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 variance of our Monte Carlo
- Inversion Sampling

Post-midterm material

- The normal distribution
  - Basic properties of the normal distribution
  - Central limit theorem
- 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?).
- Decision Theory
  - Bayes estimator,s Bayes risk and loss function definitions
  - Bayes estimator for minimizing the Bayes risk under squared error loss
  - Bayes estimator for minimizing the Bayes risk under absolute error loss
- 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)
- Gibbs sampling
  - Basic idea of Gibbs sampling
  - Never reject new proposed samples
- MCMC convergence Diagnostics
  - Run multiple chains, different initalizations
  - ACF
  - Traceplot
  - Rejection rate
  - Effective sample size
  - How the size of the "jump" proposal affects the sampler
- 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!

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.

Beta distribution is the conjugate prior of geometric distribution

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

$$Beta(\alpha, \beta) : p(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$Posterior \propto (1-\theta)^y\theta \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{(\alpha+1)-1}(1-\theta)^{(\beta+y)-1}$$

$$\propto Beta(\alpha+1, \beta+y)$$

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.

$$Prior : Beta(3, 2)$$

1c. Write an algorithm 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.

```
prior_predictive_samples <- numeric(1000)
for(i in 1:1000) {

  # BEGIN SOLUTION
  theta_value <- rbeta(1, 3, 2)
  # END SOLUTION

  prior_predictive_samples[i] <- (rgeom(1, theta_value) + 1) # SOLUTION

}

prob_three <- mean(prior_predictive_samples > 2) # SOLUTION
print(prob_three)
```

```
## [1] 0.212
```

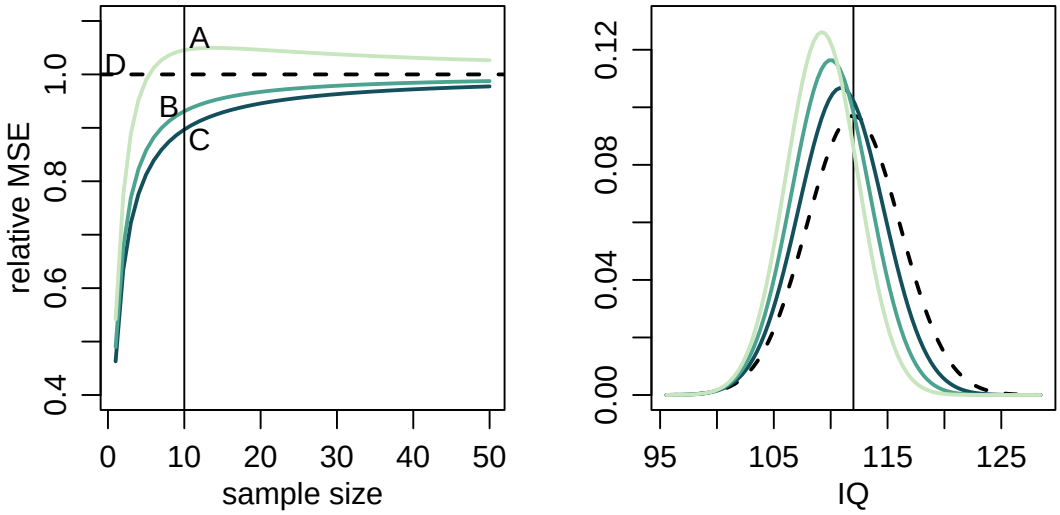
1d. As it turns out, Wally found his bone on the first try and Prof. Franks won the bet with his neighbor! Write the posterior distribution for θ, the bone-finding probability, given that you observed $y = 0$ and the prior parameters specified in the previous part. What is the posterior mean?

$$Posterior = Beta(\alpha+1, \beta+0) = Beta(4, 2)$$

The bone finding probability is 2/3. This means based the posterior probabiolity, the probability of Wally found his bone is higher than our prior assumption.

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 $\kappa_0 =$ _____.

(a)

- A $\kappa_0 = 3$
- B $\kappa_0 = 2$
- C $\kappa_0 = 1$
- D $\kappa_0 = 0$

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.

(b) Circle ii. This is a hard one. The plot shows $p(\hat{\mu}_{PM} \mid \mu)$, i.e. considering $\hat{\mu}_{PM} = w\bar{Y} + (1-w)\mu_0$ (notice capital Y not lowercase). When w = 1, *i.e.*$\kappa_0 = 0$, then $E[\mu_{PM}] = E[w\bar{Y}] = \mu = 112$ (the true population had an IQ of 112). This is why the dashed line is centered at 112. The posterior mean estimator is biased when $\kappa > 0$ but reduced variance. Note that i. is not correct, since $p(\mu \mid y)$ should be centered at $y = 120$ (the observed MLE) for $\kappa_0 = 0$ and shrink toward 100 for larger κ.

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?

From the lecture notes, we know that

$$p(\mu \mid y, \sigma^2) \sim N(\mu_n, \tau_n^2),$$

where $\mu_n = w\bar{y} + (1-w)\mu_0, \quad w = \frac{n/\sigma^2}{n/\sigma^2+1/\tau^2}$ and $\tau_n^2 = \frac{1}{n/\sigma^2+1/\tau^2}$, and $\tau^2 = \frac{\sigma^2}{\kappa_0}$. As given by the question, $n = 10$, $\sigma^2 = 150$, $\mu_0 = 20$, $\kappa_0 = 2.5$. Thus $w = 0.8$.

$$\mu_n = w\bar{y} + (1-w)\mu_0 = 0.8\bar{y} + 0.2 * 20 = 0.8\bar{y} + 4$$

$$\tau_n^2 = \frac{1}{n/\sigma^2 + 1/\tau^2} = \frac{\sigma^2}{\kappa_0 + n} = 12$$

95% HPD interval:

$$\mu_n \pm 1.96 * \tau_n = (0.8\bar{y} + 4) \pm 1.96 * \sqrt{12} = (0.8\bar{y} - 2.79,\ 0.8\bar{y} + 10.79)$$

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

Posterior mean estimate: $\mu_n = 0.8\bar{y} + 4$

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, $E[\mu \mid Y_1, ...Y_{10}]$?

Bias: $E(\mu - (0.8\bar{Y} + 4)) = \mu - 0.8\mu - 4 = 0.2\mu - 4$

Variance: $Var(0.8\bar{Y} + 4) = w^2 * Var(\bar{Y}) = 0.64 * \frac{\sigma^2}{n} = 15 * 0.64 = 9.6$

MSE: MSE = bias^2 + Var = $(0.2 * \mu - 4)^2 + 9.6$

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, $\bar{Y}$?

Let

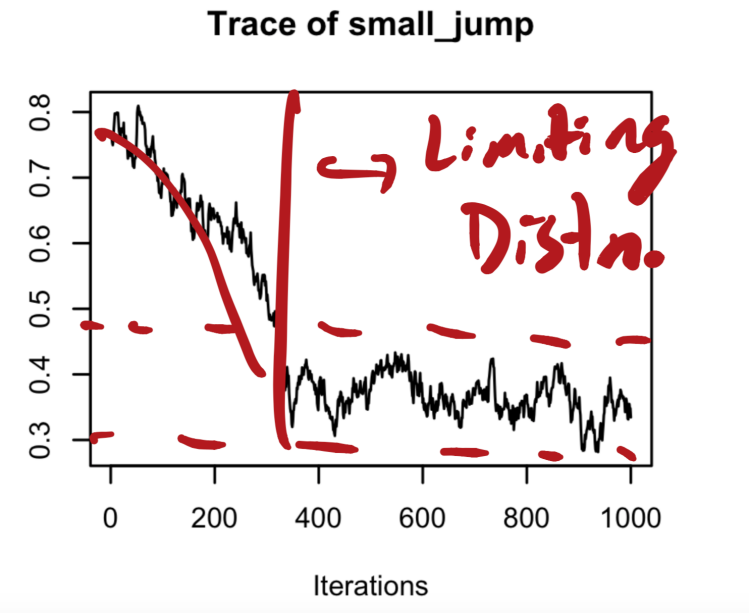
$$(0.2 * \mu - 4)^2 + 9.6 = Var(\bar{Y}) = 150/10 = 15,$$

and solve for μ. We see that there are two solutions $\mu \approx 8.4$ and $\mu \approx 31.62$ the posterior mean estimator has equal MSE to the the maximum likelihood estimator. That is, for any $\mu \in [8.4, 31.62]$ this posterior mean estimator outperforms the MLE.

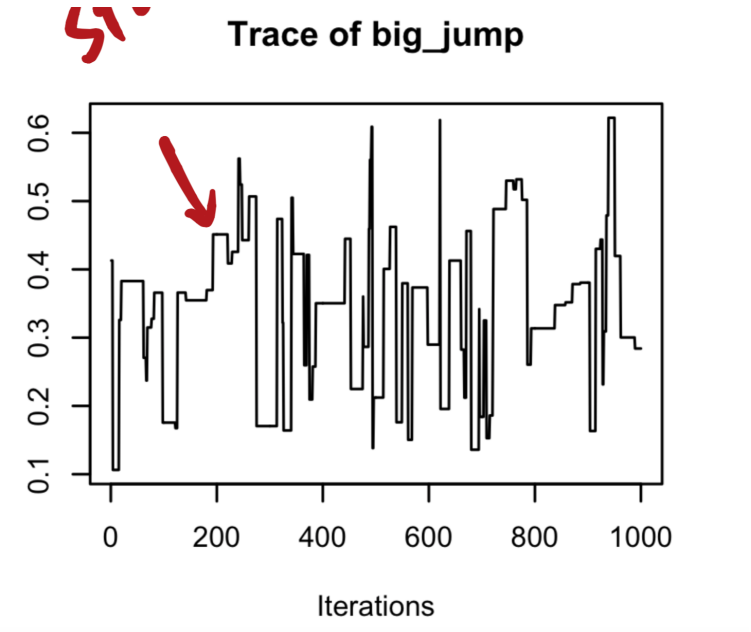
Problem 4.

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

Low Rejection



High Rejection



Problem 5.

Consider Bayesian inference for p(μ, $\frac{1}{\sigma^2}$ | y), where an $y \sim N(\mu, \sigma^2)$. Assume the non-infromative prior disribution $p(\mu, \sigma^2) \propto 1/\sigma^2$. This will lead to the diamond shaped posterior discussed in class and on the lecture notes. Argue in words (no math needed) why the diamond shaped posterior makes sense? For what values of $\frac{1}{\sigma^2}$ does μ have the most posterior variability? Lead posterior variability? Why?

- If $1/\sigma^2$ is very large, y is likely to be close to μ, meaning that the observed data is very informative about μ and we can estimate μ with more certainty.

- If $1/\sigma^2$ is very small, y can be far away from μ. Based on the observed data, we cannot estimate μ precisely. The estimation has very large uncertainty.

Multiple Choice Practice

1. The Bayes estimator (estimator which minimizes the posterior expected loss) for squared error loss is:
   (a) The posterior mean
   (b) The posterior median
   (c) The posterior mode
   (d) The posterior variance

The posterior mean

2. 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

approximating integrals computationally.

3. 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, $p(\theta \mid y)$

a. the distribution of the next state depends only on the *most recent* state

4. 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.

4a. 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 $\frac{\sigma^2}{n}$

Solution: a. The posterior mean estimator is biased, d, The posterior mean has lower variance than the MLE, e. The posterior variance is less than $\frac{\sigma^2}{n}$

4b. Write the posterior mean, $\hat{\mu}_{pm}$ as a weighted average of the MLE and the prior mean. What are the weights, w?

$\mu_{pm} = (1-w)\mu_0 + w\bar{y}$ where $w = \frac{\frac{n}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}$

4c. Compute the bias, variance and mean squared error (MSE) of $\hat{\mu}_{pm}$ the posterior mean, treating $Y_1, ...Y_n$ are random variables. Bias: $(1-w)(\mu_0 - \mu)$ MSE: $w^2\frac{\sigma^2}{n} + Bias^2$

5. 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

a. can only be used if the posterior distribution is integrable and b. the prior integrates to infinity

6. 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.

False

7. 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

a. The proposal distribution must be symmetric b. A proposed sample is always accepted if it would increase the posterior density

8. 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.

complete pooling, partial pooling, no pooling

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

Effective sample size, rejection rate and autocorrelation.