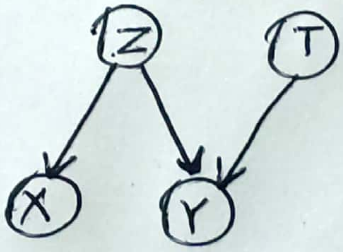
Incomplete Data



Data

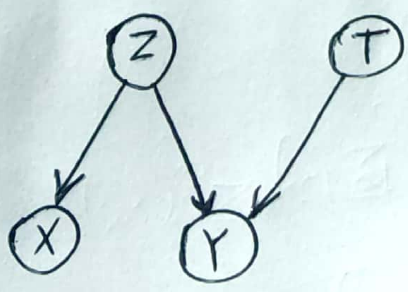
Complete (I)	Incomplete (II)
X^1, Y^1, Z^1, T^1	X^1, Y^1, Z^1, T^1
X^2, Y^2, Z^2, T^2	$X^2, Y^2, ?, T^2$
X^3, Y^3, Z^3, T^3	$X^3, Y^3, Z^3, ?$
X^4, Y^4, Z^4, T^4	$X^4, Y^4, ?, ?$

(I) complete: $\underbrace{L(\theta)}_{P_r(\text{data}|\theta)} = \prod_{i=1}^{4} P_\theta(X^i, Y^i, Z^i, T^i) = \prod_{i=1}^{4} P_{\theta_1}(Z^i)\, P_{\theta_2}(T^i)\, P_{\theta_3}(Y^i|Z^i,T^i)\, P_{\theta_4}(X^i|Z^i)$

(II) incomplete: $P_r(\text{data}|\theta) = P_\theta(X^1,Y^1,Z^1,T^1)\; P_\theta(X^2,Y^2,T^2)\; P_\theta(X^3,Y^3,Z^3)\; P_\theta(X^4,Y^4)$

$$L(\theta) = P_\theta(X^1,Y^1,Z^1,T^1)\left(\sum_Z P_\theta(X^2,Y^2,Z,T^2)\right)\left(\sum_T P_\theta(X^3,Y^3,Z^3,T)\right)\left(\sum_Z \sum_T P(X^4,Y^4,Z,T)\right)$$

$$= \cancel{P(X^4)} \sum_Z \sum_T P_{\theta_1}(Z)\, P_{\theta_2}(T)\, P_{\theta_3}(Y^4|Z,T)\, P_{\theta_4}(X^4|Z)$$



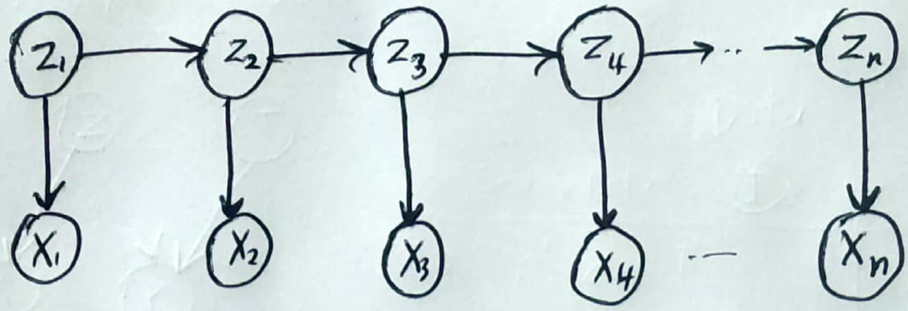
Data

X	Y	Z	T
X^1	Y^1	?	?
X^2	Y^2	?	?
⋮			
X^m	Y^m	?	?

$\underbrace{X\ Y}_{\text{observed variables}}$ $\underbrace{Z\ T}_{\text{latent (hidden) variables}}$

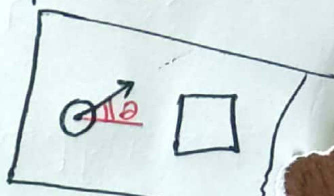
Latent variables

- present in model
- absent in data



$$P(X_1 - X_n, Z_1 - Z_n) = \prod_{i=2}^{n} P(Z_i|Z_{i-1}) \prod_{i=1}^{n} P(X_i|Z_i)$$

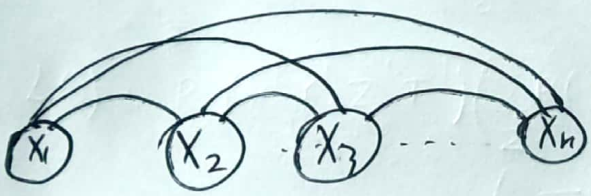
$$P(X_1 - X_n) = \sum_{Z_n} \cdots \sum_{Z_2}\sum_{Z_1} \prod_{i=2}^{n} P(Z_i|Z_{i-1}) \prod_{i=1}^{n} P(X_i|Z_i)$$



Z_i's not obsered $\Rightarrow$ all variables $X_1 - X_n$ are dependent

$Z = (P_x, P_y, \theta)$

withaut latent variables



$\Rightarrow$ form a clique of size $\underline{n}$

latent variables

$z^i \in \mathbb{R}^p$

$x^i \in \mathbb{R}^n$

directed (most common)

$$P(Z,X) = \frac{1}{Z(\theta)} \Phi(Z,X)$$

undirected (less common)

$$P_\theta(X,Z) = P_{\theta_1}(Z)\, P_{\theta_2}(X|Z)$$

prior $\leftarrow P_{\theta_1}(Z)$, $P_{\theta_2}(X|Z)$ are given (easy)

$P(X,Z)$ $p(Z)P(X|Z)$ easy!

$$P_\theta(X,Z) = P(Z|X)\; P(X) \rightarrow \text{needed for leaning}$$

posterior, hard, data likehood

$P(X)$ $p(Z|X)$ Difficult!

data $(X^1, X^2, \ldots, X^m)$

$$\cancel{ll(\theta) = \log \sum_{i=1}^{m} P_\theta(X^i)}$$

$$ll(\theta) = \sum_{i=1}^{m} \log P_\theta(X^i)$$

$$= \sum_{i=1}^{m} \log \sum_{Z} P_\theta(X^i, Z)$$

$$= \sum_{i=1}^{m} \log \underbrace{\sum_{Z} P_{\theta_1}(Z)\, P_{\theta_2}(X|Z)}_{\theta_1, \theta_2 \text{ are entangled}}$$

Solution 1

$$\frac{\partial\, ll(\theta)}{\partial \theta_1}, \frac{\partial\, ll(\theta)}{\partial \theta_2} \Rightarrow \text{gradient ascent}$$

$$\frac{\partial}{\partial \theta_1} = \sum_{i=1}^{m} \frac{\sum_Z \frac{\partial}{\partial \theta_1} P_{\theta_1}(Z)\, P_{\theta_2}(X|Z)}{\sum_Z P_{\theta_1}(Z)\, P_{\theta_2}(X|Z)}$$

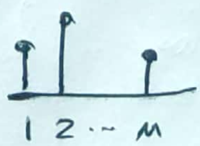
~~Solution~~ $P(X|Z)$ simple

$$P(X) = \int_{\Sigma} P(X, Z)\, d\theta = \int_{\Sigma} P(X|Z)\, P(Z)\, d\theta$$

Gaussian Mixture $Z \in \{1, 2, \ldots, M\}$

$Z \in \{1, 2, \ldots, M\}$

- $P(Z=k) = \pi_k$



$$P(X|Z) = \mathcal{N}(X; \mu_Z, \sigma_Z^2)$$

$$P(X|Z=k) = \mathcal{N}(X, \mu_k, \sigma_k)$$

parameters $\theta = \{(\pi_1, \mu_1, \sigma_1), (\pi_2, \mu_2, \sigma_2), \ldots, (\pi_M, \mu_M, \sigma_M)\}$

$$P(X) = \sum_{k=1}^{M} Pr(Z=k)\, P(X|Z=k) = \sum_{k=1}^{M} \pi_k\, \mathcal{N}(X | \mu_k, \sigma_k)$$