Final Exam Practice

Pre-midterm material

- Likelihood
  - Identify proportionality constants that can be excluded
- Cromwell's Rule
- Sufficient Statistics
- Data Generating Process
- Bias, Variance, Mean Squared Error
- Mixture Model
- Conjugate Prior
  - Pseudo-counts interpretations of conjugate priors
- Improper Priors
- Posterior Predictive Distribution
  - Integral definition involving likelihood and posterior (or prior)
- Posterior Predictive Model Checking
  - Monte Carlo approach
- Law of the unconcsious statistician (LOTUS)
- Monte Carlo error
  - How does the the Monte Carlo error change w/ number of samples?
- Inversion Sampling
- Rejection Sampling

Post-midterm material

- The normal distribution
  - Basic properties of the normal distribution
- Bayesian inference for μ when σ^2 known
  - Conjugate prior for μ is also normal
  - Posterior distribution under the conjugate prior
  - Interpretation of the prior and posterior parameters, pseudocounts
  - Add relevant formulas to cheat sheet!
- Bias-Variance tradeoff of Bayes estimators
  - Bayes estimators add bias but reduce variance (why?).
  - How to compute frequentist bias and variance of these estimators
- Bayesian inference for μ and σ^2 (both unknown)
- Sampling from the joint posterior distribution
- Markov Chains
  - Definition of a Markov Chain
  - Limiting distribution
  - Why/how they are useful in Monte Carlo sampling
- Metropolis-Hastings Algorithm
  - How to determine whether the sample should be accepted
  - Intuition of the Metropolis algorithm
  - Computational considerations
- Hastings correction (allows for non-symmetric proposals)
- MCMC convergence Diagnostics

  - Run multiple chains, different initializations
  - ACF
  - Traceplot
  - Rejection rate
  - Effective sample size
  - How the size of the "jump" proposal affects the sampler
  - Divergences (HMC only)
- Hierarchical / multilevel models
  - Complete pooling vs no pooling
  - Partial pooling
  - Relation to signal and noise variance

Practice Problems

Problem 1

Professor Franks has a dog named Wally. Every night before Wally goes to bed, he buries his favorite bone somewhere in the yard. Unfortunately, Wally is forgetful. Every day when he wakes up, he would randomly dig holes until he found his bone. Let's try to estimate how good Wally is at finding his bone!

Figure 1: Wally

Let Y be the the number of failed attempts at finding the bone. That is, Wally finds the bone on attempt $Y+1$. We will model this variable as a geometric random variable:

$$p(Y = y|\theta) = (1-\theta)^y \theta$$

where $0 \leq \theta \leq 1$ is a probability of successfully finding the bone and $Y \geq 0$ is an integer.

1a. Among the distributions discussed in this class (or on your cheat sheet), what is the conjugate prior, $p(\theta)$, for the geometric distribution? Show that your answer is correct by demonstrating that the posterior distribution is in the same family as the prior distribution. You don't need to worry about proportionality constants. Explicitly state the parameters of the posterior distribution in terms of the data and the prior parameters.

1b. One morning, before Wally started digging, Prof Franks' neighbor bet him that Wally would require at least 3 attempts to find the bone. Professor Franks knows Wally well and believed that his neighbor underestimated Wally's bone-finding ability. Find prior parameters that reflect the expectation that Wally's success rate at finding the bone any time he digs a hole is 60%. Assume the prior parameters are based on 5 prior (or pseudo) attempts.

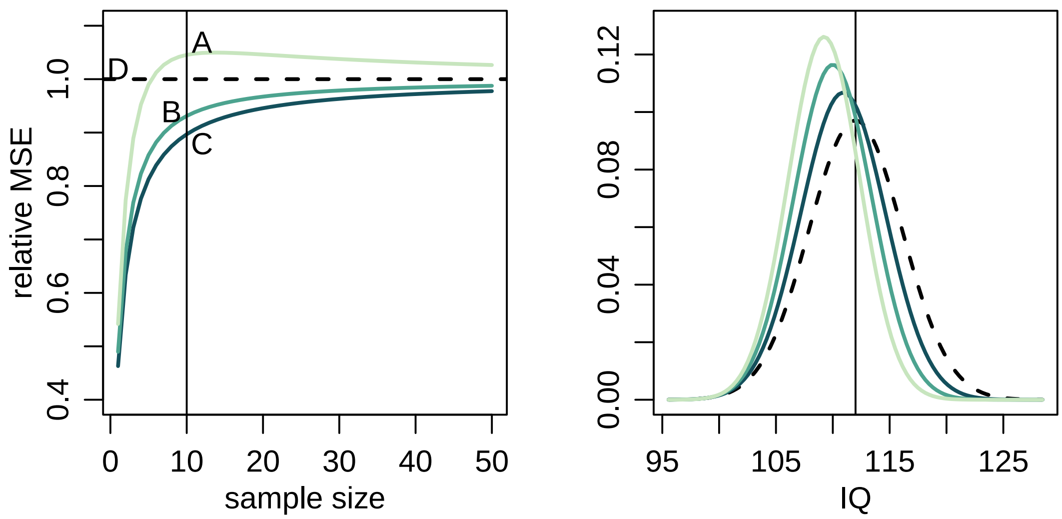
1c. Write an algorithm (pseudo-code) to compute a Monte Carlo estimate of the *prior* predictive probability of our neighbor being correct, e.g. the probability that Wally *does not* find the bone on his first two tries. Assume the prior from the previous in part. Your algorithm should include a for loop (assume 1000 samples), and list which distribution(s) you are sampling from in each step of the loop. You also need to state how you will use the Monte Carlo samples to get an estimate of the prior predictive probability that Wally digs three or more holes.

- Reminder: the prior predictive distribution is equivalent to the posterior predictive distribution except the role of the posterior is replaced with the role of the prior distribution.
- Reminder: you will need to make use of the `rgeom` function. `rgeom` is a value in $0, 1, 2, ...$, and will return the number of *failed* bone-finding attempts that Wally makes.

1d. As it turns out, Wally found his bone on the first try and Prof. Franks won the bet with his neighbor! Write the posterior distribution for θ, the bone-finding probability, given that you observed $y = 0$ and the prior parameters specified in the previous part. What is the posterior mean?

Problem 2.

First try this without referring to the lecture notes. Consider the following figure from the IQ example discussed in class. This figure is based on the following model: $p(y \mid \mu, \sigma^2) \sim N(\mu, 13^2)$ and $p(\mu) \sim N(100, \frac{13^2}{\kappa_0})$. The true unobserved value of μ is 112 and the MLE is $\bar{y} = 120$.



a. The left figure shows the mean squared error (MSE) of the posterior mean estimator relative to the maximum likelihood estimator. Fill in the blanks with the number 0, 1, 2, or 3. For line A $\kappa_0 =$ ______, for line B $\kappa_0 =$ ______, for line C, $\kappa_0 =$ ______, and for line D $\kappa_0 =$ ______.

b. Circle one. The right figure depicts:

   i. The posterior distribution for μ for each value of κ_0.
   ii. The sampling distribution of the Bayes estimator, $\hat{\mu}$ for each value of κ_0.
   iii. The likelihood of μ for each value of κ_0.
   iv. The prior distribution of μ for each value of κ_0.

Problem 3

We observe a sample of 10 observations from a normal distribution with mean μ and variance σ^2. The data, $y_1, ..., y_{10}$, are i.i.d $N(\mu, \sigma^2)$.

a). Suppose we know that the value of $\sigma^2 = 150$ and we as our prior for μ we choose $p(\mu) \sim N(20, \sigma^2/\kappa_0)$ with $\kappa = 2.5$. Find $p(\mu \mid y_1, ...y_{10})$. What is the 95% HPD interval for μ?

b). What is the posterior mean *estimate* for the observed data?

c). Now consider the posterior mean as an *estimator* by ignoring the observed values $y_1, ...y_{10}$ and treat $Y_1, ...Y_{10}$ as random variables. What is the bias, variance and mse of the posterior mean, $E[\mu \mid Y_1, ...Y_{10}]$?

d). How close must the true μ be to the prior μ_0 for the posterior mean estimator have equal MSE to the the maximum likelihood estimator, $\bar{Y}$?

Problem 4.

Draw a picture of a traceplot of a Markov Chain with high / low rejection rate.

Multiple Choice Practice

1. Monte Carlo sampling is an algorithm for...
   (a) reducing the bias of an estimator.
   (b) approximating integrals computationally.
   (c) reducing the rejection rate of the rejection sampler.
   (d) minimzing the Bayes risk

2. A sequence of random events, indexed in time, is called a *Markov Chain* if (circle one)
   (a) the distribution of the next state depends only on the *most recent* state
   (b) the distribution of the next state depends on the full history of states
   (c) the sequence has a limiting distribution
   (d) the sequence converges to the posterior distribution, $p(\theta \mid y)$

3. Let $y_1, ...y_n \sim N(\mu, \sigma^2)$ with σ^2 known. You specify the conjugate prior $\mu \sim N(\mu_0, \frac{\sigma^2}{\kappa_0})$. Assume $\kappa_0 > 0$ and that $\mu_0 \neq \mu$. Select all answers that *must* be true about estimators of μ.

4. (a) The posterior mean estimator is biased
   (b) The maximum likelihood estimator is biased
   (c) The posterior mean has lower MSE than the MLE
   (d) The posterior mean has lower variance than the MLE
   (e) The posterior variance is less than $\frac{\sigma^2}{n}$

5. Using the setup from the previous problem, write the posterior mean, $\hat{\mu}_{pm}$ as a weighted average of the MLE and the prior mean. What are the weights, w?

6. Using the setup from the previous problem, compute the bias, variance and mean squared error (MSE) of $\hat{\mu}_{pm}$ the posterior mean, treating $Y_1, ...Y_n$ are random variables. *Type your answer here, replacing this text.*

7. An improper prior distribution (select all that are true):
   (a) can't be used for valid Bayesian inference
   (b) can only be used if the posterior distribution is integrable
   (c) is another name for the uniform prior distribution
   (d) integrates to infinity

8. True or false: in the context of hierarchical models, if there are some true population differences between groups, then the complete pooling estimator will always be worse (in terms of mean squared error) than the no pooling estimator.

9. In the Metropolis Algorithm… (circle all that are true)

   (a) The proposal distribution must be symmetric
   (b) A proposed sample is always accepted if it would increase the posterior density
   (c) It's best to have high autocorrelation
   (d) The most efficient samplers have a rejection rate that is close to 0

10. Consider estimates of mean parameter, θ_i, across multiple groups of observations. When considering the variability of the resulting estimates θ_i between groups, order the following estimates from least to most variability between groups: complete pooling, no pooling and partial pooling. estimates.

11. Name two MCMC diagnostics that can be used to assess the quality of estimates derived from a sampler.

Quiz 4 questions:

1. In the Metropolis-Hastings Algorithm… (circle all that are true)
   - ◯ When a proposed sample is rejected, the previous sample is used for another iteration
   - ◯ A proposed sample is always accepted if it would increase the posterior density
   - ◯ It's best to have high autocorrelation
   - ◯ The most efficient samplers have a rejection rate that is close to 0

2. Assume $Y \sim N(\theta, \sigma^2/n)$ and $\theta \sim N(0, \sigma^2/\kappa)$. The posterior mean *estimator* of θ (treating Y as a random variable) has bias:
   - ◯ 0
   - ◯ θ
   - ◯ $-\frac{k}{(n+k)}\theta$
   - ◯ $\frac{\sigma^2}{(n+k)}$

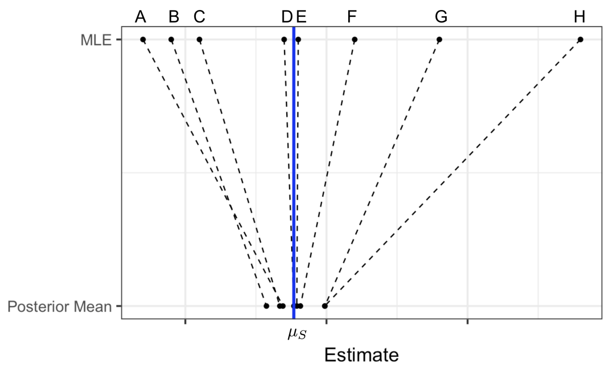
3. The Rhat diagnostic is a measure of:
   - ◯ the autocorrelation of a Markov Chain
   - ◯ how well multiple chains have mixed
   - ◯ how many chains are needed to do inference
   - ◯ the bias of the MCMC samples

4. Consider estimates of mean parameter, θ_i, across multiple groups of observations. When considering the variability of the resulting estimates θ_i between groups, order the following estimates from least to most variability between groups: complete pooling, no pooling and partial pooling. estimates.

Quiz 5 questions: A UCSB psychologist is studying reaction times to a particular stimulus in a group of 8 individuals, labeled A through H. The researcher wants to know the true average reaction time of patient, θ_i, based on observed reaction times Y_i measured in milliseconds (ms). The psychologist proposes the following hierarchical model:

$$(\mu, \tau^2) \sim p(\mu, \tau^2)$$
$$\theta_i \sim N(\mu, \tau^2)$$
$$Y_i \sim N(\theta_i, \sigma_i^2)$$

where $p(\mu, \tau^2)$ is some (unspecified) prior. The researcher fits the hierarchical model and compares the posterior mean reaction times, $\theta_A ..., \theta_H$ for subjects. The next questions relate to the following figure. The vertical blue line is at $\mu_S = E[\mu | y_1, ... y_n]$:



a. Choose one: For smaller values of τ^2 the posterior mean estimates of $\theta_A ..., \theta_H$ will be closer to
   - ◯ $\theta_i^{(MLE)}$
   - ◯ μ

b. Choose one. For smaller values of σ_A^2 the posterior mean estimate of θ_A will be closer to
   - ◯ $\hat{\theta}_A^{(MLE)}$
   - ◯ μ

c. Choose one. As τ^2 decreases the posterior variance of θ_i will
   - ◯ increase
   - ◯ decrease.

d. Note that $Y_A < Y_B$ yet after fitting the hierarchical model $\theta_A^{(PM)} > \theta_B^{(PM)}$. What would best explain this fact? Choose one.
   - ◯ $\sigma_A^2 > \sigma_B^2$
   - ◯ $\sigma_A^2 < \sigma_B^2$
   - ◯ $\sigma_A^2 > \tau^2$
   - ◯ $\sigma_B^2 > \tau^2$