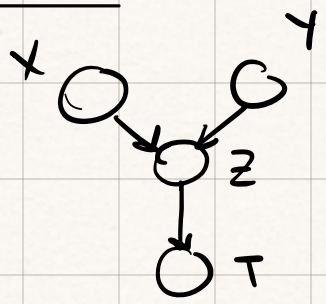
1.) DGM



Factorization is $p(X_v) = \prod_i p(x_i \mid X_{\pi_i})$

$$\Rightarrow \boxed{p(X_v) = p(x)p(y)p(z|x,y)p(t|z)}$$

Also, $X \perp\!\!\!\perp Y \mid T \iff X$ and Y are d-separated by T.

We have seen that sets X and Y are d-separated by set D if

1.) $d \in D$ and (v_{i-1}, d, v_{i+1}) is not a v-structure

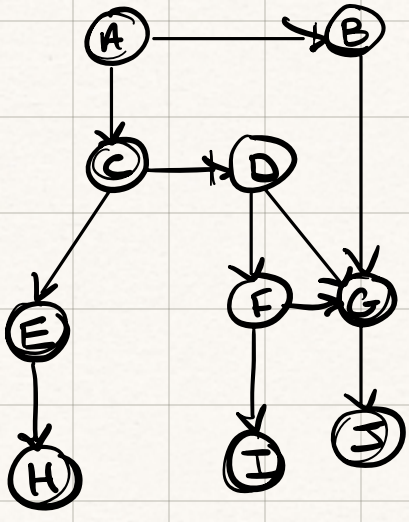
or 2.) $d \notin D$ and (v_{i-1}, d, v_{i+1}) is a v-structure but no descendants of d is in D.

1.) is not respected because X, Y, Z is a v-structure.

2.) is not respected, because although X, Y, Z is a v-structure there is a descendant of $d \in Z$ in D (it is T).

So $X \perp\!\!\!\perp Y \mid T$ is not true in general.

2) D-separation in DGM



a) $C \perp\!\!\!\perp B$ No

b) $C \perp\!\!\!\perp B \mid A$ Yes

c) $C \perp\!\!\!\perp B \mid A, J$ No

d) $C \perp\!\!\!\perp B \mid A, D, J$ Yes

e) $C \perp\!\!\!\perp G$ No

f) $C \perp\!\!\!\perp G \mid B$ No

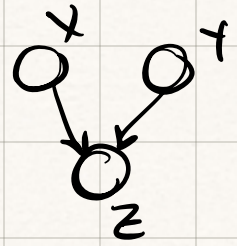
g) $C \perp\!\!\!\perp G \mid B, D$ Yes

h) $C \perp\!\!\!\perp G \mid B, D, H$ Yes

i) $C \perp\!\!\!\perp G \mid B, D, H, E$ Yes

j) $B \perp\!\!\!\perp I \mid J$ No

3. Positive interactions in V structure



$\alpha := p(X=1)$, $\beta := p(X=1 \mid Z=1)$

$\gamma := p(X=1 \mid Z=1, Y=1)$.

a)

(i) Want $\gamma < \alpha$.

With

$p(X=1)$
0.5

$p(Y=1)$
0.5

Then $\beta = p(X=1 \mid Z=1) = \dfrac{p(X=1, Z=1)}{p(Z=1)}$ (product rule, $p(x,y)=p(y)p(x|y)$)

$$= \frac{\sum_{i=0,1} p(X=1, Y=i, Z=1)}{\sum_{ij=0,1} p(X=i, Y=j, Z=1)}$$ (marginalization)

$$= \frac{\sum_{i=0,1} p(X=1)\,p(Y=i)\,p(Z=1 \mid X=1, Y=i)}{\sum_{ij=0,1} p(X=i)\,p(Y=j)\,p(Z=1 \mid X=i, Y=j)}$$ (graph factorization)

$$= \frac{0.25\left(p(Z=1 \mid X=1, Y=0) + p(Z=1 \mid X=1, Y=1)\right)}{\sum_{j=0,1} 0.5\,p(Y=j)\,p(Z=1, X=0 \mid Y=j) + 0.5\,p(Y=j)\,p(Z=1, X=1, Y=j)}.$$

(*).

$$\Rightarrow \beta = \frac{p(Z=1 \mid X=1, Y=0) + p(Z=1 \mid X=1, Y=1)}{p(Z=1 \mid X=0, Y=0) + p(Z=1 \mid X=1, Y=0) + p(Z=1 \mid X=0, Y=1) + p(Z=1 \mid X=1, Y=1)}$$

If

X	Y	P(Z=1|X,Y)
1	1	0.1
1	0	0.1
0	1	1
0	0	1

Then, $\beta = \frac{0.1+0.1}{0.1+0.1+1+1} = 0.091$

Also, $\gamma = P(X=1 \mid Z=1, Y=1) = \frac{p(X=1, Y=1, Z=1)}{p(Z=1, Y=1)}$ (product rule)

$= \frac{p(X=1)p(Y=1)p(Z=1 \mid X=1, Y=1)}{\sum_{i=0,1} p(X=i, Y=1, Z=1)}$ (graph fact & marginalization)

$= \frac{0.25 P(Z=1 \mid X=1, Y=1)}{p(X=0)p(Y=1)p(Z=1 \mid X=0, Y=1) + p(X=1)p(Y=1)p(Y=1 \mid X=1, Y=1)}$

(graph factorization)

$$\Rightarrow \boxed{\gamma = \frac{p(Z=1 \mid X=1, Y=1)}{p(Z=1 \mid X=0, Y=1) + p(Z=1 \mid X=1, Y=1)}} \quad (**)$$

In this case

$$\gamma = \frac{0.1}{1+0.1} = 0.091$$

λ	β	γ
0.5	0.091	0.091

and $\gamma < \lambda$.

(ii) $2 < \gamma < \beta$. Again let me set

$P(X=1)$	$P(Y=1)$
0.5	0.5

Equations (x) for β & (xx) for γ are valid again.

$$\beta = \frac{p(Z=1|X=1,Y=0)+p(Z=1|X=1,Y=1)}{p(Z=1|X=0,Y=0)+\ p(Z=1|X=1,Y=0)+p(Z=1|X=0,Y=1)+p(Z=1|X=1,Y=1)}$$

$$\gamma = \frac{p(Z=1|X=1,Y=1)}{p(Z=1|X=0,Y=1)+p(Z=1|X=1,Y=1)}$$

With.

X	Y	$P(Z=1\|X,Y)$
1	1	1
1	0	1
0	1	0.2
0	0	0.1

I have $\beta = \frac{1+1}{1+1+0.2+0.1} \simeq 0.87$

$\gamma = \frac{1}{1+0.2} \simeq 0.83$

So

α	β	γ
0.5	0.87	0.83

and $2 < \gamma < \beta$.

(iii) $\beta < 2 < \gamma$

X	Y	$P(Z=1\|X,Y)$
1	1	1
1	0	0.1
0	1	0.2
0	0	1

$$\beta = \frac{p(z=1|x=1,y=0) + p(z=1|x=1,y=1)}{p(z=1|x=0,y=0) + p(z=1|x=1,y=0) + p(z=1|x=0,y=1) + p(z=1|x=1,y=1)}$$

$$\gamma = \frac{p(z=1|x=1,y=1)}{p(z=1|x=0,y=1) + p(z=1|x=1,y=1)}$$

$$\beta = \frac{1+0.1}{1+0.1+0.2+1} \approx 0.478$$

So $$\gamma = \frac{1}{1+0.2} \simeq 0.833$$

α	β	γ
0.5	0.478	0.833

and $\beta < \alpha < \gamma$.

b).

i) Say Z = late? with $Z=1$ is Yes & $Z=0$ is No

X = Has watch? with $X=1$ Yes & $X=0$ No

and Y = Ate cereal? with $Y=1$ Yes & $Y=0$ No

Conditionals were

X	Y	$P(Z=1\|X,Y)$	
1	1	0.1	has watch, ate cereal
1	0	0.1	has watch, didn't eat cereal
0	1	1	forget watch, ate cereal
0	0	1	forget watch, didn't eat cereal

$P(X=1) = 0.5$, but we see that we he forgets his watch, the probability of being late is very high ($P(Z=1|X=0,Y=i) = 1$). So it makes sense that

given the fact that he is late, the probability that he has his watch is smaller, i.e.

$$p(X=1|Y=1, Z=1) < p(X=1).$$

b)

X	Y	P(Z=1|X,Y)
1	1	1
1	0	1
0	1	0.2
0	0	0.1

Say Z = late? with Z=1 is Yes & Z=0 is No

X = Forgot watch? with X=1 Yes & X=0 No

Y = Ate cereal? with Y=1 Yes & Y=0 No

Because it is certain that he is late when he forgets his watch, it is clear that the probability that he forget his watch is bigger if we knew that he is late. That is

$$p(X=1|Z=1), p(X=1|Z=1,Y=1) > p(X=1).$$

$$\beta, \gamma > \alpha.$$

Also, we he has his watch, he is more likely to be late. when Y=1, so

$$p(X=1|Z=1,Y=1) < p(X=1|Z=1)$$
$$\gamma \qquad\qquad \beta$$

$$\Rightarrow \alpha < \gamma < \beta.$$

c)

X	Y	P(Z=1|X,Y)
1	1	1
1	0	0.1
0	1	0.2
0	0	1

We knew that $Z=1$ is certain to happen if $X=Y=1$ or $X=Y=0$ at the same time, and unlikely in the other cases.

So given $Y=1$ & $Y=1$, it is much more probable that $X=1$. This is why

$$\gamma > \alpha, \beta.$$

If we just knew that $Z=1$, then the probability that $X=1$ is smaller than just $X=1$

So $$\beta < \alpha < \gamma.$$

4) Equivalence of directed tree w/ undirected tree.

Let $G(V,E)$ be a directed tree.

Since G is a directed tree, every node in it has at most one parent.

Now let $\bar{G}$ be the moralized graph of G. Then it has the same node set $J=V$ & an edge set

$$\bar{E} = \underbrace{\{\{i,j\} : (i,j) \in E\}}_{\text{same edges as } E \text{ but undirected.}} \cup \{\{k,\ell\} : k,\ell \in \pi_i, k \neq \ell, \forall i \in V\}$$

but since any node in V has at most one parent, $|\pi_i| \leq 1$, and we cannot find two nodes k & ℓ in π_i s.t. $k \neq \ell$,

so $\{\{k,\ell\} : k,\ell \in \pi_i,\ k \neq \ell,\ \forall i \in V\} = \{\phi\}$ (empty)

and

$$\bar{E} = E$$ but not undirected edges.

Now, any $p \in \mathcal{I}(G)$ factorizes as

$$p(x_V) = \prod_{i=1}^{N} p(x_i \mid x_{\pi_i}) \qquad \forall\, i \in V$$

but since $|\pi_i| \leq 1$, this is a product of factors $p(x_i \mid x_k)$ $\forall\, i \in V$ with $k \in \pi_i$ is at most one node. I.e. we just have factors between a parent & its children

Also, any $p' \in \mathcal{I}(G')$ factorizes as

$$p'(x_V) = \frac{1}{Z} \prod_{c} \phi_c(x_c)$$

but since $\bar{E} = E$ but undirected, the cliques c in $\bar{G}$ link just two nodes, so it is a product of factors

$\phi(x_1, x_2)$. where x_1 & x_2 are parent & children.

These are the same conditional independence statements as $\mathcal{I}(G)$. So if we have some $p \in \mathcal{I}(G)$ then we have $p \in \mathcal{I}(\bar{G})$, and vice versa.

$$\Rightarrow \mathcal{I}(G) = \mathcal{I}(\bar{G}).$$

5) Hammersley-Clifford Counter example

G: $X_1 - X_2 - X_3 - X_4 - X_1$

Probabilities that have $p = 1/8$:

$(0,0,0,0)$, $(1,0,0,0)$, $(1,1,0,0)$, $(1,1,1,0)$, $(0,0,0,1)$, $(0,0,1,1)$

$(0,1,1,1)$, $(1,1,1,1)$

all others have zero.

Assuming we can show that Global Markov property is satisfied,

Let's show that p cannot factorize according to G:

Usual factorization in a UGM:

$$p(x_1) = \frac{1}{Z} \prod_c \phi_c(x_c) = \frac{1}{Z} \left(\phi_{12}(x_1, x_2) \phi_{23}(x_2, x_3) \phi_{34}(x_3, x_4) \phi_{41}(x_4, x_1) \right)$$

We know

$$p(0,0,0,0) \sim \phi_{12}(0,0) \phi_{23}(0,0) \phi_{34}(0,0) \phi_{41}(0,0) = \frac{1}{8}$$

and

$$p(0,0,1,1) \sim \phi_{12}(0,0) \phi_{23}(0,1) \phi_{34}(1,1) \phi_{41}(1,0) = \frac{1}{8}.$$

This means, in particular, that $\phi_{12}(0,0)$, $\phi_{23}(0,1)$ and $\phi_{41}(0,0)$ cannot be zero (otherwise $p \neq 1/8$, but zero).

Now, consider

$$p(0,0,1,0) \sim \phi_{12}(0,0) \phi_{23}(0,1) \phi_{34}(1,0) \phi_{41}(0,0) = 0.$$

↑ not zero ↑ not zero ↑ not zero

so $\phi_{34}(1,0)$ has to be zero.

but consider $p(1,1,1,0) \sim \phi_{12}(1,1) \phi_{23}(1,1) \phi_{34}(1,0) \phi_{41}(0,1) = \frac{1}{8}$.

but if $\phi_{34}(1,0)$ is zero, $p(1,1,1,0)$ is zero... Contradiction !!

p cannot factorize according to $G \Rightarrow p \notin \mathcal{L}(G)$, which contradicts the Hammersley-Clifford theorem.

6) Bizarre conditional independence statement.

Attempt ...

$$p(x,y,z) = p(z)p(x,y|z) \qquad \text{(product rule)}$$

$$= p(z)\,p(x|z)\,p(y|z) \qquad (x \perp\!\!\!\perp y \mid z)$$

$$= p(z)\,p(x|z)\,\frac{p(y,z)}{p(z)} \qquad \text{by chain rule}$$

$$= p(x|z)\,p(y,z). \quad \text{(i)}$$

but also $p(x,y,z) = p(z)\,p(x|z)p(y|z)$

$$= p(z)\,\frac{p(x,z)}{p(z)}\,p(y|z)$$

$$= p(y|z)\,p(x,z) \quad \text{(ii)}$$

(i) = (ii)

$$p(x|z)p(y,z) = p(y|z)p(x,z)$$

Assume $x \perp\!\!\!\perp z$ but $y \not\perp\!\!\!\perp z$

Then $p(x|z) = p(x) \Rightarrow p(x)\,p(y,z) = p(y|z)p(x)p(z)$

$$\Rightarrow p(y,z) = p(y|z).$$

which is not true in general. Similarly, if

$y \perp\!\!\!\perp z$ but $x \not\perp\!\!\!\perp z$

I get $p(x,z) = p(x|z)$

which is also not true in general.

So no, it is not true that $(X \perp\!\!\!\perp Z)$ or $(Y \perp\!\!\!\perp Z)$ in general.

This is the answer to b.

7.) $p(x_i|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\vec{\mu_k}, \sigma_k I)$.

Let $\ell_c(\theta; D_c)$ be the complete log likelihood, where

D_c is our "dataset" $D_c = \{(x_i, z_i) : i = 1, \ldots, N\}$.

(z_i: latent variable)

$$\ell_c(\theta; D_c) = \sum_{i=1}^{N} \log p(x_i, z_i; \theta)$$

$$= \sum_{i=1}^{N} \log \prod_{k=1}^{K} \left(\pi_k \mathcal{N}(x_i|\vec{\mu_k}, \sigma_k^2 I)\right)^{z_i^k}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} z_i^k \log \pi_k \mathcal{N}(x_i|\vec{\mu_k}, \sigma_k^2 I)$$

E-step:

$$\langle \ell_c(\theta; D_c) \rangle_{\theta^{(t)}} = \left\langle \sum_{i=1}^{N} \sum_{k=1}^{K} z_i^k \log \pi_k \mathcal{N}(x_i|\vec{\mu_k}, \sigma_k^2 I) \right\rangle$$

$$\boxed{\langle \ell_c(\theta; D_c) \rangle_{\theta^{(t)}} = \sum_{i=1}^{N} \sum_{k=1}^{K} \langle z_i^k \rangle_{\theta^{(t)}} \log \pi_k \mathcal{N}(x_i|\vec{\mu_k}, \sigma_k^2 I)}$$

and define $\tau^i \equiv p(z^i = 1 | x, \theta)$

Bayes rule: $\tau_i^{k(t)} \equiv \dfrac{p(x_i|z^i = 1, \theta_k^{(t)}) p(z^i = 1|\pi_i)}{p(x_i|\theta_k^{(t)})}$

$$\Rightarrow \quad \tau_i^{k(t)} = \frac{\pi_k \mathcal{N}(x_i|\vec{\mu_k}, \sigma_k^2 I)}{\sum_j \pi_j \mathcal{N}(x_i|\vec{\mu_j}, \sigma_j^2 I)}$$

M-step is MLE for θ using $\tau_i^{k(t)}$ instead of $\langle z_i^k \rangle_{\theta^{(t)}}$ on $\langle \ell(\theta, D_c) \rangle_{\theta^{(t)}}$

MLE for μ_k: take derivative w.r.t. $\vec{\mu}_j$ & set to 0.

$$\frac{\partial}{\partial \vec{\mu}_j} \langle \ell_c(\theta, D_c) \rangle = \frac{\partial}{\partial \vec{\mu}_j} \sum_{k,i} \tau_i^{k(t)} \left(\log \pi_k + \log \left\{ \frac{1}{(2\pi)^{d/2} |\sigma_k^2 I|^{1/2}} \exp\left(-\frac{1}{2} (x_i - \vec{\mu}_k)^T \sigma_k^{-2} I (x_i - \vec{\mu}_k)\right) \right\} \right) = 0$$

$$\Rightarrow \frac{\partial}{\partial \vec{\mu}_j} \sum_{k,i} \tau_i^{k(t)} \left(-\frac{1}{2} \left(\log 2\pi^d + \log |\sigma_k^2 I| \right) - \frac{1}{2} (x_i - \vec{\mu}_k)^T \sigma^{-2} I (x_i - \vec{\mu}_k) \right) = 0.$$

$$\Rightarrow \frac{\partial}{\partial \vec{\mu}_j} \sum_{k,i} \tau_i^{k(t)} \left(-\frac{1}{2} (x_i - \vec{\mu}_k)^T \sigma^{-2} I (x_i - \vec{\mu}_k) \right) = 0.$$

Use $\frac{\partial}{\partial x} x^T A x = (A + A^T) x$

$$\Rightarrow \sum_{i=1}^{N} \tau_i^{j(t)} \left(-\frac{1}{2} \left(\underbrace{\sigma_j^{-2} I + (\sigma_j^{-2} I)^T}_{\sigma_j^{-2} I \text{ symmetric} \Rightarrow (\sigma_j^{-2} I)^T = \sigma_j^{-2} I} \right) (x_i - \vec{\mu}_j) \right) = 0.$$

$$\Rightarrow \sum_{i=1}^{N} \tau_i^{j(t)} \cancel{\sigma_j^{-2}} I x_i = \sum_{i=1}^{N} \tau_i^{j(t)} \cancel{\sigma_j^{-2}} I \vec{\mu}_j$$

$$\Rightarrow \boxed{\vec{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{N} \tau_i^{j(t)} x_i}{\sum_{i=1}^{N} \tau_i^{j(t)}}}$$

Now, for σ_j^2 (dropping terms that do not depend on σ_k^2)

$$\frac{\partial}{\partial \sigma_j^{-2}} \mathcal{L}(\theta, D) \big|_{\theta^{(t)}} = \frac{\partial}{\partial \sigma_j^{-2}} \sum_{i,k} \tau_i^{k(t)} \left(\frac{1}{2} \log |\sigma_k^{-2} I| - \frac{1}{2} (x_i - \vec{\mu_k})^T \right.$$

$$\left. \sigma_k^{-2} I (x_i - \vec{\mu_k}) \right) = 0$$

$$\Rightarrow -\frac{1}{2} \sum_{i,k} \tau_i^{k(t)} \left(\underbrace{\frac{\partial}{\partial \sigma_j^{-2}} \log |\sigma_k^{-2} I|}_{\det \sigma_k^{-2} I = \prod_{i=1}^{d} \lambda_{\sigma_k^{-2} I}} - \frac{\partial}{\partial \sigma_j^{-2}} \underbrace{(x_i - \vec{\mu_k})^T \sigma_k^{-2} I (x_i - \vec{\mu_k})}_{(*)} \right) = 0$$

where $\lambda_{\sigma_k^{-2} I}$ are the eigenvalues of $\sigma_k^{-2} I$. I is the $d \times d$ identity matrice, so it has d eigenvalues of 1 so $\det \sigma_k^{-2} I = \prod_{i=1}^{d} \sigma_k^{-2} = (\sigma_k^{-2})^d$

$$\Rightarrow \frac{\partial}{\partial \sigma_j^{-2}} \log |\sigma_k^{-2} I| = \frac{\partial}{\partial \sigma_j^{-2}} \log (\sigma_k^{-2})^d = d \frac{\partial}{\partial \sigma_j^{-2}} \log \sigma_k^{-2} = d \sigma_j^2$$

$(*) = (x_i - \vec{\mu_k})^T \sigma_k^{-2} I (x_i - \vec{\mu_k})$ is a scalar so it is equal to its trace

$$= \mathrm{Tr} \, (x_i - \vec{\mu_k})^T \sigma_k^{-2} I (x_i - \vec{\mu_k})$$

Cyclic property of trace two times:

$$= \mathrm{Tr} \, \sigma_k^{-2} I (x_i - \vec{\mu_k})(x_i - \vec{\mu_k})^T$$

$$= 2 \sigma_k^{-2} I, \underbrace{(x_i - \vec{\mu_k})(x_i - \vec{\mu_k})^T}_{\equiv \tilde{\Sigma}_k} >$$

$$= \langle \sigma_k^{-2} I, \tilde{\Sigma}_k \rangle$$

and $\frac{\partial}{\partial \sigma_j^2} \sum_k \langle \sigma_k^{-2} I, \tilde{\Sigma}_k \rangle = \langle I, \tilde{\Sigma}_j \rangle = \tilde{\Sigma}_j = (x_i - \vec{\mu_j})(x_i - \vec{\mu_j})^T$

So I have $\sum_{i=1}^{N} \tau_i^{j(t)} \left(d \sigma_j^2 - (x_i - \vec{\mu_j})(x_i - \vec{\mu_j})^T \right) = 0.$

$$\Rightarrow d\sum_{i=1}^{N} \tau_i^{j(t)} \sigma_j^2 = \sum_{i=1}^{N} \tau_i^{j(t)} (x_i - \vec{\mu_j})(x_i - \vec{\mu_j})^T$$

$$\Rightarrow \boxed{\sigma_j^{2(t+1)} = \frac{\sum_{i=1}^{N} \tau_i^{j(t)} (x_i - \vec{\mu_j}^{(t+1)})(x_i - \mu_j^{(t+1)})^T}{d\sum_{i=1}^{N} \tau_i^{j(t)}}}$$

Finally for π_k (dropping all terms that don't depend on π_k):

$$\frac{\partial}{\partial \pi_j} \langle \ell_c(\theta, \theta_t) \rangle_{\theta^{(t)}} = \frac{\partial}{\partial \pi_j} \sum_{i,k} \tau_i^{k(t)} \log \pi_k$$

With the constraint that $\sum_{k=1}^{K} \pi_k = 1$

$$\Rightarrow \underbrace{\sum_{k=1}^{K} \pi_k - 1}_{\equiv g(\pi_k)} = 0 \iff g(\pi_k) = 0$$

Build the Lagrangian:

$$\mathcal{L}(\pi_k, \lambda) = \sum_{i,k} \tau_i^{k(t)} \log \pi_k + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

Maximize w.r.t. π_j

$$\frac{\partial \mathcal{L}}{\partial \pi_j} = \sum_{i=1}^{N} \tau_i^{j(t)} \cdot \frac{1}{\pi_j} + \lambda = 0.$$

$$\Rightarrow \pi_j^{(t+1)} = -\sum_{i=1}^{N} \frac{\tau_i^{j(t)}}{\lambda}$$

But need $g(\pi_j^{(t+1)}) = 0 \Rightarrow -\sum_{j=1}^{K} \sum_{i=1}^{N} \frac{\tau_i^{j(t)}}{\lambda} - 1 = 0.$

$$\Rightarrow -\sum_{i=1}^{N}\sum_{j=1}^{K} \tau_i^{j(t)} = \lambda$$

$$\Rightarrow -\sum_{i=1}^{N} 1 = \lambda \quad \Rightarrow \boxed{\lambda = -N}$$

and $$\boxed{\pi_j^{(t+1)} = \frac{1}{N}\sum_i \tau_i^{k(t)}}$$

$=$

$\vec{\mu}_j^{(t+1)}$, $\sigma_j^{2\,(t+1)} I$ and $\pi_j^{(t+1)}$ are the M-step update parameters for a Gaussian mixture model.