Lecture 5: Introduction to Hierarchical Modelling, Empirical Bayes, and MCMC

Merlise Clyde

September 15

Normal Means Model

Suppose we have normal data with

$$Y_i \overset{iid}{\sim} (\mu_i, \sigma^2)$$

- separate mean for each observation!

Question: How can we possibly hope to estimate all these μ_i? One y_i per μ_i and n observations!

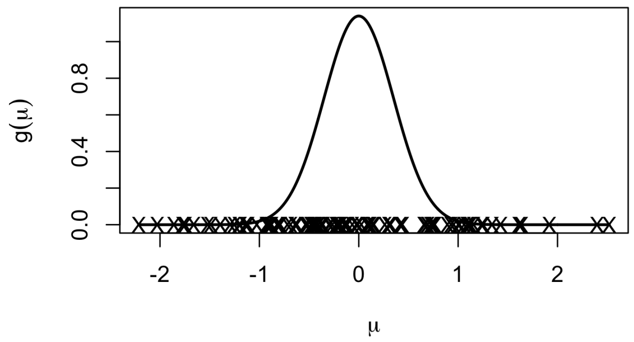
Naive estimator: just consider only using y_i in estimating and not the other observations.

- MLE $\hat{\mu}_i = y_i$

Hierarchical Viewpoint: Let's borrow information from other observations!

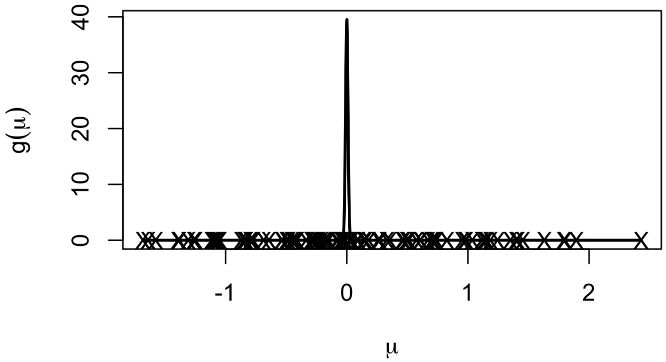
Motivation

- Example y_i is difference in gene expression for the i^{th} gene between cancer and control lines
- may be natural to think that the μ_i arise from some common distribution, $\mu_i \overset{iid}{\sim} g$



- unbiased but high variance estimators of μ_i based on one observation!

Low Variability



- little variation in μ_is so a better estimate might be $\bar{y}$
- Not forced to choose either - what about some weighted average between y_i and $\bar{y}$?

Simple Example

Data Model

$$Y_i \mid \mu_i, \sigma^2 \stackrel{iid}{\sim} (\mu_i, \sigma^2)$$

Means Model

$$\mu_i \mid \mu, \tau \stackrel{iid}{\sim} (\mu, \sigma^2_\mu)$$

- not necessarily a prior!
- Now estimate μ_i (let $\phi = 1/\sigma^2$ and $\phi_\mu = 1/\sigma^2_\mu$)
- Calculate the "posterior" $\mu_i \mid y_i, \mu, \phi, \phi_\mu$

Hiearchical Estimates

- Posterior: $\mu_i \mid y_i, \mu, \phi, \phi_\mu \overset{ind}{\sim} \textsf{N}(\tilde{\mu}_i, 1/\tilde{\phi}_\mu)$
- estimator of μ_i weighted average of data and population parameter μ

$$\tilde{\mu}_i = \frac{\phi_\mu \mu + \phi y_i}{\phi_\mu + \phi} \qquad\qquad \tilde{\phi}_\mu = \phi + \phi_\mu$$

- if ϕ_μ is large relative to ϕ all of the μ_i are close together and benefit by borrowing information
- in limit as $\sigma^2_\mu \to 0$ or $\phi_\mu \to \infty$ we have $\tilde{\mu}_i = \mu$ (all means are the same)
- if ϕ_μ is small relative to ϕ little borrowing of information
- in the limit as $\phi_\mu \to 0$ we have $\tilde{\mu}_i = y_i$

Bayes Estimators and Bias

Note: you often benefit from a hierarchical model, even if its not obvious that the μ_is are related!

- The MLE for the μ_i is just the sample y_i.
- y_i is unbiased for μ_i but can have high variability!
- the posterior mean is actually biased.
- Usually through the weighting of the sample data and prior, Bayes procedures have the tendency to pull the estimate of μ_i toward the prior or **shrinkage** mean.
- Why would we ever want to do this? Why not just stick with the MLE?
- MSE or Bias-Variance Tradeoff

Modern relevance

- The fact that a biased estimator would do a better job in many estimation/prediction problems can be proven rigorously, and is referred to as **Stein's paradox**.
- Stein's result implies, in particular, that the sample mean is an *inadmissible* estimator of the mean of a multivariate normal distribution in more than two dimensions -- i.e. there are other estimators that will come closer to the true value in expectation.
- In fact, these are Bayes point estimators (the posterior expectation of the parameter μ_i).
- Most of what we do now in high-dimensional statistics is develop biased estimators that perform better than unbiased ones.
- Examples: lasso regression, ridge regression, various kinds of hierarchical Bayesian models, etc.

Population Parameters

- we don't know μ (or σ^2 and σ^2_μ for that matter)

- Find marginal likelihood $\mathcal{L}(\mu, \sigma^2, \sigma^2_\mu)$ by integrating out μ_i with respect to g

$$\mathcal{L}(\mu, \sigma^2, \sigma^2_\mu) \propto \prod_{i=1}^{n} \int \textsf{N}(y_i; \mu_i, \sigma^2) \textsf{N}(\mu_i; \mu, \sigma^2_\mu) \, d\mu_i$$

- Product of predictive distributions for $Y_i \mid \mu, \sigma^2, \sigma^2_\mu \overset{iid}{\sim} \textsf{N}(\mu, \sigma^2 + \sigma^2_\mu)$

$$\mathcal{L}(\mu, \sigma^2, \sigma^2_\mu) \propto \prod_{i=1}^{n} (\sigma^2 + \sigma^2_\mu)^{-1/2} \exp\left\{ - \frac{1}{2} \frac{(y_i - \mu)^2}{\sigma^2 + \sigma^2_\mu} \right\}$$

- Find MLE's

MLEs

$$\mathcal{L}(\mu, \sigma^2, \sigma^2_\mu) \propto (\sigma^2 + \sigma^2_\mu)^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^2 + \sigma^2_\mu} \right\}$$

- MLE of μ: $\hat{\mu} = \bar{y}$
- Can we say anything about σ^2_μ? or σ^2 individually?
- MLE of $\sigma^2 + \sigma^2_\mu$ is

$$\widehat{\sigma^2 + \sigma^2_\mu} = \frac{\sum(y_i - \bar{y})^2}{n}$$

- Assume σ^2 is known (say 1)

$$\hat{\sigma}^2_\mu = \frac{\sum(y_i - \bar{y})^2}{n} - 1$$

Empirical Bayes Estimates

- plug in estimates of hyperparameters into the prior and pretend they are known
- resulting estimates are known as Empirical Bayes
- underestimates uncertainty
- Estimates of variances may be negative - constrain to 0 on the boundary)
- Fully Bayes would put a prior on the unknowns

Bayes and Hierarchical Models

- We know the conditional posterior distribution of μ_i given the other parameters, lets work with the marginal likelihood $\mathcal{L}(\theta)$

- need a prior $\pi(\theta)$ for unknown parameters are $\theta = (\mu, \sigma^2, \sigma^2_\mu)$ (details later)

Posterior

$$\pi(\theta \mid y) = \frac{\pi(\theta)\mathcal{L}(\theta)}{\int_\Theta \pi(\theta)\mathcal{L}(\theta)\, d\theta} = \frac{\pi(\theta)\mathcal{L}(\theta)}{m(y)}$$

- Except for simple cases (conjugate models) $m(y)$ is not available analytically

Large Sample Approximations

- Appeal to BvM (Bayesian Central Limit Theorem) and approximate $\pi(\theta \mid y)$ with a Gaussian distribution centered at the posterior mode $\hat{\theta}$ and asymptotic covariance matrix

$$V_\theta = \left[-\frac{\partial^2}{\partial\theta\partial\theta^T} \{\log(\pi(\theta)) + \log(\mathcal{L}(\theta))\} \right]^{-1}$$

- we can try to approximate $m(y)$ but this may involve a high dimensional integral

- Laplace approximation to integral (also large sample)

Stochastic methods

Stochastic Integration

$$\mathsf{E}[h(\theta) \mid y] = \int_\Theta h(\theta)\pi(\theta \mid y)\, d\theta \approx \frac{1}{T}\sum_{t=1}^{T} h(\theta^{(t)}) \qquad \theta^{(t)} \sim \pi(\theta \mid y)$$

what if we can't sample from the posterior but can sample from some distribution $q()$

$$\mathsf{E}[h(\theta) \mid y] = \int_\Theta h(\theta)\frac{\pi(\theta \mid y)}{q(\theta)} q(\theta)\, d\theta \approx \frac{1}{T}\sum_{t=1}^{T} h(\theta^{(t)})\frac{\pi(\theta^{(t)} \mid y)}{q(\theta^{(t)})}$$

where $\theta^{(t)} \sim q(\theta)$

Without the denominator in $\pi(\theta \mid y)$ we just have $\pi(\theta \mid y) \propto \pi(\theta)\mathcal{L}(\theta)$

- use twice for numerator and denominator

Important Sampling Estimate

Estimate of $m(y)$

$$m(y) \approx \frac{1}{T} \sum_{t=1}^{T} \frac{\pi(\theta^{(t)})\mathcal{L}(\theta^{(t)})}{q(\theta^{(t)})} \qquad \theta^{(t)} \sim q(\theta)$$

$$\mathsf{E}[h(\theta) \mid y] \approx \frac{\sum_{t=1}^{T} h(\theta^{(t)}) \frac{\pi(\theta^{(t)})\mathcal{L}(\theta^{(t)})}{q(\theta^{(t)})}}{\sum_{t=1}^{T} \frac{\pi(\theta^{(t)})\mathcal{L}(\theta^{(t)})}{q(\theta^{(t)})}} \qquad \theta^{(t)} \sim q(\theta)$$

$$\mathsf{E}[h(\theta) \mid y] \approx \sum_{t=1}^{T} h(\theta^{(t)}) w(\theta^{(t)}) \qquad \theta^{(t)} \sim q(\theta)$$

with un-normalized weights $w(\theta^{(t)}) \propto \frac{\pi(\theta^{(t)})\mathcal{L}(\theta^{(t)})}{q(\theta^{(t)})}$

(normalize to sum to 1)

Markov Chain Monte Carlo (MCMC)

- Typically $\pi(\theta)$ and $\mathcal{L}(\theta)$ are easy to evaluate

How do we draw samples only using evaluations of the prior and likelihood in higher dimensional settings?

- construct a Markov chain $\theta^{(t)}$ in such a way the the stationary distribution of the Markov chain is the posterior distribution $\pi(\theta \mid y)$!

$$\theta^{(0)} \xrightarrow{k} \theta^{(1)} \xrightarrow{k} \theta^{(2)} \cdots$$

- $k_t(\theta^{(t-1)}; \theta^{(t)})$ transition kernel

- initial state $\theta^{(0)}$

- choose some nice k_t such that $\theta^{(t)} \to \pi(\theta \mid y)$ as $t \to \infty$

- biased samples initially but get closer to the target

Metropolis Algorithm (1950's)

- Markov chain $\theta^{(t)}$
- propose $\theta^* \sim g(\theta^{(t-1)})$ where $g()$ is a symmetric distribution centered at $\theta^{(t-1)}$
- set $\theta^{(t)} = \theta^*$ with some probability
- otherwise set $\theta^{(t)} = \theta^{(t-1)}$

Acceptance probability is

$$\alpha = \min\left\{1, \frac{\pi(\theta^*)\mathcal{L}(\theta^*)}{\pi(\theta^{(t-1)})\mathcal{L}(\theta^{(t-1)})}\right\}$$

- ratio of posterior densities where normalizing constant cancels!

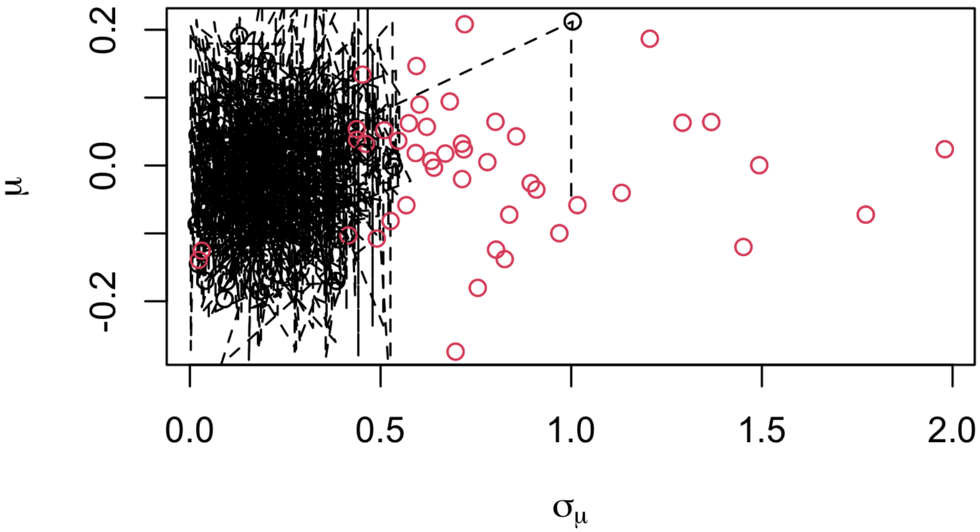
Example

- Let's use a prior for $p(\mu) \propto 1$
- Posterior for $\mu \mid \sigma^2, \sigma^2_\mu$ is $\textsf{N}\left(\bar{y}, \frac{\sigma^2+\sigma^2_\mu}{n}\right)$

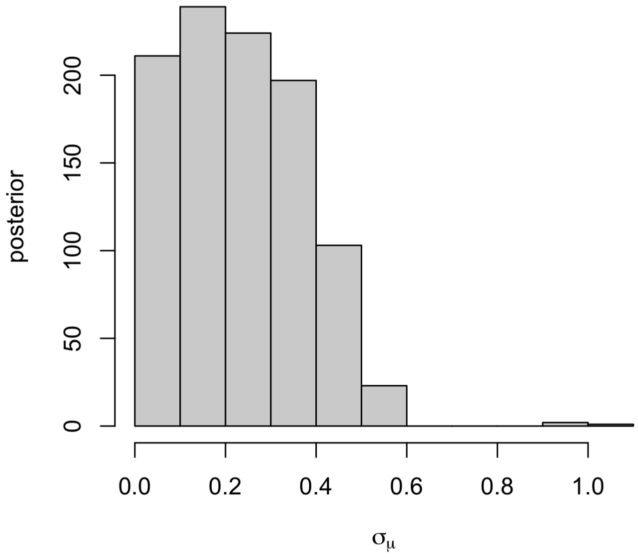
$$\mathcal{L}(\sigma^2, \sigma^2_\tau) \propto (\sigma^2 + \sigma^2_\mu)^{-\frac{n-1}{2}} \exp\left\{ -\frac{1}{2} \sum_i \frac{(y_i - \bar{y})^2}{\sigma^2 + \sigma^2_\mu)} \right\}$$

- Take $\sigma^2 = 1$
- Use a $\textsf{Cauchy}(0, 1)$ prior on σ_μ
- Symmetric proposal for σ_τ? Try a normal with variance $\frac{2.4^2}{d}\textsf{var}(\sigma_\mu)$ where d is the dimension of θ (d = 1)

Joint Posterior

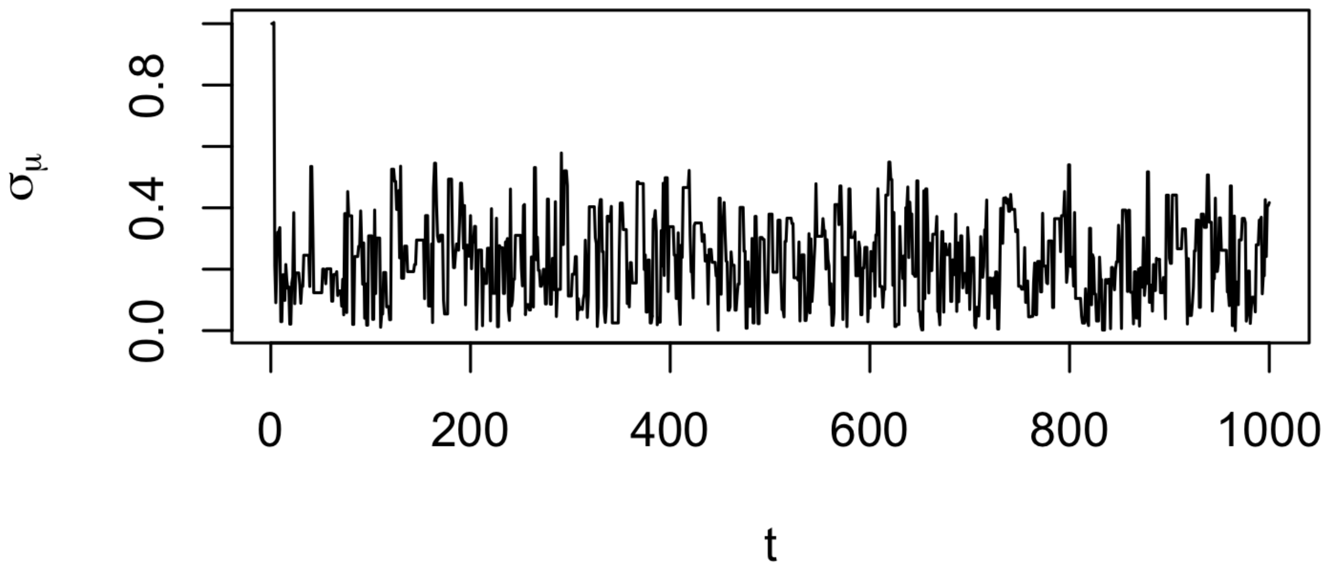


Marginal Posterior



MLE of σ_μ is 0.11

Trace Plots



- Acceptance probability is 0.57
- Goal is around 0.44 in 1 dimension to 0.23 in higher dimensions

AutoCorrelation Function