732A96/TDDE15 Advanced Machine Learning

State Space Models

Jose M. Peña
IDA, Linköping University, Sweden

Lecture 12: Learning Linear Gaussian State Space Models

Contents

- Scaling Factors for Hidden Markov Models
- $\hat{\alpha} - \gamma$ Recursion for Linear Gaussian State Space Models
- EM Algorithm for Linear Gaussian State Space Models
- Particle Filter
- Switching State Space Models

Literature

- Main source
  - Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006. Chapters 13.2.4 and 13.3.
- Additional source
  - Ghahramani Z. and Hinton, G. E. Variational Learning for Switching State-Space Models. *Neural Computation* 12, 831-864, 2000.

Scaling Factors for Hidden Markov Models

- FB algorithm = efficient way to compute the quantities required for the EM algorithm, filtering and smoothing.
- Recall that $\alpha(z_t) = p(x_{1:t}, z_t)$ and $\beta(z_t) = p(x_{t+1:T}|z_t)$.
- Problem: Arithmetic underflow for moderate lengths, e.g. 100.
- Solution: Use re-scaled or normalized versions, i.e.

$$\hat{\alpha}(z_t) = \frac{\alpha(z_t)}{p(x_{1:t})} = p(z_t|x_{1:t}) \text{ and } \hat{\beta}(z_t) = \frac{\beta(z_t)}{p(x_{t+1:T}|x_{1:t})}$$

- By defining $c_t = p(x_t|x_{1:t-1})$, we have that

$$\alpha(z_t) = \Big(\prod_{s=1}^{t} c_s\Big)\hat{\alpha}(z_t) \text{ and } \beta(z_t) = \Big(\prod_{s=t+1}^{T} c_s\Big)\hat{\beta}(z_t)$$

and thus

$$c_t\hat{\alpha}(z_t) = p(x_t|z_t)\sum_{z_{t-1}} \hat{\alpha}(z_{t-1})p(z_t|z_{t-1}) \;//\; c_t \text{ is the RHS normalization}$$

$$c_{t+1}\hat{\beta}(z_t) = \sum_{z_{t+1}} \hat{\beta}(z_{t+1})p(x_{t+1}|z_{t+1})p(z_{t+1}|z_t) \;//\; c_{t+1} \text{ computed in } \hat{\alpha} \text{ phase}$$

$$\gamma(z_t) = p(z_t|x_{1:T}) = \hat{\alpha}(z_t)\hat{\beta}(z_t)$$

$$\xi(z_{t-1}, z_t) = p(z_{t-1}, z_t|x_{1:T}) = c_t^{-1}\hat{\alpha}(z_{t-1})p(x_t|z_t)p(z_t|z_{t-1})\hat{\beta}(z_t)$$

- The $\hat{\alpha} - \hat{\beta}$ recursion is typically used with HMMs.
- The $\hat{\alpha} - \gamma$ recursion is typically used with linear Gaussian SSMs.

$\hat{\alpha} - \gamma$ Recursion for Linear Gaussian State Space Models

- We assume that the transition, emission and initial distributions are Gaussian, i.e.

$$p(z_t|z_{t-1}) = \mathcal{N}(z_t|Az_{t-1}, \Gamma)$$
$$p(x_t|z_t) = \mathcal{N}(x_t|Cz_t, \Sigma)$$
$$p(z_1) = \mathcal{N}(z_1|\mu_0, P_0)$$

or equivalently

$$z_t = Az_{t-1} + w_t \qquad \text{// Linear model}$$
$$x_t = Cz_t + v_t$$
$$z_1 = \mu_0 + u_0$$

where

$$w_t \sim \mathcal{N}(w_t|0, \Gamma) \qquad \text{// Gaussian noise}$$
$$v_t \sim \mathcal{N}(v_t|0, \Sigma)$$
$$u_0 \sim \mathcal{N}(u_0|0, P_0)$$

because recall that $E[Ax + B] = AE[x] + B$ and $cov[Ax + B] = Acov[x]A^T$.

$\hat{\alpha} - \gamma$ Recursion for Linear Gaussian State Space Models

- Note that $\hat{\alpha}(z_t) = p(z_t|x_{1:t}) = \mathcal{N}(z_t|\mu_t, V_t)$.
- In the recursion for HMMs, simply replace summation by integration, i.e.

$$\begin{aligned} c_t\hat{\alpha}(z_t) &= p(x_t|z_t) \int \hat{\alpha}(z_{t-1})p(z_t|z_{t-1})dz_{t-1} \\ &= \mathcal{N}(x_t|Cz_t, \Sigma) \int \mathcal{N}(z_{t-1}|\mu_{t-1}, V_{t-1})\mathcal{N}(z_t|Az_{t-1}, \Gamma)dz_{t-1} \end{aligned}$$

- Use the closed forms expressions to obtain the joint Gaussian distribution from the marginal and conditional, i.e.

$$\begin{aligned} \mu_t &= A\mu_{t-1} + K_t(x_t - CA\mu_{t-1}) \\ V_t &= (I - K_tC)P_{t-1} \\ c_t &= \mathcal{N}(x_t|CA\mu_{t-1}, CP_{t-1}C^T + \Sigma) \\ K_t &= P_{t-1}C^T(CP_{t-1}C^T + \Sigma)^{-1} \qquad \text{// Kalman gain matrix} \\ P_{t-1} &= AV_{t-1}A^T + \Gamma \end{aligned}$$

- For the initial conditions $t = 1$, replace $A\mu_{t-1}$ and P_{t-1} with μ_0 and P_0 in the equations above.

$\hat{\alpha} - \gamma$ Recursion for Linear Gaussian State Space Models

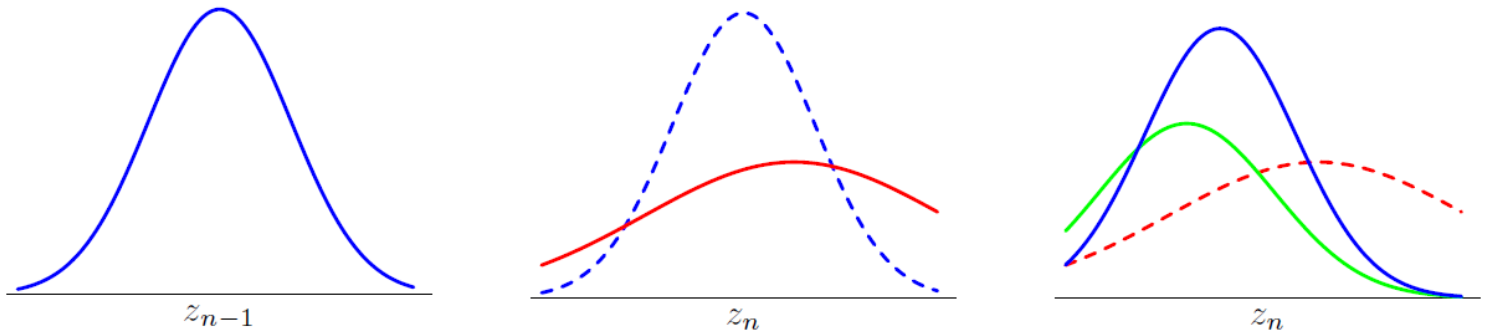


Figure 13.21 The linear dynamical system can be viewed as a sequence of steps in which increasing uncertainty in the state variable due to diffusion is compensated by the arrival of new data. In the left-hand plot, the blue curve shows the distribution $p(\mathbf{z}_{n-1}|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1})$, which incorporates all the data up to step $n-1$. The diffusion arising from the nonzero variance of the transition probability $p(\mathbf{z}_n|\mathbf{z}_{n-1})$ gives the distribution $p(\mathbf{z}_n|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1})$, shown in red in the centre plot. Note that this is broader and shifted relative to the blue curve (which is shown dashed in the centre plot for comparison). The next data observation $\mathbf{x}_n$ contributes through the emission density $p(\mathbf{x}_n|\mathbf{z}_n)$, which is shown as a function of $\mathbf{z}_n$ in green on the right-hand plot. Note that this is not a density with respect to $\mathbf{z}_n$ and so is not normalized to one. Inclusion of this new data point leads to a revised distribution $p(\mathbf{z}_n|\mathbf{x}_1, \ldots, \mathbf{x}_n)$ for the state density shown in blue. We see that observation of the data has shifted and narrowed the distribution compared to $p(\mathbf{z}_n|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1})$ (which is shown in dashed in the right-hand plot for comparison).

$\hat{\alpha} - \gamma$ Recursion for Linear Gaussian State Space Models

- Note that $\gamma(z_t) = p(z_t|x_{1:T}) = \hat{\alpha}(z_t)\hat{\beta}(z_t) = \mathcal{N}(z_t|\hat{\mu}_t, \hat{V}_t)$.
- In the recursion for HMMs, simply replace summation by integration, i.e.

$$\begin{aligned} c_{t+1}\hat{\beta}(z_t) &= \int \hat{\beta}(z_{t+1})p(x_{t+1}|z_{t+1})p(z_{t+1}|z_t)dz_{t+1} \\ &= \int \mathcal{N}(z_{t+1}|\hat{\mu}_{t+1}, \hat{V}_{t+1})\mathcal{N}(x_{t+1}|Cz_{t+1}, \Sigma)\mathcal{N}(z_{t+1}|Az_t, \Gamma)dz_{t+1} \end{aligned}$$

- Multiply both sides with $\hat{\alpha}(z_t)$ and, then, use the closed forms expressions to obtain the joint Gaussian distribution from the marginal and conditional, i.e.

$$\begin{aligned} \hat{\mu}_t &= \mu_t + J_t(\hat{\mu}_{t+1} - A\mu_t) \\ \hat{V}_t &= V_t + J_t(\hat{V}_{t+1} - P_t)J_t^T \\ J_t &= V_t A^T (P_t)^{-1} \end{aligned}$$

- Note that the $\hat{\alpha}$ or forward phase must be completed before starting with the γ or backward phase, so that μ_t and V_t are available.
- Finally, $\xi(z_{t-1}, z_t)$ is normally distributed with mean vector $(\hat{\mu}_{t-1}, \hat{\mu}_t)^T$ and covariance matrix $J_{t-1}\hat{V}_t$.

EM Algorithm for Linear Gaussian State Space Models

- Define $\theta = \{A, \Gamma, C, \Sigma, \mu_0, P_0\}$. Then

$$E_{z_{1:T}|\theta^{old}}[\ln p(x_{1:T}, z_{1:T}|\theta)] = E_{z_{1:T}|\theta^{old}}[\ln p(z_1|\mu_0, P_0) + \sum_{t=2}^{T} \ln p(z_t|z_{t-1}, A, \Gamma) + \sum_{t=1}^{T} \ln p(x_t|z_t, C, \Sigma)]$$

- Maximization step over θ, i.e.

$$\mu_0^{new} = E[z_1]$$

$$P_0^{new} = E[z_1 z_1^T] - E[z_1]E[z_1^T]$$

$$A^{new} = \Big(\sum_{t=2}^{T} E[z_t z_{t-1}^T]\Big)\Big(\sum_{t=2}^{T} E[z_{t-1} z_{t-1}^T]\Big)^{-1}\Big)$$

$$\Gamma^{new} = \frac{1}{N-1}\sum_{t=2}^{T}\Big(E[z_t z_t^T] - A^{new}E[z_{t-1}z_t^T] - E[z_t z_{t-1}^T](A^{new})^T + A^{new}E[z_{t-1}z_{t-1}^T](A^{new})^T\Big)$$

$$C^{new} = \Big(\sum_{t=1}^{T} x_t E[z_t^T]\Big)\Big(\sum_{t=1}^{T} E[z_t z_t^T]\Big)^{-1}\Big)$$

$$\Sigma^{new} = \frac{1}{N}\sum_{t=1}^{T}\Big(x_t x_t^T - (C^{new})^T E[z_t]x_t^T - x_t E[z_t^T](C^{new})^T + C^{new}E[z_t z_t^T](C^{new})^T\Big)$$

- Expectation step with respect to θ^{old}, i.e.

$$E[z_t] = \hat{\mu}_t$$

$$E[z_t z_{t-1}^T] = \hat{V}_t J_{t-1}^T + \hat{\mu}_t \hat{\mu}_{t-1}^T$$

$$E[z_t z_t^T] = \hat{V}_t + \hat{\mu}_t \hat{\mu}_t^T$$

because

$$\begin{aligned} cov[xy] &= E[(x - E[x])(y - E[y])] = E[xy] - E[xE[y]] - E[E[x]y] + E[E[x]E[y]] \\ &= E[xy] - E[x]E[y] \end{aligned}$$

Particle Filter

- For non linear Gaussian SSMs.
- For any feature function f, we have that

$$E[f(z_t)] = \int f(z_t)p(z_t|x_{1:t})dz_t = \int f(z_t)p(z_t|x_t, x_{1:t-1})dz_t$$
$$= \int f(z_t)\frac{p(x_t|z_t)p(z_t|x_{1:t-1})}{\int p(x_t|z_t)p(z_t|x_{1:t-1})dz_t}dz_t \approx \sum_{m=1}^{M} w_t^m f(z_t^m)$$

  where $\{z_t^m\}$ is a set of samples (particles) from $p(z_t|x_{1:t-1})$ and

$$w_t^m = \frac{p(x_t|z_t^m)}{\sum_{m'=1}^{M} p(x_t|z_t^{m'})}$$

- Moreover

$$p(z_{t+1}|x_{1:t}) = \int p(z_{t+1}|z_t, x_{1:t})p(z_t|x_{1:t})dz_t = \int p(z_{t+1}|z_t)p(z_t|x_{1:t})dz_t$$
$$\int p(z_{t+1}|z_t)\frac{p(x_t|z_t)p(z_t|x_{1:t-1})}{\int p(x_t|z_t)p(z_t|x_{1:t-1})dz_t}dz_t \approx \sum_{m=1}^{M} w_t^m p(z_{t+1}|z_t^m)$$

- Particle filter: $p(z_t|x_{1:t})$ is represented by $\{z_t^m\}$ and $\{w_t^m\}$. These are used to draw a sample $\{z_{t+1}^m\}$ from $p(z_{t+1}|x_{1:t})$, by sampling first a component and then from the component. Finally, the weights $\{w_{t+1}^m\}$ are calculated.

Particle Filter

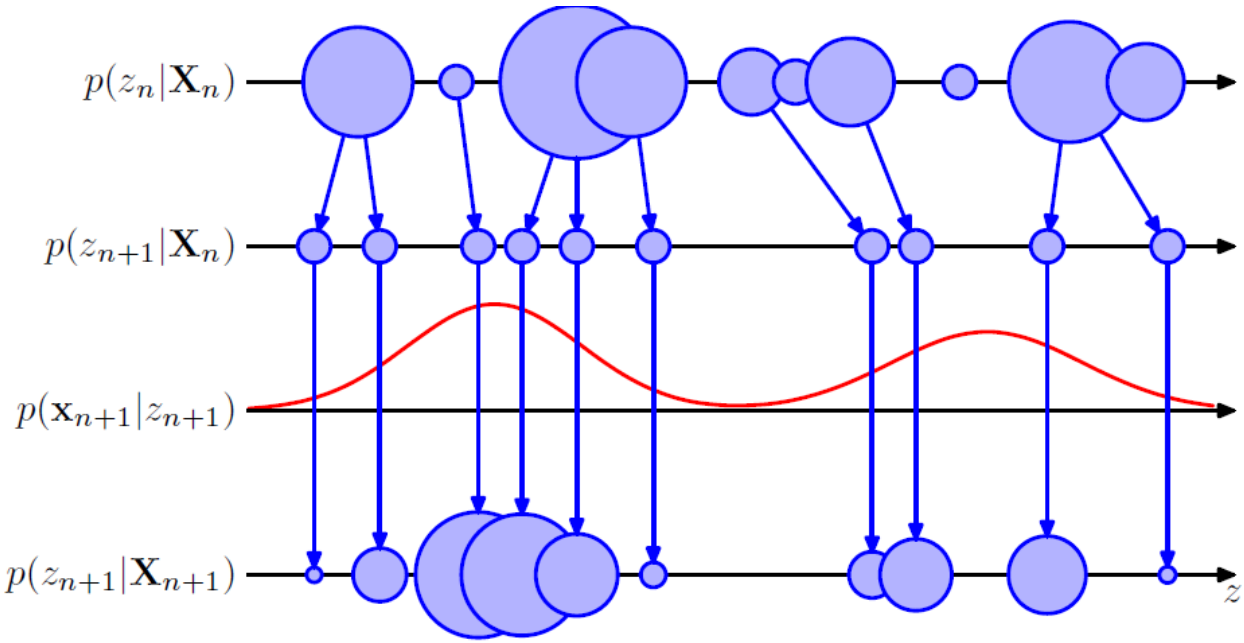
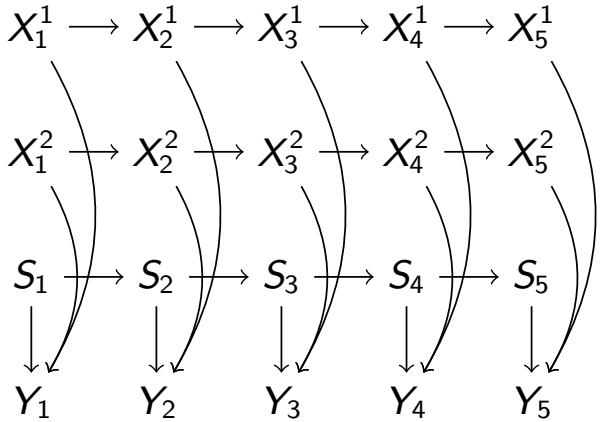


Figure 13.23 Schematic illustration of the operation of the particle filter for a one-dimensional latent space. At time step n, the posterior $p(z_n|\mathbf{x}_n)$ is represented as a mixture distribution, shown schematically as circles whose sizes are proportional to the weights $w_n^{(l)}$. A set of L samples is then drawn from this distribution and the new weights $w_{n+1}^{(l)}$ evaluated using $p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}^{(l)})$.

Switching State Space Models



- Switching SSMs = HMMs + SSMs, i.e. they model the discrete transition from one linear Gaussian regime to another.
- $s_t \in \{1, \ldots, N\}$
- $p(s_t|s_{t-1})$ = transition model for the switching variable
- $p(x_t^n|x_{t-1}^n) = \mathcal{N}(x_t^n|A^n x_{t-1}^n, Q)$ = n-th linear Gaussian regime
- $p(y_t|x_t^1, \ldots, x_t^N, s_t = n) = \mathcal{N}(y_t|C^n x_t^n, R)$ = n-th linear Gaussian regime
- Note that

$$p(\{s_t, x_t^1, \ldots, x_t^N, y_t\}) = p(s_1) \prod_{t=2}^{T} p(s_t|s_{t-1}) \cdot \prod_{n=1}^{N} [p(x_1^n) \prod_{t=2}^{T} p(x_t^n|x_{t-1}^n)] \prod_{t=1}^{T} p(y_t|x_t^1, \ldots, x_t^N, s_t)$$

Switching State Space Models

- Since $p(\{s_t, x_t^1, \ldots, x_t^N, y_t\})$ is not Gaussian, there is no closed form expressions for the E-step of the EM algorithm.
- Solution: Approximate EM algorithm.
  - E-step
    - Repeat the following steps until convergence.
    - Use the HMM to soft-assign observations to each SSM.
    - Use the prediction error of each SSM to determine the observation probabilities for the HMM.
  - M-step
    - Update the parameters of the HMM and each SSM as usual but weighing each observation accordingly.

Switching State Space Models

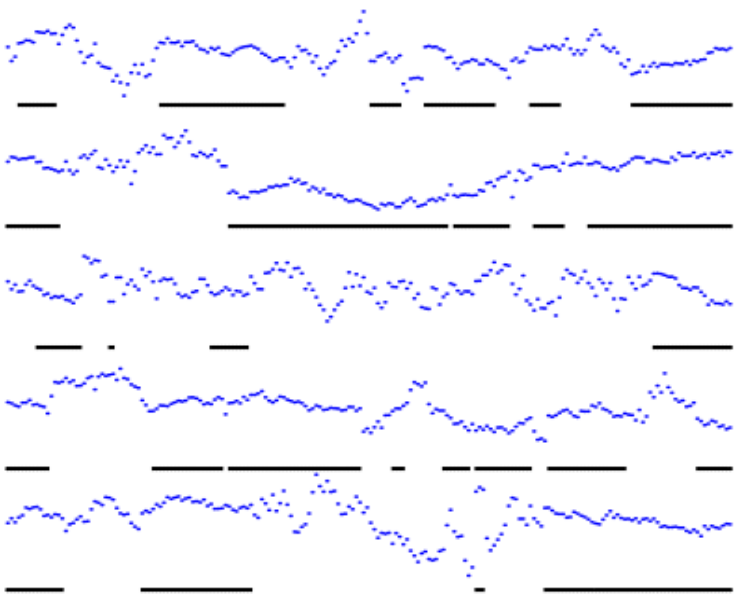


Figure 6: Five data sequences of length 200, with their true segmentations below them. In the segmentations, switch states 1 and 2 are represented with presence and absence of dots, respectively. Notice that it is difficult to segment the sequences correctly based only on knowing the dynamics of the two processes.

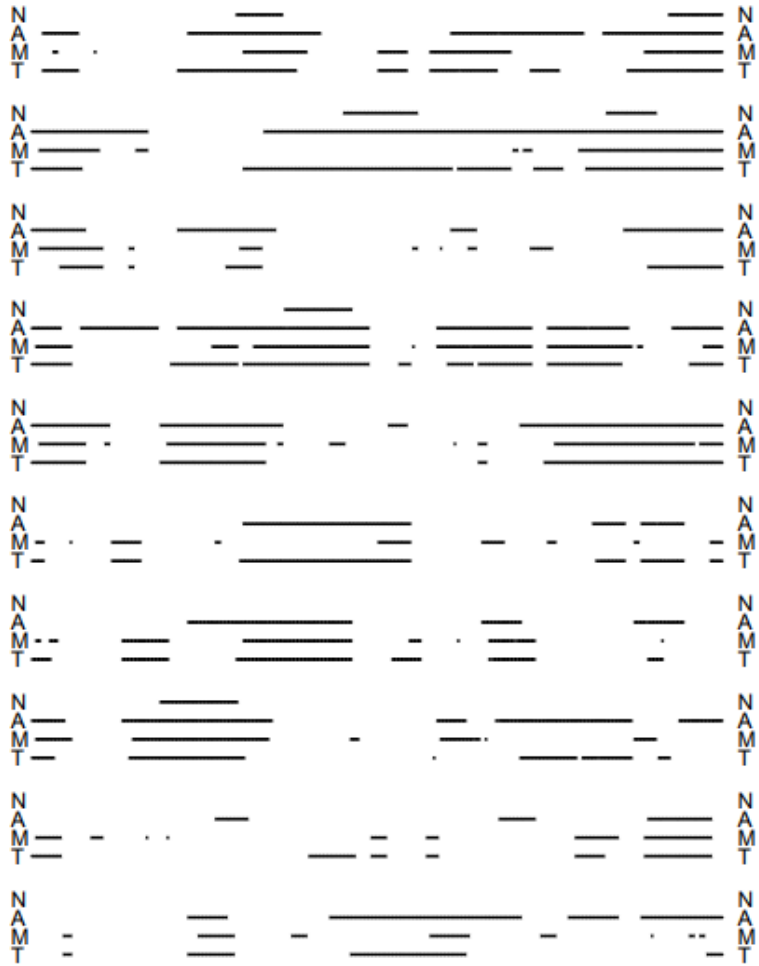


Figure 7: For 10 different sequences of length 200, segmentations are shown with presence and absence of dots corresponding to the two SSMs generating these data. The rows are the segmentations found using the variational method with no annealing (N), the variational method with deterministic annealing (A), the gaussian merging method (M), and the true segmentation (T). All three inference algorithms give real-valued $h_t^{(m)}$; hard segmentations were obtained by thresholding the final $h_t^{(m)}$ values at 0.5. The first five sequences are the ones shown in Figure 6.

Contents

- Scaling Factors for Hidden Markov Models
- $\hat{\alpha} - \gamma$ Recursion for Linear Gaussian State Space Models
- EM Algorithm for Linear Gaussian State Space Models
- Particle Filter
- Switching State Space Models

Thank you