Lecture 3: Directed Graphical Models; Hierarchical Models

STATS305C: Applied Statistics III

Scott Linderman

April 4, 2022

Last Time...

- Multivariate normal distribution
- The Wishart distribution
- Bayesian inference with Wishart, inverse Wishart, and normal inverse Wishart priors
- The multivariate Student's t distribution

Today...

Outline:

- Directed graphical models
- Hierarchical models

Reading:

- Required: Bishop, Ch 8.1-8.2
- Optional: Murphy, Ch 3.5.2, 4.2
- Optional: Gelman, Ch 5

Is the Multivariate Normal Too Simple or Too Complex?

The MVN was our first encounter with a joint distribution over multiple random variables. As a probabilistic model, you could argue that it is both too simple and too complex. Why?

Too simple:

light tails, uncorrelated ⇒ independent; only linear dependence

only mean+ cov, unimodal

Too complex:

D params from mean

+ $\frac{D(D+1)}{2}$ from cov.

cov. may be low-rank or sparse (inverse)

D^2 parametrized

to sample, $O(D^3)$ complexity (for the inverse of cov matrix); cholesky & eigen-decomposition

Solutions:

Compare to a multidimensional histogram

Now let each variable x_d be an integer in $\{1, \ldots, K\}$. (E.g. bin the real line into K bins.)

Question: How many parameters does an arbitrary distribution on $(x_1, \ldots, x_D)$ require? $K^D - 1$

Question: What if we use the **product rule** instead? How many parameters does each conditional have?

$$p(\boldsymbol{x}) = \underbrace{p(x_1)}_{K-1} \overbrace{p(x_2 \mid x_1)}^{K(K-1)} \underbrace{p(x_3 \mid x_1, x_2)}_{K^2(K-1)} \cdots \underbrace{p(x_D \mid x_1, \ldots, x_{D-1})}_{K^{D-1}(K-1)} \tag{1}$$

$$\# \text{param} = (K-1) \cdot \sum_{d=1}^{D} K^{d-1}$$

$$= (K-1) \cdot \frac{1-K^D}{1-K} = K^D - 1.$$

Question: How could we reduce complexity?

$P(X_D \mid \underbrace{\sim\sim}_{})$ ⟹ reduces complexity dramatically.

need to know only some of the variables

Directed Graphical Models

DGMs represent joint distributions as graphs.

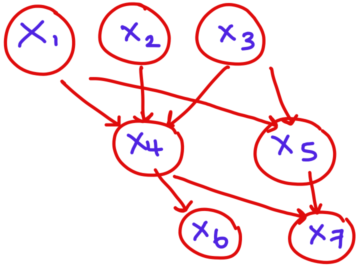
- Suppose that the conditional probability of x_d on depends on only a subset of preceding variables, $\text{pa}_d \subseteq \{1, \ldots d-1\}$.
- These are the **parents** of node d. Then,

$$p(\boldsymbol{x}) = \prod_{d=1}^{D} p(x_d \mid \boldsymbol{x}_{\text{pa}_d}) \qquad (2)$$

- We can represent the joint distribution as a **directed acyclic graph**:
  - Each **node** corresponds to a variable. It may be discrete or continuous, scalar or multidimensional.
  - Draw an **edge** from node i to j if $i \in \text{pa}_j$.

Exercise: Draw the directed graphical model for the following joint distribution,

$$p(\boldsymbol{x}) = p(x_1)p(x_2)p(x_3)p(x_4 \mid x_1, x_2, x_3) \\ \times p(x_5 \mid x_1, x_3)p(x_6 \mid x_4)p(x_7 \mid x_4, x_5)$$



doesn't tell you if x_7 is conditionally gaussian, etc.

⇒ incomplete

Directed Graphical Models II

Question: How many parameters would it take to represent the joint distribution $p(x_1, \ldots, x_D)$ if each $x_d \in \{1, \ldots, K\}$ and each node (except x_1) had exactly one parent? What type of graph is that?

x_1: $K-1$

$(D-1)$ variables: each w/ : $K(K-1)$

complexity: $O(K^2 D)$

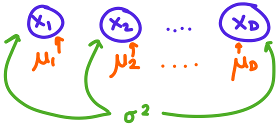
type of graph: tree

Directed Graphical Models III

Exercise: Let $\boldsymbol{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D$. Draw the graphical model for $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ with diagonal covariance.

→ no connected

→ DAG is a bunch of nodes



x_d

Plate Notation

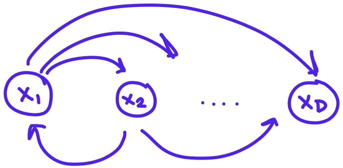
This example has **repeated structure**,

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) = \prod_{d=1}^{D} \mathcal{N}(x_d \mid \mu_d, \sigma^2). \tag{3}$$

We often use **plate notation** to such graphical models more compactly.

Directed Graphical Models IV

Exercise: Let $\boldsymbol{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D$. Draw the graphical model for $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ with arbitrary covariance.



"complete graph"

fully-connected graph

$$p(x_d \mid x_{1:d-1}) = \mathcal{N}(x_d \mid w_d^T x_{1:d-1}, \sigma_d^2)$$

Directed Graphical Models V

Note: Any joint distribution can be factored as,

$$p(\boldsymbol{x}) = p(x_1)\,p(x_2 \mid x_1)\,p(x_3 \mid x_1, x_2) \cdots p(x_D \mid x_1, \ldots, x_{D-1}), \tag{4}$$

in which case $\mathrm{pa}_d = \{1, \ldots, d-1\}$.

This is called a **fully connected graph**.

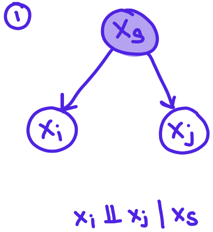
The **absence of edges** conveys independence assumptions.

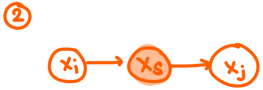
Conditional Independence

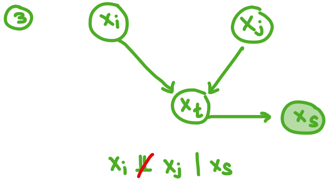
We say "*x_i is conditionally independent of x_j given $\boldsymbol{x}_s$*" if $p(x_i \mid x_j, \boldsymbol{x}_s) = p(x_i \mid \boldsymbol{x}_s)$, or equivalently, $p(x_i, x_j \mid \boldsymbol{x}_s) = p(x_i \mid \boldsymbol{x}_s)p(x_j \mid \boldsymbol{x}_s)$. We use the following shorthand,

$$x_i \perp\!\!\!\perp x_j \mid \boldsymbol{x}_s \iff p(x_i \mid x_j, \boldsymbol{x}_s) = p(x_i \mid \boldsymbol{x}_s). \tag{5}$$

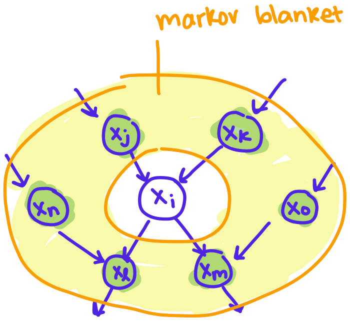
To read conditional independence relationships from a directed graphical model, we need to consider three types of motifs:







The **Markov blanket** of variable x_d consists of x_d's parents, x_d's children, and the other parents of x_d's children. (These are all the variables that appear alongside x_d in a factor of the joint distribution.) Given its Markov blanket, x_d is conditionally independent of all other variables.



markov blanket
x_j
x_k
x_n
x_i
x_o
x_l
x_m

Exchangeability

Conditional independence assumptions are natural when information is limited.

Consider modeling a collection of variables $(x_1, \ldots, x_D)$. If no information is available to order or group the variables, we must assume they are **exchangeable**:

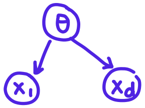
$$p(x_1, \ldots, x_D) = p(x_{\pi(1)}, \ldots, x_{\pi(D)}) \tag{6}$$

for any permutation π. The simplest exchangeable distributions assume independent and identically distributed r.v.'s,

$$p(x_1, \ldots, x_D) = \prod_{d=1}^{D} p(x_d). \tag{7}$$

More generally, we may assume the variables are conditionally independent given a parameter θ, which has been marginalized over,

$$p(x_1, \ldots, x_D) = \int \left[\prod_{d=1}^{D} p(x_d \mid \theta) \right] \overbrace{p(\theta)}^{\text{prior distribution of } \theta.} \, \mathrm{d}\theta. \tag{8}$$



Marginally, $x_1, \ldots, x_D$ are **not** independent, but they are exchangeable.

de Finetti's Theorem

de Finetti's theorem states that as $D \longrightarrow \infty$, any suitably well-behaved exchangeable distribution on $(x_1, \ldots, x_D)$ can be expressed as a mixture of independent and identical distributions, as in (8).

Though the theorem does not hold in the finite case, it is often cited as a motivation for conditional independence assumptions in Bayesian models.

Extensions of de Finetti's theorem have been proven for finite and Markov exchangeable sequences [Diaconis and Freedman, 1980a,b] and for partially exchangeable arrays, like infinite matrices, or graphs [Aldous, 1981, Hoover, 1979].

Hierarchical Models

Example: Modeling SAT scores from many schools.

Suppose we have test scores from S schools. Let N_s denote the number of students from school s and $x_{s,n} \in \mathbb{R}$ denote the score of the n-th student from the s-th school. We aim to build a probabilistic model of the scores $\boldsymbol{X} = \{\{x_{s,n}\}_{n=1}^{N_s}\}_{s=1}^{S}$ that will allow us to study relative performance across schools.

I know which school a student came from ⟹ not exchangeable!

The individual scores are not exchangeable since they are organized into groups by school. However, the schools themselves are exchangeable. This motivates the following **hierarchical model**:

$$\mu, \tau^2 \sim p(\mu, \tau^2) \quad \text{global mean/variance} \tag{9}$$

conditionally independent of other schools ←

$$\theta_s \sim \mathcal{N}(\mu, \tau^2) \qquad \text{for } s = 1, \ldots, S \quad \text{per-school effect} \tag{10}$$

$$x_{s,n} \sim \mathcal{N}(\theta_s, \underset{\text{known}}{\sigma_s^2}) \qquad \text{for } n = 1, \ldots, N_s \text{ and } s = 1, \ldots, S \tag{11}$$

Each school has its own mean θ_s, and the means are conditionally independent given the global mean and variance, μ and τ^2, respectively. Hence, the means are exchangeable.

Hierarchical Models II

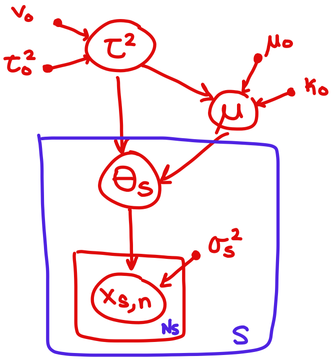
For the prior on (μ, τ^2), ~~we will assume an improper uniform distribution on the mean and a weakly informative inverse-chi-squared prior on the variance~~.

We can express this as a normal inverse-chi-squared,

$$p(\mu, \tau^2) = \mathrm{NIX}(\mu, \tau^2 \mid \mu_0, \kappa_0, \nu_0, \tau_0^2) \tag{12}$$

$$= \mathcal{N}(\mu \mid \mu_0, \tau^2/\kappa_0)\, \chi^{-2}(\tau^2 \mid \nu_0, \tau_0^2) \tag{13}$$

The hyperparameters of the full model are $\boldsymbol{\eta} = (\mu_0, \kappa_0, \nu_0, \tau_0^2, \{\sigma_s^2\}_{s=1}^S)$. (Soon we will model σ_s^2 too.)



Hierarchical Models III

Question: Consider the limit where $\kappa_0 \to 0$, $\tau_0^2 \to 0$, and $\nu_0 \to \infty$. What does that imply about $\underbrace{p(\mu, \tau^2)}$ and $\underbrace{p(\theta_s \mid \mu, \tau^2)}$?

$\chi^2(\tau^2 \mid \nu_0 \to \infty, \tau^2 \to 0) \longrightarrow \delta_0(\tau^2)$

$\mathcal{N}(\theta_s \mid \mu, \tau^2 \to 0) \to \delta_\mu(\theta_s)$

complete pooling

Question: Consider the limit where $\kappa_0 \to 0$, $\tau_0^2 \to \infty$, and $\nu_0 \to \infty$. What does that imply about $\underbrace{p(\mu, \tau^2)}$ and $\underbrace{p(\theta_s \mid \mu, \tau^2)}$?

$N(\theta_s \mid \mu, \tau^2 \to \infty) \to \text{uniform}(\mathbb{R})$

$\Longrightarrow$

no pooling

Bayesian Inference in the Hierarchical Gaussian Model I

Our goal is to compute the posterior,

$$p(\mu, \tau^2, \boldsymbol{\theta} \mid \boldsymbol{X}, \eta), \tag{14}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_S)$.

We'll take it in steps.

1. First, we can simplify the likelihood by observing that as a function of the parameter θ_s,

$$\prod_{n=1}^{N_s} \mathcal{N}(x_{s,n} \mid \theta_s, \sigma_s^2) \propto \mathcal{N}(\bar{x}_s \mid \theta_s, \bar{\sigma}_s^2) \tag{15}$$

   where $\bar{x}_s = \frac{1}{N_s} \sum_{n=1}^{N_s} x_{s,n}$ and $\bar{\sigma}_s^2 = \frac{\sigma_s^2}{N_s}$.

   Put differently, the school mean is a **sufficient statistic** of the likelihood (when variance is known).

Bayesian Inference in the Hierarchical Gaussian Model II

2. Use the product rule to write the posterior as

$$p(\mu, \tau^2, \boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{\eta}) = p(\boldsymbol{\theta} \mid \mu, \tau^2, \boldsymbol{X}, \boldsymbol{\eta})\, p(\mu \mid \tau^2, \boldsymbol{X}, \boldsymbol{\eta})\, p(\tau^2 \mid \boldsymbol{X}, \boldsymbol{\eta}) \tag{16}$$

3. The first term is the easy one:

$$p(\boldsymbol{\theta} \mid \mu, \tau^2, \boldsymbol{X}, \boldsymbol{\eta}) \propto \prod_{s=1}^{S} \left[\overbrace{\mathcal{N}(\theta_s \mid \mu, \tau^2)}^{\text{likelihood}} \underbrace{\mathcal{N}(\bar{x}_s \mid \theta_s, \bar{\sigma}_s^2)}_{\text{prior}} \right] \tag{17}$$

$$\propto \prod_{s=1}^{S} \mathcal{N}(\theta_s \mid \hat{\theta}_s, v_s) \tag{18}$$

where

$$v_s = \left(\frac{1}{\bar{\sigma}_s^2} + \frac{1}{\tau^2} \right)^{-1} \qquad \hat{\theta}_s = v_s \left(\frac{\bar{x}_s}{\bar{\sigma}_s^2} + \frac{\mu}{\tau^2} \right) \tag{19}$$

I.e. the conditional means are precision-weighted averages of the prior and sample means.

Bayesian Inference in the Hierarchical Gaussian Model III

4. To compute the second term in (16), we need to marginalize over the parameters $\boldsymbol{\theta}$. This is usually intractable, but since this model is conditionally linear and Gaussian, we can do it analytically.

$$p(\mu \mid \tau^2, \boldsymbol{X}, \eta) \propto \int p(\mu, \tau^2, \boldsymbol{\theta}, \boldsymbol{X} \mid \eta) \, \mathrm{d}\boldsymbol{\theta} \tag{20}$$

$$\propto \mathcal{N}(\mu \mid \mu_0, \tau^2/\kappa_0) \prod_{s=1}^{S} \int \mathcal{N}(\theta_s \mid \mu, \tau^2) \, \mathcal{N}(\bar{x}_s \mid \theta_s, \bar{\sigma}_s^2) \, \mathrm{d}\theta_s \tag{21}$$

$$= \mathcal{N}(\mu \mid \mu_0, \tau^2/\kappa_0) \prod_{s=1}^{S} \mathcal{N}(\bar{x}_s \mid \mu, \bar{\sigma}_s^2 + \tau^2) \tag{22}$$

$$\propto \mathcal{N}(\mu \mid \hat{\mu}, v_\mu) \tag{23}$$

where

$$v_\mu = \frac{1}{\lambda_0 + \sum_{s=1}^{S} \lambda_s} \qquad \hat{\mu} = \frac{\lambda_0 \mu_0 + \sum_{s=1}^{S} \lambda_s \bar{x}_s}{\lambda_0 + \sum_{s=1}^{S} \lambda_s} \qquad \lambda_0 = \frac{\kappa_0}{\tau^2} \qquad \lambda_s = \frac{1}{\bar{\sigma}_s^2 + \tau^2} \tag{24}$$

The posterior mean of μ is a precision-weighted average of the school means.

Bayesian Inference in the Hierarchical Gaussian Model IV

Bayesian Inference in the Hierarchical Gaussian Model V

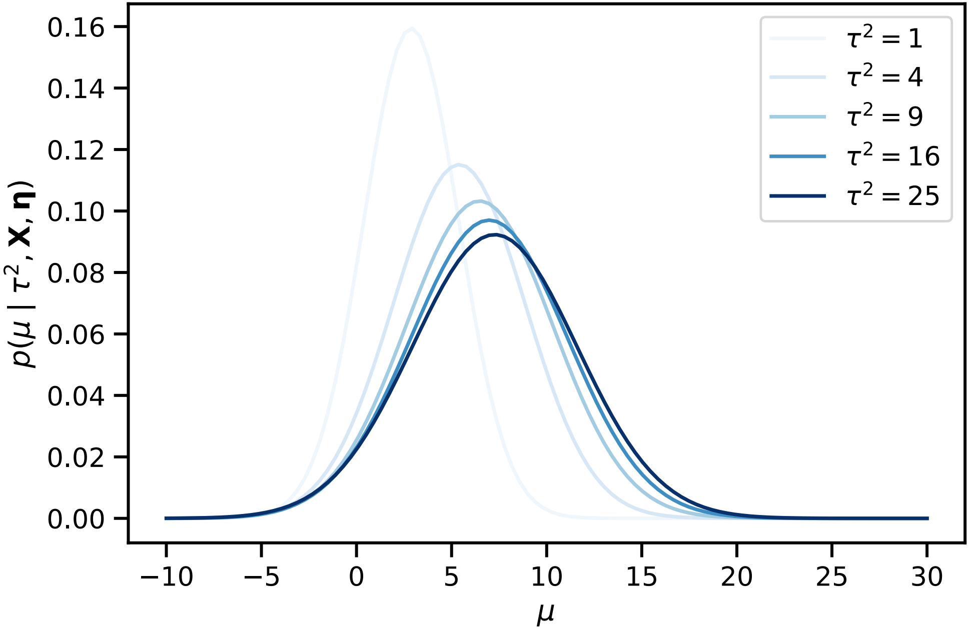


Figure: Posterior distribution on μ given the data and a range of τ^2 values.

Bayesian Inference in the Hierarchical Gaussian Model VI

5. Finally, for the last term in (16), we can integrate over μ to obtain,

$$p(\tau^2 \mid \boldsymbol{X}, \boldsymbol{\eta}) \propto \int p(\mu, \tau^2, \boldsymbol{X} \mid \boldsymbol{\eta})\, \mathrm{d}\mu \tag{25}$$

$$= p(\tau^2) \int \mathcal{N}(\mu \mid \mu_0, \tau^2/\kappa_0) \left[\prod_{s=1}^{S} \mathcal{N}(\bar{x}_s \mid \mu, \bar{\sigma}_s^2 + \tau^2)\right] \mathrm{d}\mu \tag{26}$$

The integral is very doable (a good exercise!) but it's a bit of a pain.

Alternatively, note that the following holds for any μ:

$$p(\tau^2 \mid \boldsymbol{X}, \boldsymbol{\eta}) = \frac{p(\mu, \tau^2 \mid \boldsymbol{X}, \boldsymbol{\eta})}{p(\mu \mid \tau^2, \boldsymbol{X}, \boldsymbol{\eta})} \propto \frac{p(\tau^2)\mathcal{N}(\mu \mid \mu_0, \tau^2/\kappa_0) \prod_{s=1}^{S} \mathcal{N}(\bar{x}_s \mid \mu, \bar{\sigma}_s^2 + \tau^2)}{\mathcal{N}(\mu \mid \hat{\mu}, v_\mu)}. \tag{27}$$

This is just Bayes' rule.

Bayesian Inference in the Hierarchical Gaussian Model VII

6. Plug in $\mu = \hat{\mu}$ since that will cause many terms in the denominator to disappear. Then,

$$p(\tau^2 \mid \boldsymbol{X}, \boldsymbol{\eta}) \propto p(\tau^2) \frac{\sqrt{v_\mu}}{\tau} e^{-\frac{\kappa_0}{2\tau^2}(\hat{\mu}-\mu_0)^2} \prod_{s=1}^{S} \frac{1}{\sqrt{\bar{\sigma}_s^2 + \tau^2}} e^{-\frac{1}{2}\left(\frac{\bar{x}_s - \hat{\mu}}{\sqrt{\bar{\sigma}_s^2+\tau^2}}\right)^2} \tag{28}$$

$$\triangleq f(\tau^2) \tag{29}$$

This function is complicated because both v_μ and $\hat{\mu}$ depend on τ^2.

Nevertheless, τ^2 is only a one-dimensional variable, so we can use numerical quadrature to compute the normalizing constant $\int f(\tau^2)\,\mathrm{d}\tau^2$ and draw samples from this posterior.

Bayesian Inference in the Hierarchical Gaussian Model VIII

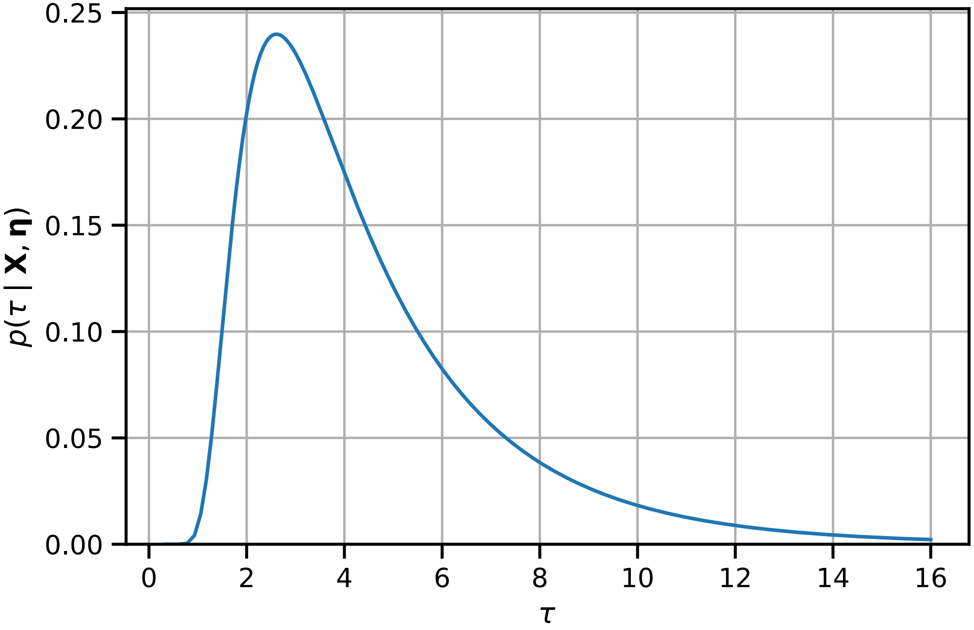


Figure: Posterior distribution on τ given the data, marginalizing out μ and $\boldsymbol{\theta}$.

One More Posterior Distribution

Last but not least, note that,

$$\begin{aligned} p(\theta_s \mid \tau, \boldsymbol{X}, \boldsymbol{\eta}) &= \int p(\theta_s \mid \mu, \tau, \boldsymbol{X}, \boldsymbol{\eta})\, p(\mu \mid \tau, \boldsymbol{X}, \boldsymbol{\eta})\, \mathrm{d}\mu \qquad (30) \\ &\propto \int \mathcal{N}(\bar{x}_s \mid \theta_s, \bar{\sigma}_s^2)\, \mathcal{N}(\theta_s \mid \mu, \tau^2)\, \mathcal{N}(\mu \mid \hat{\mu}, v_\mu)\, \mathrm{d}\mu \qquad (31) \\ &\propto \mathcal{N}(\bar{x}_s \mid \theta_s, \bar{\sigma}_s^2)\, \mathcal{N}(\theta_s \mid \hat{\mu}, \tau^2 + v_\mu) \qquad (32) \\ &= \mathcal{N}(\theta_s \mid \hat{\theta}_s, v_{\theta_s}) \qquad (33) \end{aligned}$$

where

$$v_{\theta_s} = \left(\frac{1}{\bar{\sigma}_2^2} + \frac{1}{\tau^2 + v_\mu^2} \right)^{-1} \qquad \hat{\theta}_s = v_{\theta_s} \left(\frac{\bar{x}_s}{\bar{\sigma}_s^2} + \frac{\hat{\mu}}{\tau^2 + v_\mu} \right) \tag{34}$$

Again, note that τ^2 affects all of these quantities!

One More Posterior Distribution II

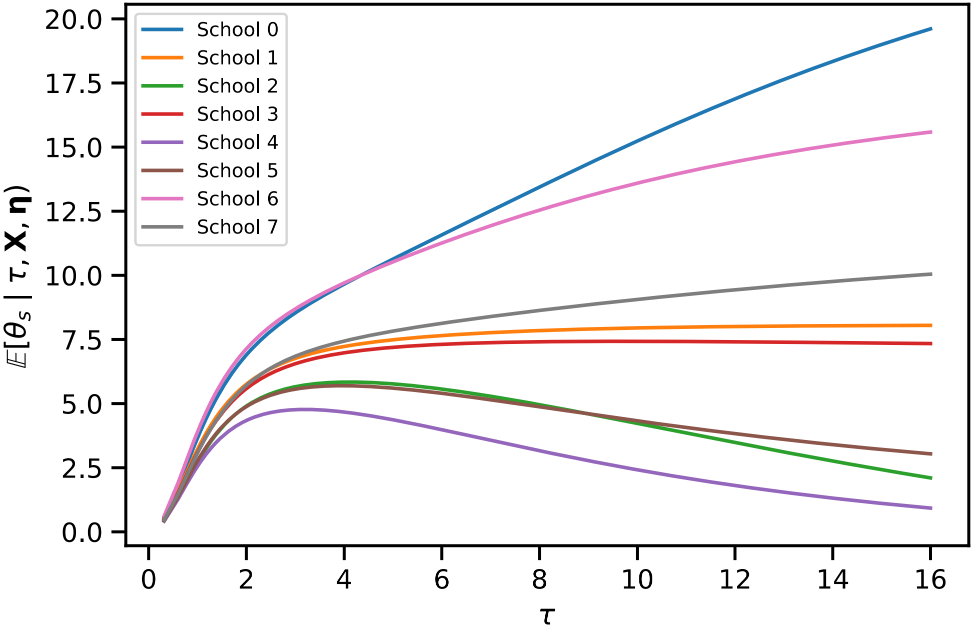


Figure: Posterior mean of $\boldsymbol{\theta}$ given the data and τ, marginalizing out μ.

Posterior Sample of Per-School Effects

We can draw samples of θ_s from their posterior marginal distribution by sampling τ^2, μ, and θ_s, then discarding the former two. This is called **ancestral sampling**.

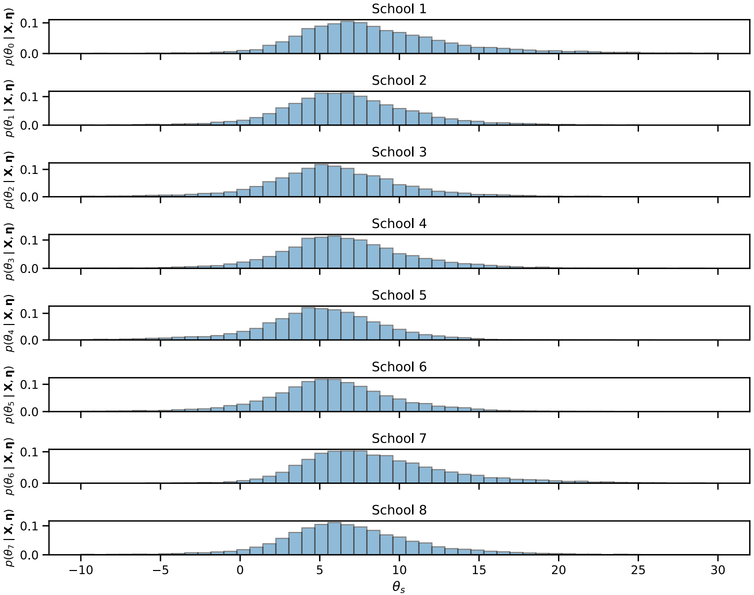


Figure: Posterior samples of $\boldsymbol{\theta}$.

Hierarchical Gaussian Model Recap

- We've derived expressions for each term in the hierarchical Gaussian posterior (16).
- With these, we can visualize the posterior marginal distribution over (μ, τ^2) since its only 2D.
- We can also simulate posterior samples of θ_s for each school.

Comparison to Classical Analysis of Variance

A classical approach to estimating θ_s is to choose between two estimators: the *unpooled* estimate, $\hat{\theta}_s = \frac{1}{N_s}\sum_{n=1}^{N_s} x_{s,n}$, or the *pooled* estimate, $\hat{\theta}_s = \frac{1}{N}\sum_{s=1}^{S}\sum_{n=1}^{N_s} x_{s,n}$. The former treats all schools as independent; the latter treats them as identical.

To choose between these two estimators, one might perform an analysis of variance.

	df	SS	MS	$\mathbb{E}[\mathrm{MS} \mid \sigma^2, \tau^2]$
Between groups	$S-1$	$\sum_s \sum_n (\bar{x}_s - \bar{x})^2$	$SS/(S-1)$	$N\tau^2 + \sigma^2$
Within groups	$S(N-1)$	$\sum_s \sum_n (x_{s,n} - \bar{x}_s)^2$	$SS/(S(N-1))$	σ^2
Total	$SN-1$	$\sum_s \sum_n (x_{s,n} - \bar{x})^2$	$SS/(SN-1)$	

If the ratio of the between to the within mean squares (MS) is significantly greater than 1 (according to an F test), then the ANOVA suggests using unpooled estimates $\hat{\theta}_s = \bar{x}_s$ for each school.

If the ratio is not significantly greater than one, then we cannot reject the null hypothesis that $\tau^2 = 0$ (i.e. all schools are identical), so we should use the pooled estimate.

The hierarchical Bayesian approach yields a posterior distribution over θ_s whose mean naturally interpolates between these two extremes.

Comparison to Classical Analysis of Variance

Another approximation uses unbiased **point estimates** of the parameters,

$$\hat{\mu} = \bar{x}, \quad \hat{\tau}^2 = (\mathrm{MS}_B - \mathrm{MS}_W)/N, \tag{35}$$

to draw inferences about θ_s from the conditional distribution,

$$p(\theta_s \mid \hat{\mu}, \hat{\tau}^2, \boldsymbol{x}_s). \tag{36}$$

However, this approach **fails to propagate uncertainty** about the global parameters and so underestimates the posterior variance of θ_s.

Moreover, the point estimate $\hat{\tau}^2$ can be negative! In this case, it's typical to set $\hat{\tau}^2 = 0$, but this is too strong a claim as well.

Next Time...

We assumed that σ_s^2 was known *a priori*, but this assumption is unwarranted in practice. An alternative is to give each school's variance a prior distribution like,

$$\sigma_s^2 \sim \chi^{-2}(\nu_0, \sigma_0^2). \tag{37}$$

Unfortunately, this further complicates the analysis to the point where it is no longer doable in closed form.

Next time we'll introduce methods to handle this added complexity.

References I

Persi Diaconis and David Freedman. Finite exchangeable sequences. *The Annals of Probability*, pages 745–764, 1980a.

Persi Diaconis and David Freedman. de Finetti's theorem for Markov chains. *The Annals of Probability*, pages 115–130, 1980b.

David J Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

Douglas Hoover. Relations on probability spaces and arrays of random variables. Technical report, Institute for Advanced Study, Princeton, NJ, 1979.