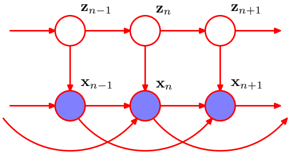
PRML
Chapter 13: Sequential Data

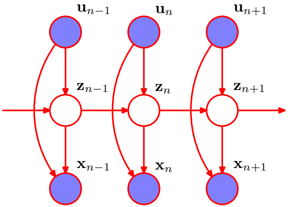
Moray Allan & Tingting Jiang

June 26, 2008

Extensions of HMM

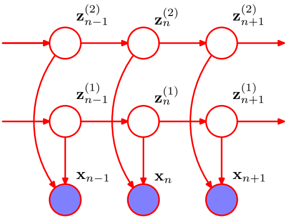
- HMM + discriminative technique (Kapadia, 1998)
  - Goal is sequence classification
  - Input: R observation sequences $\mathbf{X}_r$ with class labels
  - Output: parameters for each class which optimize the cross-entropy
- Model state duration time T at state k directly by $p(T|k)$ instead of an exponentially decaying function of T (Rabiner, 1989)
- Autoregressive HMM (Ephraim et al., 1989)
- Input-Output HMM (Bengio and Frasconi, 1995)

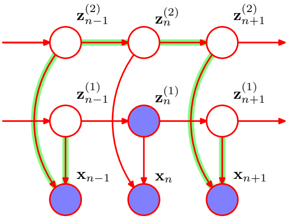




More Extensions of HMM

- Factorial HMM (Ghahramani and Jordan, 1997)





Linear Dynamical Systems

- Continuous latent variables and summations become integrals
- Consider a linear-Gaussian state space model
- Kalman Filter (Kalman, 1960)

$$p(\mathbf{z}_n|\mathbf{z}_{n-1}) = \mathcal{N}(\mathbf{z}_n|A\mathbf{z}_{n-1}, \Gamma) \tag{13.75}$$

$$p(\mathbf{x}_n|\mathbf{z}_n) = \mathcal{N}(\mathbf{x}_n|C\mathbf{z}_n, \Sigma) \tag{13.76}$$

Traditionally, it is given by

$$\mathbf{z}_n = A\mathbf{z}_{n-1} + \mathbf{w}_n \tag{13.78}$$

$$\mathbf{x}_n = C\mathbf{z}_n + \mathbf{v}_n \tag{13.79}$$

$$\mathbf{z}_1 = \boldsymbol{\mu}_0 + \boldsymbol{\mu} \tag{13.80}$$

where

$$\mathbf{w} \sim \mathcal{N}(\mathbf{w}|\mathbf{0}, \Gamma) \tag{13.81}$$

$$\mathbf{v} \sim \mathcal{N}(\mathbf{v}|\mathbf{0}, \Sigma) \tag{13.82}$$

$$\mathbf{u} \sim \mathcal{N}(\mathbf{u}|\mathbf{0}, V_0). \tag{13.83}$$

Prediction

- $p(\mathbf{z}_n|\mathbf{x}_1, \ldots, \mathbf{x}_n)$
- Joint distribution over all latent and observed variables is a Gaussian as the linear-Gaussian model
- So we can use the methods for multivariate Gaussian
- The inference process is similar to that of HMM, summations are replaced by integrations.

$$\begin{aligned}
\boldsymbol{\mu}_n &= \mathbf{A}\boldsymbol{\mu}_{n-1} + \mathbf{K}_n(\mathbf{x}_n - \mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1}) && (13.89)\\
\mathbf{V}_n &= (\mathbf{I} - \mathbf{K}_n\mathbf{C})\mathbf{P}_{n-1} && (13.90)\\
c_n &= \mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1}, \mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}} + \boldsymbol{\Sigma}). && (13.91)
\end{aligned}$$

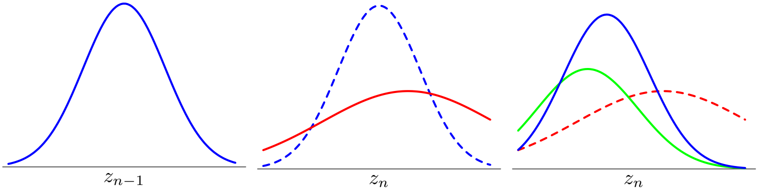
- Kalman gain matrix

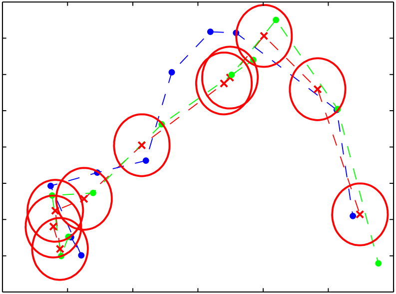
$$\mathbf{K}_n = \mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}}\left(\mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1} \tag{13.92}$$

where

$$\mathbf{P}_{n-1} = \mathbf{A}\mathbf{V}_{n-1}\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Gamma}. \tag{13.88}$$

Interpretation of Kalman Filter





Backward Inference

- $p(\mathbf{z}_n|\mathbf{x}_1,\ldots,\mathbf{x}_N)$
- Backward recursion

$$\hat{\boldsymbol{\mu}}_n = \boldsymbol{\mu}_n + J_n(\hat{\boldsymbol{\mu}}_{n+1} - A\boldsymbol{\mu}_n) \tag{13.100}$$

$$\hat{V}_n = V_n + J_n(\hat{V}_{n+1} - P_n)J_n^T \tag{13.101}$$

where $J_n = V_n A^T (P_n)^{-1}$.

- First forward pass generates $\boldsymbol{\mu}_n$ and V_n, and then backward pass.

Learning in LDS

- Model parameters $\theta = \{A, \Gamma, C, \Sigma, \boldsymbol{\mu}_0, V_0\}$
- Maximum likelihood method and EM algorithm
- E step: given θ^{old}, run the inference algorithm to determine $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$
- M step: given $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$, maximize the complete-data log likelihood function w.r.t. θ

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \mathbb{E}_{\mathbf{Z}|\boldsymbol{\theta}^{\text{old}}} \left[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})\right]. \tag{13.109}$$

$$\begin{aligned} \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) &= \ln p(\mathbf{z}_1|\boldsymbol{\mu}_0, \mathbf{V}_0) + \sum_{n=2}^{N} \ln p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}, \boldsymbol{\Gamma}) \\ &\quad + \sum_{n=1}^{N} \ln p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{C}, \boldsymbol{\Sigma}) \end{aligned} \tag{13.108}$$

Extension of LDS

- Limitation of Kalman Filter: the linear-Gaussian model assumption implies that the marginal distribution of the observed variables is a Gaussian
- Gaussian mixture: K components lead to a mixture of K Gaussians
- Extended Kalman Filter(EKF)(Zarchan and Musoff, 2005): the state transition and observation models are not linear-Gaussian, but can be approximated
- Switching State Space Model(Ghahramani and Hinton, 1998): multiple Markov chains of continuous linear-Gaussian latent variables together with a discrete Markov chain (HMM) whose output determines which continuous Markov chain to use.
- Switching Hidden Markov Model: multiple discrete Markov chains and use one of them as the switch

Particle Filter

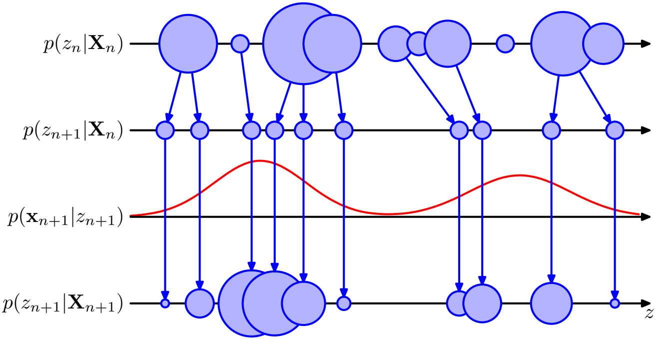
- For systems which are not linear-Gaussian, use Sampling-importance-resampling method (Sec. 11.1.5) to obtain a sequential Monte Carlo (SMC) algorithm
- Also known as
  - *Bootstrap filter* (Gordon et al., 1993)
  - *Survival of the fittest* (Kanazawa et al., 1995)
  - the *Condensention algorithm* (Isard and Blake, 1998)
- A continuous probability distribution is approximated by a set of discrete samples $\{\mathbf{z}_n^{(l)}\}$ with weights

$$w_n^{(l)} = \frac{p(\mathbf{x}_n|\mathbf{z}_n^{(l)})}{\sum_{m=1}^{L} p(\mathbf{x}_n|\mathbf{z}_n^{(m)})} \tag{13.118}$$

- How to choose the importance function $q(\mathbf{z})$?
  - Popular choice: $p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}^{(l)})$ as it is simple to evaluate
  - Optimal choice (A. Doucet et al., 2000): $p(\mathbf{z}_{n+1}|\mathbf{z}_n, \mathbf{x}_n)$ which is hard to compute

Particle Filter Operation

- Propagate: draw random samples from the approximated distribution $p(\mathbf{z}_{n+1}|\mathbf{X}_n)$
- Update: use the new observation $\mathbf{x}_{n+1}$ to evaluate the weight $w_{n+1}^{(l)} \propto p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}^{(l)})$



- Resample: it is necessary only when some samples degenerate

Questions?