AMS 207 - Homework 1

SAT Data

Mary Silva

1 Data

We use the textbook problem in which the Bayesian analysis gives conclusions that differ in important respects from other methods. We look at the data from eight schools which was part of a study to analyze the effects of special coaching programs on SAT scores. (FYI, there's many typos on the slides which was irritating). The data (from the slides) that we analyze is summarized in table below

School	Estimated Treatment effect (y_j)	Estimated Standard Error (σ_j)
A	29.39	14.90
B	7.94	10.20
C	-2.75	16.30
D	6.82	11.00
E	-0.64	9.40
F	0.63	11.40
G	18.01	10.40
H	12.16	17.60

1.1 Hierarchical Model Specifications

We have $J = 8$ independent experiments, each one is performed to estimate θ_j , the unobserved true effect, from n_j independent data points y_{ij} . We assume normality and for $i = 1, ..., n_j$; $j = 1, ..., J$;, we have

$$
\begin{aligned}
y_{ij}|\theta_j &\sim N(\theta_j, \sigma_j^2), & \sigma^2 \text{ known} \\
\overline{y_{\cdot j}} &= \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij} & \sigma_j^2 = \frac{\sigma^2}{n_j} \\
\overline{y_{\cdot\cdot}} &= \frac{\sum_{j=1}^{J} \overline{y_{\cdot j}}/\sigma_j^2}{\sum_{j=1}^{J} 1/\sigma_j^2}
\end{aligned}
$$

We construct a hierarchical model with the following specifications

$$
\begin{aligned}
p(\theta_1, \ldots, \theta_J|\mu, \tau) &= \prod_{j=1}^{J} \mathrm{N}(\theta_j|\mu, \tau^2) \\
p(\theta_1, \ldots, \theta_J) &= \int \prod_{j=1}^{J} [\mathrm{N}(\theta_j|\mu, \tau^2)]\, p(\mu, \tau) d(\mu, \tau).
\end{aligned}
$$

i.e. the θ_j are conditionally independent given (μ, τ) We also assign a non informative uniform hyperprior distribution to μ, given τ.

$$
p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto p(\tau).
$$

Our joint posterior can be expressed in terms of the sufficient statistics

$$
\begin{aligned}
p(\theta, \mu, \tau | y) &\propto p(\mu, \tau) p(\theta | \mu, \tau) p(y | \theta) \\
&\propto p(\mu, \tau) \prod_{j=1}^{J} \mathrm{N}(\theta_j | \mu, \tau^2) \prod_{j=1}^{J} \mathrm{N}(\overline{y}_{.j} | \theta_j, \sigma_j^2),
\end{aligned}
$$

The conditional posterior distributions for for θ_j becomes

$$
\theta_j | \mu, \tau, y \sim \mathrm{N}(\hat{\theta}_j, V_j),
$$

$$
\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2} \overline{y}_{.j} + \frac{1}{\tau^2} \mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \quad \text{and} \quad V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}.
$$

The remaining specifications are listed below

$$
\overline{y}_{.j} | \mu, \tau \sim \mathrm{N}(\mu, \sigma_j^2 + \tau^2).
$$

$$
p(\mu, \tau | y) \propto p(\mu, \tau) \prod_{j=1}^{J} \mathrm{N}(\overline{y}_{.j} | \mu, \sigma_j^2 + \tau^2).
$$

$$
\mu | \tau, y \sim \mathrm{N}(\hat{\mu}, V_\mu),
$$

$$
\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2} \overline{y}_{.j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}} \quad \text{and} \quad V_\mu^{-1} = \sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}.
$$

$$
\begin{aligned}
p(\tau | y) &= \frac{p(\mu, \tau | y)}{p(\mu | \tau, y)} \\
&\propto \frac{p(\tau) \prod_{j=1}^{J} \mathrm{N}(\overline{y}_{.j} | \mu, \sigma_j^2 + \tau^2)}{\mathrm{N}(\mu | \hat{\mu}, V_\mu)}.
\end{aligned}
$$

$$
\begin{aligned}
p(\tau | y) &\propto \frac{p(\tau) \prod_{j=1}^{J} \mathrm{N}(\overline{y}_{.j} | \hat{\mu}, \sigma_j^2 + \tau^2)}{\mathrm{N}(\hat{\mu} | \hat{\mu}, V_\mu)} \\
&\propto p(\tau) V_\mu^{1/2} \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(-\frac{(\overline{y}_{.j} - \hat{\mu})^2}{2(\sigma_j^2 + \tau^2)} \right),
\end{aligned}
$$

$$
p(\theta, \mu, \tau | y) = p(\tau | y) p(\mu | \tau, y) p(\theta | \mu, \tau, y).
$$

2 Direct Sampling

In simple non-hierarchical Bayesian models, it is often easy to draw from the posterior distribution directly, especially if conjugate prior distributions have been assumed. Frequently, draws from standard distributions or low-dimensional non-standard distributions are required, either as direct draws from the posterior distribution of the estimand in an easy problem, or as an intermediate step in a more complex problem.

2.1 Procedure

For the simplest discrete approximation, compute the target density, $p(\theta|y)$, at a set of evenly spaced values $\theta_1, ..., \theta_N$ that cover a broad range of the parameter space for θ, then approximate the continuous $p(\theta|y)$ by the discrete density at $\theta_1, ..., \theta_N$, with probabilities $p(\theta_i|y) / \sum_{l=1}^{N} p(\theta_l|y)$. Because the approximate density must be normalized anyway, this method will work just as well using an unnormalized density function, $q(\theta|y)$, in place of $p(\theta|y)$.

Once the grid of density values is computed, a random draw from $p(\theta|y)$ is obtained by drawing a random sample Uniform distribution on [0, 1], then transforming by the inverse cdf method to obtain a sample from the discrete approximation. When the points θ_i are spaced closely enough and miss nothing important beyond their boundaries, this method works well.

3 Results

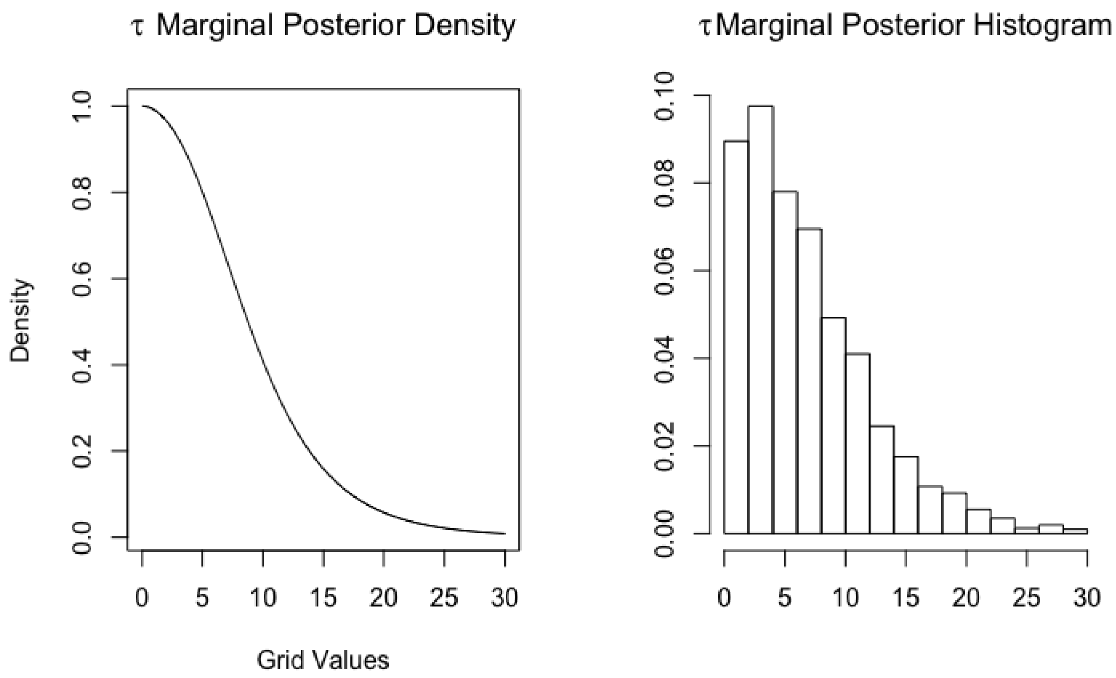


Figure 1: Direct Sampling: Marginal posterior density and histogram, $p(\tau|y)$, for standard deviation of the population of school effects θ_j in the educational testing example.

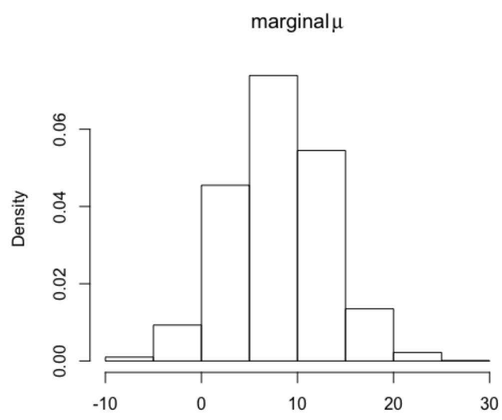


Figure 2: Marginal posterior histogram for μ.

Table 1: The Direct Sampling Summary of 2000 simulations of the treatment effects in the eight schools.

θ_j	mean	sd	0.025	0.975
1	12.37	8.58	-1.62	33.21
2	8.27	6.54	-4.45	21.08
3	6.75	8.12	-10.49	22.13
4	7.98	6.75	-5.18	21.64
5	5.81	6.50	-7.80	17.60
6	6.36	6.98	-8.50	18.91
7	10.98	6.87	-1.26	26.19
8	8.74	8.27	-7.48	25.94

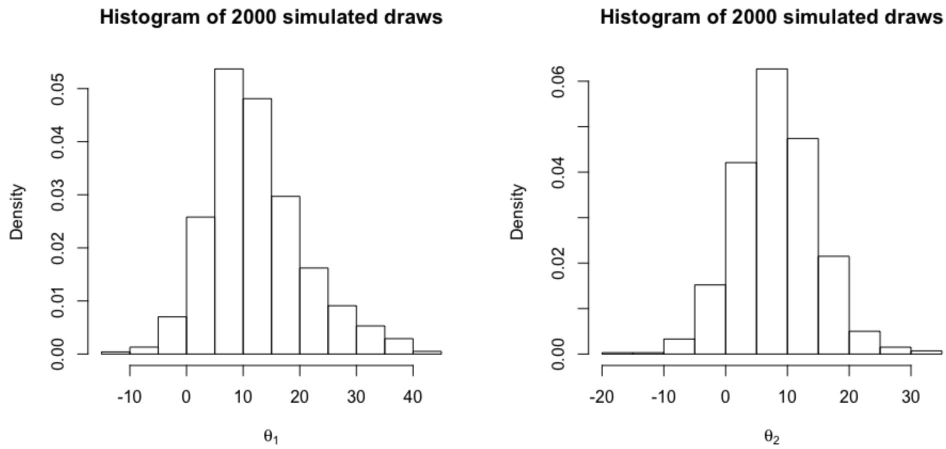


Figure 3: Histograms of two quantities of interest computed from the 2000 simulatied draws of the effect in school A, θ_1, and the effect in school B, θ_2.

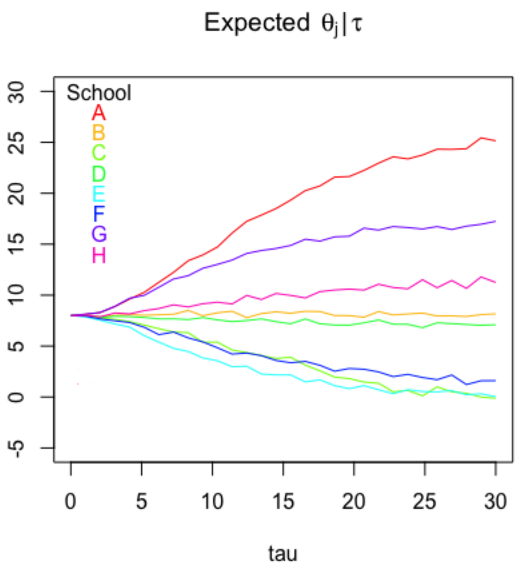


Figure 4: Conditional posterior means of treatment effects, $E(\theta_j|\tau, y)$, as functions of the between school standard deviation τ, for the educational testing example. The line for school C crosses the lines for E and F because C has a higher measurement error) and its estimate is therefore shrunk more strongly toward the overall mean in the Bayesian analysis.

4 Gibbs Sampling

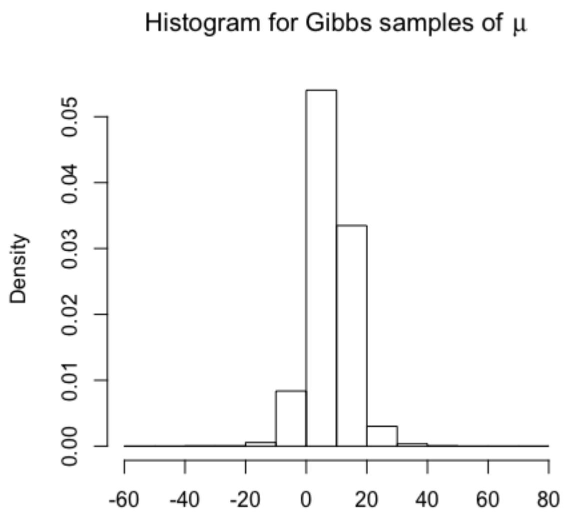
Here we wish to draw random samples from the joint posterior given by

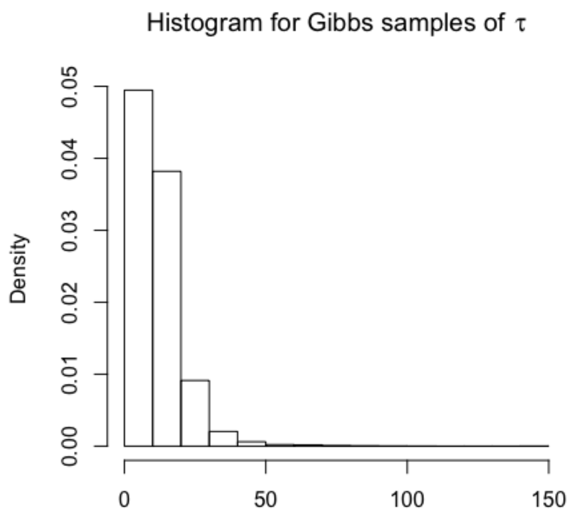
$$p(\theta, \mu, \tau | y) \quad \propto \quad p(\mu, \tau) p(\theta | \mu, \tau) p(y | \theta)$$

On the right hand side of the proportionality statement we have defined each of these in closed form, thus a Gibbs sampling algorithm is appropriate.

Table 2: The Gibbs Sampling Summary of 2000 simulations of the treatment effects in the eight schools. There is clearly a slight difference in certain values for the posterior expected mean for θ, but it is clearly comparable to Direct sampling

parameter	Mean	SD	0.025%	0.975%
θ_1	14.97	10.40	-2.20	39.04
θ_2	8.04	7.52	-7.10	23.72
θ_3	4.76	9.91	-16.78	23.74
θ_4	7.64	7.81	-8.43	23.09
θ_5	3.49	7.35	-12.13	17.41
θ_6	4.77	8.18	-12.45	20.03
θ_7	12.56	8.17	-1.94	30.05
θ_8	9.22	10.40	-11.27	30.39
μ	8.07	6.67	-4.81	21.07
τ	12.06	8.98	1.81	33.24





Most of what is seen above appears different than the direct sampling method. So we examine the trace plots and auto correlation plots to see if mixing is occuring properly.

(ADD PLOTS)

5 Abrams & Sansò

From [Abrams & Sansò 1998] we define the following model specifications:

$$
\begin{aligned}
y_i &\sim N(\theta_i, \sigma_i^2/n) \\
\theta_i &\sim N(\mu, \tau^2) \\
\sigma_i^2 &\propto 1/\sigma_i^2 \\
\tau^2 &\sim IG(a, b) \\
\mu &\sim Unif(0, 1)
\end{aligned}
$$

According to the paper, the prior on σ_i^2, chosen to be Jeffrey's prior is due to the fact that in practice little information is likely to be available about the within-study variances. A prior distribution for τ^2 has to be flexible enough to be able to easily accommodate a priori information whilst at the same time being mathematically convenient. For this reason, the paper suggests placing an inverse gamma distribution with parameters a and b. A vague Uniform prior is suggested to be placed on μ.

we have an approximation for the first moment of the effect of the ith school as described in the SAT model is given by

$$E(\theta_i|y, s^2, n) \approx y_i - \frac{(n_i-1)s_i^2 b(2a+k-1}{2n_i(n_i-3)(1+bRSS_B/2}(y_i - \bar{y})$$

where $RSS_b = \sum_i y_i^2 - k\bar{y}^2$, the residual sum of squares between studies (schools) and a and b are the hyper parameters of the inverse gamma prior on τ. We also have the following approximations

$$V(\theta_i|\boldsymbol{y}, \boldsymbol{s}^2, \boldsymbol{n}) \approx \frac{(n_i-1)s_i^2}{n_i(n_i-3)}\left\{1 + \frac{(n_i-1)s_i^2 b^2(2a+k-1)(n_i-4+2a+k)(y_i-\bar{y})^2}{2n_i(n_i-3)(n_i-5)(1+b\mathrm{RSS_B}/2)^2} + \frac{(n_i-1)s_i^2 b(2a+k-1)}{2kn_i(n_i-5)(1+b\mathrm{RSS_B}/2)}\right\}$$

$$E(\sigma_i^2|\boldsymbol{s}^2, \boldsymbol{n}) \approx \frac{n_i}{n_i-2}s_i^2.$$

$$E(\mu|\boldsymbol{y}, \boldsymbol{s}^2, \boldsymbol{n}) \approx \bar{y} \quad \text{and} \quad V(\mu|\boldsymbol{y}, \boldsymbol{s}^2, \boldsymbol{n}) \approx \frac{2(1+b\mathrm{RSS_B}/2)}{bk(2a+k-3)}.$$

$$E(\tau^2|\boldsymbol{y}, \boldsymbol{s}^2, \boldsymbol{n}) \approx \frac{2(1+b\mathrm{RSS_B}/2)}{b(2a+k-3)} \quad \text{and} \quad V(\tau^2|\boldsymbol{y}, \boldsymbol{s}^2, \boldsymbol{n}) \approx \frac{8(1+b\mathrm{RSS_B}/2)^2}{b^2(2a+k-3)^2(k+2a-5)}.$$

$$E(\mu|\boldsymbol{y}, \boldsymbol{s}^2, \boldsymbol{n}) \approx \frac{\bar{y} - \frac{b(k+2a-1)}{2(1+b\mathrm{RSS_B}/2)}\sum_{i=1}^{k}\frac{n_i s_i^2}{n_i-3}\left(\frac{\bar{y}(k-3)+y_i}{k} - \frac{(\bar{y}^2(\bar{y}-y_i)+\bar{y}y_i^2)(k+2a+1)b}{2(1+b\mathrm{RSS_B}/2)}\right)}{1 - \frac{b(k+2a-1)}{2(1+b\mathrm{RSS_B}/2)}\sum_{i=1}^{k}\frac{n_i s_i^2}{n_i-3}\left(\frac{k-1}{k} - \frac{(\bar{y}(\bar{y}-y_i)+y_i^2)(k+2a+1)b}{2(1+b\mathrm{RSS_B}/2)}\right)}.$$

These are written as in the paper, but I was not getting sensible results at first until. I was using very odd choices for the parameters of the Inverse gamma prior on τ^2. Using $a = 2$ and $b = 50$, I obtained the following (I will be asking you about this later because Some of these seem close, but the theta's are not so I am curious if there is also a typo in the paper? for this reason I exclude the approximations for thetas.)

$E(\mu\|y, s^2, n) \approx 8.5$	$V(\mu\|y, s^2, n) \approx 79.98$
$E(\tau\|y, s^2, n) \approx 18.03$	$V(\tau\|y, s, n) \approx 116.179$

6 Conclusion

I wanted to write a better conclusion and interpret the results better, but I did not manage my time well. Please enjoy this meme that I will probably attach to every homework because it never gets old!