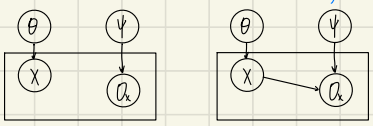
# · Likelihood of Data and Observation Models

Let $X = \{X_1 \cdots X_n\}$ be some set of random variable, and $O_X = \{O_{X_1} \cdots O_{X_n}\}$ be their observability variable

the observability model is a joint distribution $P_{missing}(X, O_X) = P(X) \cdot P_{missing}(O_X | X)$, where $P(X)$ is parameterized parameter $\theta$ and $P_{missing}(O_X | X)$ is parameterized by $\psi$




random missing values




deliberate missing values

Define a new set of variable $Y = \{Y_1 \cdots Y_n\}$, $Val(Y_i) = Val(X_i) \cup \{?\}$

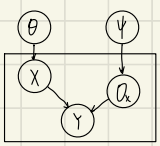
the actual observation is $Y$, which is a deterministic function of $X$ and $O_X$

$$Y_i = \begin{cases} X_i & O_{X_i} = 0' \\ ? & O_{X_i} = 0^* \end{cases}$$

"?" represents a missing value

eg. $X_i \sim Bernoulli(\theta)$ $O_{X_i} \sim Bernoulli(\psi)$

$P(Y=1) = \theta\psi$ $P(Y=0) = (1-\theta)\psi$ $P(Y=?) = 1-\psi$




Given $Y=?$ $O_X$ must be $0^*$

Given $Y \neq ?$ $O_X$ must be $0'$, $X = Y$

$(X \perp O_X | Y)$ even though knowing $Y$ activates the v-structure (context-specific independence)

$$L(\theta, \psi; D) = \psi^{M[1]+M[0]}(1-\psi)^{M[?]} \cdot \theta^{M[1]}(1-\theta)^{M[0]}$$

$$\max_{\theta,\psi} L(\theta, \psi; D) \Rightarrow \begin{cases} \max_{\psi} \psi^{M[1]+M[0]}(1-\psi)^{M[?]} \\ \max_{\theta} \theta^{M[1]}(1-\theta)^{M[0]} \end{cases}$$

$$\psi^* = \frac{M[1]+M[0]}{M[1]+M[0]+M[?]} \qquad \theta^* = \frac{M[1]}{M[1]+M[0]}$$

] separable

↓

The independence reveals in the likelihood function

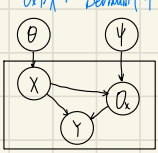
↑

eg. $X \sim Bernoulli(\theta)$ $O_X | x^0 \sim Bernoulli(\psi_{O_X|x^0})$ $O_X | x^1 \sim Bernoulli(\psi_{O_X|x^1})$

$P(Y=1) = \theta \cdot \psi(O_X|x^1)$

$P(Y=0) = (1-\theta)\psi(O_X|x^0)$

$P(Y=?) = \theta \cdot (1-\psi(O_X|x^1)) + (1-\theta)(1-\psi(O_X|x^0))$




knowing $Y=0$, we can infer $X$ from $O_X$

$\therefore X \not\perp O_X | Y$

$$L(\theta, \psi(O_X|x^0), \psi(O_X|x^1); D) = \theta^{M[1]} \cdot \psi(a|x^0)^{M[1]} \cdot (1-\theta)^{M[0]} \cdot \psi(a|x^1)^{M[0]} \cdot \left[\theta \cdot (1-\psi(O_X|x^1)) + (1-\theta)(1-\psi(a|x^0))\right]^{M[?]}$$

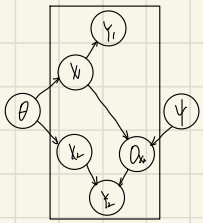
# · Decoupling of Observation Mechanism

A missing data model $P_{missing}$ is missing completely at random (MCAR) if $P_{missing} \models (X \perp O_X)$

In the case of MCAR, the likelihood function decomposes as a product

MCAR is sufficient but not necessary for decomposition of likelihood function

eg. always observe the first toss, decide whether or not hide the 2nd toss based on the first toss




$P(X_1=x_1^0, Y_1=y_1^0) = (1-\theta_{x_1}) \cdot (1-\theta_{x_2}) \cdot \psi(O_{x_2}|x_1=x_1^0)$

$P(X_1=x_1^0, Y_1=y_1^1) = (1-\theta_{x_1}) \cdot \theta_{x_2} \cdot \psi(O_{x_2}|x_1=x_1^0)$

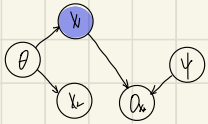
$P(X_1=x_1^0, Y_1=?) = (1-\theta_{x_1}) \cdot [1-\psi(O_{x_2}|x_1=x_1^0)]$

$\vdots$

$$L(\theta,\psi;D) = \left[(1-\theta_{x_1})(1-\theta_{x_2})\psi(O_{x_2}|x_1=x_1^0)\right]^{M[y_1^0,y_2^0]} \left[(1-\theta_{x_1})\theta_{x_2}\psi(O_{x_2}|x_1=x_1^0)\right]^{M[y_1^0,y_2^1]} \left[(1-\theta_{x_1})\cdot[1-\psi(O_{x_2}|x_1=x_1^0)]\right]^{M[y_1^0,y_2^?]}$$
$$\left[\theta_{x_1}(1-\theta_{x_2})\psi(O_{x_2}|x_1=x_1^1)\right]^{M[y_1^1,y_2^0]} \left[\theta_{x_1}\theta_{x_2}\psi(O_{x_2}|x_1=x_1^1)\right]^{M[y_1^1,y_2^1]} \left[\theta_{x_1}\cdot[1-\psi(O_{x_2}|x_1=x_1^1)]\right]^{M[y_1^1,y_2^?]}$$

$$= (1-\theta_{x_1})^{M[y_1^0]} \cdot \theta_{x_1}^{M[y_1^1]} \cdot (1-\theta_{x_2})^{M[y_2^0]} \cdot \theta_{x_2}^{M[y_2^1]} \cdot \psi(O_{x_2}|x_1=x_1^0)^{M[y_1^0,y_2^1]+M[y_1^0,y_2^0]} \left[1-\psi(O_{x_2}|x_1=x_1^0)\right]^{M[y_1^0,y_2^?]} \cdot \psi(O_{x_2}|x_1=x_1^1)^{M[y_1^1,y_2^1]+M[y_1^1,y_2^0]} \left[1-\psi(O_{x_2}|x_1=x_1^1)\right]^{M[y_1^1,y_2^?]}$$

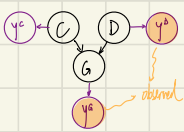
$(O_{x_2} \perp X_2 \mid X_1)$




the conditional independence helps to decouple $P(X)$ from $P(O_x|X)$

let $y$ be a tuple of observations. Variables $X$ are partitioned into 2 sets.

$X_{obs}^y = \{X_i : y_i \neq ?\}$ (observed vars) and $X_{hidden}^y = \{X_i : y_i = ?\}$ (unobserved vars)




$X_{obs}^y = \{D, G\}$

$X_{hidden}^y = \{C\}$

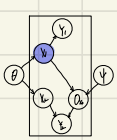
A missing data model $P_{missing}$ is missing at random (MAR) if

$\forall$ observations $y$ with $P_{missing}(y) > 0$, $\forall x_{hidden}^y \in Val(X_{hidden}^y)$, $P_{missing} \models (o_x \perp x_{hidden}^y \mid x_{obs}^y)$

where $o_x$ are specific values of observation variables given $y$

(given observed vars, whether or not other vars are observed does not depend on the actual value of those vars)

eg.




give the value of the first toss.

whether or not the second toss is observed does not depend on the actual value of second toss

$P_{missing}(x_{hidden}^y \mid x_{obs}^y, o_x) = P_{missing}(x_{hidden}^y \mid x_{obs}^y)$

given observed vars, the observation pattern does not give additional informations about unobserved vars

$$P_{missing}(y) = \sum_{x_{hidden}^y} P(x_{obs}^y, x_{hidden}^y) \cdot P_{missing}(o_x \mid x_{obs}^y, x_{hidden}^y)$$
$$= \sum_{x_{hidden}^y} P(x_{obs}^y, x_{hidden}^y) \cdot P_{missing}(o_x \mid x_{obs}^y)$$
$$= \underbrace{P_{missing}(o_x \mid x_{obs}^y)}_{\text{depends only on } \psi} \cdot \underbrace{P(x_{obs}^y)}_{\text{depends only on } \theta}$$

# The Likelihood Function

Given a BN $G$ over a set of var $\mathcal{X}$. each instance has a different set of observed vars

$O[m]$ and $o[m]$ are observed vars and values of $m$th instance

$H[m]$ is the missing vars in $m$th instance

$$\ell(\theta;D) = \log L(\theta;D) = \sum_{m=1}^{M} \log P(o[m]|\theta)$$

eg 

fully observed: $L(\theta;D) = \theta_x^{M[x]} (1-\theta_x)^{M[\bar{x}]} \theta_{y|x}^{M[x,y]} (1-\theta_{y|x})^{M[x,\bar{y}]} \theta_{y|\bar{x}}^{M[\bar{x},y]} (1-\theta_{y|\bar{x}})^{M[\bar{x},\bar{y}]}$

$L(\theta;D)$ is log-concave, has a close-form for global maxi;

$\ell(\theta;D)$ can be decomposed into function of each $\theta$

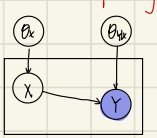
partially observed: $D = \{(x=?, y=y'), (x=x[2], y=y[2]), \cdots (x=x[M], y=y[M])\}$

$$L(\theta;D[1]) = P(x=x^0, y=y') + P(x=x^1, y=y')$$
$$= (1-\theta_x)\theta_{y|x^0} + \theta_x \cdot \theta_{y|x^1}$$

$$L(\theta;D[2,\cdots M]) = \theta_x^{M[x^1]} (1-\theta_x)^{M[x^0]} \theta_{y|x^1}^{M[x^1,y^1]} (1-\theta_{y|x^1})^{M[x^1,y^0]} \theta_{y|x^0}^{M[x^0,y^1]} (1-\theta_{y|x^0})^{M[x^0,y^0]}$$

$$\ell(\theta;D) = \ell(\theta;D[1]) + \ell(\theta;D[2,\cdots M])$$
$$= \underbrace{\log[(1-\theta_x)\theta_{y|x^0} + \theta_x \cdot \theta_{y|x^1}]}_{\text{can not be decomposed; and not necessarily log-concave}} + \ell(\theta;D[2,\cdots M])$$

each partially observed data introduce a log-plus term.



when $y$ is observed and $X$ is not

$\theta_x \to X \to Y \leftarrow \theta_{y|x}$ is an active trail

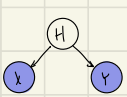
$\theta_x \not\perp \theta_{y|x}$

$$P(\theta|y) = \sum_{x \in val(x)} P(\theta, x|y) = \sum_x \frac{P(\theta)P(x,y|\theta)}{P(y)} = \frac{P(\theta)}{P(y)} \cdot \sum_x P(x,y|\theta) = \frac{P(\theta)}{P(y)} \cdot L(\theta,y)$$

can not be decomposed as $f_1(\theta_x) \cdot f_2(\theta_{y|x})$

lose the property of parameter independence

if $\theta_x$ is known, $P(\theta_{y|x}|y) = C \cdot \log[(1-\theta_x)\theta_{y|x^0} + \theta_x \theta_{y|x^1}]$ the local decomposability is lose

eg 

$X, Y$ observed

$$P(x,y) = \sum_{h \in val(H)} P(h)\, P(x|h)\, P(y|h)$$

$$L(\theta;D) = \prod_{(x,y) \in val(x,y)} \left[\sum_h P(h)\, P(x|h) \cdot P(y|h)\right]^{M[x,y]}$$

cannot be decomposed as $P(x|H)$ and $P(x|H)$ (as in MLE)

In the presence of partially observed data, we lose all important properties of likelihood function

log-concaving, close-form solution, decomposability of $\ell(\theta;D)$

## · Identifiability

ability to identify uniquely a model from data

eg randomly choose one from 2 biased coins, then toss

$(\theta_H, \theta_{X|H}, \theta_{X|T})$

$P(\theta; D) = P(X)^{M[1]} \cdot (1 - P(X))^{M[0]}$

$P(X) = \theta_H \cdot \theta_{X|H} + (1-\theta_H)\,\theta_{X|T}$ } $P(X)$ is a weighted average of $\theta_{X|H}$ and $\theta_{X|T}$

different choices of $(\theta_H, \theta_{X|H}, \theta_{X|T})$ can gives the same likelihood

Given the observation (no matter how many instances), we cannot hope to recover a unique set of parameters

Suppose we have a parametric model with $\theta \in \Theta$ that defines a distribution $P(X|\theta)$

A choice of $\theta$ is identifiable if there is no $\theta' \neq \theta$ s.t. $P(X|\theta) = P(X|\theta')$

A model is identifiable if $\theta$ is identifiable $\forall\, \theta \in \Theta$

## · Gradient Ascent

Let $\mathcal{B}$ be a Bayesian network over $\mathcal{X}$ that induces $P(\mathcal{X})$

let $o$ be a tuple of observations for some variables

$$\frac{\partial P(o)}{\partial P(x|u)} = \frac{1}{P(x|u)} P(x,u,o) \quad \text{if } P(x|u) > 0,\ x \in Val(X),\ u \in Val(Pa_X)$$

consider a full assignment $\xi$.

$$P(\xi) = \prod_i P(\xi\langle X_i\rangle \mid \xi\langle Pa_{X_i}\rangle)$$

$$\frac{\partial P(\xi)}{\partial P(x|u)} = \begin{cases} \prod_{X_i \neq X} P(\xi\langle X_i\rangle \mid \xi\langle Pa_{X_i}\rangle) = \dfrac{P(\xi,\theta)}{P(x|u,\theta)} & \text{if } \xi\langle X, Pa_X\rangle = (x,u) \\ 0 & \text{otherwise} \end{cases}$$

Consider a partial assignment $o$

$$P(o) = \sum_{\xi\langle O\rangle = o} P(\xi)$$

$$\frac{\partial P(o)}{\partial P(x|u)} = \sum_{\xi\langle O\rangle = o} \frac{\partial P(\xi,\theta)}{\partial P(x|u,\theta)} = \sum_{\xi\langle O\rangle = o,\ \xi\langle X\rangle = x,\ \xi\langle Pa_X\rangle = u} \frac{1}{P(x|u,\theta)} P(\xi,\theta)$$

$P(x|u) \neq 0$

$$= \frac{1}{P(x|u)} P(o, x, u)$$

Let $G$ be a Bayesian network over $\mathcal{X}$ $D = \{o[1] \cdots o[M]\}$ be a partially observed dataset

$$\frac{\partial l(\theta; D)}{\partial P(x|u)} = \frac{1}{P(x|u)} \sum_{m=1}^{M} P(x, u \mid o[m], \theta)$$

$$\frac{\partial l(\theta; D)}{\partial P(x|u)} = \frac{\partial}{\partial P(x|u)} \cdot \sum_{m=1}^{M} \log P(o[m]; \theta) = \sum_{m=1}^{M} \frac{\partial}{\partial P(x|u)} \log P(o[m]; \theta)$$

$$= \sum_{m=1}^{M} \frac{1}{P(o[m]; \theta)} \frac{\partial P(o[m]; \theta)}{\partial P(x|u)} = \sum_{m=1}^{M} \frac{1}{P(o[m]; \theta)} \cdot \frac{1}{P(x|u)} \cdot P(o[m], x, u, \theta)$$

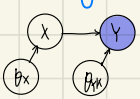
$$= \frac{1}{P(x|u)} \sum_{m=1}^{M} P(x, u \mid o[m]; \theta)$$

```
def compute_gradient(G, θ, D):
    for each φ(X) → φ(X) ∈ G:
        for each x_i, u_i ∈ val(X, U)
            M[x_i, u_i] = 0
    for m = 1 ... M:
        run clique tree calibration on ⟨G, θ⟩ with evidence o[m]
        for each φ(X) → φ(X) ∈ G:
            for each x_i, u_i ∈ val(X, U)
                M[x_i, u_i] += P(x_i, u_i | o[m])
    for each φ(X) → φ(X) ∈ G:
        for each x_i, u_i ∈ val(X, U)
            ∇ P(x_i|u_i) = 1/θ_{x_i|u_i} M[x_i, u_i]
    return {∇ P(x_i|u_i)}
```

can use similar data structure as Factor

// $\sum_{m=1}^{M} P(x,u \mid o[m]; \theta)$ for each $x, u$

## · EM Algorithm




Y is observed.

$$\sum_m \log P(y[m]; \theta) = \sum_m \log \sum_{x[m] \in Val(X)} P(y[m], x[m]; \theta)$$

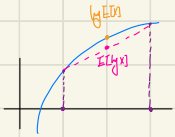
$$= \sum_m \log \sum_{x[m] \in Val(X)} Q_m(x[m]) \frac{P(y[m], x[m]; \theta)}{Q_m(x[m])}$$

construct a probability $Q_m(x)$ $\forall m = 1, \dots, M$

$$= \sum_m \log E_{x[m] \sim Q_m}\left[\frac{P(y[m], x[m]; \theta)}{Q_m(x[m])}\right]$$

$$\geq \sum_m E_{x[m] \sim Q_m}\left[\log \frac{P(y[m], x[m]; \theta)}{Q_m(x[m])}\right]$$

$\log(\cdot)$ is concave, Jenson's Inequality




$$= \sum_m \sum_{x[m] \in Val(X)} Q_m(x[m]) \log \frac{P(y[m], x[m]; \theta)}{Q_m(x[m])}$$

$$\ell(\theta; D) \geq \sum_m \sum_{x[m] \in Val(X)} Q_m(x[m]) \log \frac{P(y[m], x[m]; \theta)}{Q_m(x[m])}$$

a lower bound for $\ell(\theta; D)$


$\ell(\theta; D)$ is concave when $P(X, Y; \theta)$ is log-concave in $\theta$.

$$\sum_m \sum_{x[m] \in Val(X)} Q_m(x[m]) \log \frac{P(y[m], x[m]; \theta)}{Q_m(x[m])}$$

$$\ell(\theta; D) = \sum_m \log E_{x[m] \sim Q_m}\left[\frac{P(y[m], x[m]; \theta)}{Q_m(x[m])}\right] = \sum_m E_{x[m] \sim Q_m}\left[\log \frac{P(y[m], x[m]; \theta)}{Q_m(x[m])}\right]$$

if $\frac{P(y[m], x[m]; \theta)}{Q_m(x[m])}$ is a constant $\forall m = 1, \dots, M$

$$Q_m(x[m]) = \frac{1}{Z_m} P(y[m], x[m]; \theta)$$

$$\sum_{x[m] \in Val(X)} Q_m(x[m]) = \frac{1}{Z_m} \sum_{x[m]} P(y[m], x[m]; \theta) = \frac{1}{Z_m} P(y[m]; \theta) = 1 \qquad Z_m = P(y[m]; \theta)$$

$$Q_m(x[m]) = \frac{P(y[m], x[m]; \theta)}{P(y[m]; \theta)} = P(x[m] \mid y[m]; \theta) \quad \forall x[m] \in Val(X)$$

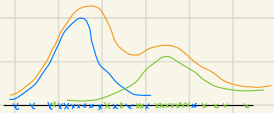
then maximize the lower bound.

$$\theta := \arg\max_\theta \sum_m \sum_{x[m] \in Val(X)} Q_m(x[m]) \log \frac{P(y[m], x[m]; \theta)}{Q_m(x[m])}$$

eg. mixture of Gaussians

$X \sim \text{Bernoulli}(\theta_X)$

$(Y|X=j) \sim \mathcal{N}(\mu_j, \Sigma_j)$



$$\ell(\theta) = \sum_{m=1}^{M} \log P(y^{[m]};\theta) = \sum_{m=1}^{M} \log \sum_{x^{[m]}\in\{0,1\}} P(x^{[m]},y^{[m]};\theta) = \sum_{m=1}^{M} \log \sum_{x^{[m]}} Q_m(x^{[m]}) \frac{P(x^{[m]},y^{[m]};\theta)}{Q_m(x^{[m]})} = \sum_{m=1}^{M} \log E_{x^{[m]}\sim Q_m}\left[\frac{P(x^{[m]},y^{[m]};\theta)}{Q_m(x^{[m]})}\right]$$

$$\geq \sum_{m=1}^{M} E_{x^{[m]}\sim Q_m}\left[\log \frac{P(x^{[m]},y^{[m]};\theta)}{Q_m(x^{[m]})}\right] = \sum_{m=1}^{M} \sum_{x^{[m]}\in\{0,1\}} Q_m(x^{[m]}) \log \frac{P(x^{[m]},y^{[m]};\theta)}{Q_m(x^{[m]})}$$

E-Step:

$$Q_m(x^{[m]}) = P(x^{[m]}|y^{[m]};\theta) = \frac{P(x^{[m]};\theta)\cdot P(y^{[m]}|x^{[m]};\theta)}{P(y^{[m]};\theta)}$$

$$Q_m(x^1) = \frac{1}{Z}\left(\theta_X \cdot \mathcal{N}(y^{[m]};\mu_1,\Sigma_1)\right) = \frac{\theta_X}{Z}\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_1|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(y^{[m]}-\mu_1)^T\Sigma_1^{-1}(y^{[m]}-\mu_1)\right)$$

$$Q_m(x^0) = \frac{1}{Z}\left((1-\theta_X)\mathcal{N}(y^{[m]};\mu_0,\Sigma_0)\right) = \frac{1-\theta_X}{Z}\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_0|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(y^{[m]}-\mu_0)^T\Sigma_0^{-1}(y^{[m]}-\mu_0)\right)$$

M-Step:

$$\max_{\theta} \sum_{m=1}^{M}\sum_{x^{[m]}} Q_m(x^{[m]})\cdot\log\frac{P(x^{[m]},y^{[m]};\theta)}{Q_m(x^{[m]})}$$

$$= \max_{\theta} \sum_{m=1}^{M}\sum_{x^{[m]}\in\{0,1\}} Q_m(x^{[m]})\cdot\left[\log\frac{\theta_{x^{[m]}}}{(2\pi)^{\frac{n}{2}}|\Sigma_{x^{[m]}}|^{\frac{1}{2}}} - \frac{1}{2}(y^{[m]}-\mu_{x^{[m]}})^T\Sigma_{x^{[m]}}^{-1}(y^{[m]}-\mu_{x^{[m]}})\right]$$

max $\mu$:

$$\nabla_{\mu_j} = \sum_{m=1}^{M} Q_m(j)\cdot -\frac{1}{2}\cdot 2\,\Sigma_j^{-1}(y^{[m]}-\mu_j)\cdot -1$$

$$= \Sigma_j^{-1}\left[\sum_{m=1}^{M} Q_m(j)y^{[m]} - \sum_{m=1}^{M} Q_m(j)\,\mu_j\right] = 0$$

$$\mu_j = \frac{\sum_{m=1}^{M} Q_m(j)\,y^{[m]}}{\sum_{m=1}^{M} Q_m(j)}$$

max $\Sigma$: let $A_j = \Sigma_j^{-1}$

$$\max_{A_j} \sum_{m=1}^{M} Q_m(j)\left[-\frac{1}{2}\log|A_j^{-1}| - \frac{1}{2}(y^{[m]}-\mu_j)^T A_j (y^{[m]}-\mu_j)\right]$$

$$= \max_{A_j} \sum_{m=1}^{M} Q_m(j)\left[\log\det(A_j) - \operatorname{tr}\left(A_j,(y^{[m]}-\mu_j)(y^{[m]}-\mu_j)^T\right)\right]$$

$$= \max_{A_j} \log\det(A_j)\sum_{m=1}^{M} Q_m(j) - \operatorname{tr}\left(A_j,\sum_{m=1}^{M} Q_m(j)(y^{[m]}-\mu_j)(y^{[m]}-\mu_j)^T\right)$$

$$\nabla_{A_j} = A_j^{-1}\sum_{m=1}^{M} Q_m(j) - \sum_{m=1}^{M} Q_m(j)(y^{[m]}-\mu_j)(y^{[m]}-\mu_j)^T = 0$$

$$\Sigma_j = A_j^{-1} = \frac{\sum_{m=1}^{M} Q_m(j)(y^{[m]}-\mu_j)(y^{[m]}-\mu_j)^T}{\sum_{m=1}^{M} Q_m(j)}$$

$\log\det(X+\Delta X) = \log\det\left[X^{\frac{1}{2}}(I+X^{\frac{1}{2}}\Delta X X^{\frac{1}{2}})X^{\frac{1}{2}}\right]$

$= \log\det\left[X(I+X^{\frac{1}{2}}\Delta X X^{\frac{1}{2}})\right]$

$= \log\det(X) + \log\det(I+X^{\frac{1}{2}}\Delta X X^{\frac{1}{2}})$

$= \log\det(X) + \sum_i \log(1+\lambda_i)$ ($\lambda_i$ is the $i$th eigenvalue of $X^{\frac{1}{2}}\Delta X X^{\frac{1}{2}}$)

$\approx \log\det(X) + \sum_i \lambda_i$ ($\lim_{\lambda\to 0}\frac{\lambda}{\log(1+\lambda)} = 1$)

$= \log\det(X) + \operatorname{tr}(X^{\frac{1}{2}}\Delta X X^{\frac{1}{2}})$

$= \log\det(X) + \operatorname{tr}(\Delta X, X^{-1})$

max $\theta$

$$\max_{\theta} \sum_{m=1}^{M}\sum_{j\in\{0,1\}} Q_m(j)\log\frac{\theta_j\cdot\mathcal{N}(y^{[m]};\mu_j,\Sigma_j)}{Q_m(j)}$$

$$\Downarrow$$

$$\min_{\theta} -\sum_{m=1}^{M}\sum_{j\in\{0,1\}} Q_m(j)\log\theta_j$$

s.t. $\sum_j \theta_j = 1$ : $\nu$

$$\mathcal{L}(\theta, U) = -\sum_{m=1}^{M}\sum_{j\in val(x)} Q_m(j)\log\theta_j + U\left(\sum_j \theta_j - 1\right)$$

$$\nabla_{\theta_j}\mathcal{L} = -\frac{1}{\theta_j}\sum_{m=1}^{M} Q_m(j) + U := 0$$

$$\theta_j = \frac{\sum_{m=1}^{M} Q_m(j)}{U}$$

$$g(U) = \min_{\theta}\mathcal{L}(\theta, U) = -\sum_{m=1}^{M}\sum_{j\in val(x)} Q_m(j)\log\frac{\sum_{m=1}^{M} Q_m(j)}{U} + \sum_j\sum_{m=1}^{M} Q_m(j) - U$$

$$= \sum_{m=1}^{M}\sum_{j\in val(x)} Q_m(j)\log U - U + C$$

$$\nabla_U g = \frac{1}{U}\sum_{m=1}^{M}\sum_{j\in val(x)} Q_m(j) - 1 := 0$$

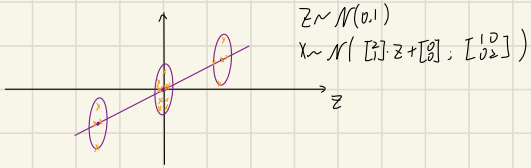
$$U^* = \sum_{m=1}^{M}\sum_j Q_m(j) = \sum_{m=1}^{M} 1 = M$$

$$\theta_j^* = \frac{\sum_{m=1}^{M} Q_m(j)}{M}$$

eg. Factor Analysis

$z \sim \mathcal{N}(0, I)$

$x|z \sim \mathcal{N}(\mu + \Lambda z, \Psi)$




$z \sim \mathcal{N}(0,1)$

$x \sim \mathcal{N}\left(\begin{bmatrix}1\\2\end{bmatrix} z + \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 2\end{bmatrix}\right)$

$$\ell(\theta) = \sum_{m=1}^{M}\log p(x_{[m]};\theta)$$

$$= \sum_{m=1}^{M}\log\int_{z_{[m]}} p(x_{[m]}, z_{[m]};\theta)\, dz_{[m]}$$

$$= \sum_{m=1}^{M}\log\int_{z_{[m]}} Q_m(z_{[m]})\cdot\frac{p(x_{[m]}, z_{[m]};\theta)}{Q_m(z_{[m]})}\, dz_{[m]} = \sum_{m=1}^{M}\log E_{z_{[m]}\sim Q_m}\left[\frac{p(x_{[m]}, z_{[m]};\theta)}{Q_m(z_{[m]})}\right]$$

$$\geq \sum_{m=1}^{M} E_{z_{[m]}\sim Q_m}\left[\log\frac{p(x_{[m]}, z_{[m]};\theta)}{Q_m(z_{[m]})}\right]$$

$$= \sum_{m=1}^{M}\int_{z_{[m]}} Q_m(z_{[m]})\cdot\log\frac{p(x_{[m]}, z_{[m]};\theta)}{Q_m(z_{[m]})}\, dz_{[m]}$$

$$Q_m(z_{[m]}) = \frac{1}{Z} p(x_{[m]}, z_{[m]};\theta) \quad \forall z_{[m]} \in val(z)$$

$$\int_{z_{[m]}} Q_m(z_{[m]})\, dz_{[m]} = \frac{1}{Z}\int_{z_{[m]}} p(x_{[m]}, z_{[m]};\theta)\, dz_{[m]} = \frac{1}{Z}\cdot p(x_{[m]};\theta) = 1$$

$$Q_m(z_{[m]}) = \frac{p(x_{[m]}, z_{[m]};\theta)}{p(x_{[m]};\theta)} = p(z_{[m]}|x_{[m]};\theta)$$

E-Step:

$$p(z_{[m]}, x_{[m]};\theta) = p(z_{[m]})\cdot p(x_{[m]}|z_{[m]};\theta)$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}}\exp\left(-\frac{1}{2} z_{[m]}^T z_{[m]}\right)\cdot\frac{1}{(2\pi)^{\frac{n}{2}}\cdot|\Psi|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}[x_{[m]}-(\mu+\Lambda z_{[m]})]^T\Psi^{-1}[x_{[m]}-(\mu+\Lambda z_{[m]})]\right)$$

$$= \frac{1}{Z}\exp\left\{-\frac{1}{2}\left[z_{[m]}^T z_{[m]} + x_{[m]}^T\Psi^{-1}x_{[m]} - 2x_{[m]}^T\Psi^{-1}(\mu+\Lambda z_{[m]}) + (\mu+\Lambda z_{[m]})^T\Psi^{-1}(\mu+\Lambda z_{[m]})\right]\right\}$$

$$= \frac{1}{Z}\exp\left\{-\frac{1}{2}\left[z_{[m]}^T z_{[m]} + x_{[m]}^T\Psi^{-1}x_{[m]} - 2x_{[m]}^T\Psi^{-1}\mu - 2x_{[m]}^T\Psi^{-1}\Lambda z_{[m]} + \mu^T\Psi^{-1}\mu + 2\mu^T\Psi^{-1}\Lambda z_{[m]} + (\Lambda z_{[m]})^T\Psi^{-1}(\Lambda z_{[m]})\right]\right\}$$

$$= \frac{1}{Z}\exp\left\{-\frac{1}{2}\begin{bmatrix}z_{[m]}, & x_{[m]}-\mu\end{bmatrix}\begin{bmatrix}I+\Lambda^T\Psi^{-1}\Lambda & -\Lambda^T\Psi^{-1}\\ -\Psi^{-1}\Lambda & \Psi^{-1}\end{bmatrix}\begin{bmatrix}z_{[m]}\\ x_{[m]}-\mu\end{bmatrix}\right\}$$

$$\begin{bmatrix}z_{[m]}\\ x_{[m]}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\ \mu\end{bmatrix}, \Sigma\right)$$

$$\Sigma = \begin{bmatrix}I+\Lambda^T\Psi^{-1}\Lambda & -\Lambda^T\Psi^{-1}\\ -\Psi^{-1}\Lambda & \Psi^{-1}\end{bmatrix}^{-1} = \begin{bmatrix}I & \Lambda^T\\ \Lambda & \Psi+\Lambda\Lambda^T\end{bmatrix}$$

$$\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}^{-1} = \begin{bmatrix}\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1} & -\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1}\Sigma_{12}\Sigma_{22}^{-1}\\ -\Sigma_{22}^{-1}\Sigma_{21}\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1} & \Sigma_{22}^{-1}+\Sigma_{22}^{-1}\Sigma_{21}\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1}\Sigma_{12}\Sigma_{22}^{-1}\end{bmatrix}$$

by block elimination

$$\left(z_{[m]} \middle| x_{[m]}\right) \sim \mathcal{N}\left(-\lambda^T(\Psi+\lambda\lambda^T)^{-1}(x_{[m]}-u),\ \Sigma-\lambda^T(\Psi+\lambda\lambda^T)^{-1}\lambda\right)$$

$$Q_m(z_{[m]}) = P(z_{[m]}|x_{[m]}) = \mathcal{N}\Big(z_{[m]};\ \underbrace{-\lambda^T(\Psi+\lambda\lambda^T)^{-1}(x_{[m]}-u)}_{E[z_{[m]}|x_{[m]}]},\ \underbrace{\Sigma-\lambda^T(\Psi+\lambda\lambda^T)^{-1}\lambda}_{Cov(z_{[m]}|x_{[m]})}\Big)$$

M-Step:

$$\max_{\lambda,u,\Psi} \sum_{m=1}^{M} \int_{z_{[m]}} Q_m(z_{[m]}) \log \frac{P(x_{[m]},z_{[m]};\theta)}{Q_m(z_{[m]})}\, dz_{[m]}$$

$$= \max_{\lambda,u,\Psi} \sum_{m=1}^{M} \int_{z_{[m]}} Q_m(z_{[m]}) \left[\log P(z_{[m]}) + \log P(x_{[m]}|z_{[m]};\theta) - \log Q_m(z_{[m]})\right] dz_{[m]}$$

$$= \max_{\lambda,u,\Psi} \sum_{m=1}^{M} \int_{z_{[m]}} Q_m(z_{[m]}) \log P(x_{[m]}|z_{[m]};\theta)\, dz_{[m]}$$

$$= \max_{\lambda,u,\Psi} \sum_{m=1}^{M} \int_{z_{[m]}} Q_m(z_{[m]}) \left[\log \frac{1}{(2\pi)^{\frac{n}{2}}|\Psi|^{\frac{1}{2}}} - \frac{1}{2}(x_{[m]}-(u+\lambda z_{[m]}))^T\Psi^{-1}(x_{[m]}-(u+\lambda z_{[m]}))\right] dz_{[m]}$$

max $\lambda$:

$$\nabla_\lambda = \sum_{m=1}^{M} \int_{z_{[m]}} Q_m(z_{[m]})\ \Psi^{-1}(x_{[m]}-u-\lambda z_{[m]})\ z_{[m]}^T\, dz_{[m]}$$

$$= \sum_{m=1}^{M} \int_{z_{[m]}} Q_m(z_{[m]})\ \Psi^{-1}(x_{[m]}-u)\ z_{[m]}^T\, dz_{[m]} - \int_{z_{[m]}} Q_m(z_{[m]})\cdot\Psi^{-1}\lambda\ z_{[m]}z_{[m]}^T\, dz_{[m]}$$

$$= \sum_{m=1}^{M} \Psi^{-1}(x_{[m]}-u)\, E_{z_{[m]}\sim Q_m}[z_{[m]}]^T - \sum_{m=1}^{M} \Psi^{-1}\lambda\, E_{z_{[m]}\sim Q_m}[z_{[m]}z_{[m]}^T] := 0$$

$$\lambda = \left(\sum_{m=1}^{M}(x_{[m]}-u)\, E_{z_{[m]}\sim Q_m}[z_{[m]}]^T\right)\left(\sum_{m=1}^{M} E_{z_{[m]}\sim Q_m}[z_{[m]}z_{[m]}^T]\right)^{-1}$$

max $u$:

$$\nabla_u = \sum_{m=1}^{M} \int_{z_{[m]}} Q_m(z_{[m]})\ \Psi^{-1}(x_{[m]}-u-\lambda z_{[m]})\, dz_{[m]}$$

$$= \sum_{m=1}^{M} \Psi^{-1}\int_{z_{[m]}} Q_m(z_{[m]})(x_{[m]}-u)\, dz_{[m]} - \sum_{m=1}^{M}\Psi^{-1}\lambda\int_{z_{[m]}} Q_m(z_{[m]})\, z_{[m]}\, dz_{[m]}$$

$$= \Psi^{-1}\sum_{m=1}^{M}(x_{[m]}-u) - \Psi^{-1}\lambda\sum_{m=1}^{M} E_{z_{[m]}\sim Q_m}[z_{[m]}]$$

$$= \Psi^{-1}\sum_{m=1}^{M}(x_{[m]}-u) - \Psi^{-1}\lambda\cdot\sum_{m=1}^{M} -\lambda^T(\Psi+\lambda\lambda^T)^{-1}(x_{[m]}-u)$$

$$= \left(\Psi^{-1}+\Psi^{-1}\lambda\lambda^T(\Psi+\lambda\lambda^T)^{-1}\right)\sum_{m=1}^{M}(x_{[m]}-u) := 0$$

$$u = \frac{1}{M}\sum_{m=1}^{M} x_{[m]}$$

max $\Psi$:

$$\max_{\Psi} \sum_{m=1}^{M} \int_{z_{[m]}} Q_m(z_{[m]}) \cdot \left\{-\frac{1}{2}\log\det(\Psi) - \frac{1}{2}(x_{[m]}-(u+\lambda z_{[m]}))^T\Psi^{-1}(x_{[m]}-(u+\lambda z_{[m]}))\right\} dz_{[m]}$$

$$= \min_{\Psi} \sum_{m=1}^{M} \int_{z_{[m]}} Q_m(z_{[m]}) \cdot \left\{\log\det(\Psi) + tr\left\langle\Psi^{-1},\ (x_{[m]}-(u+\lambda z_{[m]}))(x_{[m]}-(u+\lambda z_{[m]}))^T\right\rangle\right\} dz_{[m]}$$

let $J := \Psi^{-1}$

$$= \min_{J} \sum_{m=1}^{M} \int_{z_{[m]}} Q_m(z_{[m]}) \cdot \left\{-\log\det(J) + tr\left\langle J,\ (x_{[m]}-(u+\lambda z_{[m]}))(x_{[m]}-(u+\lambda z_{[m]}))^T\right\rangle\right\} dz_{[m]}$$

$$\nabla_J = \sum_{m=1}^{M} \int_{z_{[m]}} Q_m(z_{[m]}) \left\{-J^{-1} + (x_{[m]}-u-\lambda z_{[m]})(x_{[m]}-u-\lambda z_{[m]})^T\right\} dz_{[m]}$$

$$= \sum_{m=1}^{M} -J^{-1} + \sum_{m=1}^{M} \int_{z[m]}^{+\infty} Q_m(z[m]) \; \lambda z[m] z[m]^T \lambda^T - \lambda z[m] (x[m]-u)^T + (x[m]-u)(x[m]-u)^T \, dz[m]$$

$$= -MJ^{-1} + \sum_{m=1}^{M} \lambda E_{z[m]\sim Q_m}[z[m] z[m]^T] \lambda^T - \lambda E_{z[m]\sim Q_m}[z[m]] (x[m]-u)^T + (x[m]-u)(x[m]-u)^T \quad := 0$$

$$\Psi = J^{-1} = \frac{1}{M} \sum_{m=1}^{M} \lambda E_{z[m]\sim Q_m}[z[m] z[m]^T] \lambda^T - \lambda E_{z[m]\sim Q_m}[z[m]] (x[m]-u)^T + (x[m]-u)(x[m]-u)^T$$

## · EM Algorithm for Bayesian Network

let $o[m]$ be observed vars of $m$th data and $z[m]$ be hidden vars of $m$th data

$$l(\theta) = \sum_{m=1}^{M} \ln P(o[m]; \theta)$$

$$= \sum_{m=1}^{M} \ln \sum_{z[m] \in Val(z)} P(o[m], z[m]; \theta)$$

$$= \sum_{m=1}^{M} \ln \sum_{z[m]} Q_m(z[m]) \cdot \frac{P(o[m], z[m]; \theta)}{Q_m(z[m])}$$

$$① \geq \sum_{m=1}^{M} \sum_{z[m]} Q_m(z[m]) \ln \frac{P(o[m], z[m]; \theta)}{Q_m(z[m])}$$

$$= \sum_{m=1}^{M} \sum_{z[m]} Q_m(z[m]) \cdot \left[ \sum_{\langle x \rangle \otimes \langle u \rangle \in \theta} \ln P\left( (o[m], z[m])\langle x \rangle \,|\, (o[m], z[m])\langle u \rangle \right) - \ln Q_m(z[m]) \right]$$

### E-Step:

to achieve equality at ① $\frac{P(o[m], z[m]; \theta)}{Q_m(z[m])}$ is constant

$$Q_m(z[m]) = \frac{1}{Z} P(o[m], z[m]; \theta)$$

$$\sum_{z[m]} Q_m(z[m]) = \frac{1}{Z} \sum_{z[m]} P(o[m], z[m]; \theta) = \frac{1}{Z} P(o[m]; \theta) = 1$$

$$Q_m(z[m]) = \frac{P(o[m], z[m]; \theta)}{P(o[m]; \theta)} = P(z[m] \,|\, o[m]; \theta)$$

### M-Step:

Find $\theta_{x|u} \quad \forall x \in Val(X)$

$$\min \; -\sum_{m=1}^{M} \sum_{z[m]} Q_m(z[m]) \ln P\left( (o[m], z[m])\langle x \rangle \,|\, (o[m], z[m])\langle u \rangle \right)$$

$$s.t. \; \sum_{x \in Val(X)} \theta_{x|u} = 1$$

$$\mathcal{L}(\theta_{x|u}, v) = -\sum_{m=1}^{M} \sum_{z[m]} Q_m(z[m]) \ln P\left( (o[m], z[m])\langle x \rangle \,|\, (o[m], z[m])\langle u \rangle \right) + v\left( \sum_{x} \theta_{x|u} - 1 \right)$$

$$= -\sum_{x \in Val(X)} \ln \theta_{x|u} \sum_{m=1}^{M} \sum_{z[m] \in Val(z)} Q_m(z[m]) \cdot 1\{ (o[m], z[m])\langle x, u \rangle = (x, u) \} + v\left( \sum_{x} \theta_{x|u} - 1 \right)$$

$$\nabla_{\theta_{x|u}} \mathcal{L} = -\frac{1}{\theta_{x|u}} \cdot \sum_{m=1}^{M} \sum_{z[m] \in Val(z)} Q_m(z[m]) \, 1\{ (o[m], z[m])\langle x, u \rangle = (x, u) \} + v \quad := 0$$

$$\theta_{x|u}(v) = \frac{1}{v} \cdot \sum_{m=1}^{M} \sum_{z[m] \in Val(z)} Q_m(z[m]) \, 1\{ (o[m], z[m])\langle x, u \rangle = (x, u) \}$$

let $h(x, u) = \sum_{m=1}^{M} \sum_{z[m]} Q_m(z[m]) \, 1\{ (o[m], z[m])\langle x, u \rangle = (x, u) \}$ is constant $\forall x, u$

$\theta_{x|u}(v) = \frac{1}{v} h(x,u)$

$g(v) = \inf_{\theta_{x|u}} L(\theta_{x|u}, v)$

$= -\sum_{x,u} \log\left(\frac{1}{v} h(x,u)\right) \cdot h(x,u) + v\left(\sum_{x,u} \frac{1}{v} h(x,u) - 1\right)$

$= -\sum_{x,u} \left(\log h(x,u) - \log v\right) \cdot h(x,u) + \sum_{x,u} h(x,u) - v$

$\nabla_v g = \frac{1}{v} \cdot \sum_{x,u} h(x,u) - 1 \;:=0$

$v^* = \sum_{x,u} h(x,u)$

$\theta^*_{x|u} = \frac{h(x,u)}{\sum_x h(x,u)} = \frac{\sum_{m=1}^{M} \sum_{z[m]} Q_m(z[m]) \cdot 1\{(o[m], z[m]) \langle x,u \rangle = (x,u)\}}{\sum_x \sum_{m=1}^{M} \sum_{z[m]} Q_m(z[m]) \cdot 1\{(o[m], z[m]) \langle x,u \rangle = (x,u)\}}$

$= \frac{\sum_{m=1}^{M} \sum_{z[m]} P(z[m] | o[m]; \theta) \cdot 1\{(o[m], z[m]) \langle x,u \rangle = (x,u)\}}{\sum_x \sum_{m=1}^{M} \sum_{z[m]} P(z[m] | o[m]; \theta) \cdot 1\{(o[m], z[m]) \langle x,u \rangle = (x,u)\}}$

$= \frac{\sum_{m=1}^{M} P(x,u | o[m]; \theta)}{\sum_x \sum_{m=1}^{M} P(x,u | o[m]; \theta)}$

```
def compute_marginal_Q(G, θ, D: partially observed dataset):
    for each X → Y ∈ G
        for each (x,u) ∈ val(X,U):
            Q(x,u) = 0
    for m=1,...M,
        run inference with evidence o[m]
        for each X → Y ∈ G
            for each (x,u) ∈ val(X,U):
                Q(x,u) += P(x,u | o[m])
    return {Q(x,u)  ∀ X → Y ∈ G, ∀ (x,u) ∈ val(X,U)}

def EM_Bayesian_Net(G, θ⁰, D):
    for t=0,1,2,... until convergence
        {Q(x,u)} = Compute_Marginal_Q(G, θ^t, D)

        for each X → Y ∈ G
            for each (x,u) ∈ val(X,U):
                θ_{x|u} = Q(x,u) / Σ_x Q(x,u)
    return θ^t
```