= \left(\frac{1}{\sum_\ell \sum_{t=2}^{T} W_t^i} \right) \left(\sum_\ell \sum_{t=2}^{T} W_t^i P_t - A_i \sum_\ell \sum_{t=2}^{T} W_t^i P_{t,t-1}{}' \right) \\
\mu_i &= \frac{\sum_\ell W_1^i \hat{\mathbf{x}}_1}{\sum_\ell W_1^i} \\
\Sigma_i &= \frac{\sum_\ell W_1^i (\hat{\mathbf{x}}_1 - \mu_i)(\hat{\mathbf{x}}_1 - \mu_i)'}{\sum_\ell W_1^i} = \frac{\sum_\ell W_1^i \hat{\mathbf{x}}_1 \hat{\mathbf{x}}_1' - \mu_i (\sum_\ell W_1^i \mathbf{x}_1') - (\sum_\ell W_1^i \hat{\mathbf{x}}_1) \mu_i' + (\sum_\ell W_1^i) \mu_i \mu_i'}{\sum_\ell W_1^i} \\
Z(i,j) &= \frac{\sum_\ell \sum_{t=2}^{T} \Pr(S_{t-1} = i, S_t = j | \mathbf{y}_{1:T})}{\sum_\ell \sum_{t=1}^{T-1} W_t^i} \\
\pi_i &= \frac{1}{N} \sum_\ell W_1^i
\end{aligned}$$

The formulas for Z and π are the same as for an HMM [Rab89]; the formulas for μ_i and Σ_i are the same as for a mixture of Gaussians (see e.g., [Bis95, XJ96])[4]; and the formulas for A_i and Q_i are in fact special cases of linear regression.

These equations are the MLEs for the case in which there are no restrictions on the form of the matrices. Typically, however, we know that some entries must be 0 or 1 (or some other known value). It can be shown that the constrained MLE is obtained by first computing the unconstrained MLE as above, and then setting the constrained entries to their correct values (i.e., projecting onto the allowable subspace). For example, to estimate a covariance matrix which is constrained to be diagonal, we can compute $\hat{Q}$ as above, and then set the off-diagonal entries to 0.

[4] We have written the formula for Σ_i in the usual form, and also in a form which is easier to compute in an incremental fashion [NH98] from the sufficient statistics. Remember that $\hat{\mathbf{x}}_1$ and W_1^i are functions of ℓ.

To achieve parameter tieing, we pool the expected sufficient statistics for each parameter in the equivalence class. For example, if we have $Q_i = Q$ for all $i \in \mathcal{S}$, we replace W_t^i with $\sum_{i \in \mathcal{S}} W_t^i$ when estimating Q.

A well-known problem with mixtures-of-Gaussians models, even in the non-dynamic case, is that the covariance matrix can easily become singular. Hamilton [Ham90, Ham91] suggests using a Wishart prior [DeG70] to regularize the problem. In particular, suppose the prior is $Q_i^{-1} \sim W(\alpha_i, \Lambda_i)$, where α_i is our equivalent sample size for the precision matrix Λ_i. Then the MAP estimate of Q_i is given by

$$Q_i = \left(\frac{1}{\alpha_i + \sum_\ell \sum_{t=2}^T W_t^i} \right) \left(\Lambda_i + \sum_\ell \sum_{t=2}^T W_t^i P_t - A_i \sum_\ell \sum_{t=2}^T W_t^i P_{t,t-1}{}' \right)$$

We have found that setting $\alpha_i = 0.1NT$ (i.e., we imagine we have seen 10% of the data before) and using $\Lambda_i = \alpha_i I$ works quite well in practice.

3.2 Switching Kalman filter model

In this case, the complete data log likelihood is given by

$$\begin{aligned} L &= \log P(\mathbf{x}_{1:T}, S_{1:T}, \mathbf{y}_{1:T}) = -\tfrac{1}{2} \sum_{t=1}^T \left([\mathbf{y}_t - C_t \mathbf{x}_t]' R_t{}^{-1} [\mathbf{y}_t - C_t \mathbf{x}_t] \right) - \tfrac{1}{2} \sum_{t=1}^T \log |R_t| \\ & \quad -\tfrac{1}{2} \sum_{t=2}^T \left([\mathbf{x}_t - A_t \mathbf{x}_{t-1}]' Q_t{}^{-1} [\mathbf{x}_t - A_t \mathbf{x}_{t-1}] \right) - \tfrac{1}{2} \sum_{t=2}^T \log |Q_t| \\ & \quad -\tfrac{1}{2} [\mathbf{x}_1 - \mu_{\mathbf{1}}]' \Sigma_1^{-1} [\mathbf{x}_1 - \mu_{\mathbf{1}}] - \tfrac{1}{2} \log |\Sigma_1| - \frac{T(n+m)}{2} \log 2\pi \\ & \quad + \log \pi_{\mathbf{1}} + \sum_{t=2}^T \log Z(S_{t-1}, S_t) \end{aligned}$$

The quantity we maximise is

$$\begin{aligned} \hat{L} &= \mathrm{E}_{P(S_{1:T}, \mathbf{x}_{1:T}, \mathbf{y}_{1:T})}[L] \\ &= \mathrm{E}_{P(S_{1:T}, \mathbf{y}_{1:T})} \left[\mathrm{E}_{P(\mathbf{x}_{1:T} | S_{1:T}, \mathbf{y}_{1:T})}[L] \right] \\ &\approx \mathrm{E}_{P(S_{1:T}, \mathbf{y}_{1:T})} \left[\mathrm{E}_{P(\mathbf{x}_{1:T} | \mathbf{y}_{1:T})}[L] \right] \\ &= P(\mathbf{y}_{1:T}) \sum_{t=2}^T \sum_{S_t} \left(\sum_{\{S_\tau, \tau \neq t\}} P(S_{1:T} | \mathbf{y}_{1:T}) \right) \hat{E}[\log P(\mathbf{x}_t | \mathbf{x}_{t-1}, S_t)] + \cdots \\ &= P(\mathbf{y}_{1:T}) \sum_{t=2}^T \sum_{S_t = j} W_t^j \hat{E}[\log P(\mathbf{x}_t | \mathbf{x}_{t-1}, S_t)] + \cdots \end{aligned}$$

where $\hat{E}[\cdot] = \mathrm{E}[\cdot | \mathbf{y}_{1:T}]$. The approximation arises because we use $\mathrm{E}[\mathbf{x}_t | \mathbf{y}_{1:T}]$ instead of $\mathrm{E}[\mathbf{x}_t | \mathbf{y}_{1:T}, S_{1:T}]$, since the latter is an exponential number of vectors (one for each segmentation). The advantage is that the formula is now of the same form as Equation 1 for the SAR case (modulo the leading constant factor), with the following redefinitions:

$$\begin{aligned} W_t^j &= \Pr(S_t = j | \mathbf{y}_{1:T}] \\ \hat{\mathbf{x}}_t &= \hat{E}[\mathbf{x}_t] \end{aligned}$$

$$
\begin{aligned}
P_t &= \hat{E}[\mathbf{x}_t \mathbf{x}_t'] = V_{t|T} + \mathbf{x}_{t|T} \mathbf{x}_{t|T}' \\
P_{t,t-1} &= \hat{E}[\mathbf{x}_t \mathbf{x}_{t-1}'] = V_{t,t-1|T} + \mathbf{x}_{t|T} \mathbf{x}_{t-1|T}'
\end{aligned}
$$

Of course, since we have already computed terms like $\mathbf{x}_{t|T}^j$, V_t^j, and $V_{t,t-1}^{i(j)}$, we could use these instead. In practice, this doesn't seem to make much difference, although it does have the advantage that we don't need to collapse the cross variance terms (twice) to compute $V_{t,t-1}$ (the last four equations in Section 2.4).

Of course, the new equation for $\hat{L}$ also has terms involving C_i and R_i. Maximizing with respect to these gives (see Appendix B for the derivation):

$$
\begin{aligned}
C_i &= \left(\sum_{\ell} \sum_{t=1}^{T} W_t^i \mathbf{y}_t \hat{\mathbf{x}}_t{}' \right) \left(\sum_{\ell} \sum_{t=1}^{T} W_t^i P_t \right)^{-1} \\
R_i &= \left(\frac{1}{\sum_{\ell} \sum_{t=1}^{T} W_t^i} \right) \sum_{\ell} \sum_{t=1}^{T} W_t^i \left(\mathbf{y}_t \mathbf{y}_t' - C_i \hat{\mathbf{x}}_t \mathbf{y}_t' \right)
\end{aligned}
$$

3.3 Deterministic annealing

EM is notorious for getting stuck in local minima. This is especially common in models of the kind we are considering, which have M^T possible segmentations. One solution is to use deterministic annealing EM [UN98], as suggested in [GH96b]. In DAEM, we replace the posterior

$$
\Pr(H|o) = \frac{\Pr(H,o)}{\sum_H \Pr(H,o)}
$$

(where H are the hidden variables and o the observed values) with

$$
f(H|o) = \frac{\Pr(H,o)^{\beta}}{\sum_H \Pr(H,o)^{\beta}}
$$

where β is an inverse temperature parameter. When $\beta = 0$ (infinite temperature), the posterior becomes uniform. When $\beta = 1$ (low temperature), the posterior becomes the regular EM posterior. Applying this principle to the present case, we suggest using

$$
f_t^j = \frac{(W_t^j)^{\beta}}{\sum_j (W_t^j)^{\beta}}
$$

A Appendix: Details of the inference algorithm

A.1 Filter

The Filter operator

$$
(\mathbf{x}_{t|t}, V_{t|t}, V_{t,t-1|t}, L_t) = \mathrm{Filter}(\mathbf{x}_{t-1|t-1}, V_{t-1|t-1}, \mathbf{y}_t; F_t, H_t, Q_t, R_t)
$$

is defined as follows. First, we compute the predicted mean and variance.

$$
\begin{aligned}
\mathbf{x}_{t|t-1} &= F \mathbf{x}_{t-1|t-1} \\
V_{t|t-1} &= F V_{t-1|t-1} F' + Q
\end{aligned}
$$

Then we compute the error in our prediction (the innovation), the variance of the error, the Kalman gain matrix, and the likelihood of this observation.

$$\begin{aligned}
\mathbf{e}_t &= \mathbf{y}_t - H\mathbf{x}_{t|t-1} \\
S_t &= HV_{t|t-1}H' + R \\
K_t &= V_{t|t-1}H'S_t^{-1} \\
L_t &= N(\mathbf{e}_t; 0, S_t)
\end{aligned}$$

Finally, we update our estimates of the mean, variance, and cross variance.

$$\begin{aligned}
\mathbf{x}_{t|t} &= \mathbf{x}_{t|t-1} + K_t\mathbf{e}_t \\
V_{t|t} &= (I - K_tH)V_{t|t-1} = V_{t|t-1} - K_tS_tK_t' \\
V_{t,t-1|t} &= (I - K_tH)FV_{t-1|t-1}
\end{aligned}$$

A.2 Smoother

The Smooth operator

$$(\mathbf{x}_{t|T}, V_{t|T}, V_{t+1,t|T}) = \text{Smooth}(\mathbf{x}_{t+1|T}, V_{t+1|T}, \mathbf{x}_{t|t}, V_{t|t}, V_{t+1|t+1}, V_{t+1,t|t+1}; F_{t+1}, Q_{t+1})$$

is defined as follows. First we compute the following predicted quantities (or we could pass them in from the filtering stage):

$$\begin{aligned}
\mathbf{x}_{t+1|t} &= F_{t+1}\mathbf{x}_{t|t} \\
V_{t+1|t} &= F_{t+1}V_{t|t}F_{t+1}' + Q_{t+1}
\end{aligned}$$

Then we compute the smoother gain matrix.

$$J_t = V_{t|t}F_{t+1}'V_{t+1|t}^{-1}$$

Finally, we update our estimates of the mean, variance, and cross variance.

$$\begin{aligned}
\mathbf{x}_{t|T} &= \mathbf{x}_{t|t} + J_t\left(\mathbf{x}_{t+1|T} - \mathbf{x}_{t+1|t}\right) \\
V_{t|T} &= V_{t|t} + J_t\left(V_{t+1|T} - V_{t+1|t}\right)J_t' \\
V_{t+1,t|T} &= V_{t+1,t|t+1} + \left(V_{t+1|T} - V_{t+1|t+1}\right)V_{t+1|t+1}^{-1}V_{t+1,t|t+1}
\end{aligned}$$

We now present an alternative way to compute the smoothed estimates of the cross-variance terms, $V_{t,t-1|T}$, which does not require the corresponding filtered terms [SS91, GH96a].

The Smooth' operator

$$(\mathbf{x}_{t|T}, V_{t|T}, V_{t,t-1|T}) = \text{Smooth}'(\mathbf{x}_{t+1|T}, V_{t+1|T}, V_{t+1,t|T}, \mathbf{x}_{t|t}, V_{t|t}, V_{t-1|t-1}; F_{t+1}, Q_{t+1}, F_t, Q_t)$$

is defined as above, except

$$V_{t,t-1|T} = V_{t|t}J_{t-1}' + J_t(V_{t+1,t|T} - F_{t+1}V_{t|t})J_{t-1}'$$

where the boundary condition is

$$V_{T,T-1|T} = (I - K_TH_T)F_TV_{T-1|T-1}$$

A.3 Collapse

Consider two random variables X, Y with conditional means $\mu_X^j = E[X|S=j]$, $\mu_Y^j = E[Y|S=j]$, cross variance $V_{X,Y}^j = \text{Cov}\,[X,Y|S=j]$, and mixing coefficients $P^j = \Pr(S=j)$. Then we can compute the unconditional moments as follows. (This procedure is called moment matching.)

$$\begin{aligned} \mu_X &= \sum_j P^j \mu_X^j \\ \mu_Y &= \sum_j P^j \mu_Y^j \end{aligned}$$

$$\begin{aligned} V_{X,Y} &= \sum_j P^j \text{Cov}\,[X,Y|S=j] \\ &= \sum_j P^j \text{E}\,[(X-\mu_X)(Y-\mu_Y)'|S=j] \\ &= \sum_j P^j \text{E}\,[(X-\mu_X^j+\mu_X^j-\mu_X)(Y-\mu_Y^j+\mu_Y^j-\mu_Y)'|S=j] \\ &= \sum_j P^j \text{E}\,[(X-\mu_X^j)(Y-\mu_Y)'|S=j] + \sum_j P^j(\mu_X^j-\mu_X)(\mu_Y^j-\mu_Y)' \\ &= \sum_j P^j V_{X,Y}^j + \sum_j P^j(\mu_X^j-\mu_X)(\mu_Y^j-\mu_Y)' \end{aligned}$$

Let us introduce the following shorthand for the above operation.

$$(\mu_X, \mu_Y, V_{X,Y}) \quad = \quad \text{CollapseCross}(\mu_X^j, \mu_Y^j, V_{X,Y}^j, P^j)$$

and define

$$\text{Collapse}(\mu_X^j, V_X^j, P^j) = \text{CollapseCross}(\mu_X^j, \mu_X^j, V_{X,X}^j, P^j)$$

It can be shown [Lau96] that a Gaussian with these moments is the closest possible Gaussian (in KL distance) to the original mixture distribution.

B Derivation of the M step

We follow [GH96a], but generalize to the switching case. For simplicity of notation, we consider only a single sequence.

We use following standard identities

$$\frac{\partial\left((\mathbf{Xa}+\mathbf{b})'\mathbf{C}(\mathbf{Xa}+\mathbf{b})\right)}{\partial\mathbf{X}} = (\mathbf{C}+\mathbf{C}')(\mathbf{Xa}+\mathbf{b})\mathbf{a}' \tag{1}$$

$$\frac{\partial(\mathbf{a}'\mathbf{Xb})}{\partial\mathbf{X}} = \mathbf{ab}' \tag{2}$$

$$\frac{\partial \ln|\mathbf{X}|}{\partial\mathbf{X}} = (\mathbf{X}')^{-1} \tag{3}$$

B.1 System matrix

Using identity 1,

$$\begin{aligned} \frac{\partial}{\partial A_i}\hat{L} &= -\tfrac{1}{2}\sum_{t=2}^{T} W_t^i \hat{E}\left[2Q_i^{-1}\left(\mathbf{x}_t - A_i\mathbf{x}_{t-1}\right)\mathbf{x}_{t-1}{}'\right] \\ &= -\sum_{t=2}^{T} W_t^i Q_i^{-1} P_{t,t-1} + \sum_{t=2}^{T} W_t^i Q_i^{-1} A_i P_{t-1} = 0 \end{aligned}$$

Hence

$$A_i = \left(\sum_{t=2}^{T} W_t^i P_{t,t-1}\right)\left(\sum_{t=2}^{T} W_t^i P_{t-1}\right)^{-1} \tag{4}$$

B.2 System noise covariance

Using identities 2 and 3,

$$\begin{aligned} \frac{\partial}{\partial Q_i^{-1}}\hat{L} &= -\tfrac{1}{2}\sum_{t=2}^{T} W_t^i \hat{E}\left[\left(\mathbf{x}_t - A_i\mathbf{x}_{t-1}\right)\left(\mathbf{x}_t - A_i\mathbf{x}_{t-1}\right)'\right] + \tfrac{1}{2}\sum_{t=2}^{T} W_t^i Q_i \\ &= -\tfrac{1}{2}\sum_{t=2}^{T} W_t^i \left[P_t - A_i P_{t,t-1}{}' - P_{t,t-1}A_i' + A_i P_{t-1} A_i'\right] + \tfrac{1}{2}\sum_{t=2}^{T} W_t^i Q_i = 0 \end{aligned}$$

Using the new value of A_i and the fact that P_t is symmetric, we find

$$\begin{aligned} A_i\left(\sum_{t=2}^{T} W_t^i P_{t-1}\right)A_i' &= \left(\sum_{t=2}^{T} W_t^i P_{t,t-1}\right)\left(\sum_{t=2}^{T} W_t^i P_{t-1}\right)^{-1}\left(\sum_{t=2}^{T} W_t^i P_{t,t-1}{}'\right) \\ &= A_i \sum_{t=2}^{T} W_t^i P_{t,t-1}{}' = \left(\sum_{t=2}^{T} W_t^i P_{t,t-1}\right)A_i' \end{aligned}$$

Hence

$$Q_i = \left(\frac{1}{\sum_{t=2}^{T} W_t^i}\right)\left(\sum_{t=2}^{T} W_t^i P_t - A_i \sum_{t=2}^{T} W_t^i P_{t,t-1}{}'\right) \tag{5}$$

B.3 Observation matrix

Using identity 1,

$$\frac{\partial}{\partial C_i}\hat{L} = -\tfrac{1}{2}\sum_{t} W_t^i \hat{E}\left[2R_i^{-1}\left(-C_i\mathbf{x}_t + \mathbf{y}_t\right)\mathbf{x}_t'\right] = 0$$

Hence

$$C_i = \left(\sum_{t=1}^{T} W_t^i \mathbf{y}_t \hat{\mathbf{x}}_t{}'\right)\left(\sum_{t=1}^{T} W_t^i P_t\right)^{-1} \tag{6}$$

B.4 Observation noise covariance

Using identities 2 and 3, we find

$$\frac{\partial}{\partial R_i^{-1}} \hat{L} = \sum_{t=1}^{T} \hat{E}\left[W_t^i \tfrac{1}{2}\left(\mathbf{y}_t\mathbf{y}_t' - 2C_i\mathbf{x}_t\mathbf{y}_t' + C_i\mathbf{x}_t\mathbf{x}_t'C_i'\right)\right] + \tfrac{1}{2}R_i \sum_t W_t^i = 0$$

Using the new estimate of C_i, we have

$$\left(\sum_t W_t^i P_t\right) C_i' = \sum_t W_t^i \hat{\mathbf{x}}_t \mathbf{y}_t' \overset{\text{def}}{=} Z$$

so

$$\frac{\partial}{\partial R_i^{-1}} \hat{L} = \sum_{t=1}^{T} \tfrac{1}{2}\left(\sum_t W_t^i \mathbf{y}_t\mathbf{y}_t' - 2C_i Z + C_i Z\right) + \tfrac{1}{2}R_i \sum_t W_t^i$$

and hence

$$R_i = \left(\frac{1}{\sum_{t=1}^{T} W_t^i}\right) \sum_{t=1}^{T} W_t^i \left(\mathbf{y}_t\mathbf{y}_t' - C_i\hat{\mathbf{x}}_t\mathbf{y}_t'\right) \qquad (7)$$

B.5 Initial mean and covariance

This is the standard derivation for a mixture of Gaussians model; see e.g., [Ham90, Bis95, XJ96].

B.6 Initial state probability and transition matrix

This is a constrained maximization problem (since the probabilities must sum to 1), which can be solved using Lagrange multipliers; see e.g., [Rab89, Ham90] for a derivation.

References

[Bis95] C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, 1995.

[BK98a] X. Boyen and D. Koller. Approximate learning of dynamic models. In *Neural Info. Proc. Systems*, 1998.

[BK98b] X. Boyen and D. Koller. Tractable inference for complex stochastic processes. In *Proc. of the Conf. on Uncertainty in AI*, 1998.

[BMR98] M. Billio, A. Monfort, and C. P. Robert. Bayesian estimation of switching ARMA models. Technical report, CREST, INSEE, Paris, 1998.

[BSF88] Y. Bar-Shalom and T. Fortmann. *Tracking and data association.* Academic Press, 1988.

[BSL93] Y. Bar-Shalom and X. Li. *Estimation and Tracking: Principles, Techniques and Software.* Artech House, 1993.

[CK96] C. Carter and R. Kohn. Markov Chain Monte Carlo in conditionally Gaussian state space models. Technical report, Univ. New South Wales, Graduate School of Management, 1996.

[DeG70] M. DeGroot. *Optimal Statistical Decisions.* McGraw-Hill, 1970.

[DRO93] V. Digalakis, J. R. Rohlicek, and M. Ostendorf. ML estimation of a stochastic linear systems with the EM algorithm and its application to speech recognition. *IEEE Trans. on Speech and Audio Proc.*, 1(4):421–442, 1993.

[DW91] Thomas L. Dean and Michael P. Wellman. *Planning and Control.* Morgan Kaufmann, 1991.

[GH96a] Z. Ghahramani and G. Hinton. Parameter estimation for linear dynamical systems. Technical Report CRG-TR-96-2, Dept. Comp. Sci., Univ. Toronto, 1996.

[GH96b] Z. Ghahramani and G. Hinton. Switching state-space models. Technical Report CRG-TR-96-3, Dept. Comp. Sci., Univ. Toronto, 1996.

[Gha97] Z. Ghahramani. Learning dynamic bayesian networks. In C.L. Giles and M. Gori, editors, *Adaptive Processing of Temporal Information . Lecture Notes in Artificial Intelligence.* Springer-Verlag, 1997. To appear.

[Ham90] J. Hamilton. Analysis of time series subject to changes in regime. *J. Econometrics*, 45:39–70, 1990.

[Ham91] J. Hamilton. A quasi-bayesian approach to estimating parameters for mixtures of normal distributions. *J. Business and Economic Statistics*, 1991.

[Kim94] C-J. Kim. Dynamic linear models with Markov-switching. *J. of Econometrics*, 60:1–22, 1994.

[KRHE96] D. Kulp, M. G. Reese, D. Haussler, and F. H. Eckman. A generalized hidden Markov model for the recognition of human genes in DNA. In *International Conf. on Intelligent Systems for Molecular Biology*, 1996. To appear.

[Lau96] S. Lauritzen. *Graphical Models.* OUP, 1996.

[NH98] R. M. Neal and G. E. Hinton. A new view of the EM algorithm that justifies incremental and other variants. In M. Jordan, editor, *Learning in Graphical Models.* Kluwer, 1998.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.

[PG88] D. Pena and I. Guttman. Bayesian approach to robustifying the Kalman filter. In C. Spall, editor, *Bayesian analysis of time series and dynamic models.* Marcel Dekker, 1988.

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

[SM80] A. Smith and U. Makov. Bayesian detection and estimation of jumps in linear systems. In O. Jacobs, M. Davis, M. Dempster, C. Harris, and P. Parks, editors, *Analysis and optimization of stochastic systems.* 1980.

[SS91] R. Shumway and D. Stoffer. Dynamic linear models with switching. *J. of the Am. Stat. Assoc.*, 86:763–769, 1991.

[TW97] J. Timmer and A. S. Weigend. Modeling volatility using state space models. *Intl. J. of Neural Systems*, 8, 1997.

[UN98] N. Ueda and R. Nakano. Deterministic annealing EM algorithm. *Neural Networks*, 11:271–282, 1998.

[Wil76] A. Willsky. A survey of design methods for failure detection in dynamic systems. *Automatica*, 12:601–611, 1976.

[XJ96] L. Xu and M. I. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8:129–151, 1996.