Motivating example: "8 schools"

$$y_{j,i} \overset{iid}{\sim} \mathcal{N}(\vartheta_j, \sigma_j^2) \qquad i = 1 \ldots n_s, \quad j = 1 \ldots S$$

There are S schools, each with nj students. We assume that the test scores of the students **within** a school are iid from a shared normal.

Prior over school mean:

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

We will assume a conjugate prior over the schools' mean test score.

By conjugacy, the "**complete conditional**" for the school mean is:

$$P(\vartheta_j \mid y_{j,1:n_j}, \sigma_j^2, \tau^2, \mu) = \mathcal{N}(\vartheta_j \,;\, \hat{\vartheta}_j, V_j)$$

posterior mean

posterior variance

$$\hat{\vartheta}_j = p_j \bar{y}_j + (1 - p_j)\mu$$

$$\left(\text{Note: } \bar{y}_j = \frac{1}{n_s}\sum_{i=1}^{n_s} y_{j,i} = \hat{\vartheta}_j^{MLE}\right)$$

$$p_s = \frac{n_j/\sigma_j^2}{1/\tau^2 + n_j/\sigma_j^2}$$

$$V_j = p_j \frac{\sigma_j^2}{n_j} = \frac{1}{1/\tau^2 + n_j/\sigma_j^2}$$

Notice how the posterior mean changes as a function of the number of observations, the likelihood variance, and the prior variance of the mean.

$$n_s \to \infty \quad \text{or} \quad \sigma_s^2 \to 0 \quad \text{or} \quad \tau^2 \to \infty \quad : \quad \hat{\vartheta}_s = \hat{\vartheta}_s^{MLE}$$

$$n_s \to 0 \quad \text{or} \quad \sigma_s^2 \to \infty \quad \text{or} \quad \tau^2 \to 0 \quad : \quad \hat{\vartheta}_s = \mu$$

Notice how the posterior parameters only involve mean (and number) of observations. This is an example of **sufficiency**: the mean and number are **sufficient statistics** for the school mean—i.e.,:

$$P(\vartheta_j \mid y_{j,1:n_j}, -) = P(\vartheta_j \mid \bar{y}_j, n_j, -)$$

This means the exact same posterior would emerge if we only ever generated and conditioned on the empirical mean. The mean of iid Gaussians is Gaussian:

$$\bar{y}_j \mid \vartheta_j \overset{ind.}{\sim} \mathcal{N}(\vartheta_j, \sigma_j^2/n_j), \quad j=1\ldots S$$

This is an alternative likelihood leading to the exact same posterior.

This is a "normal means" model, with each mean having one observation. Integrating out the school-specific means, results in the following model:

$$\bar{y}_j \overset{ind.}{\sim} \mathcal{N}(\mu, \tau^2 + \sigma_j^2/n_j), \quad j=1\ldots S$$

Empirical Bayes, type-II MLE, and shrinkage estimators

Using the above model, one could imagine fitting the prior mean via MLE:

$$\hat{\mu}^{MLE} = \sum_{j=1}^{S}\left[\frac{w_j}{\sum_{j'} w_{j'}}\right]\bar{y}_j, \quad w_j = \frac{1}{\tau^2 + \sigma_j^2/n_j}$$

In this case, this is called **type-II MLE**, or maximum **marginal** likelihood estimation since we are maximizing with respect to the marignal likelihood (with the means integrated out).

Consider again the posterior mean of the school-specific means, but using the fitted prior mean:

$$\hat{\vartheta}_j = \rho_j \bar{y}_j + (1-\rho_j)\hat{\mu}^{MLE}$$

This is a **shrinkage estimator** since we are shrinking the different estimates towards the fitted mu. Such estimators can have lower risk than estimating each mean according to the local MLEs, even if estimates are shrunk towards an arbitrary location (e.g., see the James-Stein paradox).

This is also an example of an **empirical Bayes** procedure where we set the priors in a data-driven way. There is a close connection between shrinkage estimators and empirical Bayes.

The prior variance plays an important role of controlling how observations are **pooled**:

$$\tau^2 = 0 \; : \quad \hat{\mu}^{MLE} = \sum_{j=1}^{S} \frac{n_j}{n} \bar{Y}_j = \frac{1}{n} \sum_{j=1}^{S} \sum_{i=1}^{n_s} Y_{j,i}$$

no pooling

As the prior variance goes to 0, the prior Gaussian goes to a delta spike:

$$\vartheta_j \sim \delta_\mu$$

In other words, it is consistent with the assumption that all schools have the same mean (equal to mu).

$$\tau^2 \to \infty : \quad \left[\frac{w_j}{\sum_j w_j} \right] \approx \frac{1}{S} \to \hat{\mu}^{MLE} = \frac{1}{S} \sum_{j=1}^{S} \bar{Y}_j$$

complete pooling

As the variance gets very large, the weights are (nearly) equal across all schools (this isn't rigorous, the limit is technically undefined), and we get a **completely pooled** estimate. This corresponds to the prior assumption that school means do not bear any information on each other.

Hierarchical modeling

The prior mean μ plays the role of **shrinkage location** for the means across schools. The prior variance τ^2 then determines the **degree of pooling.**

One data-adative way to set these would be to do type-II **MLE** (i.e., empirical Bayes). Another is to place priors and infer them as latent variables (i.e., hierarchical modeling)

$$\hat{\theta}_j = \mathbb{E}\left[\theta_j \mid y_j, \hat{\mu}_{MLE}, \tau^2_{MLE}\right]$$

posterior mean w/ empirical Bayes

vs.

$$\hat{\theta}_j = \mathbb{E}_{(\mu,\tau^2)\sim p(\mu,\tau^2 \mid y)}\left[\mathbb{E}\left[\theta_j \mid y_j, \mu, \tau^2\right]\right]$$

posterior mean w/ hierarchical model

posterior of hyperparameters

A conjugate prior would let us get the posterior of hyperparameters in closed form. A conjugate prior in this case would have the following form:

$$(\mu, \tau^2) \sim g(\eta_0)$$

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

$$p(\mu, \tau^2 \mid y_{1:S}) = g(\mu, \tau^2; \eta_S)$$

The conjugate prior is the **normal-inverse (scaled) chi-squared (NIX) distribution:**

$$g(\mu,\tau^2;\eta_0) = \text{NIX}(\mu,\tau^2; \underbrace{\nu_0, \tau_0^2, \mu_0, \kappa_0}_{=\eta_0}) = \underbrace{\mathcal{N}(\mu; \mu_0, \tau^2/\kappa_0)}_{p(\mu \mid \tau^2)} \underbrace{\chi^{-2}(\tau^2; \nu_0, \tau_0^2)}_{p(\tau^2)}$$

This is a **bivariate** distribution equal to an inverse scaled chi-squared times a **conditional** Gaussian.

Due to conjugacy, the **joint** posterior is another NIX distribution:

$$p(\mu,\tau^2 \mid y, \eta_0) = g(\mu,\tau^2; \eta_S) = \text{NIX}(\mu, \tau^2; \nu_S, \tau_S^2, \mu_S, \kappa_S)$$

for certain **posterior parameters** whose form is left unstated. (You will derive this in HW1.)

Conjugate prior(s) for variance of Gaussian likelihoods with known mean

The conjugate prior for the mean of a Gaussian likelihood with **known** variance is Gaussian.
How about for the reverse? The variance is must be positive, so the conjugate prior must be some continuous distribution with support >0. The answer is the **inverse scaled chi-squared distribution.**

$$y_i \sim \mathcal{N}(\mu, \tau^2)$$ likelihood (with known mean)

$$\tau^2 \sim \chi^{-2}(\underset{\uparrow}{\nu_0}, \tau_0^2)$$ conjugate prior for variance

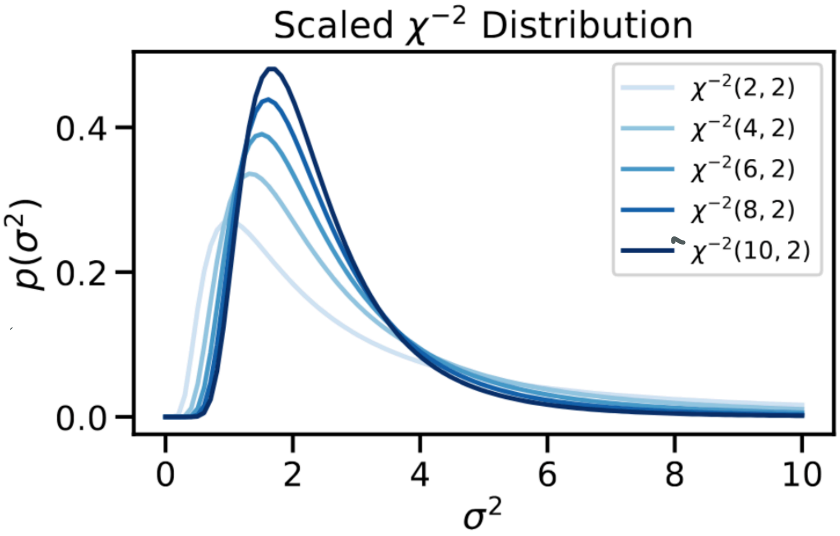
"scale" (sometimes called "mean" loosely)

"degrees of freedom"

The inverse chi-squared is asymmetric with mode < mean, and scale between them.

$$\underbrace{\frac{\nu_0}{\nu_0+2}\tau_0^2}_{\text{mode}} \leq \tau_0^2 \leq \underbrace{\frac{\nu_0}{\nu_0-2}\tau_0^2}_{\text{mean } \mathbb{E}[\tau^2]}$$

$$p(\tau^2) = \frac{\left(\frac{\nu_0\tau_0^2}{2}\right)}{\Gamma\left(\frac{\nu_0}{2}\right)} \left(\tau^2\right)^{-\frac{\nu_0}{2}-1} \exp\left(-\frac{1}{2\tau^2}\nu_0\tau_0^2\right)$$



It is a reparameterization of the **inverse gamma distribution:**

$$\chi^{-2}(\nu_0, \tau_0^2) \equiv \mathrm{InvGamma}\left(\overset{\text{"shape"}}{\frac{\nu_0}{2}}, \overset{\text{"rate"}}{\frac{\nu_0\tau_0^2}{2}}\right)$$

The inverse of an inverse chi-squared random variable is a chi-square random variable.
This is a conjugate prior for the **precision** of a Gaussian likelihood with known mean.

$$\lambda = 1/\tau^2 \qquad \lambda_0 = 1/\tau_0^2 \implies \lambda \sim \chi^2(\underset{\uparrow}{\nu_0}, \underset{\uparrow}{\lambda_0})$$

"degrees of freedom" "mean" (the actual mean)

The (scaled) chi-squared distribution is a reparameterization of the **gamma distribution.**

$$\chi^2(\lambda; \nu_0, \lambda_0) \equiv \mathrm{Ga}\left(\lambda; \underset{\uparrow}{\frac{\nu_0}{2}}, \underset{\uparrow}{\frac{\nu_0}{2\lambda_0}}\right)$$

"shape" "rate"

Inverse chi-squared—Gaussian conjugacy:

$$\tau^2 \sim \chi^2(\nu_0, \tau_0^2), \quad y_i \overset{iid}{\sim} \mathcal{N}(\mu, \tau^2), \quad i = 1 \ldots n$$

$$P(\tau^2 \mid -) \propto \chi^{-2}(\tau^2; \nu_0, \tau_0^2) \prod_{i=1} \mathcal{N}(y_i; \mu, \tau^2)$$

$$\propto [\tau^2]^{-\frac{\nu_0}{2} - 1} \exp\left(-\frac{1}{2\tau^2} \nu_0 \tau_0^2\right)$$

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}(y_i - \mu)^2\right)$$

$$\propto [\tau^2]^{-\frac{\nu_0}{2} - \frac{n}{2} - 1} \exp\left(-\frac{1}{2\tau}\underbrace{\left[\nu_0\tau_0^2 + \sum_{i=1}^{n}(y_i - \mu)^2\right]}_{= \nu_n \tau_n^2}\right)$$

$$\propto \chi^{-2}(\tau^2; \nu_n, \tau_n^2)$$

$$\nu_n = \nu_0 + n$$

$$\tau_n^2 = \left[\nu_0 \tau_0^2 + \sum_{i=1}^{n}(y_i - \mu)^2\right] \Big/ [\nu_0 + n]$$

WLOG say $\mu = 0$:

$$\tau_n^2 = \left(\frac{\nu_0}{\nu_0+n}\right)\tau_0^2 + \left(\frac{n}{\nu_0+n}\right)\underbrace{\frac{1}{n}\sum_i y_i^2}_{= \overline{y^2} = \hat{\tau}^2_{MLE}}$$

Notice that only the sum-of-squares enters the posterior parameters. This is another instance of sufficiency.

$$P(\tau^2 \mid y_{1:n}, -) = P(\tau^2 \mid \overline{y^2}, -)$$

This means if we only generated and conditioned on the sum-of-squares, the posterior would be the same. The distribution of this statistic is chi-squared:

$$\overline{y^2} \sim \chi^2(n, \tau^2) \equiv \text{Gamma}\left(\frac{n}{2}, \frac{n}{2\tau^2}\right)$$

Let's confirm we would get the same posterior:

$$p(\tau^2 \mid \overline{y^2}, -)$$

$$\propto \chi^{-2}(\tau^2; \nu_0, \tau_0^2)\, \chi^2(\overline{y^2}; n, \tau^2)$$

$$\propto \propto [\tau^2]^{-\frac{\nu_0}{2}-1} \exp\left(-\frac{1}{2\tau^2}\nu_0\tau_0^2\right)$$

$$\left(\frac{n}{2\tau^2}\right)^{\frac{n}{2}} \exp\left(-\overline{y^2}\frac{n}{2\tau^2}\right)$$

$$\propto [\tau^2]^{-\frac{\nu_0}{2}-\frac{n}{2}-1} \exp\left(-\frac{1}{2\tau^2}\left[\nu_0\tau_0 + \sum_{i=1}^{n} y_i^2) \right]\right)$$

This not only confirms that this is a sufficient statistic, but also shows that the inverse chi-squared is also a conjugate prior to the chi-squared distribution.

Intro to probabilistic graphical models (PGMs):

Consider the following hierarchical model for the 8 schools data:

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

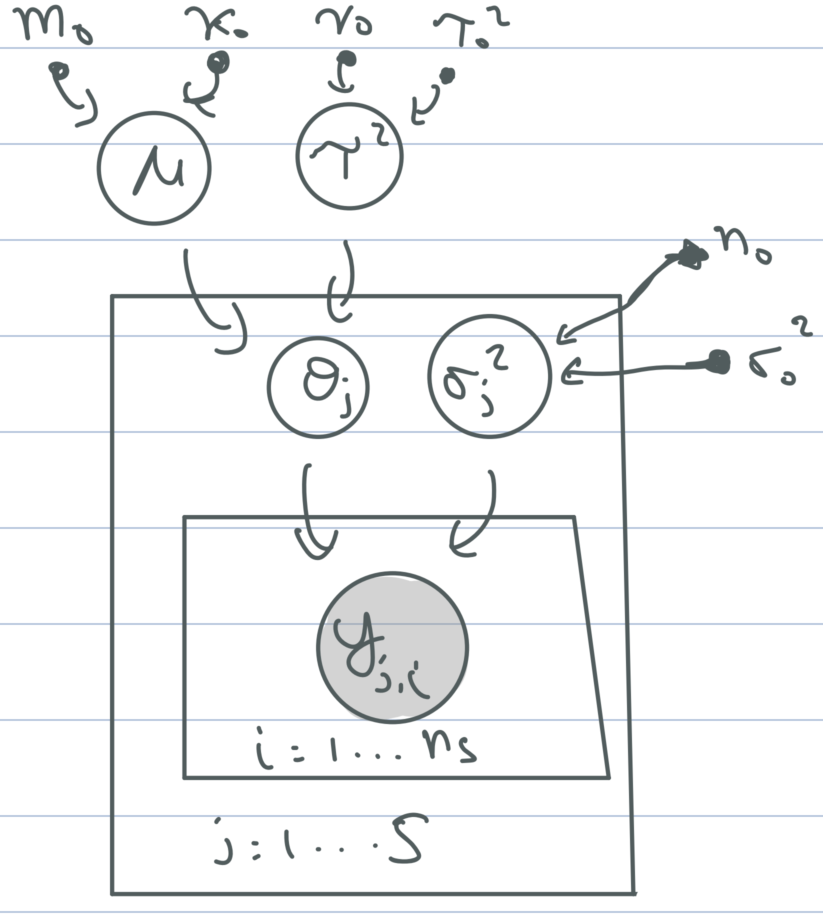
$$\mu \sim \mathcal{N}(m_0, 1/\kappa_0)$$

$$\sigma_j^2 \overset{iid}{\sim} \chi^{-2}(n_0, \sigma_0^2) \qquad j=1\dots S$$

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

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

This is actually still a simple model, but it's already hard to read. PGMs help us express and visualize hierarchical models.



This is a **probabilistic graphical model (PGM)** corresponding to the generative process above.

The **shaded nodes** are **observed** variables, while the **unshaded nodes** are **latent** variables.

The filled nodes are **hyperparameters** which are not random (i.e., fixed and usually known).

The **plates** denote **repeated sampling.**

$$Z = \{\mu, \tau^2, \theta_{1:S}, \sigma^2_{1:S}\}$$

the set of all latent variables in the model

The joint distribution over all latent and observed variables can always be written as:

$$p(Z, y) = \left[\prod_{d=1}^{D} p(Z_d \mid Z_{<d})\right] p(y \mid Z)$$

However, this does not take into account any conditional independences in the model.

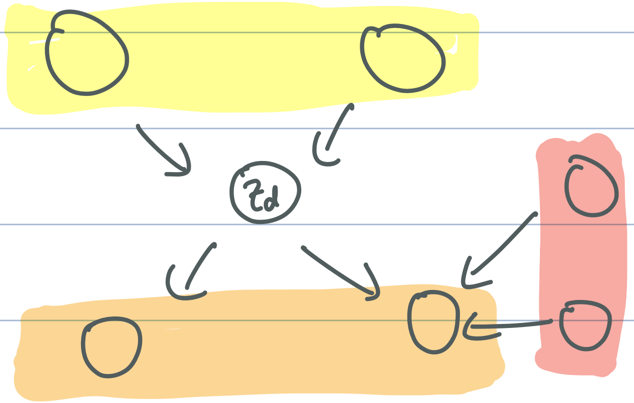
The joint distribution described by a PGM can always be factorized as:

$$= \left[\prod_{d=1}^{D} p(Z_d \mid Z_{par(Z_d)})\right] p(y \mid Z_{par(y)})$$

where **par(Zd)** is the set of all nodes in the graph that are parents of node Zd.

More generally, PGMs describe the set of all conditional independences in a joint distribution. A node is conditionally independent of all other nodes given its **Markov blanket.**

$$MB(Z_d) = \{ Par(Z_d) \cup Child(Z_d) \cup Coparents(Z_d) \}$$



$$Z_d \perp\!\!\!\perp Z_{d'} \mid MB(Z_d) \quad \forall \; Z_{d'} \notin MB(Z_d)$$

Preview to Gibbs sampling and MCMC:

Typically we interact with the posterior via **posterior expectations.**

For example, above we motivated the posterior expectation of the

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

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

$$\underset{\uparrow \text{ posterior marginal}}{P(\theta_j | Y)} = \underbrace{\int P(\theta_j, z_{\setminus \theta_j} | Y) \, dz_{\setminus \theta_j}}_{\text{marginalizes out all other latent}}$$

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

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

$$\mathbb{E}_{P(\theta_j | Y)} [f(\theta_j)] = \frac{1}{M} \sum_{m=1}^{M} f(\theta_j^m),$$

$$\theta_j^m \overset{iid}{\sim} P(\theta_j | Y), \quad m = 1 \ldots M$$

If we cannot analytically compute the marginal posterior, how do we sample from it?

Gibbs sampling:

```
for m = 1 ... M:
    for d = 1 ... D:
        $z_d \sim \underbrace{P(z_d | Y, z_{\setminus d})}_{\text{"complete conditional"}}$

        save $z_d^m \leftarrow z_d$
```

Claim: $\lim_{m \to \infty} P_r(z_d^m) = P(z_d | Y)$

The takeaway: even if the joint posterior over all latent variables is not tractable, or the desired posterior marginal is not tractable, we can approximate the marginal using Gibbs sampling as long as the **complete conditional are tractable.**

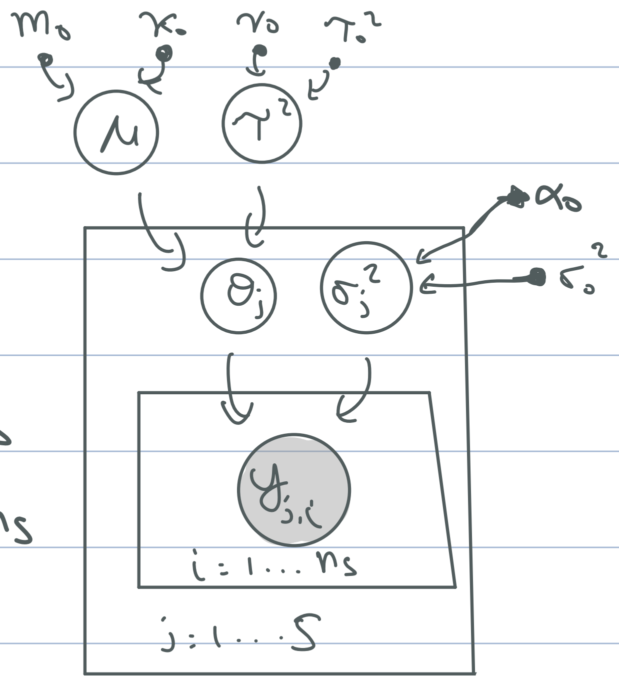
$$\tau^2 \sim \chi^{-2}(\nu_0, \tau_0^2)$$

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

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

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

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



For example, the complete conditional for the likelihood variance follows from conjugacy, since conditioning on all other variables turns the problem into an inverse chi-squared prior over the variance of a Gaussian likelihood with **known** mean.

e.g., $$p(\sigma_j^2 \mid -) = \chi^{-2}(\alpha_{n_j}, \sigma_{n_j}^2)$$

$$\alpha_{n_j} = \alpha_0 + n_j$$

$$\sigma_{n_j}^2 = \frac{\alpha_0 \sigma_0^2 + \sum_{i=1}^{n_j} (y_i - \theta_j)^2}{\alpha_0 + n_j}$$

As another example, the complete conditional for mu follows from Gaussian-Gaussian conjugacy, where here the school means are treated like data:

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

$$m_S = \rho_S \bar{\theta} + (1-\rho_S) m_0$$

$$\rho_S = \frac{1/\tau^2}{\kappa_0 + 1/\tau^2}$$

$$\kappa_S = \kappa_0 + 1/\tau^2$$

Notice that the complete conditional for mu involves tau^2, but not sigma^2. Why is that? This has to do with conditional independences in the model. PGMs make these easy to see.