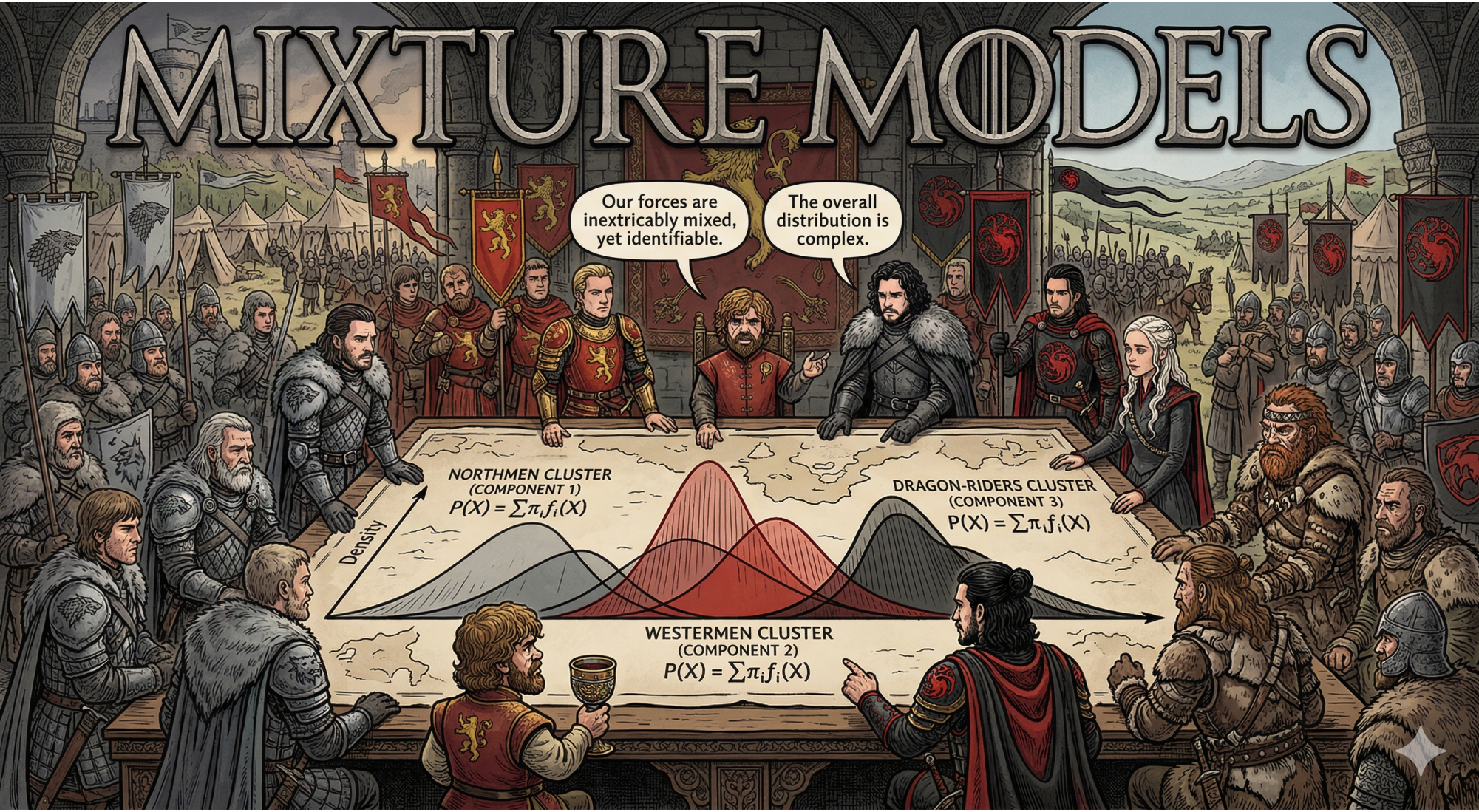
MIXTURE MODELS
Our forces are inextricably mixed, yet identifiable.
The overall distribution is complex.
NORTHMEN CLUSTER
(COMPONENT 1)
$P(X) = \sum \pi_i f_i(X)$
DRAGON-RIDERS CLUSTER
(COMPONENT 3)
$P(X) = \sum \pi_i f_i(X)$
Density
WESTERMEN CLUSTER
(COMPONENT 2)
$P(X) = \sum \pi_i f_i(X)$

Data augmentation

- In the last lecture, we discussed missing data induced by (usually) physical mechanism that led to data being not collected.
- In many problems without explicit missing data issue, we may still introduce reasonable auxiliary variables to help us make inference.
- Consider a simple probit regression problem: Let y_i be a binary response variable, and suppose we have a K level categorical covariate x_i. We want to estimate the prevalence of response in each level of the covariate using a model

$$p(y_i = 1 | x_i = k) = \Phi(\mu_k)$$

- The distribution of y can be thought of a mixture of K different distributions determined by the value of covariate x.
- With some prior for μ, the posterior distribution given the observed data is

$$p(\mu|x, y) \quad \propto \quad p(\mu) \prod_i \Phi(\mu_{x_i})^{y_i} (1 - \Phi(\mu_{x_i}))^{1-y_i}$$

  which is difficult to sample from.

Data augmentation in general

- We can reparameterize the problem and introduce a latent variable z_i such that

$$z_i|x_i = k \sim N(\mu_k, 1)$$

  and that $y_i = 1_{z_i>0}$.
- The (x, y, z) can be treated as the complete data and z the missing data.
- Inference can then be conducted with Gibbs sampling using

$$\begin{aligned} p(\mu, z|x, y) &\propto p(\mu) \prod_i p(z_i|\mu, x_i) p(y_i|z_i) \\ &= p(\mu) \prod_i N(z_i; \mu_{x_i}, 1) 1_{sign(z_i)=y_i} \end{aligned}$$

1. Sample $\mu_k|z, x, y \sim p(\mu_k|z) = N(\mu_k; \cdot, \cdot)$.
2. Sample $z_i|\mu, x, y \sim TruncNormal(\mu_{x_i}, 1, (0, \infty))$ if $y_i = 0$ and $TruncNormal(\mu_{x_i}, 1, (-\infty, 0))$ if otherwise.

Overview

- In the previous example, the marginal probability of y is

$$p(y = 1|\mu) = \sum_k p(x_i = k)\Phi(\mu_k)$$

- What if which group each observation come from, i.e., x_i in this example, is are unobserved, and we only know $p(x_i = k) = \pi_k$?
- More generally, we are dealing with mixture distributions in the form of

$$p(y|\pi, \theta) = \sum_{k=1}^{K} \pi_k f(y|\theta_k)$$

- This is known as the mixture model.

Gaussian mixtures

- Consider the model where $f(y|\theta_k)$ is a normal distribution with mean μ_k and variance Σ_k.
- Equivalently, we can introduce latent variable $z \sim Cat(\pi)$, and

$$y|z \sim N(\mu_z, \Sigma_z)$$

- If we marginalize out z, we obtain the mixture density

$$p(y|\pi, \theta) = \sum_{k=1}^{K} \pi_k N(y; \mu_k, \Sigma_k)$$

- Given $y_1, ..., y_n$ and a fixed K, the estimation of $\{\mu_k, \Sigma_k\}$ can be carried out using maximum likelihood estimation or Bayesian inference.

Gaussian mixture model (GMM)

- Consider the full model

$$
\begin{aligned}
z_i|\pi &\sim Cat(\pi) \\
\pi &\sim Dir(\alpha) \\
y_i|z_i,\pi,\mu,\Sigma &\sim N(\mu_{z_i},\Sigma_{z_i}) \\
\mu_1,...,\mu_K &\sim_{iid} N(\mu_0,\Sigma_0) \\
\Sigma_1,...,\Sigma_K &\sim_{iid} InvWishart(v,S)
\end{aligned}
$$

- Posterior sampling can be carried out using Gibbs steps:
  - $z_i|\cdots \sim Cat(\omega_i)$ where
$$\omega_{ik} = \frac{\pi_k N(y_i|\mu_k,\Sigma_k)}{\sum_k \pi_k N(y_i|\mu_k,\Sigma_k)}$$
  - $\pi|\cdots \sim Dir(\alpha_1+n_1,...,\alpha_K+n_K)$ where $n_k = \sum_i 1_{z_i=k}$.
  - $\mu_k|\cdots \sim N(m_k,V_k)$ where
$$V_k^{-1} = \Sigma_0^{-1} + n_k\Sigma_k^{-1}, \quad m_k = V_k(\Sigma_0^{-1}\mu_0 + \Sigma_k^{-1}\sum_{i:z_i=k} y_i)$$
  - $\Sigma_k|\cdots \sim InvWishart(v+n_k, S+\sum_{i,z_i=k}(y_i-\mu_k)(y_i-\mu_k)^T)$

Extensions and other inferential tools

- In certain situations, we might also consider working directly with the likelihood

$$p(y|\pi, \theta) = \sum_{k=1}^{K} \pi_k N(y; \mu_k, \Sigma_k)$$

  e.g., if you use HMC for inference.
- When the goal is clustering instead of paramter inference, it is also possible to integrate out all parameters (μ, Σ, π) and just sample he indicators z if the prior is fully conjugate.
- For high-dimensional data, it is usually useful to constrain the covariance matrix to be diagonal, as they are hard to estimate.
- We can in general plug in other non-Gaussian distributions for the mixture components.

Label switching

- The mixture likelihood is invariant to permutations of the labels. So there are $K!$ regions of parameter space with the same high posterior density.
- Suppose MCMC is mixing well enough, it should move between different permutations, right?
  - What would the sample average for μ_1 be in this case?
- MCMC sample averages of permutation-dependent quantities are usually meaningless if the chain mixes between different permutations.
- Gibbs sampler usually gets stuck to one of the $K!$ modes. One can argue that is the desired behavior.

Possible solutions

- Only average label-invariant quantities, e.g., instead of $p(z_i = j)$, consider $p(z_i = z_{i'})$.
- Relabel each MCMC to minimize a loss function that encourages similar points to be together.
- If some labels are available, use them as anchors.
- Informative priors.
- Target the mode instead of the distribution.
- ...

Examples: Mixture model with partially missing group membership

- Suppose we are interested in the prevalence of COVID-19 in a given population with N people. We randomly sampled n individuals to perform a PCR test. We then asked everyone to conduct a symptom questionnaire.
- Denote $s_i = 1$ if person i was selected for the PCR test. For notation simplicity, let $s_1, ..., s_n$ to be 1 and $s_{n+1}, ..., s_N$ to be 0.
- Denote $y_i = 1$ if person i was tested positive. We observe $y_1, ..., y_n$ only.
- Denote x_{ij} as the binary response to question j for person i.
- Consider the following model:

$$S_i \sim Bern(\lambda), \quad i = 1, ..., N$$
$$Y_i \sim Bern(\pi), \quad i = 1, ..., N$$
$$X_{ij}|Y_i = k \sim Bern(\theta_{kj}), \quad i = 1, ..., N, j = 1, ..., J, k = 0, 1$$

with independent $Beta(1, 1)$ priors on all parameters.

Example: ignorable missing in y

- The posterior distribution of the missing data and parameters is

$$\begin{aligned} p(Y_{mis}, \lambda, \pi, \theta | Y_{obs}, X, S) \quad \propto \quad & \prod_{i=1}^{N} \pi^{Y_i} (1-\pi)^{1-Y_i} \\ & \prod_{i=1}^{N} \prod_{j} \left(\theta_{0j}^{(1-Y_i)X_{ij}} (1-\theta_{0j})^{(1-Y_i)(1-X_{ij})} \right) \\ & \prod_{i=1}^{N} \prod_{j} \left(\theta_{1j}^{Y_i X_{ij}} (1-\theta_{1j})^{Y_i(1-X_{ij})} \right) \\ & \prod_{i} \left(\lambda^{S_i} (1-\lambda)^{1-S_i} \right) p(\lambda) p(\pi) p(\theta) \end{aligned}$$

- Notice λ can be integrated out from the posterior easily, so we do not need to worry about s and λ in general.
- The posterior conditionals can be derived analytically for a Gibbs sampler
  - Sample $Y_i | X_i, \pi, \theta \sim Bern(\cdots)$ for $i = n+1, ..., N$
  - Sample $\theta | X, Y \sim Beta(\cdots)$
  - Sample $\pi | Y \sim Beta(\cdots)$

Example: ignorable missing in x and y

- Suppose it turns out that the questionnaire has a null field for each question and some people mistakenly selected that as the answer.
- Let I_{ij} denote whether x_{ij} is recorded, and $O_i = \{j : I_{ij} = 1\}$. If we assume

$$p(I_i|X_i, Y_i, S_i, \rho) = p(I_i|X_{obs}, \rho)$$

with some distinct parameter ρ, we have

$$\begin{aligned} & p(Y_{mis}, \lambda, \pi, \theta|Y_{obs}, X_{obs}, S, I) \\ \propto & \prod_{i=1}^{N} \pi^{Y_i}(1-\pi)^{1-Y_i} \\ & \prod_{i=1}^{N} \prod_{j \in O_i} \left(\theta_{0j}^{(1-Y_i)X_{ij}} (1-\theta_{0j})^{(1-Y_i)(1-X_{ij})} \theta_{1j}^{Y_i X_{ij}} (1-\theta_{1j})^{Y_i(1-X_{ij})} \right) \\ & \prod_i \left(\lambda^{S_i}(1-\lambda)^{1-S_i} \right) p(\lambda)p(\pi)p(\theta) \prod_i p(I_i|X_{obs}, \rho) \end{aligned}$$

We can again ignore the components involving λ, S, I.

- The posterior conditionals are similar:
  - Sample $Y_i|X_i, \pi, \theta \sim Bern(\cdots)$ for $i = n+1, ..., N$
  - Sample $\theta|X, Y \sim Beta(\cdots)$
  - Sample $\pi|Y \sim Beta(\cdots)$

Example: ignorable missing in x and y

- Why we did not have X_{mis} in the previous posterior joint distribution? It is because

$$p(Y_{mis}, X_{mis}, \lambda, \pi, \theta | Y_{obs}, X_{obs}, S, I) =$$
$$p(Y_{mis}, \lambda, \pi, \theta | Y_{obs}, X_{obs}, S, I) p(X_{mis} | Y_{mis}, Y_{obs}, \theta)$$

- So we can integrate out X_{mis} from LHS analytically by discarding the second term on the RHS (as it integrates to 1.)
- Imputing X_{mis} is not needed for the inference of Y_{mis} and π because we can sample $Y_i | X_{obs}, \pi, \theta$ exactly. If the sampling of Y_{mis} is only computationally feasible conditional on X_{mis} then we should include the sampling of X_{mis}.
- Operationally, we can always sample X_{mis} at the end of our Gibbs steps based on the above factorization, but it will not contribute to the sampling of other parameters.
- Sampling X_{mis} can be a separate task of interest if our goal is imputation.

Example: Non-ignorable missing in y

- Now suppose that the data collector let you know that instead of a simple random sample, they conducted PCR tests only to university students in the population.
- Realizing that the prevalence within the observed samples and unobserved samples could differ due to the selection bias, we may change our model

$$\begin{aligned} S_i &\sim Bern(\lambda), \quad i = 1, ..., N \\ Y_i|S_i = s &\sim Bern(\pi_s), \quad i = 1, ..., N \\ X_{ij}|Y_i = k &\sim Bern(\theta_{kj}), \quad i = 1, ..., N, j = 1, ..., J, k = 0, 1 \end{aligned}$$

  with independent priors on π_0 and π_1.
- The population prevalence can be computed with

$$p(Y_i = 1) = \sum_s p(Y_i = 1|S_i = s)p(S_i = s) = (1 - \lambda)\pi_0 + \lambda\pi_1$$

  where λ is (usually) a fixed fraction that does not need to be estimated.
- However, we do not observe Y_i for any i with $S_i = 0$, so there is no information to make inference about π_0. Additional assumptions are necessary to relate π_0 to other parameters, e.g., π_1, and the inference depend critically on such assumptions.

Example: Non-ignorable missing in y

- The model we considered previously is usually referred to as pattern-mixture model in missing data literature.
- Alternatively, in more general settings, we may also consider a selection model

$$Y_i \sim Bern(\pi), \quad i = 1, ..., N$$
$$S_i|Y_i, X_i \sim Bern(\lambda(y_i, x_i)), \quad i = 1, ..., N$$

- Again, inference of this model depend critically on the parametric assumptions for $\lambda(y, z)$.
- Finally, some warnings:
  - There is no free lunch: MAR is an untestable assumption with observed data.
  - Modeling the missing data mechanism require a lot of care. We should always be careful what modeling assumptions we are making (e.g., $S \perp X|Y$ on the previous slide and parametric forms specifying $\lambda(y, z)$ above).
  - Usually sensitivity analysis is necessary when strong assumptions are made. There are many recent-ish literature on designing effective sensitivity analysis.