Statistical Inference

Classical and Bayesian Methods

Class 16

AMS-UCSC

Tu Mar 13, 2012

Topics

We will talk about...

1. Introduction to Hierarchical Models

Posterior distributions

Consider the joint prior: $p(\phi, \theta) = p(\phi)p(\theta|\phi)$. ϕ are the model hyperparameters. θ are the model parameters.
The joint posterior is given by:

$$\begin{aligned} p(\phi, \theta|y) &\propto p(\phi, \theta)p(y|\phi, \theta) \\ &= p(\phi, \theta)p(y|\theta) \end{aligned}$$

the sampling model only depends on θ. The hyperparameters ϕ affect y through θ.
Previously the hyperparameters were considered known. Now we take into account the uncertainty on ϕ.

Predictive distributions

We might be interested in the following quantities:

- Distribution of future observations $\tilde{y}$ given $\theta_j's$. We can simulate $\tilde{y}$ based in the posterior distribution of θ_j.
- Distribution of future observations $\tilde{y}$ corresponding to future values of θ_j ($\tilde{\theta}$). In this case we can simulate $\tilde{\theta}$ conditional on the posterior simulation of ϕ and then the values of $\tilde{y}$ are simulated given the simulated values of $\tilde{\theta}$.

Marginal and conditional distributions

- 1. Write $p(\theta, \phi|y)$ in an unormalized form. This implies to calculate:

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

- Determine $p(\theta|\phi, y)$ analytically give hyperparameters ϕ.
- 2. Determine ϕ finding the posterior marginal (Bayesian paradigm). This implies to find the integral

$$p(\phi|y) = \int p(\theta, \phi|y)d\theta.$$

For some models the following formula can be used:

$$p(\phi|y) = \frac{p(\theta, \phi|y)}{p(\theta|\phi, y)},$$

- 3. Simulate the hyperparameters ϕ from the marginal $p(\phi|y)$.

Marginal and conditional distributions

- 4. Simulate θ from $p(\theta|\phi, y)$. We can consider $p(\theta|\phi, y) = \prod_j p(\theta_j|\phi, y)$. The components of θ_j can be simulated independently one at a time.
- 5. Simulate predictive values $\tilde{y}$ from the posterior predictive distribution given the values of θ. Depending on the problem it can be necessary to simulate $\tilde{\theta}$ given ϕ as previously discussed.

The previous steps are repeated L time to get L samples from al parameters.

Educational testing example

We have data from a normal distribution with a different means for each group or experiment; observational variance is known and a normal distribution is assumed for the mean of each group. This model is known as the *one way normal model with random effects*.

Example 5.5 (GCSR)

A study is carried out to investigate the impacts of a special coaching program on the test scores of SAT-V (Scholastic Aptitude Test-Verbal) for 8 schools. Test are applied to more than 30 students at each school. The response variable is the score of the test.

Educational testing example

Data structure: J independent experiments are assumed. Parameter θ_j measures the impact of the coaching program on school j,from n_j observations y_{ij}, assumed independent and normally distributed with error variance σ^2 known; this is:

$$y_{ij}|\theta_j \sim N(\theta_j, \sigma^2), \quad \text{para} \;\; i = 1, \dots, n_j; \;\; j = 1, \dots, J$$

Let $\bar{y}_{.j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$ the sample mean of each group and $\sigma_j^2 = \frac{\sigma^2}{n_j}$ the sample variance of group j. The likelihood can be written in terms of $\bar{y}_{.j}$ such that $\bar{y}_{.j} \sim N(\theta_j, \sigma_j^2)$.

θ_j could be estimated from $\bar{y}_{.j}$ which is the average result for group j or a common weighted average could be used:

$$\bar{y}_{..} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}\bar{y}_{.j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}}$$

Educational testing example

What is a good estimator for parameters $\theta_1, \ldots, \theta_J$?. The traditional method to answer this question uses an analysis of variance with an F test to proof if the means are different. If $n_j = n$ and $\sigma_j^2 = \sigma^2$ for all j we have the ANOVA table 1. If the ratio of MS between groups and MS within groups is significantly greater than 1, then $\hat{\theta}_j = \bar{y}_{.j}$. Otherwise $\hat{\theta}_j = \bar{y}_{..}$.

Table: Classic ANOVA table for the one way model

	df	SS	MS	$E(MS \mid \sigma^2, \tau)$
Between Groups	$J-1$	$\sum_i \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2$	SS$/(J-1)$	$n\tau^2 + \sigma^2$
Within Groups	$J(n-1)$	$\sum_i \sum_j (y_{ij} - \bar{y}_{.j})^2$	SS$/((J(n-1))$	σ^2
Total	$Jn-1$	$\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$	SS$/(nJ-1)$	

Educational testing example

Another alternative:

$$\hat{\theta}_j = \lambda_j \bar{y}_{.j} + (1 - \lambda_j)\bar{y}_{..}$$

where λ_j lies between 0 and 1. We can or not combine all data

Educational testing example

Hierarchical model

Parameters θ_i are assumed samples from a Normal distribution with hyperparameters (μ, τ),

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

A non-informative prior for the hyperparameters is given by:

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

(the prior density for μ is uniform).

Educational testing example

Joint posterior distribution:

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

Educational testing example

Conditional posteriors

The $\theta_j's$ are conditionally independent given (μ, τ) and the rest of terms depending on y and σ_j can be ignored because they are known.
We have J independent normal means, therefore:

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

where

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

.
Note that $\hat{\theta}_j$ and V_j are functions of μ, τ and the data.

Educational testing example

Marginal posterior distribution

For the hyperparameters one can write:

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

The marginal distributions of $\bar{y}_{.j}$ (group means averaged over θ are independent normals:

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

Then the marginal posterior is:

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

Educational testing example

From this equation we can find:

- Posterior distribution of μ conditional on τ, by factorizing

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

where $p(\mu|\tau, y)$ is the posterior distribution of μ when τ is known. From the posterior distribution $p(\mu, \tau|y)$ it is found that the log is a quadratic function of μ, therefore $p(\mu|\tau, y)$ is a normal distribution. With a uniform prior for $p(\mu|\tau)$ we have:

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

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

- The posterior distribution of τ, is:

$$p(\tau|y) = \frac{p(\mu, \tau|y)}{p(\mu|\tau, y)}$$

Educational testing example

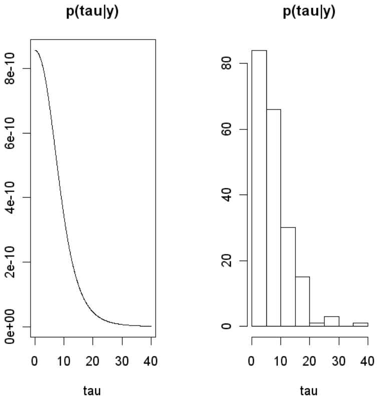
- A uniform prior $p(\tau) \propto 1$ produces a proper posterior prior.
- A posterior prior $p(\log\ \tau) \propto 1$produces an improper prior.
- If a variance estimate τ is available and an upper bound for τ is known, it is possible to find a prior from an inverse-χ^2 trying to match the distribution mean and the upper bound with the 99% quantile.

Educational testing example

Posterior simulations

The following factorization can be used:

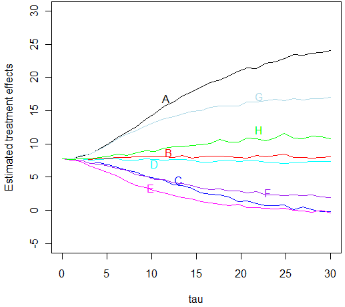
$$p(\theta, \mu, \tau | y) \propto p(\tau | y) p(\mu | \tau, y) p(\theta | \mu, \tau, y)$$



Educational testing example

Simulation from τ can be done from a uniform grid in the τ values with the function $p(\tau|y)$. μ and θ can be simulated from the corresponding normals. Density and histogram plots of the marginal marginal posterior distribution for τ, and the effect expected values conditional on τ are shown based in 5,000 samples.

Educational testing example



It can be observed that τ values close to zero are more plausible, and the effect among schools are very similar when τ is small. When τ increases, (greater variability among schools) effect estimates are far apart from each other.

Educational testing example

Simulation from posterior predictive distributions

Given samples from the posterior distribution the options are:

- Future observations $\tilde{y}$ with means $\theta = (\theta_1, \ldots, \theta_J)$. In this case to get samples for future observation $\tilde{y}$, get samples from $p(\theta, \mu, \tau | y)$ first, and then samples from $y_{ij} \sim N(\theta_j, \sigma^2)$.
- Future observations $\tilde{y}$ from $\tilde{J}$ future values with means $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_{\tilde{J}})$. In this case we can specify $\tilde{J}$ future individual sample size $\tilde{n}_j$. The simulation steps are as follows:
  - Simulate (μ, τ) from the posterior.
  - Simulate $\tilde{J}$ new parameter values $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_{\tilde{J}})$ from the population distribution $p(\tilde{\theta}_j | \mu, \tau)$ which is the θ prior distribution given the hyperparameters.
  - Simulate $\tilde{y}$ given $\tilde{\theta}$ from the sampling distribution

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

Thanks for your attention ...