Bayesian Data Analysis
Lectures Notes (Stat 220, Fall 2014)

Won I. Lee

These notes focus on summaries of the theoretical development of conjugate models (single- and multi-parameter), hierarchical models, mixture models (and Dirichlet processes), and Bayesian regression. It then tackles computational methods via MCMC, and addresses some topics in missing data for Bayesian data analysis.[1] In general, we assume exchangeability of observations.

1 Single-Parameter Models

1.1 Normal-Normal (Known Variance)

The quintessential conjugate model is the Normal-Normal model, with known variance (i.e. conditional on the variance). That is:

$$\begin{aligned} y_1, \ldots, y_n | \theta, \sigma^2 &\sim \mathcal{N}(\theta, \sigma^2) \\ \theta | \sigma^2 &\sim \mathcal{N}(\mu_0, \tau_0^2) \end{aligned}$$

Assuming that "we know σ^2", we can break down the posterior into:

$$p(\theta | \mathbf{y}, \sigma^2) \propto p(\mathbf{y} | \theta, \sigma^2) p(\theta | \sigma^2)$$

where $p(\mathbf{y}|\theta, \sigma^2) = \prod_{i=1}^n p(y_i|\theta, \sigma^2)$ by exchangeability. We also know the prior $p(\theta|\sigma^2)$, since it is specified by the Normal.

The posterior then can be simplified into a Normal (implied by the conjugacy) using completion of squares:

$$\begin{aligned} p(\theta | \mathbf{y}, \sigma^2) &\propto \prod_{i=1}^n p(y_i | \theta, \sigma^2) p(\theta | \sigma^2) \\ &\propto \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta)^2 \right] \exp\left[-\frac{1}{2\tau_0^2} (\theta - \mu_0)^2 \right] \\ &= \exp\left[-\frac{1}{2} \left(\frac{\sum y_i^2 - 2n\bar{y}\theta + n\theta^2}{\sigma^2} + \frac{\theta^2 - 2\theta\mu_0 + \mu_0^2}{\tau_0^2} \right) \right] \\ &\propto \exp\left\{ -\frac{1}{2} \left[\left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2} \right) \theta^2 - 2 \left(\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau_0^2} \theta \right) \right] \right\} \\ &\propto \exp\left[-\frac{n/\sigma^2 + 1/\tau_0^2}{2} \left(\theta - \frac{n\bar{y}/\sigma^2 + \mu_0/\tau_0^2}{n/\sigma^2 + 1/\tau_0^2} \right)^2 \right] \end{aligned}$$

$$\Rightarrow \boxed{\theta | \mathbf{y}, \sigma^2 \sim \mathcal{N}(\mu_n, \tau_n^2)}$$

where:

$$\boxed{\mu_n = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\tau_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}} \text{ and } \tau_n^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}}$$

In addition, the posterior predictive distribution of $\tilde{y}$ is also Normal:

[1] These notes are loosely based on topics covered in Stat 220 (Bayesian Data Analysis) at Harvard University, but most material is drawn from the reference material.

$$
\begin{aligned}
p(\tilde{y}|\mathbf{y},\sigma^2) &= \int p(\tilde{y}|\theta,\mathbf{y},\sigma^2)p(\theta|\mathbf{y},\sigma^2)d\theta \\
&= \int p(\tilde{y}|\theta,\sigma^2)p(\theta|\mathbf{y},\sigma^2)d\theta \\
&\propto \int \exp\left[-\frac{1}{2\sigma^2}(\tilde{y}-\theta)^2\right]\exp\left[-\frac{1}{2\tau_n^2}(\theta-\mu_n)^2\right]d\theta
\end{aligned}
$$

Because both are quadratic exponentials, we have that the posterior predictive distribution is also Normal. To find the moments, we use:

$$
E[\tilde{y}|\mathbf{y},\sigma^2] = E[E(\tilde{y}|\theta,\mathbf{y},\sigma^2)] = E[E(\tilde{y}|\theta,\sigma^2)] = E[\theta] = \mu_n
$$

$$
\mathrm{var}(\tilde{y}|\mathbf{y},\sigma^2) = E[\mathrm{var}(\tilde{y}|\theta,\mathbf{y},\sigma^2)] + \mathrm{var}[E(\tilde{y}|\theta,\mathbf{y},\sigma^2)] = E[\sigma^2] + \mathrm{var}(\theta) = \sigma^2 + \tau_n^2
$$

$$
\Rightarrow \boxed{\tilde{y}|\mathbf{y},\sigma^2 \sim \mathcal{N}(\mu_n, \tau_n^2 + \sigma^2)}
$$

1.2 Beta-Binomial

Suppose:

$$
\begin{aligned}
y|\pi &\sim \mathrm{Bin}(n,\pi) \\
\pi &\sim \mathrm{Beta}(a,b)
\end{aligned}
$$

And we observe some y success in the n trials. Then the posterior is:

$$
\begin{aligned}
p(\pi|y) &\propto p(y|\pi)p(\pi) \\
&\propto \left[\pi^y(1-\pi)^{n-y}\right]\left[\pi^{a-1}(1-\pi)^{b-1}\right] \\
&= \pi^{a+y-1}(1-\pi)^{b+n-y-1}
\end{aligned}
$$

$$
\Rightarrow \boxed{\pi|y \sim \mathrm{Beta}(a+y, b+n-y)}
$$

The posterior predictive distribution is beta-binomial:

$$
\begin{aligned}
p(\tilde{y}|y) &= \int p(\tilde{y}|\pi,y)p(\pi|y)d\pi \\
&= \int p(\tilde{y}|\pi)p(\pi|y)d\pi \\
&= \int \binom{n}{\tilde{y}}\pi^{\tilde{y}}(1-\pi)^{n-\tilde{y}}\frac{\Gamma(a+b+n)}{\Gamma(a+y)\Gamma(b+n-y)}\pi^{a+y-1}(1-\pi)^{b+n-y-1}d\pi \\
&= \binom{n}{\tilde{y}}\frac{\Gamma(a+b+n)}{\Gamma(a+y)\Gamma(b+n-y)}\int \pi^{a+y+\tilde{y}-1}(1-\pi)^{b+2n-y-\tilde{y}-1}d\pi \\
&= \binom{n}{\tilde{y}}\frac{\Gamma(a+b+n)}{\Gamma(a+y)\Gamma(b+n-y)}\frac{\Gamma(a+y+\tilde{y})\Gamma(b+2n-y-\tilde{y})}{\Gamma(a+b+2n)}
\end{aligned}
$$

$$
\Rightarrow \boxed{\tilde{y}|y \sim \text{Beta-Bin}(a+y, b+n-y)}
$$

1.3 Gamma-Poisson

Suppose:

$$
\begin{aligned}
y_1,\ldots,y_n|\theta &\sim \mathrm{Pois}(\theta) \\
\theta &\sim \mathrm{Gamma}(a,b)
\end{aligned}
$$

Then, assuming exchangeability, we have:

$$\begin{aligned} p(\theta|\mathbf{y}) &\propto p(\mathbf{y}|\theta)p(\theta) \\ &\propto \left[\prod_{i=1}^{n} \theta^{y_i} e^{-\theta}\right] \theta^{a-1} e^{-b\theta} \\ &= \theta^{a+n\bar{y}-1} e^{-(b+n)\theta} \end{aligned}$$

$$\Rightarrow \boxed{\theta|\mathbf{y} \sim \text{Gamma}(a+n\bar{y}, b+n)}$$

The posterior predictive distribution for $\tilde{y}$ is negative binomial:

$$\begin{aligned} p(\tilde{y}|\mathbf{y}) &= \int p(\tilde{y}|\theta, \mathbf{y}) p(\theta|\mathbf{y}) d\theta \\ &= \int p(\tilde{y}|\theta), p(\theta|\mathbf{y}) d\theta \\ &= \int \frac{\theta^{\tilde{y}} e^{-\theta}}{\tilde{y}!} \frac{(b+n)^{a+n\bar{y}}}{\Gamma(a+n\bar{y})} \theta^{a+n\bar{y}-1} e^{-(b+n)\theta} d\theta \\ &= \frac{(b+n)^{a+n\bar{y}}}{\tilde{y}!\Gamma(a+n\bar{y})} \int \theta^{a+n\bar{y}+\tilde{y}-1} e^{-(b+n+1)\theta} d\theta \\ &= \frac{(b+n)^{a+n\bar{y}}}{\tilde{y}!\Gamma(a+n\bar{y})} \frac{\Gamma(a+n\bar{y}+\tilde{y})}{(b+n+1)^{a+n\bar{y}+\tilde{y}}} \\ &= \frac{\Gamma(a+n\bar{y}+\tilde{y})}{\Gamma(\tilde{y}!)\Gamma(a+n\bar{y})} \left(\frac{b+n}{b+n+1}\right)^{a+n\bar{y}} \left(\frac{1}{b+n+1}\right)^{\tilde{y}} \end{aligned}$$

$$\Rightarrow \boxed{\tilde{y}|\mathbf{y} \sim \text{NBin}\left(a+n\bar{y}, \frac{1}{b+n+1}\right)}$$

2 Multiparameter Models

2.1 Normal-Inv-χ^2 (Unknown Variance)

Conjugate Prior/Posterior : To do a full Bayesian analysis of Normally-distributed data, we need to infer both θ, σ^2 rather than considering each separately. However, we can break up the joint probability by conditioning on σ^2, that is:

$$p(\theta, \sigma^2|\mathbf{y}) \propto p(\mathbf{y}|\theta, \sigma^2)p(\theta, \sigma^2) = p(\mathbf{y}|\theta, \sigma^2)p(\theta|\sigma^2)p(\sigma^2)$$

The conjugate prior for the σ^2 parameter is the **Inverse-χ^2** distribution, so that if we let $\tau_0^2 = \sigma^2/\kappa_0$:

$$\begin{aligned}\theta|\sigma^2 &\sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0)\\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)\end{aligned}$$

This yields the joint prior density:

$$p(\theta, \sigma^2) = p(\theta|\sigma^2)p(\sigma^2) \propto \frac{1}{\sigma}\exp\left[-\frac{(\theta-\mu_0)^2}{2\sigma^2/\kappa_0}\right](\sigma^2)^{-(\nu_0/2+1)}\exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right)$$

The likelihood for the exchangeable Normal observations is:

$$\begin{aligned}p(\mathbf{y}|\theta, \sigma^2) &\propto \frac{1}{\sigma^n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right]\\ &= \frac{1}{\sigma^n}\exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}(y_i-\bar{y})^2 + n(\bar{y}-\theta)^2\right)\right]\\ &= \frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y}^2 - 2\bar{y}\theta + \theta^2)\right]\right\}\end{aligned}$$

Thus, the joint posterior is:

$$\begin{aligned}p(\theta, \sigma^2|\mathbf{y}) &\propto p(\mathbf{y}|\theta, \sigma^2)p(\theta|\sigma^2)p(\sigma^2)\\ &= \frac{1}{\sigma}\exp\left[-\frac{1}{2\sigma^2/\kappa_n}(\theta-\mu_n)^2\right](\sigma^2)^{-(\frac{\nu_0+n}{2}+1)}\exp\left(-\frac{\nu_n\sigma_n^2}{2\sigma^2}\right)\\ &\propto p(\theta|\mathbf{y}, \sigma^2)p(\sigma^2|\mathbf{y})\end{aligned}$$

where:

$$\begin{aligned}\mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}\\ \kappa_n &= \kappa_0 + n\\ \nu_n &= \nu_0 + n\\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{k_0+n}(\bar{y}-\mu_0)^2\end{aligned}$$

$$\begin{aligned}\Rightarrow \theta|\mathbf{y}, \sigma^2 &\sim \mathcal{N}(\mu_n, \sigma^2/\kappa_n)\\ \sigma^2|\mathbf{y} &\sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2)\end{aligned}$$

We write this finally as:

$$\Rightarrow \boxed{\theta, \sigma|\mathbf{y} \sim \mathcal{N}\text{-Inv-}\chi^2(\mu_n, \sigma_n^2/\kappa_n; \nu_n, \sigma_n^2)}$$

Noninformative Priors : The primary noninformative prior for the joint (θ, σ^2) model is:

$$p(\theta, \sigma^2) \propto (\sigma^2)^{-1}$$

which results in a uniform distribution over $(\theta, \log \sigma)$.

Simulating from Posterior : In order to simulate from the posterior, we sample in stages:

1. Draw σ^2 from the posterior: $\sigma^2|\mathbf{y} \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2)$
2. Draw θ from posterior given σ^2: $\theta|\mathbf{y}, \sigma^2 \sim \mathcal{N}(\mu_n, \sigma^2/\kappa_n)$

2.2 Dirichlet-Multinomial

Conjugate Prior/Posterior : The Multinomial distribution is the generalization of the Binomial; suppose we have k categories, with n observations, and y being the vector of counts in each category. Then:

$$p(y|\pi) \propto \prod_{j=1}^{k} \pi_j^{y_j}$$

The conjugate prior for the Multinomial is the Dirichlet (generalization of the Beta):

$$p(\pi) \propto \prod_{j=1}^{k} \pi_j^{\alpha_j - 1}$$

with parameters $\alpha = (\alpha_1, \ldots, \alpha_k)$. That is:

$$\begin{aligned} y|\pi &\sim \text{Mult}_k(n, (\pi_1, \ldots, \pi_k)) \\ \pi &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k) \end{aligned}$$

Clearly, the posterior is then:

$$p(\pi|y) \propto p(y|\pi)p(\pi) \propto \prod_{j=1}^{k} \pi_j^{\alpha_j + y_j - 1}$$

$$\Rightarrow \boxed{\pi|y \sim \text{Dirichlet}(\alpha_1 + y_1, \ldots, \alpha_k + y_k)}$$

The $(\alpha_1, \ldots, \alpha_k)$ represent "prior observations" in each of the categories.

Noninformative Priors : There are multiple options, including:

1. Noninformative: $\alpha = (1, 1, \ldots, 1)$; this distribution yields equal probability to end vector π s.t. $\sum_{i=1}^{k} \pi_i = 1$.
2. Improper: $\alpha_j = 0$ for every j; uniform over $\log \pi_j$; the resulting posterior is proper iff every category has at least one observation.

2.3 Multivariate Normal (Known Variance)

Conjugate Prior/Posterior : The conjugate analysis of the Multivariate Normal (MVN) is similar to the univariate case. We have, for known covariance Σ:

$$\begin{aligned} y_1, \ldots, y_n|\theta, \Sigma &\sim \mathcal{N}(\theta, \Sigma) \\ \theta|\Sigma &\sim \mathcal{N}(\mu_0, \Lambda_0) \end{aligned}$$

Then the posterior for θ becomes:

$$\begin{aligned} p(\theta|\mathbf{y}, \Sigma) &\propto p(\mathbf{y}|\theta, \Sigma)p(\theta|\Sigma) \\ &\propto |\Sigma|^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta)^T\Sigma^{-1}(y_i - \theta)\right] \exp\left[-\frac{1}{2}(\theta - \mu_0)^T\Lambda_0^{-1}(\theta - \mu_0)\right] \end{aligned}$$

Again, completing the square for the quadratic form in θ, we use:

$$\Rightarrow \boxed{\theta|\mathbf{y}, \Sigma \sim \mathcal{N}(\mu_n, \Lambda_n)}$$

where:

$$\begin{aligned} \mu_n &= (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}) \\ \Lambda_n^{-1} &= \Lambda_0^{-1} + n\Sigma^{-1} \end{aligned}$$

Posterior Predictive Distribution : The posterior predictive distribution has:

$$p(\tilde{y}|\mathbf{y}, \Sigma) = \int p(\tilde{y}|\theta, \Sigma)p(\theta|\mathbf{y}, \Sigma)d\theta$$

which is exponential of quadratic in $\tilde{y}$, so $\tilde{y}$ must have MVN distribution, that is:

$$\tilde{y}|\mathbf{y} \sim \mathcal{N}(\mu_n, \Sigma + \Lambda_n)$$

2.4 Multivariate Normal (Unknown Mean/Variance)

As in the univariate case, full Bayesian analysis of the MVN model requires modeling of the Σ parameter as well. We use the Inv-Wishart distribution, which is the multivariate generalization of the Inv-χ^2 distribution, as conjugate prior for Σ. Thus:

$$\begin{aligned} y_1, \ldots, y_n|\theta, \Sigma &\sim \mathcal{N}(\theta, \Sigma) \\ \theta|\Sigma &\sim \mathcal{N}(\mu_0, \Sigma/\kappa_0) \\ \Sigma &\sim \text{Inv-Wishart}(\nu_0, \Lambda_0^{-1}) \end{aligned}$$

This is described in Gelman et al. (2014) as the "N-Inv-Wishart" (Normal Inverse-Wishart) distribution, parameterized in terms of: $(\mu_0, \kappa_0; \nu_0, \Lambda_0)$, where ν_0 is the degrees of freedom for the Inv-Wishart and Λ_0 is the scale matrix.

$$p(\theta, \Sigma) = p(\theta|\Sigma)p(\Sigma) \propto |\Sigma|^{-(\frac{\nu_0+d}{2}+1)} \exp\left[-\frac{1}{2}\text{trace}(\Lambda_0\Sigma^{-1}) - \frac{\kappa_0}{2}(\theta - \mu_0)^T\Sigma^{-1}(\theta - \mu_0)\right]$$

Using the MVN likelihood with this prior density yields:

$$\boxed{\theta, \Sigma|\mathbf{y} \sim \mathcal{N}\text{-Inv-Wishart}(\mu_n, \kappa_n; \nu_n, \Lambda_n)}$$

that is:

$$\begin{aligned} \theta|\mathbf{y}, \Sigma &\sim \mathcal{N}(\mu_n, \Sigma/\kappa_n) \\ \Sigma|\mathbf{y} &\sim \text{Inv-Wishart}(\nu_n, \Lambda_n) \end{aligned}$$

with updated hyperparameters:

$$\begin{aligned} \mu_n &= \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y} \\ \kappa_n &= \kappa_0 + n \\ \nu_n &= \nu_0 + n \\ \Lambda_n &= \lambda_0 + S + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T \end{aligned}$$

where $S = \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T$ is the sum of squares matrix.

3 Hierarchical Models

Hierarchical models provide a useful way of "learning from the whole"; that is, such models suppose that there exist underlying parameters for each of the observations/groups, which are drawn from a hyper-distribution. This allows the number of parameters to grow with the data, improving fit, while keeping the number of overall hyperparameters low, avoiding overfitting. That is:

$$\begin{aligned} y_i|\theta_i &\sim p(\cdot|\theta_i) \\ \theta_i|\alpha &\sim p(\cdot|\alpha) \\ \alpha &\sim p(\cdot) \end{aligned}$$

This sort of model is best illustrated with a concrete example, this one using the normal hierarchical model. The notation/example is drawn from Gelman et al. (2014).

Consider $j = 1, \ldots, J$ independent groups of n_j normally-distributed observations. Each group of observations, $(y_{1j}, \ldots, y_{n_j,j})$ has separate parameter θ_j and known common error variance σ^2, i.e.:

$$y_{ij}|\theta_j \sim \mathcal{N}(\theta_j, \sigma^2)$$

This yields the sufficient statistics:

$$\begin{aligned} \bar{y}_{\cdot j} &= \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij} \\ \sigma_j^2 &= \frac{\sigma^2}{n_j} \end{aligned}$$

and so the likelihood becomes:

$$\bar{y}_{\cdot j}|\theta_j \sim \mathcal{N}(\theta_j, \sigma_j^2)$$

It is noted here that the "classical" method of dealing with these issues is to consider: 1) $\theta_j = \bar{y}_{\cdot j}$, i.e. setting the expected value equal to the sufficient statistic; 2) $\theta_j = \theta = \bar{y}_{\cdot\cdot}$, i.e. use one overall pooled estimate; 3) use a linear combination $\theta_j = \lambda_j \bar{y}_{\cdot j} + (1 - \lambda_j)\bar{y}_{\cdot\cdot}$. (Note: (1) and (2) are special cases of (3) with $\lambda_j = 1, \lambda_j = 0$, respectively.)

These all arise from potential prior distributions, but the point that interests us most at the moment is setting up an actual prior on $\theta_1, \ldots, \theta_J$. We use Normal for conjugacy:

$$\theta_j|\mu, \tau^2 \sim \mathcal{N}(\mu, \tau^2)$$

where θ_j are conditionally independent (exchangeable) given μ, τ^2).

We also need a hyperprior distribution on (μ, τ); for noninformative uniform hyperprior on μ:

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

The resulting joint posterior on all the parameters and hyperparameters is:

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

We now analyze the parameters and hyperparameters separately.

1. **Conditional Posterior of Parameters**: We first consider the conditional posterior of parameters θ_j given the hyperparameters μ, τ^2. But this is exactly the Normal-Normal conjugate model with known variance, with J independent such models with n_j observations. Thus:

   $$\theta_j|\mathbf{y}, \mu, \tau \sim \mathcal{N}(\mu_j, V_j)$$

   where:

   $$\begin{aligned} \mu_j &= \frac{\frac{1}{\sigma_j^2}\bar{y}_{\cdot j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \\ V_j &= \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \end{aligned}$$

2. **Marginal Posterior of Hyperparameters**: In general, we cannot marginalize out θ in analyzing the posterior of μ, τ^2, that is:

$$p(\mu, \tau|\mathbf{y}) \propto p(\mathbf{y}|\mu, \tau)p(\mu, \tau)$$

but generally $p(\mathbf{y}|\mu, \tau)$ is not a closed-form, known density. It is, however, for the Normal hierarchical model; in fact, it is:

$$\bar{y}_{\cdot j}|\mu, \tau \sim \mathcal{N}(\mu, \sigma_j^2 + \tau^2)$$

and so the marginal posterior for the hyperparameters is:

$$p(\mu, \tau|\mathbf{y}) \propto p(\mu, \tau) \prod_{j=1}^{J} \mathcal{N}(\bar{y}_{\cdot j}|\mu, \sigma_j^2 + \tau^2)$$

3. **Posterior of μ Given τ**: While we can directly compute using the marginal posterior of μ, τ, the Normal model allows a further simplification by considering τ fixed (i.e. known variance) and $\bar{y}_{\cdot j}$ as independent "observations" of μ. Again using the single-parameter conjugacy:

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

where:

$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2} \bar{y}_{\cdot j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}}$$

$$V_\mu = \frac{1}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}}$$

that is, $\hat{\mu}$ is the precision-weighted average of the observed $\bar{y}_{\cdot j}$, and the overall precision is the sum of the observation precision values.

4. **Posterior of τ**: Finally, we can consider the posterior of τ given $\mathbf{y}$. Note that the marginal posterior from (2) splits into:

$$p(\mu, \tau|\mathbf{y}) = p(\mu|\tau, \mathbf{y})p(\tau|\mathbf{y}) \Rightarrow p(\tau|\mathbf{y}) = \frac{p(\mu, \tau|\mathbf{y})}{p(\mu|\tau, \mathbf{y})} \propto \frac{p(\mu, \tau) \prod_{j=1}^{J} \mathcal{N}(\bar{y}_{\cdot j}|\mu, \sigma_j^2 + \tau^2)}{\mathcal{N}(\mu|\hat{\mu}, V_\mu)}$$

where $p(\mu, \tau) \propto p(\tau)$ for the noninformative prior. Moreover, this identity must hold for every value of μ, since there is no μ on the left-hand side; thus, we can set $\mu = \hat{\mu}$ to cancel the denominator, yielding:

$$p(\tau|\mathbf{y}) \propto p(\tau) V_\mu^{1/2} \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} \exp\left[-\frac{1}{2}(\bar{y}_{\cdot j} - \hat{\mu})^2\right]$$

where $\hat{\mu}, V_\mu$ are given in (3). This is a complex function in τ regardless of the prior, so computational methods must be used (i.e. grid approximation).

Noninformative priors on τ can be used, i.e. $p(\tau) \propto 1$ or improper priors such as $p(\tau) \propto \frac{1}{\tau}$, which is uniform on $p(\log \tau)$. Often, an empirical Bayes approach is used for an upper bound on τ.

Computation + Simulation. We draw each hyperparameter one at a time, then draw parameters based on the hyperparameter draws:

1. Simulate $\tau|\mathbf{y}$ using inverse CDF on grid, rejection sampling, etc.
2. Simulate $\mu|\mathbf{y}, \tau$ using Normal model with $\hat{\mu}, V_\mu$ as in (3).
3. Simulate $\theta_1, \ldots, \theta_J|\mathbf{y}, \mu, \tau$ independently based on Normal model in (1), with μ_j, V_j as defined.

4 Mixture Models

Mixture models are another form of more sophisticated models that can be used to model complex and heterogeneous data while still retaining some parsimony. In general, we assume that $\mathbf{y}$ is drawn from a mixture of distributions, and that this mixture category is a latent (unobserved) variable.

4.1 Finite Mixtures, Examples

Suppose that we assume $y = (y_1, \ldots, y_n)$ is drawn from a mixture of a finite number, H, of component distributions. We assume that the h^{th} component distribution depends on a parameter vector θ_h, with λ_h denoting the proportion of the population from component h. It is generally assumed that the component distributions are all from the same parametric family of distributions, simply with different parameters. We let $z_i = (z_{i1}, \ldots, z_{iH})$ denote the latent indicator variable describing the component that observation i belongs to. That is:

$$z_{ih} = \begin{cases} 1 & \text{if observation } i \text{ in component } h \\ 0 & \text{otherwise} \end{cases}$$

$$y_i | z_{ih} = 1 \sim f(y_i|\theta_h)$$

This yields the overall marginal density of y_i as:

$$\boxed{p(y_i|\theta, \lambda) = \lambda_1 f(y_i|\theta_1) + \cdots + \lambda_H f(y_i|\theta_H)}$$

Note that given λ:

$$z_i|\lambda \sim \text{Mult}_H(1, (\lambda_1, \ldots, \lambda_H))$$

and thus the likelihood of y, z jointly is:

$$p(y, z|\theta, \lambda) = p(y|z, \theta)p(z|\lambda) = \left[\prod_{i=1}^{n}\prod_{h=1}^{H} \lambda_h^{z_{ih}}\right]\left[\prod_{i=1}^{n}\prod_{h=1}^{H} f(y_i|\theta_h)^{z_{ih}}\right] = \prod_{i=1}^{n}\prod_{h=1}^{H}[\lambda_h f(y_i|\theta_h)]^{z_{ih}}$$

For the prior on θ, λ, we could assign Normal draws for θ and Dirichlet for λ, yielding conjugacy, or use a non-informative distribution. In general, $p(\theta, \lambda) = p(\theta)p(\lambda)$ in the prior. In either case, the posterior is given by:

$$p(\theta, \lambda|y) = \int p(\theta, \lambda|y, z)p(z|y)dz \propto \int p(y, z|\theta, \lambda)p(z|y)p(\theta, \lambda)dz$$

where the likelihood $p(y, z|\theta, \lambda)$ is given as above, $p(\theta, \lambda)$ is the prior, and the integral over z is the expected value over the indicators.

Thus, an EM-based approach over the unobserved z is directly implied by the formulation, in that a convenient computational approach to simulating the posterior would follow:

1. Initialize λ, θ
2. Update λ based on expectation on z given the densities
3. Draw θ from the posterior using the updated λ

4.2 Dirichlet Process Mixtures

Finite mixtures can be extended to infinite components using the Dirichlet process. We can conceive of the mixture weights λ as a discrete probability distribution over the component distributions; thus, an infinite-component mixture is equivalent to using a probability distribution with an infinite support over the mixture components. Suppose that given component and parameter θ, the distribution of y is given by $\mathcal{K}(y|\theta)$, and the parameters θ are drawn from some distribution $P(\theta)$:

$$y|\theta \sim \mathcal{K}(\cdot|\theta)$$
$$\theta \sim P(\cdot)$$

Then, we can represent the distribution of y over the mixture distribution P as:

$$\boxed{f(y|P) = \int \mathcal{K}(y|\theta)dP(\theta)}$$

The component distributions in this case are referred to as "kernels" $\mathcal{K}(\cdot|\theta)$, while the mixture weights are represented by the "mixing measure" P.

We require a prior on P, that is:

$$P \sim \pi_{\mathcal{P}}$$

where $\mathcal{P}$ is the space of all probability measures on the parameters, and $\pi_{\mathcal{P}}$ is a prior over that space.

We obtain a "Dirichlet process mixture model" if we let $\pi_{\mathcal{P}}$ correspond to the Dirichlet process. A **Dirichlet process** (DP) is the generalization of the Dirichlet distribution such that for every finite partition $(B_1, \ldots, B_k)$ of the sample space Ω, we have:

$$(P(B_1), \ldots, P(B_k)) \sim \text{Dirichlet}(\alpha P_0(B_1), \ldots, \alpha P_0(B_k))$$

We then say that:

$$P \sim \text{DP}(\alpha P_0)$$

where α is a precision parameter and P_0 is the reference probability measure on Ω.

The convenient aspect of the DP model is that if we observe:

$$y_i | P \sim P(\cdot)$$

and $P \sim \text{DP}(\alpha P_0)$, then the posterior is:

$$P|y_1, \ldots, y_n \sim \text{DP}\left(\alpha P_0 + \sum_i \delta_{y_i}\right)$$

that is, we simply add point masses to the reference measure. In short, this means:

$$(P(B_1), \ldots, P(B_k)) \sim \text{Dirichlet}\left(\alpha P_0(B_1) + \sum_i 1(y_i \in B_1), \ldots, \alpha P_0(B_k) + \sum_i 1(y_i \in B_k)\right)$$

While DP models are often defined axiomatically (and implicitly) as above, they are most usefully conceptualized in terms of the **stick-breaking process**:

1. Draw a countably infinite set of atoms $\theta_1, \theta_2, \ldots$ from the reference distribution P_0.
2. For each i, draw $V_i \sim \text{Beta}(1, \alpha)$.
3. Define $\pi_i = V_i \prod_{j<i}(1 - V_j)$ as the probability weight on each atom θ_i.
4. The induced measure P is a draw from the DP prior, given by:

$$\boxed{P(\cdot) = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}(\cdot)}$$

Thus, the DP prior is a *discrete distribution*, except over a countably infinite support. We can then use this stick-breaking process to write the marginal probability of y as:

$$\boxed{p(y) = \sum_{i=1}^{\infty} \pi_i \mathcal{K}(y|\theta_i)}$$

where:

$$\pi_i \sim \text{Stick-breaking}(\alpha)$$
$$\theta_i \sim P_0$$

5 Simulations + MCMC Methods

In general, if we are not in the case of the strictly conjugate settings describe above (and even if we are), which usually arises in the case of sophisticated modeling using hierarchical or mixture models, we must resort to computational and simulation-based methods to work with the posterior (i.e. calculate the expected value). There are a number of methods by which to achieve this; most direct is using the inverse CDF or grid-based approximation methods to compute the posterior, but this quickly becomes cumbersome and even intractable in higher dimensions. This leads to MCMC-based methods, which do not run into dimensionality problems in most cases.

In general, methods are divided into: 1) deterministic; 2) simulation-based. In the first category are numerical integration and grid methods, whereas the second category includes rejection sampling and MCMC.

5.1 Grid-Based Approximation

In general, we can simulate an (unnormalized) posterior distribution on a grid. Suppose that we are in parameter space with two dimensions x, y. By numerical optimization, it is often straightforward to find modes for most reasonable posteriors that can be dealt with using grid-based methods. We can use a grid about the mode, with intervals depending on the precision required and the variability of the posterior.

Suppose that x^*, y^* denote the mode of the posterior $p(x, y)$; one approximation is to use

$$\left(x^* - L, x^* - \frac{L-1}{N}, \ldots, x^*, x^* + \frac{1}{N}, \ldots, x^* + L\right)$$

and equivalently for y. We then form a grid by using the array of all combinations of points, and evaluate the posterior at each point (x, y). We can then simply numerically average over the entire grid to find the expectation of any function, say $g(X, Y)$:

$$E[g(X,Y)] \approx \sum_{i=1}^{N}\sum_{i=1}^{N} g(x_i, y_i)p(x_i, y_i)$$

We can also obtain a draw from the posterior by using the **inverse CDF method**. To do so:

1. Draw $u \sim \text{Unif}[0, 1]$.
2. After sorting the probabilities $p(x_i, y_i)$ in magnitude, find the point (x_i, y_i) that yields the largest $p(x_i, y_i) \leq u$, that is:
$$(x_i, y_i) = \underset{(x,y):p(x,y)\leq u}{\arg\max}\ p(x, y)$$

Repeating the above procedure yields samples of x, y distributed according to the posterior.

5.2 Rejection Sampling

If we can find another density that dominates the desired posterior (i.e. $p(x, y) \leq Mq(x, y)$ for some finite constant M and every x, y), and we can easily sample from this density $q(x, y)$. Then, we can conduct **rejection sampling** by using:

1. Sample (x, y) from $q(x, y)$.
2. Accept (x, y) with probability $\frac{p(x,y)}{Mq(x,y)}$; that is, draw $u \sim \text{Unif}[0, 1]$ and accept (x, y) if $u \leq \frac{p(x,y)}{Mq(x,y)}$.

5.3 Importance Sampling

Related to rejection sampling is the method of **importance sampling**. Again, suppose that we can simulate from $q(x, y)$ and we are interested in the posterior expectation of some function $g(x, y)$. Then:

$$E[g(X,Y)] = \int g(x,y)p(x,y)dxdy = \frac{\int g(x,y)p(x,y)dxdy}{\int p(x,y)dxdy} = \frac{\int \frac{g(x,y)p(x,y)}{q(x,y)}q(x,y)dxdy}{\int \frac{p(x,y)}{q(x,y)}q(x,y)dxdy}$$

We can treat this as an expectation over $q(x, y)$, which we can simulate from, with weights given by:

$$w(x, y) = \frac{p(x, y)}{q(x, y)}$$

$$\Rightarrow E[g(X, Y)] = \frac{\int [w(x, y)g(x, y)]q(x, y)dxdy}{\int w(x, y)q(x, y)dxdy}$$

Thus, if we can draw (x_i, y_i) from $q(x, y)$, then we can approximate this posterior expectation:

$$E[g(x, y)] \approx \frac{\frac{1}{N}\sum_{i=1}^{N} w(x_i, y_i)h(x_i, y_i)}{\frac{1}{N}\sum_{i=1}^{N} w(x_i, y_i)}$$

5.4 Gibbs Sampling

Often the most commonly-employed simulation-based method is the Gibbs sampler, which is a special case of the Metropolis-Hastings algorithm, discussed in the next section. The Gibbs sampler is based on the premise that we can easily simulate from the "full conditional" of the posterior. Suppose that we want to simulate from the posterior of θ given observed data y, and $\theta = (\theta_1, \ldots, \theta_d)$ is a d-dimensional parameter vector. If we assume that $\theta_j|\theta_{-j}, y$ (that is, the conditional distribution of θ_j given all the other θ values and the data) is a distribution from which we can easily simulate, then we can use the following procedure:

1. Initialize some values $(\theta_1^0, \ldots, \theta_d^0)$
2. Draw θ_j^t conditional on $\theta_{-j}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1})$ (i.e. use all preceding updated values and old values for subsequent):
$$\theta_j^t \sim p(\theta_j|\theta_{-j}^{t-1}, y)$$

5.5 Metropolis-Hastings

The Metropolis-Hastings algorithm is one in which we "jump" from a sampled value to the next, and accept/reject it according to a given ratio. The next step θ^t is given by a **proposal distribution** $J_t(\theta^t|\theta^{t-1})$, and accepted according to the Hastings ratio. The algorithm is as follows:

1. Initialize θ^0.
2. Draw a proposal θ^* from the proposal distribution $J_t(\theta^*|\theta^{t-1})$.
3. Calculate the Hastings ratio:
$$r = \frac{p(\theta^*|y)/J_t(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|y)/J_t(\theta^{t-1}|\theta^*)}$$
(in effect, we compute the ratio of the densities at the current and proposed values, and normalize by the proposal distributions)
4. Accept the proposed value θ^* with probability r, and stay at the current value otherwise, that is:
$$\theta^t = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{t-1} & \text{otherwise} \end{cases}$$
(In practice, draw $u \sim \text{Unif}[0, 1]$ and accept θ^* if $u \leq r$ and use $\theta^t = \theta^{t-1}$ otherwise.)

6 Missing Data

A significant aspect of modern data analysis not covered under the "standard" Bayesian modeling repertoire is accounting for missing data, and more broadly for different methods of data collection. For example, data can arise in experiments or observational studies, with randomized or deterministically assigned units, and after or before processing. All of these aspects can play a significant role in the final statistical result, and thus must be accounted for in a "full" Bayesian analysis of the data.

6.1 Ignorability, Missing-at-Random

There do exist cases in which the "standard" machinery works perfectly fine for drawing inferences, and we can simply ignore the process by which the data was collected. This occurs under **ignorable** data collection settings, in which - as the name implies - the data analyst can ignore the process.To motivate this definition, we first consider the full analysis and see how it can be simplified under the proper conditions.

Suppose that the complete data set is given by $y = (y_{obs}, y_{mis})$, where we observe y_{obs} and do not observe y_{mis}. I is then the inclusion indicator, where $I_i = 1$ if y_i is completely observed. Let the data be governed by parameters θ, and the data inclusion process by parameters ϕ, which may or may not be the same as θ. The likelihood is then given by:

$$p(y, I|\theta, \phi) \propto p(I|y, \theta, \phi)p(y|\theta, \phi) = p(I|y, \phi)p(y|\theta)$$

Note that this "complete likelihood" is not what is observed; we only observe y_{obs}, though we observe the full I. Thus, the "observed likelihood is for (y_{obs}, I) and is given by:

$$p(y_{obs}, I|\theta, \phi) = \int p(I|y_{obs}, y_{mis}, \phi)p(y_{obs}, y_{mis}|\theta)dy_{mis}$$

The posterior is then given by:

$$p(\theta, \phi|y_{obs}, I) \propto p(y_{obs}, I|\theta, \phi)p(\theta, \phi) = p(\theta, \phi) \int p(I|y_{obs}, y_{mis}, \phi)p(y_{obs}, y_{mis}|\theta)dy_{mis}$$

Since the parameter of interest is θ (in general), the posterior of θ alone is given by:

$$\begin{aligned} p(\theta|y_{obs}, I) &= \int p(\theta, \phi|y_{obs}, I)d\phi \\ &\propto \int p(y_{obs}, I|\theta, \phi)p(\theta, \phi)d\phi \\ &= \int \int p(\theta, \phi)p(I|y_{obs}, y_{mis}, \phi)p(y_{obs}, y_{mis}|\theta)dy_{mis}d\phi \end{aligned}$$

The full integral is cumbersome, but we can simplify it under certain conditions. **Ignorability** provides exactly these conditions.

A data collection process is said to be **ignorable** under the following conditions:

1. **Missing-at-random (MAR)**: The distribution of the missing-data mechanism does not depend on the missing values themselves:
$$p(I|y_{obs}, y_{mis}, \phi) = p(I|y_{obs}, \phi)$$
2. **Distinct parameters**: The parameters for the data and the missing-data mechanism are distinct (independent in the prior):
$$p(\theta, \phi) = p(\theta)p(\phi)$$

If the data collection process is in fact ignorable, the full integral simplifies to:

$$p(\theta|y_{obs}, I) = p(\theta)p(y_{obs}|\theta) \int p(\phi)p(I|y_{obs}, \phi)d\phi \propto p(\theta|y_{obs}) \int p(\phi|y_{obs}, I)d\phi = p(\theta|y_{obs})$$

which is exactly the form of the "standard" Bayesian analysis; the data inclusion indicators I have disappeared!

6.2 Data Collection Models

Gelman et al. (2014) review a number of different data collection models that arise in common practice (Sec. 8.3 - 8.6), including surveys, experiments, and observational studies. We provide a brief overview.

Sample Surveys The problem of finite-population inference (i.e. unobserved but potentially observable, such as the population mean) and superpopulation inference (i.e. for parameters of probabilistic model that can never be observed, even in theory) occur in different sampling mechanisms. Prominent ones include:

1. **Simple random sampling**: For a finite population of N individuals, exchangeable, we let $I = (I_1, \ldots, I_N)$ be the indicators for whether i is in the sample. This data inclusion model is **strongly ignorable**, in the sense that:

$$p(I|y, \phi) = p(I) = \frac{1}{\binom{N}{n}}$$

as long as $\sum_{i=1}^{N} I_i = n$. The inference is very straightforward in this case, and the standard approach ignoring data collection can be utilized for superpopulation inference. For finite-population estimands, say the population mean $\bar{y}$, we need to average over for the observed and missing y, that is:

$$\bar{y} = \frac{n}{N}\bar{y}_{obs} + \frac{N-n}{N}\bar{y}_{mis}$$

In general, we impute/simulate the missing y_{mis} using the posterior of $\theta|y_{obs}$, which we can compute given the ignorability of the design.

2. **Stratified sampling**: Rather than treating all N individuals exchangeably, we stratify them into J strata. Within each stratum, simple random sampling is conducted, drawing n_j from stratum j. The design is also ignorable as long as we have the strata indicators, i.e. $x_1, \ldots, x_J$ with $x_j = (x_{1j}, \ldots, x_{n_j,j})$ such that:

$$x_{ij} = \begin{cases} 1 & \text{if unit } i \text{ in stratum } j \\ 0 & \text{otherwise} \end{cases}$$

For stratified samples, in which units within strata are expected to be more similar than individuals across strata, it is natural to employ a hierarchical model with a separate parameter for each stratum and a hyper-distribution over the strata parameters. For example, results of educational studies collected across schools leads to a naturally stratified sample with each school constituting a separate stratum. Proper Bayesian analysis would yield a separate parameter for each school, governed by an overall distribution whose hyperparameters are the subject of inference.

3. **Cluster sampling**: This is almost identical to the stratified sampling method, except that there are now K clusters, and a random sample of J of these clusters is drawn. Within each cluster j, another random sample of n_j individuals is drawn and observed. Analysis under this design proceeds almost identically to the stratified sampling design, and is ignorable given the indicators for how individuals in the population are segregated by cluster. One exception is that the missing clusters must now be imputed using the posterior of the hyperdistribution, and individuals within these missing clusters also wholly imputed.

Experiments Experiments are very common in data analysis, and help to clarify situations in which the design is ignorable despite considerable "non-randomness" in assignment. Also, experiments are the bedrocks of causal inference, which is intricately related to the missing data framework.

1. **Completely randomized experiment**: The complete data consist of (y_i^A, y_i^B) of outcomes for individual i under treatment A and B, respectively. We assume SUTVA - that is, the treatment of individual i has no effect on the outcomes of the other $n-1$ units. For causal inference, we are interested in $y_i^A - y_i^B$, and the superpopulation average causal effect of interest is:

$$E(y_i^A - y_i^B|\theta) = E(y_i^A|\theta) - E(y_i^B|\theta)$$

over the entire complete data, i.e. $p(y_i^A, y_i^B|\theta)$. The finite-population causal effect is: $\bar{y}^A - \bar{y}^B$. As in the random sampling model, we have a strongly ignorable design with $p(I|y,\phi) = p(I) = 1/\binom{n}{n/2}$. For full Bayesian inference, we again have to impute the missing data (i.e. the non-observed y_i^T for $T = A, B$, since we only observe one of the two outcomes for each individual) then compute the finite-population average.

2. **Sequential design**: An interesting polar opposite to the completely randomized experiment is the sequential design, in which the treatment assignment for individual i is deterministically related to the treatment assignments or outcomes of the previous individuals $1, \ldots, i-1$. The design is obviously not strongly ignorable here (i.e. $p(I|\mathbf{y},\phi) \neq p(I)$ since it can be determined by y) but it is ignorable given all the variables used for determining future allocations. For example, suppose that the assignment of i depends on whether more of the previous $i-1$ individuals were assigned to A or B; then i is assigned the minority treatment. In this case, the design is ignorable conditional on I as a covariate since the observed treatment assignments reveal everything about the future assignments.

Observational Studies While data collection methods and inference in observational studies is generally too vast an area, Gelman et al. (2014) cover principal stratification as an interesting example in which Bayesian analysis can play a significant role. Generally, observational studies can be very problematic because both treatment and treatment assignment variables are out of the analyst's control.

Principal stratification can play an interesting role when there can be **non-compliance**: that is, a subject can be assigned to a treatment but refuse it in favor of another. This is most often the case in drug-related studies, in which individuals assigned to the treatment can refuse to take the drug and receive a placebo instead (*a priori*). More generally, there can be:

- Never-takers: Individual always takes placebo; that is, he/she can be assigned to treatment but refuses to take it, whereas comply when assigned to placebo
- Always-takers: Individual is assigned to placebo but takes treatment; complies under treatment
- Defiers: Individual takes treatment when assigned to placebo and placebo when assigned to treatment
- Compliers: Individual takes treatment/placebo according to assignment

The idea of principal stratification is that we can use the observed compliance status of the individuals to define strata within the population, and proceed according to a stratified sampling paradigm. Unfortunately, it is often the case that we don't observe some of the compliance statuses of the individuals (i.e. in the treatment/placebo case, the placebo group is usually not given the option to switch to the treatment). This leads to an imputation setting in which the compliance status of the placebo (or other unobserved) group has to be imputed in finite-population inference.

References

A. Gelman *et al.* (2014). *Bayesian Data Analysis.*

P. D. Hoff (2010). *A First Course in Bayesian Statistical Methods.*