Bayesian Modeling

Group comparisons and hierarchical modeling

Yixin Wang

8 Group comparisons and hierarchical modeling

In this section we will study questions related to comparisons of different populations. While group comparison may conjure up the question of ranking, a thorough treatment will inevitably require thinking of notions such as within-group variability and between-group variability. Such notions will be best addressed by employing (Bayesian) hierarchical modeling. In this sense, this section is also good entry point to hierarchical modeling, which is applicable far beyond the basic group comparison problems. In fact, hierarchical modeling is also one of the most powerful tools in the arsenal of Bayesian statistics.

8.1 Comparing two groups

Example 8.1. Given a sample of 10th grade students from two public U.S. high schools. $n_1 = 31$ and $n_2 = 28$ are the two sample sizes from school 1 and 2, respectively. Both schools have a total enrollment of around 600 10th graders and both are in a similar environment (urban neighborhoods).

- Suppose we are interested in comparing the population means θ_1 and θ_2.
- Sample means: $\bar{y}_1 = 50.81$ and $\bar{y}_2 = 46.15$ suggesting that $\theta_1 > \theta_2$.
- Let's take a look at the box plots. There are evidently different levels of variability in the two groups. A standard approach is to consider the t-statistic:

$$t(\boldsymbol{y}_1, \boldsymbol{y}_2) \quad = \quad \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{1/n_1 + 1/n_2}} = \frac{50.81 - 46.15}{10.44\sqrt{1/31 + 1/28}} = 1.74,$$

where $s_p = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$, the pooled estimate of the population variance of the two groups.

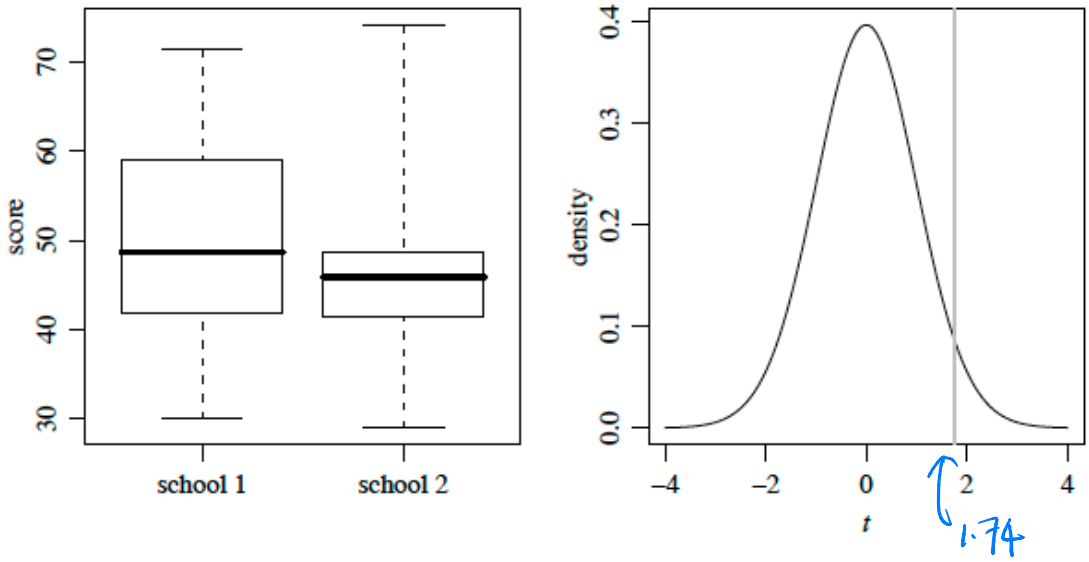


Figure 8.1: Left panel: Boxplots of samples of math scores from two schools. Right panel: gray line indicates the observed value of the t-statistic.

A basic frequentist technique (the t-test) proceeds as follows.

null hypothesis

- Exploit the fact that if the two populations are normal distributions with the same mean and variance, then the t-statistic $t(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ is a t-distribution with $n_1+n_2-2 = 57$ degrees of freedom. The density of this distribution is plotted in the second panel of Fig. 8.2. Under this distribution, the probability that $|t(\boldsymbol{Y}_1, \boldsymbol{Y}_2)| > 1.74$ is $p = 0.087$. This is called the (two-sided) p-value of the obtained statistic.

  - Although not completely justified in theory, p-values are widely used and easily misused and abused in parameter estimation and model selection. A small p-value is considered as indicating the evidence supporting the rejection of the null hypothesis/ model $\theta_1 = \theta_2$. Thus, a small p value is construed with a strong evidence that the two populations are different ($\theta_1 \neq \theta_2$). Customarily, p is considered small if $p < 0.05$ (or a smaller positive threshold number).
  - Mathematically,

    $$p = \Pr(|t(\boldsymbol{Y}_1, \boldsymbol{Y}_2)| > t(\boldsymbol{y}_1, \boldsymbol{y}_2)|\theta_1 = \theta_2).$$

    This is a (pre-experiment) probability statement on the unseen data represented by $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$, even though the observed statistic $t(\boldsymbol{y}_1, \boldsymbol{y}_2)$ supplies part of the equation that defines p-value. This is a source of confusion for many practitioners of frequentist tests. It should not be the case for a student of Bayesian statistics. Clearly, p is *not* the (post-experiment) probability that $\theta_1 = \theta_2$ is true given the data evidence provided by $t(\boldsymbol{y}_1, \boldsymbol{y}_2)$:

    $$\Pr(\theta_1 = \theta_2|t(\boldsymbol{y}_1, \boldsymbol{y}_2)).$$

- The t-test commonly taught in statistic classes continues as follows:
  - if $p < 0.05$: reject the null hypothesis/model that the two groups have the same distributions; conclude that $\theta_1 \neq \theta_2$. Moreover, use the estimates:

$$\hat{\theta}_1 = \bar{y}_1; \quad \hat{\theta}_2 = \bar{y}_2.$$

  - if $p \geq 0.05$: accept the null hypothesis/model, and conclude that $\theta_1 = \theta_2$. Moreover, use the estimate

$$\hat{\theta}_1 = \hat{\theta}_2 = (\sum y_{i,1} + \sum y_{i,2})/(n_1 + n_2).$$

- In our present example: $p \geq 0.05$, so we accept that $\theta_1 = \theta_2$, even though there seems to be some evidence to the contrary.

$p = 0.051$

vs $p = 0.049$

- Imagine a scenario where the sample from school 1 might have included a few more high-performing students, and the sample from school 2 a few more low-performing students. Then we could have observed a p-value of 0.04 or so, in which case we would have treated the two populations as different, and resorted to using only data from school 1 for estimating θ_1, and data from school 2 for estimating θ_2. It seems such estimates for θ_1 and θ_2 are not robust with respect to changes to the samples. [5]

- Estimating θ_1 and θ_2 and the difference $\theta_1 - \theta_2$ is perhaps more important than determining in binary the question whether $\theta_1 \neq \theta_2$ or not when the difference between the two is relatively small. The above frequentist approach results in taking two extreme positions for the estimation of θ_1 and θ_2:

$$\hat{\theta}_1 = w_1 \bar{y}_1 + (1 - w_1)\bar{y}_2$$
$$\hat{\theta}_2 = (1 - w_2)\bar{y}_1 + w_2 \bar{y}_2,$$

where $w_1 = w_2 = 1$ if $p < 0.05$ and $w_1 = n_1/(n_1 + n_2); w_2 = n_2/(n_1 + n_2)$ otherwise.

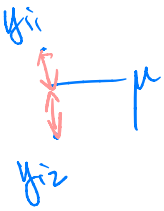
- It might make more sense to allow w to vary continuously and have a value that depends on quantities such as sample sizes n_1, n_2 and other quantities that determine population variabilities. In other words, we want to allow the borrowing of information across groups: the data from group 1 may influence the estimate for group 2 and vice versa.

[5] In the t-test, as is the case with most frequentist tests, we are on a firm mathematical ground when we happen to reject. I.e., the rejection is mathematically justified. However, in such a scenario for the t test, our estimates may not be robust for the issue mentioned. When we happen to *not* reject, i.e., we remain with the null hypothesis/model, then the issue becomes whether the null model is too simplistic and heavily misspecified; the estimates would be suspect as a result.

$\begin{cases} y_{i1} - \delta \mid \delta, \mu, \sigma^2 \sim N(\mu, \sigma^2) \\ y_{i2} + \delta \mid \delta, \mu, \sigma^2 \sim N(\mu, \sigma^2) \end{cases}$

difference between θ_1 & $\theta_2 = 2\delta$

μ: overall mean of 2 groups

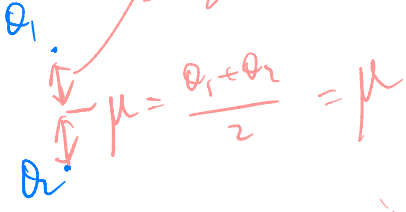


Enabling information sharing across groups Consider the following sampling model for two groups:

$$\begin{array}{lrcll} \text{group 1} \rightarrow & Y_{i,1} & = & \mu + \delta + \epsilon_{i,1}, & i = 1, \cdots, n_1 \\ \text{group 2} \rightarrow & Y_{i,2} & = & \mu - \delta + \epsilon_{i,2}, & i = 1, \cdots, n_2 \\ & \{\epsilon_{i,j}\} & \overset{iid}{\sim} & \text{normal}(0, \sigma^2). & \end{array} \Rightarrow \begin{array}{l} y_{i1} - \mu \mid \delta \sim N(\delta, \sigma^2) \\ \mu - y_{i2} \mid \delta \sim N(\delta, \sigma^2) \end{array}$$

We have utilized a (re)parameterization trick: under this parameterization, $\theta_1 = \mu + \delta$ and $\theta_2 = \mu - \delta$, so $\mu = (\theta_1 + \theta_2)/2$ and $\delta = (\theta_1 - \theta_2)/2$. The intention is to enable the coupling (dependence) of the two groups via variables μ and δ, which will be made random by a prior distribution. The fact that these two are random is enough to allow the coupling and subsequent information sharing in posterior inference. The specific prior choice given below is for computational convenience:

$$\begin{array}{rcll} p(\mu, \delta, \sigma^2) & = & p(\mu) \times p(\delta) \times p(\sigma^2) & \\ \mu & \sim & \text{normal}(\mu_0, \gamma_0^2) & (*) \\ \delta & \sim & \text{normal}(\delta_0, \tau_0^2) & \\ \sigma^2 & \sim & \text{inverse-gamma}(\nu_0/2, \nu_0\sigma_0^2/2). & \end{array}$$



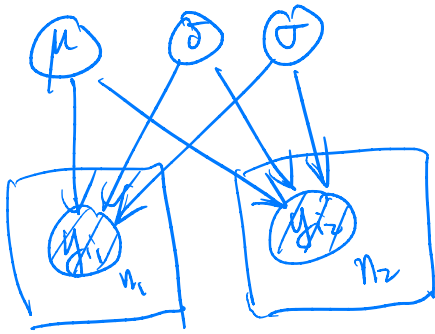
For any pair of θ_1, θ_2, there exist μ, δ s.t. $\theta_1 = \mu + \delta$, $\theta_2 = \mu - \delta$.

how? $\mu = \frac{\theta_1 + \theta_2}{2}$

$\delta = \frac{\theta_1 - \theta_2}{2}$

Sampling model: priors in $(*)$

$$\begin{array}{l} y_{i1} \mid \mu, \delta, \sigma \overset{iid}{\sim} N(\theta_1 = \mu + \delta, \sigma^2) \\ y_{i2} \mid \mu, \delta, \sigma \overset{iid}{\sim} N(\theta_2 = \mu - \delta, \sigma^2) \end{array}$$



- If μ, δ, σ are fixed (non-random)

$$\{y_{i,1}\}_{i=1}^{n_1} \perp\!\!\!\perp \{y_{j,2}\}_{j=1}^{n_2}$$

- Now if we allow μ, δ, σ to be random

then $\{y_{i,1}\}_{i=1}^{n_1} \not\perp\!\!\!\perp \{y_{j,2}\}_{j=1}^{n_2}$

posterior : $p(\mu, \delta, \sigma^2 \mid \{y_{i1}\}_{i=1}^{n_1}, \{y_{j2}\}_{j=1}^{n_2})$

Based on our previous calculations for the (univariate) normal model, it should be an easy exercise to derive the full conditional distributions of these parameters as follows

$$
\begin{aligned}
\{\mu|\boldsymbol{y}_1, \boldsymbol{y}_2, \delta, \sigma^2\} &\sim \text{normal}(\mu_n \gamma_n^2), \text{where} \\
\left\{ \begin{array}{rcl} \gamma_n^2 &=& [1/\gamma_0^2 + (n_1+n_2)/\sigma^2]^{-1}, \\ \mu_n &=& \gamma_n^2 \times [\mu_0/\gamma_0^2 + \sum_{i=1}^{n_1}(y_{i,1}-\delta)/\sigma^2 + \sum_{i=1}^{n_2}(y_{i,2}+\delta)/\sigma^2]; \end{array} \right. \\
\{\delta|\boldsymbol{y}_1, \boldsymbol{y}_2, \mu, \sigma^2\} &\sim \text{normal}(\delta_n, \tau_n^2), \text{where} \\
\tau_n^2 &= [1/\tau_0^2 + (n_1+n_2)/\sigma^2]^{-1}, \\
\delta_n &= \tau_n^2 \times [\delta_0/\tau_0^2 + \sum(y_{i,1}-\mu)/\sigma^2 - \sum(y_{i,2}-\mu)/\sigma^2]; \\
\{\sigma^2|\boldsymbol{y}_1, \boldsymbol{y}_2, \mu, \delta\} &\sim \text{inverse-gamma}(\nu_n/2, \nu_n\sigma_n^2/2), \text{where} \\
\nu_n &= \nu_0 + n_1 + n_2, \\
\nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + \sum(y_{i,1}-[\mu+\delta])^2 + \sum(y_{i,2}-[\mu-\delta])^2.
\end{aligned}
$$

$\frac{1}{\gamma_n^2} = \frac{1}{\gamma_0^2} + \frac{n_1+n_2}{\sigma^2}$

$\mu = \frac{\theta_1+\theta_2}{2}$

$\delta = \frac{\theta_1-\theta_2}{2}$

prior

data

prior

data

$y_{i1} - (\mu+\delta) \mid \mu, \delta \overset{iid}{\sim} N(0, \sigma^2)$

$y_{i2} - (\mu-\delta) \mid \mu, \delta \overset{iid}{\sim} N(0, \sigma^2)$

$\theta_1 = \mu+\delta$

$\theta_2 = \mu-\delta$

data-driven way to decide

how much information to share.

Let us go back to our example of comparing math test scores of students from two high schools.

Example 8.2. As for prior distribution parameter for $\mu \sim \text{normal}(\mu_0, \gamma_0^2)$, we put $\mu_0 = 50, \gamma_0 = 50/2 = 25$ to get a reasonably diffuse prior. For the prior on δ, set $\delta_0 = 0, \tau_0 = 25$. For the prior for σ^2, set $\nu_0 = 1, \sigma_0 = 10$ (this latter choice is due to the setup that the math scores were standardized to produce a nationalwide mean of 50 and a standard deviation of 10).

- the following figure shows the posterior distribution for μ and δ. In particular, the 95% quantile-based posterior confidence interval for 2δ, the difference of average scores between the two schools, is (-.61, 9.98), indicating a strong evidence that the posterior mean for school 1 is higher than that of school 2.
- In addition, $\Pr(\theta_1 > \theta_2 | \boldsymbol{y}_1, \boldsymbol{y}_2) = \Pr(\delta > 0 | \boldsymbol{y}_1, \boldsymbol{y}_2) \approx 0.96$, even though the prior probability is such that $\Pr(\delta > 0) = .50$.
- As for posterior predictive probability that a randomly selected student from school 1 has a higher score than a randomly selected student from school 2:

$$\Pr(Y_1 > Y_2 | \boldsymbol{y}_1, \boldsymbol{y}_2) \approx 0.62.$$

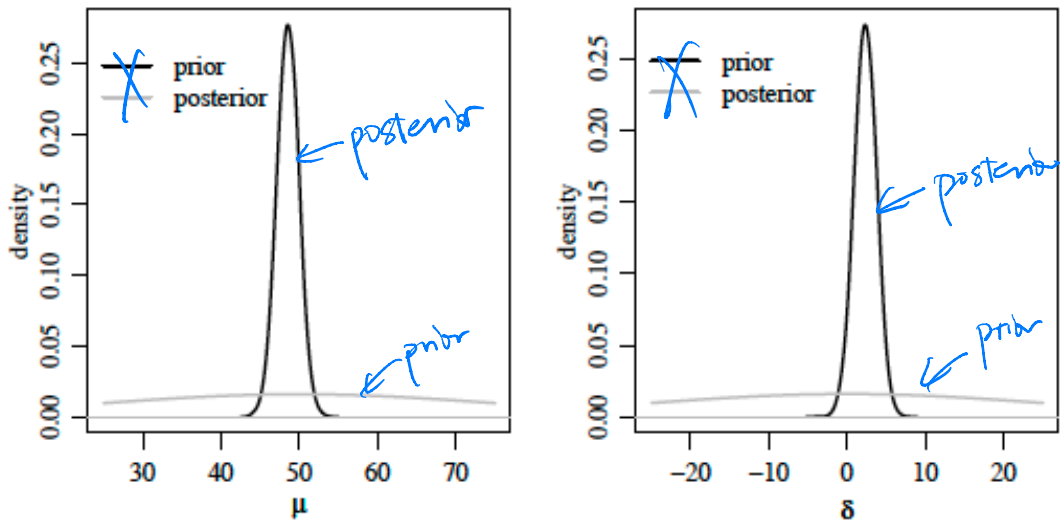


Figure 8.2: Posterior distributions for μ and δ.

8.2 Comparing multiple groups

It is very common to organize data or data sets in a hierarchy of nested populations. Such data sets are often called hierarchical or multilevel data. For example

- there are multiple hospitals, each hospital has many patients
- there are different animals, each animal carry a set of genes
- different countries, each of which is organized into regions, each of which is organized into counties, with residents in each of them
- ”activity recognition problem”: a collection of computer users, each user is associated with a collection of computer related activities (organized by days), each day has a collection of activities (apps run)
- a collection of text corpora, each text corpus is a collection of documents, each document is a collection of words
- a database of images divided by groups, each image is a collection of image patches, each patch a collection of pixels or other specific computer vision elements

multiple groups

multiple levels of groups

We are interested in learning about these groups: what are the shared features among them, what make different groups different and how. In most applications, it does not make great sense to assume that the groups are independent. It makes sense to assume that they are dependent, and to exploit such dependence to learn about global aspects of all groups, as well as locally distinct aspects of each group. In other words, we wish to borrow information from one group to inform about the others, as well as the whole. The question is how.

8.3 Exchangeability and hierarchical models

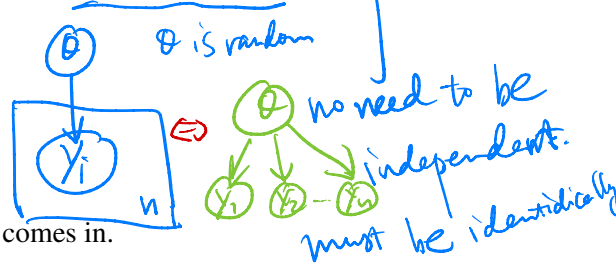
Hierarchical models are a general method for describing dependence for grouped data. They can be motivated by a theorem of Bruno de Finetti. At a high level, de Finetti's theorem says that a collection of exchangeable sequence of random variables must be conditionally i.i.d., and as a consequence, an exchangeable collection of groups of random variables must be distributed according to a hierarchical model. Let us make this statement more precise.

Definition 8.1. (Exchangeable). Let $p(y_1, \ldots, y_n)$ be the joint density of random variables $Y_1, \ldots, Y_n$. If $p(y_1, \ldots, y_n) = p(y_{\pi_1}, \ldots, y_{\pi_n})$ for all permutation π of $1, \ldots, n$.[6] Equivalently, the joint distribution of $(Y_{\pi_1}, \ldots, Y_{\pi_n})$ remains invariant under any permutation π. Then, we say that $Y_1, \ldots, Y_n$ are exchangeable.

Intuitively, when $Y_1, \ldots, Y_n$ are exchangeable, then the subscript labels of these n variables convey no additional information about them.

It is simple to see that if a collection of random variables $Y_1, \ldots, Y_n$ are conditionally i.i.d. given some random variable θ, i.e.,

$$\begin{cases} \theta & \sim & \pi(\theta) \\ Y_1, \ldots, Y_n | \theta & \overset{i.i.d.}{\sim} & p(\cdot|\theta), \end{cases}$$



then $Y_1, \ldots, Y_n$ are exchangeable.

What about the other direction? This is where de Finetti's theorem comes in.

[6] At this point, it may be helpful to express the identity explicitly: $p_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = p_{Y_1,\ldots,Y_n}(y_{\pi_1}, \ldots, y_{\pi_n})$.

Theorem 8.1. *Let $Y_1, Y_2, \ldots$ be an infinite sequence of random variables all having a common sample space $\mathcal{Y}$. Suppose that $Y_1, \ldots, Y_n$ are exchangeable for any sequence size n. Then $Y_1, Y_2, \ldots$ must be conditionally i.i.d. That is, the joint distribution of $Y_1, \ldots, Y_n$ for any n must be of the form (provided that a density function exists): for all n and $y_1, \ldots, y_n$*

$$p(y_1, \ldots, y_n) = \int \left\{ \prod_{i=1}^{n} p(y_i|\theta) \right\} \pi(\theta) d\theta \tag{32}$$

for some parameter θ, some distribution π over θ, and some sampling model $p(y|\theta)$.

Remark 8.1.

- The "infinite" part in the statement is necessary, along with the condition of exchangeability for any n.
- de Finetti's theorem is one of the great theorems in probability theory. It also gives us probability models that can be written as Eq. (32), as well as hierarchical versions of this, as we will see.
- It has a foundational role in Bayesian statistics, because it provides a mathematical justification for the existence of the notion of *random* parameter θ:
  - whereas a frequentist statistician may be content with making an i.i.d. assumption about an unknown sampling mechanism such as

    $$Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} p(\cdot|\theta),$$

    de Finetti's theorem says that if the observation sequence is in fact exchangeable, then the unknown θ must be *random*. Bayesian statisticians proceed by placing a prior distribution π on such θ.

- Exchangeability makes sense in many practical situations:
  - the math scores from n randomly selected students from a particular school, in absence of other information about the students, may be treated as exchangeable.
  - the collection of U.S. high schools in similar environments (e.g., large urban areas).
  - The computer-related activities by a user collected on Monday mornings in the past year.
  - What are not exchangeable? The collection of time-stamped computer-related activities in the past 24 hours, is not exchangeable.

    The words in a document, read from the beginning to the end, are not exchangeable, either. But if we print out the document into a piece of paper, and cut the paper into small pieces, one for each word, which are then placed into a bag and shuffled well. Then we have a bag of exchangeable words. ~ bag-of-words representation of text.

Now, let us consider a model to describe our information about a hierarchical data structure: there are m groups $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m\}$; each group $\boldsymbol{Y}_j = \{Y_{j1}, \ldots, Y_{jn_j}\}$ has n_j elements, for some $n_j \geq 1$.

Suppose that the elements within each group $\boldsymbol{Y}_j$ may be treated as exchangeable. Then, by de Finetti's theorem we may model the observations from each group as conditionally i.i.d. given some parameter:

$$Y_{j1}, \ldots, Y_{jn_j} | \phi_j \overset{i.i.d.}{\sim} p(y|\phi_j). \tag{33}$$

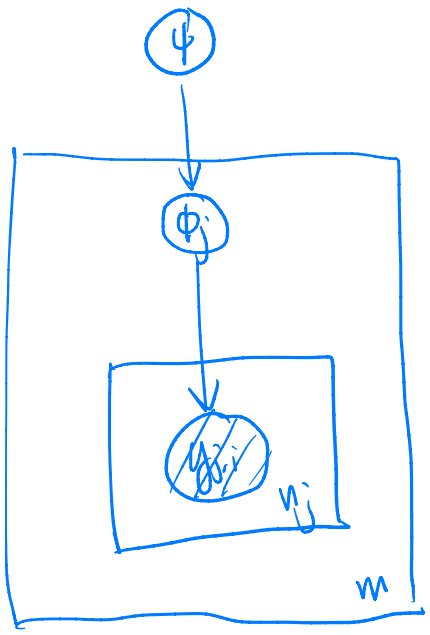
What about the collection of parameters $\phi_1, \ldots, \phi_m$? If we assume that the m groups are exchangeable, then, applying de Finetti's theorem once more, we have

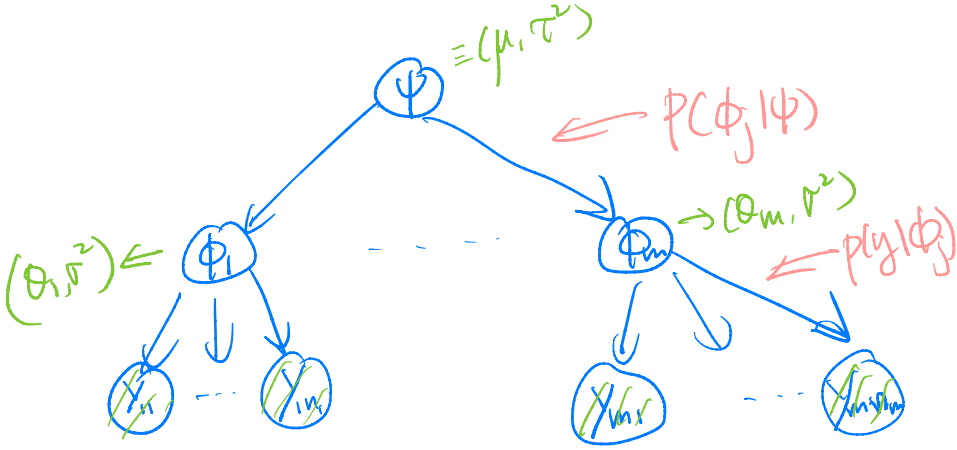
$$\phi_1, \ldots, \phi_m | \psi \overset{i.i.d.}{\sim} p(\phi|\psi), \tag{34}$$

for some random parameter ψ. Collecting the above specifications, we arrive at the following hierarchical model

$$\begin{aligned}
\psi &\sim p(\psi) \quad \text{(prior distribution)} \\
\phi_1, \ldots, \phi_m | \psi &\overset{i.i.d.}{\sim} p(\phi|\psi) \quad \text{(between-group sampling variability)} \\
Y_{j1}, \ldots, Y_{jn_j} | \phi_j &\overset{i.i.d.}{\sim} p(y|\phi_j), j = 1, \ldots, m \quad \text{(within-group sampling variability).}
\end{aligned}$$

This hierarchical model has three levels that representing different aspects of randomness/ random variability: $p(y|\phi)$ represents the sampling variability among measurements within a group, and $p(\phi|\psi)$ represents the sampling variability across groups. Finally, $p(\psi)$ represents prior information about unknown parameter ψ. Depending on data structure and the modeler's knowledge, there may be more levels in the hierarchy of sampling distributions and prior distributions that can be constructed.





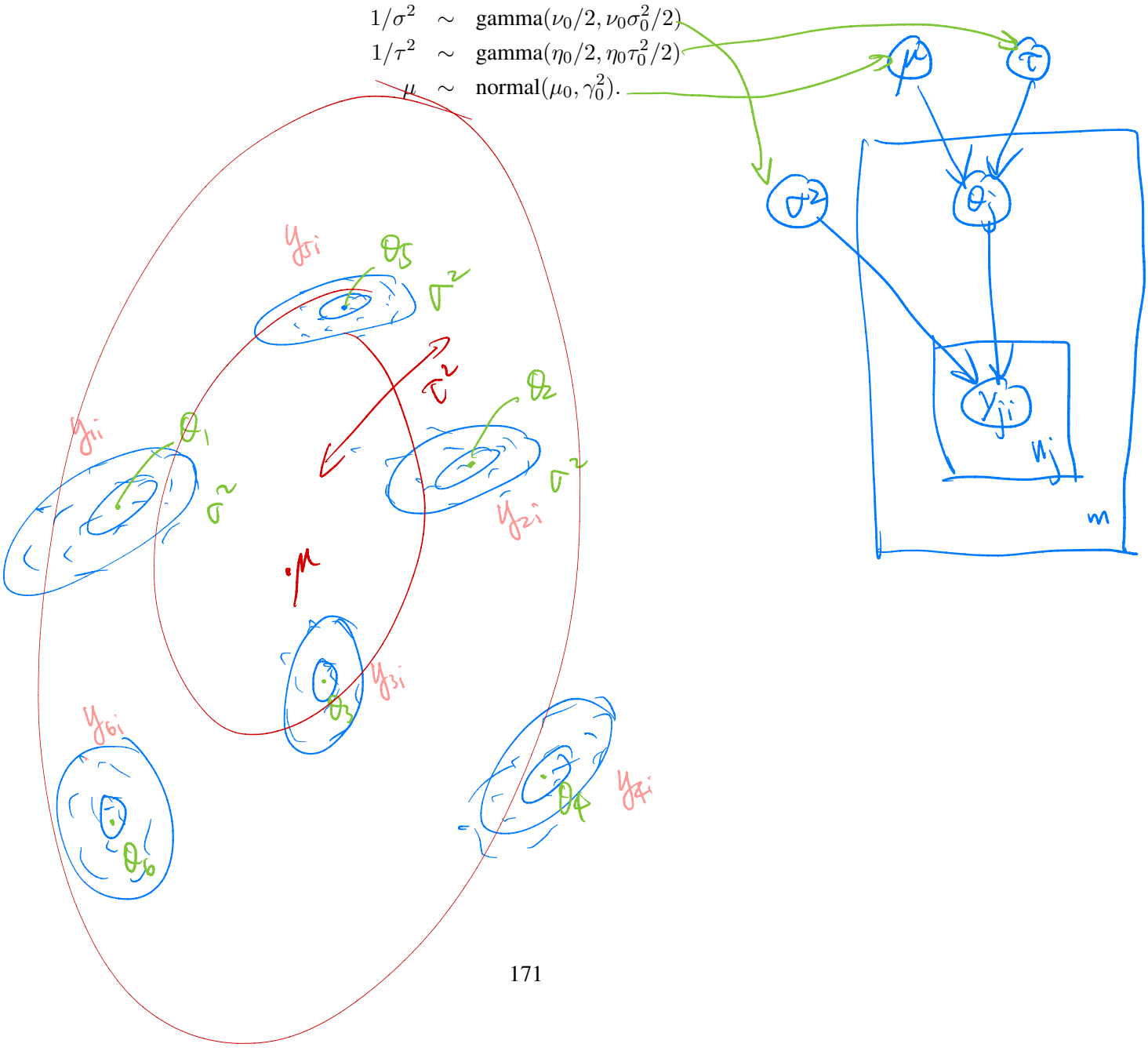
8.4 Hierarchical normal models

A popular model for describing the heterogeneity of means across several populations is the hierarchical normal models: here, each group is endowed with a normal sampling model; the mean parameters across groups are endowed with another normal sampling model further up in the hierarchy.

$$\begin{aligned} \phi_j = (\theta_j, \sigma^2),\ p(y|\phi_j) &= \text{normal}(\theta_j, \sigma^2) \quad \text{(within-group model)} && \text{(35a)} \\ \psi = (\mu, \tau^2),\ p(\theta_j|\psi) &= \text{normal}(\mu, \tau^2) \quad \text{(between-group model).} && \text{(35b)} \end{aligned}$$

Note that in this model, we allow different groups to have different means, but they share the same variance σ^2 (this assumption may be relaxed). The parameters for the given sampling model are μ, τ^2, σ^2. For convenience we may give them standard semi-conjugate priors:

$$\begin{aligned} 1/\sigma^2 &\sim \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2) \\ 1/\tau^2 &\sim \text{gamma}(\eta_0/2, \eta_0\tau_0^2/2) \\ \mu &\sim \text{normal}(\mu_0, \gamma_0^2). \end{aligned}$$



8.4.1 Posterior inference

The unknown quantities in our model include the group-specific means $(\theta_1, \ldots, \theta_m)$, within-group sampling variability σ^2, the mean and variance μ, τ^2 of the population of group-specific means. Joint posterior inference for these parameters may be made by an MCMC approximation for the posterior distribution

Bayes rule

$$\begin{aligned} & p(\theta_1, \ldots, \theta_m, \sigma^2, \mu, \tau^2 | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m) \\ \propto \; & p(\mu, \tau^2, \sigma^2) \times p(\theta_1, \ldots, \theta_m | \mu, \tau^2, \sigma^2) \times p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | \theta_1, \ldots, \theta_m, \mu, \tau^2, \sigma^2) \\ = \; & p(\mu)p(\tau^2)p(\sigma^2) \Big\{ \prod_{j=1}^{m} p(\theta_j | \mu, \tau^2) \Big\} \Big\{ \prod_{j=1}^{m} \prod_{i=1}^{n_j} p(y_{ji} | \theta_j, \sigma^2) \Big\}. \end{aligned}$$

by assumption

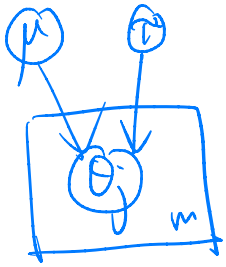
Although this may look daunting, we will see shortly that it is not difficult to derive full conditional distributions for all parameters of interest, which will enable us to run a Gibbs sampler. The key is to observe that the joint distribution of all parameters and observed is expressed in factorized form (i.e., product form) given above. This is a reflection of the conditional independence relations inherent in our hierarchical modeling assumption. It is also the conditional independence that we exploits in deriving the full conditional distributions comfortably.

Full conditional distributions of μ and τ^2:

It is useful to note that μ and τ^2 are conditionally independent of all other variables in the joint model when given $\theta_1, \ldots, \theta_m$. Collecting only relevant terms from the joint distribution, we find that

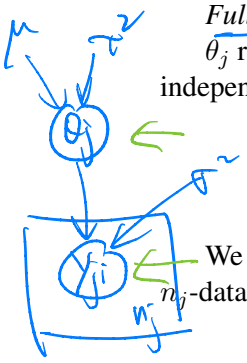
$$\begin{aligned} p(\mu | \theta_1, \ldots, \theta_m, \tau^2, \sigma^2, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m) &\propto p(\mu) \prod p(\theta_j | \mu, \tau^2) \\ p(\tau^2 | \theta_1, \ldots, \theta_m, \mu, \sigma^2, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m) &\propto p(\tau^2) \prod p(\theta_j | \mu, \tau^2). \end{aligned}$$

The right hand side of the two equations in the above display allow us to look at only "submodels" for μ and τ^2. For example: in the first equation we can treat θ_j as the m-data sample for normal submodel with mean parameter μ, so we need to compute the posterior distribution of μ for this submodel. We have seen such submodels before, in Section 5. Thus,



$$\begin{aligned} \mu | \theta_1, \ldots, \theta_m, \tau^2 &\sim \text{normal}\left(\frac{m\bar{\theta}/\tau^2 + \mu_0/\gamma_0^2}{m/\tau^2 + 1/\gamma_0^2}, (m/\tau^2 + 1/\gamma_0^2)^{-1} \right), \\ 1/\tau^2 | \theta_1, \ldots, \theta_m, \mu &\sim \text{gamma}((\eta_0 + m)/2, \eta_0\tau_0^2/2 + \textstyle\sum (\theta_j - \mu)^2/2). \end{aligned}$$

same as normal conjugacy



Full conditional distribution of $\theta_j, j = 1, \ldots, m$:

θ_j represents the mean for group j. It is useful to note that, given $\mu, \tau^2, \sigma^2, \boldsymbol{y}_j$, θ_j must be conditionally independent of all other mean parameters θ's, as well as the data from groups other than j. In fact,

$$p(\theta_j|\mu, \tau^2, \sigma^2, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m) \propto p(\theta_j|\mu, \tau^2) \prod_{i=1}^{n_j} p(y_{ji}|\theta_j, \sigma^2).$$

We can view this as the posterior distribution for the normal sampling model for group j, given the n_j-data sample from this group only. Let $\bar{y}_j$ denote the sample mean for group j, then

$$\theta_j|\sigma^2, y_{j1}, \ldots, y_{jn_j} \sim \text{normal}\left(\frac{n_j\bar{y}_j/\sigma^2 + \mu/\tau^2}{n_j/\sigma^2 + 1/\tau^2}, (n_j/\sigma^2 + 1/\tau^2)^{-1}\right). \tag{36}$$

Full conditional distribution of σ^2:

σ^2 represents the shared within-group variance for all groups. Note that σ^2 is conditionally independent of μ, τ^2 given $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m, \theta_1, \ldots, \theta_m$. We find that

$$\begin{aligned} p(\sigma^2|\theta_1, \ldots, \theta_m, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m) &\propto p(\sigma^2) \prod_{j=1}^{m}\prod_{i=1}^{n_j} p(y_{ji}|\theta_j, \sigma^2) \\ &\propto (\sigma^2)^{-\nu_0/2+1} e^{-\frac{\nu_0\sigma_0^2}{2\sigma^2}} (\sigma^2)^{-\sum n_j/2} \exp -\frac{1}{2\sigma^2}\sum_j\sum_i (y_{ji} - \theta_j)^2, \end{aligned}$$

so

$$1/\sigma^2|\boldsymbol{\theta}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m \sim \text{gamma}((\nu_0 + \sum_{j=1}^{m} n_j)/2, \nu_0\sigma_0^2/2 + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ji} - \theta_j)^2/2).$$

Note that the double sum term is the sum of squared residuals across all groups, conditional on the within-group means, so the (full) conditional distribution of σ^2 concentrates probability around a pooled-sample estimate of the variance. This makes sense, because σ^2 is the same variance parameter shared across all groups according to our model.

8.4.2 Example: Math scores in U.S. public schools

We return to the analysis of math scores examined in Hoff (2009). The setting is as follows

- there are 100 large urban public high schools, all having a 10th grade enrollment of 400 or larger.

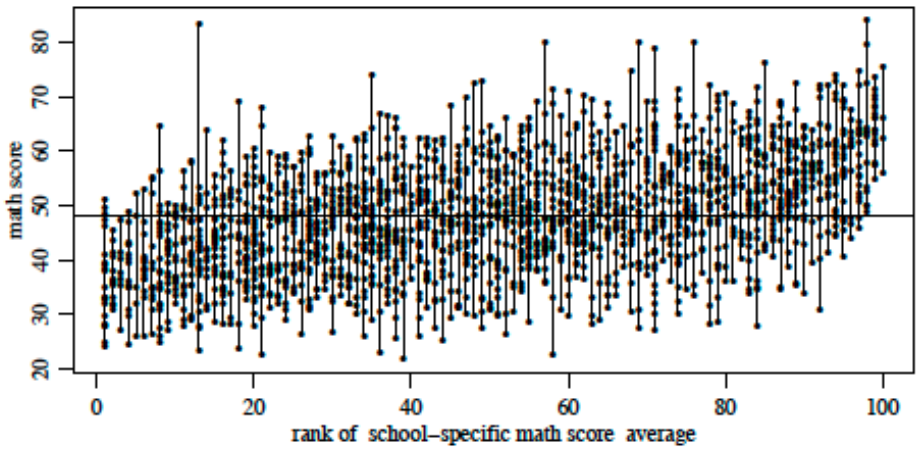


Figure 8.3: ELS data.

- average score per school ranges from 36.6 to 65.0.

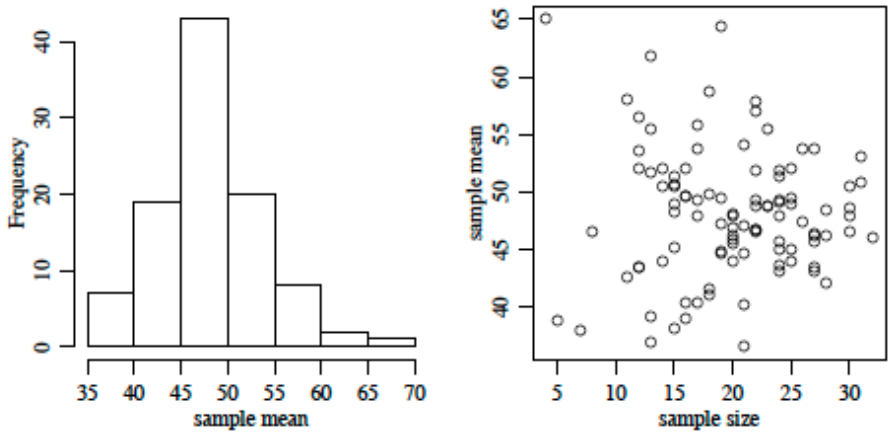


Figure 8.4: Empirical distribution of sample means and relationship with sample size.

- extreme average scores tend to be associated with low sample sizes. This is a common phenomenon for hierarchical data sets (how?)

Prior specification and posterior approximation

- recall our hierarchical model

$$\begin{aligned}
\mu, \tau^2, \sigma^2 &\sim p(\psi)p(\tau^2)p(\sigma^2) \quad \text{(prior distribution)} \\
\theta_1, \ldots, \theta_m | \psi &\overset{i.i.d.}{\sim} \text{normal}(\mu, \tau^2) \quad \text{(between-group sampling variability)} \\
Y_{j1}, \ldots, Y_{jn_j} | \theta_j &\overset{i.i.d.}{\sim} \text{normal}(\theta_j, \sigma^2), j = 1, \ldots, m \quad \text{(within-group sampling variability).}
\end{aligned}$$

- we need to provide hyperparameters for the semi-conjugate priors

$$\begin{aligned}
1/\sigma^2 &\sim \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2) \\
1/\tau^2 &\sim \text{gamma}(\eta_0/2, \eta_0\tau_0^2/2) \\
\mu &\sim \text{normal}(\mu_0, \gamma_0^2).
\end{aligned}$$

  - the math exam was designed to give a nationwide variance of 100, so we set $\sigma_0^2 = 100$. For a diffuse prior for the variance, we set $\nu_0 = 1$.
  - for between-group variance: we set $\tau_0^2 = 100$, and $\eta_0 = 1$.
  - for the global mean: we set $\mu_0 = 50, \gamma^2 = 25$ (so the prior probability that μ is in $(\mu_0 - 2\gamma, \mu_0 + 2\gamma) = (40, 60)$ is about 95%).

- the previous subsection gave the derivations of all full conditional distributions required for the implementation of a Gibbs sampler.

MCMC diagnostic

- run the Gibbs sampler for 5000 iterations. Fig. 8.5 shows the boxplots for batch of 500 consecutive MCMC samples (e.g., $\{1, \ldots, 500\}$, $\{510, \ldots, 1000\}$, and so on). There does not seem to be any evidence that the chain has not achieved stationarity.

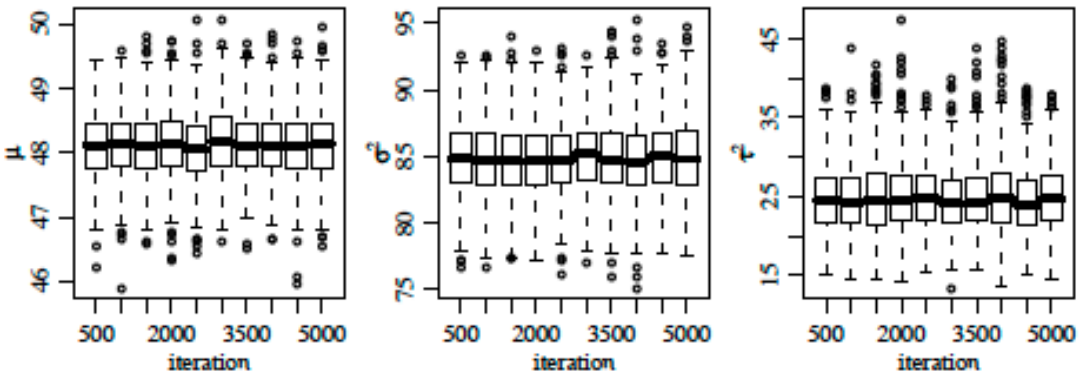


Figure 8.5: Stationarity plots of the MCMC samples of μ, σ^2, τ^2.

- lag-1 autocorrelations for the sequences of μ, σ^2 and τ^2 are 0.15, 0.053, and 0.312, respectively.
- the effective sample sizes are 3706, 4499, and 2503, respectively.
- the approximate MC std can be obtained by dividing the approximated posterior std by the square root of the effective sample sizes, resulting in 0.009, 0.004, 0.09 for μ, σ^2, τ^2, resp. These are small compared to the posterior means of these parameters (Fig. 8.6).

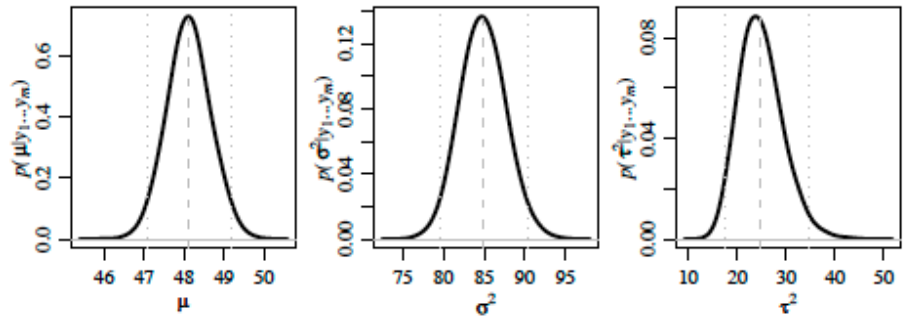


Figure 8.6: Marginal posteriors with 2.5%, 50% and 97.5% quantiles.

- for θ: we found the ESS for the 100 sequences of θ-values ranged between 3,500 and 6,000, with the MC std ranging between 0.02 and 0.05.

Posterior summaries and shrinkage

- The posterior means of μ, σ and τ are $48.12, 9.21$ and 4.97, respectively. Recalling the meaning of these parameters; this indicates that roughly 95% of scores *within* a class room are within $4 \times 9.21 \approx 37$ points of each other, whereas 95% of the average classroom scores (across schools) are within $4 \times 4.97 \approx 20$ points of each other.

- The shrinkage effect: recall that, conditional on μ, τ^2, σ^2, the expected value of θ_j is a weighted average of $\bar{y}_j$ and μ (cf. Eq. (36)):

$$\mathbb{E}[\theta_j | \boldsymbol{y}_j, \mu, \tau, \sigma] = \frac{n_j \bar{y}_j / \sigma^2 + \mu / \tau^2}{n_j / \sigma^2 + 1 / \tau^2}.$$

  As a result, the expected value of θ_j is from the sample mean $\bar{y}_j$ toward the global mean μ. This is called the *shrinkage effect*: the parameter estimates are "shrinked" toward the global mean. How strong this effect is dependent partially on the sample size n_j.

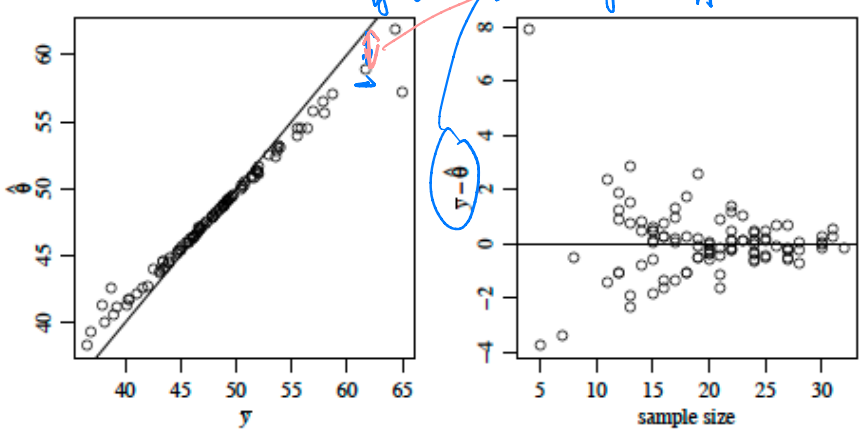


Figure 8.7: Shrinkage as a function of sample size.

- Fig. 8.7 illustrates the amount of shrinkage for different groups. Left panel shows that the groups with large sample means are "pulled down" a bit, while the groups with low sample means are "pushed up". The right panel shows that groups with small sample size receives the largest amount of shrinkage $|\bar{y}_j - \hat{\theta}|$.
  - for this reason we say that hierarchical modeling faciliates the "borrowing of strength": in particular, the groups with small sample size borrow information from the groups with large sample size. In theory, it has been shown that the borrowing of strength (also, sharing of information) results in more robust and efficient inference.

Back to the question of ranking

- We may rank all schools according to the posterior expectations

$$\{\mathbb{E}[\theta_1|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m],\ldots,\mathbb{E}[\theta_m|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m]\}$$

  Alternatively, one may simply rank all schools according to the sample means $\bar{y}_1,\ldots,\bar{y}_m$

- Although these two rankings would be quite similar, there are differences.
- Let's consider two schools: school 46 and school 82; these two schools are at the bottom 10% of the 100 schools in the data set. The sample means are

$$\bar{y}_{46} = 40.18 > \bar{y}_{82} = 38.76$$

  However, in terms of posterior expectation, the ranking would be different:

$$\mathbb{E}[\theta_{46}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m] = 41.31 < \mathbb{E}[\theta_{82}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m] = 42.53.$$

- We observe the effects of shrinkage: $n_{46} = 21$, while $n_{82} = 5$. School 82 receives a larger amount of shrinkage toward global mean ($\mathbb{E}[\mu|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m] = 48.11$) than that of school 46, resulting in a "reversal" in the ranking.

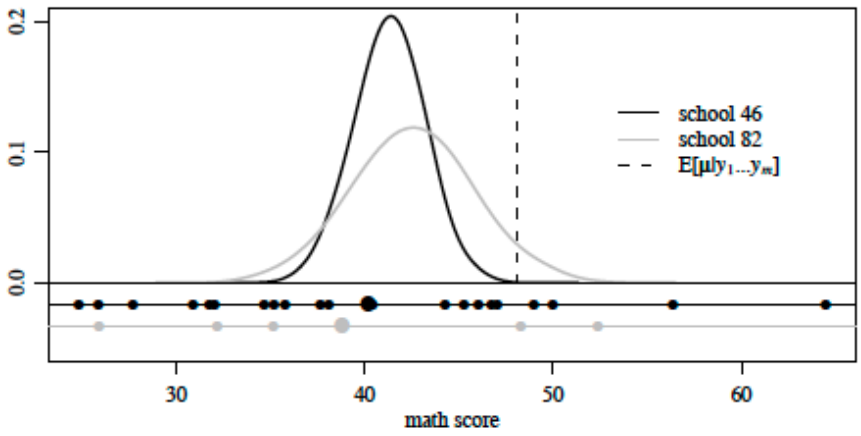


Figure 8.8: Data and posterior distributions for two schools

- Does this make sense?
  - there are more uncertainty about school 82's average scores due to its low sample size.
  - suppose on the day of the exam, the student who got the lowest exam score from school 82 doesn't show up, then the sample mean would have been 41.99, a change of more than three points from 38.76. In the case of school 46, the sample mean would have been 40.9, a change of only three quarters of a point. So, while we are more certain about the average score of school 46, we are less certain about that of school 82, which results in a larger amount of shrinkage toward the global mean.

- to some, this ranking may seem unfair. However, it reflects an objective fact that there is more evidence that θ_{46} is exceptionally low than there is for θ_{82}.
- An example in sport: on any basketball team, there are "bench" players who play very little play time, many of whom have taken only a few free throws in their entire career, resulting in very high free throw shooting percentage, e.g., 100%. Yet, the coach when given an opportunity for a free throw (during a technical foul) will likely choose a veteran player, despite having a lower shooting percentage, say 87%. This is because coaches recognize that the bench player's true free throw percentage is nowhere near the "sample mean" 100%.

implicitly shrink toward global mean

8.5 Topic models

We will study hierarchical models for discrete data, such as texts, images and biological data. The class of models that we consider is known as topic models [7] and finite admixtures. [8] The paper by Blei and co-authors was motivated from the information retrieval/machine learning of texts and images. It also develops variational inference for this particular class of model. The paper by Pritchard and co-authors was motivated by population genetics applications and makes use of Gibbs sampling for the posterior inference. Both are extremely well-known (and combine for more than 60,000 citations on Google Scholar).

8.5.1 Model formulation

First come some notations.

- Random variable $W \in \{1, ..., V\}$ represents words in a vocabulary, where V is the length of the vocabulary.
- A *document* is a collection of words denoted by $\overline{W} = (W_1, ..., W_N)$.

  Although we write $\overline{W}$ as if it is a sequence, the ordering of the words does not matter in the modeling that we introduce here.
- A *corpus* is a collection of documents $(\overline{W}_1, ..., \overline{W}_m)$. For each document m, let N_m be the document length.

Topic model is essentially a hierarchical model for discrete data that can be viewed as a hierarchical mixture model (for discrete random variables). Each mixing component of the model will be referred to as a topic. Thus a topic is a particular distribution over words, and a document can be described as a mixture of topics.

[7] D. Blei, A. Ng and M. I. Jordan. Latent Dirichlet allocation, Journal of Machine Learning Research, 3:993–1022, 2003.

[8] J. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multi-locus genotype data. Genetics, 155:945–959, 2000.

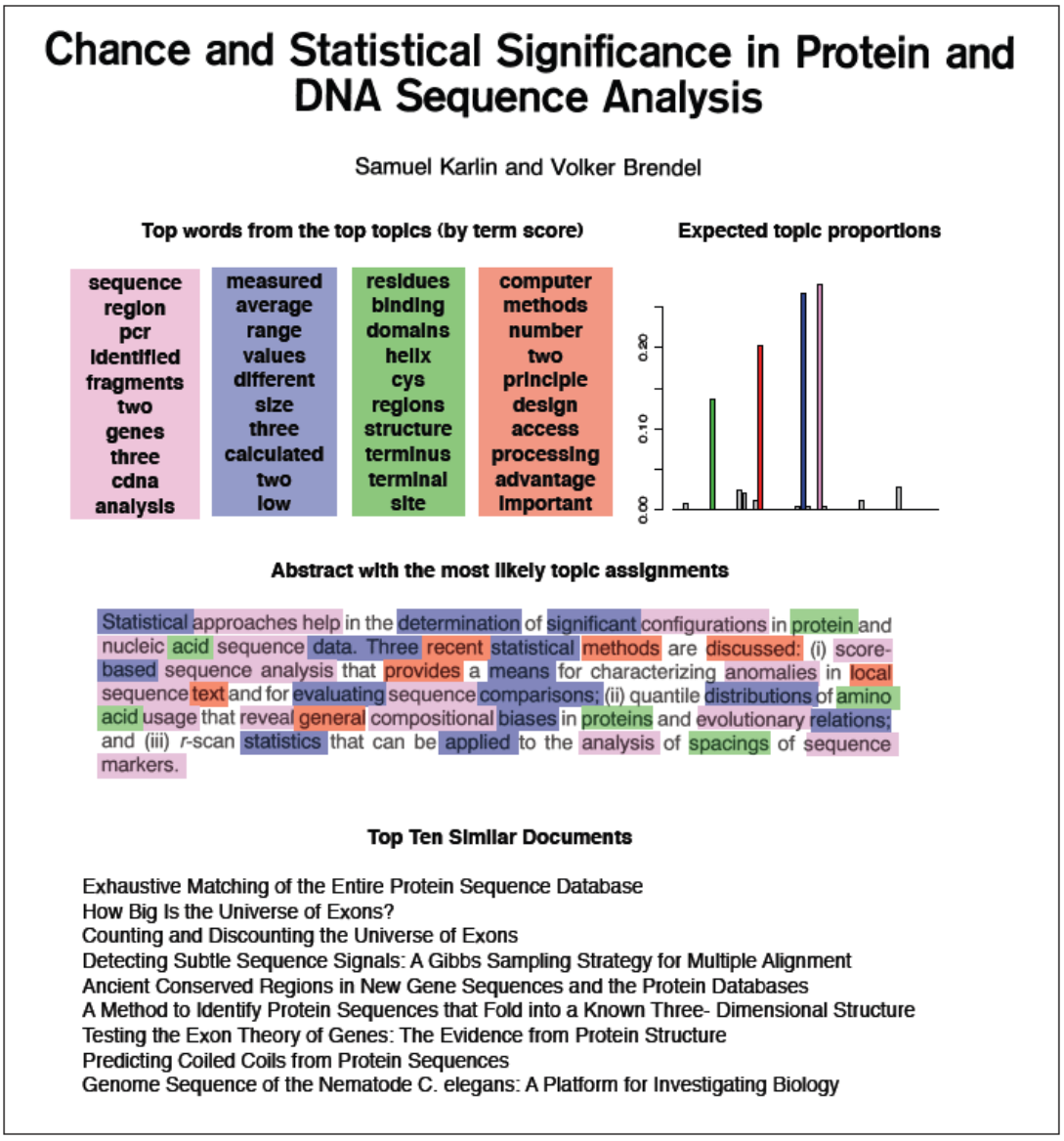
An example document from the AP corpus (Blei, Ng, Jordan, 2003)

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

After feeding such documents to Latent Dirichlet Allocation (LDA) model:

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Another example document from *Science* corpus (1880–2002) (Blei & Lafferty, 2009)



Topic models – such as Latent Dirichlet allocation and its variants – are a popular tool for modeling and mining patterns from texts in news articles, scientific papers, blogs, but also tweets, query logs, digital books, metadata records...

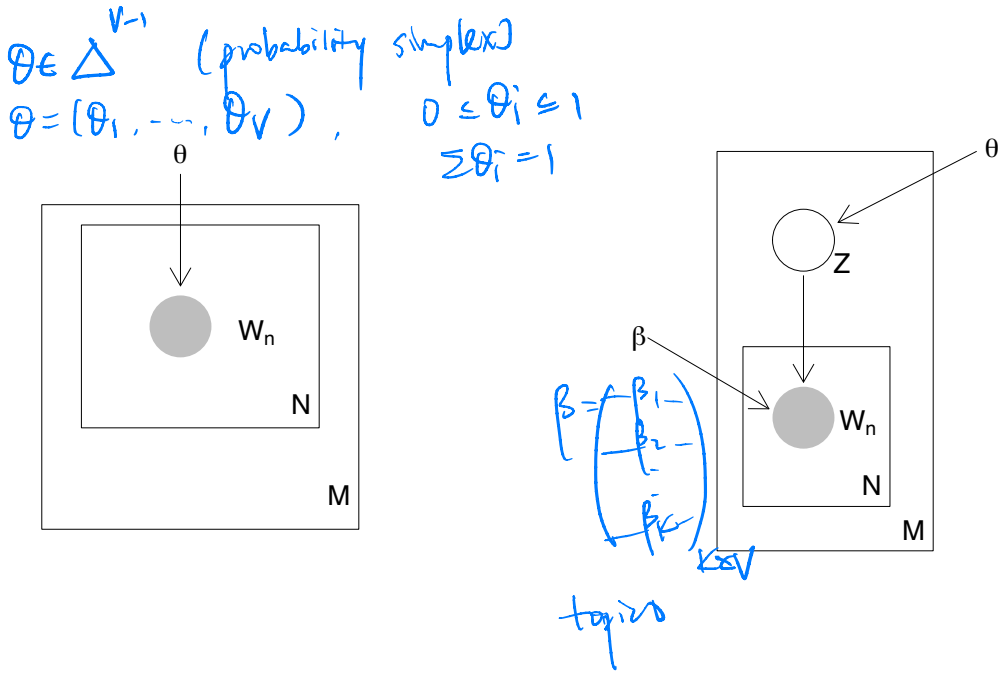


Figure 8.9: Graphical representation of the unigram (left) and mixture of unigrams (right).

Before we describe the latent Dirichlet allocation model, let us start with simpler precursors.

Unigram model For any document $\overline{W}$, assume

$$\overline{W} = (W_1, ..., W_N)|\theta \overset{i.i.d.}{\sim} \text{Cat}(\theta).$$

In other words, θ is the (same) word frequency that characterizes each document in the corpus. Thus, the corpus generated this way is implicitly assumed to have only one topic. See Fig. 8.9.

Mixture of unigrams Each document $\overline{W}_d$ is associated with a latent topic variable Z_d. Suppose that there are K topics, where K is given. Assume that

$$Z_d|\theta \sim \text{Cat}(\theta).$$

Now given Z_d, we assume

$$\overline{W}_d|Z_d = k \sim \text{Cat}(\beta_k),$$

where parameter $\beta_k \in \Delta^{V-1}$ is the frequency vector associated with topic k.

This is nothing but a mixture of discrete distributions. The parameters of interest are $\{\beta_k\}_{k=1}^K$ and θ.

In both models, the documents are assumed to be an i.i.d. sample from fairly simple distributions on vocabulary of words. Both above models were utilized in the early days of "natural language processing" (NLP), a field in artificial intelligence that focuses on analysis of texts.

Latent Dirichlet Allocation (LDA) LDA is an instance of hierarchical modeling. It was in fact motivated from de Finetti's theorem. Given the hierarchical view of the text corpus, we assume that the documents are exchangeable. Moreover within each document the words are assumed to be exchangeable.

Exchangeability assumption can be questioned, as we discussed in a previous subsection. However, this is an important step-up from the previous i.i.d. assumption. Moreover, exchangeability is not an unreasonable assumption if we do not want to capture aspects of the data the violates exchangeability (such as the ordering of words, or documents).

From de Finetti's theorem, we expect a hierarchical model specification for the words and then for the documents. Originally the LDA is described as a generative process: To generate document $\overline{W}$, one proceeds as follows

- For a document, generate N of $\overline{W}$ from a Poisson distribution: $N \sim \text{Poisson}(\lambda)$.
- For some parameter $\alpha_1, \ldots, \alpha_K > 0$, let θ represent "topic proportion" for document $\overline{W}$:

$$\theta|\alpha \sim \text{Dir}(\alpha_1, ..., \alpha_K).$$

- Given N and θ associated with the document, for each word index $n = 1, \ldots, N$,

$$\begin{aligned} Z_n|\theta &\overset{i.i.d.}{\sim} \text{Cat}(\theta), \\ W_n|Z_n = k, \boldsymbol{\beta} &\overset{iid}{\sim} \text{Cat}(\beta_k) \end{aligned}$$

In the above, we use β_k to denote row vector k of $K \times V$ matrix $\boldsymbol{\beta}$. In particular β_k represents the distribution over the vocabulary for topic k. This means $\Pr(W_n = j|Z_n = k, \boldsymbol{\beta}) = \beta_{kj}$.

A graphical representation of this model is given Fig. 8.10

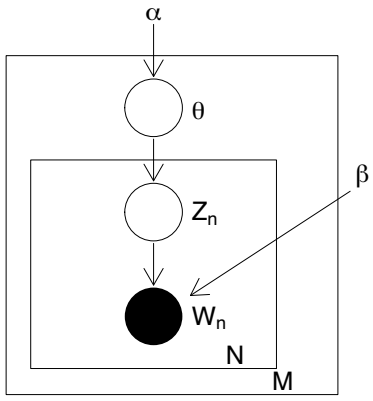
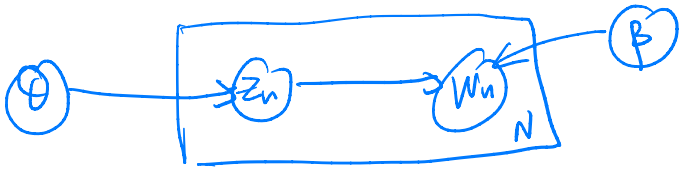


Figure 8.10: Latent Dirichlet Allocation Model.



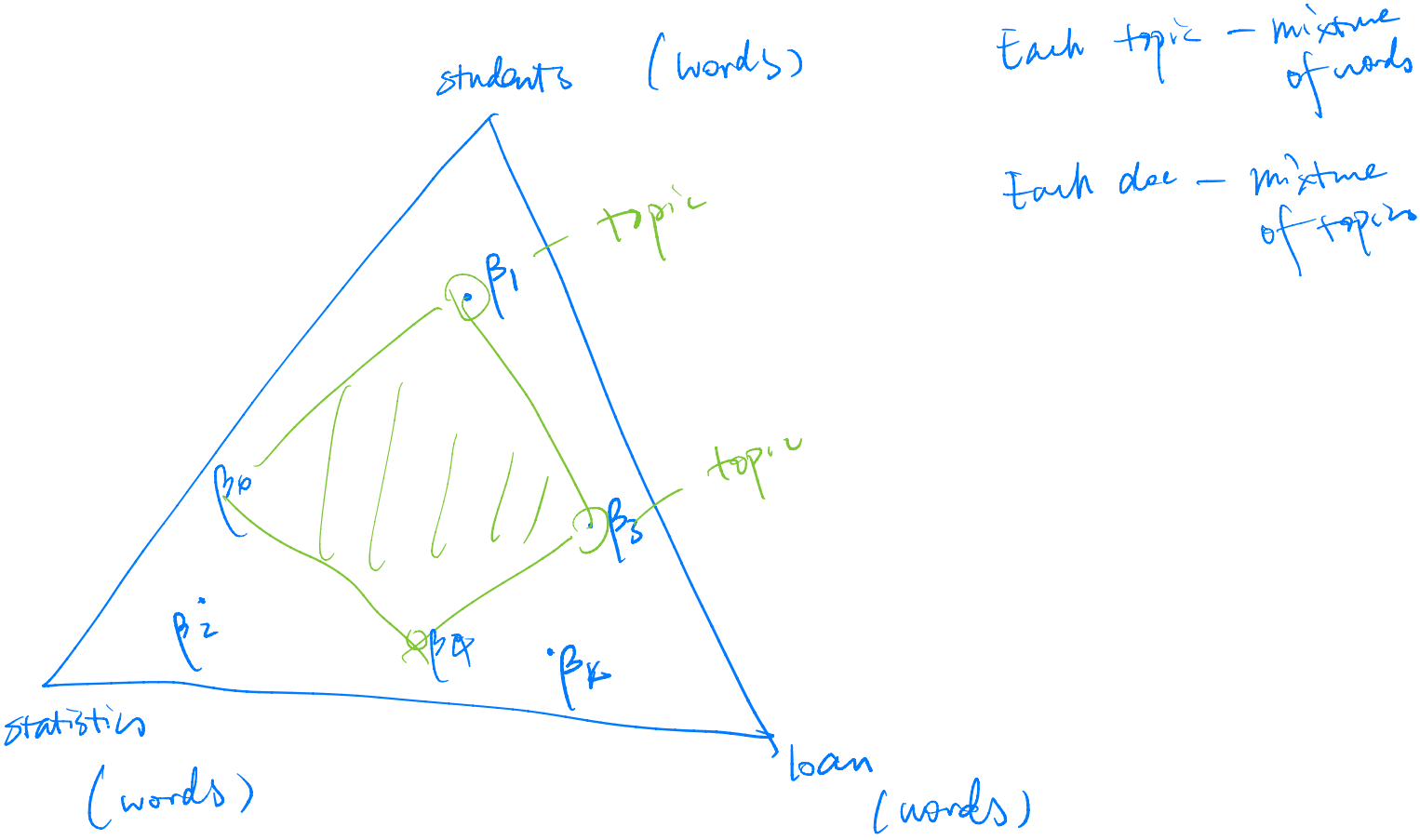
There is a simpler geometric reformulation for the LDA. It goes like this.[9]

Each document $\overline{W} = (W_1, \ldots, W_N)$ consists of words that are generated i.i.d. according to the following probability

$$\Pr(W_n = j|\theta, \boldsymbol{\beta}) = \sum_{k=1}^{K} \Pr(Z_n = k|\theta) \times \Pr(W_n = j|Z_n = k, \boldsymbol{\beta}) = \sum_{k=1}^{K} \theta_k \beta_{kj}.$$

That is, the vector of word frequency for document $\overline{W}$ is $\sum_{k=1}^{K} \theta_k \beta_k \in \Delta^{V-1}$. This is a point that lies in the convex hull $G = \text{conv}(\beta_1, \ldots, \beta_K)$.

Each extreme point $\beta_1, \ldots, \beta_K$ corresponds to a word frequency vector of a topic (e.g., "education", "politics", "sports"). Given the convex hull G, a document corresponds to a point randomly drawn from the convex hull G. The randomness is due to the random weight vector $\theta \in \Delta^{K-1}$, which is distributed by a Dirichlet distribution.



[9] J. Tang, Z. Meng, X. Nguyen, Q. Mei and M. Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. Proceedings of the 31st International Conference on Machine Learning (ICML), 2014.

8.5.2 Posterior inference

We have seen how the LDA is composed of familiar building blocks, Poisson for the document length, multinomial/categorical distributions for topic-specific distribution over words, Dirichlet for topic proportion, as well as suitable prior distributions for parameters of interest.

Posterior inference is computationally challenging, due to the presence of mixed data type (categorical and continuous-valued). Moreover, the model is typically applied to large collection of documents, and later on images, genomes and all sort of large-scale data types. There are two computational tasks:

1. Compute the posterior distribution, $P(\theta, Z|\overline{W}, \alpha, \beta)$.
2. Estimating α, β from the data.

The posterior distribution can be rewritten as

$$P(\theta, Z|\overline{W}, \alpha, \beta) = \frac{P(\theta, Z, \overline{W}|\alpha, \beta)}{P(\overline{W}|\alpha, \beta)} \tag{37}$$

The numerator in the above display is easy to compute

$$P(\theta, Z, \overline{W}|\alpha, \beta) = P(\theta|\alpha) \prod_{n=1}^{N} P(Z_n|\theta) P(W_n|Z_n, \beta). \tag{38}$$

However, the denominator

$$p(\overline{W}|\alpha, \beta) = \int_{\theta} \sum_{Z_1, \ldots, Z_n} P(\theta, Z, \overline{W}|\alpha, \beta) \mathrm{d}\theta = \int \frac{\Gamma(\sum \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \prod_{n=1}^{N} \left[\sum_{k=1}^{K} \prod_{j=1}^{V} (\theta_k \beta_{kj})^{\mathbb{I}\{W_n = j\}} \right] \mathrm{d}\theta \tag{39}$$

is much harder to compute because we must integrate out all the latent variables of mixed types.

Exercise. Derive a Gibbs sampling algorithm for the LDA model. For this purpose, we need to endow prior distributions for parameters α and $\boldsymbol{\beta}$.

Although the Gibbs sampler is easy to derive, the Markov chains it produces may take a long time to mix (due to the large number of latent variables to be sampled). An alternative is variational inference — a general method for approximating posterior distributions based on optimization. We will introduce this method in the context of LDA next. Note that the state of the art method $\overline{W}$ for learning specifically the LDA model and its extensions, both in terms of parameter estimation accuracy and computational efficiency, appears to be a geometric algorithm of Yurochkin et al.[10]

[10] Dirichlet simplex nest and geometric inference. M. Yurochkin, A. Guha, Y. Sun and X. Nguyen. Proceedings of the 36th International Conference on Machine Learning (ICML), 2019.

Goal: approximate the posterior

$p(\theta, z | \bar{w}, \alpha, \beta)$

8.5.3 Variational Bayes

Variational inference is a general computational technique for inference with complex models in which the problem of model fitting and probabilistic inference (problem 1 and 2 in the previous page) can be reformulated as an optimization problem.

When applied to the approximate computation of the posterior distribution, we call this "variational Bayes". The strength of variational Bayes is that it's generally applicable to all (complex) Bayesian models; it's fast compared to sampling based techniques such as MCMC. While fast, it may not be as accurate as MCMC if the latter is run for sufficiently long time.

We shall now illustrate the variational Bayes technique to topic models. The basic idea is as follows:

(1) Consider a family of simplified distribution $Q = \{q(\theta, Z|\overline{W})\}$

(2) Choose the one in Q, that is closest to the true posterior

$$q^* := \text{argmin}_{q \in Q} \, KL(q||p(\theta, Z|\overline{W}, \alpha, \beta)) \tag{40}$$

is calculable even if we don't known exact posterior, but only know it up to normalizing constant

(3) Use q^* as the surrogate for the true posterior $p(\theta, Z|\overline{W}, \alpha, \beta)$ for subsequent inferential purposes.

In the above display, KL denotes the Kullback-Leibler divergence: given two distributions with corresponding probability density functions f and g on some common space, the KL divergence is given by

$$KL(f||g) = \mathbb{E}_f \log(f(X)/g(X)) = \int f(x) \log(f(x)/g(x)) dx.$$

Although Kullback-Lebler divergence is not symmetric, it is always non-negative. Moreover, $KL(f||g) = 0$ iff $f(x) = g(x)$ for almost all x. The KL is a fundamental quantity that measures how far g is from f.

$KL(f||g) \neq KL(g||f)$

It is somewhat surprising but not difficult to verify that the optimization problem given in Eq. (40) becomes relatively tractable the class of approximating distribution Q takes a sufficiently simple form.

The simple choice for Q is the family of "factorized" distributions: each $q \in Q$ satisfies

$$q(\theta, Z|\overline{W}, \gamma, \boldsymbol{\phi}) = q(\theta|\gamma)\Pi_{n=1}^{N} q(Z_n|\phi_n). \quad \text{mean field} \tag{41}$$

Here, the parameters γ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)$ are called *variational parameters* to be optimized according to the KL objective so as to obtain as tight as possible an approximation to the true posterior

$$(\gamma^*, \phi^*) := \operatorname{argmin} KL(q(\theta, Z|\gamma, \phi)||p(\theta, Z|\overline{W}, \alpha, \beta)). \tag{42}$$

A few words about the roles of variational parameters γ and $\boldsymbol{\phi}$: recall that $\theta \in \Delta^{K-1}$. Here, we shall take $q(\theta|\gamma)$ to be Dirichlet with parameters $\gamma \in R_+^K$.

Similarly, for each $n = 1, \ldots, N$, $q(Z_n|\phi_n)$ is taken to be categorical distribution, where parameter ϕ_n is composed of $\phi_n = (\phi_{n1}, \ldots, \phi_{nK})$ so that under q:

$$q(Z_n = i|\phi_n) = \phi_{ni}, \ n = 1, ..., N, i = 1, \ldots, K. \tag{43}$$

Optimization algorithm for variational Bayes We will show that the optimization in Eq. (42) can be solved by coordinate descent via iteratively applying the updating equations as follows: for $n = 1, \ldots, N$, $i = 1, \ldots, K$,

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}, \tag{44}$$

$$\phi_{ni} \propto \beta_{iW_n} \exp\{\mathbb{E}_q[\log \theta_i | \gamma]\}. \tag{45}$$

Thus, the algorithm is fairly simple to implement: initialize the variational parameters $\gamma, \boldsymbol{\phi}$ in some fashion, and then keep updating them via above equations until convergence.

Some remarks

(1) In the updating equation for ϕ_{ni}, since $\theta_i | \gamma \sim \text{Dirichlet}(\gamma)$ it is a simple fact of the Dirichlet distribution that

$$\mathbb{E}\left[\log \theta_i | \gamma\right] = \Psi(\gamma_i) - \Psi(\sum_{i=1}^{K} \gamma_i), \tag{46}$$

where Ψ is called digamma function $\Psi(x) = \frac{d \log \Gamma}{dx} = \frac{\Gamma'(x)}{\Gamma(x)}$.

(2) Note the roles of data W_n in the two updating equations.

(3) The updating equations are reminiscent of Gibbs sampler's updates for semi-conjugate priors, except here the updates are deterministic (subject to initialization). The fact that we are optimizing rather than sampling makes this approximate inference technique computationally more efficient than MCMC.

The remaining pages in this section will be devoted the to derivation of the algorithm and can be skipped at the first reading.

The first step is to note that the minimization of the KL divergence in Eq. (42) is equivalently viewed as the maximization of a lower bound to the log likelihood function of the original LDA model. Indeed, by Jensen's inequality

$$\begin{aligned}
\log p(\overline{W}|\alpha,\beta) &= \log \int_\theta \sum_Z p(\theta, Z, \overline{W}|\alpha,\beta) d\theta \\
&= \log \int_\theta \sum_Z \frac{p(\theta, Z, \overline{W}|\alpha,\beta)}{q(\theta, Z)} q\left(\theta, Z\right) d\theta \\
&\geq \int_\theta \sum_Z q\left(\theta, Z\right) \log \frac{p(\theta, Z, \overline{W}|\alpha,\beta)}{q(\theta, Z)} d\theta \\
&= \int_\theta \sum_Z q\left(\theta, Z\right) \log p(\theta, Z, \overline{W}|\alpha,\beta) d\theta - \int_\theta \sum_Z q\left(\theta, Z\right) \log q(\theta, Z) d\theta \\
&= \mathbb{E}_q \log p(\theta, Z, \overline{W}|\alpha,\beta) - \mathbb{E}_q \log q(\theta, Z) \\
&=: L(\gamma, \phi; \alpha, \beta).
\end{aligned}$$

We immediately see that the difference between the two sides of the above inequality is

$$\begin{aligned}
\log p(\overline{W}|\alpha,\beta) - L(\gamma, \phi; \alpha, \beta) &= \mathbb{E}_q \left\{ \log q(\theta, Z) - \log \frac{p(\theta, Z, \overline{W}|\alpha,\beta)}{p(\overline{W}|\alpha,\beta)} \right\} \\
&= \mathbb{E}_q \left\{ \log q(\theta, Z) - \log p(\theta, Z|\overline{W}, \alpha,\beta) \right\} \\
&= KL(q(\theta, Z)||p(\theta, Z|\overline{W}, \alpha,\beta)),
\end{aligned}$$

so minimizing the KL divergence in Eq. (42) is equivalent to

$$\max_{\gamma,\phi} L(\gamma, \phi; \alpha, \beta).$$

The second step is to note that the quantities in $L(\gamma, \phi; \alpha, \beta)$ are relatively easy to compute and optimize, due to the fact that the (full) joint probability distribution $p(\theta, Z, \overline{W}|\alpha, \beta)$ factorizes into marginal and conditional distributions, while q also factorizes by our choice of approximation. Indeed,

$$\log p(\theta, Z, \overline{W}|\alpha, \beta) = \log p(\theta|\alpha) + \sum_{n=1}^{N} \{\log p(Z_n|\theta) + \log p(W_n|Z_n, \beta)\},$$

so taking expectation with respect to the q distribution we obtain

$$\begin{aligned} L(\gamma, \phi; \alpha, \beta) &= \mathbb{E}_q \log p\left(\theta|\alpha\right) + \sum_{n=1}^{N} \{\mathbb{E}_q \log p(Z_n|\theta) + \mathbb{E}_q \log p(W_n|Z_n, \beta)\} \\ &\quad -\mathbb{E}_q \log q(\theta|\gamma) - \sum_{n=1}^{N} \mathbb{E}_q \log q\left(Z_n|\phi_n\right). \end{aligned} \tag{47}$$

Now, we proceed to compute each of the quantities in the above display.

$$\begin{aligned} p\left(\theta|\alpha\right) &= \frac{\Gamma\left(\sum \alpha_i\right)}{\prod_{i=1}^{K} \Gamma\left(\alpha_i\right)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1}, \text{so} \\ \log p\left(\theta|\alpha\right) &= \sum_{i=1}^{K} \left(\alpha_i - 1\right) \log \theta_i + \log \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - \sum_{i=1}^{K} \log \Gamma\left(\alpha_i\right) \\ \mathbb{E}_q \log p\left(\theta|\alpha\right) &= \sum_{i=1}^{K} \left(\alpha_i - 1\right) \left(\Psi(\gamma_i) - \Psi\left(\sum_{i=1}^{K} \gamma_i\right)\right) + \log \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - \sum_{i=1}^{K} \log \Gamma\left(\alpha_i\right). \end{aligned}$$

Next up, we consider $\sum_{n=1}^{N} \mathbb{E}_q \log p(Z_n|\theta)$:

$$\begin{aligned} p(Z_n|\theta) &= \prod_{i=1}^{K} \theta_i^{I(Z_n = i)}, \text{so} \\ \log p(Z_n|\theta) &= \sum_{i=1}^{K} I\left(Z_n = i\right) \log \theta_i \\ \mathbb{E}_q \log p(Z_n|\theta) &= \sum_{i=1}^{K} \phi_{ni} \left(\Psi\left(\gamma_i\right) - \Psi\left(\sum_{i=1}^{K} \gamma_i\right)\right), \end{aligned}$$

where the last equality is due to (46).

Continuing along,

$$\begin{aligned} \log p\left(W_n|Z_n, \beta\right) &= \log \prod_{i=1}^{K} \prod_{j=1}^{V} \left(\beta_{ij}\right)^{I(W_n = j, Z_n = i)}, \text{so} \\ \mathbb{E}_q \log p(Z_n|\theta) &= \sum_{i=1}^{K} \sum_{j=1}^{V} I(W_n = j) \phi_{ni} \log \beta_{ij}. \end{aligned}$$

In addition, we take care of $q(\theta|\gamma)$ and $q(Z_n|\phi_n)$

$$\begin{aligned} q(\theta|\gamma) &= \frac{\Gamma(\sum \gamma_i)}{\sum_{i=1}^{K} \Gamma(\gamma_i)} \prod_{i=1}^{K} \theta_i^{\gamma_i - 1}, \text{so} \\ E_q \log q(\theta|\gamma) &= \sum_{i=1}^{K} (\gamma_i - 1) \left(\Psi(\gamma_i) - \Psi\left(\sum_{i=1}^{K} \gamma_i \right) \right) + \log \Gamma(\sum_{i=1}^{K} \gamma_i) - \sum_{i=1}^{K} \log \Gamma(\gamma_i) \end{aligned}$$

as well,

$$\begin{aligned} q(Z_n|\phi_n) &= \prod_{i=1}^{K} \phi_{ni}^{I(Z_n=i)}, \text{so} \\ \mathbb{E}_q \log q(Z_n|\phi_n) &= \sum_{i=1}^{K} \phi_{ni} \log \phi_{ni}. \end{aligned}$$

The final step: with all components in the expression (47) for $L(\gamma, \boldsymbol{\phi}; \alpha, \beta)$ computed, it remains to optimize L with respect to the unknown variational parameters γ and $\boldsymbol{\phi}$.

$$\max_{\gamma, \boldsymbol{\phi}} \quad L(\gamma, \boldsymbol{\phi}; \alpha, \beta) \tag{48}$$

$$\text{subject to} \quad \sum_{i=1}^{K} \phi_{ni} = 1 \;\; n = 1, \ldots, N. \tag{49}$$

Differentiate with the γ and set to zero to obtain the updating equation (45) for γ. Differentiate with respect to the Lagrangian (by accounting for the equality constraints for ϕ_n) and set to zero to obtain the updating equation (44) for ϕ_n. Iterating these algorithms upon convergence for the estimates $\gamma^*, \boldsymbol{\phi}^*$).

Thus, we have accomplished the task of approximating the true posterior $p(\theta, Z|\overline{W}, \alpha, \beta)$ by means of the surrogate $q(\theta, Z|\gamma^*, \boldsymbol{\phi}^*)$. The second task of estimating the parameter α, β can also be done in a similar fashion. See Blei et al (2003) for details.

The materials in this course are adapted from materials created by David Blei, Yang Chen, Andrew Gelman, Scott Linderman, Long Nguyen, and the 3blue1brown channel.