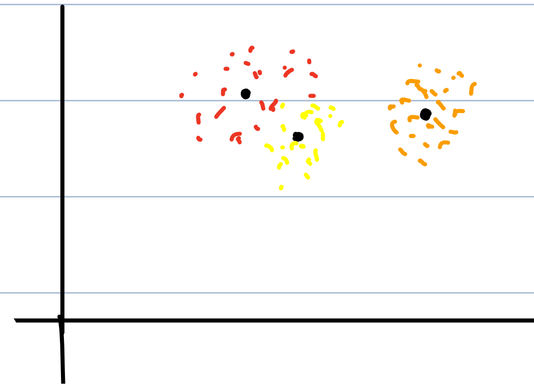
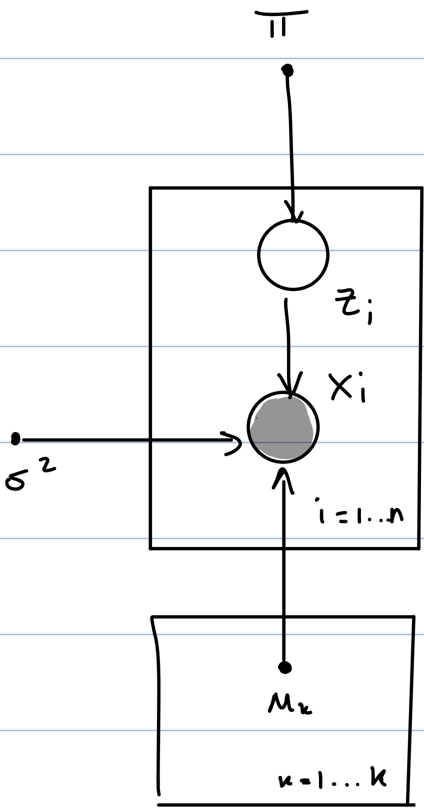
(Bayesian) Mixture Models



Gaussian mixture model

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

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



Marginal likelihood

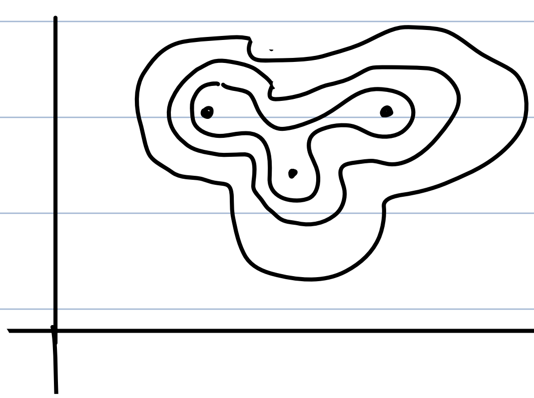
$$\theta \triangleq \{\pi, \sigma^2, \mu_1 \ldots \mu_K\}$$

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

(will this be tractable?)

$$= \sum \cdots \sum \prod_i p(x_i, z_i)$$

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



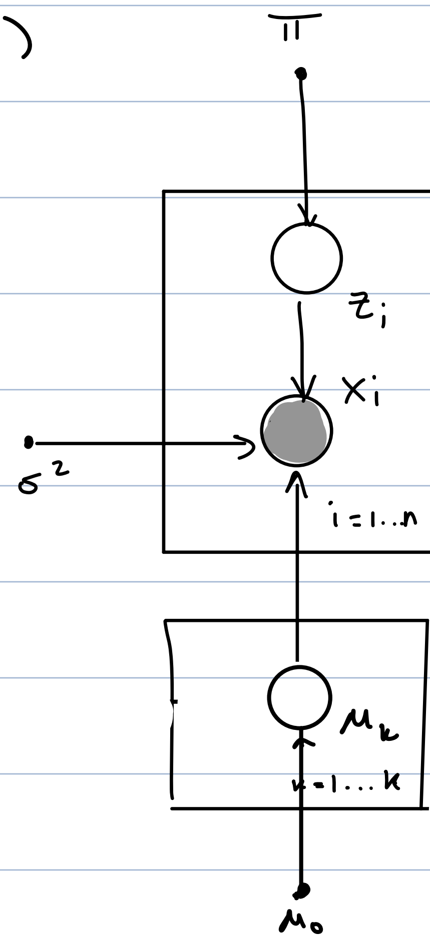
$$= \prod_i \sum_k \pi_k \mathcal{N}(x_i ; \mu_k, \sigma^2)$$

(yes, why? graph = tree.)

The posterior will be tractable too.

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

Now $\mu_1 \dots \mu_K$ are <u>unknown</u>. Say we have a prior:



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

Marginal likelihood

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

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

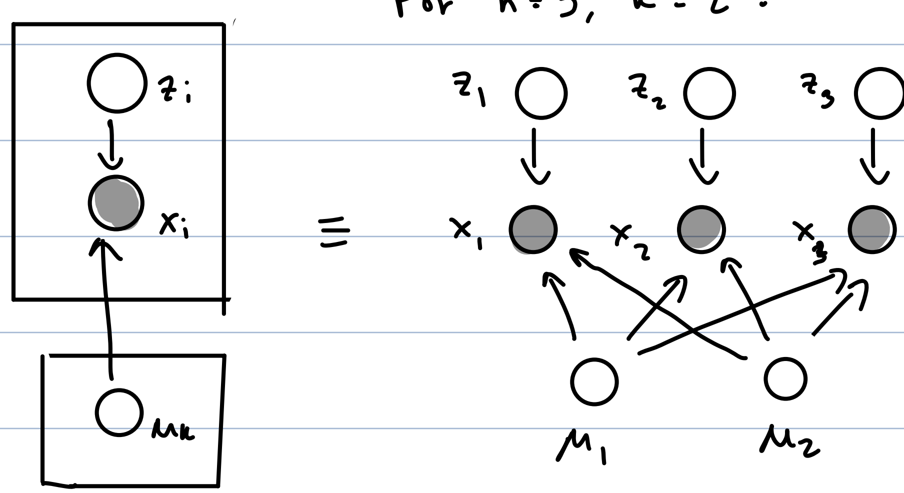
Will this be tractable?

$$\neq \sum \cdots \sum \prod_i p(x_i, z_i ; \theta)$$

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

How do we see that in the graph?

For $n=3$, $K=2$:



← <u>not</u> a tree.

$$p(x_{1:n} ; \theta) = \sum_{z_1} p(z_1) \cdots \sum_{z_n} p(z_n) \underbrace{p(x_{1:n} \mid z_{1:n} ; \theta)}$$

$$= \sum_{z_1} p(z_1) \cdots \sum_{z_n} p(z_n) \left[\int \cdots \int \prod_i \mathcal{N}(x_i ; \mu_{z_i}, \sigma^2)\, d\mu_1 \cdots d\mu_K \right]$$

We can rewrite the integrals, since z partitions x:

$$= \sum_{z_{1:n}} P(z_{1:n}) \prod_{k=1}^{K} \underbrace{\left[\int p(\mu_k) \prod_{i: z_i = k} \mathcal{N}(x_i ; \mu_k)\, d\mu_k \right]}_{\text{this term is tractable due to conjugacy}}$$

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

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

EM for MAP estimation

Goal: $\hat{\mu}_{1:k} = \underset{\mu_{1:k}}{\arg\max}\, P(\mu_{1:k} \mid x_{1:n})$

$$P(\mu \mid x) \underset{\mu}{\propto} p(x, \mu) = \sum_{z} P(x, z, \mu)$$

$$\underset{\mu_{1:k}}{\arg\max}\, P(\mu_{1:k} \mid x_{1:n}) = \underset{\mu_{1:k}}{\arg\max} \sum_{z_{1:n}} P(x_{1:n}, z_{1:n}, \mu_{1:n})$$

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 : $Q(z) = p(z \mid x, \mu)$

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

M-step

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

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

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

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

objective for MLE w/ EM

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

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

in mixture models the beliefs are called "responsibilities"

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

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

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

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

Consider only one:

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

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

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

$$0 = \cdots$$

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

$$\mu_k = \frac{\sigma^2 \mu_0 + \sum_i r_{ik} x_i}{\sigma^2 + \sum_i r_{ik}}$$

E-step:

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

$$= \frac{\pi_k \mathcal{N}(x_i ; \mu_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(x_i ; \mu_k)}$$

Another way to derive the M-step:

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

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

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 μ_k

<u>Normal-normal conjugacy</u>

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

$x_i \sim \mathcal{N}(\mu, \sigma^2)$

$i = 1 \dots n$

$$p(\mu \mid \cdots) = \mathcal{N}(\mu_k ; \mu_n, \sigma_n^2)$$

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

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

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

Conjugate Expfam Mixture model

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

$$G(x_i \mid z_i = k, \eta_k) = h_\ell(x_i) \exp\left(\eta_k^T t_\ell(x_i) - a_\ell(\eta_k) \right)$$

$$F(\eta_k \mid \lambda) = h_c(\eta_k) \exp\left(\lambda_1^T \eta_k + \lambda_2(-a_\ell(\eta_k)) - a_c(\lambda) \right)$$

Complete conditional:

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

$$p(\eta_k \mid x_{1:n}, z_{1:n}, \lambda) \propto$$

$$h_c(\eta_k) \exp\left(\lambda_1^T \eta_k + \lambda_2 (- a_\ell(\eta_k)) \right)$$

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

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

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

M-step:

$$\underset{\eta_k}{\arg\max}\ \mathbb{E}_Q[\log p(\eta, x_{1:n}, z_{1:n} \mid \lambda)]$$

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

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

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

$$\mathbb{E}_Q[\lambda_{k,n,1}] = \lambda_1 + \sum_i t_e(x_i)\, \underbrace{\mathbb{E}_Q[z_{ik}]}_{\equiv\, r_{ik}}$$

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

Even better way:

$$\underset{\eta_k}{\arg\max}\ \mathbb{E}_Q\left[\log p(\eta, x_{1:n}, z_{1:n} \mid \lambda)\right]$$

$$= \underset{\eta}{\arg\max}\ \exp\left(\mathbb{E}_Q[p(\eta \mid x_{1:n}, z_{1:n}, \lambda)]\right)$$

$$= \underset{\eta}{\arg\max}\ \exp\left(\log h_c(\eta_k) + \dots\right)$$

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

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

usually we know the mode

e.g. Gaussian $\mathcal{N}(\mu, \sigma^2)$

$[\mu\sigma^{-2}, \sigma^{-2}]$ is the nat. param

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

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

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

$$\mu\left(\lambda_2 + \sum_i r_{ik}\right) = \lambda_1 + \sum_i x_i r_{ik}$$

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

$$\boxed{\begin{aligned} \lambda_1 &= \mu_0\sigma_0^{-2} \\ \lambda_2 &= \sigma_0^{-2} \end{aligned}}$$

What does this tell us?

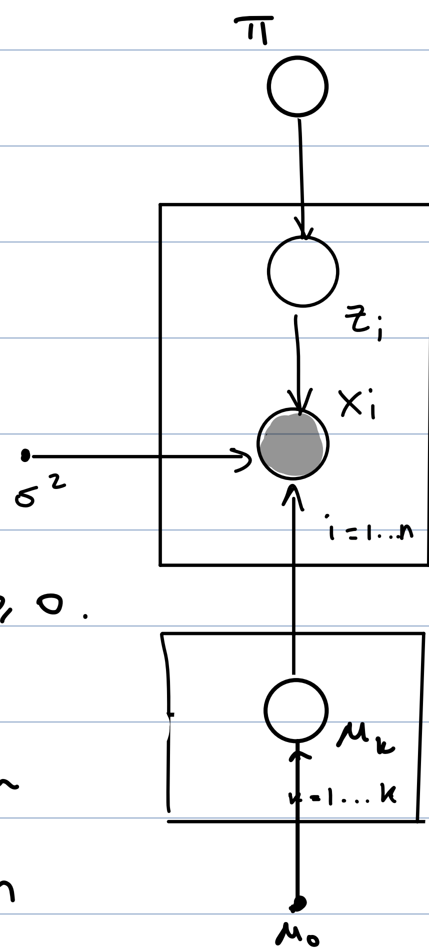
$$\underset{\mu_k}{\text{argmax}} \ \mathbb{E}_q \left[\log p(x_{1:n}, z_{1:n}, \mu_k) \right]$$

$$= \text{argmax} \ \exp\left(\mathbb{E}_q \left[\log p(\mu_k \mid \cdots) \right] \right)$$

$$= \text{argmax} \ \mathcal{N}\left(\mu_k ; \mu = \frac{\mu_0 \sigma_0^{-2} \sum_i x_i r_{ik}}{\sigma_0^{-2} + \sum_i r_{ik}}, \sigma = \cdots \right)$$

$$= \frac{\mu_0 \sigma_0^{-2} \sum_i x_i r_{ik}}{\sigma_0^{-2} + \sum_i r_{ik}} \quad (\text{easy} \ \ddot\smile)$$

(same answer as above for $\sigma_0 = 1$).



We typically don't know π.

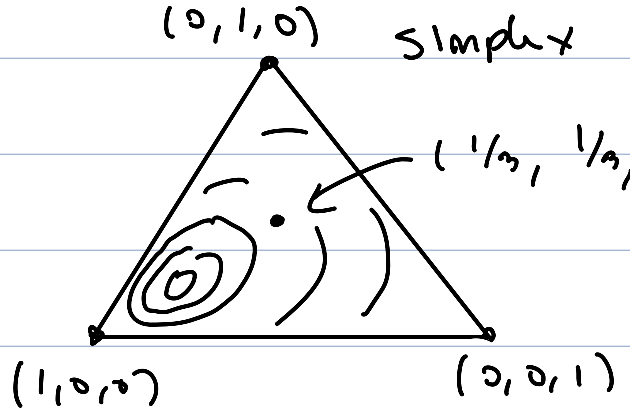
What is an appropriate prior?

π is a <u>simplex</u> vector.

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

Distribution with support on the simplex: <u>Dirichlet</u> distribution

Dirichlet



simplex for $k=3$

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

$$\Downarrow$$

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

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

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

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

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

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

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

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

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

kernel of $\text{Dirichlet}(\pi; \alpha + y)$

Multinomial and Dirichlet are both exp-form.

Multinomial vs. Multinoulli vs. Categorical

|| $p(z = k) = \pi_k \Rightarrow z \sim \text{Categorical}(\pi)$

|| $y_k = 1(z = k)$, $y = (y_1 \dots y_k)$ $\Rightarrow y \sim \text{Multinoulli}(\pi)$

|| $Y = \sum_{j=1}^{N} y_j$, $y_j \overset{iid}{\sim}$ Multinoulli $\Rightarrow Y \sim \text{Multinomial}(\pi)$

Expfam mixture, unknown π

$\pi \sim \mathrm{Dir}(\alpha)$ $\qquad$ $\eta_k \sim F(\lambda)$

$z_i \sim \mathrm{Multinoulli}(\pi)$ $\qquad$ $x_i \mid z_i = k \sim G(\eta_k)$

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

M-step for π:

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

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

$$= \underset{\pi}{\operatorname{argmax}} \; \exp\left(\mathbb{E}_Q [\mathrm{Dir}(\alpha + \textstyle\sum_i z_i)] \right)$$

Dirichlet's natural parameter is just α ...

$$= \underset{\pi}{\operatorname{argmax}} \; \mathrm{Dir}\left(\alpha + \textstyle\sum_i \mathbb{E}_Q [z_i]\right)$$

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

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

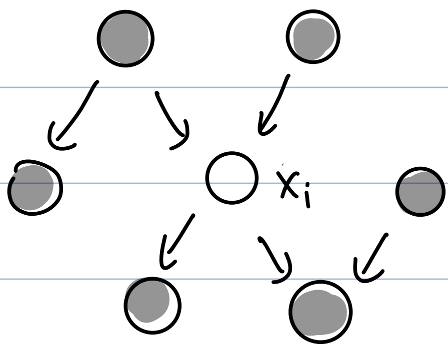
Conditional conjugacy

We've just seen how <u>conditional conjugacy</u> is useful.

e.g. $p(\mu)$ and $p(x \mid \mu)$ are not conjugate

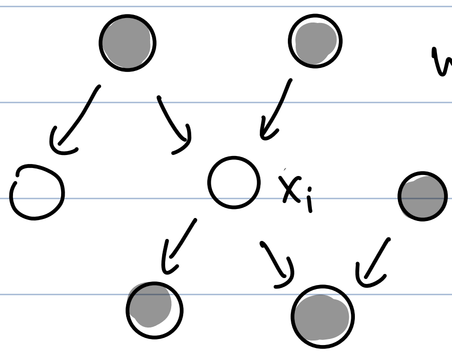
$p(\mu)$ and $p(x \mid \mu, z)$ <u>are</u> conjugate

A related idea is the <u>complete conditional</u>



$p(x_i \mid x_{\setminus i})$ dist of a variable x_i given values of <u>all</u> other variables $x_{\setminus i}$.

$$p(x_i \mid x_{\setminus i}) = p(x_i \mid x_{\text{Markov}(i)})$$



We only need the <u>Markov blanket</u> of x_i

Markov(i) =

Parents(i) $\cup$ Children(i) Coparents(i)

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

Example:

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

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

$$p(\gamma_1 \mid y) \propto p(\gamma_1) \int \mathrm{Pois}(y; \gamma_1 \gamma_2)\, p(\gamma_2)\, d\gamma_2$$

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

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

Not conjugate...

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

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

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

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

conditionally conjugate!

How about:

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

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

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

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

Claim:

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

for any subset A

Another way of saying this is that:

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

More generally:

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

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

Posterior

Conceptually our algorithm is a <u>Markov chain</u>.

State: $S^m \triangleq (z^s_{1:n}, \mu^s_{1:k}, \pi^s)$

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

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

in this case:

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

$$P_r(S^m) = \int T(S^{m-1} \to S^m)\, p(S^{m-1})\, dS^{m-1}$$

The important aspect of this Markov chain is its <u>stationary distribution</u> $p^*(s)$.

$$p^*(s) = \int T(s' \to s)\, p^*(s')\, ds'$$

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

"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...