Graphical Models

Gaussian Network Models

Siamak Ravanbakhsh

Winter 2018

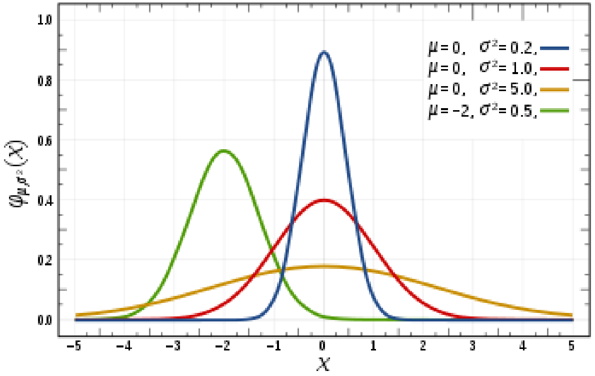
Learning objectives

- multivariate Gaussian density:
  - different parametrizations
  - marginalization and conditioning
  - expression as Markov & Bayesian networks

Univariate Gaussian density

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- motivated by central limit theorem
- max-entropy dist. with a fixed variance



$\mu \in \Re, \sigma^2 > 0$

$\mathbb{E}[X] = \mu,$

$\mathbb{E}[X^2] - \mathbb{E}[X]^2 = \sigma^2$

Multivariate Gaussian

$\mathbf{x} \in \Re^n$ is a column vector *(convention)*

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

$$(2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}}$$

compre to $p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Multivariate Gaussian; sufficient statistics

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

$$\mu = \mathbb{E}[\mathbf{X}]$$

$$\Sigma = \underbrace{\mathbb{E}[\mathbf{X}\mathbf{X}^T]}_{n \times n} - \underbrace{\mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}^T]}_{n \times n}$$ the covariance matrix

$$\Sigma_{i,i} = Var(X_i)$$

$$\Sigma_{i,j} = Cov(X_i, X_j)$$

only captures these two statistics

Multivariate Gaussian; covariance matrix

since $\mathbf{y}^T \Sigma \mathbf{y} = (\mathbf{y}^T \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T]\mathbf{y}) = a^2 > 0$

$\uparrow$

move this expectation out

Multivariate Gaussian; covariance matrix

since $\mathbf{y}^T \Sigma \mathbf{y} = (\mathbf{y}^T \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T]\mathbf{y}) = a^2 > 0$

↑ move this expectation out

$\Sigma \succ 0$

- is symmetric positive definite (PD) $\mathbf{y}^T \Sigma \mathbf{y} > 0 \quad \forall \mathbf{y}; \|\mathbf{y}\| > 0$
- the inverse of a PD matrix is PD
  - the precision matrix $\Lambda = \Sigma^{-1} \succ 0$
- is diagnoalized by orthonormal matrices

Multivariate Gaussian; covariance matrix

since $\mathbf{y}^T \Sigma \mathbf{y} = (\mathbf{y}^T \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T]\mathbf{y}) = a^2 > 0$

↑ move this expectation out

$\Sigma \succ 0$

- is symmetric positive definite (PD) $\mathbf{y}^T \Sigma \mathbf{y} > 0 \quad \forall \mathbf{y}; \|\mathbf{y}\| > 0$
- the inverse of a PD matrix is PD
  - the precision matrix $\Lambda = \Sigma^{-1} \succ 0$
- is diagnoalized by orthonormal matrices

$$\Sigma = QDQ^T$$

D ↓ diagonal

Q ↓

- orthogonal rows & columns of unit norm $QQ^T = Q^T Q = I$
- rotation and reflection

Multivariate Gaussian; covariance matrix

$$\Sigma = QDQ^T$$

D → diagonal (scaling)

Q →
- orthogonal rows & columns of unit norm $QQ^T = Q^TQ = I$
- rotation and reflection

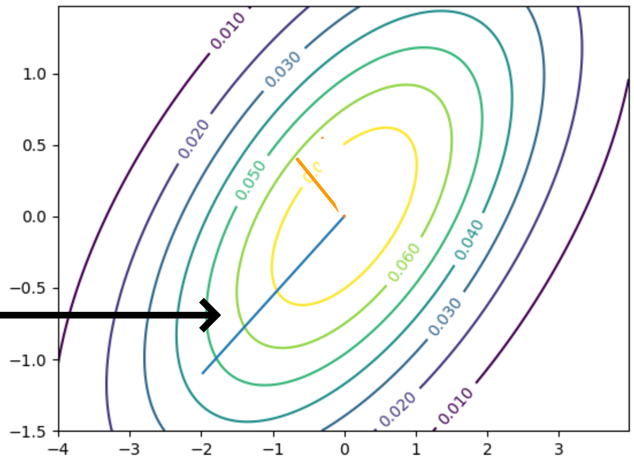
Scaling along axes in some rotated/reflected coordinate system

Multivariate Gaussian; example

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

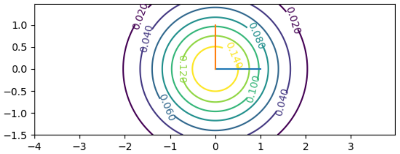
$$\Sigma = \begin{bmatrix} 4, 2 \\ 2, \frac{1}{2} \end{bmatrix} \approx \underbrace{\begin{bmatrix} -.87, -.48 \\ -.48, .87 \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} 5.1, 0 \\ 0, .39 \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} -.87, -.48 \\ -.48, .87 \end{bmatrix}^T}_{Q^T}$$

$$\sqrt{5.1} \times [-.87, .48] \longrightarrow$$



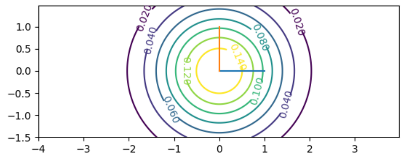
Multivariate Gaussian; from univariates

given n univariate Gaussians $\mathbf{X} \sim \mathcal{N}(0, I)$

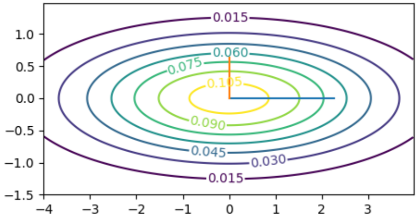


Multivariate Gaussian; from univariates

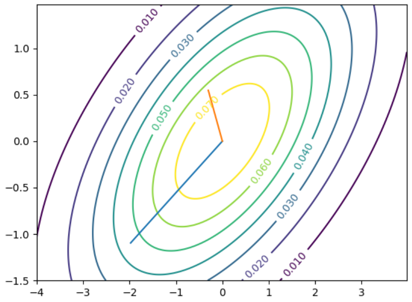
given n univariate Gaussians $\mathbf{X} \sim \mathcal{N}(0, I)$



scale them by $\sqrt{D_{ii}}$ $D^{\frac{1}{2}}\mathbf{X} \sim \mathcal{N}(0, D)$

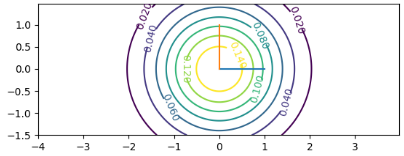


rotate/reflect using Q $QD^{\frac{1}{2}}\mathbf{X} \sim \mathcal{N}(0, QDQ^T) = \mathcal{N}(0, \Sigma)$

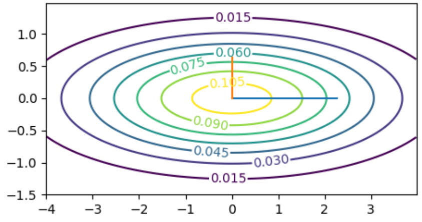


Multivariate Gaussian; from univariates

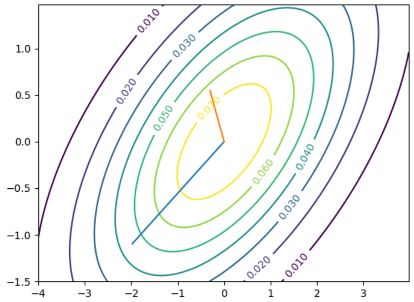
given n univariate Gaussians $\mathbf{X} \sim \mathcal{N}(0, I)$



scale them by $\sqrt{D_{ii}}$ $D^{\frac{1}{2}}\mathbf{X} \sim \mathcal{N}(0, D)$



rotate/reflect using Q $QD^{\frac{1}{2}}\mathbf{X} \sim \mathcal{N}(0, QDQ^T) = \mathcal{N}(0, \Sigma)$



more generally

$$\mathbf{X} \sim \mathcal{N}(\mu, \Sigma) \Rightarrow A\mathbf{X} + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$$

parametrization

moment form *(mean parametrization)*

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

parametrization

moment form *(mean parametrization)*

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

$\eta = \Sigma^{-1}\mu$: local potential

$\Lambda = \Sigma^{-1}$: precision matrix

parametrization

moment form *(mean parametrization)*

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

$\eta=\Sigma^{-1}\mu$: local potential

$\Lambda=\Sigma^{-1}$: precision matrix

information form *(cannonical parametrization)*

$$p(\mathbf{x};\eta,\Lambda)=\sqrt{\frac{|\Lambda|}{(2\pi)^n}}\exp\left(-\frac{1}{2}\mathbf{x}^T\Lambda\mathbf{x}+\eta\mathbf{x}-\frac{1}{2}\eta^T\Lambda\eta\right)$$

parametrization

moment form *(mean parametrization)*

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

$\eta = \Sigma^{-1}\mu$: local potential

$\Lambda = \Sigma^{-1}$: precision matrix

$\mu = \Lambda^{-1}\eta$

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

information form *(cannonical parametrization)*

$$p(\mathbf{x};\eta,\Lambda) = \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \exp\left(-\frac{1}{2}\mathbf{x}^T\Lambda\mathbf{x} + \eta\mathbf{x} - \frac{1}{2}\eta^T\Lambda\eta\right)$$

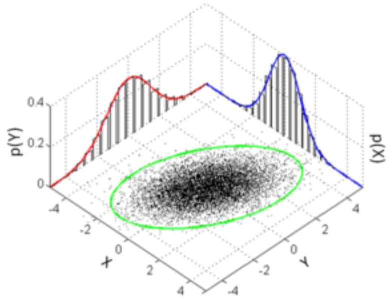
the relationship between the two types goes beyond Gaussians

(exp-family lecture)

marginalization

moment form $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$

is useful for marginalization:



$$\mathbf{X} = [\mathbf{X}_A, \mathbf{X}_B]^T$$

$$\mathbf{X}_A \sim \mathcal{N}(\mu_m, \Sigma_m)$$

$$\mu = [\mu_A, \mu_B]^T$$

$$\Sigma = \begin{bmatrix} \Sigma_{AA}, \Sigma_{AB} \\ \Sigma_{BA}, \Sigma_{BB} \end{bmatrix}$$

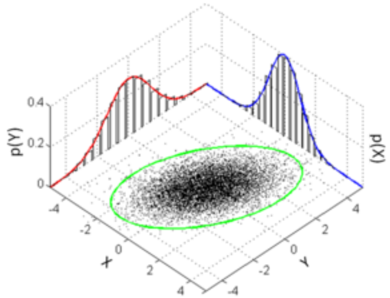
marginalization

moment form $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$

is useful for marginalization:

$$\mu = [\mu_A, \mu_B]^T$$

$$\Sigma = \begin{bmatrix} \Sigma_{AA}, \Sigma_{AB} \\ \Sigma_{BA}, \Sigma_{BB} \end{bmatrix}$$



$$\mathbf{X} = [\mathbf{X}_A, \mathbf{X}_B]^T$$

$$\mathbf{X}_A \sim \mathcal{N}(\mu_m, \Sigma_m)$$

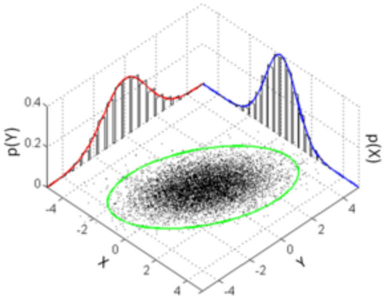
$$\mu_m = \mu_A$$

$$\Sigma_m = \Sigma_A$$

marginalization

moment form $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$

is useful for marginalization:



$$\mathbf{X} = [\mathbf{X}_A, \mathbf{X}_B]^T$$

$$\mathbf{X}_A \sim \mathcal{N}(\mu_m, \Sigma_m)$$

$$\mu_m = \mu_A$$

$$\Sigma_m = \Sigma_A$$

$$\mu = [\mu_A, \mu_B]^T$$

$$\Sigma = \begin{bmatrix} \Sigma_{AA}, \Sigma_{AB} \\ \Sigma_{BA}, \Sigma_{BB} \end{bmatrix}$$

marginalization as a linear transformation: $A = \begin{bmatrix} I_{AA}, & 0 \end{bmatrix}$

$$\mathbf{X} \sim \mathcal{N}(\mu, \Sigma) \Rightarrow A\mathbf{X} \sim \mathcal{N}\left(\mu_A, \Sigma_{AA}\right)$$

Marginal independencies; **moment form**

covariance means dependence & vise versa

$$X_i \perp X_j \mid \emptyset \quad \Leftrightarrow \quad \Sigma_{i,j} = Cov(X_i, X_j) = 0$$

why?

marginalize $\mathcal{N}(\mu, \Sigma)$ to get $\begin{bmatrix} X_i \\ X_j \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} \mu_i \\ \mu_j \end{bmatrix} \begin{bmatrix} \sigma_i^2, 0 \\ 0, \sigma_j^2 \end{bmatrix}) = \mathcal{N}(x_i; \mu_i, \sigma_i^2)\mathcal{N}(x_j; \mu_j, \sigma_j^2)$

Marginal independencies; moment form

covariance means dependence & vise versa

$$X_i \perp X_j \mid \emptyset \quad \Leftrightarrow \quad \Sigma_{i,j} = Cov(X_i, X_j) = 0$$

why?

marginalize $\mathcal{N}(\mu, \Sigma)$ to get $\begin{bmatrix} X_i \\ X_j \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} \mu_i \\ \mu_j \end{bmatrix} \begin{bmatrix} \sigma_i^2, 0 \\ 0, \sigma_j^2 \end{bmatrix}) = \mathcal{N}(x_i; \mu_i, \sigma_i^2)\mathcal{N}(x_j; \mu_j, \sigma_j^2)$

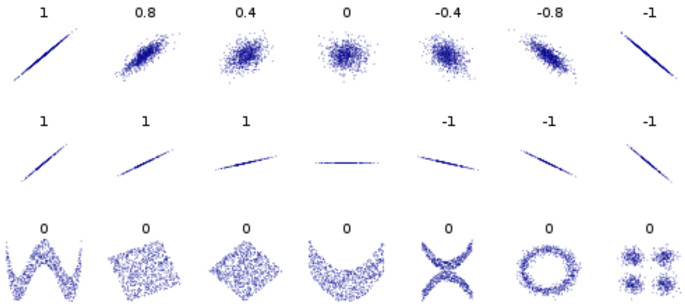


image from wikipedia

Gaussian is special in this sense

correlation: normalized covariance

$$\rho(X_i, X_j) = \frac{Cov(X_i,X_j)}{\sqrt{Var(X_i)Var(X_j)}}$$

Conditional independencies; *information form*

zeros of the precision matrix mean conditional independence

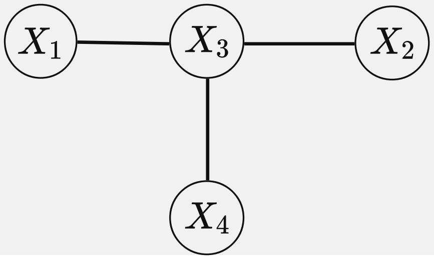
$$X_i \perp X_j \mid \mathbf{X} - \{X_i, X_j\} \quad \Leftrightarrow \quad \Lambda_{i,j} = 0$$

$\Lambda \neq 0$ adjacency matrix in the *Markov network* (**Gaussian MRF**)

$$p(\mathbf{x}; \eta, \Lambda) = \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \exp\left(-\frac{1}{2}\mathbf{x}^T\Lambda\mathbf{x} + \eta\mathbf{x} - \frac{1}{2}\eta^T\Lambda\eta\right)$$

why?

$$\Lambda = \begin{bmatrix} \Lambda_{11}, & 0, & \Lambda_{1,3}, & 0 \\ 0, & \Lambda_{2,2}, & \Lambda_{2,3}, & 0 \\ \Lambda_{3,1}, & \Lambda_{3,2}, & \Lambda_{3,3}, & \Lambda_{3,4} \\ 0, & 0, & \Lambda_{4,3}, & \Lambda_{4,4} \end{bmatrix}$$



Conditional independencies; *information form*

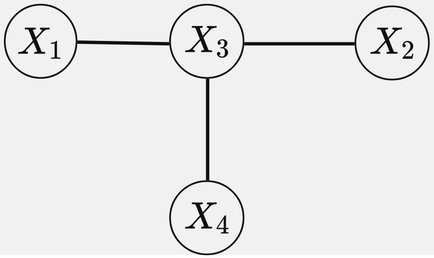
zeros of the precision matrix mean conditional independence

$$X_i \perp X_j \mid \mathbf{X} - \{X_i, X_j\} \quad \Leftrightarrow \quad \Lambda_{i,j} = 0$$

$\Lambda \neq 0$ adjacency matrix in the *Markov network* (**Gaussian MRF**)

$$p(\mathbf{x}; \eta, \Lambda) = \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \exp\left(-\frac{1}{2}\mathbf{x}^T \Lambda \mathbf{x} + \eta \mathbf{x} - \frac{1}{2}\eta^T \Lambda \eta\right)$$

$$\Lambda = \begin{bmatrix} \Lambda_{11}, & 0, & \Lambda_{1,3}, & 0 \\ 0, & \Lambda_{2,2}, & \Lambda_{2,3}, & 0 \\ \Lambda_{3,1}, & \Lambda_{3,2}, & \Lambda_{3,3}, & \Lambda_{3,4} \\ 0, & 0, & \Lambda_{4,3}, & \Lambda_{4,4} \end{bmatrix}$$



why? write it as the product of factors:

corresponding potentials

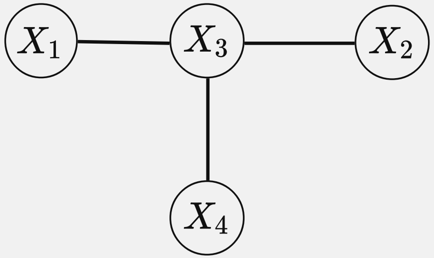
$$\psi_{i,j}(x_i, x_j) = -x_i \Lambda_{i,j} x_j$$

$$\psi_i(x_i) = -\frac{1}{2}\Lambda_{i,i} x_i^2 + \eta_i x_i$$

Gaussian MRF; *information form*

$$p(\mathbf{x};\eta,\Lambda) = \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \exp\left(-\frac{1}{2}\mathbf{x}^T\Lambda\mathbf{x} + \eta\mathbf{x} - \frac{1}{2}\eta^T\Lambda\eta\right)$$

$$\Lambda = \begin{bmatrix} \Lambda_{11}, & 0, & \Lambda_{1,3}, & 0 \\ 0, & \Lambda_{2,2}, & \Lambda_{2,3}, & 0 \\ \Lambda_{3,1}, & \Lambda_{3,2}, & \Lambda_{3,3}, & \Lambda_{3,4} \\ 0, & 0, & \Lambda_{4,3}, & \Lambda_{4,4} \end{bmatrix}$$



corresponding potentials

$$\psi_{i,j}(X_i, X_j) = -x_i\Lambda_{i,j}x_j$$

$$\psi_i(X_i) = -\Lambda_{i,i}x_i^2 + \eta_i x_i$$

Λ should be positive definite

- otherwise the partition function

$$Z = \int_{-\infty}^{\infty} \exp(-\frac{1}{2}\mathbf{x}^T\Lambda\mathbf{x} + \eta^T\mathbf{x})\mathrm{d}\mathbf{x}$$ is not well-defined

Conditioning; information form

marginalization: easy in the moment form

conditioning: easy in the information form

$$\mathbf{X}_A \mid \mathbf{X}_B \sim \mathcal{N}(\eta_{A|B}, \Lambda_{A|B})$$

why?

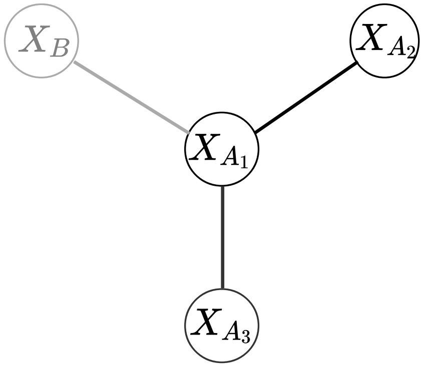
$$\Lambda_{A|B} = \Lambda_{AA}$$

$$\eta_{A|B} = \eta_A + \Lambda_{AB}\mathbf{X}_B$$

$$\mathbf{X} = [\mathbf{X}_A, \mathbf{X}_B]^T$$

$$\eta = [\eta_A, \eta_B]^T$$

$$\Lambda = \begin{bmatrix} \Lambda_{AA}, \Lambda_{AB} \\ \Lambda_{BA}, \Lambda_{BB} \end{bmatrix}$$



Conditioning; information form

marginalization: easy in the moment form

conditioning: easy in the information form

$$\mathbf{X}_A \mid \mathbf{X}_B \sim \mathcal{N}(\eta_{A|B}, \Lambda_{A|B})$$

$$\mathbf{X} = [\mathbf{X}_A, \mathbf{X}_B]^T$$

$$\eta = [\eta_A, \eta_B]^T$$

$$\Lambda = \begin{bmatrix} \Lambda_{AA}, \Lambda_{AB} \\ \Lambda_{BA}, \Lambda_{BB} \end{bmatrix}$$

why?

$$\Lambda_{A|B} = \Lambda_{AA}$$

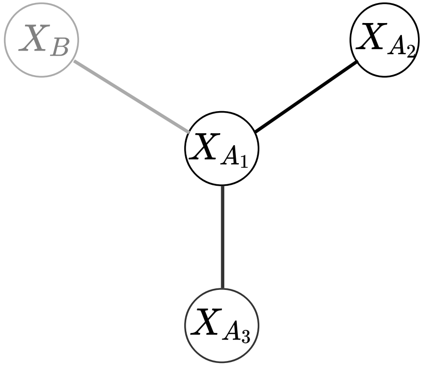
$$\eta_{A|B} = \eta_A + \Lambda_{AB}\mathbf{X}_B$$

not so easy in the moment form!

$$\mathbf{X}_A \mid \mathbf{X}_B \sim \mathcal{N}(\mu_{A|B}, \Sigma_{A|B})$$

$$\Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$$

$$\mu_{A|B} = \mu_A - \Sigma_{AB}\Sigma_{BB}^{-1}(\mathbf{X}_B - \mu_B)$$

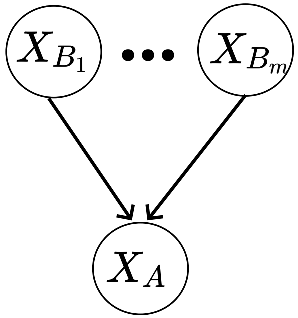


Gaussian Bayesian network

an alternative representation for multivariate Gaussian

linear Gaussian CPD $X_A \mid \mathbf{X}_B \sim \mathbf{w}^T \mathbf{X}_B + \mathcal{N}(\mu_0, \sigma_0^2)$

and $\mathbf{X}_B \sim \mathcal{N}(\mu_B, \Sigma_B)$



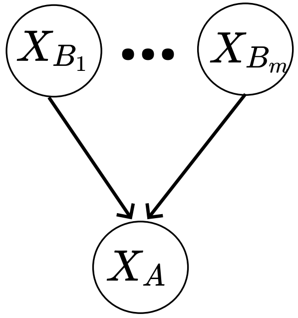
Gaussian Bayesian network

an alternative representation for multivariate Gaussian

linear Gaussian CPD $X_A \mid \mathbf{X}_B \sim \mathbf{w}^T \mathbf{X}_B + \mathcal{N}(\mu_0, \sigma_0^2)$

and $\mathbf{X}_B \sim \mathcal{N}(\mu_B, \Sigma_B)$

joint dist. $\Longleftrightarrow$ conditional form (CPD)

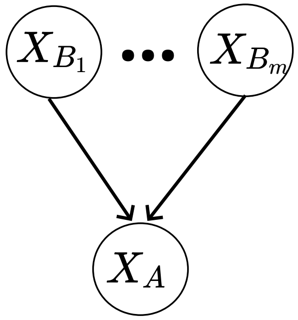


Gaussian Bayesian network

an alternative representation for multivariate Gaussian

linear Gaussian CPD $X_A \mid \mathbf{X}_B \sim \mathbf{w}^T \mathbf{X}_B + \mathcal{N}(\mu_0, \sigma_0^2)$

and $\mathbf{X}_B \sim \mathcal{N}(\mu_B, \Sigma_B)$



joint dist. $\Leftrightarrow$ conditional form (CPD)

$$X_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$$
$$X_1 = \mathbf{w}^T \mathbf{X}_B \sim \mathcal{N}(\mathbf{w}^T \mu_B, \mathbf{w}^T \Sigma_B \mathbf{w})$$

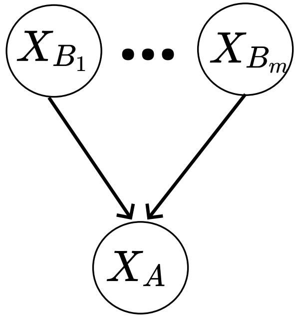
$$X_A = X_0 + X_1 \sim \mathcal{N}(\mu_0 + \mathbf{w}^T \mu_B, \sigma_0^2 + \mathbf{w}^T \Sigma_B \mathbf{w})$$

Gaussian Bayesian network

an alternative representation for multivariate Gaussian

linear Gaussian CPD $X_A \mid \mathbf{X}_B \sim \mathbf{w}^T \mathbf{X}_B + \mathcal{N}(\mu_0, \sigma_0^2)$

and $\mathbf{X}_B \sim \mathcal{N}(\mu_B, \Sigma_B)$



joint dist. $\Leftrightarrow$ conditional form (CPD)

sum of two Gaussian RVs is a Gaussian RV

the pdf of the sum of RVs from the convolution of pdfs

$$X_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$$
$$X_1 = \mathbf{w}^T \mathbf{X}_B \sim \mathcal{N}(\mathbf{w}^T \mu_B, \mathbf{w}^T \Sigma_B \mathbf{w})$$

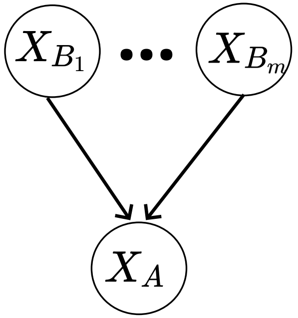
$$X_A = X_0 + X_1 \sim \mathcal{N}(\mu_0 + \mathbf{w}^T \mu_B, \sigma_0^2 + \mathbf{w}^T \Sigma_B \mathbf{w})$$

Gaussian Bayesian network

an alternative representation for multivariate Gaussian

linear Gaussian CPD $X_A \mid \mathbf{X}_B \sim \mathbf{w}^T \mathbf{X}_B + \mathcal{N}(\mu_0, \sigma_0^2)$

and $\mathbf{X}_B \sim \mathcal{N}(\mu_B, \Sigma_B)$



joint dist. $\Leftrightarrow$ conditional form (CPD)

marginal over X_A is $X_A \sim \mathcal{N}(\mu_0 + \mathbf{w}^T \mu_B, \sigma_0^2 + \mathbf{w}^T \Sigma_B \mathbf{w})$

joint dist. is $$(X_A, \mathbf{X}_B) \sim \mathcal{N}\left(\begin{bmatrix} \mu_0 + \mathbf{w}^T \mu_B \\ \mu_B \end{bmatrix}, \begin{bmatrix} \sigma_0^2 + \mathbf{w}^T \Sigma_B \mathbf{w}, & \mathbf{w}^T \Sigma_B \\ \Sigma_B \mathbf{w}, & \Sigma_B \end{bmatrix}\right)$$

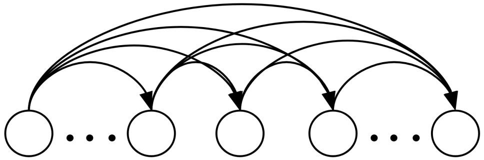
all the other elements follow from the marginals

$Cov(X_A, X_{B,i}) = \sum_j w_j Cov(X_{B,j}, X_{B,i})$

Gaussian Bayesian network

an alternative representation for multivariate Gaussian

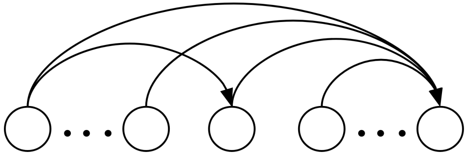
linear Gaussian CPD $X_i \mid Pa_{X_i} \sim \mathbf{w_i}^T Pa_{X_i} + \mathcal{N}(\mu_i, \sigma_i^2)$



worst case: $\mathcal{O}(n)$ parameters per node

- even if Σ is sparse!

generally:

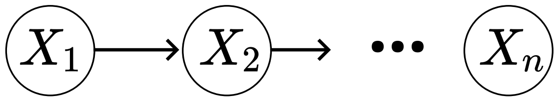


- DAG structure depends on the ordering
- v-structures $X_k \perp X_j \Rightarrow \Sigma_{k,j} = 0$
- d-sepration to find the sparsity of Σ

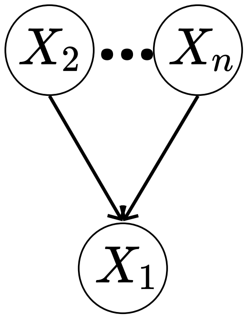
quiz

what is the sparsity patterns of $\Sigma, \Lambda,$ and $\mathbf{w}^i$ in Gaussian BN?

case 1



case 2



Summary

- multivariate Gaussian:
  - mean param. (moment form) Σ, μ
    - useful for marginalization
    - sparsity $\Leftrightarrow$ *marginal* independence
  - canonical param. (information form) Λ, η
    - useful for conditioning
    - sparsity $\Leftrightarrow$ *conditional* independence
- Gaussian Bayesian network (linear Gaussian CPD)
- Gaussian MRF (using information form)