STA 602 - Intro to Bayesian Statistics

Lecture 16

Li Ma

Duke University

Exchangeability

- The rationale behind modeling a group of observations as i.i.d. given some unknown parameter is *exchangeability*, which represents our lack of knowledge to differentiate the observations *a priori*.

$$y_1, y_2, \ldots, y_n \,|\, \theta \overset{\text{iid}}{\sim} p(\cdot|\theta).$$

- In practice, observations often form groups that are exchangeable within but not across. Examples:
  - Test scores of students from 7 high schools in a school district.
  - Blood pressure measurements from 10 patients. $y_i|\theta_i \sim$ normal, $\theta_i \sim$ Normal
- In these settings, we do have some ability to differentiate the observations, and clearly those observations shouldn't be modeled as i.i.d. given some common sampling model.

Conditional exchangeability

- Note that other than the grouping information, sometimes we also have additional information about the observations, such as the age of the 10 patients, and the average income level of the school district, which will also destroy the exchangeability.
- We can assume *conditional exchangeability* rather than exchangeability.
- That is, we assume that observations are exchangeable given that their additional information (e.g., grouping information, covariates) is the same.
- More on this later when we talk about regression models.

Hierarchical modeling

* narrower sense: modeling multiple layers of exchangability.

* broader sense: specifying a multivariate joint dist. through multiplying a sequence of conditional prob.

- It is often reasonable to model such data using several layers of exchangeability.
  - Students within each high school exchangeable, and the high schools with the school district exchangeable, and the school districts in a state exchangeable, etc.
- By this logic, we can model
  - the average score for the school distrcts as i.i.d. from some distribution of school districts for the state
  - the average score for the schools in a school district as i.i.d. from some distribution of shools for the school distrct (which should depend on the average score of the district)
  - the scores of individual students in a school as i.i.d. from some distribution of students for the school (which should depend on the average score of the school).
- So we can attempt to model the sampling distribution of the students using a few layers of conditional distributions.

Mathematical formulation

If we only consider one district (kill all k)
$\phi | \psi \sim p(\cdot | \psi)$ is prior.

- The *sampling model* for the test scores can be represented as

$$\begin{cases} y_{ijk} \,|\, \theta_{jk}, \psi \overset{\text{iid}}{\sim} p(\cdot | \theta_{jk}, \psi) & \text{Students} \\ \theta_{jk} \,|\, \phi_k, \psi \overset{\text{iid}}{\sim} p(\cdot | \phi_k, \psi) & \text{Schools} \\ \phi_k \,|\, \psi \overset{\text{iid}}{\sim} p(\cdot | \psi) & \text{Districts} \end{cases}$$

  where θ_{jk} is the parameter for School j in District k; ϕ_k is the parameter for District k.
- Note that up to this point, we have *not* been a Bayesian yet.
- The above simply specifies the sampling model.
- The modeling assumption on $\theta_{jk} \,|\, \phi_k, \psi$ for example, does not quantify our *subjective prior belief*, but rather models the sampling mechanism from which the school averages arise in District i.
- Finally, we can place a *prior—quantifying our prior belief—*on the (global) parameter ψ,

hyper prior

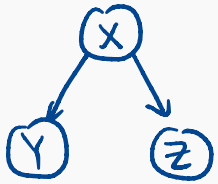
$$\psi \sim p(\psi).$$

Why bother? Look around and borrow information

- Why don't we take the easier route to simply analyze the scores from each school *separately*?
- Then we can perhaps more comfortably use a simple i.i.d. sampling model.
- Borrowing of information.
  - In modeling a group of students as i.i.d. from some sampling model, we are attempting to use the more than one observation to help infer the infer the score/ability of a single student.
  - Modeling the schools as i.i.d. from a distribution for the district, the scores from the other schools in the same district will help us infer the average score in each school, which in turn will improve our inference on each student in each school.
  - Similarly, sharing information among districts can help infer the average score in each district, which in turn will help infer average scores of the schools in each district, and in turn help infer the scores/ability of students in all schools within each district.
- This will be especially useful if some or all groups of observations are small.
  - For example: Suppose we only have 1 score from a school.

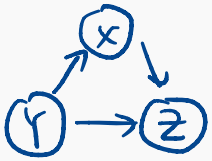
Relationship to graphical models

- More generally speaking, a *hierarchical model* is a probabilistic model for the data that are specified in terms of *multiple layers of conditional models*.
- These models can be represented using a so-called *"directed acyclic graph"* (DAG) or *"Bayesian network"* (BN).
- *Graphical models* is a visual way of representing joint probability distribution of multiple random variables.
- "Directed" implies that we can express the relationship using directed edges between variables, which indicates an ordering.
- "Acyclic" means that the graphical models are specified without any "loops' '. That is, there is a clear notion of hierarchy in the model specification:
  - We specify y_{ijk} given θ_{jk}, and θ_{jk} given ϕ_k and ϕ_k given ψ. We don't need to and shouldn't simultaneously specify ϕ_k or ψ given θ_{jk}, which will create a loop.



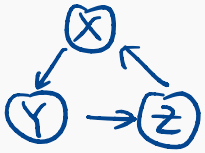
$$p(x,y,z) = p(Y|x) \cdot p(z|x) \cdot p(x)$$

Allowed:



$$p(x,y,z) = p(y) \cdot p(z|y) \cdot p(z|x,y)$$

Not allowed:



cycle

Graphical model helps us specify the model (joint distribution)

Full conditional helps to capture the upstream information flow and make inference.

→ X → Y ← Z Collider

* Without observing Y, X and Z are conditionally independent.
* Once observing Y, X and Z are negatively correlated.

→ X → Y → Z Path

* Without observing Y, X and Z are positively correlated.
* Once observing Y, X and Z are conditionally independent.

→ X ← Y → Z Common Cause

* Without observing Y, X and Z are positively correlated.
* Once observing Y, X and Z are conditionally independent.

Relationship to graphical models

- In other words, under a DAG or BN, we specify our model by factorizing the joint probability of all random variables in a *given order*. For example

$$\begin{aligned} & p(\{y_{ijk}\},\{\theta_{jk}\},\{\phi_k\},\psi) \\ =& p(\{y_{ijk}\}|\{\theta_{jk}\},\{\phi_k\},\psi)\cdot p(\theta_{jk})|\{\phi_k\},\psi)\cdot p(\{\phi_k\}|\psi)p(\psi). \end{aligned}$$

- More generally, suppose $\mathbf{X} = (X_1, X_2 \ldots, X_d)$ is a joint random vector. Then under DAG we model it under a factorization

$$p(\mathbf{X}) = \prod_i p(X_i|\text{pa}(X_i))$$

where $\text{pa}(X_i)$ represents the "parents" of X_i according to the graph.

Notations for graphical models

- Each random quantity is a vertex in the graph.
- Directional arrows indicate parent-child relationships.
- A circle around an X_i indicates it is random in the model.
- A square around a quantity indicates it is a "hyperparameter" that's fixed under the model.
- A plate notation indicates repeated observations (often used along with exchangeability not always).
- Exercise: When reviewing all the notes from the past, draw the graphical model for every model we have covered.

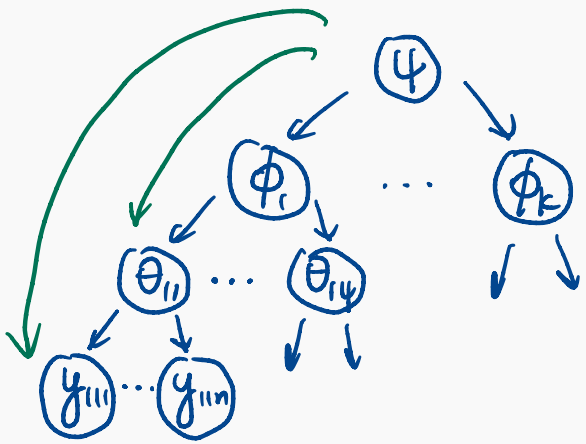
Rationale for hierarchical (i.e., DAG) modeling

- Specifying our model this way allows us to focus on thinking about what are the appropriate *conditional* distributions!
- We can incorporate conditional indpendence in some of these conditional models and use i.i.d. assumptions to reflect exchangeability at different levels.
- For example, it is easy to incorporate exchangeability through conditional independence relationships such as

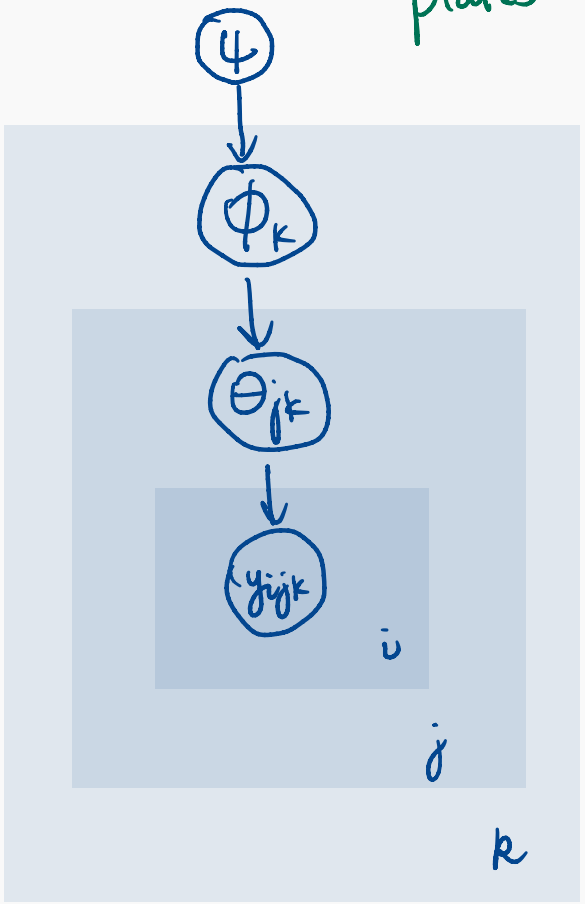
make simplification

$$p(\{y_{ijk}\}|\{\theta_{jk}\},\{\phi_k\},\psi) = p(\{y_{ijk}\}|\{\theta_{jk}\},\psi)$$

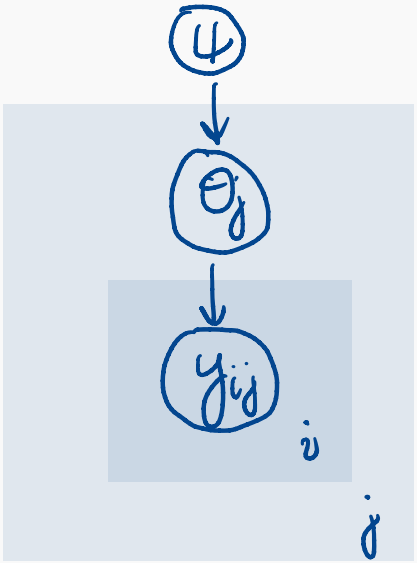
  Reasoning: "A student's score only depends on the school district through the average score in that school."
- It also makes the derivation of the posterior, or more specifically the full conditionals, much easier.

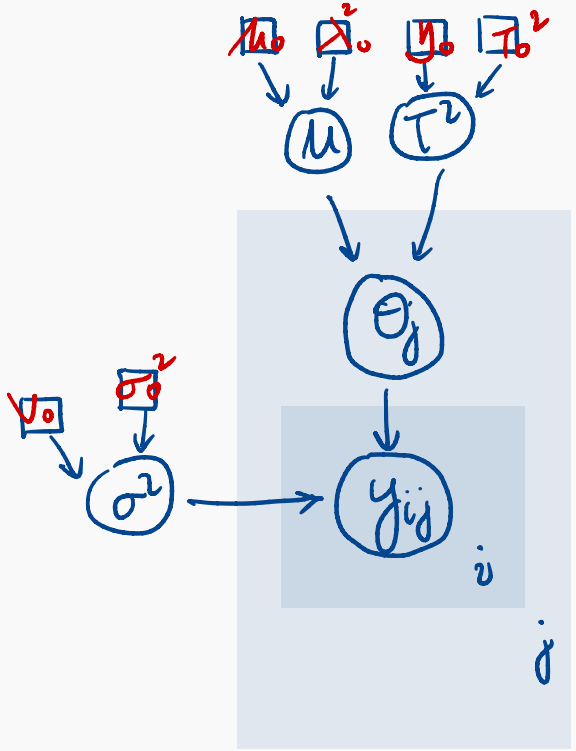


plates



Simplified





$$\theta_j \mid \mu, \tau^2 \overset{iid}{\sim} N(\mu, \tau^2)$$

$$y_{ij} \mid \theta_j, \sigma^2 \overset{iid}{\sim} N(\theta_j, \sigma^2)$$

Differentiate hierarchical model from Bayesian strategy:

① prior: Bayesian treats top layers as random variables and specify priors.

② inference: Bayesian gets joint distribution, and samples (by mcmc) from posterior.

Example: Hierarchical normal

- For illustration, let's consider a model with one level of grouping and the conditional distribution of each layer is a univariate normal model.
- The sampling model is

$$y_{ij} \mid \theta_i, \sigma^2 \overset{\text{iid}}{\sim} \text{N}(\theta_i, \sigma^2)$$
$$\theta_i \mid \mu, \tau \overset{\text{iid}}{\sim} \text{N}(\mu, \tau^2)$$

- For example, y_{ij} is the test score for student j in school i, and θ_i is the average score for school i.
- Note that there are two variance parameters:
  - Within-school variance σ^2. Here we assume σ^2 are the same around schools.
  - Between-school variance τ^2. If τ^2 is large, then borrowing info is hard.
- Here we are assuming that the within-school variance is all equal across schools.
- Is this reasonable?
- Draw the graphical model.

Prior on the parameters

- The (global) parameters for the sampling models are
  - μ: grand mean
  - σ^2: within-group variance
  - τ^2: between-group variance
- For Bayesian inference, we can specify priors on these parameters.
- For example, we decide to adopt a prior that assumes prior independence:

$$p(\mu, \sigma^2, \tau^2) = p(\mu)p(\sigma^2)p(\tau^2)$$

- Note that the θ_i's are also parameters but their distribution is determined by the sampling model given (μ, σ^2, τ^2).
  - The θ_i's are often of prime interest, for example, if I am to evaluate the performance of a particular school i or to predict the score for a student from the school.
  - Some like to refer to the model for θ_i also as prior rather than sampling model.
  - These are just different names — no practical differences!

Hierarchical normal prior specification

- For the hierarchical normal model, we can adopt the standard normal-inverse-Gamma priors:

$$
\begin{aligned}
p(\mu) &\sim \mathrm{N}(\mu_0, \gamma_0^2) \\
p(\sigma^2) &\sim \mathrm{IG}(\nu_0/2, \nu_0\sigma_0^2/2) \\
p(\tau^2) &\sim \mathrm{IG}(\eta_0/2, \eta_0\tau_0^2/2).
\end{aligned}
$$

- As usual, we can choose σ_0^2 and τ_0^2 based on our prior expectation of σ^2 and τ^2, and choose ν_0 and η_0 based on the strength of our prior belief.

Applying Bayes theorem

- To carry out Bayesian inference, we need to find or sample from the joint posterior of all the unknown quantities

$$(\boldsymbol{\theta}, \mu, \sigma^2, \tau^2)$$

  where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_J)$.
- The joint posterior of all these unknowns is proportional to the joint probability of all random quantities

$$\begin{aligned} & p(\boldsymbol{\theta}, \mu, \sigma^2, \tau^2 \,|\, \mathbf{y}) \\ \propto & p(\mathbf{y}|\boldsymbol{\theta}, \sigma^2) p(\boldsymbol{\theta}|\mu, \tau^2) p(\mu) p(\sigma^2) p(\tau^2) \\ = & \left[\prod_{j=1}^{J} \prod_{i=1}^{n_j} p(y_{ij}|\theta_j, \sigma^2) \right] \times \left[\prod_{j=1}^{J} p(\theta_j|\mu, \tau^2) \right] \times p(\mu) p(\sigma^2) p(\tau^2). \end{aligned}$$

Finding the full conditionals

- The full conditional of θ_j:

$$p(\theta_j|\boldsymbol{\theta}_{-j},\mu,\sigma^2,\tau^2,\mathbf{y}) \propto \left[\prod_{i=1}^{n_j} p(y_{ij}|\theta_j,\sigma^2)\right] p(\theta_j|\mu,\tau^2)$$

  where $\boldsymbol{\theta}_{-j}$ represents all θ_k's except for θ_j.
- Observation: This is exactly the same full conditional we would have gotten had we only have the data from school j as our data and placed a $\mathrm{N}(\mu,\tau^2)$ prior on θ_j! In particular, the θ_j is conditionally independent of $\boldsymbol{\theta}_{-j}$ given (μ,σ^2,τ^2).
- Intuition: the information regarding θ_j from observations in other schools are all summarized in the value of μ, σ^2, and τ^2.
- The full conditional for θ_j is having the same form as before

$$[\theta_j|\boldsymbol{\theta}_{-j},\mu,\sigma^2,\tau^2,\mathbf{y}] \sim \mathrm{N}(\mu_{j,n_j},\tau^2_{j,n_j})$$

  where

Don't involve prior paras.

$$\frac{1}{\tau^2_{j,n_j}} = \frac{n_j}{\sigma^2} + \frac{1}{\tau^2} \quad \text{and} \quad \mu_{j,n_j} = \tau^2_{j,n_j}\left(\frac{n_j}{\sigma^2}\cdot\bar{y}_j + \frac{1}{\tau^2}\cdot\mu\right).$$

What does the graphical model tell us

- This observation is clearly reflected in the graphical model!
- A path from a variable X_i to a variable X_j is a sequence of variables linked by arrows from X_i to X_j that does not include a *collider*.
  - E.g., $X_2 \to X_5 \to X_3 \to X_6$ is a path from X_2 to X_6.
  - E.g., $X_2 \leftarrow X_5 \to X_3 \to X_6$ *is* a path from X_2 to X_6.
  - E.g., $X_2 \to X_5 \leftarrow X_3 \to X_6$ is *not* a path from X_2 to X_6. X_5 is a *"collider"*, which blocked the flow of information.
- Conditioning on a variable along a path "blocks" the path, i.e., the information flow.
- Conditioning on a collider variable does the opposite: it *"unblocks"* the information flow!

Relationship of paths blocking and conditional independence

- Let A, B, C be three sets of variables.
- If conditioning on the variables in B, all "paths" from variables in A to variables in C are "blocked" then

$$A \perp\!\!\!\perp C \,|\, B.$$

- Presence of an unblocked path indicates correlation between A and B.
- For the current example, what "paths" are blocked for the full conditional of θ_j?
- Be aware of "colliders" among children!
  - If no colliders the full conditional will depend just on parents and children.
  - If there are colliders among children, i.e., there are other parents for the children, the full conditional will also depend on those parents.

Finding the full conditionals

- The full conditional for (μ, τ^2) is

$$p(\mu, \tau^2 | \boldsymbol{\theta}, \sigma^2, \mathbf{y}) \propto \left[\prod_{j=1}^{J} p(\theta_j | \mu, \tau^2) \right] p(\mu) p(\tau^2)$$

- This is exactly the posterior for a normal "sample" $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ with priors $p(\mu)p(\tau^2)$.
- So the full conditional on μ and τ^2 will be exactly those for a single normal sample $\boldsymbol{\theta}$.
- Check the graphical model.

- Specifically, the full conditional for μ given by

$$p(\mu|\boldsymbol{\theta},\sigma^2,\tau^2,\mathbf{y}) \propto \left[\prod_{j=1}^{J} p(\theta_j|\mu,\tau^2)\right] p(\mu)$$

is

$$\mu|\boldsymbol{\theta},\sigma^2,\tau^2,\mathbf{y} \sim \mathrm{N}(\mu_J,\lambda_J^2)$$

where

$$\frac{1}{\lambda_J^2} = \frac{1}{\lambda_0^2} + \frac{J}{\tau^2} \quad \text{and} \quad \mu_J = \lambda_J^2\left(\frac{\mu_0}{\lambda_0^2} + \frac{J}{\tau^2}\bar{\theta}\right).$$

- Check graphical model.

- The full conditional for τ^2 given by

$$p(\tau^2|\boldsymbol{\theta},\sigma^2,\mu,\mathbf{y}) \propto \left[\prod_{j=1}^{J} p(\theta_j|\mu,\tau^2)\right] p(\tau^2)$$

is

$$\tau^2|\boldsymbol{\theta},\sigma^2,\mu,\mathbf{y} \sim \text{IG}(\eta_J/2, \eta_J\tau_J^2/2)$$

where

$$\eta_J = \eta_0 + J \quad \text{and} \quad \eta_J\tau_J^2 = \eta_0\tau_0^2 + \sum_{j=1}^{J}(\theta_j - \mu)^2.$$

- Check the graphical model.

- Finally, the full conditional for σ^2 is given by

$$
\begin{aligned}
p(\sigma^2|\boldsymbol{\theta},\mu,\tau^2,\mathbf{y}) &\propto \left[\prod_{j=1}^{J}\prod_{i=1}^{n_j} p(y_{ij}|\theta_j,\sigma^2)\right]\cdot p(\sigma^2) \\
&\propto \left[\prod_{j=1}^{J}\prod_{i=1}^{n_j} (\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(y_{ij}-\theta_j)^2}\right]\cdot (\sigma^2)^{-\frac{\nu_0}{2}+1} e^{-\frac{\nu_0\sigma_0^2}{2\sigma^2}} \\
&= (\sigma^2)^{-\frac{\nu_0+\sum_{j=1}^{J} n_j}{2}+1} e^{-\frac{\nu_0\sigma_0^2+\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\theta_j)^2}{2\sigma^2}}.
\end{aligned}
$$

- That is,

$$\sigma^2\,|\,\boldsymbol{\theta},\mu,\tau^2,\mathbf{y} \sim \mathrm{IG}(\nu_N/2, \nu_N\sigma_N^2/2)$$

where for $N=\sum_{j=1}^{J} n_j$,

$$\nu_N = \nu_0 + N \quad \text{and} \quad \nu_N\sigma_N^2 = \nu_0\sigma_0^2 + \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\theta_j)^2.$$

- Check the graphical model.

Some observations

* Why observation of students in school 1 can be used to make inference of students in school 2 ?
* High $y_{i1} \longrightarrow$ High $\bar{y}_1 \longrightarrow$ Update $\theta_1 \uparrow \longrightarrow$ High $\bar{\theta}$ $\longrightarrow$ Update $\mu \uparrow \longrightarrow$ Update $\theta_2 \uparrow$.

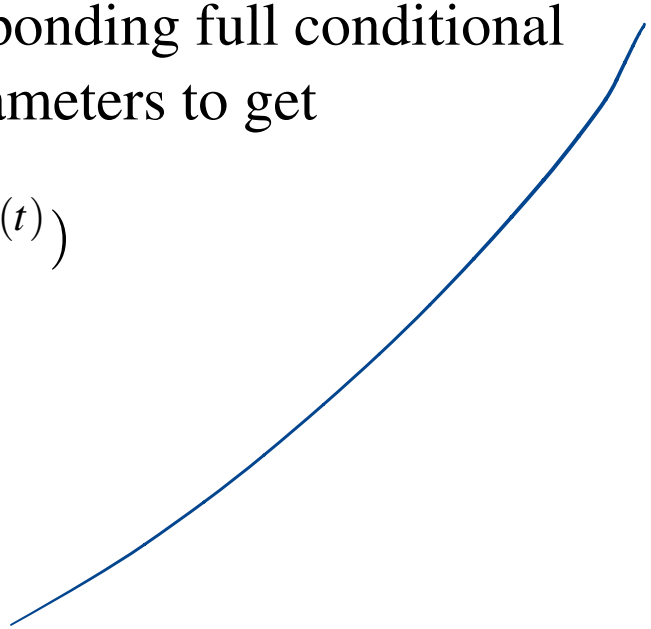
- In this example, despite the complexity of the entire model, the full conditionals can all be found by conjugate updates.
- Is this a coincidence?
- In Bayesian hierarchical modeling, if the conditional models are specified in a conjugate way, that is, then the full conditionals are simple conjugate updates.
- This can be readily seen by the graphical model representation.

Gibbs sampling

- Given the full conditionals, we can implement a Gibbs sampler as usual.
- After initializing $(\boldsymbol{\theta}^{(0)}, \mu^{(0)}, \sigma^{2\,(0)}, \tau^{2\,(0)})$, draw by iteratively update each parameter from their corresponding full conditional given the current values of the other parameters to get

$$(\boldsymbol{\theta}^{(t)}, \mu^{(t)}, \sigma^{2\,(t)}, \tau^{2\,(t)})$$

for $t = 1, 2, \ldots$.



Gibbs sampling

- Initializing $(\boldsymbol{\theta}^{(0)}, \mu^{(0)}, \sigma^{2\,(0)}, \tau^{2\,(0)})$.
- For $t = 1, 2, \ldots, T$
  - Update μ:
    - Compute

$$\lambda_J^{2\,(t)} = \left(\frac{1}{\lambda_0^2} + \frac{J}{\tau^{2\,(t-1)}} \right)^{-1} \quad \text{and} \quad \mu_J^{(t)} = \lambda_J^{2\,(t)} \left(\frac{\mu_0}{\lambda_0^2} + \frac{J\bar{\theta}^{(t-1)}}{\tau^{2\,(t-1)}} \right)$$

    - Draw

$$\mu^{(t)} \sim \mathrm{N}\left(\mu_J^{(t)}, \lambda_J^{2\,(t)} \right)$$

  - Update τ^2:
    - Compute

$$\eta_J = \eta_0 + J \quad \text{and} \quad \eta_J \tau_J^{2\,(t)} = \eta_0 \tau_0^2 + \sum_{j=1}^{J} (\theta_j^{(t-1)} - \mu^{(t)})^2.$$

    - Draw

$$\tau^{2\,(t)} \sim \mathrm{IG}\left(\frac{\eta_J}{2}, \frac{\eta_J \tau_J^{2\,(t)}}{2} \right)$$

Gibbs sampling

- Update σ^2:
  - Compute

$$\nu_N = \nu_0 + N \quad \text{and} \quad \nu_N \sigma_N^{2\,(t)} = \nu_0 \sigma_0^2 + \sum_j \sum_i (y_{ij} - \theta_j^{(t-1)})^2$$

  - Draw

$$\sigma^{2\,(t)} \sim \text{IG}(\nu_N/2, \nu_N \sigma_N^{2\,(t)}/2)$$

- For $j = 1, 2, \ldots, J$
  - Compute

$$\tau_{j,n_j}^{(t)} = \left(\frac{n_j}{\sigma^{2\,(t)}} + \frac{1}{\tau^{2\,(t)}} \right)^{-1} \quad \text{and} \quad \mu_{j,n_j}^{(t)} = \tau_{j,n_j}^{2\,(t)} \left(\frac{n_j \bar{y}_j}{\sigma^{2\,(t)}} + \frac{\mu^{(t)}}{\tau^{2\,(t)}} \right).$$

  - Draw

$$\theta_j^{(t)} \sim \text{N}(\mu_{j,n_j}^{(t)}, \tau_{j,n_j}^{2\,(t)}).$$

Borrowing information through Bayesian Shrinkage

- Now let's again look at the full conditional for θ_j:
- So the full conditional for θ_j is having the same form as before

$$[\theta_j|\boldsymbol{\theta}_{-j}, \mu, \sigma^2, \tau^2, \mathbf{y}] \sim \mathrm{N}(\mu_{j,n_j}, \tau^2_{j,n_j})$$

where

* If τ^2 is small: schools are similar, the guess is precise, and μ_{j,n_j} is shrunk to grand mean μ.

$$\frac{1}{\tau^2_{j,n_j}} = \frac{n_j}{\sigma^2} + \frac{1}{\tau^2} \quad \text{and} \quad \mu_{j,n_j} = \tau^2_{j,n_j}\left(\frac{n_j}{\sigma^2}\cdot \bar{y}_j + \frac{1}{\tau^2}\cdot\mu\right).$$

- How does an observations in school j' contribute to the inference on θ_j?

* If σ^2 is small: students in school j are homogeneous, then guess about the school is precise.

* If n_j is large: more empirical evidence, and the guess is also precise.

μ_{j,n_j} is more shrunk to empirical mean $\bar{y}_j$.

Borrowing information

- The Gibbs update for θ_j will shrink $\bar{y}_j$ toward μ, the grand mean, whose value is influenced by $\bar{\theta}_{-j}$, each of which is influenced by the average scores in the other schools $\mathbf{y}_{-j}$.
- The amount of shrinkage is determined by n_j/σ^2 and τ^2.
  - In praticular, τ^2 is the between-group variance, which characgterize how similar those schools are to each other.
  - If τ^2 is large, then the schools are very different, in which case we do not shrinkage $\bar{y}_j$ toward the average of all schools by much.
  - If τ^2 is small, (in the extreme $\tau = 0$ so all schools are equal), then we shrinkage more toward the other schools.
  - If n_j/σ^2 is large, e.g., the sample size for school j, n_j is huge, then we shrink less toward other schools.
- It is exactly through this mechanism that "borrowing of information" across groups is achieved.

The ESL example

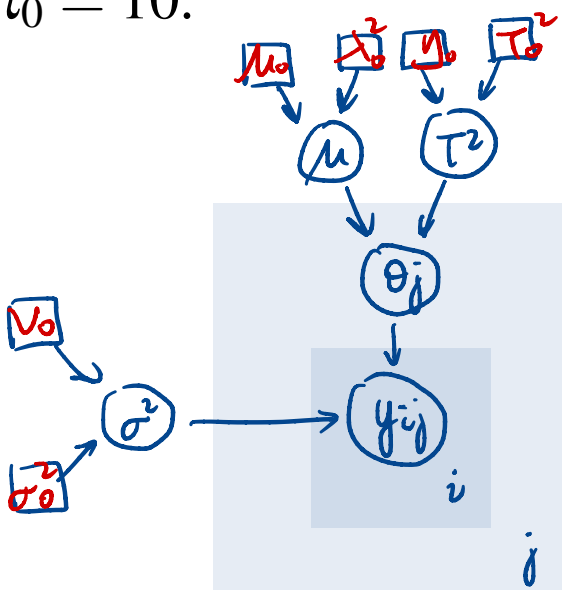
empirical bayes is fundamentally frequentist (treat to layer μ, τ^2 as fixed)

Sharing information is also allowed, but you lose some uncertainty quantification.

If you look at the "posterior" of θ_j, due to the fixed truth of μ, τ^2, the posterior uncertainty will be less spread out.

(It would be okay if the fixed top layer is much higher.)

- Let's look at the ESL data set from the textbook.
- Here $J = 100$ high schools, and the observations y_{ij}'s are the math scores for a sample of students from each school.
- Prior specification: $\mu_0 = 50$, $\lambda_0 = 10$, $\nu_0 = 1$, $\sigma_0 = 10$, $\eta_0 = 1$, $\tau_0 = 10$.



If you pre-determine μ, τ^2 before even looking at data, then there is no information sharing.

Summary of the data

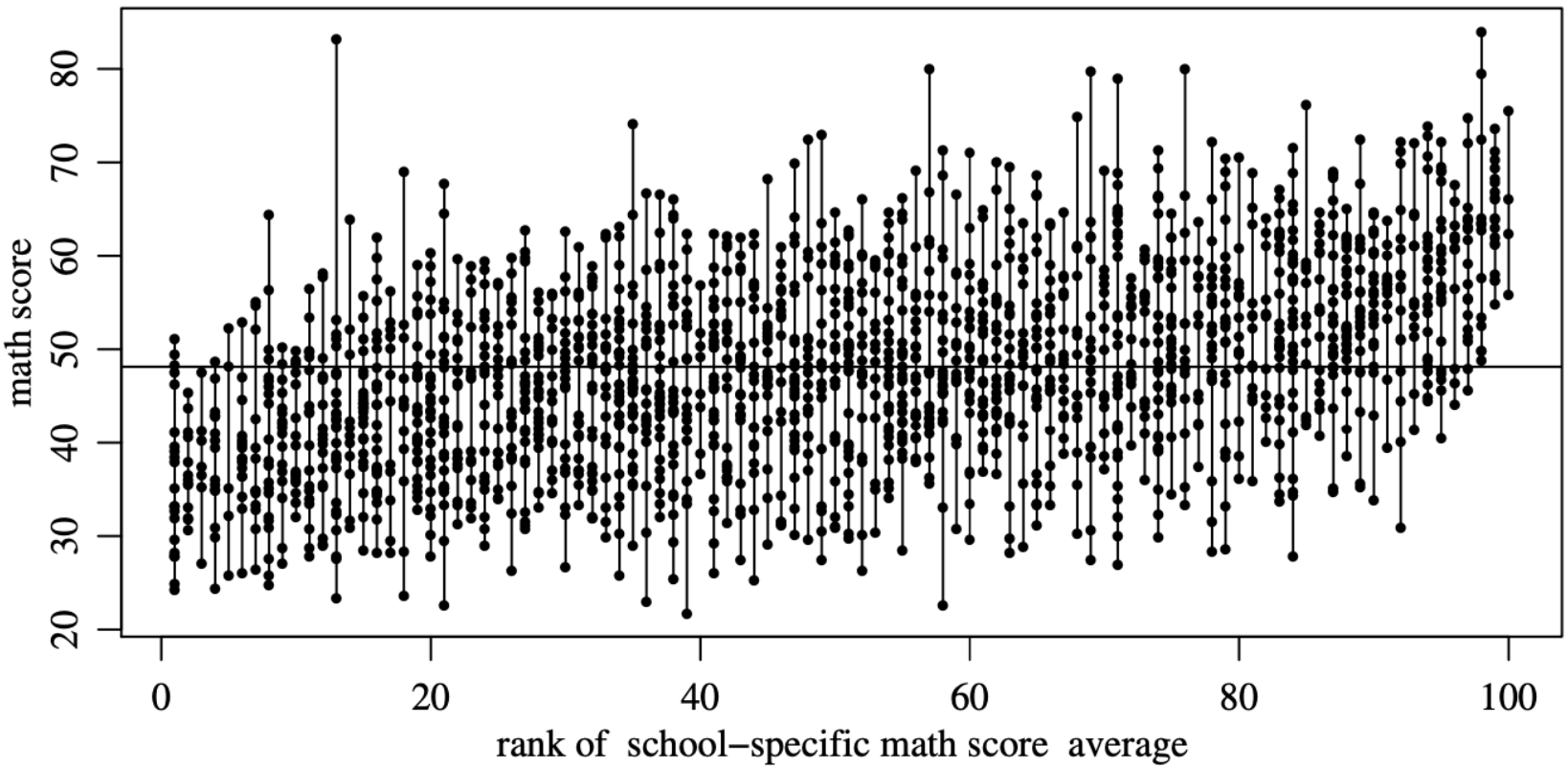
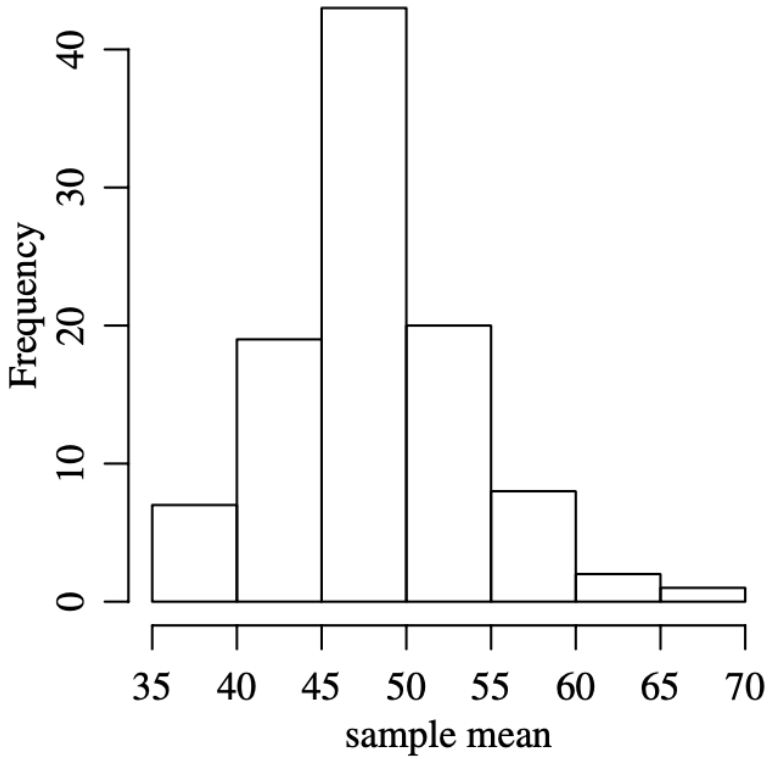


Fig. 8.4. A graphical representation of the ELS data.

Summary of the data

variability decreases as sample sizes increases.

unreliable: don't want to make inference based on only 4 or 5 data points.



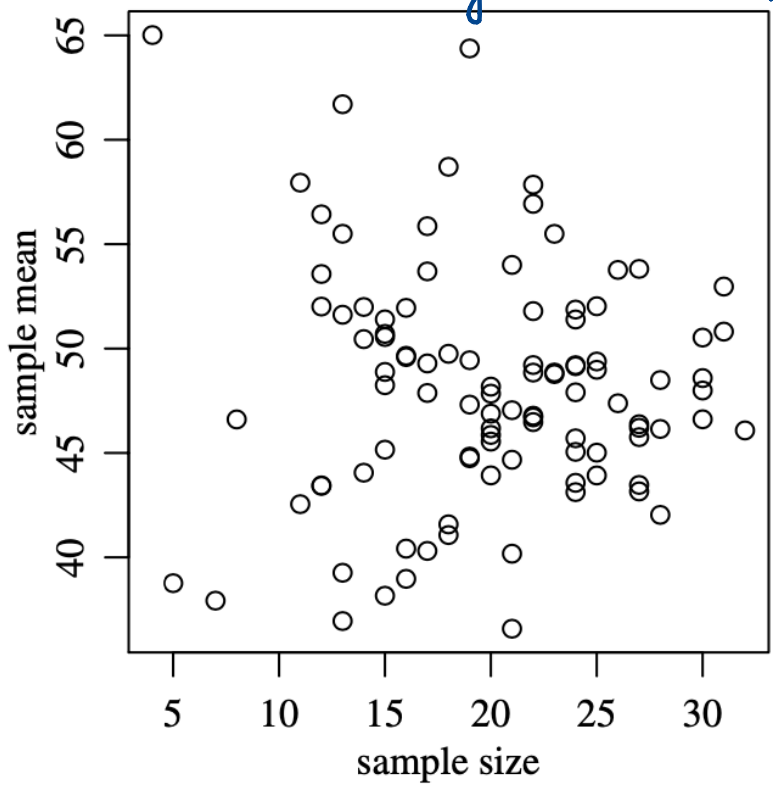


Fig. 8.5. Empirical distribution of sample means, and the relationship between sample mean and sample size.

Marginal posterior densities

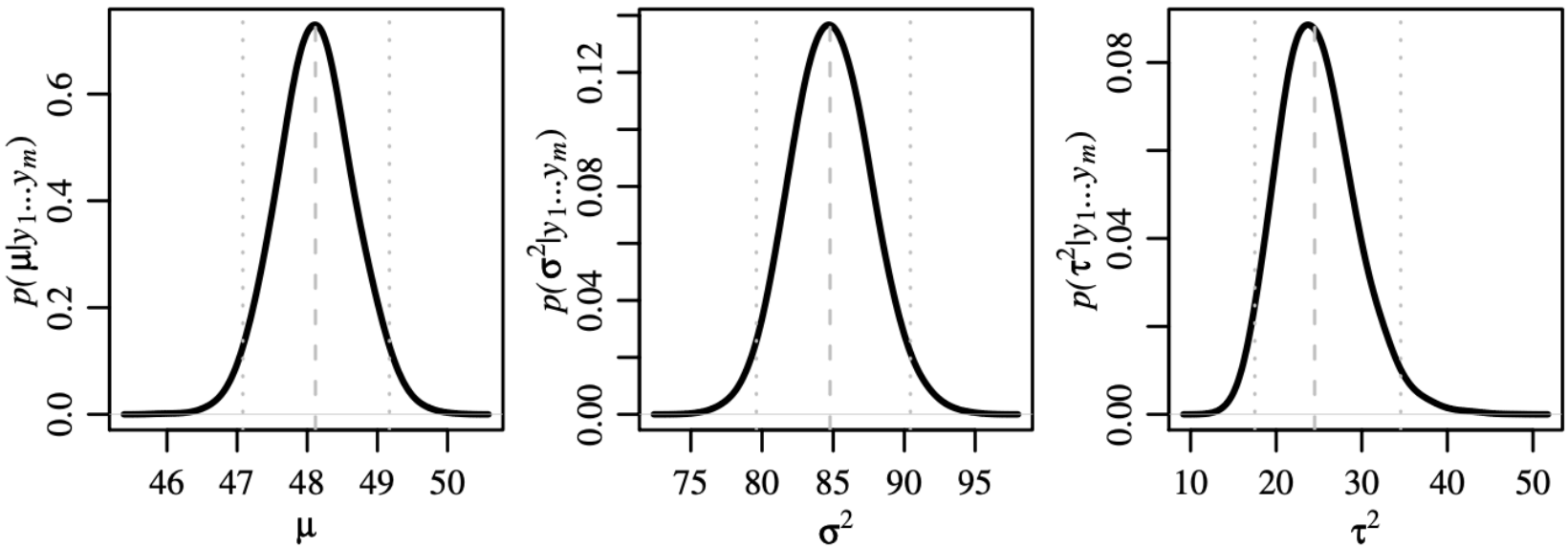


Fig. 8.7. Marginal posterior distributions, with 2.5%, 50% and 97.5% quantiles given by vertical lines.

Bayesian shrinkage

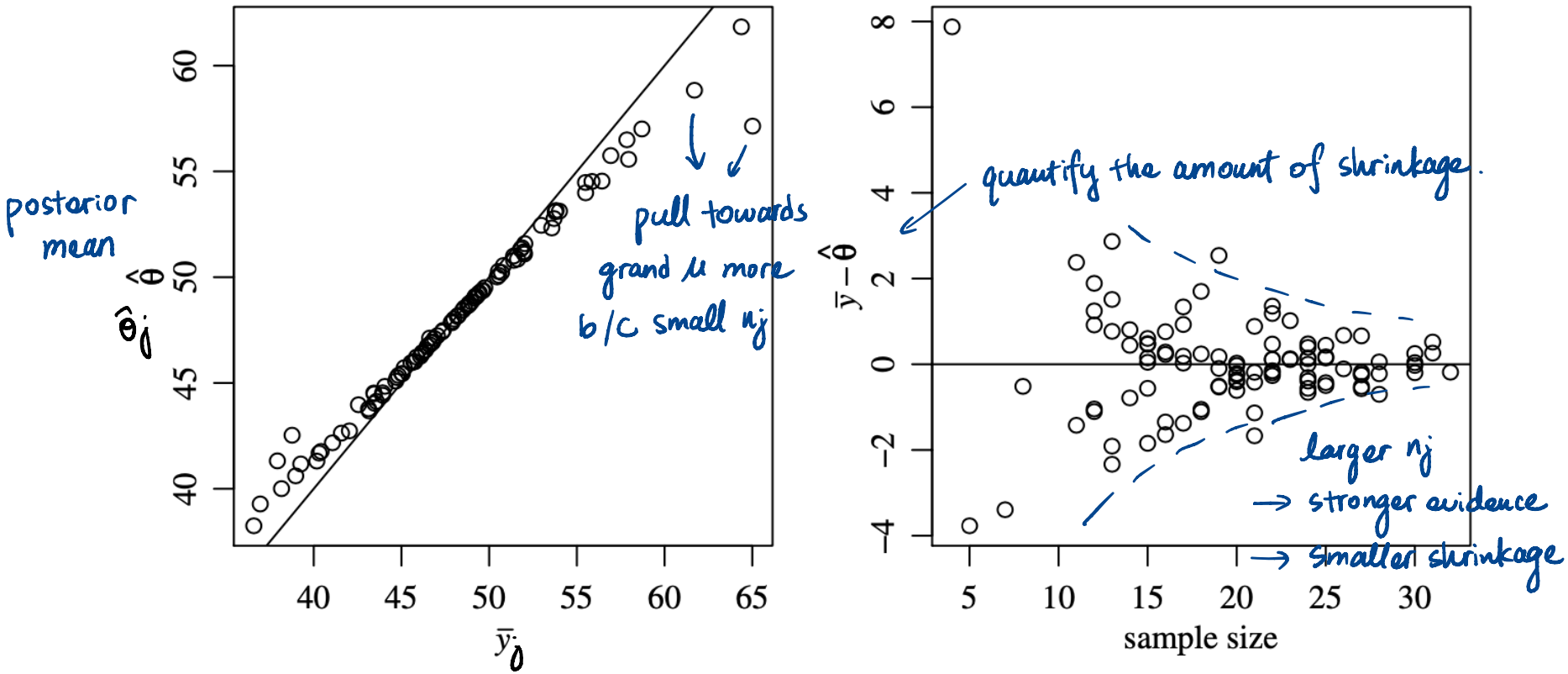


Fig. 8.8. Shrinkage as a function of sample size.

Comparison of two schools

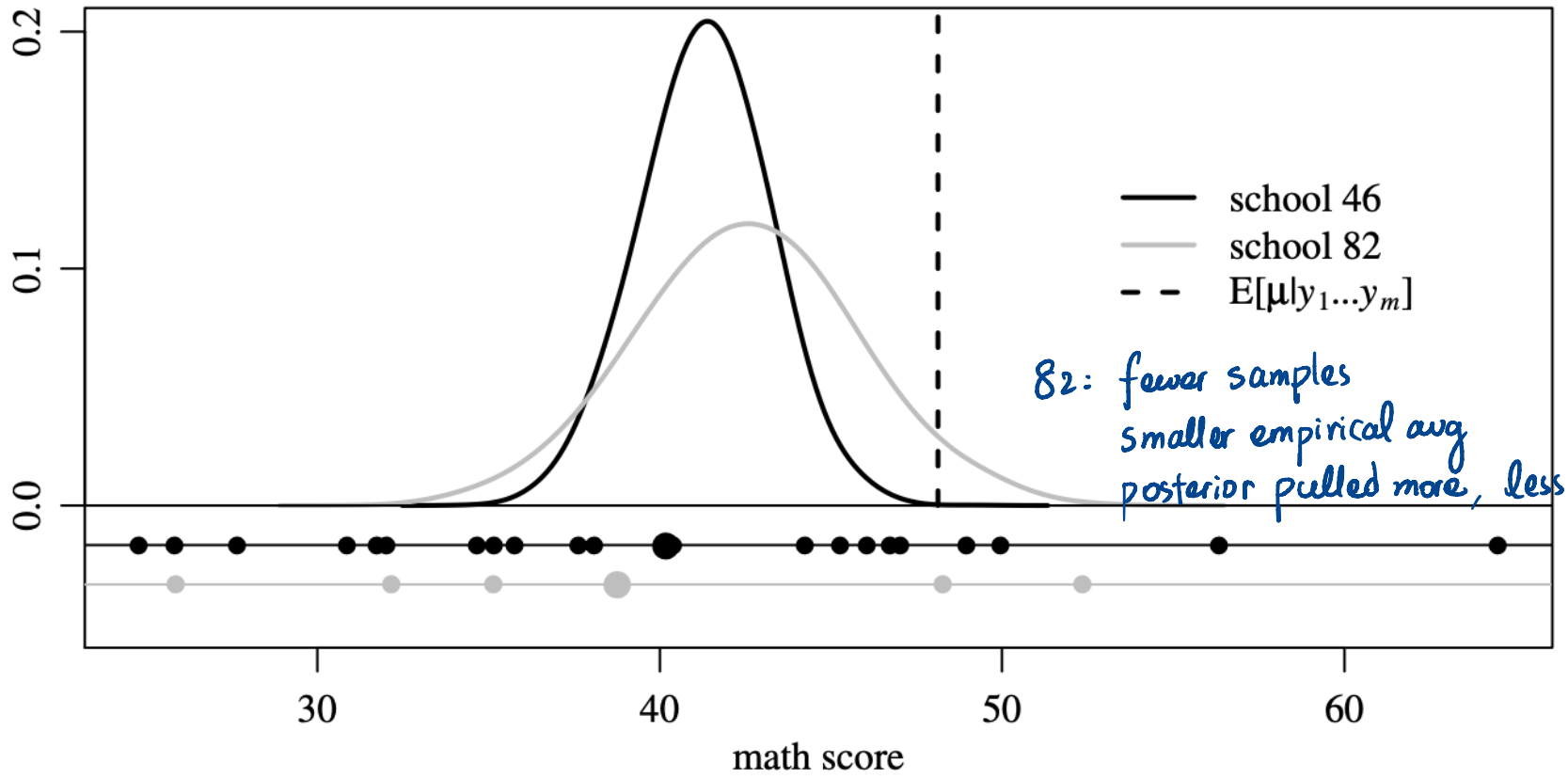


Fig. 8.9. Data and posterior distributions for two schools.

Group-specific within-group variance

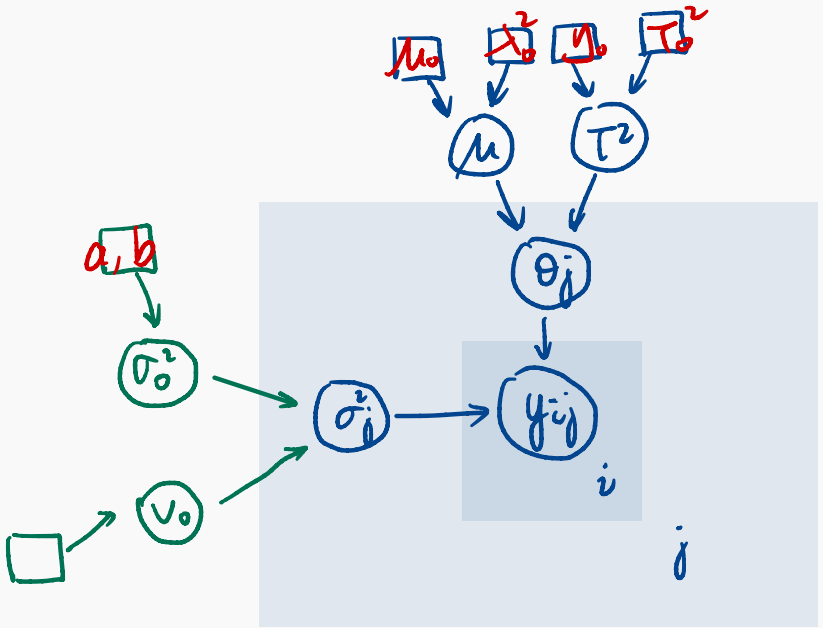
- In the above model we have use σ^2 to model the within-school variance for all schools.
- In reality this assumption is probably unrealistic. Some schools will have wider performance spread among its students than others.
- A natural extension of the above model is to let σ_j^2 be a group-specific variance parameter for each school, that is, in our hierarchical model, we let

$$y_{ij} \,|\, \theta_j, \sigma_j^2 \overset{\text{iid}}{\sim} \text{N}(\theta_j, \sigma_j^2).$$

- Then σ_j^2 can also be modeled exchangeability as

$$\sigma_j^2 \,|\, \nu_0, \sigma_0^2 \overset{\text{iid}}{\sim} \text{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right).$$

- Draw a graphical model representation.



$$X \rightarrow Z \leftarrow Y$$

A priori: $x \perp\!\!\!\perp y$

A posteriori: $x \not\perp\!\!\!\perp y$.

$\sigma_j^2 \mid \sigma_0^2, \nu_0 \overset{iid}{\sim} IG\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$

Downstream: IG

Upstream: $\sigma_0^2 \sim Gamma(a, b)$

$\sigma_0^2 \mid \sigma_j^2, \nu_0, a, b \sim Gamma(a^*, b^*)$

Borrowing of information

- Then in order to allow borrowing information among the j groups in inferring the within-group variances, we can place priors on (ν_0, σ_0^2) as well.
- Note that only when we allow ν_0 and σ_0^2 to be inferred from the data can we achieve borrowing of information across groups for inferring σ_j's.
- This will allow us to borrow information across the samples in a "soft" way.
- Note that previously, we have assumed that all σ_j^2 are equal, and so borrowing of infomration was in a "hard" way, all variances are forced to take exactly the same value σ^2, and all observations y_{ij} in all groups contribute to the inference on σ.
- Here, σ_0^2 play the role of σ as before, which characterizes the average of σ_j^2, which are now allowed to deviate from σ_0^2.

Influence on shrinkage

- Note that allowing σ_j^2 to differ for each sample not only changes our inference on the within-school variance but will also further influence the amount of shrinkage on θ_j toward μ.
- Schools with larger within-school variance σ_j^2 will be pulled toward the grand mean more than those that have smaller within-school variance.
- To see this, note that the full conditional for θ_j becomes

$$\theta_j | \boldsymbol{\theta}_{-j}, \mu, \tau^2, \sigma_j^2, \mathbf{y} \sim \mathrm{N}(\mu_{j,n_j}, \tau_{j,n_j}^2)$$

where

$$\tau_{j,n_j}^2 = \left(\frac{n_j}{\sigma_j^2} + \frac{1}{\tau^2} \right)^{-1}$$

and

$$\mu_{j,n_j} = \tau_{j,n_j}^2 \left(\frac{n_j}{\sigma_j^2} \bar{y}_j + \frac{1}{\tau^2} \mu \right).$$

Full conditional for σ_j^2

- The full condtional for τ^2 is exactly the same as before, which can be seen from the graphical model representation easily.
- The full conditional for σ_j^2

$$p(\sigma_j^2|\boldsymbol{\theta},\mu,\tau^2,\nu_0,\sigma_0^2,\mathbf{y}) \propto \left[\prod_{i=1}^{n_j} p(y_{ij}|\theta_j,\sigma_j^2)\right] \cdot p(\sigma_j^2|\nu_0,\sigma_0^2).$$

- Thus

$$\sigma_j^2\,|\,\boldsymbol{\theta},\mu,\tau^2,\mathbf{y} \sim \text{IG}\left(\frac{\nu_0+n_j}{2}, \frac{\nu_0\sigma_0^2+\sum_i(y_{ij}-\theta_j)^2}{2}\right)$$

- The mean of the full conditional is

$$\text{E}(\sigma_j^2\,|\,\boldsymbol{\theta},\mu,\tau^2,\mathbf{y}) = \frac{\nu_0\sigma_0^2+\sum_i(y_{ij}-\theta_j)^2}{\nu_0+n_j-2},$$

which shrinks $\sum_i(y_{ij}-\theta_j)^2/n_j$ toward the prior mean $\nu_0\sigma_0^2/(\nu_0-2)$.

Prior and full conditional of σ_0^2

- To get conjugate full conditional on σ_0^2, as before we can use an either a Gamma ~~or an inverse-Gamma~~ prior σ_0^2.
- For example, if we use a Gamma prior

$$\sigma_0^2 \sim \text{Gamma}(a, b)$$

- Check that the full conditional of σ_0^2 is

$$\begin{aligned} p(\sigma_0^2 \mid \nu_0, \sigma_1^2, \ldots, \sigma_J^2) &\propto \prod_j p(\sigma_j^2 | \nu_0, \sigma_0^2) \cdot p(\sigma_0^2) \\ &\propto \prod_j \left(\frac{\nu_0 \sigma_0^2}{2} \right)^{\nu_0/2} e^{-\frac{\nu_0 \sigma_0^2}{2\sigma_j^2}} \cdot \left(\sigma_0^2\right)^{a-1} e^{-b\sigma_0^2} \\ &\propto \left(\sigma_0^2\right)^{J\nu_0/2+a-1} e^{-(b+\frac{\nu_0}{2}\sum_j \frac{1}{\sigma_j^2})\sigma_0^2} \end{aligned}$$

which is

$$\text{Gamma}\left(a + \frac{J\nu_0}{2}, b + \frac{\nu_0}{2} \sum_j \frac{1}{\sigma_j^2} \right).$$

Prior and full conditional on ν_0

- So far we are able to get conjugate forms in the full conditionals.
- But this is no longer possible for ν_0.
- One strategy is to approximate ν_0 using a grid of integer values, which will allow us to compute the posterior on that grid using brute-force computation (i.e., enumeration).
- For example, the textbook suggests using

$$p(\nu_0) \propto e^{-\alpha\nu_0}$$

for some $\alpha > 0$ on $\{1, 2, \ldots\}$.

- Then the full conditional of ν_0 is also a discrete distribution on $\{1,2,\ldots\}$, and it is given by Bayes theorem

$$
\begin{aligned}
& p(\nu_0 \mid \sigma_1^2, \ldots, \sigma_J^2, \sigma_0^2) \\
& \propto \left[\prod_{j=1}^{J} p(\sigma_j^2 | \sigma_0^2, \nu_0) \right] \cdot p(\nu_0) \\
& \propto \left[\prod_{j=1}^{J} \frac{\left(\frac{\nu_0 \sigma_0^2}{2} \right)^{\nu_0/2}}{\Gamma\left(\frac{\nu_0}{2}\right)} \left(\sigma_j^2\right)^{-\nu_0/2-1} e^{-\frac{\nu_0 \sigma_0^2}{2\sigma_j^2}} \right] \cdot e^{-\alpha \nu_0} \\
& \propto \left(\frac{\nu_0 \sigma_0^2}{2} \right)^{J\nu_0/2} \Gamma\left(\frac{\nu_0}{2}\right)^{-J} \left(\prod_{j=1}^{J} \sigma_j^2 \right)^{-\frac{\nu_0}{2}} e^{-\nu_0 \left(\alpha + \frac{\sigma_0^2}{2} \sum_{j=1}^{J} \frac{1}{\sigma_j^2} \right)}.
\end{aligned}
$$

- While this looks daunting, it's a piece of cake for a modern computer to calculate this on a large grid, e.g., $\{1,2,\ldots,10000\}$.
- We can then draw from the multinomial distribution on this grid to (approximately) sample from the full coditional of ν_0.
- We can also use more general MCMC algorithms (e.g., Metropolis-Hastings). More on this later.

Posterior densities

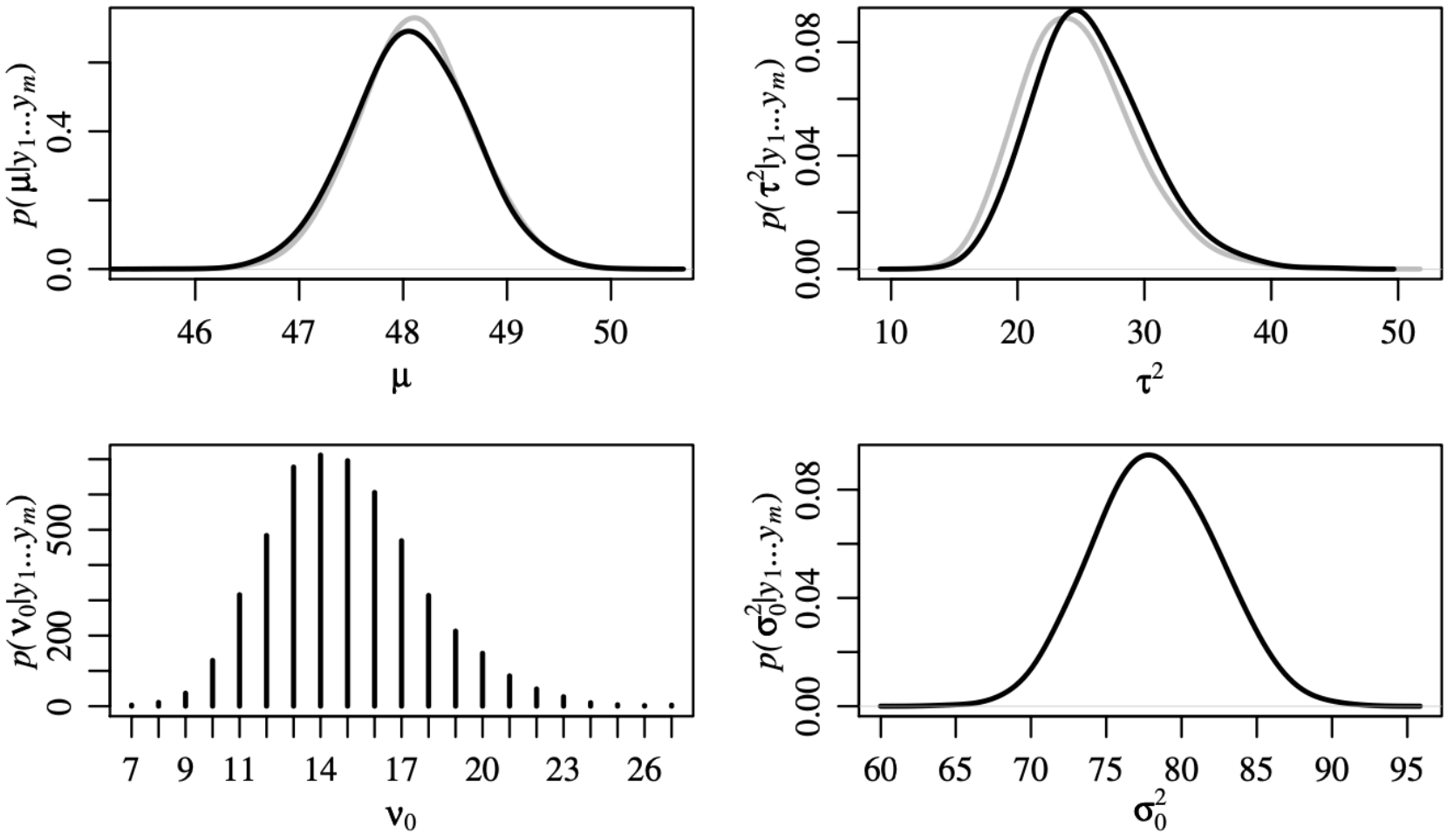


Fig. 8.11. Posterior distributions of between-group heterogeneity parameters.

Posterior shrinkage on σ_j^2

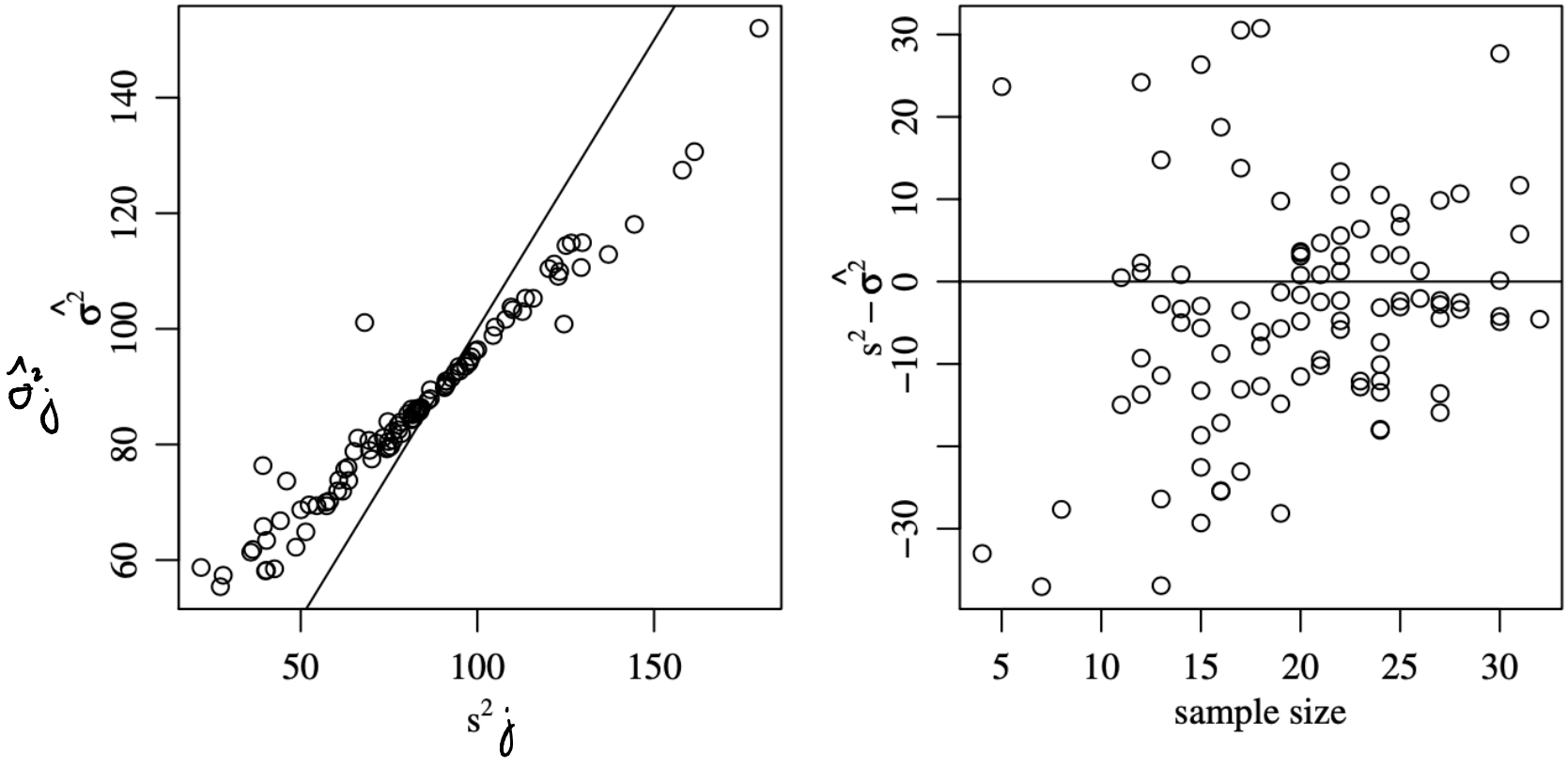


Fig. 8.12. Shrinkage as a function of sample size.