Lecture 13: March 4, 2015
cs 573: Probabilistic Reasoning
Professor Nevatia
Spring 2015

Review

- Assignment # 4 due March 9
- Exam 1may be graded by March 11
- Last lecture:
  - Loopy Belief Propagation Algorithm
  - Very brief intro to Gaussian Distributions
- Today's objective
  - Inference in Gaussian Networks

Multi-variate Gaussian Distribution

- Eq. 7.1

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

  – Mean vector, $\boldsymbol{\mu}$; $\boldsymbol{\mu} = E(\mathbf{X})$; μ_i is mean of X_i
  – Σ is n x n covariance matrix,
    - $\Sigma = E[\mathbf{X}\,\mathbf{X}^T] - \mathbf{E}[\mathbf{X}]\,\mathbf{E}[\mathbf{X}^T]$
    - $\Sigma_{i,i}$ is the variance of X_i;
    - $\Sigma_{i,j} = \Sigma_{j,i}$ is the *covariance* between X_i and X_j
      – $\mathbf{C}\text{ov}[X_i ; X_j] = \mathbf{E}[X_i X_j] - \mathbf{E}[X_i]\,\mathbf{E}[X_j]$
  – Σ must be positive definite, $\mathbf{x}^T\Sigma\,\mathbf{x} > 0$, $\mathbf{x} \neq 0$, for density to be well defined; Equivalent property: all eigenvalues are > 0
- *Standard* Multivariate Gaussian: $\boldsymbol{\mu} = \mathbf{0}$ (vector), $\Sigma = I$ (identity matrix) (1s on diagonal, 0s elsewhere)

Multi-variate Gaussian Distribution

- 2-D example

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}.$$

- 3-D example

$$\mu = \begin{pmatrix} 1 \\ -3 \\ 4 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 4 & 2 & -2 \\ 2 & 5 & -5 \\ -2 & -5 & 8 \end{pmatrix}$$

Marginalize a Gaussian Distribution

- For multiple variables, best not to expand the terms

$$p(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{N}\left(\begin{pmatrix} \mu_{\boldsymbol{X}} \\ \mu_{\boldsymbol{Y}} \end{pmatrix}; \begin{bmatrix} \Sigma_{\boldsymbol{XX}} & \Sigma_{\boldsymbol{XY}} \\ \Sigma_{\boldsymbol{YX}} & \Sigma_{\boldsymbol{YY}} \end{bmatrix}\right)$$

  – 1^{st} term in matrix above is n x n, 2^{nd} is n x m, and 3^{rd} is m x n, 4^{th} is n x n (**X** has n elements, **Y** has m elements)

  – Can be shown that marginal over **Y** (sum out **X**) is Gaussian, given by $N(\boldsymbol{\mu}_{\mathbf{Y}}, \Sigma_{\mathbf{YY}})$

    - Derivation is straight-forward but requires some manipulation of matrix terms; for a derivation see http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html

Computing Conditional Distribution

- Say **Y** = **y**
- Substitute in the density formula; consider 2 variable case

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right),$$

- Result is p(X), y is treated as a constant (parameter)
- Equation represents a valid Gaussian
- For multi-variate case, expressions for the new **μ** and Σ are complex and require some matrix inversions
  - See next slide

Conditional Distribution Formulas

From Wikipedia

If μ and Σ are partitioned as follows

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \text{with sizes} \begin{bmatrix} q \times 1 \\ (N-q) \times 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \text{with sizes} \begin{bmatrix} q \times q & q \times (N-q) \\ (N-q) \times q & (N-q) \times (N-q) \end{bmatrix}$$

then, the distribution of $\mathbf{x}_1$ conditional on $\mathbf{x}_2 = a$ is multivariate normal $(\mathbf{x}_1|\mathbf{x}_2 = \mathbf{a}) \sim N(\overline{\mu}, \overline{\Sigma})$ where

$$\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(\mathbf{a} - \mu_2\right)$$

and covariance matrix

$$\overline{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$ [8]

Information Form

- Information matrix, J, (also called precision matrix)

$$J = \Sigma^{-1}$$

$$-1/2\ (x - \boldsymbol{\mu})^T \Sigma^{-1} (x - \boldsymbol{\mu}) = -1/2\ (x - \boldsymbol{\mu})^T J (x - \boldsymbol{\mu})$$

$$= -1/2\ [x^T J x - 2x^T J \boldsymbol{\mu} + \boldsymbol{\mu}^T J\, \boldsymbol{\mu}]$$

$$p(\boldsymbol{x}) \propto \exp\left[-\frac{1}{2}\boldsymbol{x}^T J \boldsymbol{x} + (J\boldsymbol{\mu})^T \boldsymbol{x}\right]$$

- **h** = J**μ** is called the potential vector
- J must be positive definite (also symmetric)
- Conditioning:
  - Info form has terms such as $(1/2)*J_{ii}x_i^2 - J_{ij}x_ix_j + h_ix_i +$ constant
  - If X_i is in evidence (conditioning set), linear term and square terms (containing only x_i) become constants, product term becomes linear
  - Resulting form is same as original information form

Canonical Form

- Similar to information form
- Expand the Gaussian density expression as:

$$\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$
$$= \exp\left(-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x} + \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} - \log\left((2\pi)^{n/2}|\Sigma|^{1/2}\right)\right)$$

- Let:

$$\begin{aligned} K &= \Sigma^{-1} \\ \boldsymbol{h} &= \Sigma^{-1}\boldsymbol{\mu} \\ g &= -\frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} - \log\left((2\pi)^{n/2}|\Sigma|^{1/2}\right) \end{aligned}$$

$$\mathcal{C}(\boldsymbol{X}; K, \boldsymbol{h}, g) = \exp\left(-\frac{1}{2}\boldsymbol{X}^T K \boldsymbol{X} + \boldsymbol{h}^T\boldsymbol{X} + g\right) \quad \text{Eq 14.1}$$

- Note K is same as J in information form

Conditional Distribution in Canonical Form

- In canonical form, marginalize over Y (eq 14.6)

$$K = \begin{bmatrix} K_{XX} & K_{XY} \\ K_{YX} & K_{YY} \end{bmatrix} \quad ; \quad h = \begin{pmatrix} h_X \\ h_Y \end{pmatrix}$$

$$\begin{aligned} K' &= K_{\boldsymbol{XX}} \\ h' &= \boldsymbol{h}_{\boldsymbol{X}} - K_{\boldsymbol{XY}}\boldsymbol{y} \\ g' &= g + \boldsymbol{h}_{\boldsymbol{Y}}^T\boldsymbol{y} - \frac{1}{2}\boldsymbol{y}^T K_{\boldsymbol{YY}}\boldsymbol{y} \end{aligned}$$

Independencies

- X_i and X_j =are independent *iff* $\Sigma_{i,j} = 0$ (no edge in directed graph)
- $J_{i,j} = 0$ *iff* $(X_i \perp X_j \mid X - \{X_i, X_j\})$
  - Information matrix directly defines a minimal I-map
  - If $J_{i,j} \neq 0$, edge between *i* and *j* nodes (in undirected graph)
- Ex 7.2

$$J = \begin{pmatrix} 0.3125 & -0.125 & 0 \\ -0.125 & 0.5833 & 0.3333 \\ 0 & 0.3333 & 0.3333 \end{pmatrix}$$

$$\mu = \begin{pmatrix} 1 \\ -3 \\ 4 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 4 & 2 & -2 \\ 2 & 5 & -5 \\ -2 & -5 & 8 \end{pmatrix}$$

Gaussian Bayesian Networks (GBN)

- In a GBN, all variables are continuous; all CPDs are linear Gaussian

Let Y be a continuous variable with continuous parents $X_1, \dots, X_k$. We say that Y has a linear Gaussian *model if there are parameters $\beta_0, \dots, \beta_k$ and σ^2 such that*

$$p(Y \mid x_1, \dots, x_k) = \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k; \sigma^2\right).$$

In vector notation,

$$p(Y \mid \boldsymbol{x}) = \mathcal{N}\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}; \sigma^2\right).$$

– Thm 7.3

Given: $p(Y \mid x) = \mathcal{N}\left(\beta_0 + \beta^T x; \sigma^2\right)$; $X_1, .., X_k$ distributed as $N(\boldsymbol{\mu}, \Sigma)$

We can show that: Y is a normal distribution $N(\mu_Y; \sigma_Y^2)$

$$\begin{aligned} \mu_Y &= \beta_0 + \beta^T \mu \\ \sigma_Y^2 &= \sigma^2 + \beta^T \Sigma \beta. \end{aligned}$$

Also that {X,Y} is a normal distribution where $Cov[X_i; Y] = \sum_{j=1}^{k} \beta_j \Sigma_{i,j}.$

Distribution to GBN

- Previous slide shows how given a GBN (network parameters), we can get joint distribution parameters.
- Reverse: given the joint distribution, recover the linear model (thm 7.4)

Given:

$$p(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{N}\left(\left(\begin{array}{c} \boldsymbol{\mu}_{\boldsymbol{X}} \\ \boldsymbol{\mu}_{\boldsymbol{Y}} \end{array}\right); \left[\begin{array}{cc} \Sigma_{\boldsymbol{X}\boldsymbol{X}} & \Sigma_{\boldsymbol{X}\boldsymbol{Y}} \\ \Sigma_{\boldsymbol{Y}\boldsymbol{X}} & \Sigma_{\boldsymbol{Y}\boldsymbol{Y}} \end{array}\right]\right)$$

Derive:

$$p(Y \mid \boldsymbol{X}) = \mathcal{N}\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{X}; \sigma^2\right)$$

$$\begin{aligned} \beta_0 &= \mu_Y - \Sigma_{Y\boldsymbol{X}} \Sigma_{\boldsymbol{X}\boldsymbol{X}}^{-1} \boldsymbol{\mu}_{\boldsymbol{X}} \\ \boldsymbol{\beta} &= \Sigma_{\boldsymbol{X}\boldsymbol{X}}^{-1} \Sigma_{Y\boldsymbol{X}} \\ \sigma^2 &= \Sigma_{YY} - \Sigma_{Y\boldsymbol{X}} \Sigma_{\boldsymbol{X}\boldsymbol{X}}^{-1} \Sigma_{\boldsymbol{X}Y} \end{aligned}$$

Distribution to GBN

- Given a distribution over *n* variables, we can construct a GBN (BN with linear Gaussian Models) that is an I-map of the distribution
- A key difference with discrete case: number of parameters in GBN is not necessarily smaller than in the joint distribution itself as the joint distribution is compact by itself.
- We can also go from distributions to Markov networks
  - J_{ij} help define pairwise (log) potentials
  - However, additional complexity in case of MN; not every set of potentials induces a valid Gaussian distributions
    - Sufficient but not necessary condition is that each edge potential be normalizable (corresponding information matrix is positive definite)
  - We skip other details of Gaussian Markov Random Fields (sec 7.3)

Inference in Networks with Continuous Variables

- Essentially, all the algorithms for discrete case apply
- Sum-Product algorithm
  - Steps consist of multiplying factors (product) and marginalizing over some variables (sum) and passing messages to other nodes
  - Iterate until convergence (two passes suffice for trees)
  - We already know how to marginalize Gaussians
  - Now consider product and division

Marginalize in Canonical Form

- More complex then in covariance form (eq 14.5)

$$K = \begin{bmatrix} K_{\boldsymbol{XX}} & K_{\boldsymbol{XY}} \\ K_{\boldsymbol{YX}} & K_{\boldsymbol{YY}} \end{bmatrix} \quad ; \quad h = \begin{pmatrix} h_{\boldsymbol{X}} \\ h_{\boldsymbol{Y}} \end{pmatrix}$$

$$\begin{aligned} K' &= K_{\boldsymbol{XX}} - K_{\boldsymbol{XY}} K_{\boldsymbol{YY}}^{-1} K_{\boldsymbol{YX}} \\ h' &= \boldsymbol{h}_{\boldsymbol{X}} - K_{\boldsymbol{XY}} K_{\boldsymbol{YY}}^{-1} \boldsymbol{h}_{\boldsymbol{Y}} \\ g' &= g + \frac{1}{2}\left(\log\left|2\pi K_{\boldsymbol{YY}}^{-1}\right| + \boldsymbol{h}_{\boldsymbol{Y}}^{T} K_{\boldsymbol{YY}}^{-1} \boldsymbol{h}_{\boldsymbol{Y}}\right) \end{aligned}$$

- We can switch between the covariance and canonical forms, depending on the desired computation
  - However, switching requires inverting K or Σ matrices as well.

Operations on Canonical Forms

- Product of two canonical forms over the same set of variables:
- $C(\mathbf{X}, K_1, \mathbf{h}_1, g_1) . C(\mathbf{X}, K_2, \mathbf{h}_2, g_2) = C(\mathbf{X}, K_1+K_2, \mathbf{h}_1+ \mathbf{h}_2, g_1 + g_2)$
  - Formula follows from the definition of the canonical form; we are just multiplying two Gaussians over **X**
  - If scopes of two factors are different, expand each by including zeros entries in K and **h** for entries corresponding to absent variables
- Example 14.1

$$\phi_1(X,Y) = \mathcal{C}\left(X,Y; \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, -3\right)$$

$$\phi_2(Y,Z) = \mathcal{C}\left(Y,Z; \begin{bmatrix} 3 & -2 \\ -2 & 4 \end{bmatrix}, \begin{pmatrix} 5 \\ -1 \end{pmatrix}, 1\right).$$

$$\phi_1(X,Y,Z) = \mathcal{C}\left(X,Y,Z; \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, -3\right)$$

Example 14.2

- After multiplication

$$\mathcal{C}\left(X, Y, Z; \begin{bmatrix} 1 & -1 & 0 \\ -1 & 4 & -2 \\ 0 & -2 & 4 \end{bmatrix}, \begin{pmatrix} 1 \\ 4 \\ -1 \end{pmatrix}, -2\right)$$

- Division

$$\frac{\mathcal{C}(K_1, \boldsymbol{h}_1, g_1)}{\mathcal{C}(K_2, \boldsymbol{h}_2, g_2)} = \mathcal{C}(K_1 - K_2, \boldsymbol{h}_1 - \boldsymbol{h}_2, g_1 - g_2)$$

- Vacuous Canonical Form
  - K = 0, and **h** = 0, g =0
  - Similar to factor with all 1 in discrete case
  - Multiplying by it has no effect, but can be used to initialize potentials and make them "ready" for passing messages

SP and BP Algorithms

- Sum-Product algorithm
  - As before; need to show that resulting factors maintain K matrices to be positive definite so integration is not infinite
  - Shown in Proposition 14.1
- Gaussian Belief Propagation
  - Similar steps as in discrete case but the observation that the potentials are quadratic simplifies the equations (eqs 14.7 to 14.9)
  - **Interesting property**: If BP converges, resulting beliefs encode the correct means but estimated variances are underestimates (overconfident estimates)
  - Also convergence guaranteed if pairwise normalizability condition holds.
- We skip other details

Next Class

- Read section 11.5.1 of the KF book