

L13 – Continuous factored models

Barber 21.1–4; AIMA 13.2.3, 14.4

What you should know after this lecture

- Multivariate Gaussians!
- Continuous and hybrid Bayes nets
- Idea of Gaussian belief propagation
- Kalman filtering

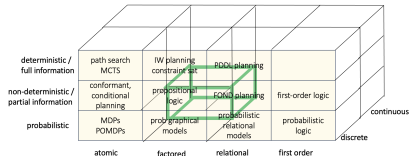
Probabilistic reasoning about partially-specified world states

Factored states

Continuous-valued factors

Exact inference

Temporal models



Conjugate families of probability distributions

In order for exact probabilistic inference to be tractable, we generally need for the joint and conditional distributions of factors to be conjugate:¹

- Let $f(\theta_A)(a)$ be the pdf of a random variable A and $f(\theta_B)(b)$ be the pdf of a random variable B , where f has some fixed parametric form and θ specifies a particular pdf in that family.
- Then the product of the pdfs on A and B has the form $f(\theta_{AB})(a, b)$ where θ_{AB} is a function of θ_A and θ_B .

$$f(\theta_A)(a) \cdot f(\theta_B)(b) = f(\theta_{AB})(a, b)$$

¹The actual definition is more general and specifically relates a prior distribution and an observation distribution, but this basic idea is what we need for now.

Categorical distribution is conjugate family

We have been using the categorical distribution²

- $\Omega = \{x_1, \dots, x_M\}$

- $\theta^A = (\theta_1^A, \dots, \theta_M^A)$

- $f_A(\theta^A)(x_i) = \theta_i^A$

$$\theta^B = (\theta_1^B, \dots, \theta_M^B)$$

$$f_B(\theta^B)(x_i) = \theta_i^B$$

If we multiply these distributions on the same variable (e.g. during message passing), then we get

- $f_{AB}(\theta_{AB})(x_i) = \theta_i^{AB} = \frac{1}{Z} \theta_i^A \cdot \theta_i^B$

where $Z = \sum_{i=1}^M \theta_i^A \theta_i^B$

²We like the name “multinoulli” better, though!

Categorical distribution is conjugate for joint

Combining two categorical distributions on different variables:

- $\Omega_A = \{a_1, \dots, a_M\}$
 - $\theta^A = (\theta_1^A, \dots, \theta_M^A)$
 - $f_A(\theta^A)(a_i) = \theta_i^A$
- $\Omega_B = \{b_1, \dots, b_N\}$
 - $\theta^B = (\theta_1^B, \dots, \theta_N^B)$
 - $f_B(\theta^B)(b_i) = \theta_i^B$

If we multiply these distributions on different variables (e.g. computing the joint when A and B are independent), then we get

- $\Omega_{AB} = \Omega_A \times \Omega_B$
- $f_{AB}(\theta^{AB})(a_i, b_j) = \theta^{AB}(a_i, b_j) = \theta_i^A \cdot \theta_j^B$

Univariate Gaussian is conjugate family

- $\Omega = \mathbb{R}$
- $\theta_A = (\mu_A, \sigma_A^2)$ $\theta_B = (\mu_B, \sigma_B^2)$
- $f_A(\theta_A)(x) = \frac{1}{\sqrt{2\pi}\sigma_A} \exp\left\{-\frac{1}{2\sigma_A^2}(x - \mu_A)^2\right\}$
- $f_B(\theta_B)(x) = \frac{1}{\sqrt{2\pi}\sigma_B} \exp\left\{-\frac{1}{2\sigma_B^2}(x - \mu_B)^2\right\}$

If we multiply these distributions on the same variable (e.g. during Bayes rule), then

- Observe that multiplying f 's yields

$$f_{AB}(\theta_{AB})(x) = \frac{1}{\sqrt{2\pi}\sigma_A} \frac{1}{\sqrt{2\pi}\sigma_B} \exp\left\{-\frac{1}{2\sigma_A^2}(x - \mu_A)^2 - \frac{1}{2\sigma_B^2}(x - \mu_B)^2\right\}$$

- After completing the square and some algebra, we find that $f_{AB}(\theta_{AB})(x) = \frac{1}{\sqrt{2\pi}\sigma_{AB}} \exp\left\{-\frac{1}{2\sigma_{AB}^2}(x - \mu_{AB})^2\right\}$ where

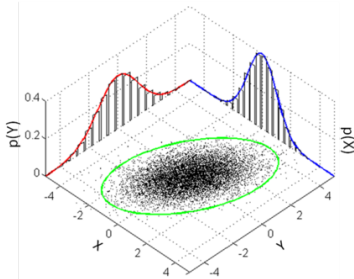
$$\mu_{AB} = \frac{\mu_A \sigma_B^2 + \mu_B \sigma_A^2}{\sigma_A^2 + \sigma_B^2} \quad \sigma_{AB}^2 = \frac{\sigma_A^2 \sigma_B^2}{\sigma_A^2 + \sigma_B^2}$$

Multivariate Gaussian

- $\Omega = \mathbb{R}^D$
- $\theta = (\mu \in \mathbb{R}^D, \Sigma \in \mathbb{R}^{D \times D})$ // Σ is positive definite

$$f(\mu, \Sigma)(x) = \frac{1}{\sqrt{2\pi^D |\Sigma|}} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right\}$$

$|\Sigma|$ is the determinant; figure from Wikipedia



- Axes are eigenvectors of Σ
- Axis-aligned if Σ is diagonal
- Round if Σ is identity

Fun facts about the multivariate Gaussian

Let's say our MVG has dimensions $1..D$, but we are interested in marginalizing some of them out, or conditioning some of them on particular values. Let's divide them into one set of dimensions $A = 1..K$ and another $B = K + 1..D$. So, we can think of the parameters as

$$\mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}$$

Marginalizing out dimensions A yields Gaussian on B with

$$\mu_B^m = \mu_B \quad \Sigma_B^m = \Sigma_B$$

Conditioning on $B = b$ yields a Gaussian on A with

$$\mu_{A|B}^c = \mu_A + \Sigma_{AB} \Sigma_{BB}^{-1} (b - \mu_B) \quad \Sigma_{A|B}^c = \Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA}$$

Multivariate Gaussian is conjugate family

Product of MVGs:

- $\Omega_A = \mathbb{R}^D$ $\Omega_B = \mathbb{R}^D$
- $\theta_A = (\mu_A, \Sigma_A)$ $\theta_B = (\mu_B, \Sigma_B)$

If we multiply these distributions on the same variable (e.g. during Bayes rule), then we get an MVG with

$$\mu_{AB} = (\Sigma_A^{-1} + \Sigma_B^{-1})^{-1} (\Sigma_A^{-1} \mu_A + \Sigma_B^{-1} \mu_B) \quad \Sigma_{AB} = (\Sigma_A^{-1} + \Sigma_B^{-1})^{-1}$$

Multivariate Gaussian is conjugate for joint

Product of MVGs on different domains

- $\Omega_A = \mathbb{R}^{D_A}$ $\Omega_B = \mathbb{R}^{D_B}$
- $\theta_A = (\mu_A, \Sigma_A)$ $\theta_B = (\mu_B, \Sigma_B)$

We get an MVG with dimension $D = D_A + D_B$, and

$$\mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_A & 0 \\ 0 & \Sigma_B \end{pmatrix}$$

Gaussian Bayesian networks

Assume the conditional probability distribution for each node V has the form $V \sim \text{Normal}(w_V^0 + w_V^T \cdot \text{pa}(V), \eta_V^2)$ where

- w_V is a vector of real-valued weights of length number of parents of V
- η_V^2 is the variance of added noise

Then the joint distribution over all variables V_1, \dots, V_N is Gaussian. You can construct the distribution by induction:

- Assume the parents of node V are normally distributed with mean μ_P, Σ_P
- Then $\mu_V = w_V^0 + W_V^T \mu_P$
- $\sigma_V^2 = \eta_V^2 + w^T \Sigma_P w$
- Also have to compute rest of entries in Σ for the joint

Hybrid networks

Some standard cases:

- Discrete parent of Gaussian nodes: mixture-of-Gaussians models
- Continuous parent of discrete node: apply sigmoid or softmax to get categorical distribution

Gaussian Factor graphs

Make a factor graph in which all potentials are described using μ, Σ over their neighbor variables.

- Joint distribution (suitably normalized) is a multivariate Gaussian
- If the graph is a tree, you can do belief propagation, using exactly the same algorithmic structure as sum-product, but using operations on Gaussian-PDF-form functions:
 - Multiply
 - Marginalize
- It turns out that it's usually easier to do it with messages representing the same information as μ, Σ but in a different ("canonical") form. We're not going to look at it in detail.

Linear Gaussian Hidden Markov Models

$$\text{Process step: } \mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{w}_{t-1}$$

$$\text{Measurement step: } \mathbf{y}_t = H\mathbf{x}_t + \mathbf{v}_t$$

where

- \mathbf{x}_t — state vector at time step t , a random process
- \mathbf{y}_t — observation received at time step t , a random process
- \mathbf{w}_t — process noise $\sim N(\mathbf{0}, W)$
- \mathbf{v}_t — measurement noise $\sim N(\mathbf{0}, R)$
- A — process model (Note: this is not the same kind of matrix as A in the HMM, although it plays a similar role.)
- H — measurement model
- π — initial distribution $N(\mathbf{x}_0, Q_{0|0})$
- We are ignoring the control term $B\mathbf{u}_{t-1}$ (ignore this comment if it doesn't worry you)
- A and H are assumed known and constant, but could vary
- Continuous time version possible, but hairy

Filtering

Want to compute $P(\mathbf{x}_t | \mathbf{y}_{0:t})$.

- We know it's Gaussian because this is a linear Gaussian Bayesian network!
- So $P(\mathbf{x}_t | \mathbf{y}_{0:t}) = \mathcal{N}(\hat{\mathbf{x}}_{t|t}, Q_{t|t})$
- Assume we know parameters of distribution at previous step $\hat{\mathbf{x}}_{t|t-1}, Q_{t|t-1}$. Note that π is our base case.
- Recursively compute
 1. Transition update finds

$$P(\mathbf{x}_t | \mathbf{y}_{0:t-1}) = \mathcal{N}(\hat{\mathbf{x}}_{t|t-1}, Q_{t|t-1})$$

2. Observation update finds

$$P(\mathbf{x}_t | \mathbf{y}_{0:t}) = \mathcal{N}(\hat{\mathbf{x}}_{t|t}, Q_{t|t})$$

- Can be understood as sum-product on associated Gaussian factor graph

Transition update

- Current belief $P(\mathbf{x}_{t-1} \mid \mathbf{y}_{0:t-1}) = \mathcal{N}(\hat{\mathbf{x}}_{t-1|t-1}, Q_{t-1|t-1})$
- Transition $\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{w}_t$ where $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, W)$
- Construct the joint on \mathbf{x}_{t-1} and \mathbf{x}_t :

$$\mu = \begin{pmatrix} \hat{\mathbf{x}}_{t-1|t-1} \\ A\hat{\mathbf{x}}_{t-1|t-1} \end{pmatrix} \quad \Sigma = \begin{pmatrix} Q_{t-1|t-1} & Q_{t-1|t-1}A^T \\ AQ_{t-1|t-1} & AQ_{t-1|t-1}A^T + W \end{pmatrix}$$

- Marginalize out \mathbf{x}_{t-1}

$$\begin{aligned} P(\mathbf{x}_t \mid \mathbf{y}_{0:t-1}) &= \mathcal{N}(\hat{\mathbf{x}}_{t|t-1}, Q_{t|t-1}) \\ \hat{\mathbf{x}}_{t|t-1} &= A\hat{\mathbf{x}}_{t-1|t-1} \\ Q_{t|t-1} &= AQ_{t-1|t-1}A^T + W \end{aligned}$$

Note that $\text{Var}[A + B] = \text{Var}[A] + \text{Var}[B]$ when A and B are independent. Here \mathbf{x}_{t-1} and \mathbf{w}_t are independent. Also $\text{Var}[CA + c]$, where C and c are constant, is

Observation update

- Current belief $P(\mathbf{x}_t | \mathbf{y}_{0:t-1}) = \mathcal{N}(\hat{\mathbf{x}}_{t|t-1}, \mathbf{Q}_{t|t-1})$
- Observation model $\mathbf{y}_t = \mathbf{H}\mathbf{x}_t + \mathbf{v}_t$ where $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$
- Construct the joint on \mathbf{x}_t and \mathbf{y}_t

$$\mu = \begin{pmatrix} \hat{\mathbf{x}}_{t|t-1} \\ \mathbf{H}\hat{\mathbf{x}}_{t|t-1} \end{pmatrix} \quad \Sigma = \begin{pmatrix} \mathbf{Q}_{t|t-1} & \mathbf{Q}_{t|t-1}\mathbf{H}^T \\ \mathbf{H}\mathbf{Q}_{t|t-1} & \mathbf{H}\mathbf{Q}_{t|t-1}\mathbf{H}^T + \mathbf{R} \end{pmatrix}$$

- Condition on actual observation $\mathbf{y}_t = y_t$

$$P(\mathbf{x}_t | \mathbf{y}_{0:t}) = \mathcal{N}(\hat{\mathbf{x}}_{t|t}, \mathbf{Q}_{t|t})$$

$$\mathbf{Q}_{t|t} = \mathbf{Q}_{t|t-1} - \mathbf{Q}_{t|t-1}\mathbf{H}^T (\mathbf{H}\mathbf{Q}_{t|t-1}\mathbf{H}^T + \mathbf{R})^{-1} \mathbf{H}\mathbf{Q}_{t|t-1}$$

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{Q}_{t|t-1}\mathbf{H}^T (\mathbf{H}\mathbf{Q}_{t|t-1}\mathbf{H}^T + \mathbf{R})^{-1} (\mathbf{y}_t - \mathbf{H}\hat{\mathbf{x}}_{t|t-1})$$

Observation update: simplified

- Define Kalman gain $K_t = Q_{t|t-1}H^T (HQ_{t|t-1}H^T + R)^{-1}$.
- Use (tricky!) matrix algebra-fu to get useful relationships:

$$K_t = Q_{t|t}H^T R^{-1}$$

$$Q_{t|t} = Q_{t|t-1} - K_t H Q_{t|t-1}$$

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t (y_t - H\hat{x}_{t|t-1})$$

- Call $y_t - H\hat{x}_{t|t-1}$ the innovation: how surprising is our observation?
- K_t maps y_t into an opinion about x_t : Big if observations are accurate and prior on x_t is weak.
- Intuition-building rewrite:

$$\hat{x}_{t|t} = (I - K_t H) \hat{x}_{t|t-1} + K_t y_t$$

Some important properties of the Kalman filter:

- Transition adds uncertainty: $Q_{t|t-1}$ is always “larger” than $Q_{t-1|t-1}$
- Observation reduces uncertainty: $Q_{t|t}$ is always “smaller” than $Q_{t|t-1}$

Kalman smoothing

Just as in discrete HMMs, we can run a similar belief-propagation pass backward to compute $P(\mathbf{x}_t \mid \mathbf{y}_{0:T})$

In Gaussian systems, the max of the individual marginals is the max of the joint!!!

Handling non-linear and non-Gaussian systems

- Extended / unscented Kalman filter
- Particle filter