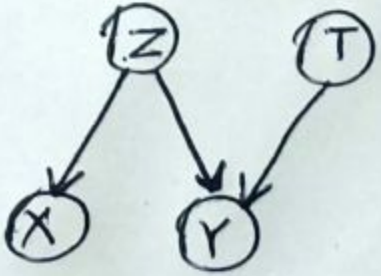
Probabilistic Graphical Models

Lecture 25,26

Learning with incomplete data
Latent Variables Models
Expectation Maximization

Learning with incomplete data

K. N. Toosi
University of Technology



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	$X^4, Y^4, ?, ?$

$Pr(data|\theta)$

4

Learning with incomplete data

Incomplete (II)

X^1, Y^1, Z^1, T^1

$X^2, Y^2, ?, T^2$

$X^3, Y^3, Z^3, ?$

$X^4, Y^4, ?, ?$

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

(II) incomplete: $Pr(data|\theta) = P_\theta(X^1, Y^1, Z^1, T^1)\ P_\theta(X^2, Y^2, T^2)\ P_\theta(X^3, Y^3, Z^3)\ P_\theta(X^4, Y^4)$

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

Learning with incomplete data

K. N. Toosi

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

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

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

Incomplete II

X^1, Y^1, Z^1, T^1

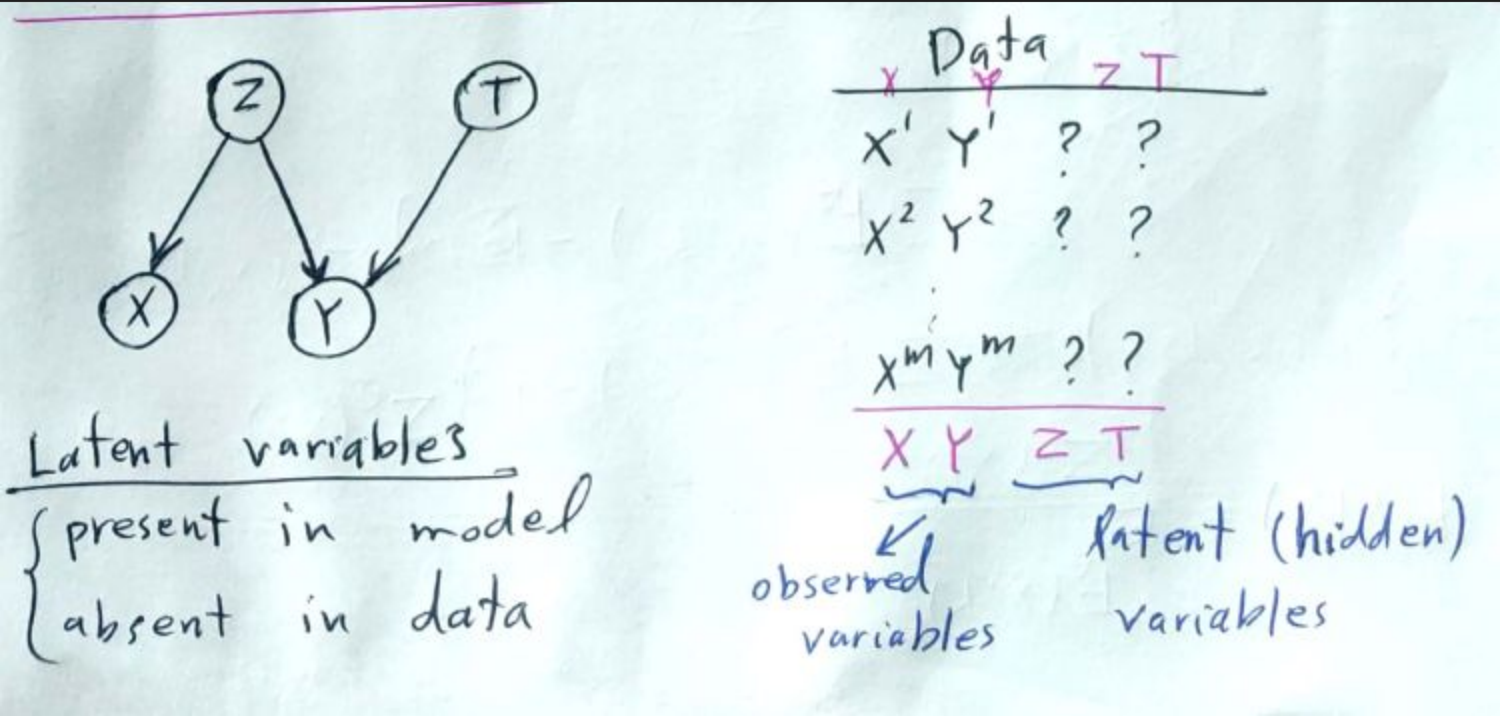
$X^2, Y^2, ?, T^2$

$X^3, Y^3, Z^3, ?$

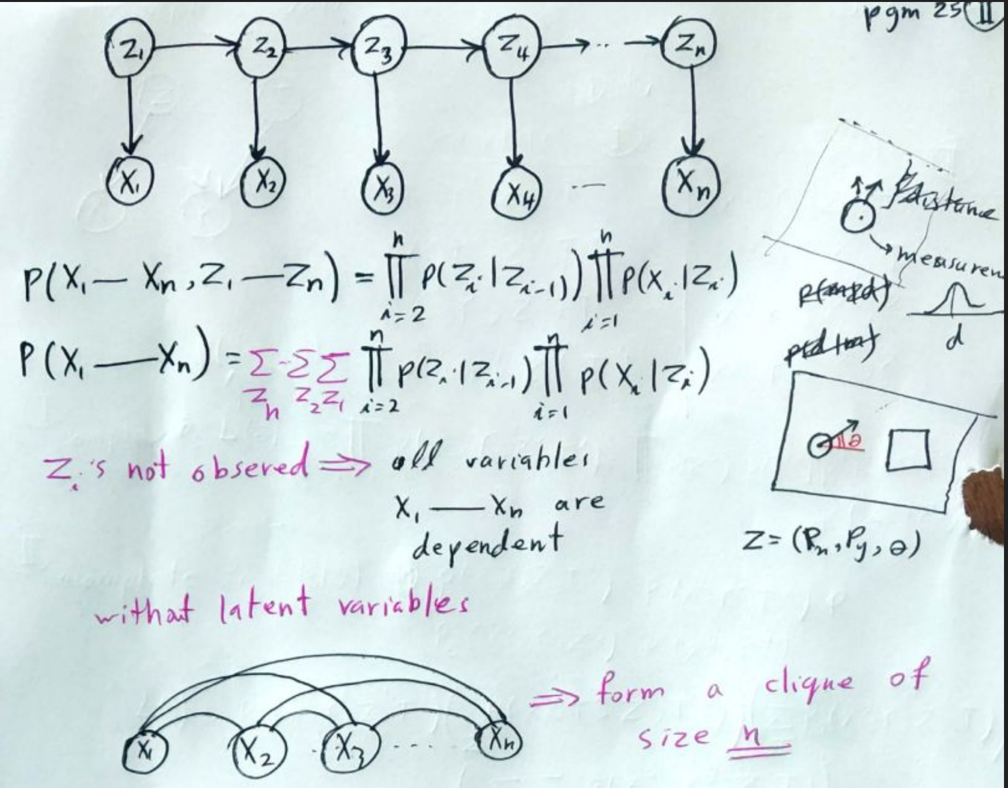
$X^4, Y^4, ?, ?$

Latent Variables

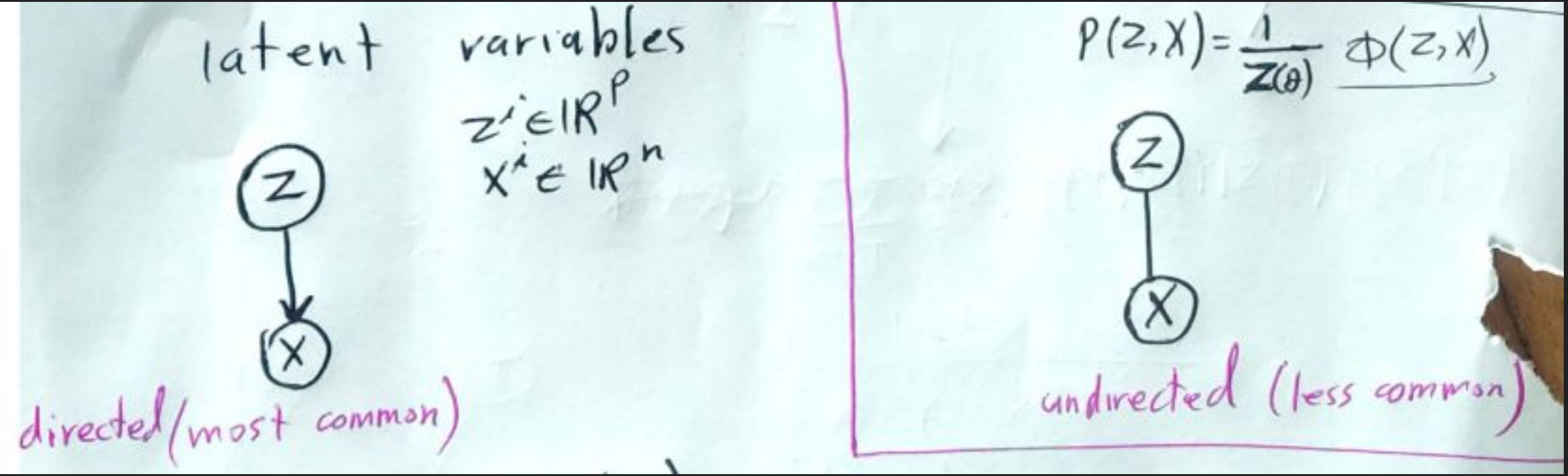
K. N. Toosi
University of Technology



Latent Variables: Example



Latent Variables



Latent Variables

latent variables

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

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

directed (most 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)

$$P_\theta(x,z) = \underbrace{p(z|x)}_{\text{posterior}}\; p(x) \rightarrow \text{needed for learning}$$

$p(z|x)$, $p(x)$ → hard

$p(x)$ → data likehood

$P(x,z)$ $p(z)\,p(x|z)$ easy!

$p(x)$ $p(z|x)$ Difficult!

Incomplete data introduces complexities

K. N. Toosi
University of Technology

pg

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

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

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

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

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

Learning with Incomplete data

Solution 1

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

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

Example Gaussian Mixture Models

K. N. Toosi
University of Technology

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

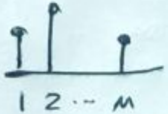
$$p(X) = \int_{\Sigma} p(X,Z)\, d\theta = \int_{\Sigma} p(X|Z)\, p(Z)\, d\theta$$

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

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

$$Z \rightarrow X$$

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

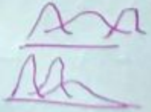


~~p(Z)~~ $p(X|Z) = \mathcal{N}(X; \mu_Z, \sigma_Z^2)$

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

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

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



How to learn latent variable models?

K. N. Toosi
University of Technology

$\boxed{P_\theta(X)}$ $\quad x^1, x^2, \ldots, x^m$

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

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

How to learn latent variable models?

$$ll(\theta_0) \leqslant ll(\theta_1) \leqslant ll(\theta_2) \leqslant ll(\theta_3) \leqslant \cdots \leqslant ll(\theta_t) \leqslant \cdots$$

$\boxed{\theta_t} \rightarrow$ current θ

We dont have access to z!

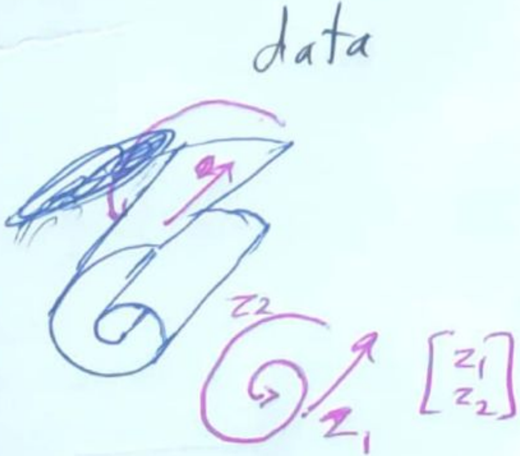
heuristic:

$$z^i = \underset{z}{\operatorname{argmax}}\, P_{\theta_t}(z \mid x^i)$$ Maximize → posterior distribution

OR

$$z^i \sim P_{\theta_t}(z \mid x^i)$$ take a sample

$$\theta_{t+1} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{m} \log P_\theta(x^i; z^i)$$



How to learn latent variable models?

How to compute the posterior?

$$P_\theta(X, Z) = P_\alpha(X \mid Z)\, P_\beta(Z) \qquad \theta = (\alpha, \beta)$$

$$\theta_t = (\alpha_t, \beta_t)$$

$$P_{\theta_t}(Z \mid X^i) = \frac{P_{\theta_t}(X^i, Z)}{\sum_Z P_{\theta_t}(X^i, Z)} = \frac{P_{\alpha_t}(X^i \mid Z)\, P_{\beta_t}(Z)}{\sum_Z P_{\alpha_t}(X^i \mid Z)\, P_{\beta_t}(Z)}$$

Expectation Maximization (EM)

Expectation-Maximization

Compute $P_{\theta_t}(Z|x^i)$ for all z. Expectation step (E-step)

$$\theta_{t+1} = \arg\max_{\theta} \sum_{i=1}^{m} E_{P_{\theta_t}(z|x^i)}\{\log P_{\theta}(x^i; z)\}$$

→ expected log-likelihood

$$\theta_{t+1} = \arg\max_{\theta} \sum_{i=1} \sum_{z} P_{\theta_t}(z|x^i) \log P_{\theta}(x^i, z)$$

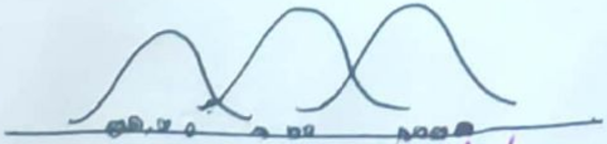
Maximization-Step ~~M-S~~ (M-step)

EM (Expectation-Maximization): Alternate betweem E-step and M-step.

Can Prove: $ll(\theta_0) \le ll(\theta_1) \le ll(\theta_2) \le \cdots$

Example: Mixture of Gaussians

pgm 26 (I)



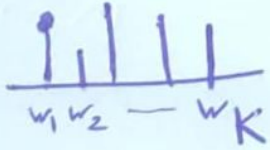
→ traing data $x^1, x^2, \dots, x^m$

Mixture of Gaussians

$$P(x) = \sum_{k=1}^{K} w_k \, N(x \mid \mu_k, \sigma_k^2) \qquad \sum_{k=1}^{K} w_k = 1$$

$$p(x,z) = \underbrace{p(x \mid z)}_{N(x;\, \mu_k, \sigma_k^2)} \; p(z)$$

$$P(z) = Pr(Z=z) = w_z$$



Example: Mixture of Gaussians

K. N. Toosi
University of Technology

$$ll(\theta) = \cancel{\log} \sum_{i=1}^{m} \log P(x^i) = \sum_{i=1}^{m} \log \sum_{k=1}^{K} \underbrace{Pr(Z=k)}_{w_k} N(X \mid \mu_k, \sigma_k^2)$$

$$\theta = (\{\mu_k\}, \{\sigma_k^2\}, w_1, \dots, w_K)$$

$$\theta^t \qquad P(Z \mid x^i) = \frac{P(x^i \mid z)\, p(z)}{\sum_{z} p(x^i \mid z)\, p(z)}$$

E-step

find $\quad Pr(Z=k \mid x^i) = \dfrac{N(x^i; \mu_k^t, \sigma_k^{2\,t})\, w_k^t}{\sum_{k'=1}^{K} N(x^i; \mu_{k'}^t, \sigma_{k'}^{2\,t})\, w_{k'}^t} = \alpha_{ik}^t$

for $x^1, x^2, \dots, x^m$

for $k = 1, 2, \dots, K$

Example: Mixture of Gaussians

K. N. Toosi
University of Technology

M-step $E_{p_{\theta_t}(z|x^i)}\left\{\sum_{i=1}^{m}\log p_\theta(x^i,z)\right\} = \sum_{i=1}^{m}\sum_{z} p_{\theta^t}(z|x^i)\log p_\theta(x^i,z)$

$$= \sum_{i=1}^{m}\sum_{z=1}^{K} p_{\theta^t}(z|x^i)\left[\log p_\theta(x^i|z) + \log p_\theta(z)\right]$$

$$\theta^{t+1} = \arg\max_\theta \underbrace{\sum_{i=1}^{m}\sum_{k=1}^{K}\alpha_{ik}^{t}\left[\log N(x^i|\mu_k,\sigma_k) + \log w_k\right]}$$

$\theta \swarrow$ $\{\mu_k\},\{\sigma_k\},\{w_k\}$

$$\frac{\partial}{\partial w_j}\sum_{k=1}^{K}\underbrace{\left(\sum_{i=1}^{m}\alpha_{ik}^{t}\right)}_{\beta_k}\log w_k \;\underline{+\lambda\left(\sum w_k - 1\right)}$$

$$\boxed{w_j = \frac{\beta_j}{\sum_k \beta_k}}$$

$\frac{\beta_j}{w_j} = -\lambda \Rightarrow \frac{w_j}{\beta_j}$ cost. $\quad \sum w_j = 1$

$\sum \frac{\beta_j}{-\lambda} = 1$

Limitations of EM

$P_\theta(x,z)$ data $x^1, x^2, \dots, x^m$ pgm 2

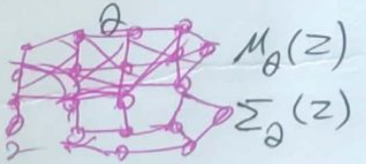
$= P_{\theta_1}(z)\, P_{\theta_2}(x|z)$

$P(z) = \mathcal{N}(z; \mu=0, \Sigma=I)$

$P_\theta(x|z) = \mathcal{N}(x; \mu_\theta(z), \Sigma_\theta(z))$

deep neural network

z



Assume θ is known

take a sample from $P_\theta(x,z)$

z^i

Image of

Limitations of EM: Computing the posterior

take a sample from $P_\theta(x,z)$

z^i  Image of noise

1- $z^i \sim p(z) = \mathcal{N}(z; \mu=0, \Sigma=I)$

2- Feed z^i to neural nets to get $\mu_\theta(z^i)$, $\Sigma_\theta(z^i)$

3- sample $x^i \sim \mathcal{N}(x; \mu_\theta(z), \Sigma_\theta(z))$

$$p(x) = \sum_z p_\theta(x,z) = \sum_z p(x|z)\, p(z)$$

$$= \sum_z \mathcal{N}(x | \mu_\theta(z), \Sigma_\theta(z))\, \mathcal{N}(z; 0, I)$$

Very hard to compute the posterior! $p(z|x^i)$

Data $x^1, x^2, \ldots, x^m$