Stat260: Bayesian Modeling and Inference **Lecture Date: March 29, 2010**

Lecture 15

Lecturer: Michael I. Jordan *Scribe: Joshua G. Schraiber*

1 Hierarchical models

Recall that we were discussing hierarchical models last lecture. As an example, we were examining a random effects model. Figure 1 shows the graphical model, which indicates that data is generated J Gaussians with means $\theta_1, \ldots, \theta_J$, which themselves were generated by a Gaussian with parameter μ higher up in the hierarchy. An example might be that the y_{ij} are the SAT scores of student i at school j and the θ_j is the mean SAT score at school j.

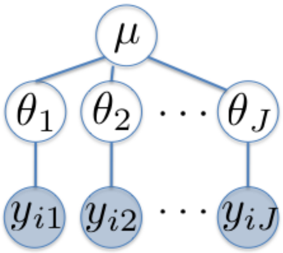

Figure 1: Graphical depiction of the random effects model

For simplicity, we considered the sufficient statistics of the model:

$$\bar{Y}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}, \quad \bar{Y}_{\cdot j} \sim N(\theta_j, \sigma_j^2), \quad \sigma_j^2 = \frac{\sigma^2}{n_j} \tag{1}$$

$$\theta_j \sim N(\mu, \tau^2), \quad \pi(\mu, \tau^2) \propto \pi(\tau). \tag{2}$$

Other than the $\pi(\mu, \tau^2)$ term, this would be a classical random effects model. However, the prior on μ and τ makes it a Bayesian model. Some standard calculations reveal that the posterior mean of $\theta_j \,|\, \mu, \tau^2, y$ is

$$\frac{\sigma_j^{-2}\bar{Y}_{\cdot j} + \tau^{-2}\mu}{\sigma_j^{-2} + \tau^{-2}}. \tag{3}$$

This shows that this estimator of θ shrinks back towards μ, which is a desirable property. Unfortunately, we don't know what μ is! Nevertheless, as Bayesians we can observe that $\mu \,|\, \tau, y \sim N(\hat{\mu}, V_\mu)$ where

$$\hat{\mu} = \frac{\sum_j (\sigma_j^2 + \tau^2)^{-1}\bar{Y}_{\cdot j}}{\sum_j (\sigma_j^2 + \tau^2)^{-1}} \tag{4}$$

$$V_\mu = \sum_j (\sigma_j^2 + \tau^2)^{-1}. \tag{5}$$

Moreover, to obtain $p(\tau \mid y)$ we use the definition of conditional probability

$$p(\tau \mid y) = \frac{p(\mu, \tau \mid y)}{p(\mu \mid \tau, y)}. \tag{6}$$

The left side does not depend on μ while the right side does, hence it must be true for any μ. In particular we can make a natural choice of $\mu = \hat{\mu}$. So for the posterior distribution of τ we have

$$p(\tau \mid y) \propto \frac{\pi(\tau) \prod_j p(\bar{Y}_{\cdot j} \mid \hat{\mu}, \sigma^2, \tau^2)}{N(\hat{\mu}, V_\mu)}. \tag{7}$$

Working through the algebra shows that we *cannot* use the Jeffreys prior $\pi(\tau) \propto \tau^{-1}$ because it leads to an improper posterior distribution. So, how do we choose a prior for τ? So far, no systematic theory has been developed and it is a bit ad-hoc. One common approach is to choose a variety of priors that "work" in the sense of giving proper posterior distributions and running the analysis multiple times to see if the prior affects the inference that much. Some options that do "work" in this case are $\pi(\tau) \propto 1$ and $\pi(\tau)$ conjugate.

Example 1. SAT prep tests This example comes from Gelman et al. (2003). Students at different schools were given SAT prep classes and then they took the SAT (along with a control group who had not received prep classes). The data are the difference between the treatment group and the control group at each school. The data are provided in Table 1.

School	Difference	Standard error
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

Table 1: SAT score difference data from schools that participated in the experiment

To do an inference on this data, we have to first assume that schools are exchangeable. This is a reasonable assumption for now simply because we don't have any other information about the schools. However, if we were to obtain information that the schools were different in some way, we could easily apply another level to the hierarchy and the general framework would still hold.

The classical (frequentist) method of analysis would rely on either assuming all the schools are different or accepting the null hypothesis that all the schools are the same. Under the former, each school has a completely different effect and thus the MLE is simply the sample mean within each school. Accepting the null hypothesis would indicate the mean improvement is the overall mean (which happens to be 8.75).

Both of these analyses seem somewhat unsatisfying. For instance, since the data are assumed to be normally distributed, assuming each school is different would suggest that 50% of the students at school A improved by more than 28 points! But based on the rest of the data, that seems unlikely. Assuming the null hypothesis would mean that half the probability mass is below 8.75, which seems to not be reasonable considering average score in school A being so high. What we are looking for here is *shrinkage*, and we know that Bayes approaches give us shrinkage.

If we use Bayes, we must pick a prior for τ, as discussed previously. Using a uniform prior on τ, we find that the posterior distribution of τ has the shape shown in Figure 2. Moreover, it reduces the estimate of the median score in school A from 28 to 19, which seems more reasonable based on the rest of the data. This is a bit arbitrary but it does have nice features.

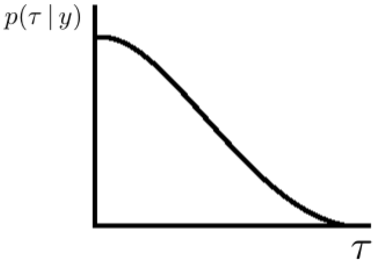

Figure 2: The posterior distribution of τ.

What if we were to analyze this data using empirical Bayes? From Figure 2, it is clear that $\hat{\tau}_{EB} = 0$, which is equivalent to accepting the null hypothesis that there is no variation between schools. This does not seem desirable compared to the full Bayes approach. In a future lecture we will see that this is because the number of samples per school was not large enough compared to the number of schools sampled.

Another benefit of using full Bayes is that allows us to compute probabilities from the data. For instance, $P(\max_j \theta_j > 28 \mid y) \approx .1$, which suggests that the high school for school A is not as much of an outlier as we would initially suspect. One can also compute, for example, $P(\theta_i = \theta_j)$ for any i, j. In general these computations will require numerical methods, but they are possible.

1.1 Meta-analysis

We briefly mention that this same approach can be applied to a meta-analysis. In a meta-analysis, a statistician will comb through the literature on a subject (say clinical trials of a drug) and aggregate data in order to gain more power.

2 Computational aspects

We now change focus and discuss computational aspects. Much of this class has been focused on obtaining analytical expressions for quantities of interest. However, many problems are high dimensional and don't yield to analytical computation very easily. In particular, we may be interested in 1) computing expectations and other high dimensional integrals under the posterior distribution, 2) obtaining samples from the posterior distribution, 3)doing sensitivity analysis, etc.

2.1 Examples of methods

Some methods that we may or may not discuss in this class are

- Numerical integration (e.g. quadrature)
- Importance sampling (generates IID samples)
- Markov Chain Monte Carlo (samples from a Markov chain)

- Sequential Monte Carlo (a special case of Importance Sampling, used initially with time-series data)
- Variational inference (optimization-based approaches)
- Laplace expansions (which approximate integrals)

The methods of numerical integration, variational inference and Laplace expansion are closely related and are deterministic. The other methods are stochastic.

We will now consider the Laplace expansion method.

3 Laplace approximation

Laplace approximation provides a general way to approach marginalization problems. The basic setting for Laplace approximation is an integral of the form:

$$I(t) = \int e^{-th(x)} dx, \tag{8}$$

where we are interested in the value of $I(t)$ as t approaches an asymptotic limit (which for concreteness we will take to be infinity). The method aims specifically at problems in which $h(x)$ is unimodal with strictly positive second derivative at its modal value $\hat{x}$ so that the integrand peaks sharply around $\hat{x}$ as t approaches infinity.

Laplace approximation belongs to the field of asymptotic analysis. This field is concerned with the general problem of obtaining approximate solutions to sets of parameterized problems (sums, integrals, differential equations, etc) as a parameter approaches an asymptotic limit. The results generally take the form of an *asymptotic series*. An asymptotic series is not necessarily a convergent series—for any specific value of t the series may diverge. But as t goes to infinity, the series converges to the solution, and converges increasingly rapidly as more terms are included in the series.

In statistical applications of Laplace approximation, the asymptotic parameter is generally—but not always—the number of data points N. To suggest this application we will use N in place of t in the remainder of this section.

Laplace approximation is essentially an application of Taylor series expansion—we expand both $h(x)$ and the exponential function in Taylor series, collect terms and compute the resulting elementary integrals. This yields a sum of terms that are functions of N—an asymptotic expansion. The mathematical conditions that are required for Laplace approximation to yield an honest-to-goodness asymptotic expansion are essentially those under which this linked pair of Taylor series expansions can be justified. We will proceed without concern for these regularity conditions.

We begin by expanding $h(x)$ in a Taylor series up to fourth order:

$$h(x) \approx h(\hat{x}) + \frac{1}{2}h_2(\hat{x})(x-\hat{x})^2 + \frac{1}{6}h_3(\hat{x})(x-\hat{x})^3 + \frac{1}{24}h_4(\hat{x})(x-\hat{x})^4, \tag{9}$$

where $h_i(\hat{x})$ denotes the ith derivative of $h(x)$ evaluated at $\hat{x}$, and where $\hat{x}$ is the value of x such that

$h_1(\hat{x}) = 0$. Plugging the latter expansion into our integral, we obtain:

$$\begin{aligned} I(N) &= \int e^{-Nh(x)}dx \\ &\approx \int \exp\left\{-N\left(h(\hat{x}) + \frac{1}{2}h_2(\hat{x})(x-\hat{x})^2 + \frac{1}{6}h_3(\hat{x})(x-\hat{x})^3 + \frac{1}{24}h_4(\hat{x})(x-\hat{x})^4\right)\right\}dx \\ &= e^{-Nh(\hat{x})}\int e^{-\frac{N}{2}\hat{h}_2u^2}\exp\left\{-\frac{N}{6}\hat{h}_3u^3 - \frac{N}{24}\hat{h}_4u^4\right\}du, \end{aligned} \tag{10}$$

where we have changed variables and introduced the shorthand $\hat{h}_i = h_i(\hat{x})$. We see that we need to perform a Gaussian integral. To make further progress we expand the exponential in a second-order Taylor series:

$$\begin{aligned} I(N) &\approx e^{-Nh(\hat{x})}\int e^{-\frac{N}{2}\hat{h}_2u^2}\left(1 - \frac{N}{6}\hat{h}_3u^3 - \frac{N}{24}\hat{h}_4u^4 + \frac{1}{2}\left(\frac{N^2}{36}\hat{h}_3^2u^6 + \frac{2N^2}{144}\hat{h}_3\hat{h}_4u^7 + \frac{N^2}{576}\hat{h}_4^2u^8\right)\right)du \\ &= e^{-Nh(\hat{x})}\int e^{-\frac{N}{2}\hat{h}_2u^2}\left(1 - \frac{N}{24}\hat{h}_4u^4 + \frac{N^2}{72}\hat{h}_3^2u^6 + \frac{N^2}{1052}\hat{h}_4^2u^8\right)du, \end{aligned} \tag{11}$$

where the integrals of odd powers vanish given that the Gaussian is an even function. Each of the remaining Gaussian integrals can be readily computed. Indeed, the general form of a Gaussian integral of a power is obtained via integration by parts and the definition of the gamma function[1]

$$\int e^{-sx^2}x^{2m}dx = \frac{2m!}{m!\ 2^{2m}}\pi^{1/2}s^{-(m+1)/2}. \tag{12}$$

Substituting into Eq. (11), we obtain the final result:

$$I(N) \approx e^{-Nh(\hat{x})}\sqrt{2\pi}\sigma N^{-1/2}\left(1 - \frac{\hat{h}_4\sigma^4}{8N} + \frac{5\hat{h}_3^2\sigma^6}{24N}\right), \tag{13}$$

where $\sigma^2 = 1/\hat{h}_2$. The neglected terms in this expansion are readily seen to be of order $O(1/N^2)$; in fact we have obtained an asymptotic expansion to this order. The first term in the expansion:

$$I(N) \approx e^{-Nh(\hat{x})}\sqrt{2\pi}\sigma N^{-1/2}. \tag{14}$$

is accurate to order $O(1/N)$.

NOTE: the following was actually covered in lecture 16

The multivariate case goes through in essentially the same way. Letting x denote a d-dimensional vector and $h(x)$ a scalar function of x, we find:

$$\int e^{-Nh(x)}dx \approx e^{-Nh(\hat{x})}(2\pi)^{d/2}|\Sigma|^{1/2}N^{-d/2}, \tag{15}$$

where $\Sigma = (D^2h(\hat{x}))^{-1}$ is the inverse of the Hessian of h evaluated at $\hat{x}$. This expansion is accurate to order $O(1/N)$. Moreover, by analogy with Eq. (13), the expansion can readily be continued to order $O(1/N^2)$.

A major application of Laplace approximation is to the computation of the marginal likelihood. It is also worth noting that Laplace approximation can be used for approximation of conditional probabilities and conditional expectations. Consider in particular the conditional expectation

$$E[g(X)\,|\,y] = \int g(x)p(x\,|\,y)dx \tag{16}$$

$$= \frac{\int g(x)p(y\,|\,x)p(x)dx}{\int p(y\,|\,x)p(x)dx}. \tag{17}$$

[1] To check this, you will need to make use of the following properties of the gamma function: $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(1/2) = \sqrt{\pi}$.

This ratio of integrals can be approximated as a ratio of Laplace approximations. One way to do this is to extend the Laplace approximation to integrals of the form $\int b(x)e^{-Nh(x)}dx$. In this case, in the numerator integral, the function $h(x)$ is defined as $-(\log p(y \,|\, x) + \log p(x))/N$. Another possibility, applicable when g is positive, is to define $h(x) = -(\log p(y \,|\, x) + \log p(x) + \log g(x))/N$ in the numerator integral. In taking the ratio it turns out that the $O(1/N)$ term vanishes (more accurately, it is $O(1/N^2)$), so that the overall approximation is of order $O(1/N^2)$.

References

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis.* Chapman and Hall.