

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

# Review

- Assignment # 4 due March 9
- Exam 1 may 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

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

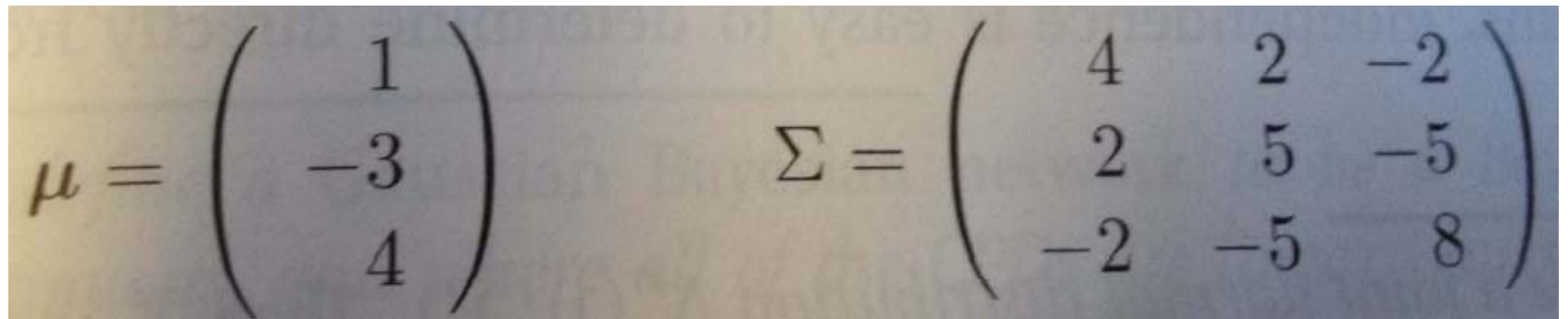
- Mean vector,  $\boldsymbol{\mu}$  ;  $\boldsymbol{\mu} = \mathbf{E}(\mathbf{X})$ ;  $\mu_i$  is mean of  $X_i$
- $\Sigma$  is  $n \times n$  covariance matrix,
  - $\Sigma = \mathbf{E}[\mathbf{X} \mathbf{X}^T] - \mathbf{E}[\mathbf{X}] \mathbf{E}[\mathbf{X}^T]$
  - $\Sigma_{i,i}$  is the variance of  $X_i$ ;
  - $\Sigma_{i,j} = \Sigma_{j,i}$  is the *covariance* between  $X_i$  and  $X_j$ 
    - $\text{Cov}[X_i; X_j] = \mathbf{E}[X_i X_j] - \mathbf{E}[X_i] \mathbf{E}[X_j]$
- $\Sigma$  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

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

- 3-D example



A photograph of handwritten mathematical expressions for a 3-D Gaussian distribution. The mean vector  $\mu$  is shown as a column vector with elements 1, -3, and 4. The covariance matrix  $\Sigma$  is shown as a 3x3 symmetric matrix with elements 4, 2, -2 in the first row; 2, 5, -5 in the second row; and -2, -5, 8 in the third row.

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

# Marginalize a Gaussian Distribution

- For multiple variables, best not to expand the terms

$$p(\mathbf{X}, \mathbf{Y}) = \mathcal{N} \left( \begin{pmatrix} \mu_{\mathbf{X}} \\ \mu_{\mathbf{Y}} \end{pmatrix}; \begin{bmatrix} \Sigma_{\mathbf{X}\mathbf{X}} & \Sigma_{\mathbf{X}\mathbf{Y}} \\ \Sigma_{\mathbf{Y}\mathbf{X}} & \Sigma_{\mathbf{Y}\mathbf{Y}} \end{bmatrix} \right)$$

- 1<sup>st</sup> term in matrix above is  $n \times n$ , 2<sup>nd</sup> is  $n \times m$ , and 3<sup>rd</sup> is  $m \times n$ , 4<sup>th</sup> is  $m \times m$  ( $\mathbf{X}$  has  $n$  elements,  $\mathbf{Y}$  has  $m$  elements)
- Can be shown that marginal over  $\mathbf{Y}$  (sum out  $\mathbf{X}$ ) is Gaussian, given by  $\mathcal{N}(\mu_{\mathbf{Y}}, \Sigma_{\mathbf{Y}\mathbf{Y}})$ 
  - 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  $\mathbf{Y} = \mathbf{y}$
- 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  $\boldsymbol{\mu}$  and  $\Sigma$  are complex and require some matrix inversions
  - See next slide

# Conditional Distribution Formulas

From Wikipedia

If  $\mu$  and  $\Sigma$  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  $\mathbf{x}_2 = \mathbf{a}$  is multivariate normal  $(\mathbf{x}_1 | \mathbf{x}_2 = \mathbf{a}) \sim N(\bar{\mu}, \bar{\Sigma})$  where

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

and covariance matrix

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

## Information Form

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

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

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

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

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

- $\mathbf{h} = J\boldsymbol{\mu}$  is called the potential vector
- $J$  must be positive definite (also symmetric)
- Conditioning:
  - Info form has terms such as  $(1/2) * J_{ii} x_i^2 - J_{ij} x_i x_j + h_i x_i + \text{constant}$
  - 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:

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

- Let:

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

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

- Note K is same as J in information form

# Conditional Distribution in Canonical Form

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

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

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

# 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

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

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

# 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  $\sigma^2$  such that*

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

*In vector notation,*

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

## – Thm 7.3

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

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

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

Also that  $\{X, Y\}$  is a normal distribution where

$$\text{Cov}[X_i; Y] = \sum_{j=1}^k \beta_j \Sigma_{i,j}.$$

## 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:

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

Derive:

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

$$\begin{aligned} \beta_0 &= \mu_Y - \Sigma_{YX} \Sigma_{XX}^{-1} \mu_X \\ \boldsymbol{\beta} &= \Sigma_{XX}^{-1} \Sigma_{YX} \\ \sigma^2 &= \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} \end{aligned}$$

## 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 than in covariance form (eq 14.5)

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

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

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



# Operations on Canonical Forms

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

$$\begin{aligned}\phi_1(X, Y) &= C\left(X, Y; \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, -3\right) \\ \phi_2(Y, Z) &= C\left(Y, Z; \begin{bmatrix} 3 & -2 \\ -2 & 4 \end{bmatrix}, \begin{pmatrix} 5 \\ -1 \end{pmatrix}, 1\right).\end{aligned}$$

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

## Example 14.2

- After multiplication

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

- Division

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

- Vacuous Canonical Form
  - $K = 0$ , and  $\mathbf{h} = 0$ ,  $g = 0$
  - 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