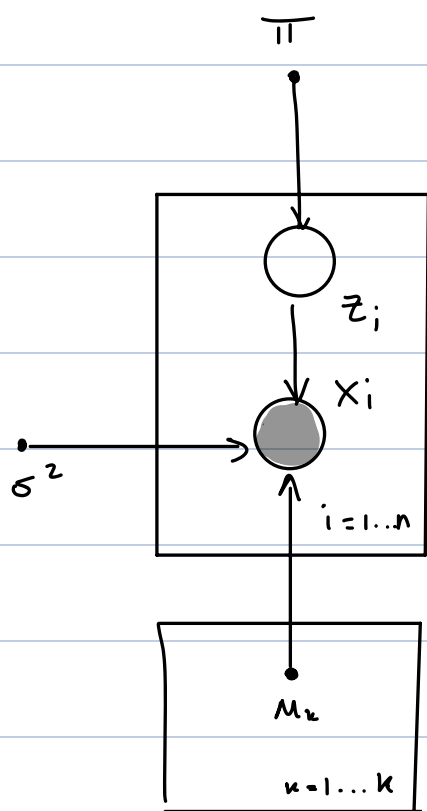
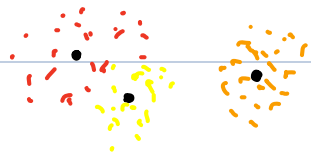


# (Bayesian) Mixture Models

Gaussian mixture model

$$P(x_i | z_i = k) = \mathcal{N}(x_i; \mu_k, \sigma^2)$$

$$P(z_i = k) = \pi_k$$



Marginal likelihood

$$\theta \triangleq \{\pi, \sigma^2, \mu_1, \dots, \mu_K\}$$

$$P(x_{1:n}; \theta) = \sum_{z_1=1}^K \dots \sum_{z_n=1}^K P(x_{1:n}, z_{1:n}; \theta)$$

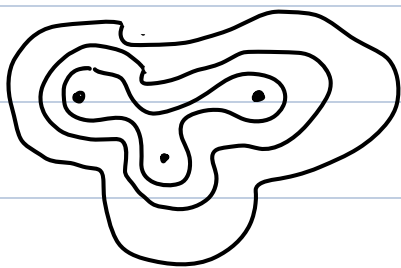
(Will this be tractable?)

$$= \sum \dots \sum \prod_i P(x_i, z_i)$$

$$= \prod_i \sum_{z_i} P(x_i, z_i)$$

$$= \prod_i \sum_k \pi_k \mathcal{N}(x_i; \mu_k, \sigma^2)$$

(Yes, why? graph = tree.)



The posterior will be tractable too.

$$P(z_{1:n} | x_{1:n}) = \frac{P(z_{1:n}, x_{1:n})}{P(x_{1:n})} = \frac{\prod_i P(z_i, x_i)}{\prod_i P(x_i)} = \prod_i P(z_i | x_i)$$

Now  $\mu_1 \dots \mu_k$  are unknown. Say we have a prior:

$$p(\mu_k | \mu_0) = \mathcal{N}(\mu_k; \mu_0, 1)$$

Marginal likelihood

$$\theta \triangleq \{\pi, \mu_0, \sigma^2\}$$

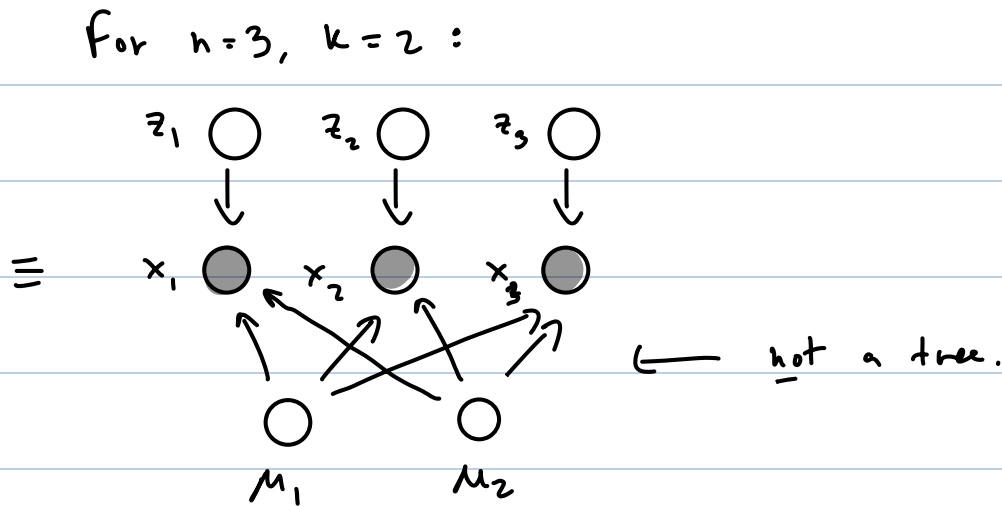
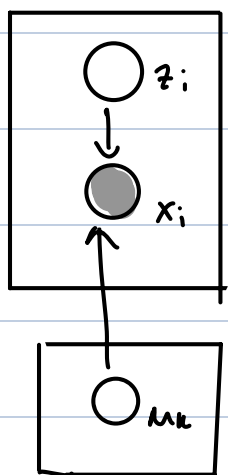
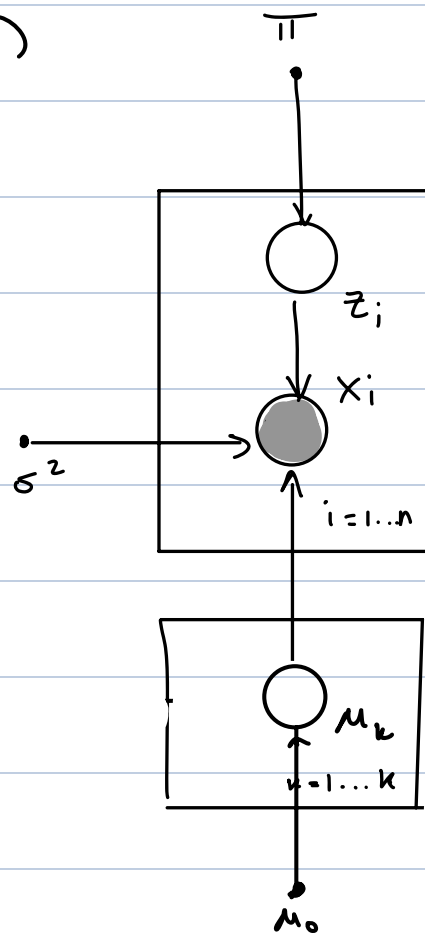
$$p(x_{1:n} | \theta) = \sum_{z_1} \dots \sum_{z_n} p(x_{1:n}, z_{1:n} | \theta)$$

Will this be tractable?

$$\neq \sum \dots \sum \prod_i p(x_i, z_i | \theta)$$

why?  $x_i \not\perp x_j | z_{1:n}, \theta$

How do we see that in the graph?



$$p(x_{1:n} | \theta) = \sum_{z_1} p(z_1) \dots \sum_{z_n} p(z_n) \underbrace{p(x_{1:n} | z_{1:n} | \theta)}$$

$$= \sum_{z_1} p(z_1) \dots \sum_{z_n} p(z_n) \left( \int \dots \int \prod_i \pi_i \mathcal{N}(x_i; \mu_{z_i}, \sigma^2) d\mu_1 \dots d\mu_k \right)$$

We can rewrite the integrals, since  $z$  partitions  $x$ :

$$= \sum_{z_{1:n}} P(z_{1:n}) \prod_{k=1}^K \left[ \int P(\mu_k) \prod_{i: z_i = k} \mathcal{N}(x_i; \mu_k) d\mu_k \right]$$

this term is tractable due to conjugacy

... however the sums over  $z_1, \dots, z_n$  do not push in, so we have to consider all  $x^n$  assignments.

Consequence: the joint posterior  $P(z_{1:n}, \mu_{1:n} | x_{1:n})$  is intractable.

---

## EM for MAP estimation

Goal:  $\hat{\mu}_{1:k} = \operatorname{argmax}_{\mu_{1:k}} P(\mu_{1:k} | x_{1:n})$

$$P(\mu | x) \propto_{\mu} P(x, \mu) = \sum_z P(x, z, \mu)$$

$$\operatorname{argmax}_{\mu_{1:k}} P(\mu_{1:k} | x_{1:n}) = \operatorname{argmax}_{\mu_{1:k}} \sum_{z_{1:n}} P(x_{1:n}, z_{1:n}, \mu_{1:n})$$

ELBO:

$$\log P(x, \mu) \geq \mathbb{E}_{Q(z)} \left[ \log \frac{P(x, z, \mu)}{Q(z)} \right]$$

$$= B(Q, \mu)$$

① E-step :  $Q(z) = P(z | x, \mu)$

② M-step :  $\mu = \underset{\mu}{\operatorname{argmax}} \mathbb{E}_Q [\log P(x, z, \mu)]$

M-step

$$\mathbb{E}_Q [\log P(x, z, \mu)]$$

$$= \mathbb{E}_Q [\log P(z) P(\mu) P(x | z, \mu)]$$

$$\propto_{\mu} \mathbb{E}_Q [\log P(\mu) P(x | z, \mu)]$$

$$= \log P(\mu) + \mathbb{E}_Q [\log P(x | z, \mu)]$$

objective for MLE w/ EM

$$= \log P(\mu) + \mathbb{E}_Q [\log \prod_i \prod_k P(x_i | \mu_k)^{1(z_i=k)}]$$

$$= \log P(\mu) + \sum_i \sum_k \underbrace{\mathbb{E}_Q [1(z_i=k)]}_{\text{"belief"}} \log P(x_i | \mu_k)$$

in mixture models the beliefs are called "responsibilities"

$$r_{ik} = \mathbb{E}_Q [1(z_i=k)] = Q(z_i=k)$$

$$= \log \prod_k \mathcal{N}(\mu_k; \mu_0, 1) + \sum_i \sum_k r_{ik} \log \mathcal{M}(x_i; \mu_k, \sigma^2)$$

$$\propto \sum_{1:k} \left[ -\frac{1}{2} (\mu_k - \mu_0)^2 \right] + \sum_i \sum_k r_{ik} \left[ -\frac{1}{2\sigma} (x_i - \mu_k)^2 \right]$$

$$\propto \sum_{1:k} \left[ -\frac{1}{2} (\mu_k^2 - 2\mu_k \mu_0) \right] + \sum_i \sum_k r_{ik} \left[ -\frac{1}{2\sigma} (\mu_k^2 - 2\mu_k x_i) \right]$$

Consider only one:

$$\propto \mu_k - \frac{1}{2} \mu_k^2 + \mu_k \mu_0 - \frac{1}{2\sigma} \mu_k^2 \sum_i r_{ik} + \frac{1}{\sigma} \mu_k \sum_i r_{ik} x_i$$

$$\propto \mu_k \left( \mu_0 + \frac{1}{\sigma} \sum_i r_{ik} x_i \right) - \frac{1}{2} \mu_k^2 \left( 1 + \frac{1}{\sigma} \sum_i r_{ik} \right)$$

$$\frac{\partial}{\partial \mu_k} \dots = \mu_0 + \frac{1}{\sigma} \sum_i r_{ik} x_i - \mu_k \left( 1 + \frac{1}{\sigma} \sum_i r_{ik} \right)$$

$$0 = \dots$$

$$\mu_k \left( 1 + \frac{1}{\sigma} \sum_i r_{ik} \right) = \mu_0 + \frac{1}{\sigma} \sum_i r_{ik} x_i$$

$$\mu_k = \frac{\mu_0 + \frac{1}{\sigma} \sum_i r_{ik} x_i}{1 + \frac{1}{\sigma} \sum_i r_{ik}}$$

$$= \frac{\sigma \mu_0 + \sum_i r_{ik} x_i}{\sigma + \sum_i r_{ik}}$$

E-step:

$$\sigma + \sum_i r_{ik}$$

$$r_{ik} = Q(z_i = k) = P(z_i = k \mid x, \mu)$$

$$\begin{aligned} & \frac{\pi_k \mathcal{N}(x_i; \mu_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(x_i; \mu_{k'})} \\ &= \end{aligned}$$

Another way to derive the M-step:

$$\mathbb{E}_Q [\log P(\mu, x, z)]$$

$$\propto_{\mu} \mathbb{E}_Q [\log P(\mu \mid x, z)]$$

Since  $z$  partitions  $x$ :

$$P(\mu_{1:k} \mid x_{1:n}, z_{1:n}) = \prod_k P(\mu_k \mid \{x_i : z_i = k\})$$

complete conditional of  $\mu_k$

Normal-normal conjugacy

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

$$x_i \sim \mathcal{N}(\mu, \sigma^2) \\ i = 1 \dots n$$

$$P(\mu \mid \dots) = \mathcal{N}(\mu_n; \mu_n, \sigma_n^2)$$

$$c_n \triangleq \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}$$

$$\mu_n \triangleq c_n \mu_0 + (1 - c_n) \bar{x}$$

$$\sigma_n^2 \triangleq \left( 1/\sigma_0^2 + n/\sigma^2 \right)^{-1}$$

## Conjugate Exponential Mixture model

$$P(z_i = k) = \pi_k$$

$$G(x_i | z_i = k, \eta_k) = h_{\ell}(x_i) \exp(\eta_k^T t_{\ell}(x_i) - a_{\ell}(\eta_k))$$

$$F(\eta_k | \lambda) = h_c(\eta_k) \exp(\lambda_1^T \eta_k + \lambda_2(-a_{\ell}(\eta_k)) - a_c(\lambda))$$

Complete conditional:

$$g_{ik} \triangleq 1(z_i = k)$$

$$P(\eta_k | x_{1:n}, z_{1:n}, \lambda) \propto$$

$$h_c(\eta_k) \exp(\lambda_1^T \eta_k + \lambda_2(-a_{\ell}(\eta_k)))$$

$$\prod_i \exp(\eta_k^T t_{\ell}(x_i) - a_{\ell}(\eta_k))^{g_{ik}}$$

$$\propto h_c(\eta_k) \exp(\eta_k^T [\lambda_1 + \sum_i t_{\ell}(x_i) g_{ik}] - a_{\ell}(\eta_k) [\lambda_2 + \sum_i g_{ik}])$$

$$\propto F(\eta_k; \lambda_{k,n}), \quad \lambda_{k,n} = \begin{bmatrix} \lambda_1 + \sum_i t_{\ell}(x_i) g_{ik} \\ \lambda_2 + \sum_i g_{ik} \end{bmatrix}$$

M-step:

$$\operatorname{argmax}_{\eta_k} \mathbb{E}_Q [\log p(\eta, x_{1:n}, z_{1:n} | \lambda)]$$

$$\propto \mathbb{E}_Q [\log p(\eta_k | x_{1:n}, z_{1:n}, \lambda)]$$

$$= \mathbb{E}_Q \left[ \log h_c(\eta_k) \exp \left( \eta_k^T \lambda_{x,n,1} - a_c(\eta_k) \lambda_{z,n,2} \right) \right]$$

$$= \mathbb{E}_Q [\log h_c(\eta_k)] + \eta_k^T \mathbb{E}_Q [\lambda_{x,n,1}] - a_c(\eta_k) \mathbb{E}_Q [\lambda_{z,n,2}]$$

$$\mathbb{E}_Q [\lambda_{x,n,1}] = \lambda_1 + \sum_i t_{x,i} \mathbb{E}_Q [f_{ik}]$$

"  $r_{ik}$

$$\mathbb{E}_Q [\lambda_{z,n,2}] = \lambda_2 + \sum_i \mathbb{E}_Q [p_{ik}]$$

Even better way:

$$\operatorname{argmax}_{\eta_k} \mathbb{E}_Q [\log p(\eta, x_{1:n}, z_{1:n} | \lambda)]$$

$$= \operatorname{argmax}_{\eta} \exp \left( \mathbb{E}_Q [\log p(\eta | x_{1:n}, z_{1:n}, \lambda)] \right)$$

$$= \operatorname{argmax}_{\eta} \exp \left( \log h_c(\eta_k) + \dots \right)$$



$$= \underset{\eta}{\text{argmax}} h_c(\eta_k) \exp \left( t_c(\eta_k)^T \mathbb{E}_Q[\lambda_{n,k}] \right)$$

$$= \underset{\eta}{\text{argmax}} F(\eta_k; \mathbb{E}_Q[\lambda_{n,k}])$$

↑  
usually we know the mode

e.g. Gaussian  $\mathcal{N}(m, s^2)$

$[m s^{-2}, s^{-2}]$  is the nat. param

Solve for the  $m, \sigma$  corresponding to  $\mathbb{E}_Q[\lambda_{n,k}]$

$$m \sigma^{-2} = \mathbb{E}_Q[\lambda_{n,k,1}] = \lambda_1 + \sum_i t(x_i) r_{ik}$$

$$\sigma^{-2} = \mathbb{E}_Q[\lambda_{n,k,2}] = \lambda_2 + \sum_i r_{ik}$$

$$m (\lambda_2 + \sum_i r_{ik}) = \lambda_1 + \sum_i t(x_i) r_{ik}$$

$$m = \frac{\mu_0 \sigma_0^{-2} + \sigma^{-1} \sum_i x_i r_{ik}}{\sigma_0^{-2} + \sum_i r_{ik}}$$

$$\lambda_1 = \mu_0 \sigma_0^{-2}$$

$$\lambda_2 = \sigma_0^{-2}$$

$$t(x_i) = x_i / \sigma$$

What does this tell us?

$$\begin{aligned}
 & \text{avgmax}_{\mu_k} \mathbb{E}_{\varphi} [\log p(x_{1:n}, z_{1:n}, \mu_k)] \\
 &= \text{avgmax}_{\mu_k} \exp \left( \mathbb{E}_{\varphi} [\log p(\mu_k | \dots)] \right) \\
 &= \text{avgmax}_{\mu_k} \mathcal{N}(\mu_k; \mu = \frac{\mu_0 \sigma_0^{-2} \sum_i x_i r_{ik}}{\sigma_0^{-2} + \sum_i r_{ik}}, \sigma^2 = \dots) \\
 &= \frac{\mu_0 \sigma_0^{-2} \sum_i x_i r_{ik}}{\sigma_0^{-2} + \sum_i r_{ik}} \quad (\text{easy :}) \\
 & \quad (\text{same answer as above for } \sigma_0 = 1).
 \end{aligned}$$

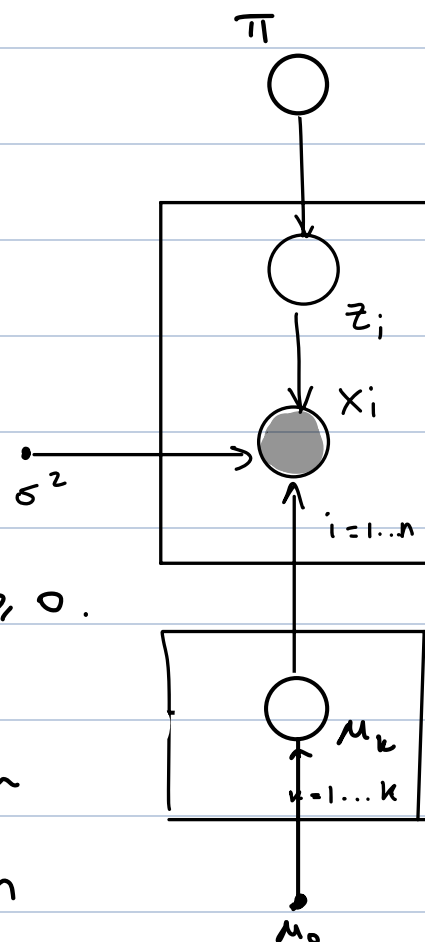
We typically don't know  $\pi$ .

What is an appropriate prior?

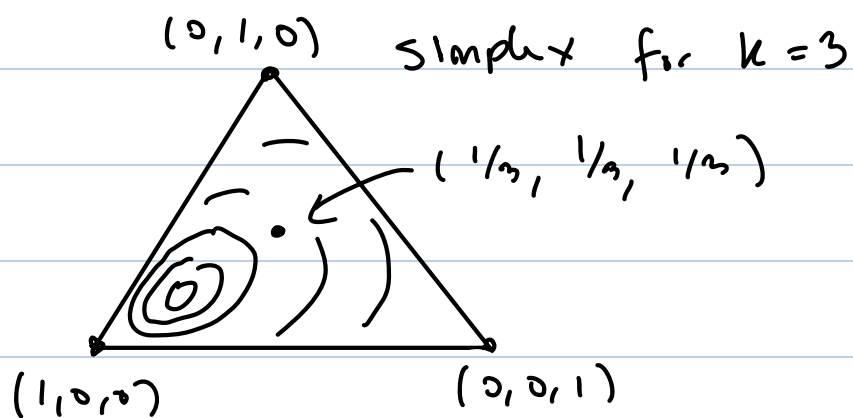
$\pi$  is a simplex vector.

$$\pi \in \Delta_k : \sum_k \pi_k = 1, \pi_k \geq 0.$$

Distribution with support on the simplex: Dirichlet distribution



## Dirichlet



$$X \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k), \alpha_k > 0$$

↓

$$X \equiv (x_1, \dots, x_k), \quad \sum_k x_k = 1, \quad x_k > 0$$

$$\mu = \left( \frac{\alpha_1}{\sum_k \alpha_k}, \dots, \frac{\alpha_k}{\sum_k \alpha_k} \right) \text{ mean}$$

$$c = \sum_k \alpha_k \quad \text{"concentration"}$$

$$P(X | \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k x_k^{\alpha_k - 1}$$

For  $k=2$ :

$$\frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) \Gamma(\alpha_2)} x_1^{\alpha_1 - 1} (1 - x_1)^{\alpha_2 - 1} \equiv \text{Beta}(x_1; \alpha_1, \alpha_2)$$

Generalization of the Beta distribution.

## Dirichlet - Multinomial conjugacy

$$\vec{\pi} \sim \text{Dir}(\vec{\alpha}) \quad (\vec{\alpha} \text{ is a } k\text{-vector})$$

$$\vec{y} \sim \text{Mult}(N, \vec{\pi})$$

$$P(\pi | \dots) \propto_{\pi} \text{Dir}(\pi, \alpha) \text{Mult}(y; N, \pi)$$

$$\propto_{\pi} \prod_k \pi_k^{y_k + \alpha_k - 1}$$

↑  
kernel of Dirichlet( $\pi; \alpha + y$ )

Multinomial and Dirichlet are both exp-fam.

The marginal generates the Beta-binomial:

$$P(y) = \text{DirMult}(N, \vec{\alpha}) = \frac{\Gamma(\sum_j \alpha_j) N!}{\Gamma(N + \sum_j \alpha_j) \prod_j \Gamma(\alpha_j)} \prod_j \frac{\Gamma(y_j + \alpha_j)}{\Gamma(\alpha_j) y_j!}$$

## Multinomial vs. Multinoulli vs. categorical

$$\parallel P(Z = k) = \pi_k \Rightarrow Z \sim \text{categorical}(\pi)$$

$$\parallel \begin{array}{l} g_k = 1(Z=k) \\ g = (g_1, \dots, g_k) \end{array} \Rightarrow g \sim \text{Multinoulli}(\pi)$$

$$\parallel \begin{array}{l} y = \sum_{j=1}^N g_j \\ g_j \text{ iid Multinoulli} \end{array} \Rightarrow y \sim \text{Multinomial}(\pi)$$

## Expfam mixture, unknown $\pi$

$$\pi \sim \text{Dir}(\alpha)$$

$$\eta_k \sim F(\lambda)$$

$$j_i \sim \text{Multinoulli}(\pi)$$

$$X_i | z_i = k \sim G(\eta_k)$$

MAP for  $\pi, \eta_1, \dots, \eta_k$  with EM.

M-step for  $\pi$ :

$$\underset{\pi}{\text{argmax}} \mathbb{E}_Q [\log p(x, z, \eta, \pi)]$$

$$= \underset{\pi}{\text{argmax}} \exp \left( \mathbb{E}_Q [\log p(\pi | \dots)] \right)$$

$$= \underset{\pi}{\text{argmax}} \exp \left( \mathbb{E}_Q [\text{Dir}(\alpha + \sum_i j_i)] \right)$$

Dirichlet's natural parameter is just  $\alpha \dots$

$$= \underset{\pi}{\text{argmax}} \text{Dir}(\alpha + \sum_i \mathbb{E}_Q [j_i])$$

The mode of a Dirichlet is well-defined if  $\alpha > 1$ :

$$\pi_k^* = \alpha_k + \sum_i \mathbb{E}_Q [j_{ik}] - 1$$

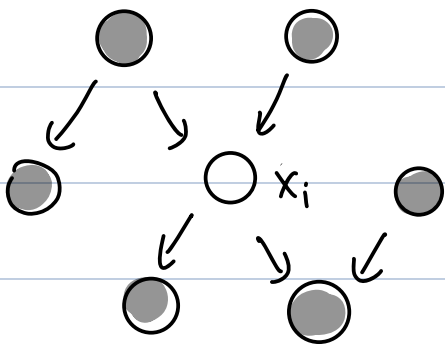
## Conditional conjugacy

We've just seen how conditional conjugacy is useful.

e.g.  $P(\mu)$  and  $P(x | \mu)$  are not conjugate

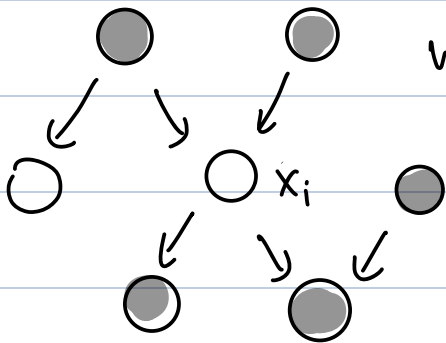
$P(\mu)$  and  $P(x | \mu, z)$  are conjugate

A related idea is the complete conditional



$P(x_i | x_{\setminus i})$  dist of a variable  $x_i$   
given values of all other variables  $x_{\setminus i}$ .

$$P(x_i | x_{\setminus i}) = P(x_i | x_{\text{Markov}(i)})$$



We only need the Markov blanket of  $x_i$

$\text{Markov}(i) =$

$$\text{Parents}(i) \cup \text{Children}(i) \cup \text{Co-parents}(i)$$

We can more easily define this in the  
undirected ("moralized") graph.

Example :

$$\gamma_1 \sim \Gamma(a, b), \quad \gamma_2 \sim \Gamma(a, b)$$

$$y \sim \text{Pois}(\gamma_1 \gamma_2)$$

$$p(\gamma_1 | y) \propto p(\gamma_1) \int \text{Pois}(y; \gamma_1 \gamma_2) p(\gamma_2) d\gamma_2$$

$$\propto \Gamma(\gamma_1; a, b) \text{NB}(y; a, \frac{\gamma_1}{\gamma_1 + b})$$

$$\propto \gamma_1^{a-1} \exp(-b\gamma_1) \left(1 - \frac{\gamma_1}{\gamma_1 + b}\right)^a \left(\frac{\gamma_1}{\gamma_1 + b}\right)^y$$

Not conjugate...

$$p(\gamma_1 | y, \gamma_2) \propto \Gamma(\gamma_1; a, b) \text{Pois}(y; \gamma_1 \gamma_2)$$

$$\propto \gamma_1^{a-1} \exp(-b\gamma_1) (\gamma_1 \gamma_2)^y \exp(-\gamma_1 \gamma_2)$$

$$\propto \gamma_1^{y+a-1} \exp(-\gamma_1(b + \gamma_2))$$

$$\propto \Gamma(\gamma_1; a + y, b + \gamma_2)$$

Conditionally conjugate!

How about :

$$p_1 \sim \text{Beta}(a, b), \quad p_2 \sim \text{Beta}(a, b)$$

$$y \sim \text{Binom}(n, p_1 p_2)$$

Will this be conditionally conjugate?

# Gibbs Sampling

initialize  $z_1^0, \dots, z_n^0, \mu_{1:k}^0, \pi^0$

for iteration  $m = 1, 2, \dots, M$ :

$$z_{1:n}^m \sim p(z_{1:n} \mid \mu_{1:k}^{m-1}, \pi^{m-1}, x_{1:n})$$

$$\mu_{1:k}^m \sim p(\mu_{1:k} \mid z_{1:n}^m, x_{1:n})$$

$$\pi^m \sim p(\pi \mid z_{1:n}^m, x_{1:n})$$

(This should feel like EM.)

This returns a set of samples

$$\left\{ z_{1:n}^m, \mu_{1:k}^m, \pi^m \right\}_{m=1}^M$$

Claim:

$$\lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M \mathbb{1}(\mu_k^m \in A) = p(\mu_k \in A \mid x_{1:n})$$

for any subset  $A$

Another way of saying this is that:

$$\lim_{M \rightarrow \infty} P_r(\mu_k^M) = p(\mu_k \mid x_{1:n})$$

More generally:

$$\lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M f(z_{1:n}^m, \mu_{1:k}^m, \pi^m)$$

$$= \mathbb{E} \left[ f(z_{1:n}, \mu_{1:k}, \pi) \mid x_{1:n} \right]$$

Posterior  $\nearrow$



Conceptually our algorithm is a Markov chain.

$$\text{State: } S^m \triangleq (z_{1:n}^s, \mu_{1:k}^s, \pi^s)$$

$$\text{Transition operator: } T(S^{m-1} \rightarrow S^m)$$

$$S^m \sim T(S^{m-1} \rightarrow S^m)$$

in this case:

$$(z^m, \mu^m, \pi^m) \sim \underbrace{P(z, \mu, \pi \mid \overbrace{z^{m-1}, \mu^{m-1}, \pi^{m-1}}^{\equiv S^{m-1}}, x)}_{\equiv T(S^{m-1} \rightarrow S^m)}$$

$$P_r(S^m) = \int T(S^{m-1} \rightarrow S^m) P(S^{m-1}) dS^{m-1}$$

The important aspect of this Markov chain is

its stationary distribution  $P^*(s)$ .

$$P^*(s) = \int T(s' \rightarrow s) P^*(s') ds'$$

$$\equiv \mathbb{E}_{s' \sim P^*} [T(s' \rightarrow s)]$$

"if you start in  $P^*$ , you never leave!"

In this case: (the exact posterior)

$$P^*(\cdot) = P(\cdot \mid x_{1:n})$$

We'll see why next time...