Hierarchical Models

Lin Zhang

Department of Biostatistics
School of Public Health
University of Minnesota

Previously

- We have discussed the Bayesian model with multiple unknown parameters using the normal example
  - Specifying a joint prior by assuming independence a *priori*
  - Deriving the marginal posterior for inference on the parameter of interest
  - Approximating the marginal posterior using sampling-based algorithm
  - Considering the impact of estimating nuisance parameters on the posterior for the parameter of interest
- **Now,** we consider a hierarchical prior.

Hierarchical prior for normal data

- **Likelihood:** Let $y_1, \ldots, y_n$ be iid normal random variables with mean θ and variance σ^2. The likelihood is

$$\begin{aligned} f(\mathbf{y}|\theta,\sigma^2) &= \prod_{i=1}^{n} N(y_i|\theta,\sigma^2) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{(n-1)s^2 + n(\bar{y}-\theta)^2}{2\sigma^2}\right\} \end{aligned}$$

- Consider a hierarchical prior

$$\begin{aligned} \pi(\theta,\sigma^2) &= \pi(\theta|\sigma^2)\pi(\sigma^2) \\ &= N(\theta|\mu,\sigma^2\tau^2)\cdot IG(\sigma^2|\alpha,\beta) \end{aligned}$$

  where μ, τ, α, β are known/pre-specified parameters.

- This specification results in a hierarchical model

$$\begin{aligned} y_i|\theta,\sigma^2 &\overset{iid}{\sim} N(\theta,\sigma^2) \\ \theta|\sigma^2 &\sim N(\mu,\sigma^2\tau^2) \\ \sigma^2 &\sim IG(\alpha,\beta) \end{aligned}$$

Deriving the posterior

- The joint posterior is thus
$$p(\theta,\sigma^2|\mathbf{y}) \propto f(\mathbf{y}|\theta,\sigma^2)\pi(\theta,\sigma^2) = f(\mathbf{y}|\theta,\sigma^2)\pi(\theta|\sigma^2)\pi(\sigma^2)$$
- Again, our interest is inference of θ based on its marginal posterior.
- The marginal posterior of θ can be obtained analytically
$$\begin{aligned} p(\theta|\mathbf{y}) &= \int p(\theta,\sigma^2|\mathbf{y})d\sigma^2 \\ &\propto \int f(\mathbf{y}|\theta,\sigma^2)\pi(\theta|\sigma^2)\pi(\sigma^2)d\sigma^2 \quad , \end{aligned}$$
- Or alternatively, be approximated numerically by that
$$p(\theta|\mathbf{y}) = \int p(\theta|\sigma^2,\mathbf{y})p(\sigma^2|\mathbf{y})d\sigma^2$$

Deriving the components

- First, we consider the conditional posterior of θ:

$$
\begin{aligned}
p(\theta|\sigma^2, \mathbf{y}) &= \frac{p(\theta, \sigma^2|\mathbf{y})}{p(\sigma^2|\mathbf{y})} \\
&\propto \frac{f(\mathbf{y}|\theta, \sigma^2)\pi(\theta|\sigma^2)\pi(\sigma^2)}{p(\sigma^2|\mathbf{y})} \\
&\propto N(\bar{y}|\theta, \sigma^2/n) \cdot N(\theta|\mu, \sigma^2\tau^2) \\
&\propto N\left(\frac{\frac{1}{n}\mu + \tau^2\bar{y}}{\frac{1}{n} + \tau^2}, \frac{\sigma^2}{n + \frac{1}{\tau^2}}\right)
\end{aligned}
$$

- We **drop** the prior and marginal posterior of σ^2 as they are constant as a function of θ.
- The last step is obtained due to the conjugacy of the prior conditional on σ^2.

Deriving the components

- Now, we consider the marginal posterior of σ^2:

$$
\begin{aligned}
p(\sigma^2|\mathbf{y}) &= p(\theta, \sigma^2|\mathbf{y})d\theta \\
&\propto \int f(\mathbf{y}|\theta, \sigma^2)\pi(\theta|\sigma^2)\pi(\sigma^2)d\theta \\
&\propto \int \prod_{i=1}^{n} N(y_i|\theta, \sigma^2)N(\theta|\mu, \sigma^2\tau^2)d\theta \cdot IG(\sigma^2|\alpha, \beta) \\
&\propto (\sigma^2)^{-\alpha+n/2+1} \exp\left\{-\frac{1}{\sigma^2}\left(\sum_{i=1}^{n} y_i^2 + \frac{\mu^2}{\tau^2} - \frac{(n\bar{y} + \mu/\tau^2)^2}{n + 1/\tau^2}\right)\right\}
\end{aligned}
$$

- Therefore, $p(\sigma^2|\mathbf{y}) = IG(\alpha^*, \beta^*)$ where

$$
\alpha^* = \alpha + n/2, \quad \beta^* = \sum_{i=1}^{n} y_i^2 + \frac{\mu^2}{\tau^2} - \frac{(n\bar{y} + \mu/\tau^2)^2}{n + 1/\tau^2}.
$$

Sampling-based approximation of the posterior

- We can then use the same sequential sampling algorithm described in last lecture to approximate the posterior.
- For each $i = 1, \ldots, M$,
  1. draw $\sigma^2_{(i)} \sim p(\sigma^2|\mathbf{y}) = IG(\alpha^*, \beta^*)$
  2. draw $\theta_{(i)} \sim p(\theta|\sigma^2, \mathbf{y}) = N\left(\frac{\frac{1}{n}\mu+\tau^2\bar{y}}{\frac{1}{n}+\tau^2},\ \frac{\sigma^2}{n+\frac{1}{\tau^2}}\right)$.
- Note that the Normal-InvGamma hierarchical prior is conjugate for the mean and variance of a normal distribution.
- Question: Guess what is the marginal posterior of θ with the hierarchical prior?

Interim Summary

- We have considered a hierarchical prior for a two-parameter distribution which leads to a hierarchical model.
- It is often difficult to derive the joint posterior in a hierarchical model.

Interim Summary

- We have considered a hierarchical prior for a two-parameter distribution which leads to a hierarchical model.
- It is often difficult to derive the joint posterior in a hierarchical model.
- We have discussed a general approach to approximating the posterior using conditional conjugate prior.
- The approach is based on **factorizing** the joint posterior as product of marginal and conditional posteriors.
  - First, draw hyperparameter(s) from the marginal posterior.
  - Then, draw parameter(s) from the conditional posterior.

Interim Summary

- We have considered a hierarchical prior for a two-parameter distribution which leads to a hierarchical model.
- It is often difficult to derive the joint posterior in a hierarchical model.
- We have discussed a general approach to approximating the posterior using conditional conjugate prior.
- The approach is based on **factorizing** the joint posterior as product of marginal and conditional posteriors.
  - First, draw hyperparameter(s) from the marginal posterior.
  - Then, draw parameter(s) from the conditional posterior.
- Conditional conjugacate priors are **normally** used in hierarchical model so that the conditional posteriors are in closed form!

Another Hierarchical Model

- Now, let's consider another hierarchical model with multiple parameters.

- Consider independent observations $y_1, \ldots, y_k$, each from a normal distribution with different unknown means θ_i and a common known variance σ^2, and the unknown means are assume to come from a common population i.e.

$$\begin{aligned} y_i|\theta_i &\overset{ind}{\sim} N(\theta_i, \sigma^2),\ i = 1, \ldots, k, \quad \sigma^2 \text{ known };\\ \theta_i &\overset{iid}{\sim} N(\mu, \tau^2),\ i = 1, \ldots, k, \quad (\mu, \tau^2) \text{ both unknown} \end{aligned}$$

Another Hierarchical Model

- Now, let's consider another hierarchical model with multiple parameters.

- Consider independent observations $y_1, \ldots, y_k$, each from a normal distribution with different unknown means θ_i and a common known variance σ^2, and the unknown means are assume to come from a common population i.e.

$$y_i|\theta_i \overset{ind}{\sim} N(\theta_i, \sigma^2),\ i = 1, \ldots, k, \quad \sigma^2 \text{ known };$$
$$\theta_i \overset{iid}{\sim} N(\mu, \tau^2),\ i = 1, \ldots, k, \quad (\mu, \tau^2) \text{ both } \underline{\text{un}}\text{known}$$

- This is essentially a random effects model.

Motivation

- This hierarchical modeling naturally arises from many applications:
  - Cardiac treatments within hospital
  - Same cancer treatment with different tumor sites/genetic markers (basket trails)
  - Tobacco control interventions in different sub-populations
- In these applications, the individual-level parameters are considered similar and can be viewed as a sample from a common population.
- Our interest lie in the population-level parameters or individual subgroups.
- The specification of hierarchical model allows us to borrow strength across subgroups, resulting in more efficient estimation.

Specifying the Hyperprior

- We have defined the hierarchy:

$$\begin{array}{rccll} \textbf{Likelihood:} & y_i|\theta_i & \stackrel{ind}{\sim} & N(\theta_i, \sigma^2), & \sigma^2 \text{ known} \\ \textbf{Prior:} & \theta_i & \stackrel{iid}{\sim} & N(\mu, \tau^2), & \mu, \tau^2 \text{ unknown} \end{array}$$

  Note that μ and τ^2 affects y_i only through θ_i.

- We need to specify a hyperprior for the two unknown hyperparameters, μ and τ^2.

- Consider the improper (flat) hyperprior

$$\pi(\mu, \tau^2) = 1.$$

- This results in a multi-dimensional posterior

$$\begin{aligned} p(\boldsymbol{\theta}, \mu, \tau^2|\mathbf{y}) &\propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mu, \tau^2)\pi(\mu, \tau^2) \\ &= \prod_{i=1}^{k} N(y_i|\theta_i, \sigma^2) \cdot \prod_{i=1}^{k} N(\theta_i|\mu, \theta) \cdot 1 \end{aligned}$$

Approximating the posterior

- It is NOT easy to derive the multi-dimensional posterior.
- But again, we can approximate the posterior numerically by factorizing the joint posterior as we did previously:

$$\begin{aligned} p(\boldsymbol{\theta}, \mu, \tau^2|\mathbf{y}) &= p(\boldsymbol{\theta}|\mathbf{y}, \mu, \tau^2)p(\mu, \tau^2|\mathbf{y}) \\ &= \prod_{i=1}^{k} p(\theta_i|y_i, \mu, \tau^2)p(\mu, \tau^2|\mathbf{y}) \end{aligned}$$

  The last step is due to the conditional independence of θ_i given μ and τ^2.
- Now, we only need two components
  - $p(\theta_i|y_i, \mu, \tau^2)$: the conditional posterior of θ_i given the data and hyperpamaeters
  - $p(\mu, \tau^2|\mathbf{y})$: the marginal posterior of the hyperparameters

  Each is a univariate or bi-variate distribution!

Conditional posterior of θ_i

- The conditional posterior of θ_i conditional on other parameters is
$$\begin{aligned} p(\theta_i|y_i,\mu,\tau^2) &\propto f(y_i|\theta_i,\mu,\tau^2)\pi(\theta_i|\mu,\tau^2) \\ &= N(y_i|\theta_i,\sigma^2)N(\theta_i|\mu,\tau^2) \end{aligned}$$
- The derivation does not involve the hyperprior because it is constant with respect to θ_i.
- Again, we have a conjugate normal-normal model for each θ_i conditional on the hyperparameters!
- Therefore, the conditional posterior for each θ_i is easily obtained:
$$\theta_i|y_i,\mu,\tau^2 \sim N\left(\frac{\sigma^2\mu+\tau^2 y_i}{\sigma^2+\tau^2},\ \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)$$

Marginal posterior of the hyperparameters

- The marginal posterior of hyperparameters is

$$
\begin{aligned}
p(\mu, \tau^2 | \mathbf{y}) &\propto m(\mathbf{y}|\mu, \tau^2) \cdot \pi(\mu, \tau^2) \\
&= \prod_{i=1}^{k} m(y_i|\mu, \tau^2) \cdot \pi(\mu, \tau^2) \\
&= \prod_{i=1}^{k} \int \underline{N(y_i|\theta_i, \sigma^2) N(\theta_i|\mu, \tau^2) d\theta_i} \cdot \pi(\mu, \tau^2) \\
&= \prod_{i=1}^{k} N(y_i|\mu, \sigma^2 + \tau^2) \cdot 1
\end{aligned}
$$

$\Rightarrow$ This is equivalent to the posterior from a Bayesian model with n iid observations from $N(\mu, \sigma^2 + \tau^2)$ and a flat prior $\pi(\mu, \tau^2) = 1$.

Marginal posterior of the hyperparameters

- The marginal posterior of (μ, τ^2) is NOT in closed-form.
- But we can further factorize it into
$$p(\mu, \tau^2|\mathbf{y}) = p(\mu|\mathbf{y}, \tau^2)p(\tau^2|\mathbf{y}),$$
- The posterior of μ conditional on τ^2 is easily seen as
$$\mu|\mathbf{y}, \tau^2 \sim N\left(\bar{y}, \frac{\sigma^2 + \tau^2}{k}\right)$$
- The posterior distribution of τ^2 is
$$p(\tau^2|\mathbf{y}) \propto (\sigma^2 + \tau^2)^{-\frac{k-1}{2}} \exp\left\{-\frac{(k-1)s^2}{2(\sigma^2 + \tau^2)}\right\}$$
  τ^2 does NOT have a closed-form posterior, but $\sigma^2 + \tau^2$ does!

  $\Rightarrow \sigma^2 + \tau^2|\mathbf{y} \sim IG(\frac{k-3}{2}, \frac{(k-1)s^2}{2})$

Approximating the joint posterior

- We now can approximate the joint posterior given the factorization

$$p(\boldsymbol{\theta}, \mu, \tau^2 | \mathbf{y}) = \prod_{i=1}^{k} p(\theta_i | y_i, \mu, \tau^2) p(\mu | \mathbf{y}, \tau^2) p(\tau^2 | \mathbf{y})$$

- For each $t = 1, \ldots, M$,
  1. draw $\tau^2_{(t)}$ conditional on $\mathbf{y}$ from its marginal posterior.
  2. draw $\mu_{(t)}$ conditional on τ^2 and $\mathbf{y}$ from $p(\mu | \mathbf{y}, \tau^2)$
  3. draw $\theta_{i(t)}$'s conditional on μ, τ^2 and $\mathbf{y}$ **independently** from their conditional posteriors.

- We are lucky so far that we were able to draw hyperparameters directly somehow from a closed-form marginal posterior. **However,** it normally is not the case.
  - To be discussed next week ...

Bayesian Prediction

- In the above hierarchical model, two posterior predictive distributions of interest:
  - y^* for an existing θ_i
  - y^* for a "future" θ_{i*}
- In the first case, we use same sampling algorithm approximating the posterior predictive distribution, with one extra step
  4. draw "future" observation y^* conditional on the existing θ_i from the likelihood model
- In the second case, we need two extra steps
  4. draw "future" individual θ_{i*} conditional on μ and τ^2 from the population distribution $\pi(\theta_{i*}|\mu, \tau^2)$
  5. draw "future" observation y^* conditional on the new θ_{i*} from the likelihood model

Illustration: Morris' Baseball Data

i	player	y_i	i	player	y_i
1	Clemente	.400	10	Swoboda	.244
2	F. Robinson	.378	11	Unser	.222
3	F. Howard	.356	12	Williams	.222
4	Johnstone	.333	13	Scott	.222
5	Berry	.311	14	Petrocelli	.222
6	Spencer	.311	15	E. Rodriguez	.222
7	Kessinger	.289	16	Campaneris	.200
8	L. Alvarado	.267	17	Munson	.178
9	Santo	.244	18	Alvis	.156

For players $i = 1, \ldots, 18$,

$$\begin{aligned} y_i &= \text{batting average after first 45 at bats in 1970,} \\ \theta_i &= \text{true 1970 batting ability (measured by final 1970 averages)} \end{aligned}$$

Illustration: Morris' Baseball Data

- Model: For $i = 1, \ldots, 18$

$$\begin{aligned} y_i|\theta_i &\sim N(\theta_i, \sigma^2 = 0.6^2) \\ \theta_i|\mu, \tau^2 &\sim N(\mu, \tau^2) \\ \pi(\mu, \tau^2) &= 1 \end{aligned}$$

- Data: $\bar{y} = .265$ and $s^2 = 0.0048$.

- Result: next slide ...

Illustration: Morris' Baseball Data

i	player	θ_i	y_i	2.5%	Median	97.5%
1	Clemente	.346	.400	0.239	0.309	0.418
2	F. Robinson	.298	.378	0.234	0.299	0.403
3	F. Howard	.276	.356	0.227	0.294	0.391
4	Johnstone	.222	.333	0.219	0.286	0.380
5	Berry	.273	.311	0.209	0.279	0.365
6	Spencer	.270	.311	0.211	0.279	0.367
7	Kessinger	.263	.289	0.200	0.272	0.355
8	L. Alvarado	.210	.267	0.190	0.266	0.343
9	Santo	.269	.244	0.178	0.259	0.332
10	Swoboda	.230	.244	0.177	0.259	0.330
11	Unser	.264	.222	0.166	0.253	0.322
12	Williams	.256	.222	0.164	0.253	0.321
13	Scott	.303	.222	0.165	0.252	0.321
14	Petrocelli	.264	.222	0.165	0.253	0.322
15	E. Rodriguez	.226	.222	0.166	0.253	0.323
16	Campaneris	.285	.200	0.152	0.246	0.313
17	Munson	.316	.178	0.143	0.240	0.304
18	Alvis	.200	.156	0.130	0.231	0.298

Illustration: Morris' Baseball Data

- Note that the usual MLE estimator is $\widehat{\theta}_i^{MLE} = y_i$.
- The Bayesian point estimator $\widehat{\theta}_i^B$ is "shrunk back" toward the grand mean $\bar{y}$ from its original MLE estimator y_i.
  - Intuitively, shinkage makes sense here: problems are independent, but similar.
- The amount of shrinkage is controlled by the estimated heterogeneity in the data.
  - The smaller τ^2 is relative to σ^2, the closer $\widehat{\theta}_i^B$ is to $\hat{\mu}^B = \bar{y}$.
- $\widehat{\theta}_i^B$ have better performance than $\widehat{\theta}_i^{MLE}$:
  - individually: in 16 of the 18 cases, $(\widehat{\theta}_i^{PEB} - \theta_i)^2 < (y_i - \theta_i)^2$
  - overall: aggregate MSE numbers are:
    - ◊ $MSE(\mathbf{y}) = \sum_{i=1}^{18}(y_i - \theta_i)^2 = .075$
    - ◊ $MSE(\widehat{\boldsymbol{\theta}}^{PEB}) = \sum_{i=1}^{18}(\hat{\theta}_i^{PEB} - \theta_i)^2 = .022$

Summary of the hierarchical model

- This is a classical example where we have multiple studies/experiments with similar endpoints, etc.
- Hierarchical modeling "shrunk" the point estimates to a common population mean.
  - i.e. We borrow strength across individual studies, resulting in more efficient estimation $\Rightarrow$ less variance and MSE.
- This is an example of bias-variance trade-off.
  - Independent analyses are unbiased but inefficient.
  - Hierarchical model introduces bias but is far more efficient.

Adding another level of hyperprior?

- We have looked at the **3-level** hierarchical model

$$\begin{aligned} y_i|\theta_i &\sim N(\theta_i, \sigma^2) \\ \theta_i|\mu, \tau^2 &\sim N(\mu, \tau^2) \\ \pi(\mu, \tau^2) &= 1 \end{aligned}$$

- We can continually add randomness as we move down the hierarchy. For example, test scores Y_{ijk} for student k in classroom j of school i :

$$\begin{aligned} y_{ij}|\theta_{ij} &\sim N(\theta_{ij}, \sigma^2) \\ \theta_{ij}|\mu_i &\sim N(\mu_i, \tau^2) \\ \mu_i|\lambda, \kappa^2 &\sim N(\lambda, \kappa^2) \\ \pi(\lambda, \kappa^2) &= 1 \end{aligned}$$

- Subtle changes to levels near the top are **NOT** likely to have much of an impact on the bottom or data level.