Hierarchical models – motivation

James-Stein inference

- Suppose $X \sim N(\theta, 1)$
  - X is admissible (not dominated) for estimating θ with squared error loss
- Now $X_i \sim N(\theta_i, 1),\ i = 1, \ldots, r$
  - $X = (X_1, \ldots, X_r)$ is admissible if $r = 1, 2$ but not $r \geq 3$
  - for $r \geq 3$

$$\delta_i = (1 - \frac{r-2}{\sum_i X_i^2})X_i$$

    yields better estimates
  - known as James-Stein estimation

Hierarchical models – motivation

James-Stein inference (cont'd)

- Bayes view: $X_i \sim N(\theta_i, 1)$ and $\theta_i \sim N(0, a)$
  - posterior distn: $\theta_i | X_i \sim N$
  - posterior mean is $(1 - \frac{1}{a+1})X_i$
  - need to estimate a; one natural approach yields James-Stein
- Summary
  - estimation results depend on loss function
  - squared-error loss do well on avg but maybe poor for one component
  - powerful lesson about combining related problems to get improved inferences

Hierarchical Models

Suppose we have data

$$\begin{aligned} Y_{ij} \quad & j = 1, \ldots, J \\ & i = 1, \ldots, n_j \end{aligned}$$

such that $Y_{ij} \quad i = 1, \ldots, n_j$ are independent given θ_j with distribution $p(Y|\theta_j)$. e.g.
$\underbrace{\textit{scores}}_{Y}$ for $\underbrace{\textit{students}}_{(i)}$ in $\underbrace{\textit{classrooms}}_{(j)}$ It might be reasonable to expect θ_j's to be "similar" (but not necessarily identical). Therefore, we may perhaps try to estimate population distribution of θ_j's. This is achieved in a natural way if we use a prior distribution in which the θ_j's are viewed as a sample from a common *population distribution.*

Hierarchical Models

- **Key:** The observed data, y_{ij}, with units indexed by i within groups indexed by j, can be used to estimate aspects of the population distribution of the θ_j's even though the values of θ_j are not themselves observed.
- **How?** It is natural to model such a problem hierarchically
  - observable outcomes modeled conditionally on parameters θ
  - θ given a probabilistic specification in terms of other parameters, ϕ, known as *hyperparameters*.

Hierarchical Models

- Nonhierarchical models are usually inappropiate for hierarchical data.
  - a single θ (i.e. $\theta_j \equiv \theta \;\; \forall j$) may be inadequate to fit a combined data set.
  - separate unrelated θ_j are likely to "overfit" data.
  - information about one θ_j can be obtained from others' data.
- Hierarchical model uses many parameters but population distribution induces enough structure to avoid overfitting.

Setting up hierarchical models

Exchangeability

Recall: A set of random variables $(\theta_1, \ldots, \theta_k)$ is **exchangeable** if the joint distribution is invariant to permutations of the indexes $(1, \ldots, k)$. The indexes contain no information about the values of the random variables.

- hierarchical models often use exchangeable models for the prior distribution of model parameters
- iid random variables are one example
- seemingly non-exchangeable r.v.'s may become exchangeable if we condition on all available information (e.g., regression analysis)

Setting up hierarchical models

Exchangeable models

- Basic form of exchangeable model
  - $\theta = (\theta_1, \ldots, \theta_k)$ are independent conditional on additional parameters ϕ (known as hyperparameters)

$$p(\theta|\phi) = \prod_{j=1}^{k} p(\theta_j|\phi)$$

  - ϕ referred to as hyperparameter(s) with hyperprior distn $p(\phi)$
  - implies $p(\theta) = \int p(\theta|\phi)p(\phi)d\phi$
  - work with joint posterior distribution, $p(\theta, \phi|y)$
- One objection to exchangeable model is that we may have other information, say (X_j). In that case may take

$$p(\theta_1, \ldots, \theta_J|X_1, \ldots, X_J) = \prod_{i=1}^{J} p(\theta_i|\phi, X_i)$$

Setting up hierarchical models

- Model is specified in nested stages
  - sampling distribution $p(y|\theta)$
    (first level of hierarchy)
  - prior (or population) distribution for θ: $p(\theta|\phi)$
    (second level of hierarchy)
  - prior distribution for ϕ (hyperprior): $p(\phi)$
  - Note: more levels are possible
  - hyperprior at highest level is often diffuse but improper priors must be checked carefully to avoid improper posterior distributions.

Setting up hierarchical models

- Inference
  - Joint distn:

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

  - Posterior distribution

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

    * often $p(\theta|\phi)$ is conjugate for $p(y|\theta)$
    * if we know (or fix) ϕ: $p(\theta|y, \phi)$ follows from conjugacy
    * then need inference for ϕ: $p(\phi|y)$

Computational approaches for hierarchical models

- Marginal model

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

  do inference only for ϕ (e.g. marginal maximum likelihood)

- Empirical Bayes

$$p(\theta|y, \hat{\phi}) \propto p(y|\theta)p(\theta|\hat{\phi})$$

  do inference for θ

- Hierarchical Bayes (a.k.a. full Bayes)

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

  inference for θ and ϕ

Hierarchical models and random effects

Animal breeding example

Consider the following mixed linear model commonly used in animal breeding studies

$$Y = X\beta + Zu + e$$

X = design matrix for fixed effects
Z = design matrix for random effects
β = fixed effects parameters
u = random effects parameters
e = individual variation $\sim N(0, \sigma_e^2 I)$

$$Y|\beta, u, \sigma_e^2 \ \sim \ N(X\beta + Zu, \sigma_e^2 I)$$

$$u|\sigma_a^2 \sim N(0, \sigma_a^2 A)$$

(can also think of β as random with $p(\beta) \propto 1$)

Hierarchical models and random effects

Animal breeding example

- Marginal model (after integrating out u)

$$Y|\beta, \sigma_a^2, \sigma_e^2 \sim N(X\beta, \sigma_a^2 ZAZ' + \sigma_e^2 I)$$

- Note: the separation of parameters into θ and ϕ is somewhat ambiguous here:
  - model specification suggests $\phi = \{\sigma_a^2\}$ and $\theta = \{\beta, u, \sigma^e\}$
  - marginal model suggests $\phi = \{\beta, \sigma_a^2, \sigma_e^2\}$ and $\theta = \{u\}$

Hierarchical models and random effects

Animal breeding example

- Empirical Bayes (known as REML/BLUP)

  We can estimate σ_a^2, σ_e^2 by marginal (restricted?) maximum likelihood ($\hat{\sigma}_a^2$, $\hat{\sigma}_e^2$). Then

  $$p(u, \beta | y, \hat{\sigma}_a^2, \hat{\sigma}_e^2) \propto p(y | \beta, u, \hat{\sigma}_e^2) p(u | \hat{\sigma}_a^2)$$

  (a joint normal distn)

- Hierarchical Bayes

  $$p(\beta, \sigma_a^2, \sigma_e^2, \mu | y) \propto p(y | \beta, u, \sigma_e^2) P(u | \sigma_a^2) p(\beta, \sigma_a^2, \sigma_e^2)$$

Computation with hierarchical models

- Two cases
  - conjugate case ($p(\theta|\phi)$ conjugate prior for $p(y|\theta)$)
    - approach described below
  - non-conjugate case
    - requires more advanced computing
    - problem-specific implementations
- Computational strategy for conjugate case
  - write $p(\theta, \phi|y) = p(\phi|y)p(\theta|\phi, y)$
  - identify conditional posterior density of θ given ϕ, $p(\theta|\phi, y)$ (easy for conjugate models)
  - obtain marginal posterior distribution of ϕ, $p(\phi|y)$
  - simulate from $p(\phi|y)$ and then $p(\theta|\phi, y)$

Computation with hierarchical models

The marginal posterior distribution $p(\phi|y)$

- Approaches for obtaining $p(\phi|y)$
  - integration $p(\phi|y) = \int p(\theta, \phi|y)d\theta$
  - algebra - for a convenient value of θ

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

- Sampling from $p(\phi|y)$
  - easy if known distribution
  - grid if ϕ is low-dimensional
  - more sophisticated methods (later)

Beta-binomial example

- Series of toxicology studies
- Study j: n_j exchangeable individuals
  y_j develop tumors
- Model specification:
  - $y_j|\theta_j \sim \text{Bin}(n_j, \theta_j), j = 1, \ldots, J$ (indep)
  - $\theta_j, j = 1, \ldots, J \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta)$ (iid)
  - $p(\alpha, \beta)$ – to be specified later, hopefully ”non-informative”
- Marginal model:
  - can integrate out $\theta_j, j = 1, \ldots, J$ in this case

$$
\begin{aligned}
p(y|\alpha,\beta) &= \int \cdot \int \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1} \binom{n_j}{y_j} \theta_j^{y_j}(1-\theta_j)^{n_j-y_j}\, d\theta_1 \cdot d\theta_J \\
&= \prod_{j=1}^{J} \binom{n_j}{y_j} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}
\end{aligned}
$$

  - $y_j, j = 1, \ldots, J$ are ind
  - distn of y_j is known as beta-binomial distn

Beta-binomial example

- Conditional distn of θ's given α, β, y
  - $p(\theta|\alpha, \beta, y) = \prod_j \text{Beta}(\alpha + y_j, \beta + n_j - y_j)$
  - independent conjugate analyses
  - find this by algebra or by inspection of $p(\theta, \alpha, \beta|y)$
  - analysis is thus reduced to finding (and simulating from) $p(\alpha, \beta|y)$
- Marginal posterior distn of α, β

$$p(\alpha, \beta|y) \propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + y_j)\Gamma(\beta + n_j - y_j)}{\Gamma(\alpha + \beta + n_j)}$$

  - could derive from marginal distn on previous slide
  - could also derive from joint posterior distn
  - not a known distn (on α, β) but easy to evaluate

Beta-binomial example

- Hyperprior distn $p(\alpha, \beta)$
  - First try: $p(\alpha, \beta) \propto 1$ (flat, noninformative?)
    - $*$ equivalent to $p(\alpha/(\alpha+\beta), \alpha+\beta) \propto (\alpha+\beta)$
    - $*$ equivalent to $p(\log(\alpha/\beta), \log(\alpha+\beta)) \propto \alpha\beta$
    - $*$ check to see if posterior is proper
      - · consider diff't cases (e.g., $\alpha \to 0, \beta$ fixed)
      - · if $\alpha, \beta \to \infty$ with $\alpha/(\alpha+\beta) = c$, then $p(\alpha, \beta|y) \propto$ constant (not integrable)
      - · this is an improper distn
      - · contour plot would also show this (lots of probability extending out towards infinity)

Beta-binomial example

- Hyperprior distn $p(\alpha, \beta)$
  - Second try: $p(\alpha/(\alpha+\beta), \alpha+\beta) \propto 1$
    (flat on prior mean and precision)
    * more intuitive, these two params are plausibly independent
    * equivalent to $p(\alpha, \beta) \propto 1/(\alpha+\beta)$
    * still leads to improper posterior distn
  - Third try: $p(\log(\alpha/\beta), \log(\alpha+\beta)) \propto 1$
    (flat on natural transformation of prior mean and variance)
    * equivalent to $p(\alpha, \beta) \propto 1/(\alpha\beta)$
    * still leads to improper posterior distn
  - Fourth try: $p(\alpha/(\alpha+\beta), (\alpha+\beta)^{-1/2}) \propto 1$
    (flat on prior s.d. and prior mean)
    * equivalent to $p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}$
    * "final answer" - proper posterior distn
    * equivalent to
      $p(\log(\alpha/\beta), \log(\alpha+\beta)) \propto \alpha\beta(\alpha+\beta)^{-5/2}$ (this will come up later)

Beta-binomial example

- Computing
  - later consider more sophisticated approaches
  - for now, use grid approach
    * simulate α, β from grid approx to posterior distn
    * then simulate θ's using conjugate beta posterior distn
  - convenient to use $(\log(\alpha/\beta), \log(\alpha + \beta))$ scale because contours "look better" and we can get away with smaller grid
- Illustrate with rat tumor data (separate handout)

Normal-normal hierarchical model

- Data model
  - $y_j|\theta_j \sim N(\theta_j, \sigma_j^2), j = 1, \ldots, J$ (indep)
  - σ_j^2's are assumed known
    (can release this assumption later)
  - motivation: y_j could be a summary statistic with (approx) normal distn from the j-th study (e.g., regression coefficient, sample mean)
- Prior distn
  - need a prior distn $p(\theta_1, \ldots, \theta_J)$
  - if exchangeable, then model θ's as iid given parameters ϕ
  - some additional comments follow

Normal-normal hierarchical model

- **Constructing a prior distribution**
  Can think of this data model as a one-way ANOVA model (especially if y_j is a sample mean of n_j obs in group j). Typical ANOVA analysis begins by testing:

$$H_0: \ \theta_1 = \ldots = \theta_J$$
$$H_a: \ \text{not } H_0$$

  - If we don't reject H_0, we might prefer to estimate each θ_j by the pooled estimate,

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

  - If we reject H_0, we might use separate estimates, $\hat{\theta}_j = y_j$ for each j.
  - Alternative: compromise between complete pooling and none at all,
    e.g., a weighted combination,

$$\theta_j = \lambda_j y_j + (1 - \lambda)\bar{y}_{..} \text{ where } \lambda_j \in (0, 1)$$

Normal-normal hierarchical model

- **Constructing a prior distribution** (Cont'd)

  (a) The pooled estimate $\hat{\theta} = \bar{y}_{\cdot\cdot}$ is the posterior mean if the J values θ_j are restricted to be equal, with a uniform prior density on the common θ; i.e. $p(\theta) \propto 1$.

  (b) The unpooled estimate $\hat{\theta}_j = y_j$ is the posterior mean if the J values θ_j have independent uniform prior densities on $(-\infty, \infty)$;
  i.e. $p(\theta_1, \ldots, \theta_J) \propto 1$.

  (c) The weighted combination is the posterior mean if the J values θ_j are iid $N(\mu, \tau^2)$.

  Note: (a) corresponds to (c) with $\tau^2 = 0$
  (b) corresponds to (c) with $\tau^2 \to \infty$

Normal-normal hierarchical model

- Data model $p(y_j|\theta_j) \sim N(\theta_j, \sigma_j^2), j = 1, \ldots, J$

  σ_j^2's assumed known

- Prior model for θ_j's is normal (conjugate)

$$p(\theta_1, \ldots, \theta_J|\mu, \tau) = \prod_{j=1}^{J} N(\theta_j|\mu, \tau^2)$$

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

  i.e. θ_j's conditionally independent given (μ, τ)

- Hyperprior distribution
  - noninformative distribution for μ given τ, i.e., $p(\mu|\tau) \propto 1$ (this won't matter much because the combined data from all J experiments are highly informative about μ)
  - more on $p(\tau)$ later
  - $p(\mu, \tau) = p(\tau)p(\mu|\tau) \propto p(\tau)$

Normal-normal model: computation

- Joint posterior distribution:

$$
\begin{aligned}
&p(\theta, \mu, \tau | y) \\
&\quad \propto \; p(\mu, \tau) p(\theta | \mu, \tau) p(y | \theta) \\
&\quad \propto \; p(\tau) \prod_{j=1}^{J} N(\theta_j | \mu, \tau^2) \prod_{j=1}^{J} N(y_j | \theta_j, \sigma_j^2) \\
&\quad \propto \; p(\tau) \frac{1}{\tau^J} \exp\left[-\frac{1}{2} \sum_j \frac{1}{\tau^2} (\theta_j - \mu)^2 \right] \exp\left[-\frac{1}{2} \sum_j \frac{1}{\sigma_j^2} (y_j - \theta_j)^2 \right]
\end{aligned}
$$

- Factors that depend only on y and $\{\sigma_j\}$ are treated as constants because they are known
- Posterior distn is a distn on $J + 2$ parameters
- Can compute using MCMC (later) or
- Hierarchical computation:
  1. $p(\theta_1, \ldots, \theta_J | \mu, \tau, y)$
  2. $p(\mu | \tau, y)$
  3. $p(\tau | y)$

Normal-normal model: computation Conditional posterior distn of θ given μ, τ, y

- Treat (μ, τ) as fixed in previous expression
- Given (μ, τ), the J separate parameters θ_j are independent in their posterior distribution
- $\theta_j | y, \mu, \tau \sim N(\hat{\theta}_j, V_j)$ with

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\ y_j + \frac{1}{\tau^2}\ \mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \quad \text{and} \quad V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

- Result from simple normal-normal conjugate analysis
- $\hat{\theta}_j$ is weighted average of hyperprior mean and data

Normal-normal model: computation

Marginal posterior distribution of μ, τ given y

- We can analytically integrate the full posterior distn $p(\theta, \mu, \tau|y)$ over θ

$$p(\mu, \tau|y) = \int p(\theta, \mu, \tau|y)\, d\theta$$

- An alternative is to use the marginal model $p(\mu, \tau|y) \propto p(y|\mu, \tau)p(\mu, \tau)$
- Marginal model

$$p(y|\mu, \tau) = \prod_{j=1}^{J} \int \underbrace{N(\theta_j|\mu, \tau)N(\bar{y}_{.j}|\theta_j, \sigma_j^2)}_{\text{quadratic in } y_j}\, d\theta_j$$

$$\Rightarrow \quad y_j|\mu, \tau \sim \text{Normal}$$

$$E(y_j|\mu, \tau) = E(E(y_j|\theta_j, \mu, \tau)) = E(\theta_j) = \mu$$

$$\begin{aligned} Var(y_j|\mu, \tau) &= E(Var(y_j|\mu, \tau, \theta_j)) + \\ &\quad + Var(E(y_j|\mu, \tau, \theta_j)) = \\ &= E(\sigma_j^2) + Var(\theta_j) = \sigma_j^2 + \tau^2 \end{aligned}$$

Normal-normal model: computation

Marginal posterior distribution of μ, τ given y

- End result is

$$
\begin{aligned}
p(\mu, \tau | y) &\propto p(\tau) \prod_{j=1}^{J} N(y_j | \mu, \sigma_j^2 + \tau^2) \\
&\propto p(\tau) \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(-\frac{(y_j - \mu)^2}{2(\sigma_j^2 + \tau^2)}\right)
\end{aligned}
$$

- Note: in non-normal models, it is not generally possible to integrate over θ and rely on the marginal model, so that more elaborate computational methods are needed

Normal-normal model: computation

Posterior distribution of μ given τ, y

- Instead of sampling (μ, τ) on a grid, factor the distribution: $p(\mu, \tau|y) = p(\tau|y)p(\mu|\tau, y)$
- $p(\mu|\tau, y)$ is obtained by looking at $p(\mu, \tau|y)$ and thinking of τ as known:

$$\Rightarrow \quad p(\mu|\tau, y) \propto \prod_{j=1}^{J} N(y_j|\mu, \sigma_j^2 + \tau^2)$$

- This is the posterior distn corresponding to a normal sampling distribution with a noninformative prior density on μ
- Result: $\mu|\tau, y \sim N(\hat{\mu}, V_\mu)$ with

$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2} y_j}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}} \quad \text{and} \quad V_\mu = \frac{1}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}}$$

Normal-normal model: computation

Posterior distribution of τ given y

- $p(\tau|y)$ can be found in two equivalent ways
  - integrate $p(\mu, \tau|y)$ over μ
  - use algebraic form $p(\tau|y) = p(\mu, \tau|y)/p(\mu|\tau, y)$, which must hold for any μ
- Choose the second option, and evaluate at $\mu = \hat{\mu}$ (for simplicity):

$$
\begin{aligned}
p(\tau|y) &\propto \frac{\prod_{j=1}^{J} N(y_j|\hat{\mu}, \sigma_j^2 + \tau^2)}{N(\hat{\mu}|\hat{\mu}, V_\mu)} \\
&\propto V_\mu^{1/2} \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(-\frac{(y_j - \hat{\mu})^2}{2(\sigma_j^2 + \tau^2)} \right)
\end{aligned}
$$

- Note that V_μ and $\hat{\mu}$ are both functions of τ
- Compute $p(\tau|y)$ on a grid of values of τ

Normal-normal model: computation Summary

- To simulate from joint posterior distribution $p(\theta, \mu, \tau|y)$:
  1. draw τ from $p(\tau|y)$ (grid approximation)
  2. draw μ from $p(\mu|\tau, y)$ (normal distribution)
  3. draw $\theta = (\theta_1, \ldots, \theta_J)$ from $p(\theta|\tau, y)$ (independent normal distribution for each θ_j)
- Choice of $p(\tau)$
  - $p(\tau) \propto 1$ - proper posterior distribution
  - $p(\log \tau) \propto 1$ - improper posterior distribution (equivalent to $p(\tau^2) \propto 1/\tau^2$)
  - alternative family for informative prior distn is scaled inverse-χ^2 family
- Illustrate with SAT coaching example (separate handout)

Computation

Introduction

- Goal: Posterior inference for parameters, missing data (if any), and predictions
- Thus far:
  - analytic results or exact simulation in small problems
  - normal approximation
  - grid approximation
  - use hierarchical structure (e.g., $\mu, \tau | y$, then $\theta | y, \mu, \tau$, then $\tilde{y} | y, \theta, \mu, \tau$)
- Now consider additional tools:
  - iterative simulation (Markov chain Monte Carlo)
  - importance sampling (covered later in robust models section)
- A mini statistical computing course

Computation

Motivation for additional tools

- Examples from first part of the course have obvious extensions for which computation becomes difficult:
  - logistic regression
    - more than one predictor
    - incorporating random effects
  - normal-normal hierarchical model
    - may use non-normal distn at either level (t-distn for population distn or Poisson distn for count data)
    - nontrivial covariance matrix in prior distn (spatial models, time series models)

Computation

- An overall computation strategy
  - initial (perhaps crude) estimates of parameters
  - direct simulation when possible
  - if direct simulation is not possible
    - approximations based at the posterior mode
    - iterative simulation (e.g., Gibbs sampler, Metropolis algorithm)
  - importance sampling for robustness checks and sensitivity analysis
- Next
  - review/discuss these ideas

Computation

Some helpful ideas we have met

- Compute posterior distn on log scale (to avoid underflows or overflows)
- Factoring the posterior distribution (e.g., $p(\theta_1, \theta_2|y) = p(\theta_1|\theta_2, y)p(\theta_2|y)$)
  - reduce to easier, lower-dimensional problems
  - isolate the parameters most influenced by prior distribution (e.g., τ in 8 schools example)
  - difficulties:
    * can't generally find marginal distn easily
    * hard to use a grid with a high-dimensional marginal distn
- Transformations
  - create more understandable parameters
  - make prior independence plausible
  - improve normal approximation (e.g., log of scale parameter)
  - speed/simplify iterative simulation

Computation

Notation/Notes

- $p(\theta|y)$ is the posterior distn
  - θ now includes all parameters (even in hierarchical model)
  - often we only know the unnormalized posterior distn $q(\theta|y)$
    - i.e., $p(\theta|y) \propto p(y|\theta)p(\theta) = q(\theta|y)$
    - more formally, $p(\theta|y) = c(y)q(\theta|y)$
  - our computation discussion will generally use $p(\theta|y)$ and I will point out whether it matters whether the posterior distn is normalized

Computation

Initial estimation

- Starting point for subsequent approaches
- Serves as a check for other approaches
- Problem-specific methods are required
  - use results from other methods (e.g., maximum likelihood estimates in bioassay logistic regression)
  - fix hyperparameters at crude estimates (e.g., separate and pooled estimates for the 8 schools are equivalent to $\tau = \infty$ and $\tau = 0$)

Computation

Direct simulation

- We have already seen that simulation is a powerful approach for studying the posterior distn in a Bayesian analysis
- Brief discussion of simulation tools
  - useful in simpler (low dimensional) problems
  - same tools are useful as components for more advanced simulations
- Simulation analysis
  - report number of draws
  - report summary statistics (mean, sd, percentiles)
  - graphs
  - how many draws? depends on desired accuracy (e.g., if we have iid simulations then std error of posterior mean is equal to posterior s.d. divided by $\sqrt{n}$)
- Direct simulation is not usually possible in high dimensions but direct simulation techniques can be useful tools within more sophisticated algorithms

Computation

Direct simulation approaches

- Exact simulation
  - standard algorithms for drawing from standard distns (uniform, normal, Poisson, gamma, etc.)
  - available in most software including S-plus
- Grid approximation
  - discrete (evenly spaced) grid $\theta_1, \theta_2, \ldots, \theta_N$,

$$\Pr_{grid}(\theta = \theta_j) = p(\theta_j|y)/(\sum_i p(\theta_i|y))$$

  - we have already seen this approach
  - works for normalized or unnormalized posterior distn
  - hard in 2 or more dimensions
  - choice of grid can affect the answer

Computation

Direct simulation approaches

- Probability integral transform
  - consider posterior distn $p(\theta|y)$ with corresponding cdf $F(\theta|y)$
  - recall probability result: if $U \sim \text{Unif}(0,1)$, then $\theta = F^{-1}(U)$ is a r.v. with distn $p(\theta|y)$
  - e.g., if $\theta|y \sim N(\mu, \tau^2)$, then $\theta = \mu + \tau\Phi^{-1}(U)$
  - discrete r.v.'s are possible but harder to program
  - can use this to improve grid to trapezoidal approximation

Computation

Direct simulation approaches

- Rejection sampling
  - suppose we find $g(\theta)$we can sample from with $p(\theta|y)/g(\theta) \leq M$ (with M known)
  - algorithm:
    * draw $\theta \sim g(\theta)$
    * accept θ with prob $p(\theta|y)/(Mg(\theta))$, otherwise reject and draw a new candidate
    * for log-concave densities this approach can be used with trapezoids defining rejection function (Gilks and Wild, 1992, Applied Statistics)
- Many other useful methods for direct simulation that we don't have time to discuss here

Computation

Iterative simulation

- Basic idea: to sample from $p(\theta|y)$ create a Markov chain with $p(\theta|y)$ as stationary distribution
- Algorithms:
  - Gibbs sampler (full conditionals)
  - Metropolis-Hastings algorithm (jumping distn)
  - combinations of Gibbs and M-H
- Implementation issues (later)

Iterative simulation

Gibbs sampler

- Key features
  - break problem into lower-dimensional pieces using conditional distributions
  - conditional posterior distributions often have simple form
- Start by drawing an initial $\theta = (\theta_1, \ldots, \theta_k)$ from an approximation to $p(\theta|y)$.
- Repeat the following steps using most recently drawn values for variables in conditioning set:
  - draw θ_1 from $p(\theta_1 \mid \theta_2, \ldots, \theta_k, y)$
  - draw θ_2 from $p(\theta_2 \mid \theta_1, \theta_3, \ldots, \theta_k, y)$

    $\ldots$
  - draw θ_k from $p(\theta_k \mid \theta_1, \ldots, \theta_{k-1}, y)$
- Can update parameters one at a time (as above) or in blocks

Iterative simulation

Plan of attack

- We have glossed over some details
  - non-standard distributions come up in Gibbs sampling
  - starting values
  - monitoring convergence
  - inference from iterative simulation
  - software availability
  - efficiency considerations
- Return to these after an example

Iterative simulation

Non-standard distributions

- It may happen that one or more of the Gibbs sampling distns is not a known distn
- What then?
  - can go back to previous direct simulation discussion
    - grid approximation
    - rejection sampling, etc.
  - Metropolis (or (Metropolis-Hastings) algorithm
    - let's meet this important subject now

Iterative simulation

Metropolis-Hastings (M-H) algorithm

- Replaces "conditional draws" of Gibbs sampler with "jumps" around the parameter space
- Algorithm:
  - given current draw θ (scalar or vector)
  - sample a candidate point θ^* from jumping distribution $J(\theta^*|\theta)$
  - accept candidate or stay in place with probabilities determined by importance ratio

$$r = \frac{p(\theta^*|y)/J(\theta^*|\theta)}{p(\theta|y)/J(\theta|\theta^*)}$$

- Simplifies if J is symmetric (Metropolis algorithm)
- Combining M-H and Gibbs: M-H steps can be used in place of one conditional distn in a Gibbs sampler, or a single M-H step can replace several (or even all) of the conditional distns

Iterative simulation

Starting values

- Markov chain will converge to stationary distribution from **any** starting value assuming
  - chain has a nonzero probability of eventually getting from any point to any other point (i.e., parameter space is not divided into separate regions)
  - chain does not drift off to infinity (can happen if the posterior distribution is improper – which means the model is wrong!)
- Assessing when this convergence has occurred is best done using multiple chains with overdispersed starting points

Iterative simulation

Starting values

- An algorithm for choosing starting values:
  - find posterior mode (or modes)
    (marginal distn usually better than joint distn)
  - create overdispersed approximation to posterior
    (e.g., t_4 instead of normal)
  - sample 1000 points from approximation
  - resample 5 or 10 starting values
    (using importance ratios as described later)

Iterative simulation

Monitoring convergence

- Run several sequences in parallel
- Can use graphical displays to monitor convergence or semi-formal approach of Gelman and Rubin (described now)
- Two estimates of $\mathrm{sd}(\theta|y)$
  - underestimate from sd within each sequence
  - overestimate from sd of mixture of sequences
- Potential scale reduction factor:

$$\sqrt{\widehat{R}} = \frac{\text{mixture-of-sequences estimate of } \mathrm{sd}(\theta|y)}{\text{within-sequence estimate of } \mathrm{sd}(\theta|y)}$$

- Initially $\sqrt{\widehat{R}}$ is large (because we use overdispersed starting points)
- At convergence, $\sqrt{\widehat{R}} = 1$ (each sequence has made a complete tour)
- Monitor $\sqrt{\widehat{R}}$ for all parameters and quantities of interest; stop simulations when they are all near 1 (e.g., below 1.2)

Iterative simulation

Inference from posterior simulations

- At approximate convergence we have many draws from the posterior distribution
- The draws are **not** independent
- This means that obtaining standard errors to assess simulation noise is difficult
  (can use between-chain info, batching, ...)
- Note there is a distinction here between posterior uncertainty about θ and Monte Carlo uncertainty about some summary of the posterior distn (e.g., std error of $E(\theta|y)$)
- Good news: Simulation noise is generally minor compared to posterior uncertainty about θ

Iterative simulation

Software availability

- Variety of packages (more in development)
- One popular package is BUGS/CODA
  - BUGS (Bayesian analysis Using Gibbs Sampling)
  - CODA (Convergence Diagnosis and Output Analysis)
  - available on the web at
    http://www.mrc-bsu.cam.ac.uk/
    bugs/welcome.shtml
- Other software described by Carlin and Louis (1996)
- Create new models – write your own software

Iterative simulation

Efficiency considerations

- Theory under construction but some things are known:
  - Gibbs sampling
    - works best if we can create independent or nearly independent blocks of parameters
    - partition parameters into groups
    - transform parameters
  - Metropolis-Hastings algorithms
    - choice of jumping distn is key

Iterative simulation

Efficiency considerations - M-H

- How do we choose the jumping distribution $J(\theta|\theta^{(t-1)})$?
- Optimal J is $p(\theta|y)$ independent of current value $\theta^{(t-1)}$
  - this always accepts ($r = 1$)
  - but if we could do this we wouldn't need M-H
- Goals in choosing J:
  - J should be easy to sample from
  - it should be easy to compute r
  - jumps should go far (so we move around the parameter space) but not too far (so they are not always rejected)

Iterative simulation

Efficiency considerations - M-H

- Three primary approaches
  - independence M-H
  - random walk M-H (used most often)
  - approximation M-H
- Independence M-H
  - find a distribution $g(\theta)$ independent of current $\theta^{(t-1)}$ and keep generating candidates from $g(\theta)$
  - requires g be a reasonably good approximation
  - hard to do for M-H within Gibbs

Iterative simulation

Efficiency considerations - M-H

- Random Walk M-H
  - generate candidate using random walk (often normal) centered at current value
  - $J(\theta|\theta^{(t-1)}) = N(\theta|\theta^{(t-1)}, cV)$
  - note this is symmetric so M-H acceptance calculation simplifies
  - works well if V is chosen to be posterior variance (don't know this but can use a pilot run to get some idea)
  - c is a constant chosen to optimize efficiency
  - theory results indicate optimal acceptance rate for this kind of jumping distn is between .2 and .5 (decreases with dimension)

Iterative simulation

Efficiency considerations - M-H

- Approximation M-H
  - generate candidate using an approximation to target distn (varying from iteration to iteration)
  - e.g., $J(\theta|\theta^{(t-1)}) = N(\theta|\theta^{(t-1)}, V_{\theta^{(t-1)}}$
  - now variance matrix depends on current value this is no longer symmetric
  - idea is to make this a good approximation (high acceptance rate)

Computation

Debugging iterative simulation methods

- Checking that programs are correct is crucial (especially if you write your own)
- Can be difficult to check because
  - output is a distribution not a point estimate
  - incorrect output may look reasonable
- Some useful debugging ideas:
  - build up from simple (debugged) models
  - when adding a new parameter, start by setting it to a fixed value, then let it vary
  - simulate fake data (repeat the following steps)
    * draw "true parameters" from prior distn (must be proper)
    * simulate data from the model
    * obtain draws from posterior distn
    * compare distns of posterior draws and "true parameters"

Computation

Debugging iterative simulation methods

- Common problems
  - conceptual flaw in part of model
  - prior is too vague
    - this may give improper posterior distn
    - detect by looking for values that don't make substantive sense

Computation

Approximation

- Recall results of Chapter 4 ... for large samples $p(\theta|y)$ is approx $N(\theta|\hat{\theta}, I(\hat{\theta})^{-1})$ where $\hat{\theta}$ is the posterior mode
- Often use inverse curvature matrix of log posterior density, $V_\theta = \left[-\frac{d^2}{d\theta^2} \log p(\theta|y)|_{\theta=\hat{\theta}}\right]^{-1}_{,}$ as variance matrix for approximation
- Transformations are often used to improve quality of normal approx
- May use t distn with few degrees of freedom in place of normal distn (to protect against long tails)
- Multiple modes can be a problem: $\mathrm{N}(\hat{\theta}, V_\theta)$ or $t_4(\hat{\theta}, V_\theta)$ approx at each mode (i.e., a mixture)
- Reasons **not** to approximate based on modes:
  - misleading in some problems
    (e.g., in 8 schools example, mode is $\tau = 0$
    which is at edge of parameter space)
  - advances in algorithms have made inference
    from exact posterior distn possible

Computation

Approximation - mode finding

- To apply normal approximation, need posterior mode
- Review traditional stat computing topic of mode finding (optimization)
- Iterative conditional modes
  - start at $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})$
  - for $i = 1, \ldots$
    - for $j = 1, \ldots, d$
      - choose $\theta_j^{(i)}$ as the value that maximizes (or even just increases) $p(\theta_1^{(i)}, .., \theta_{j-1}^{(i)}, \theta, \theta_{j+1}^{(i-1)}, .., \theta_d^{(i-1)})$
  - leads to a local maximum

Computation

Approximation - mode finding

- Newton's method ($L = \log p(\theta|y)$)
  - start at $\theta^{(0)}$
  - iterate with $\theta^{(t)} = \theta^{(t-1)} - [L''(\theta^{(t-1)})^{-1}L'(\theta^{(t-1)}$
  - converges fast but is sensitive to starting value
  - can use numerical derivatives
- Other optimization methods
  - steepest ascent $\theta^{(t)} = \theta^{(t-1)} + \alpha L'(\theta^{(t-1)})$
  - quasi-Newton methods
  - simplex/polytope (no derivative methods)

Computation

Approximation

- For many problems, especially hierarchical models, the joint mode is not very useful
- Instead may focus on factorization $p(\theta, \phi|y) = p(\phi|y)p(\theta|\phi, y)$
- Often $p(\theta|\phi, y)$ is easy (e.g., conjugate family)
- Normal approximation for marginal posterior distn $p(\phi|y)$
- But need mode of $p(\phi|y)$
  - sometimes this function can be identified and maximized analytically
  - for other situations EM algorithm is helpful

Computation

Approximation - The EM algorithm

- EM is an iterative algorithm for maximizing functions (likelihoods or posterior distns) when there is missing data
- Applied here in maximizing $p(\phi|y)$ treating θ as missing data
- Idea:
  - start with initial guess for ϕ
  - given ϕ we can estimate "missing data" θ
  - given estimated θ it may be easy to now maximize for improved ϕ
  - repeat last two steps

Computation

Approximation - The EM algorithm

- Iterative algorithm with two steps
- Suppose current value of ϕ is $\phi^{(t)}$
  - E-step
    - compute $Q(\phi) = E(\log(p(\theta, \phi|y)|\phi = \phi^{(t)}) = \int \log(p(\theta, \phi|y))p(\theta|\phi^{(t)}, y)d\theta$
    - essentially computes expected value of needed functions of θ rather than estimating the "missing" θ
  - M-step
    - choose $\phi^{(t+1)}$ as the value of ϕ that maximizes $Q(\phi)$
- Can show that $p(\phi|y)$ increases after each E-M pair of steps

Computation

Numerical integration

- Historically people often used numerical integration to study posterior distn
- Many quantities of interest can be written as $E(h(\theta)|y) = \int h(\theta)p(\theta|y)d\theta$ (e.g., posterior mean)
- In modern world, simulation is often preferred (but numerical integration still used)
- We focus on useful tools developed in this context

Computation

Numerical integration

- Traditional quadrature
  - trapezoidal rule (piecewise linear approximation)
  - Simpson's rule (piecewise quadratic)
  - algorithms for iterating
  - Gaussian quadrature

Computation

Numerical integration

- Integration via direct simulation
  - if we can generate $\theta_1, \ldots, \theta_N$ from $p(\theta|y)$ then we can estimate integral as $\sum_i h(\theta_i)/N$
  - of course, this is just our direct simulation approach!
- Importance sampling
  - can write $E(h(\theta)|y) = \int \frac{h(\theta)p(\theta|y)}{g(\theta)} g(\theta) d\theta$
  - if we can generate $\theta_1, \ldots, \theta_N$ from $g(\theta)$, then we can estimate integral as $\frac{1}{N}\sum_i \frac{h(\theta_i)p(\theta_i|y)}{g(\theta_i)}$
  - $w(\theta_i) = p(\theta_i|y)/g(\theta_i)$ is known as the importance ratio
  - improves upon simple MC if we can find g yielding low variability weights

Computation

Numerical integration

- Importance sampling (cont'd)
  - won't work at all if g's tails are too short
  - can work for unnormalized distn
- Many other techniques for improving Monte Carlo (e.g., antithetic variables) ... see statistical computing texts

Computation

Numerical integration

- Analytical approximation (Laplace's method)
  - can write $E(h(\theta)|y) = \int e^{\log(h(\theta)p(\theta|y))} d\theta$
  - approximate $u(\theta) = \log(h(\theta)p(\theta|y))$ using a quadratic expansion around the mode θ_o
  - find $E(h(\theta)|y) \approx h(\theta_o)p(\theta_o|y)(2\pi)^{-d/2}|-u''(\theta_o)|^{1/2}$
  - requires large samples
  - need two approximations for unnormalized posterior distn
    $(E(h(\theta)|y) = \int h(\theta)q(\theta|y)d\theta / \int q(\theta|y)d\theta)$

Computation

Summary

- Goal: posterior inference concerning the vector of parameters (and any missing data)
- Simulation is an extremely powerful tool, especially so in complex models
- Basic approach
  - initial estimates
  - direct simulation (if possible)
  - if direct simulation is not possible:
    - normal or t approximation about posterior mode
    - iterative simulation (Gibbs, Metropolis-Hastings)
- For iterative simulation
  - inference is conditional on the starting points
  - use multiple sequences and run until they mix

Model checking

Introduction

- So far:
  - build probability models
  - compute/simulate posterior distn
- Now:
  - model checking (does the model fit the data)
  - sensitivity analysis (are conclusions sensitive to assumptions)
  - model selection (which is the best model)
  - robust analysis (are conclusions sensitive to data)

Model checking

General ideas

- Don't ask if the model is true
- Does the model fit and provide useful inferences
- Remember the model includes
  - sampling distribution
  - prior distribution
  - hierarchical structure
  - explanatory variables
- More than one model can fit (sensitivity analysis)

Model checking: types of checks

- Classical ideas
  - Check whether parameter estimates make sense
  - Check whether predictions make sense
  - Does the model generate data like "my data" (simulation approach, residual analysis)
  - Embed in a larger model
- Bayesian ideas
  - Compare posterior distribution of parameters to substantive knowledge
  - Compare posterior predictive distribution of future data to substantive knowledge
  - Compare posterior predictive distribution of future data to observed data
  - Evaluate sensitivity of inferences to other model specifications (e.g., alternate priors or sampling distributions, embed in larger model)

Posterior predictive model checking

- y^{rep} = replicate data that might have occurred
- Replicated under same model as original data (e.g., same covariate values) with same values for unknown parameters θ
- Posterior predictive distribution of y^{rep}

$$\begin{aligned} p(y^{rep}|y) &= \int p(y^{rep}, \theta|y)\ d\theta \\ &= \int p(y^{rep}|\theta, y)p(\theta|y)d\theta \\ &=? \int p(y^{rep}|\theta)p(\theta|y)d\theta \end{aligned}$$

- Last equality is generally (but not always) true
- Easy to obtain simulations of y^{rep} given posterior simulations of θ
- Other possible definitions of replications (more on this later)

Posterior predictive model checking

- $T(y, \theta)$ is a test quantity or discrepancy measure
- Compare posterior predictive distribution of $T(y^{rep}, \theta)$ to posterior distribution of $T(y, \theta)$
- One possible summary (but not the only one) is the posterior predictive P-value

$$\begin{aligned} P_b &= \Pr(T(y^{rep}, \theta) > T(y, \theta) | y) \\ &= \int \int I_{[T(y^{rep}, \theta) > T(y, \theta)]} p(y^{rep}|\theta) p(\theta|y) dy^{rep} d\theta \end{aligned}$$

- Special case $T(y, \theta) = T(y)$ is a test statistic
  - compare posterior predictive distribution of $T(y^{rep})$ to observed $T(y)$
- Diagnostics such as plots of residuals are special cases of posterior predictive checks

Posterior predictive model checking

Relation to traditional tests

- Example:
  - suppose $y_1, \ldots, y_n$ are iid $N(\mu, \sigma^2)$
  - believe $\mu = 0$, so fit $N(0, \sigma^2)$ model
  - want to check fit of $N(0, \sigma^2)$ model
  - weak example because obvious model checking approach is to fit the "bigger" $N(\mu, \sigma^2)$ model and check whether $\mu = 0$ is plausible
- Frequentist approach
  - test statistic: $T(y) = \bar{y}$
  - begin by assuming σ^2 is fixed

$$\begin{aligned}
\text{p-value} &= P(\overbrace{T(y^{rep})}^{r.v.} \geq \overbrace{T(y)}^{\text{obs.value}} \ |\sigma^2) \\
&= P(\bar{y}^{rep} \geq \bar{y}|\sigma^2) \\
&= P\left(\frac{\sqrt{n}\bar{y}^{rep}}{S} \geq \frac{\sqrt{n}\bar{y}}{S}|\sigma^2\right) = P\left(t_{n-1} \geq \frac{\sqrt{n}\bar{Y}}{S}\right)
\end{aligned}$$

  - last equality because distn no longer depends on σ^2
  - it is not always possible to get rid of nuisance parameters in this way

Posterior predictive model checking

Relation to traditional tests (cont'd)

- Posterior predictive approach

$$
\begin{aligned}
\text{p-value} = \quad & P(T(y^{rep}) \geq T(y)|y) \\
= \quad & \int \int I_{[T(Y^{rep}) \geq T(y)]} p(Y^{rep}|\sigma^2) p(\sigma^2|y) dy^{rep} d\sigma^2 \\
= \quad & \int \underbrace{P(T(y^{rep}) \geq T(y)|\sigma^2)}_{\text{classical p-value}} p(\sigma^2|y) d\sigma^2
\end{aligned}
$$

  - if the classical p-value is independent of σ^2, as for $T(y) = \bar{y}$ in the example, then the posterior predictive p-value is equal to classical p-value
  - if not, then formula above shows how the Bayesian approach handles nuisance parameters

Posterior predictive model checking

Defining replications

- Defining replications y^{rep}
  - usually keep features of original data fixed (e.g., sample size)
  - different definitions are possible in hierarchical models
    - replications of the same units

$$p(\phi|y) \to p(\theta|\phi, y) \to p(y^{rep}|\theta)$$

    - replicate data for new units

$$p(\phi|y) \to p(\theta|\phi) \to p(y^{rep}|\theta)$$

Posterior predictive model checking

Defining test measures

- Defining test statistics or discrepancies
  - measure features of data not directly included in the model (bad to use $T(y) = \bar{y}$ if the model includes a location parameter)
  - may define a number of test measures
  - difficult to speak in general terms because good test measures depend on context
  - examples
    * to check for autocorrelation in a sequence of Bernoulli trials, use a count of the number of runs
    * to check for new predictor in regression model, use $\mathrm{corr}(y - X\beta, x_{new})$
    * to check for asymmetry in a normal model, use $|y_{.9} - \theta| - |y_{.1} - \theta|$
    * to check overall fix in a complex model, use $T(y;\theta) = \sum \left[(y_i - E(y_i|\theta))^2 / \mathrm{Var}(y_i|\theta) \right]$ (Note: asympt χ^2 for known θ but here no reliance on asymptotic distn)

Related ideas

- Parametric bootstrap (e.g., Efron, 1979)
  - plug in point estimate $\hat{\theta}$
  - simulated replicate data sets from $p(y|\hat{\theta})$
- Marginal distribution (Box, 1980)
  - reference distribution is $p(y) = \int p(y|\theta)p(\theta)d\theta$
  - note this is prior predictive distribution
  - requires proper prior distribution
    (even a bit more than that)

Criticisms of pp model checks

- Too conservative ("double-counting(?)" the data)
- pp p-values are not uniform under the null
- Difficult to interpret because of above ... what is an unusually high or low value in practice
- Unobserved data y^{rep} is not relevant for some Bayesians
- Conditional predictive distn or partial posterior predictive distn (Bayarri and Berger in JASA 2000)
  - avoid some of the criticisms by conditioning on "some" of the data but not all
  - can be hard to compute
- Summary: post. pred. checks are conservative but easy to use and easy to interpret

On the conservatism of pp model checks

- Suppose that $Y \sim N(\mu, 1)$ and $\mu \sim N(0, 9)$
- Observe $Y_{obs} = 10$. Is this unusual?
- Prior predictive approach
  - marginal distn of Y is $N(0, 10)$
  - p-value $= 1 - \Phi(10/\sqrt{10}) = .008$
  - don't believe model
  - the observed value 10 is not consistent with this prior distn and data model
- Posterior predictive approach
  - posterior distn of μ is $N(0.9Y_{obs}, 0.9) = N(9, .9)$
  - posterior predictive distn of Y is $N(9, 1.9)$
  - p-value $= .23$
  - model cares about posterior fit
    (this minimizes the effect of the prior)
  - would this approach ever reject the model
    (yes, $Y_{obs} = 23$)

Sensitivity analysis

- Generally true that many models can be fit to the same data
- Question is how sensitive the inferences we draw are to the different models
- Different types of inferences may have different sensitivity
  - posterior mean or median for parameter of interest is typically not sensitive
  - extreme percentiles are more sensitive
- Approaches
  - fit different models
  - expand model/embed model in larger family
    - examp: consider normal distn as part of $t_\nu(\mu, \sigma^2)$ family (normal distn corresponds to $\nu = \infty$)

Bayes factors

- Suppose there are two competings models M_1 and M_2 for a data set
  - different prior distns $p_1(\theta_1)$ and $p_2(\theta_2)$
  - different data models $p_1(y|\theta_1)$ and $p_2(y|\theta_2)$
  - note θ_1 and θ_2 may be of different dimension
- Consider a full Bayesian analysis
  - begin with prior probability $p(M_1) = 1 - p(M_2)$
  - then posterior odds of M_1 relative to M_2 are

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)} \frac{p(M_1)}{p(M_2)}$$

  - posterior odds are the product of prior odds and a form of likelihood ratio $p(y|M_1)/p(y|M_2)$
  - the ratio $p(y|M_1)/p(y|M_2)$ is known as the Bayes factor
  - it is a measure of how much the data changes the odds in favor of M_1 vs M_2

Bayes Factors

- Bayes factor of model 1 relative to model 2

$$BF_{12} = \frac{p(y|M_1)}{p(y|M_2)} = \frac{\int p(y|\theta_1, M_1)p(\theta_1|M_1)\, d\theta_1}{\int p(y|\theta_2, M_2)p(\theta_2|M_2)\, d\theta_2}$$

  - notation: M_1 and M_2 are not events they merely identify models
  - Bayes factor is only defined when the marginal density of y under each model is proper (requires a proper prior distn)

Bayes Factors

Bayes factors and model averaging

- Given m models with prior probabilities $P(M_1), \ldots, P(M_m)$
- Posterior probability for model j is

$$p(M_j|y) = \frac{p(y|M_j)p(M_j)}{\sum_k p(y|M_k)p(M_k)}$$

- Note: $p(M_j|y)/p(M_i|y) = BF_{ji}\frac{p(M_j)}{p(M_i)}$
  $p(M_j|y) = p(M_j)/\left(\sum_k BF_{kj}p(M_k)\right)$
- Model averaging - instead of relying on a single model we can use all of the models
  (essentially a "super" model)
  - then to make a prediction $\tilde{y}$, use
    $p(\tilde{y}|y) = \sum_j p(M_j|y)p(\tilde{y}|M_j, y)$
  - computation - a single MCMC incorporating all models (reversible jump MCMC)

Bayes Factor

Computation

- To compute Bayes factors we need to be able to compute marginal likelihoods

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

- There are a number of approaches
- Simple Monte Carlo approach
  - simplest concept but doesn't work very well
  - draw G values of θ from $p(\theta)$,
    call them $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(G)}$
  - $\hat{p}(y) = \frac{1}{G}\sum_{g=1}^{G} p(y|\theta^{(g)})$
  - problem: prior distn may not have probability where $p(y|\theta)$ is substantial $\rightarrow$ poor estimate

Bayes Factor

Computation (cont'd)

- Alternative Monte Carlo approach
  - consider following identity (true for any pdf $h(\theta)$)

$$p(y)^{-1} = \int \frac{h(\theta)}{p(y|\theta)p(\theta)} p(\theta|y) d\theta$$

  - draw G values of θ from $p(\theta|y)$
  - $\hat{p}(y) = \left[\frac{1}{G} \sum_{g=1}^{G} \frac{h(\theta^{(g)})}{p(y|\theta^{(g)})p(\theta^{(g)})} \right]^{-1}$
  - $h(\theta)$ could be prior distribution or normal approx to the posterior distn
  - problem: not a stable calculation because of the possibility of small numbers in the denom
- Chib's marginal likelihood method
  - note that $p(y) = p(y|\theta)p(\theta)/p(\theta|y)$
  - idea: evaluate above at one value of θ, say the posterior mean or the posterior mode
  - numerator terms are easy
  - need to estimate denominator at chosen θ (crude density estimation approach or Chib's MCMC approach)

Bayes Factor

Improper prior distributions

- Consider $y|\theta \sim N(\theta, 1)$ with $p(\theta) \propto 1$

$$p(y) \propto \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\theta)^2} d\theta = 1$$

- Looks OK **but** $p(y) = 1$ for $y \in (-\infty, \infty)$ is not a valid marginal distn
- Ideas:
  - approx improper prior with proper prior (Unif$(-c, c)$) but BF is sensitive to choice of c
  - partial Bayes factor: use part of the data to build a proper prior distn and then compute BF on the rest of the data, e.g., use y_1 and flat prior to define "new" prior

$$p(\theta) = N(\theta|y_1, 1)$$

    and then can define a Bayes Factor for $y_2, \ldots, y_n$
  - fractional Bayes factor

Bayes Factor

Asymptotic approximation

- If sample size n is large, then

$$\begin{aligned} \log(BF) \quad &\approx \quad log(p(y|\hat{\theta}_2, M_2)) - log(p(y|\hat{\theta}_1, M_1)) \\ & \quad - \frac{1}{2}(d_1 - d_2)log(n) \end{aligned}$$

  where

  – $\hat{\theta}_i$ = posterior mode under M_i $(i = 1, 2)$

  – d_i = dimension of the parameter space of M_i

- Equivalent to ranking models based on the BIC (Bayes information criterion)

$$\text{BIC} = -log(p(y|\hat{\theta}, M) + \frac{1}{2}d\ log(n)$$

- Common non-Bayesian criterion is AIC (Akaike information criterion)

$$\text{AIC} = -log(p(y|\hat{\theta}, M) + d$$

- Both criteria start with log-likelihood and then penalize for additional parameters

Classical ideas and Bayesian Inference

- Classical/Bayesian
  - Bayesian = classical for some problems (large samples, small number of parameters with noninformative prior distns)
  - Standard methods often correspond to a Bayesian model for some prior (will see this in discussion of hierarchical models)
  - Big differences on some issues (e.g., p-values)

Classical ideas and Bayesian Inference

- Asymptotics
  - $\hat{\theta}_{MLE}$ is asymptotic efficient and consistent
  - $\hat{\theta}_{post.mode}$ is asymptotic efficient and consistent
- Point estimation
  - optimal Bayes point estimates depend on the specification of a loss function
  - classical inference relies on MLE
  - Bayes estimators are not generally unbiased .... neither are MLEs
    (recall defn of unbiasedness: $E(\hat{\theta}(y)|\theta) = \theta$)

Classical ideas and Bayesian Inference

- Confidence intervals
  - interpretation of Bayes and frequentist intervals
  - central posterior intervals or highest posterior density intervals
- Hypothesis testing
  - Frequentist setup:

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_a : \theta > \theta_0$$

$$\text{p-value} \quad = \quad P(\bar{Y} \text{ is unusually large} | H_0 \text{ is true})$$

    * only assessing H_0 vs data
    * p-value depends on unobserved values
    * likelihood ratio tests work for nested models only

Classical ideas and Bayesian Inference

- Hypothesis testing (cont'd)
  - Bayesian view:
    - * need a prior distn $p(\theta)$ under both hypotheses
    - * Bayes factor $BF = p(y|H_0)/p(y|H_a)$ where $p(y|H) = \int p(y|\theta, H)p(\theta|H)d\theta$
    - * more on Bayes factors later
    - * alternative for simple situation (like previous slide), just compute $\Pr(\theta > \theta_o|y)$

Classical ideas and Bayesian Inference

Hypothesis testing - an interesting example

- Discussion due to Morris (JASA 1987)
- Consider binomial sampling: $y|\theta \sim \text{Bin}(n, \theta)$

$$H_0 : \theta \leq 0.5 \qquad H_a)\theta > 0.5$$

n	y	$\hat{\theta}$	t	p-value
20	15	0.750	2.03	0.02
200	115	0.575	2.05	0.02
2000	1064	0.523	2.03	0.02

- Simple Bayesian analysis
  - model: $\hat{\theta} \sim N(\theta, 0.25/n)$ (normal approximation to binomial)
  - prior: $\theta \sim N(0.5, (0.5)^2)$

$$p(\theta > 0.5|y) = \begin{cases} 0.796 & (n = 20) \\ 0.953 & (n = 200) \\ 0.976 & (n = 2000) \end{cases}$$

Classical ideas and Bayesian Inference

- Multiple comparisons
  - e.g., effect of performing many hypothesis tests
  - tempting to say that Bayesian's don't care about multiple comparisons but there is a price to modeling many parameters
- Stopping rules/data collections
  - recall binomial/neg.binomial example
  - more on this towards the end of semester
- Nonparametrics
  - many nonparametric tests/procedures have been developed
  - Bayesian non-parametrics is complex