Probabilistic Graphical Models for Image Analysis - Lecture 6

Stefan Bauer

26 October 2018

Max Planck ETH Center for Learning Systems

Overview

1. Recall HMM
2. HMM Extensions
3. Inference
4. Factor Analysis

Conference on Robot Learning

Conference on Robot Learning (CoRL) - 2018 Edition

The Conference on Robot Learning (CoRL) is a new annual international conference focusing on the intersection of robotics and machine learning. The first meeting (CoRL 2017) was held in Mountain View, California on November 13 - 15, 2017, and brought together about 350 of the best researchers working on robotics and machine learning.

CoRL 2018 will be held on October 29th-31st, 2018, in Zürich, Switzerland.

Announcement Guest Lecture: Olivier Bachem - Generative Adversarial Networks, 9th of November*

*Image from https://www.christies.com/Features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.

Recall HMM

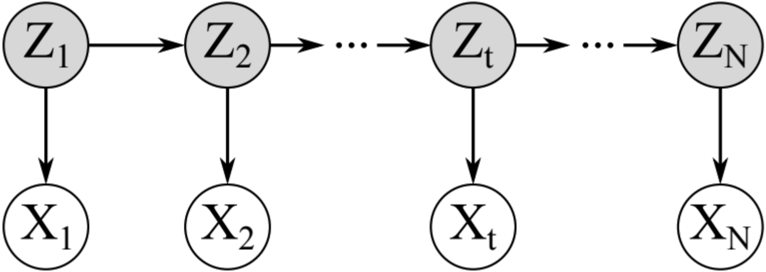
Dynamic Systems - State Space Models

$$\dot{x}(t) = f(x(t), u(t)), \quad \text{state evolution}$$
$$y(t) = g(x(t), u(t)), \quad \text{observations}$$

Most often used in practice are linear, discrete Systems

$$x(t+1) = Ax(t) + Bu(t)$$
$$y(t) = Cx(t) + Du(t)$$

Inference tasks

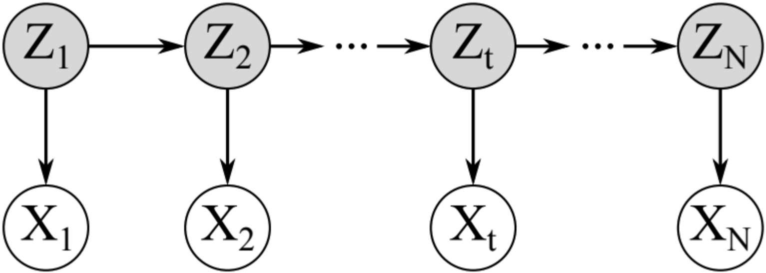


Filtering $P(Z_t|X_{1:t})$

Prediction $P(Z_{t+\tau}|X_{1:t})$

Smoothing $P(Z_t|X_{1:t})$ for $1 \leq t \leq T$

HMM and Kalman Filter



HMM: Z_i Multinomial, X_i arbitrary

Kalman: Z_i, X_i Gaussian

Extended Kalman: Z_i Gaussian, X_i arbitrary

Recall EM Algorithm (Lecture 2)

Need to maximize

$$\log p(\mathcal{D}) = \sum_{x \in D} \log p(x) = \sum_{x \in D} \log \left(\sum_{z} p(x|z) p(z) \right)$$

Problem: Only x is observed but we have parameters θ and latent variables z

The Expectation Maximization (EM) algorithm:

- **Expectation**: Assign values to hidden/missing variables i.e. compute $p(z|x;\theta_t)$
- **Maximization**: Maximize parameter log likelihood via $\theta_{t+1} = \arg\max_\theta \sum_{x \in D} \mathbb{E}_{z \sim p(z|x,\theta_t)} \log p(x, z, \theta)$
- Repeat until convergence for $t = 1, 2, \cdots$, starting with θ_0

EM using Jensen

Y observations, X latent states, θ parameters.

$$\begin{aligned}
\log P(Y|\theta) &= \log \sum_{\mathcal{X}} P(Y, X|\theta) \\
&= \log \sum_{\mathcal{X}} p(X, Y|\theta) \frac{q(X)}{q(X)} \\
&\geq \sum_{\mathcal{X}} q(X) \log \frac{p(X, Y|\theta)}{q(X)} \\
&= \sum_{\mathcal{X}} q(X) \log p(X, Y|\theta) - \sum_{\mathcal{X}} q(X) \log q(X) \\
&= \mathcal{L}(q, \theta)
\end{aligned}$$

Learning HMMs using the EM algorithm

$$\log P(X_{1:T}, Y_{1:t}) = \log P(X_1) + \sum_{t=1}^{T} \log P(Y_t|X_t) + \sum_{t=2}^{T} \log P(X_t|X_{t-1})$$

Hidden Markov Model i.e. X_t categorical (with K values). Thus we can represent X_t as a K dimensional unit vector e.g. for taking on the second value:

$$X_t = [010 \cdots 0]^\intercal$$

The transition probability can then be written as:

$$P(X_t|X_{t-1}) = \prod_{i=1}^{K} \prod_{j=1}^{K} A_{ij}^{X_{t,i}, X_{t-1,j}}$$

where A_{ij} is the transition matrix, with non-negative entries and each row sums to 1.

State transition models

$$\log P(X_t|X_{t-1}) = \sum_{i=1}^{K}\sum_{j=1}^{K} X_{t,i}X_{t-1,j}\log A_{ij} = X_t^\intercal(\log A)X_{t-1}$$

Similarly if initial state probabilities are arranged in a vector π, of dimension $K \times 1$ with $\pi_i = P(X_{1i=1})$, then

$$P(X_1|\pi) = \prod_{i=1}^{K} \pi_i^{X_{1i}}$$

and

$$\log P(X_1) = X_1^\intercal \log \pi$$

Observation model

If Y_t is discrete and can take on D values, we can again write

$$\log P(Y_t|X_t) = Y_t^\intercal (\log B) X_t$$

where B is a $D \times K$ dimensional emission probability matrix.

The final parameter set of the model is then

$$\theta = (A, B, \pi)$$

Goal: $\arg\max_\theta \log P(Y)$

Expectation Maximization for HMM

M-Step

$$A_{ij} \propto \sum_{t=2}^{T} \mathbb{E}[X_{t,i} X_{t-1,j}] \leftarrow \frac{\sum_{t=2}^{T} \mathbb{E}[X_{t,i} X_{t-1,j}]}{\sum_{t=2}^{T} \mathbb{E}[X_{t-1,j}]} \tag{1}$$

$$\pi \leftarrow \mathbb{E}[X_{1,i}] \tag{2}$$

$$B_{di} \leftarrow \frac{\sum_{t=1}^{T} Y_{t,d} \mathbb{E}[X_{t,i}]}{\sum_{t=1}^{T} \mathbb{E}[X_{t,i}]} \tag{3}$$

E-Step Calculate Expectations using forward-backward algorithm.

$$\mathbb{E}[X_{t,i}] = \gamma_{ti} = \frac{\alpha_{t,i} \beta_{t,i}}{\sum_j \alpha_{t,j} \beta_{t,j}}$$

$$\mathbb{E}[X_{t,i} X_{t-1,j}] = \zeta_{tij} = \frac{\alpha_{t-1,j} A_{ij} P(Y_t | X_{t,i}) \beta_{t,i}}{\sum_{k,l} \alpha_{t-1,k} A_{kl} P(Y_t | X_{t,l}) \beta_{t,l}}$$

Linear Gaussian State Space Models

Assumption: Initial states are Gaussian distributed:

$$x_1 \sim \mathcal{N}(\mu_1, Q_1)$$

With linear dynamics all future states x_t and observations will be Gaussian distributed:

$$P(x_{t+1}|x_t) = \mathcal{N}(Ax_t, Q)$$
$$P(y_t|x_t) = \mathcal{N}(Cx_t, R)$$

With Markov property it follows:

$$P(X_{1:T}, Y_{1:T}) = P(x_1) \prod_{t=2}^{T} P(x_t|x_{t-1}) \prod_{t=1}^{T} P(y_t|x_t)$$

Linear Gaussian State Space Models II

From before

$$P(X_{1:T}, Y_{1:T}) = P(x_1) \prod_{t=2}^{T} P(x_t | x_{t-1}) \prod_{t=1}^{T} P(y_t | x_t)$$

Each of the above densities is Gaussian, thus:

$$\begin{aligned} -2 \log P(X_{1:T}, Y_{1:T}) = & \sum_{t=1}^{T} [(y_t - Cx_t)^\intercal R^{-1} (y_t - Cx_t) + \log |R|] \\ & + \sum_{t=1}^{T-1} [(x_{t+1} - Ax_t)^\intercal Q^{-1} (x_{t+1} - Ax_t) + log|Q|] \\ & + (x_1 - \mu_1)^\intercal Q_1^{-1} (x_1 - \mu_1) + \text{const.} \end{aligned}$$

Method: Again EM, M-Step e.g. $C \leftarrow (\sum_t y_t x_t^\intercal)(\sum_t x_t x_t^\intercal)^{-1}$

Problem x is hidden <- use expectations! (kalman smoother)

HMM Extensions

Recall from last time

Problems for LDS and HMM:

- state dynamics can be non-linear
- relations between observed and latent states can be non-linear
- noise can be non-Gaussian
- HMM are dynamic extensions to Mixture Models -> in theory (with enough components) they can model any distribution.
- However HMMs are inefficient wrt. number of required states and a high number of states might result in severe over-fitting!

Factorial HMM

Generalize HMM by representing state as collection of discrete state variables

$$X_t = X_t^{(1)}, \cdots, X_t^{(m)}, \cdots, X_t^{(M)}$$

each can take $K^{(m)}$ values. Assume $K^{(m)} = K$ for simplicity for all m.

Then (from before) the transition matrix would be of size $K^M \times K^M$!

Problem:

- equivalent to HMM with K^M states
- time and sample complexity of estimation are exponential in M.
- unlikely to discover interesting structure since all variables can arbitrarily interact.

Factorial HMM

Idea: Constrain underlying state transitions - each state variable evolves according to its own dynamics and is a prior uncoupled from the other states:

$$P(X_t|X_{t-1}) = \prod_{m=1}^{M} P(X_t^{(m)}|X_{t-1}^{(m)})$$

Motivation for FHMM:

- transition structure can now be described using M distinct $K \times K$ matrices
- richer modeling tool
- inclusion of prior structural information about state variables underlying the dynamics of the system generating the data.

Factorial HMM

Observation at time t can depend on all states at that time step!

Idea Assume linear Gaussian dependence!

$$P(Y_t|X_t) = |R|^{-\frac{1}{2}}(2\pi)^{\frac{-D}{2}} \exp(-\frac{1}{2}(Y_t - \mu_t)^\intercal R^{-1}(Y_t - \mu_t))$$

where $\mu_t = \sum_{m=1}^{M} W^{(m)} X_t^{(m)}$

- $W^{(m)}$ is a $D \times K$ matrix, where columns are contributions to the means for each setting of $X_t^{(m)}$
- R is a $D \times D$ covariance matrix

Interpretation: GMM with K^M mixture components, each having constant covariance matrix R and underlying markov dynamics.

Switching State Space Models

Recall (last week): HMM discrete latent variables, state space model (continuous).

idea: Model time series with continuous but nonlinear dynamics by combining HMM and SSM!

Switching State Space Models

- Y_t is modelled using latent space comprising M real valued state vectors $X_t^{(M)}$ and one discrete state S_t
- S_t is discrete and can take on Mvalues, so called *Switch*.

$$P(S, X^{(1)}, \cdots, X^{(M)}, Y) = P(S_1) \prod_{t=2}^{T} P(S_t|S_{t-1}) \prod_{m=1}^{M} [P(X_1^{(m)}) \prod_{t=2}^{T} P(X_t^m|X_{t-1}^{(m)})]$$
$$\times \prod_{t=1}^{T} P(Y_t|S, X^{(1)}, \cdots, X^{(M)})$$

Switching State Space Models

$$P(S, X^{(1)}, \cdots, X^{(M)}, Y) = P(S_1) \prod_{t=2}^{T} P(S_t|S_{t-1}) \prod_{m=1}^{M} [P(X_1^{(m)}) \prod_{t=2}^{T} P(X_t^m|X_{t-1}^{(m)})]$$
$$\times \prod_{t=1}^{T} P(Y_t|S, X^{(1)}, \cdots, X^{(M)})$$

Conditioned on the switch state i.e. $S_t = m$, the observable is a multivariate Gaussian with output equation given by state space model m.

$$P(Y|X^{(1)}, \cdots, X^{(M)}, S = m) = |R|^{-\frac{1}{2}} (2\pi)^{\frac{-D}{2}} \exp\left[-\frac{1}{2}\left(Y_t - C^{(m)} x_t^{(m)}\right)^{\intercal} R^{-1} \left(Y_t - C^{(m)} x_t^{(m)}\right)\right]$$

where

- D is the dimension of the observation vector
- R is the observation noise covariance matrix
- C^m is the output matrix for state space model m $(Y_t = CX_t + \text{noise})$

Example - Still highly used

Article | OPEN | Published: 27 June 2018

Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition

Jalil Taghia, Weidong Cai, Srikanth Ryali, John Kochalka, Jonathan Nicholas, Tianwen Chen & Vinod Menon

Nature Communications **9**, Article number: 2505 (2018) | Download Citation

Example - Still highly used*

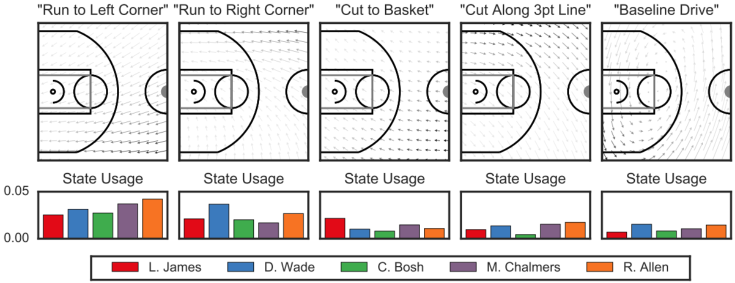


Figure 5: Exploratory analysis of NBA player trajectories from the Nov. 1, 2013 game between the Miami Heat and the Brooklyn Nets. **(Top)** When applied to trajectories of five Heat players, the recurrent AR-HMM (ro) discovers $K = 30$ discrete states with linear dynamics; five hand-picked states are shown here along with our names. Speed of motion is proportional to length of arrow. Location-dependent state probability is proportional to opacity of the arrow. **(Bottom)** The probability with which each player uses the corresponding state under the posterior.

*figure from: Lindermann et.al. Recurrent Switching Linear Dynamical Systems, tech report 2016

Inference

Mean-filed for factorial HMM

Variational approximation:

$$Q(X|\varphi) = \prod_{t=1}^{T} \prod_{m=1}^{M} Q(X_t^{(m)}|\varphi_t^{(m)})$$

where $\varphi = \{\varphi_t^{(m)}\}$ are the variational parameters and the means of the state variables $X_t^{(m)}$, which is represented as a K-dimensional vector.

Assuming independence we can thus write:

$$Q(X_t^{(m)}|\varphi_t^{(m)}) = \prod_{k=1}^{K} \left(\varphi_{t,k}^{(m)}\right)^{X_{t,k}^{(m)}}$$

where $X_{t,k}^{(m)} \in \{0, 1\}$ and $\sum_{k=1}^{K} X_{t,k}^{(m)} = 1$.

Mean-filed for factorial HMM

Exercise[*]

Update for variational parameter

$$\varphi_t^{(m),new} = \mathrm{softmax}\left(W^{m\prime} R^{-1} \widetilde{Y}_t^{(m)} - \frac{1}{2}\Delta^{(m)} + (\log \varphi^{(m)})\varphi_{t-1}^{(m)} + (\log \varphi^{(m)})^\intercal \varphi_{t+1}^{(m)} \right)$$

where

- $\widetilde{Y}_t^{(m)} = Y_t - \sum_{l \neq m} W^{(l)} \varphi_t^{(l)}$
- $\Delta^{(m)}$ is the vector of diagonal elements of $W^{(m)\prime} R^{-1} W^{(m)}$
- $\log \varphi^{(m)}$ denotes the elementwise logarithm of the transition matrix $\varphi^{(m)}$

[*]Solution in: Ghahramani and Jordan, Factorial Hidden Markov Models, NIPS 1996

Mean-filed for factorial HMM

Intuition

- Given one particular observation sequence, the hidden state variables for the M Markov chains at time step t are stochastically coupled.
- Stochastic coupling is approximated by a system in which hidden variables are uncorrelated but have coupled means.
- The mean-field approximation solves for the deterministic coupling of the means that best approximate the stochastically coupled system.

Extension: Recall (Structured mean-field) and Bayesian HMMs.

Algorithm Switching State Space[*]

Algorithm 1 Learning Switching State Space Models

Input:
Initialisations of all parameters

Until Convergence:
E-Step
Compute $q_t^{(m)}$ for state space model m
Compute $h_t^{(m)}$ using forward-backward algorithm on HMM with observations prob. $q_t^{(m)}$
Run Kalman smoothing for each state.
M-Step
Re-estimate parameters for each state space model using the data weighted by $h_t^{(m)}$
Re-estimate parameters for the switching process using forward backward algorithm.

[*] Natural Approximation: make M-state space models and switch variable independent-> Can use tractable inference from last week / exercise.

Factor Analysis

Example - Recall GMM

- Assume data $x^{(i)} \in \mathbb{R}^n$ that comes from several Gaussians.
- So far: Assumed training set size m is larger than n. We then used
- Here we used the EM-algorithm for inference.

Question: What can we do if $n >> m$?

Motivation

- Often there are some unknown *underlying causes* of the data.
- Continuous factors which control the data we observe *data manifold* (or subspace).
- Training continuous latent variable models is often called *dimensionality reduction*, since there are typically many fewer latent dimensions.
- Examples (see reference) PCA, Factor Analysis, ICA
- Reason for choosing continuous representation is often motivated by efficiency.
- Mixture models uses discrete class variable: clustering
- Simplest case: *linear* subspace and underlying latent variable with a Gaussian distribution.

Factor Analysis Model

$z \in \mathbb{R}^k$ is a latent variable and y is the observed data:

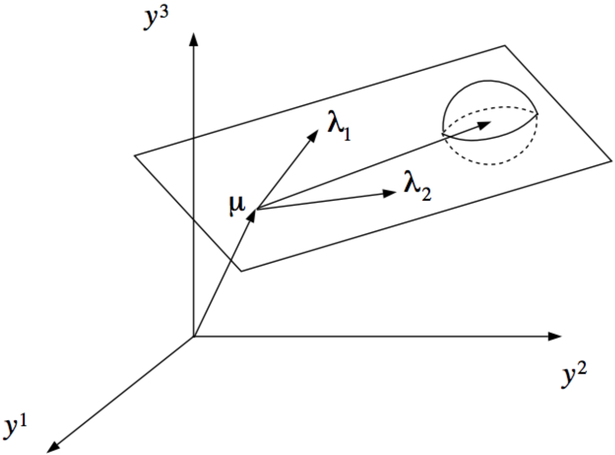
$$z \sim \mathcal{N}(0, 1)$$
$$x|z \sim \mathcal{N}(\mu + \Lambda z, \Psi)$$

Parameters of our model are thus:

- $\mu \in \mathbb{R}^n$
- $\Lambda \in \mathbb{R}^{n \times k}$
- Diagonal matrix $\Psi \in \mathbb{R}^{n \times n}$

Note: Dimensionality reduction since k is chosen smaller than n.

Illustration



where y are the observations.

Equivalent formulation

$$z \sim \mathcal{N}(0, 1)$$
$$\varepsilon \sim \mathcal{N}(0, \Psi)$$
$$x = \mu + \Lambda z + \varepsilon$$

where ε and z are independent.

Joint model:

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}(\mu_{zx}, \Sigma)$$

Goal: Identify μ_{zx} and Σ.

Factor Analysis

Joint model:

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} 1 & \Lambda^\intercal \\ \Lambda & \Lambda\Lambda^\intercal + \Psi \end{bmatrix} \right)$$

Marginal Distribution

$$x \sim \mathcal{N}(\mu, \Lambda\Lambda^\intercal + \Psi)$$

Log-Likelihood of parameters

$$l(\mu, \Lambda, \Psi) = \log \prod_{i=1}^{m} \frac{1}{(2\pi)^{\frac{n}{2}} |\Lambda\Lambda^\intercal + \Psi|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu)^\intercal (\Lambda\Lambda^\intercal + \Psi)^{-1} (x^{(i)} - \mu) \right)$$

EM for Factor Analysis

Maximum likelihood learning using EM:

- **E-Step**: $q^{t+1} = p(z|x, \theta^t)$
- **M-Step**: $\theta^{t+1} = \arg\max_\theta \sum_n \int_z q^{t+1}(z|x) \log p(x, z|\theta) dz$

where $\theta = (\mu, \Lambda, \Psi)$. Results for both steps:

- **E-Step**
  $q^{t+1} = p(z|x, \theta^t) = \mathcal{N}(z|m^{(i)}, V^{(i)})$ where
  $V^{(i)} = (1 + \Lambda^\intercal \Psi^{-1} \Lambda)^{-1}$ and $m^{(i)} = V^{(i)} \Lambda^\intercal \Psi^{-1}(x - \mu)$.
- **M-Step** $\Lambda^{t+1} = \left(\sum_i x^{(i)} m^{(i)\intercal} \right) \left(\sum_i V^{(i)} \right)^{-1}$
  $\Psi^{t+1} = \frac{1}{n} \mathrm{diag} \left[\sum_i x^{(i)} x^{(i)\intercal} + \Lambda^{t+1} \sum_i m^{(i)} x^{(i)\intercal} \right]$

Connection with State Space Models

State space models are dynamical generalizations of FA model.

$$x_t = Ax_{t-1} + Gw_t$$

whee $w_t = \mathcal{N}(0, Q)$

- Linear combinations of Gaussians is Gaussian i.e. added white noise w_t does not affect linearity.
- at each point in time t, we use a FA model to represent the output
- C is the *loading* matrix, shared across all (x_t, y_t) pairs.
- We assume all data points lie in the same low-dimensional space.

Summary - so far

- Factor analysis implies latent variable is assumed to lie on low-dimensional linear subspace
- Similar to mixture model, now just continuous
- Dimensionality reduction technique

- State space models (last week) are chain of Factor analysis models
- latent variables are connected sequentially in chains.
- HMM as dynamic generalization of mixture model
- Linear state space models are dynamic generalization of Factor analysis models.

Next week

Plan for next week:

- Continue with Dimensionality Reduction i.e. unifying framework and PCA, ICA
- Summary Linear State Space Models
- Missing: Recurrency i.e. non-Markovian dynamics (November)

Questions?