$\boxed{\text{Multivariate Normal Distributions}}$ captures correlation among r.v.s.

<u>Multivariate Distribution</u>: distributions of >1 r.v. (vector of r.v.s).

Consider a random variable $X = [X_1, X_2, \dots, X_d]^T \in \mathbb{R}^d$.
It has pdf $p(x) = p([x_1, \dots, x_d])$, and it satisfies
the axioms of probability 1) $p(x) \geq 0$, 2) $\int_{\mathbb{R}^d} p(x)\,dx = 1$.

$\mathbb{R}^d$ ← d dimensional space.

For any function $h(x)$, its expectation under X is

$$E[h(X)] = \int h(x)\,p(x)\,dx.$$

Mean vector is:

$$\mu = E[X] = \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_d] \end{bmatrix} = \begin{bmatrix} \int x_1 p(x)\,dx \\ \vdots \\ \int x_d p(x)\,dx \end{bmatrix} \in \mathbb{R}^d$$

Covariance matrix is:

$$\Sigma = \mathrm{cov}(X) = [\mathrm{cov}(X_i, X_j)]_{ij} =$$

d x d matrix.

$$= E[(X-\mu)(X-\mu)^T] \quad \text{(definition)}$$

row vector

column vector

$=$ [ ]

matrix

$$
= \begin{bmatrix} var(X_1) & \cdots & cov(X_1, X_d) \\ cov(X_2, X_1) & \ddots & cov(X_2, X_d) \\ \vdots & \ddots & \vdots \\ cov(X_d, X_1) & \cdots & var(X_d) \end{bmatrix}
$$

where $cov(X_i, X_j) = E\left[(X_i - E[X_i])(X_j - E[X_j])\right]$

$$= E[X_i X_j] - E[X_i]E[X_j].$$

$$= E[\mathbf{X}\mathbf{X}^T] - E[\mathbf{X}](E[\mathbf{X}])^T$$

$var(X_i) = cov(X_i, X_i) = E\left[(X_i - E[X_i])^2\right]$

$$= E[X_i^2] - (E[X_i])^2.$$

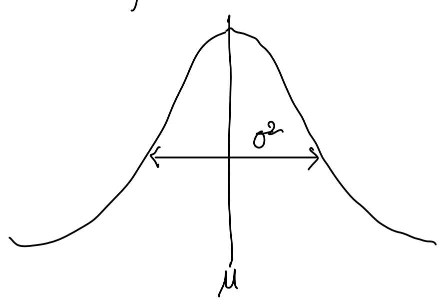
Univariate Normal Distribution:

- Recall that a random variable X on $\mathbb{R}$ is univariate normal $N(\mu, \sigma^2)$ if its pdf is:

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \quad // \text{univariate normal.}$$

with 1) mean $\mu = E(X) = \int x\, p(x)\, dx$ // scalar

2) $\sigma^2 = E\left[(x-\mu)^2\right]$. // scalar



variance = covariance with itself.

Given $\{x_i\}_{i=1}^n \sim N(\mu, \sigma^2)$ we can estimate μ, σ: using MLE:

$$\underbrace{\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i}_{\text{empirical mean}} \qquad \underbrace{\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \hat{\mu})^2}_{\text{empirical variance}}.$$

proof in later pages.

<u>Multivariate Normal Distribution</u> : Extension of univariate normal.

- X becomes a vector of length d, instead by being a scalar.
- A d-dimensional multi-variate distribution is called

$$N(\mu, \Sigma) \text{ if } p(x) = \frac{1}{Z}\exp\left[-\frac{1}{2}\underbrace{(x-\mu)^T}_{1\times d}\underbrace{\Sigma^{-1}}_{d\times d}\underbrace{(x-\mu)}_{d\times 1}\right],$$

$$Z = \sqrt{(2\pi)^d \det(\Sigma)}$$

<u>Mean</u> : $\mu = E[X] = \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_d] \end{bmatrix}$, <u>variance</u> : $\Sigma = E[(X-\mu)(X-\mu)^T]$

$d \times d$: $\begin{bmatrix} & \cdot \\ & \end{bmatrix}$ $\sigma_{ij} = \mathrm{cov}(x_i, x_j)$

Independent Normal Distribution:

Let $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \dots, d$.

Assume $\{X_i\}$ are independent with one other ($X_i \perp X_j$ for $i \neq j$).

Then their concatenation $X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{bmatrix}$ follows a multivariate normal.

$$\mu = E[X] = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix}, \quad \Sigma = E\left[(X-\mu)(X-\mu)^T\right] = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2 \end{bmatrix}$$

all the X_i's are independent from one another.

Independence implies 0 covariance, and vice versa,

$$X_i \perp X_j \iff \text{cov}(X_i, X_j) = 0.$$

Marginal Distributions of Multivariate Normal:

Assume $x \sim N(\mu, \Sigma)$. We can divide the vectors into 2 independent blocks $(x_1 \cap x_2 = 0)$:

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad \text{where}$$

$x_1 \in R^{d_1}$ $\quad\quad$ $\Sigma_{11} \in R^{d_1 \times d_2}$

$x_2 \in R^{d_2}$ $\quad\quad$ $\Sigma_{12} \in R^{d_1 \times d_2}$

$d_1 + d_2 = d$ $\quad\quad$ $\Sigma_{12} = \Sigma_{21}^T$

Then the marginal distribution of x_1 is also Gaussian:

$$x_1 \sim N(\mu_1, \Sigma_{11}).$$

Correlation: $\Sigma_{12} = 0$ iif $x_1 \perp x_2$.

$$P_{x_1}(x_1) = \frac{1}{D_1} \exp\left[-\frac{1}{2}(x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1)\right]$$

$$D_1 = \sqrt{(2\pi)^{d_1} \det(\Sigma_{11})}$$

Conditional Distributions:

Assume $X \sim N(\mu, \Sigma)$. We can divide the vectors into 2 blocks:

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

The conditional distribution $p(X_1 \mid X_2 = a)$ is also Gaussian:

$$X_1 \mid (X_2 = a) \sim N(\mu_{1|2}, \Sigma_{1|2})$$

where 1) $\mu_{1|2} = E[X_1 \mid X_2 = a]$

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

2) $\Sigma_{1|2} = \mathrm{cov}(X_1 \mid X_2 = a)$

$$= \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$$

$\Sigma_{1|2}$ is called the Schur complement of Σ_{22} in Σ

There is a better way (see later pages).

Linear Transformation of Multivariate Normal:

Let $Z \sim N(\mu_0, \Sigma_0)$ // multivariate normal random vector

where $Z = [z_1, \cdots, z_d]^T \in R^d$

Applying linear transformation: $X = AZ + b$,

where $A \in R^{d_1 \times d}$,

$b \in R^{d_1 \times 1}$,

$Z = [Z_1, \cdots, Z_d]^T$ is a set of independent standard normal random variables, that is, $Z_i \sim N(0, 1)$ and $Z_i \perp Z_j$ for $i \neq j$.

then X is also a multivariate normal, $\in R^{d_1 \times 1}$.

$X \sim N(\mu_X, \Sigma_X)$ where:

Mean is:

$$\mu_X = E[X] = E[AZ + b]$$

$$= A\underbrace{E[z]}_{\mu_0} + b \quad \text{// linearity of expectation.}$$

$$= \boxed{A\mu_0 + b.}$$

Variance:

$$\Sigma_X = E\Big[\underbrace{(X-\mu_X)}_{\text{column vector}}\underbrace{(X-\mu_X)^T}_{\text{row vector}}\Big]$$

$$= E\Big[(AZ+\cancel{b}-A\mu_0-\cancel{b})(AZ+\cancel{b}-A\mu_0-\cancel{b})^T\Big]$$

$$= E\Big[A(Z-\mu_0)(A(Z-\mu_0))^T\Big]$$

$$= E\Big[A(Z-\mu_0)(Z-\mu_0)^T A^T\Big]$$

$$= A\,E\Big[(Z-\mu_0)(Z-\mu_0)^T\Big]A^T$$

$$= \underbrace{\boxed{A\Sigma_0 A^T}}_{d_1 \times d_1 \text{ matrix.}}$$

Summary:

Given $X = AZ+b$, where $Z \sim N(\mu_0, \Sigma_0)$,

the $X \sim N(\mu_X, \Sigma_X)$ where

$$\mu_X = A\mu_0 + b$$

$$\Sigma_X = A\Sigma_0 A^T.$$

$d_1 \times d_1$ $d_1 \times d$ $d \times d$ $d \times d_1$

Multivariate Normal and PCA:

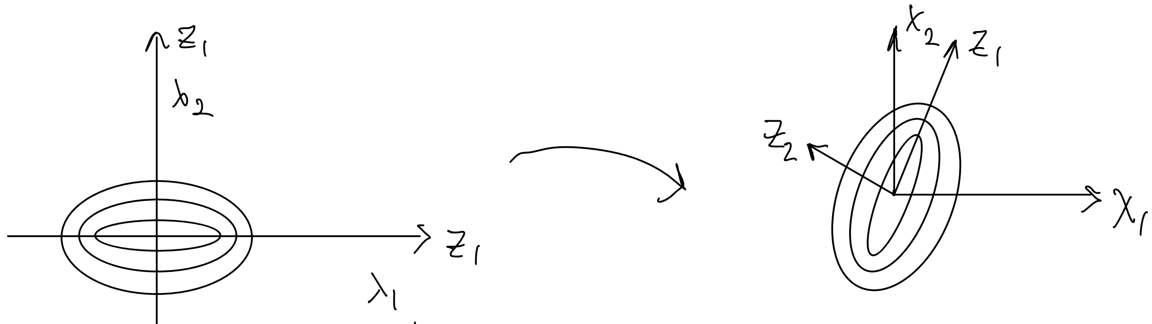
we can construct a probabilistic interpretation of PCA using multivariate normal distributions

Assume $Z \sim N(0, \Sigma)$, $\Sigma = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ 0 & \ddots & \lambda_d \end{bmatrix}$

$X = Az + b.$

Observed $\{x^{(i)}\}_{i=1}^{n}$, iid. Estimate Z?

Geometrically:



$(\lambda_1 > \lambda_2)$

variance of z_1 variance of z_2

$$\underbrace{\Sigma_X}_{\text{eigendecompostion}} = A \Sigma A^T$$

diagonal matrix

MLE of Multivariate Normal:

- Given an observation $\{x_i\}_{i=1}^n$ drawn iid from $N(\mu, \Sigma)$, we can show that the MLE of μ and Σ equals the empirical mean and variance.

$$\max_{\mu,\Sigma} \sum_{i=1}^n \log p(x_i \mid \mu, \Sigma) \qquad \text{// log-likelihood}$$

$$= \sum_{i=1}^n \log\left(\frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}} \exp\left[\frac{-1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right]\right)$$

$$= \underbrace{-n\log\left((2\pi)^{d/2}\right)}_{\text{constant}} - \frac{n}{2}\log\left(\det\Sigma\right) - \frac{1}{2}\sum_{i=1}^n\left[(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right]$$

For μ: $-\frac{1}{2}\left(n\mu^T\Sigma^{-1}\mu - 2\left(\sum_{i=1}^n x_i\right)^T\Sigma^{-1}\mu + \text{const}\right)$

$$\xRightarrow{\text{derivative}} \hat{\mu} = \left(n\cancel{\Sigma^{-1}}\right)^{-1}\left(\cancel{\Sigma^{-1}}\left(\sum_{i=1}^n x_i\right)\right) \Rightarrow \boxed{\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i}$$

For Σ: Let $Q = \Sigma^{-1}$. Optimize Q ($d \times d$ matrix).

$$l(Q) = \frac{n}{2}\log\left(\det Q\right) - \frac{1}{2}\sum_{i=1}^n (x_i-\hat{\mu})^T Q (x_i-\hat{\mu})$$

$$\underset{d\times d}{\nabla_Q l(Q)} = \frac{n}{2}Q^{-1} - \frac{1}{2}\sum_{i=1}^n (x_i-\mu)(x_i-\mu)^T$$

If Q is a number, $\nabla\left(\log\det Q\right) = \nabla\left(\log Q\right) = Q^{-1}$

$$\Rightarrow \boxed{Q^{-1} = \frac{1}{n}\sum_{i=1}^n (x_i-\mu)(x_i-\mu)^T = \text{empirical covariance.}}$$

Natural Form of Multinormal Distribution

- **Standard Form:** $p(x) = \frac{1}{Z} \exp\left[\frac{-1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right]$

$$= \frac{1}{D} \exp\Big(\underbrace{\tfrac{1}{2} x^T \Sigma^{-1} x}_{\text{quadratic term}} - \underbrace{\mu^T \Sigma^{-1} x}_{\text{linear term}} + \mu^T \Sigma^{-1} \mu \Big) \quad \text{// expanded form.}$$

- **Natural/Information Form:** $p(x) = \frac{1}{C} \exp\left[\frac{-1}{2} x^T Q x + b^T x \right]$

$$N(\mu, \Sigma) \rightarrow \bar{N}(\underset{\text{mean}}{b}, \underset{\text{inverse variance/covariance.}}{Q})$$

- **Natural parameters:**
  - $Q = \Sigma^{-1}$, (inverse covariance matrix / precision matrix)
  - $b = \Sigma^{-1} \mu = Q\mu$ (?)
  - $C = \exp\left[-\frac{1}{2} b^T Q^{-1} b \right] / Z$ (normalization constant.)

The natural form provides a convenient form to study conditional distributions and conditional independence.

In contrast, standard form is convenient for marginal distributions and marginal independence.

Conditional Distribution via Natural Form:

- The natural form makes it easy to derive conditional distributions.
- For $X \sim N(\mu, \Sigma)$,

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

$Q = \Sigma^{-1}$

$$b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \qquad Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$$

- Then the conditional distribution is:

$$X_1 \mid X_2 \sim \bar{N}\Big(b_1 - \underbrace{Q_{12} X_2}_{\text{depends on } X_2\,!}\;,\; \underbrace{Q_{11}}_{\text{does not depend on } X_2,}\Big).$$

Equivalently: $Cov(X_1 \mid X_2) = Q_{11}^{-1}$.

We can then convert back to moment parameters following matrix identities (derivation not provided):

- $\Sigma_{1|2} = Cov(X_1 \mid X_2) = Q_{11}^{-1} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$
- $\mu_{1|2} = E[X_1 \mid X_2]$

$$= Q_{11}^{-1}(b_1 - Q_{12} X_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2).$$

Proof for conditional probability:

assume we only have X_1 and X_2.

$$p(\overbrace{[X_1, X_2]}) \propto \exp\left[\frac{-1}{2}X^T Q X + b^T X\right], \quad X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

$$\propto \exp\left[\frac{-1}{2}\underbrace{\left(Q_{11}X_1^2 + 2Q_{12}X_1X_2 + \cancel{Q_{22}X_2^2}\right)}_{\text{quadratic term}} \cdots \underbrace{+ b_1X_1 + \cancel{b_2X_2}}_{\text{linear term}}\right]$$

treat as constant (fixed).

$$p(X_1 \mid X_2) \propto \exp\left(\underbrace{\frac{-1}{2}Q_{11}X_1^2}_{\text{quadratic term}} + \underbrace{(b_1 - Q_{12}X_2)}_{\text{linear term}}X_1\right)$$

// dropped terms related to only X_2

$$\boxed{\sim N(b_1 - Q_{12}X_2, \; Q_{11})}.$$

$\therefore$ using natural form allows us to derive conditional distributions very conveniently.

Theorem: Assuming X is multivariate normal with inverse covariance matrix Q.

$$\boxed{\text{If } q_{ij} = 0 \iff X_i \perp X_j \mid X_{\neg ij}}$$

Independence and Conditional Independence:

Note: $\sigma_{ij} = \mathrm{cov}(x_i, x_j)$

- Covariance matrix $\Sigma = [\sigma_{ij}]_{ij}$ measures (marginal) independence.

  $\sigma_{ij} = 0$, $x_i \perp x_j \iff p([x_i, x_j]) = p(x_i)\, p(x_j)$

- Precision matrix $Q = [q_{ij}]_{ij}$ measures conditional independence.

  $$q_{ij} = 0 \iff x_i \perp x_j \mid \underbrace{x_{\neg ij}}_{\text{all variables except } i \text{ and } j:}.$$

  all variables except i and j:

  $x_{\neg ij} = x_{[d]/\{i,j\}}$.

  Example: If x_1, x_2, x_3:

  $x_1 \perp x_2 \mid x_{\neg 12} =$

  $x_1 \perp x_2 \mid x_3$

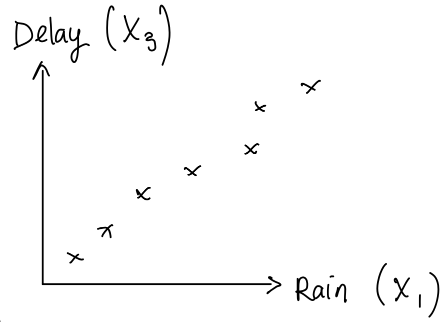
Definition of conditional independence.

$$x_i \perp x_j \mid x_{\neg ij} \iff p([x_i, x_j] \mid x_{\neg ij}) = p(x_i \mid x_{\neg ij})\, p(x_j \mid x_{\neg ij})$$

∴ The zero elements of the covariance matrix encodes the marginal independence,
The zero elements of inverse covariance encodes the conditional independence.

Example of conditional independence

$$\underset{\text{Rain}}{X_1} \longrightarrow \underset{\text{Traffic}}{X_2} \longrightarrow \underset{\text{Delay}}{X_3}$$



$\sigma_{13} = \mathrm{cov}(X_1, X_3)$ = positive nonzero.

$q_{13} = a_{13}\,\mathrm{cov}(X_1, X_3 \mid X_2) = 0$

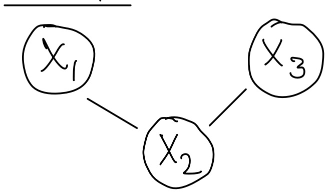
if I fix traffic constant, then rain and delay are not correlated.

$$a_{ij} = \frac{-1}{(q_{ii} q_{jj} - q_{ij}^2)}$$

$$X_1 \perp X_3 \mid X_2 \;\not\Longrightarrow\; X_1 \perp X_3$$
$$\not\Longleftarrow$$

Conditional independence does not imply marginal independence.
Marginal independence does not imply conditional independence.

Example:



$X_2 = X_1 + X_3$ (marginal independence of X_1 and X_3):

$X_3 = X_2 - X_1$ (if X_2 is fixed, then X_1 and X_3 are not independent).

Marginal Independence does not imply conditional independence (and vice versa)

Proof: Assuming 3 variables:

$$P([X_1, X_2, X_3]) \sim \bar{N}(b, Q)$$

column 1:2   row 1:2

$$P([X_1, X_2] \mid X_3) \sim \bar{N}\Big(\underbrace{b_{1:2} - Q_{1:2,3} X_3}_{\tilde{b}},\ \underbrace{Q_{1:2,1:2}}_{\tilde{Q}}\Big).$$

$$Q = \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix} \qquad \mathrm{cov}([X_1\ X_3] \mid X_3)^{-1} = \underbrace{Q_{1:2,1:2}}_{\begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix}}$$

$$P([X_1\ X_2] \mid X_3) \propto \exp\Big[\frac{-1}{2}\big(x_1 q_{11} x_1 + 2x_1 q_{12} x_2 + q_{21} x_2^2\big) + \tilde{b}_1 x_1 + \tilde{b}_2 x_2\Big]$$

If $q_{12} = 0$:

$$\propto \underbrace{\exp\Big(\frac{-1}{2} x_1 q_{11} x_1 + \tilde{b}_1 x_1\Big)}_{P(X_1 \mid X_3)} \cdot \underbrace{\exp\Big(\frac{-1}{2} x_2 q_{22} x_2 + \tilde{b}_2 x_2\Big)}_{P(X_2 \mid X_3)}$$

// separated out X_1 and X_2.

Term depends only on X_1.   Term depends only on X_2.

$$\propto P(X_1 \mid X_3) \cdot P(X_2 \mid X_3) \Rightarrow X_1 \perp X_2 \mid X_3.$$

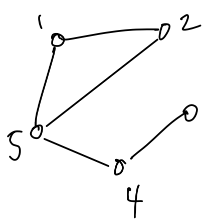
Correlation:

$$\mathrm{Corr}(X_i, X_j \mid X_{\neg ij}) \overset{\text{def}}{=} \frac{\mathrm{cov}(X_i, X_j \mid X_{\neg ij})}{\sqrt{\mathrm{var}(X_i \mid X_{\neg ij})\,\mathrm{var}(X_i \mid X_{\neg ij})}}$$

$$= \frac{-q_{ij}}{\sqrt{q_{ii}\, q_{jj}}}.$$

Gaussian Graphical Models / Gaussian Markov random fields (MRF)

- $N(\mu, \Sigma)$ with spare precision matrix $Q = \Sigma^{-1}$
- use graph to represent the dependency structure of random variables.



$$
\begin{array}{c|ccccc|}
 & 1 & 2 & 3 & 4 & 5 \\
\hline
1 & \times & & 0 & 0 & \\
2 & & \times & 0 & 0 & \\
3 & 0 & 0 & \times & & 0 \\
4 & 0 & 0 & & \times & \\
5 & & & 0 & & \times \\
\hline
\end{array} = Q
$$

- we can read Markov (conditional independence) properties from the graph.

Algorithm:

1) Each random variable is represented by a node.
2) If the precision element of pairs of nodes is non-zero, add an edge btwn the nodes.

Properties:

① If i, j are not directly connected $\Rightarrow x_i \perp x_j \mid x_{\neg ij}$.

For example:

x_1 and x_2 are NOT conditionally independent, in above graph.

② If conditioning on a node's neighbors being fixed, then the node is independent from all nodes outside the neighbors.
The neighbors of the node = Markov blanket.

For example;

$x_1 \perp (x_3, x_4) \mid \underbrace{(x_2 \mid x_5)}$.

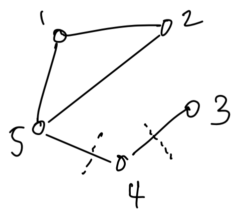
X_2 and X_5 = markov blanket.

③ If 2 nodes are connected through some path, being able to find a bottomneck node implies conditional independence, given the bottomneck node;

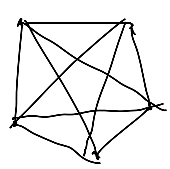
$X_A \perp X_B \mid X_C$ if C is a bottleneck between A and B.

Example:

$X_2 \perp X_3 \mid X_4$



If you look at covariance graph: it does not reveal the dependencies.



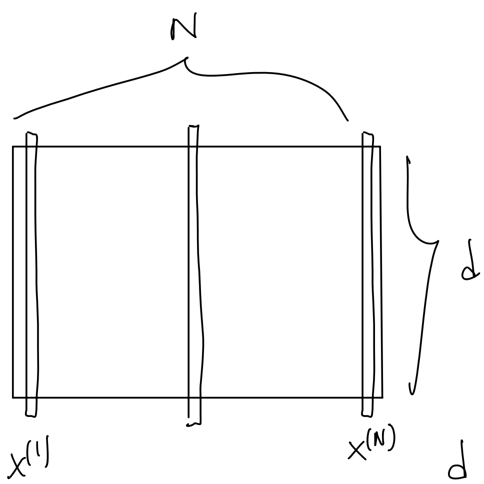
Learning Gaussian Markov Random Fields:

- Given observation $D = \{x^k\}$ drawn from $N(\mu, \Sigma)$.
- Assume the precision matrix $Q = \Sigma^{-1}$ is sparse.
- Goal: Estimate Q and its graph structure.

Graphical Lasso:

- Estimate sparse Q from data $D = \{x^k\}$
- Idea: use MLE with some penalty to encourage sparsity.

$D =$



$\sim N(\mu, \Sigma)$.

$d \times N$

Estimate $Q = \Sigma^{-1}$, with Q being sparse (many elements $= 0$).

1) Assume $\mu = 0$ (if not, centralize the data).

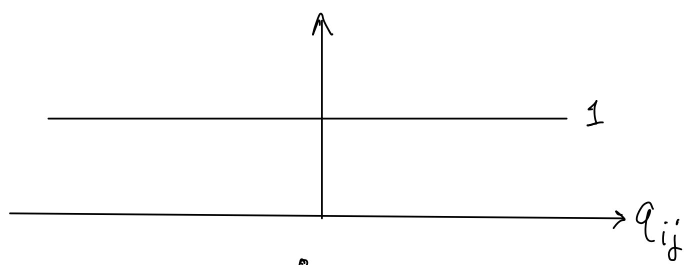
2) $$\max_Q \underbrace{\log P(D \mid Q)}_{\text{log likelihood function}} - \underbrace{\lambda \Phi(Q)}_{\text{penalty to encourage sparsity.}}$$

Want to max the likelihood but minimize the edges. How?

Define $\Phi(Q) \equiv \sum_{i \neq j} \mathbb{I}(q_{ij} \neq 0) = \#$ of edges.

$$= \sum_{i \neq j} q_{ij}^0 \triangleq \|Q\|_0$$ (L_0 norm).

Zero norm.



This is a L_0 regularized MLE. Difficult.

$0^0 = 0$

$x^0 = 1$ (if $x \neq 0$).

Fix:

Relax it by using L_1 norm, to approximate L_0 norm.

L1 norm:

$\Phi(Q) = \sum_{i \neq j} |q_{ij}| \triangleq \|Q\|_1$ // replace indicator function with absolute value.

This is a convex function now.

$$\Rightarrow \max_{Q} \underbrace{\log P(D \mid Q)}_{\shortparallel} - \lambda \|Q\|_1 \quad \text{(Graphical Lasso)}.$$

$$- \sum_{k=1}^{N} \log P(x^k \mid Q)$$

$$= \sum_{k=1}^{N} \left(-\frac{1}{2} (x^k)^T Q (x^k) + \frac{1}{2} \log(\det Q) \right)$$