L06 – Continuous factored and temporal models

Barber 24.1,3,4; 27.6 AIMA 13.2.3, 14.4, 14.5.3

What you should know after this lecture

- Gaussian graphical models
- Kalman filtering

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.

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

[1] The 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 distribution[2]

- $\Omega = \{x_1, \ldots, x_M\}$
- $\theta^A = (\theta_1^A, \ldots, \theta_M^A)$ $\qquad$ $\theta^B = (\theta_1^B, \ldots, \theta_M^B)$
- $f_A(\theta^A)(x_i) = \theta_i^A$ $\qquad$ $f_B(\theta^B)(x_i) = \theta_i^B$

If we multiply these functions 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 $Z = \sum_{i=1}^{M} \theta_i^A \theta_i^B$

[2] We like the name "multinoulli" better, though!

Categorical distribution is conjugate for joint

Combining two categorical distributions on different variables:

- $\Omega_A = \{a_1, \ldots, a_M\}$ $\quad$ $\Omega_B = \{b_1, \ldots, b_N\}$
- $\theta^A = (\theta^A_1, \ldots, \theta^A_M)$ $\quad$ $\theta^B = (\theta^B_1, \ldots, \theta^B_N)$
- $f_A(\theta^A)(a_i) = \theta^A_i$ $\quad$ $f_B(\theta^B)(b_i) = \theta^B_i$

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

- $\Omega_{AB} = \Omega_A \times \Omega_B$
- $f_{AB}(\theta^{AB})(a_i, b_j) = \theta^{AB}(a_i, b_j) = \theta^A_i \cdot \theta^B_j$

Univariate Gaussian is conjugate family

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

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

- Observe that multiplying f's yields

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

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

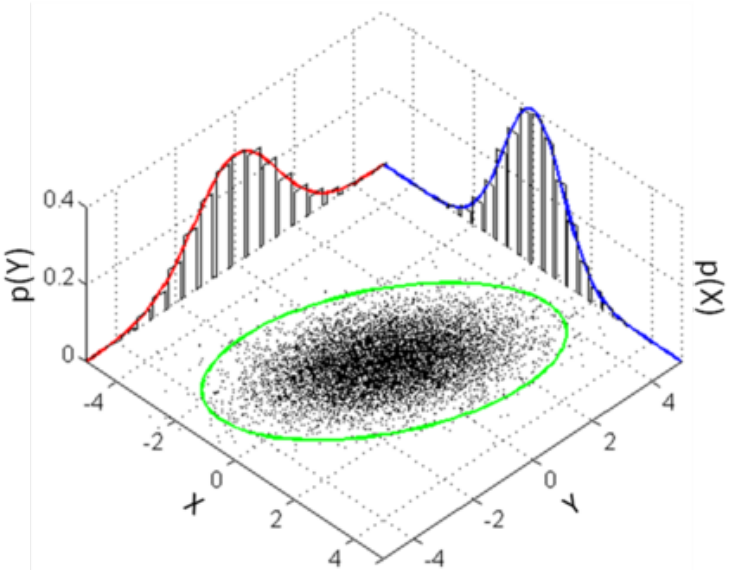
$$\mu_{AB} = \frac{\mu_A\sigma_B^2 + \mu_B\sigma_A^2}{\sigma_A^2 + \sigma_B^2} \qquad \sigma_{AB}^2 = \frac{\sigma_A^2\sigma_B^2}{\sigma_A^2 + \sigma_B^2}$$

Multivariate Gaussian

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

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

$|\Sigma|$ 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

$$\mu_B^m = \mu_B \quad \Sigma_B^m = \Sigma_{BB}$$

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

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

For random variables $X_1, \ldots, X_n$ that are jointly Gaussian with parameters μ, Σ:

- The mean of $c_0 + \sum_i c_i X_i$, where the c_i are constants, is $c_0 + \sum_i c_i \mu_i$
- The variance of $c_0 + \sum_i c_i X_i$ is $c^T \Sigma c$

Multivariate Gaussian is conjugate family

Product of MVGs:

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

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

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

Can be useful to define precision : $\Lambda = \Sigma^{-1}$
Then $\Lambda_{AB} = \Lambda_A + \Lambda_B$ and

$$\mu_{AB} = (\Lambda_A + \Lambda_B)^{-1}(\Lambda_A\mu_A + \Lambda_B\mu_b)$$

Multivariate Gaussian is conjugate for joint

Product of MVGs on different domains

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

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 $N - 1$ (number of parents of V) and w_0 is a scalar offset
- η_V^2 is the variance of added noise at this node

then the joint distribution over all variables $V_1, \ldots, V_N, V$ is Gaussian.

- Assume the parents of node V are normally distributed with mean μ_P, Σ_P the distribution over V is normal with
- $\mu_V = w_V^0 + W_V^T \mu_P$
- $\sigma_V^2 = \eta_V^2 + w^T \Sigma_P w$

Gaussian Bayesian networks

- Assume distribution $V \sim \text{Normal}(w_V^0 + w_V^T \cdot \text{pa}(V), \eta_V^2)$
- Assume the parents of V are normally distributed with mean μ_P, Σ_P

then the joint distribution over all variables $V_1, \ldots, V_N, V$ is Gaussian with

- Mean: μ_P, μ_V
- Cov:

$$\begin{bmatrix} \Sigma_P & \Sigma_{PV} \\ \Sigma_{PV}^T & \sigma_V^2 \end{bmatrix}$$

  where $\Sigma_{PV}[i] = \sum_j \Sigma_P[i, j]$

By induction, you can show that a whole Bayes net with this linear Gaussian structure defines a joint Gaussian distribution!

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 propgation, 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

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

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

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 $\sim N(0, W)$
- $\mathbf{v}_t$ — measurement noise $\sim N(0, R)$
- 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 $P(\mathbf{x}_t \mid \mathbf{y}_{0:t})$.

- We know it's Gaussian because this is a linear Gaussian Bayesian network!
- So $P(\mathbf{x}_t \mid \mathbf{y}_{0:t}) = \mathcal{N}(\hat{\mathbf{x}}_{t|t}, Q_{t|t})$
- 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

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

  2. Observation update finds

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

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

Transition update

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

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

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

Observation update

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

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

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

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

$$Q_{t|t} = Q_{t|t-1} - Q_{t|t-1}H^T \left(HQ_{t|t-1}H^T + R\right)^{-1} HQ_{t|t-1}$$

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

Observation update: simplified

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

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

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

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

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

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

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 $Q_{t|t-1}$ (or same)

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!!!

What if your system isn't conjugate?

- Gaussian errors, but non-linear dynamics: extended Kalman filter
- Somewhat non-Gaussian errors, non-linear dynamics: unscented Kalman filter
- Arbitrary model: particle filter : next time!