Hierarchical model for "8 schools" (from last time)

The generative process:

$$\tau^2 \sim \text{Inv-}\chi^2(\nu_0, \tau_0^2)$$

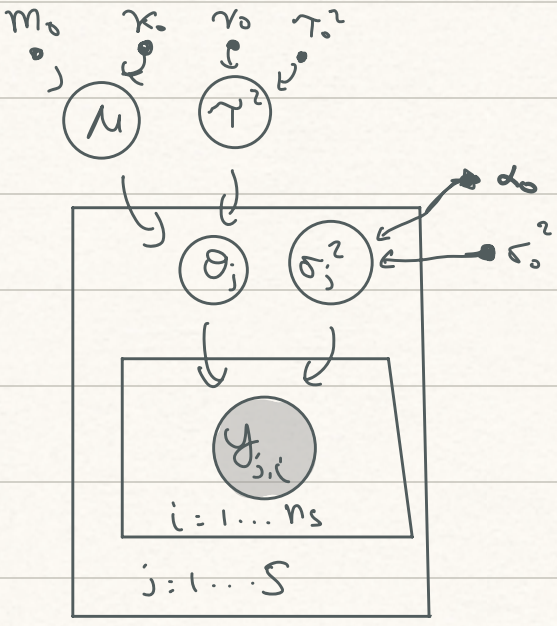
$$\mu \sim \mathcal{N}(m_0, 1/\kappa_0)$$

$$\sigma_j^2 \overset{iid}{\sim} \text{Inv-}\chi^2(\alpha_0, \sigma_0^2) \qquad j = 1 \dots S$$

$$\theta_j \overset{iid}{\sim} \mathcal{N}(\mu, \tau^2) \qquad j = 1 \dots S$$

$$y_{j,i} \overset{iid}{\sim} \mathcal{N}(\theta_j, \sigma_j^2) \qquad i = 1 \dots n_S$$

The probabilistic graphical model (PGM):



The set of latent variables is:

$$Z = \left\{ (\theta_j, \sigma_j^2)_{j=1}^S, \mu, \tau^2, \right\}, \qquad |Z| = D \; (= 2S + 2, \text{ in this case})$$

The set of hyperparameters is:

$$\eta_0 = \{ m_0, \kappa_0, \nu_0, \tau_0^2, \alpha_0, \sigma_0^2 \}$$

The joint posterior in this model is intractable

$$p(Z \mid Y, \eta_0) \quad \times$$

...but complete conditionals are tractable

$$p(Z_d \mid Z_{1d}, Y, \eta_0) \quad \checkmark$$

For example, by Gaussian-Gaussian conjugacy:

$$p(\mu \mid -) = \mathcal{N}(m_S, 1/\kappa_S)$$

$$m_S = w_S \frac{1}{S} \sum_{j=1}^{S} \theta_j + (1 - w_S) m_0$$

$$w_S = \frac{S/\tau^2}{\kappa_0 + S/\tau^2}$$

$$\kappa_S = \kappa_0 + S/\tau^2$$

The Gibbs sampler (today):

for $m = 1 \dots M$:

  for $d = 1 \dots D$:

    $Z_d^m \sim p(Z_d \mid -)$

Claim:

$$\lim_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} f(Z^m) = \mathbb{E}_{p(Z \mid Y)}[f(Z)]$$

PGM semantics

Notice that the complete conditional for mu depends on tau^2, theta_j, but not Y. Why?

$$p(\mu \mid -) = p(\mu \mid z_{\mu}, Y, \eta_0) = \frac{p(\mu, z_{\mu}, Y \mid \eta_0)}{\int \text{"} \; d\mu}$$

$$= \frac{p(\mu)\, p(\tau^2) \left[\prod_j p(\theta_j \mid \mu, \tau^2)\, p(\sigma_j^2)\right] \prod_{i,j} p(y_{ij} \mid \theta_j, \sigma_j^2)}{\int \text{"} \; d\mu}$$

$$= \frac{\cancel{\text{[blob]}}\ \text{[blob]}}{\cancel{\int \text{[blob]}\, d\mu}} = \frac{\text{[blob]}}{\int \text{[blob]}\, d\mu}$$

The kernel of the posterior involves tau^2 and theta_j but not Y; this is due to the conditional independence assumptions of the model.

e.g. $y_{ij} \perp\!\!\!\perp \mu \mid \theta_j, \sigma_j^2$

PGMs encode the conditional independences in a model. In general, the joint distribution is:

$$p(z, y) = \left[\prod_{d=1}^{D} p(z_d \mid z_{par(z_d)})\right] p(y \mid z_{par(y)})$$

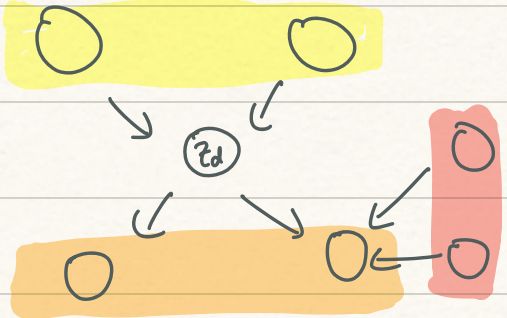
$$par(z_d) = \{ z_{d'} : z_{d'} \to z_d \in graph \}$$

"parents" of Zd

A variable is conditionally independent of all other variables given its Markov blanket:

$$MB(z_d) = \{ par(z_d) \cup child(z_d) \cup coparents(z_d) \}$$

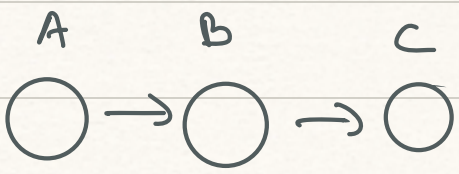
"parents" **"children"** **"co-parents"**



$$z_d \perp\!\!\!\perp z_{d'} \mid MB(z_d)$$

$$\forall\ z_{d'} \notin MB(z_d)$$

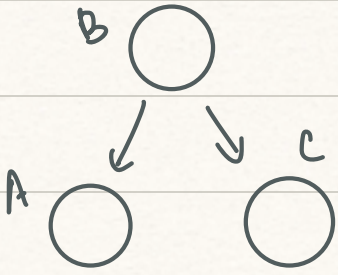
Directional separation (D-separation)



① chain

$A \perp\!\!\!\perp B \mid C$

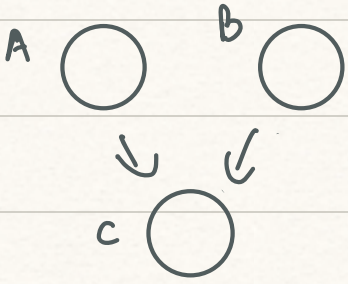
$A \not\perp\!\!\!\perp B$



② tree

$A \perp\!\!\!\perp C \mid B$

$A \not\perp\!\!\!\perp C$



③ V-structure

$A \not\perp\!\!\!\perp B \mid C$

$A \perp\!\!\!\perp B$

There are 3 basic "motifs" (i.e., mini-graphs) that encode fundamentally different conditional independence semantics.

1) Chain: e.g., "the past is independent of the future given the present".
2) Tree: e.g., iid sampling: student 1 independent of student 2 given school mean/variance
3) V-structure: "explaining away", e.g., COVID **not** independent of cold given coughing

We will talk more about PGM semantics in week 5.

Monte carlo approximations

Last time, we motivated the posterior expectation of the school means:

$$\hat{\theta} = \mathbb{E}_{p(\theta|y)}[\theta]$$

where the expectation is with respect to the **posterior marginal**:

$$p(\theta|y) = \int p(\theta, z_{\setminus\theta}|y)\, dz_{\setminus\theta}$$

↑ posterior marginal

marginalizes out all **other** latent

In complicated enough models, we cannot analytically compute these marginals.

However, we can still **approximate posterior expectations** using Monte Carlo.

$$\mathbb{E}_{p(\theta|y)}[f(\theta)] = \frac{1}{M}\sum_{m=1}^{M} f(\theta^m),$$

$$\theta^m \overset{iid}{\sim} p(\theta|y), \quad m = 1 \dots M$$

Clearly the Monte Carlo estimator is unbiased:

$$\mathbb{E}_{p(\theta|y)}\left[\frac{1}{M}\sum f(\theta^m)\right] = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{p(\theta|y)}[f(\theta)] = \mathbb{E}[f(\theta)]$$

It has variance:

$$\mathrm{Var}\left[\frac{1}{M} f(\theta^m_j)\right] = \frac{1}{M^2}\left(\sum_m \mathrm{Var}(f(\theta)) + 2\sum_{m<m'} \mathrm{Cov}\left[f(\theta^m), f(\theta^{m'})\right]\right)$$

If the samples are uncorrelated, the root mean squared error (RMSE) of the Monte Carlo estimator is then:

$$\mathrm{RMSE}\left(\frac{1}{M}\sum_m f(\theta^m)\right) = \sqrt{\frac{1}{M}\mathrm{Var}(f(\theta))} = O\left(\frac{1}{\sqrt{M}}\right)$$

Notice that this **does not depend on the dimension** of the parameter.

Takeaway: if we can draw (uncorrected) iid samples from the posterior, we can get unbiased estimates of posterior expectations whose error does not depend on the dimension of our parameter set. This is the basis for MCMC methods.

Markov chains and MCMC:

The Gibbs sampler is a **Markov chain** with state equal to the latent variables:

$$\text{State}: \quad S^m = (S_1^m \cdots S_D^m)$$

A Markov chain defines a joint distribution over a sequence of states, which factorizes:

$$\pi(S^0, S^1, \ldots, S^m) = \pi(S^0) \prod_{m=1}^{M} \pi(S^m \mid S^{m-1})$$

"initial distribution" ↑ ↑ "transition operator"

A Markov chain is "**homogenous**" if the transition operator is the same at all steps:

$$\pi(S^m = s \mid S^{m-1} = s') \equiv \pi(s \mid s') \quad \text{for "homogenous" chains}$$

The marginal distribution over the state at step m is:

$$\pi(S^m = s) = \int \pi(S^{m-1} = s')\, \pi(S^m = s \mid S^{m-1} = s')\, ds'$$
$$= \int \pi(S^{m-1} = s')\, \pi(s \mid s')\, ds'$$

A "**stationary distribution**" of the Markov chain is one with the following property:

$$\pi^*(s) = \int \pi^*(s')\, \pi(s \mid s')\, ds' \equiv \mathop{\mathbb{E}}_{s' \sim \pi^*}[\pi(s \mid s')]$$

A stationary distribution is "**invariant**" to the transition operator.
More informally, if you start the chain from a stationary distribution, you stay there.

A Gibbs sampler is one kind of **Markov Chain Monte Carlo (MCMC)** method.
MCMC rely on **Markov chains whose stationary distribution is the exact posterior.**

How do we know if a distribution is a stationary distribution of a Markov chain?
A sufficient condition is "**detailed balance**":

$$\pi^*(s')\, \pi(s \mid s') = \pi^*(s)\, \pi(s' \mid s)$$

To see why, we integrate both sides:

$$\int \pi^*(s')\, \pi(s \mid s')\, ds' = \int \pi^*(s)\, \pi(s' \mid s)\, ds'$$

$$= \pi^*(s) \text{ (by definition)} \qquad = \pi^*(s) \int \pi(s' \mid s)\, ds'$$

The target distribution is thus a stationary distribution of the Gibbs Markov chain.

In general, a Markov chain may have multiple stationary distributions. A Markov chain has a **unique** stationary distribution if it is "**ergodic**", and if it is ergodic, it will eventually converge to its unique stationary distribution:

$$\lim_{m \to \infty} \pi(S^m = s) = \pi^*(s), \quad \text{for any } S^0$$

A sufficient condition for **ergodicity** is that all states are reachable from all other states under the transition operator:

$$\pi(s \mid s') > 0$$

More generally, a chain is ergodic if it is 1) **irreducible** and 2) **aperiodic**. (We will not define these in lecture, unless there is time.)

Gibbs sampler:

A Gibbs sampler is a Markov chain whose transition operator updates one coordinate at a time from the conditional distribution of the target distribution.

$$\pi_{Gibbs}(S^m = s \mid S^{m-1} = s') = \prod_{d=1}^{D} p^*(S_d = s_d \mid S_{\setminus d} = s_{\setminus d}, S_{>d} = s'_{>d})$$

The marginal distribution over the dth coordinate (i.e., latent variable) is:

$$\pi_{Gibbs}(S_d^m = s_d) = \int p^*(s_d \mid s'_{\setminus d}) \, \pi_{Gibbs}(S^{m-1} = s') \, ds'$$

Say that at (m-1)th marginal is the target distribution:

$$\text{Say } \pi_{Gibbs}(S^{m-1} = s') = p^*(s'_1 \cdots s'_D).$$

Then the mth marginal over the dth coordinate is the marginal under the target:

$$\begin{aligned} &= \int ds' \; p^*(s_d \mid s'_{\setminus d}) \, p^*(s'_1 \cdots s'_D) \\ &= \int ds'_{\setminus d} \; p^*(s_d \mid s'_{\setminus d}) \int ds'_d \; p(s'_1 \cdots s'_D) \\ &= p^*(s_d) \end{aligned}$$

So: $p^*(s_1 \cdots s_D)$ is a stationary dist.

In Gibbs sampling, our target distribution is the posterior, and our conditionals are the complete conditionals. If our chain is ergodic then, the posterior is the unique stationary distribution, and the chain will eventually converge to it.

$$\lim_{m \to \infty} \pi_{Gibbs}(S^m = s) = p^*(s) \equiv p(z \mid y)$$

Non-ergodic chains usually arise when there are deterministic steps in the generative process. Consider the following model:

$$b \sim \text{Bernoulli}(p)$$

$$p \sim \begin{cases} \delta_0 \text{ if } b = 0 \\ \text{Beta}(\alpha, \beta) \text{ if } b = 1 \end{cases}$$

$$y \sim \begin{cases} \delta_0 \text{ if } p = 0 \\ \text{Bernoulli}(p) \text{ if } p > 0 \end{cases}$$

Consider the following complete conditionals for p and b given y=0:

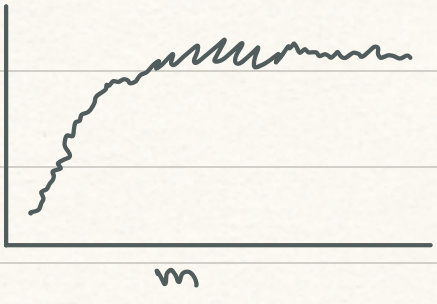
$$p(p = 0 \mid b = 0, y = 0) = 1$$

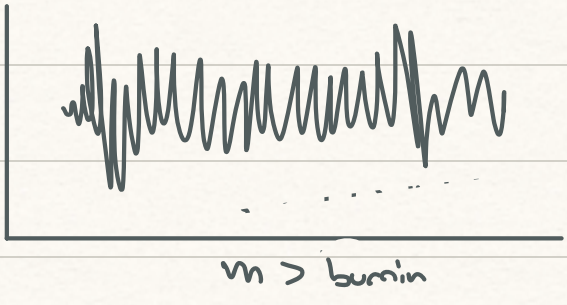
$$p(b = 0 \mid p = 0, y = 0) = 1$$

So if the chain starts at (b=0, p=0), it will never leave, and it won't converge to the posterior. As a rule of thumb, if all complete conditionals in your model always place non-zero probability on all possible outcomes, then the Gibbs sampler will be ergodic.

Practical considerations:

The chain will eventually be independent of the initial state, but that is only after enough transitions. In practice, we set a "**burn-in**" or "**warm-up**" period where we run the chain until it appears to converge; then we start collecting samples.





"trace plots"

As mentioned above, our Monte Carlo estimates will have higher error if samples are correlated. Samples in a Gibbs chain are indeed correlated, since they are directly conditioned on each other. To reduce our chain's "**autocorrelation**" we will "**thin**" the chain by retaining only every H samples (e.g., every 50th sample).

Missing data:

What if we have missing / heldout data?

$$y = y_{obs} \cup y_{miss}, \qquad S = \{z, y_{miss}\}$$

MCMC has a natural way to handle this: by simply treating missing observations like any of the other latent variables. The Gibbs sampler then becomes:

for $m = 1 \ldots M$:

  for $d = 1 \ldots D$:

$$z_d^m \sim p(z_d \mid z_{1d}^m, z_{2d}^{m-1}, y_{miss}^{m-1}, y_{obs})$$

$$y_{miss}^m \sim p(y_{miss} \mid z^m, y_{obs})$$

This last step samples the missing data from its complete conditional.

Question: in the "8 schools" model, what is this conditional?

Question: What expectations do the following Monte Carlo estimates approximate:

$$\lim_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} y_{miss}^m = ?$$

$$\lim_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} p(y_{miss} \mid z^m, y_{obs}) = ?$$

Question: What would the Gibbs stationary distribution be in this case?

$$\pi_{Gibbs}^{\infty}(S) = ?$$

Testing MCMC:

What if we treated all our data as missing? This is strange but worthwhile:

$$y = y_{miss}, \qquad S = \{y, z\}$$

What would the Gibbs stationary distribution be in this case?

$$\pi^*_{Gibbs}(S) = p(y_{miss}, z \mid y_{obs}) = p(y, z)$$

full joint!

This observation is the basis for a very useful trick called "Geweke testing" which lets you test whether your implementation of an MCMC algorithm is correct.

Define two different sampling algorithms called the forward and backward sampler.

Forward sampler

for $m = 1 \ldots M$:

$$z_f^m \sim p(z)$$

$$y_f^m \sim p(y \mid z)$$

$$\Downarrow$$

$$(z_f^m, y_f^m) \sim p(z, y)$$

Backward sampler

for $m = 1 \ldots M$:

$$z_b^m \sim \pi_{Gibbs}(z \mid z_b^{m-1}, y_b^{m-1})$$

$$y_b^m \sim p(y \mid z_b^m)$$

$$\Downarrow$$

$$(z_b^m, y_b^m) \sim p(z, y)$$

The key idea is that both should be drawing samples from the joint. The forward sampler simply samples M independent joint samples by sampling from the prior and then the likelihood. The backward sampler begins at some initialization, then it iteratively resamples the latents given data and data given latents. The backward sampler involves the Gibbs transition operator while the forward does not. Thus you can test whether your transition operator is correct by comparing the joint samples from each (they should be drawn from the same distribution!).