Gaussian Bayesian Networks

Sargur Srihari srihari@cedar.buffalo.edu

Topics

- Linear Gaussian Model
- Learning Parameters of a Gaussian Bayesian Network
- Multivariate Gaussian representations
 - 1. Covariance Form
 - 2. Information Form:
 - 1. Bayesian Networks
 - Markov Random Fields

Linear Gaussian Model

• Y has a linear Gaussian model with continuous parents $X_1,...X_k$

with parameters β_0 ... β_k and σ^2 such that

$$P(Y|X_1,..X_k) \sim N(\beta_0 + \beta_1 X_1 + ... \beta_k X_k; \sigma^2)$$

In vector notation

$$P(Y|X) \sim N(\beta_0 + \beta^T X; \sigma^2)$$

- Can be viewed as $Y=\beta_0+\beta_1X_1+...\beta_kX_k+\varepsilon$
 - i.e., Y is a linear function of the variables with the addition of Gaussian noise with mean θ and variance σ^2

Linear regression could be used to solve for the parameters

 X_k

Gaussian Bayesian Network

- Definition: All variables are Gaussian and all CPDs are linear Gaussian
- Let Y be a linear Gaussian of its parents $X_1,...,X_k$
 - i.e., mean of Y is a linear combination of means of Gaussian parents

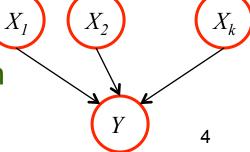
$$P(Y|X) \sim N(\beta_0 + \beta^t X; \sigma^2)$$

• If $X_1,...X_k$ are jointly normal $N(\mu;\Sigma)$ then the distribution of Y is normal $N(\mu_Y;\sigma^2_Y)$ where

$$\mu_Y = \beta_0 + \beta^t \mu \text{ and } \sigma^2_Y = \sigma^2 + \beta^t \Sigma \beta$$

The joint distribution over (X,Y) is normal with

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



Learning parameters of Gaussian Bayesian Network

- All variables are Gaussian
- $P(X|u_1,...u_k) \sim N(\beta_0 + \beta_1 u_1 + ... \beta_k u_k; \sigma^2)$
- Task is to learn the parameters

$$\boldsymbol{\theta}_{X|U} = \{\beta_0, \beta_1, \dots \beta_k, \sigma^2\}$$

 To find ML estimate of parameters, define loglikelihood

$$\log L_{\boldsymbol{X}}(\boldsymbol{\theta}_{\boldsymbol{X}|\boldsymbol{U}}:D) = \sum_{\boldsymbol{m}} \left[-\frac{1}{2} \log \left(2\pi \sigma^2 \right) - \frac{1}{2\sigma^2} \left(\beta_0 + \beta_1 u_{\scriptscriptstyle 1}[\boldsymbol{m}] + ..\beta_k u_{\scriptscriptstyle k}[\boldsymbol{m}] - \boldsymbol{x}[\boldsymbol{m}] \right)^2 \right]$$

Closed-form solutions for parameters

Solving for the parameters

• Gradient of log-likelihood wrt β_0

$$-\frac{1}{\sigma^2}\!\!\left[\!M\beta_0+\beta_1\!\sum_m\!u_{\!\scriptscriptstyle 1}\!\left[m\right]\!+\ldots\!+\beta_k\!\sum_m\!u_{\!\scriptscriptstyle k}\!\left[m\right]\!-x\!\left[m\right]\!\right]$$

Equating to zero and rearranging we get

$$\frac{1}{M} \sum_{\boldsymbol{m}} \boldsymbol{x} \left[\boldsymbol{m} \right] = \beta_{\scriptscriptstyle 0} + \beta_{\scriptscriptstyle 1} \frac{1}{M} \sum_{\boldsymbol{m}} \boldsymbol{u}_{\scriptscriptstyle 1} \left[\boldsymbol{m} \right] + \ldots + \beta_{\scriptscriptstyle k} \frac{1}{M} \sum_{\boldsymbol{m}} \boldsymbol{u}_{\scriptscriptstyle k} \left[\boldsymbol{m} \right]$$

- All the summations can be obtained from data thus giving us a linear equation
- Similarly we get k more linear equations by taking derivatives wrt β_i
- Standard linear algebra techniques are used to solve k+1 simultaneous equations

Srihari Machine Learning

Solving Linear Equations

$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n = b_m$$

$$\begin{vmatrix} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2 \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n = b_m \end{vmatrix} + x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \dots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

1. Matrix Solution: Ax=b Therefore x=A-1b
$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

2. Gaussian Elimination

$$x + 3y - 2z = 5$$
$$3x + 5y + 6z = 7$$
$$2x + 4y + 3z = 8$$

$$\begin{bmatrix} 3x + 5y + 6z = 7 \\ 2x + 4y + 3z = 8 \end{bmatrix} \begin{bmatrix} 1 & 3 & -2 & | & 5 \\ 3 & 5 & 6 & | & 7 \\ 2 & 4 & 3 & | & 8 \end{bmatrix} \sim \begin{bmatrix} 1 & 3 & -2 & | & 5 \\ 0 & -4 & 12 & | & -8 \\ 2 & 4 & 3 & | & 8 \end{bmatrix} \sim \begin{bmatrix} 1 & 3 & -2 & | & 5 \\ 0 & -4 & 12 & | & -8 \\ 2 & 4 & 3 & | & 8 \end{bmatrix} \sim \begin{bmatrix} 1 & 3 & -2 & | & 5 \\ 0 & -4 & 12 & | & -8 \\ 0 & -2 & 7 & | & -2 \end{bmatrix} \sim \begin{bmatrix} 1 & 3 & -2 & | & 5 \\ 0 & 1 & -3 & | & 2 \\ 0 & 0 & 1 & | & 2 \end{bmatrix} \sim \begin{bmatrix} 1 & 3 & -2 & | & 5 \\ 0 & 1 & 3 & | & 2 \\ 0 & 0 & 1 & | & 2 \end{bmatrix} \sim \begin{bmatrix} 1 & 3 & 0 & | & 9 \\ 0 & 1 & 0 & | & 8 \\ 0 & 0 & 1 & | & 2 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & | & -15 \\ 0 & 1 & 0 & | & 8 \\ 0 & 0 & 1 & | & 2 \end{bmatrix}$$

Estimating variance

• Taking derivative of likelihood and setting to zero, we get U_1

$$\sigma^2 = Cov_{_{D}} \big[X; X \big] - \sum_i \sum_j \beta_{_i} \beta_{_j} Cov_{_{D}} \big[U_{_i}; U_{_j} \big]$$

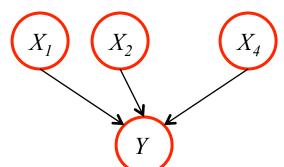
- where

$$\begin{split} Cov_{_{D}}\!\left[\boldsymbol{X};\boldsymbol{Y}\right] &= E_{_{D}}\!\left[\boldsymbol{X}\cdot\boldsymbol{Y}\right] - E_{_{D}}\!\left[\boldsymbol{X}\right]\!\cdot\boldsymbol{E}_{_{D}}\!\left[\boldsymbol{Y}\right] \\ &E_{_{D}}\!\left[\boldsymbol{X}\cdot\boldsymbol{Y}\right] = \frac{1}{M}\!\sum_{\boldsymbol{m}}\!\boldsymbol{x}\!\left[\boldsymbol{m}\right]\!\boldsymbol{y}\!\left[\boldsymbol{m}\right] \\ &E_{_{D}}\!\left[\boldsymbol{X}\right] = \frac{1}{M}\!\sum_{\boldsymbol{m}}\!\boldsymbol{x}\!\left[\boldsymbol{m}\right]\!\boldsymbol{y}\!\left[\boldsymbol{m}\right] \end{split}$$

- First term is the empirical variance of X
- Other terms are empirical covariances of inputs

Properties of Linear Gaussian

- Advantage: Simple model that captures many dependencies
- Disadvantage: cannot capture dependence of variance of child on values of parents
 - Can be extended
 - E.g., mean of Y is $sin(x_1)^{x_2}$ variance is $(x_3/x_4)^2$



- But linear Gaussian is a good approximation
- Provides an alternative for representing multivariate Gaussian distributions

Multivariate Gaussian

- Three equivalent representations
 - 1. Basic parameterization (covariance form)

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left[(\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] \qquad \begin{bmatrix} \boldsymbol{\mu} = E[\mathbf{x}] \\ \boldsymbol{\Sigma} = E[\mathbf{x}\mathbf{x}^t] - E[\mathbf{x}]E[\mathbf{x}]^t \end{bmatrix}$$

$$\mu = E[\mathbf{x}]$$

$$\Sigma = E[\mathbf{x}\mathbf{x}^t] - E[\mathbf{x}]E[\mathbf{x}]^t$$

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

 x_1 is negatively correlated with x_3 x_2 is negatively correlated with x_3

- 2. Gaussian Bayesian network (information form)
 - Conversion to Information form (precision matrix) captures conditional independencies
- 3. Gaussian Markov Random Field
 - **Easiest conversion**

Gaussian Distribution Operations

- There are two main operations we wish to perform on a distribution:
 - 1. Marginal distribution over some subset Y
 - 2. Conditioning the distribution on some Z=z
- Marginalization is trivially performed in the covariance form
- Conditioning a Gaussian is very easy to perform in the information form

Marginalization in Covariance Form

- Can be simply read from the mean and covariance matrix
- Decompose the mean and covariance as:

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

• Marginal distribution over Y is a normal distribution $N(\mu_Y; \Sigma_{YY})$

Determining Conditional Distribution using Covariance Form

Let (X,Y) have a normal distribution defined by

$$p(X,Y) = N \left(\left(\begin{array}{c} \mu_X \\ \mu_Y \end{array} \right); \left[\begin{array}{cc} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{array} \right] \right)$$

• Then marginal distribution P(Y|X) is given by $P(Y|X) \sim N(\beta_0 + \beta^T X; \sigma^2)$ where

$$\beta_0 = \mu_Y - \Sigma_{YX} \Sigma_{XX}^{-1} \mu_X$$
$$\beta = \Sigma_{XX}^{-1} \Sigma_{YX}$$
$$\sigma^2 = \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}$$

 This allows us to take a joint Gaussian distribution and construct a Bayesian network

Constructing a BN from a Gaussian

- Definition of I-map
 - I is set of independence assertions of form $(X \perp Y \mid Z)$, which are true of distribution P
 - $-\mathcal{G}$ is an I-map of P implies: $I_G \subseteq I$
 - Minimal I-map: removal of a single edge renders the graph not an I-map
- Algorithm uses the principle that for Z to be parents of node X, $(X \perp Y \mid Z)$ where $Y = \chi \cdot Z$
 - need to check if P(X|Z) is independent of $P(Y_i|Z)$
 - Can be done using deviance measure: χ^2 or mutual info (KL-div) $\begin{bmatrix} d_{\chi^2}(\mathcal{D}) = \sum_{x_i,x_j} \frac{\left(M[x_i,x_j] M \cdot \hat{P}(x_i) \cdot \hat{P}(x_j)\right)^2}{M \cdot \hat{P}(x_i) \cdot \hat{P}(x_j)} \end{bmatrix}$ $\begin{bmatrix} d_{\chi}(\mathcal{D}) = \frac{1}{M} \sum_{x_i,x_j} M[x_i,x_j] \log \frac{M[x_i,x_j]}{M[x_i]M[x_j]} \end{bmatrix}$
 - Simpler Method: use Information form of Gaussian

BN from conditional independencies

Algorithm computes an I-Map which may not be unique

- $\chi = \{X_1, ... X_n\}$: set of random variables
- *I*: set of independencies

```
Set \mathcal{G} to an empty graph over \mathcal{X}
         for i = 1, \ldots, n
3
           U \leftarrow \{X_1, \dots, X_{i-1}\} // U is the current candidate for parents of X_i
           for U' \subseteq \{X_1, \ldots, X_{i-1}\}
5
              if U' \subset U and (X_i \perp \{X_1, \dots, X_{i-1}\} - U' \mid U') \in \mathcal{I} then
                 U \leftarrow U'
6
              // At this stage U is a minimal set satisfying (X_i \perp
                 \{X_1,\ldots,X_{i-1}\}-U\mid U
8
               // Now set U to be the parents of X_i
9
           for X_i \in U
                                                                            Nodes other than U'
10
              Add X_i \to X_i to \mathcal{G}
11
         return \mathcal{G}
```

Information form gives edges of BN

• Standard Form
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp[(\mathbf{x} - \mathbf{\mu})^t \Sigma^{-1} (\mathbf{x} - \mathbf{\mu})]$$

• Using Precision $J=\Sigma^{-1}$

$$\left| -\frac{1}{2} (x - \mu)^t \Sigma^{-1} (x - \mu) = -\frac{1}{2} (x - \mu)^t J(x - \mu) \right| = -\frac{1}{2} \left[x^t J x - 2x^t J \mu + \mu^t J \mu \right]$$

$$= -\frac{1}{2} \left[x^t J x - 2 x^t J \mu + \mu^t J \mu \right]$$

- Since last term is constant, we get information form

$$p(x)$$
 α $\exp\left(-\frac{1}{2}x^{t}Jx + (J\mu)^{t}x\right)$

- $J\mu$ is called the potential vector
- $J_{i,i}=0$ iff X_i is independent of X_i given χ - $\{X_i,X_i\}$
- Example

$$\mu = \begin{bmatrix} 1 \\ -3 \\ 4 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 4 & 2 & -2 \\ 2 & 5 & -5 \\ -2 & -5 & 8 \end{bmatrix} \qquad J = \begin{bmatrix} 0.3125 & -0.125 & 0 \\ -0.125 & 0.5833 & 0.3333 \\ 0 & 0.3333 & 0.3333 \end{bmatrix}$$

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

- Since $J_{1,3}=0$, x_1 and x_3 are independent given x_2
- Non-zero entries define edges (non-correlations)

Gaussian Markov Random Fields

Follows directly from information form

$$p(\mathbf{x}) \quad \alpha \quad \exp\left[-\frac{1}{2}\mathbf{x}^t J \mathbf{x} + \left(J \mathbf{\mu}\right)^t \mathbf{x}\right]$$

– which is obtained from covariance form with $J=\Sigma^{-1}$

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

- Break-up exponent into two types of terms
 - Using the potential vector $h=J\mu$
 - Terms involving single variable X_i

$$-\frac{1}{2}J_{i,i}x_i^2 + h_i x_i$$

Terms involving pairs of variables X_i , X_j

$$\left[-\frac{1}{2}\left[J_{i,j}x_{i}x_{j}+J_{j,i}x_{j}x_{i}\right]=-J_{i,j}x_{i}x_{j}\right]$$

Called edge potentials (when $J_{i,i}=0$, there is no edge)